An efficient sieving based secant method for sparse optimization problems with least-squares constraints

Defeng Sun

Department of Applied Mathematics

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

East China Normal University, Shanghai; April 26, 2024

Joint work with Qian Li (PolyU), Yancheng Yuan (PolyU)

Outline

Least-squares constrained optimization problem

Level-set : Properties of the value function $\varphi(\cdot)$

The HS-Jacobian of $\varphi(\cdot)$ for polyhedral gauge functions $p(\cdot)$

The convergence properties of the secant method

Adaptive sieving

Numerical experiments

Conclusion

Outline

Least-squares constrained optimization problem

Level-set : Properties of the value function $\varphi(\cdot)$

The HS-Jacobian of $\varphi(\cdot)$ for polyhedral gauge functions $p(\cdot)$

The convergence properties of the secant method

Adaptive sieving

Numerical experiments

Conclusion

Least-squares constrained optimization problem

We consider the following least-squares constrained optimization problem

$$\min_{x \in \mathbb{R}^n} \ \{p(x) \mid \|Ax - b\| \leq \varrho\}, \qquad (\mathrm{CP}(\varrho))$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are given data, ϱ (noise level) is a given parameter satisfying $0 < \varrho < \|b\|$, and $p : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper closed convex function with $p(0) = 0$.

We assume that (CP(ϱ)) admits an active solution.

Examples :

- The ℓ_1 penalty : $p(x) = \|x\|_1$, $x \in \mathbb{R}^n$.
- The sorted ℓ_1 penalty : $p(x) = \sum_{i=1}^n \gamma_i |x|_{(i)}$, $x \in \mathbb{R}^n$ with given parameters $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n \geq 0$ and $\gamma_1 > 0$, where $|x|_{(1)} \geq |x|_{(2)} \geq \cdots \geq |x|_{(n)}$.
- The fused lasso penalty, ...

The level set methods

- Method 1 [Van den Berg-Friedlander 2008, 2011] solves (CP(ϱ)) by finding a root of the following univariate nonlinear equation

$$\phi(\tau) = \varrho, \tag{E_ϕ}$$

  where $\phi(\cdot)$ is the value function of the following level-set problem

$$\phi(\tau) := \min_{x \in \mathbb{R}^n} \{\|Ax - b\| \mid p(x) \leq \tau\}, \quad \tau \geq 0. \tag{1}$$

  Feasibility issue with a dimension reduction technique applied to (1) **?**

- Method 2 [Li-Sun-Toh 2018] solves (CP(ϱ)) by finding a root of the following equation :

$$\varphi(\lambda) := \|Ax(\lambda) - b\| = \varrho, \tag{E_φ}$$

  where $x(\lambda) \in \Omega(\lambda)$ is any solution to

$$\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2}\|Ax - b\|^2 + \lambda p(x) \right\}, \quad \lambda > 0. \tag{$\mathrm{P_{LS}}(\lambda)$}$$

The secant method

Let $f : \mathbb{R} \to \mathbb{R}$ be a locally Lipschitz continuous function which is semismooth at a solution x^* to the equation $f(x) = 0$.

The secant method :

Step 1. Given $x^0, x^{-1} \in \mathbb{R}$. Let $k = 0$.

Step 2. Let

$$x^{k+1} = x^k - \left(\frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} \right)^{-1} f(x^k).$$

Step 3. $k := k + 1$. Go to Step 2.

- If f is smooth, the secant method is superlinearly convergent with Q-order at least $(1 + \sqrt{5})/2$ [Traub 1964] .
- If f is (strongly) semismooth, then the secant method is 3-step Q-superlinearly (Q-quadratically) convergent [Potra-Qi-Sun 1998].

Outline

Properties of the value function $\varphi(\cdot)$

The dual of $(\mathrm{P}_{\mathrm{LS}}(\lambda))$ can be written as

$$\max_{y \in \mathbb{R}^m, u \in \mathbb{R}^n} \left\{ -\frac{1}{2}\|y\|^2 + \langle b, y \rangle - \lambda p^*(u) \,|\, A^T y - \lambda u = 0 \right\}. \tag{$\mathrm{D}_{\mathrm{LS}}(\lambda)$}$$

We assume

$$\lambda_\infty := \Upsilon(A^T b \,|\, \partial p(0)) > 0 \tag{2}$$

and that for any $\lambda > 0$, there exists $(y(\lambda), u(\lambda), x(\lambda)) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n$ satisfying the following Karush–Kuhn–Tucker (KKT) system

$$x \in \partial p^*(u), \quad y = b - Ax, \quad A^T y - \lambda u = 0. \tag{KKT}$$

Proposition

Assume that $\lambda_\infty > 0$. It holds that

- for all $\lambda \geq \lambda_\infty$, $y(\lambda) = b$ and $0 \in \Omega(\lambda)$;
- the value function $\varphi(\cdot)$ is nondecreasing on $(0, +\infty)$ and for any $\lambda_1 > \lambda_2 > 0$, $\varphi(\lambda_1) = \varphi(\lambda_2)$ implies $p(x(\lambda_1)) = p(x(\lambda_2))$, where for any $\lambda > 0$, $x(\lambda)$ is an optimal solution to $(\mathrm{P}_{\mathrm{LS}}(\lambda))$.

Properties of $\varphi(\cdot)$ when p is a gauge function

When $p(\cdot)$ is a gauge function, $p^*(\cdot) = \delta(\cdot \,|\, \partial p(0))$ and the optimization problem $(\mathrm{D_{LS}}(\lambda))$ is equivalent to

$$\max_{y\in\mathbb{R}^m} \left\{ -\frac{1}{2}\|y\|^2 + \langle b, y\rangle \mid \lambda^{-1} y \in Q \right\}, \quad Q := \{z \in \mathbb{R}^m \mid A^T z \in \partial p(0)\}. \tag{3}$$

The unique solution to (3) is

$$y = -\lambda \Pi_Q(\lambda^{-1} b).$$

Proposition

Let $p(\cdot)$ be a gauge function. Assume that $\lambda_\infty > 0$. It holds that

- (i) the functions $y(\cdot)$ and $\varphi(\cdot)$ are locally Lipschitz continuous on $(0, +\infty)$;
- (ii) the function $\varphi(\cdot)$ is strictly increasing on $(0, \lambda_\infty]$;
- (iii) if the set Q is tame, then $\varphi(\cdot)$ is semismooth on $(0, +\infty)$;
- (iv) if Q is globally subanalytic, then $\varphi(\cdot)$ is γ-order semismooth on $(0, +\infty)$ for some $\gamma > 0$.

- Let $p(\cdot) = \|\cdot\|_*$ be the nuclear norm function defined on $\mathbb{R}^{d\times n}$. Then $Q = \{z \in \mathbb{R}^m \,|\, \mathcal{A}^* z \in \partial p(0)\}$ is a tame set and $\Pi_Q(\cdot)$ is semismooth.

Properties of $\varphi(\cdot)$ when p is a gauge function Cont.

Proposition

Let $p(\cdot)$ be a gauge function. Define $\Phi(x) := \frac{1}{2}\|Ax - b\|^2$, $x \in \mathbb{R}^n$ and

$$H(x, \lambda) := x - \mathrm{Prox}_p(x - \lambda^{-1}\nabla\Phi(x)), \quad (x, \lambda) \in \mathbb{R}^n \times \mathbb{R}_{++}.$$

For any $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}_{++}$, denote $\partial_x H(x, \lambda)$ as the Canonical projection of $\partial H(x, \lambda)$ onto $\mathbb{R}^n$. It holds that

- if $\Pi_{\partial p(0)}(\cdot)$ is strongly semismooth and $\partial_x H(\bar{x}, \bar{\lambda})$ is nondegenerate at some $(\bar{x}, \bar{\lambda})$ satisfying $H(\bar{x}, \bar{\lambda}) = 0$, then $y(\cdot)$ and $\varphi(\cdot)$ are strongly semismooth at $\bar{\lambda}$;
- if $p(\cdot)$ is further assumed to be polyhedral, the function $y(\cdot)$ is piecewise affine and $\varphi(\cdot)$ is strongly semismooth on $\mathbb{R}_{++}$.

Outline

The HS-Jacobian of $\varphi(\cdot)$

- Assume that $p(\cdot)$ is a polyhedral gauge function. Then the set $\partial p(0)$ is polyhedral, which can be assumed to take the form of

$$\partial p(0) := \{u \in \mathbb{R}^n \mid Bu \leq d\} \tag{4}$$

for some $B \in \mathbb{R}^{q \times n}$ and $d \in \mathbb{R}^q$.

- We will derive the HS-Jacobian [Han-Sun 1997] of the function $\varphi(\cdot)$ to prove that the Clarke Jacobian of $\varphi(\cdot)$ at any $\lambda \in (0, \lambda_\infty)$ is positive.

- Let $\lambda \in (0, \lambda_\infty)$ be arbitrarily chosen. Let $(y(\lambda), u(\lambda))$ be the unique solution to

$$\max_{y \in \mathbb{R}^m, u \in \mathbb{R}^n} \left\{ -\frac{1}{2}\|y\|^2 + \langle b, y \rangle - \lambda p^*(u) \mid A^T y - \lambda u = 0 \right\} \tag{$\mathrm{D_{LS}}(\lambda)$}$$

with the parameter λ. We denote $(y, u) = (y(\lambda), u(\lambda))$ to simplify our notation.

The HS-Jacobian of $\varphi(\cdot)$ Cont.

- There exists $x \in \Omega(\lambda)$ such that (y, u, x) satisfies the following KKT system :

$$u = \Pi_{\partial p(0)}(u + x), \quad y - b + Ax = 0, \quad A^T y - \lambda u = 0. \tag{5}$$

$$u = \Pi_{\partial p(0)}(u + x) \Leftrightarrow u = \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2}\|z - (u + x)\|^2 \mid Bz \leq d \right\}. \tag{6}$$

- The augmented KKT system :

$$\begin{cases} B^T \xi - x = 0, \quad Bu - d \leq 0, \quad \xi \geq 0, \quad \xi^T (Bu - d) = 0, \\ y - b + Ax = 0, \quad A^T y - \lambda u = 0. \end{cases} \tag{7}$$

Let $M(\lambda)$ be the set of Lagrange multipliers associated with (y, u) defined as

$$M(\lambda) := \left\{ (x, \xi) \in \mathbb{R}^n \times \mathbb{R}^l \mid (y, u, x, \xi) \text{ satisfies (7)} \right\}.$$

The HS-Jacobian of $\varphi(\cdot)$ Cont.

- Since $x = B^T\xi$, we obtain the following system by eliminating the variable x in (7) :

$$\begin{cases} Bu - d \leq 0, \quad \xi \geq 0, \quad \xi^T(Bu - d) = 0, \\ y - b + \widehat{A}\xi = 0, \quad A^T y - \lambda u = 0, \end{cases} \tag{8}$$

  where $\widehat{A} = AB^T \in \mathbb{R}^{m\times q}$. Denote

$$\widehat{M}(\lambda) := \{\xi \in \mathbb{R}^q \mid (y, u, \xi) \text{ satisfies (8)}\}. \tag{9}$$

- Denote the active set of u as

$$I(u) := \{i \in l \mid B_{i:}u - d_i = 0\}. \tag{10}$$

  For any $\lambda \in (0, \lambda_\infty)$, we define

$$\mathcal{B}(\lambda) := \left\{K \subseteq [q] \mid \exists\ \xi \in \widehat{M}(\lambda) \text{ s.t. } \mathrm{supp}(\xi) \subseteq K \subseteq I(u) \text{ and } \mathrm{rank}(\widehat{A}_{:K}) = |K|\right\}. \tag{11}$$

The HS-Jacobian of $\varphi(\cdot)$ Cont.

- Since the polyhedral set $\widehat{M}(\lambda)$ does not contain a line, this implies that $\widehat{M}(\lambda)$ has at least one extreme point. Note that $0 < \lambda < \lambda_\infty$ and $x \neq 0$, which implies that $\bar{\xi} \neq 0$ and $\mathcal{B}(\lambda)$ is nonempty.

- Define the HS-Jacobian of $y(\cdot)$ as

$$\mathcal{H}(\lambda) := \left\{ h^K \in \mathbb{R}^m \,|\, h^K = \widehat{A}_{:K}(\widehat{A}_{:K}^T \widehat{A}_{:K})^{-1} d_K,\ K \in \mathcal{B}(\lambda) \right\}, \quad \lambda \in (0, \lambda_\infty), \tag{12}$$

where d_K is the subvector of d indexed by K. For notational convenience, for any $\lambda \in (0, \lambda_\infty)$ and $K \in \mathcal{B}(\lambda)$, denote

$$P^K = I - \widehat{A}_{:K}(\widehat{A}_{:K}^T \widehat{A}_{:K})^{-1} \widehat{A}_{:K}^T. \tag{13}$$

Define

$$\mathcal{V}(\lambda) := \left\{ t \in \mathbb{R} \,|\, t = \lambda \|h\|^2 / \varphi(\lambda),\ h \in \mathcal{H}(\lambda) \right\}, \quad \lambda \in \mathcal{D}, \tag{14}$$

where $\mathcal{D} = \{\lambda \in (0, \lambda_\infty) \,|\, \varphi(\lambda) > 0\}$.

Nondegeneracy of $\partial\varphi(\bar{\lambda})$ for any $\bar{\lambda} \in (0, \lambda_\infty)$

Lemma

Let $\bar{\lambda} \in (0, \lambda_\infty)$ be arbitrarily chosen. It holds that

$$y(\bar{\lambda}) = P^K b + \bar{\lambda} h^K, \quad \forall\, h^K \in \mathcal{H}(\bar{\lambda}). \tag{15}$$

Moreover, there exists a positive scalar ς such that $\mathcal{N}(\bar{\lambda}) := (\bar{\lambda} - \varsigma, \bar{\lambda} + \varsigma) \subseteq (0, \lambda_\infty)$ and for all $\lambda \in \mathcal{N}(\bar{\lambda})$,

- $\mathcal{B}(\lambda) \subseteq \mathcal{B}(\bar{\lambda})$ and $\mathcal{H}(\lambda) \subseteq \mathcal{H}(\bar{\lambda})$;
- $y(\lambda) = y(\bar{\lambda}) + (\lambda - \bar{\lambda})h, \quad \forall\, h \in \mathcal{H}(\lambda)$.

Theorem

For any $\bar{\lambda} \in (0, \lambda_\infty)$, it holds that

- for any positive integer $k \geq 1$, the function $\varphi(\cdot)$ is piecewise C^k in an open interval containing $\bar{\lambda}$;
- all $v \in \partial\varphi(\bar{\lambda})$ are positive.

Nondegeneracy of HS-Jacobian of $\varphi(\cdot)$ for polyhedral gauge functions

Proposition

Suppose that $p(\cdot)$ is a polyhedral gauge function and $\partial p(0)$ has the expression as in (4). Let $\bar{\lambda} \in (0, \lambda_\infty)$ be arbitrarily chosen. Let $\mathcal{B}(\bar{\lambda})$ and $\mathcal{V}(\bar{\lambda})$ be the sets defined as in (11) and (14) for $\lambda = \bar{\lambda}$. If $d_K \neq 0$ for all $K \in \mathcal{B}(\bar{\lambda})$, then $v > 0$ for all $v \in \mathcal{V}(\bar{\lambda})$. Moreover, $d_K \neq 0$ for all $K \in \mathcal{B}(\bar{\lambda})$ when $p(\cdot) = \|\cdot\|_1$.

- This proposition shows that for the least-squares constrained Lasso problem, $\partial_{\mathrm{HS}}\varphi(\bar{\lambda})$ is positive for any $\bar{\lambda} \in (0, \lambda_\infty)$.

Outline

The convergence properties of the secant method

Let $f : \mathbb{R} \to \mathbb{R}$ be a locally Lipschitz continuous function which is semismooth at a solution x^* to the following equation

$$f(x) = 0. \tag{16}$$

The secant method :

Step 1. Given $x^0, x^{-1} \in \mathbb{R}$. Let $k = 0$.

Step 2. Let

$$x^{k+1} = x^k - \left(\frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} \right)^{-1} f(x^k).$$

Step 3. $k := k + 1$. Go to step 2.

Denote

$$\bar{d}^- := -f'(\bar{x}; -1) \quad \text{and} \quad \bar{d}^+ := f'(\bar{x}; 1), \tag{17}$$

The convergence properties of the secant method

Proposition

Suppose that $f : \mathbb{R} \to \mathbb{R}$ is semismooth at a solution x^* to (16). Let d^- and d^+ be the lateral derivatives of f at x^* as defined in (17). If d^- and d^+ are both positive (or negative), then there are two neighborhoods $\mathcal{U}$ and $\mathcal{N}$ of x^*, $\mathcal{U} \subseteq \mathcal{N}$, such that for $x^{-1}, x^0 \in \mathcal{U}$, The secant method is well defined and produces a sequence of iterates $\{x^k\}$ such that $\{x^k\} \subseteq \mathcal{N}$. The sequence $\{x^k\}$ converges to x^* 3-step Q-superlinearly, i.e., $|x^{k+3} - x^*| = o(|x^k - x^*|)$. Moreover, it holds that

(i) $|x^{k+1} - x^*| \leq \frac{|d^+ - d^- + o(1)|}{\min\{|d^+|, |d^-|\} + o(1)} |x^k - x^*|$ for $k \geq 0$;

(ii) if $\alpha := \frac{|d^+ - d^-|}{\min\{|d^+|, |d^-|\}} < 1$, then $\{x^k\}$ converges to x^* Q-linearly with Q-factor α;

(iii) if f is γ-order semismooth at x^* for some $\gamma > 0$, then $|x^{k+3} - x^*| = O(|x^k - x^*|^{1+\gamma})$ for sufficiently large k; the sequence $\{x^k\}$ converges to x^* 3-step quadratically if f is strongly semismooth at x^*.

- When $|d^+ - d^-|$ is small and f is strongly semimsooth, we know from the above proposition that the secant method converges with a fast Q-linear rate and 3-step Q-quadratic rate.

A numerical example for the secant method

We test the secant method with $x^{-1} = 0.01$ and $x^{0} = 0.005$ for finding the zero $x^{*} = 0$ of

$$f(x) = \begin{cases} x(x+1) & \text{if } x < 0, \\ -\beta x(x-1) & \text{if } x \geq 0, \end{cases} \tag{18}$$

where β is chosen from $\{1.1, 1.5, 2.1\}$.

- Case I : $\beta = 1.1,\ d^{+} = 1.1,\ d^{-} = 1,$ and $\alpha = 0.1$;
- Case II : $\beta = 1.5,\ d^{+} = 1.5,\ d^{-} = 1,$ and $\alpha = 0.5$;
- Case III : $\beta = 2.1,\ d^{+} = 2.1,\ d^{-} = 1,$ and $\alpha = 1.1$.

Table – The numerical performance of finding the zero of (18).

Case	Iter	1	2	3	4	5	6	7	8
I	x	-5.1e-5	-4.3e-6	2.2e-10	-2.2e-11	-1.8e-12	4.1e-23	-4.1e-24	-3.4e-25
II	x	-5.1e-5	-1.7e-5	8.4e-10	-4.2e-10	-1.1e-10	4.5e-20	-2.2e-20	-5.6e-21
III	x	-5.1e-5	-2.6e-5	1.3e-9	-1.5e-9	-5.1e-10	7.4e-19	-8.2e-19	-2.8e-19

The convergence properties of the secant method cont.

Proposition

Let $p(\cdot)$ be a polyhedral gauge function and λ^* be the solution to (E_φ). Assume that $0 < \lambda_\infty < +\infty$. If $\partial\varphi(\lambda^*)$ is a singleton, the sequence $\{\lambda_k\}$ generated by the secant method for solving (E_φ) converges to λ^* Q-superlinearly with Q-order at least $(1+\sqrt{5})/2$.

- A strongly semismooth function is not necessarily piecewise smooth. For example

$$f(x) = \begin{cases} \kappa x, & \text{if } x < 0, \\ -\frac{1}{3}\left(\frac{1}{4^k}\right) + (1 + \frac{1}{2^k})x, & \text{if } x \in \left[\frac{1}{2^{k+1}}, \frac{1}{2^k}\right] \quad \forall k \geq 0, \\ 2x - \frac{1}{3} & \text{if } x > 1, \end{cases} \tag{19}$$

where κ is a given constant.

A numerical example for the secant method cont.

Set $\kappa = 1$. Note that $x^* = 0$ is the unique solution of (19). In the secant method, we choose $x^0 = 0.5$ and $x^{-1} = x^0 + 0.1 \times f(0.5)^2 = 0.545$. The numerical results are shown in the following table.

Table – The numerical performance of the secant method on finding the zero of (19).

Iter	1	2	3	4	5	6	7	8
x	1.7e-1	3.6e-2	4.0e-3	1.0e-4	2.7e-7	2.0e-11	4.0e-18	6.1e-29
$f(x)$	1.9e-1	3.7e-2	4.0e-3	1.0e-4	2.7e-7	2.0e-11	4.0e-18	6.1e-29

We can observe that the generated sequence $\{x_k\}$ converges to the solution $x^* = 0$ superlinearly with Q-order $(1 + \sqrt{5})/2$.

A globally convergent secant method for (CP(ϱ))

The globally convergent secant method for (CP(ϱ)) :

- Step 1. Given $\mu \in (0,1)$, $\lambda_{-1}, \lambda_0, \lambda_1$ in $(0, \lambda_\infty)$ satisfying $\varphi(\lambda_0) > \varrho$, and $\varphi(\lambda_{-1}) < \varrho$. Set $i = 0$, $\underline{\lambda} = \lambda_{-1}$, and $\overline{\lambda} = \lambda_0$. Let $k = 0$.
- Step 2. Compute

$$\hat{\lambda}_{k+1} = \lambda_k - \frac{\lambda_k - \lambda_{k-1}}{\varphi(\lambda_k) - \varphi(\lambda_{k-1})}(\varphi(\lambda_k) - \varrho). \tag{20}$$

- Step 3. **If** $\hat{\lambda}_{k+1} \in [\lambda_{-1}, \lambda_0]$, **then** continue, **else**, go to Step 4.
  1. Compute $x(\hat{\lambda}_{k+1})$ and $\varphi(\hat{\lambda}_{k+1})$. Set $i = i + 1$.
  2. **If** either (i) or (ii) holds : (i) $i \geq 3$ and $|\varphi(\hat{\lambda}_{k+1}) - \varrho| \leq \mu|\varphi(\lambda_{k-2}) - \varrho|$ (ii) $i < 3$, **then** set $\lambda_{k+1} = \hat{\lambda}_{k+1}$, $x(\lambda_{k+1}) = x(\hat{\lambda}_{k+1})$; **else** go to Step 4.
  3. Go to Step 5.
- Step 4. **If** $\varphi(\hat{\lambda}_{k+1}) > \varrho$, **then** set $\overline{\lambda} = \min\{\overline{\lambda}, \hat{\lambda}_{k+1}\}$; **else** set $\underline{\lambda} = \max\{\underline{\lambda}, \hat{\lambda}_{k+1}\}$. Set $\lambda_{k+1} = 1/2(\overline{\lambda} + \underline{\lambda})$. Compute $x(\lambda_{k+1})$ and $\varphi(\lambda_{k+1})$. Set $i = 0$.
- Step 5. **if** $\varphi(\lambda_{k+1}) > \varrho$, **then** set $\overline{\lambda} = \min\{\overline{\lambda}, \lambda_{k+1}\}$; **else** set $\underline{\lambda} = \max\{\underline{\lambda}, \lambda_{k+1}\}$.
- Step 6. $k = k + 1$. Go to Step 2.

Outline

Least-squares constrained optimization problem

Level-set : Properties of the value function $\varphi(\cdot)$

The HS-Jacobian of $\varphi(\cdot)$ for polyhedral gauge functions $p(\cdot)$

The convergence properties of the secant method

Adaptive sieving

Numerical experiments

Conclusion

The adaptive sieving technique [Yuan-Lin-Sun-Toh 2023]

Consider the problem

$$\min_{x\in\mathbb{R}^n} \{\Phi(x) + P(x)\}, \tag{21}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable convex function, and $P : \mathbb{R}^n \to (-\infty, +\infty]$ is a closed proper convex function. We define the proximal residual function $R : \mathbb{R}^n \to \mathbb{R}^n$ as

$$R(x) := x - \mathrm{Prox}_P(x - \nabla\Phi(x)), \quad x \in \mathbb{R}^n. \tag{22}$$

Algorithm AS for (21) (simplified form) :

- Step 1. Given an initial index set $I_0 \subseteq [n]$, a given tolerance $\epsilon \geq 0$ and a given positive integer $k_{\max}$. Find an approximate solution x^0 to (21) with the constraint $x_{I_0^c} = 0$. Let $s = 0$.
- Step 2. Create $J_{s+1} = \left\{ j \in I_s^c \;\middle|\; (R(x^s))_j \neq 0 \right\}$. If $J_{s+1} = \emptyset$, let $I_{s+1} \leftarrow I_s$; otherwise, set a integer $0 < k \leq \min\{|J_{s+1}|, k_{\max}\}$ and define

  $$\widehat{J}_{s+1} = \left\{ j \in J_{s+1} \;\middle|\; |(R(x^s))_j| \text{ is among the first } k \text{ largest values in } \{|(R(x^s))_i|\}_{i\in J_{s+1}} \right\}.$$

  Update $I_{s+1} \leftarrow I_s \cup \widehat{J}_{s+1}$.
- Step 3. Find an approximate solution x^{s+1} to (21) with the constraint $x_{I_{s+1}^c} = 0$.
- Step 5. Set $s = s + 1$. Go to Step 2.

SMOP : A root finding based Secant Method for solving the Optimization Problem (CP(ϱ))

SMOP : A root finding based secant method for (CP(ϱ)) **:**

- Step 1. Given $0 < \underline{\lambda} < \lambda_1 < \lambda_0 \leq \overline{\lambda} \leq \lambda_\infty$ satisfying $\varphi(\underline{\lambda}) < \varrho < \varphi(\overline{\lambda})$. Call Algorithm AS with $I_0 = \emptyset$ to solve $(\mathrm{P_{LS}}(\lambda))$ with $\lambda = \lambda_0$ and obtain the solution $x(\lambda_0)$. Compute $\varphi(\lambda_0)$. Let $k = 1$.
- Step 2. Set $I_0^k = \{i \in [n] \,|\, (x(\lambda_{k-1}))_i \neq 0\}$.
- Step 3. Call Algorithm AS with $I_0 = I_0^k$ to solve $(\mathrm{P_{LS}}(\lambda))$ with $\lambda = \lambda_k$ to obtain $x(\lambda_k)$ and compute $\varphi(\lambda_k)$.
- Step 4. Generate λ_{k+1} by the globally convergent secant method.
- Step 5. Set $k = k + 1$. Go to Step 2.

Outline

Least-squares constrained optimization problem

Level-set : Properties of the value function $\varphi(\cdot)$

The HS-Jacobian of $\varphi(\cdot)$ for polyhedral gauge functions $p(\cdot)$

The convergence properties of the secant method

Adaptive sieving

Numerical experiments

Conclusion

Numerical experiments

Table – Statistics of the UCI test instances.

Problem idx	Name	m	n	Sparsity(A)	norm(b)
1	E2006.train	16087	150360	0.0083	452.8605
2	log1p.E2006.train	16087	4272227	0.0014	452.8605
3	E2006.test	3308	150358	0.0092	221.8758
4	log1p.E2006.test	3308	4272226	0.0016	221.8758
5	pyrim5	74	201376	0.5405	5.7768
6	triazines4	186	635376	0.6569	9.1455
7	bodyfat7	252	116280	1.0000	16.7594
8	housing7	506	77520	1.0000	547.3813

Table – The values of c to obtain $\varrho = c\|b\|$ when $p(\cdot) = \|\cdot\|_1$.

Test	idx	1	2	3	4	5	6	7	8
I	c	0.1	0.1	0.08	0.08	0.05	0.1	0.001	0.1
	nnz(x)	339	110	246	405	79	655	107	148
	c_{LS}	2.6-7	2.8-4	4.2-7	2.1-4	5.7-3	2.8-3	1.1-6	1.3-3
II	c	0.09	0.09	0.06	0.06	0.015	0.03	0.0001	0.04
	nnz(x)	1387	1475	884	1196	92	497	231	377
	c_{LS}	1.1-7	6.2-5	1.7-7	9.6-5	3.0-4	5.6-5	3.8-8	3.0-5

ℓ_1 penalty, Test I

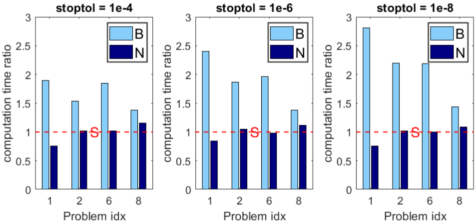
Table – The performance of SMOP (A1), SSNAL-LSM (A2), SPGL1 (A3) and ADMM (A4), in solving CP(ϱ) with $\varrho = c\|b\|$.

idx	time (s)				η				outermost iter			
	A1	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**	**A1**	**A2**	**A3**	**A4**
	Test I with stoptol $= 10^{-4}$											
1	1.39+0	2.18+2	3.51+2	4.22+2	2.3-5	4.9-5	1.0-4	1.0-4	24	29	7342	2049
2	2.29+0	5.12+2	1.45+3	6.84+2	3.1-6	7.8-5	9.0-5	8.7-5	12	16	3445	1470
3	4.02−1	5.83+1	3.21+2	8.87+1	9.4-6	2.6-5	1.0-4	1.0-4	24	30	21094	4918
4	1.59+0	2.06+2	7.19+2	9.90+1	1.2-5	7.3-5	9.5-5	1.3-5	13	15	3174	854
5	2.73−1	1.20+1	9.81+0	5.63+0	6.9-6	5.4-6	7.4-5	2.2-5	6	14	498	273
6	2.32+0	1.74+2	3.35+2	1.01+2	5.8-6	4.4-5	9.1-5	7.5-5	9	17	1987	571
7	4.35−1	9.12+0	8.98+0	8.59+0	2.8-5	5.9-5	9.8-5	9.9-5	15	18	539	583
8	2.99−1	9.07+0	1.29+1	7.94+0	2.6-5	8.6-5	1.0-4	9.0-5	10	14	515	424
	Test I with stoptol $= 10^{-6}$											
1	1.45+0	3.22+2	1.51+3	7.06+2	2.5-7	6.1-8	9.9-7	1.0-6	25	36	28172	3539
2	2.52+0	6.68+2	1.75+3	3.42+3	9.9-8	3.5-8	9.2-7	9.9-7	13	24	4155	8725
3	4.12−1	7.40+1	2.11+3	1.81+2	1.1-8	2.3-7	6.2-6	1.0-6	25	35	100000	10100
4	1.72+0	3.40+2	1.04+3	4.03+2	1.3-9	5.7-7	7.2-7	7.9-7	14	26	4584	3820
5	2.93−1	1.61+1	4.58+1	3.95+2	1.0-7	6.0-8	9.1-7	9.8-7	7	19	2468	20155
6	2.47+0	2.13+2	8.24+2	2.31+3	3.0-7	4.0-7	8.2-7	3.4-7	10	23	5578	13672
7	4.68−1	1.18+1	9.11+0	1.85+1	1.9-9	9.6-7	2.7-7	9.9-7	17	22	544	1250
8	3.28−1	1.45+1	3.84+1	4.40+1	2.4-7	8.4-8	4.0-7	8.7-7	11	24	1539	2427

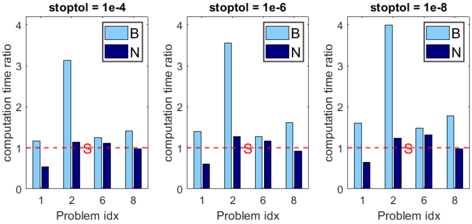
Table – The performance of SMOP (A1), SSNAL-LSM (A2), SPGL1 (A3) and ADMM (A4), in solving CP(ϱ) with $\varrho = c\|b\|$.

idx	time (s)				η				outermost iter			
	A1	A2	A3	A4	A1	A2	A3	A4	A1	A2	A3	A4
	Test II with stoptol $= 10^{-4}$											
1	7.26+0	4.51+2	1.38+3	6.12+2	3.0-6	4.6-5	1.0-4	1.0-4	26	30	27775	3014
2	6.79+0	1.54+3	1.32+3	4.01+2	1.8-5	3.6-5	9.7-5	6.8-5	14	21	3000	733
3	3.51+0	1.84+2	1.50+3	1.34+2	1.3-5	2.3-5	8.7-2	1.0-4	25	29	100000	7333
4	2.91+0	6.91+2	6.23+2	4.94+1	7.5-6	3.6-6	9.6-5	5.8-5	14	22	2694	385
5	6.23−1	1.53+1	8.65+0	2.01+1	2.8-5	7.9-6	6.6-5	9.5-5	9	13	395	1000
6	9.02+0	3.46+2	3.60+3	3.82+2	6.8-6	3.7-5	7.6-2	9.9-5	12	17	24924	2232
7	1.50+0	1.59+1	3.06+2	3.39+1	1.6-5	8.7-6	9.9-5	9.8-5	12	18	19820	2340
8	2.37+0	1.90+1	1.69+2	1.19+1	1.4-6	8.9-5	9.1-5	9.8-5	13	18	5914	644
	Test II with stoptol $= 10^{-6}$											
1	7.23+0	5.96+2	3.60+3	8.82+2	3.7−9	2.9-7	3.6-2	1.0-6	27	35	62384	4453
2	7.37+0	1.85+3	2.04+3	1.46+3	1.4−7	3.9-7	9.7-7	1.0-7	15	27	4688	3464
3	3.59+0	2.36+2	1.49+3	1.99+2	8.1-10	8.3-7	8.7-2	1.0-6	26	36	100000	11051
4	3.02+0	8.44+2	1.37+3	2.18+2	3.1−9	4.3-7	9.9-7	6.0-7	15	28	5912	1980
5	6.37−1	2.49+1	4.14+2	1.48+2	2.4−7	3.0-8	8.7-7	9.7-7	10	22	22091	7592
6	9.37+0	4.25+2	3.60+3	3.60+3	5.4-11	6.7-7	7.5-2	1.2-7	14	22	25158	21556
7	1.59+0	2.09+1	3.37+2	8.54+1	3.2−7	1.9-8	8.8-7	9.7-7	13	23	21523	5817
8	2.39+0	2.68+1	1.65+3	3.34+1	4.5−7	6.9-7	8.8-7	9.8-7	14	26	59147	1834

The ratio of computation time between BMOP (B) and NMOP (N) to the computation time of SMOP in solving (CP(ϱ))



Test I



Test II

Generating a solution path for $(\mathrm{CP}(\varrho))$.

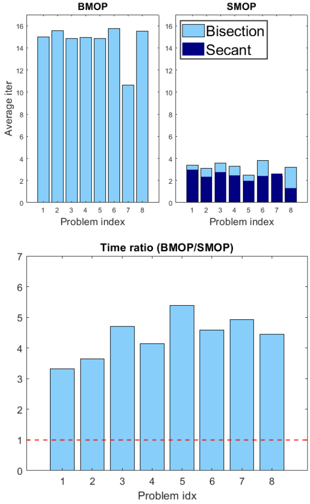


Fig. Test I

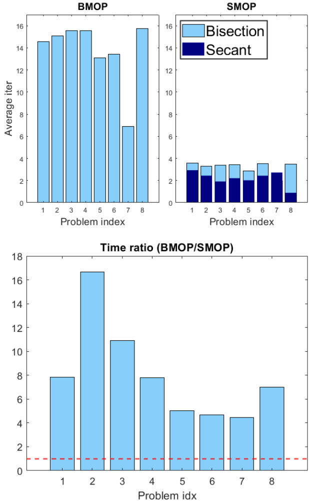


Fig. Test II

sorted ℓ_1 penalty

Table – The performance of SMOP (A1), Newt-ALM-LSM (A2) and ADMM (A4), in solving the sorted ℓ_1 penalized problems with least-squares constraints (CP(ϱ)) with $\varrho = c\|b\|$. The stopping tolerance is set to 10^{-6}.

idx	c	nnz(x)	c_{LS}	time (s)			η			outermost iter		
				A1	A2	A4	A1	A2	A4	A1	A2	A4
Test I												
2	0.15	3	2.4-2	3.84+0	1.34+2	$\underline{3.60+3}$	1.1-7	5.3-7	$\underline{2.8-1}$	8	21	$\underline{8637}$
4	0.1	3	4.8-3	4.79+0	1.35+2	$\underline{3.60+3}$	6.0-7	8.9-7	$\underline{2.9-4}$	10	17	$\underline{28891}$
5	0.1	113	1.9-2	6.29-1	4.98+1	4.23+2	1.0-7	4.5-7	1.5-7	7	22	17974
6	0.15	413	1.0-2	3.10+0	2.43+2	$\underline{3.60+3}$	2.7-7	1.6-7	$\underline{1.9-4}$	9	21	$\underline{19071}$
7	0.002	22	1.9-5	3.56-1	1.67+1	2.44+1	3.6-9	6.0-7	9.9-7	14	22	1616
8	0.15	95	6.9-3	6.06-1	2.55+1	1.57+2	1.3-7	7.7-7	9.0-7	10	23	8329
Test II												
1	0.1	339	2.6-7	2.53+1	1.40+2	5.13+2	2.9-7	5.6-7	1.0-6	25	34	2490
2	0.095	629	1.0-4	5.39+1	4.82+2	2.87+3	1.7-7	2.9-7	9.4-7	17	27	6770
3	0.08	246	4.2-7	4.98+0	6.54+1	1.60+2	2.0-8	7.1-7	1.0-6	25	36	8491
4	0.07	758	1.4-4	2.26+1	4.26+2	5.86+2	4.0-8	9.0-7	9.8-7	16	27	4550
5	0.02	95	5.7-4	2.05+0	9.87+1	3.58+2	3.2-8	5.6-7	7.6-7	11	20	15582
6	0.05	997	5.5-4	2.32+1	1.04+3	$\underline{3.60+3}$	8.4-7	2.1-7	$\underline{3.5-6}$	10	23	$\underline{19159}$
7	0.001	107	1.1-6	1.02+0	2.85+1	1.30+1	5.9-8	6.9-9	9.5-7	17	22	826
8	0.08	206	4.3-4	3.38+0	1.03+2	5.58+1	5.7-9	7.4-7	3.8-7	13	25	2842

sorted ℓ_1 penalty

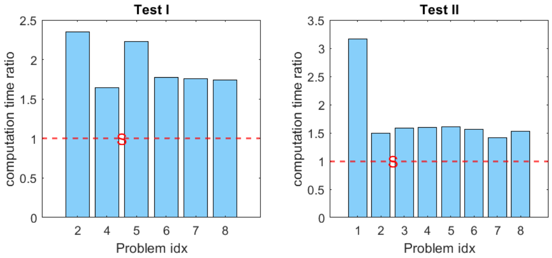


Fig. The computation time : SMOP & BMOP.

A non-polyhedral case

We consider the following group lasso penalty function

$$p(x) = \sum_{t=1}^{l} \sqrt{x_{2t-1}^2 + x_{2t}^2}, \quad x \in \mathbb{R}^{2l}. \tag{23}$$

Table – The values of c.

	idx	c	nnz(x)	c_{LS}
Test I	4	0.1	6	4.4-3
	5	0.1	50	2.4-2
	6	0.15	138	1.3-2
	7	0.002	28	2.4-5
	8	0.15	66	8.4-3
Test II	1	0.105	95	7.5-7
	3	0.08	403	4.3-7
	4	0.08	731	2.2-4
	5	0.02	120	9.1-4
	6	0.05	372	6.3-4
	7	0.001	186	1.3-6
	8	0.08	260	4.9-4

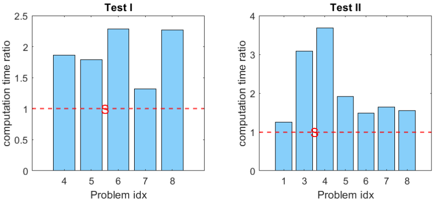


Fig. The computation time : SMOP & BMOP.

A non-polyhedral case

Table – The performance of SMOP (A1), SSNAL-LSM (A2), SPGL1 (A3) and ADMM (A4), in solving the group lasso penalized problems with least-squares constraints (CP(ϱ)) with $\varrho = c\|b\|$. The stopping tolerance is set to 10^{-6}.

idx	time (s)				η				outermost iter			
	A1	A2	A3	A4	A1	A2	A3	A4	A1	A2	A3	A4
Test I												
4	3.75+0	1.16+2	8.49+2	3.60+3	1.3−7	3.1-7	6.37-7	7.2-5	11	21	3024	22125
5	8.14−1	2.74+2	2.96+1	9.16+2	1.4−9	3.5-7	6.05-7	1.0-6	11	21	1319	38530
6	5.19+0	1.46+3	1.70+2	3.02+3	3.2-10	4.5-7	5.98-7	9.8-7	10	22	1086	15768
7	5.98−1	8.80+0	3.02+1	2.59+1	3.7−8	5.0-7	2.07-7	1.0-6	14	19	2102	1627
8	6.88−1	1.41+2	8.30+0	1.19+2	1.8−8	2.6-7	2.46-7	9.6-7	9	22	334	6211
Test II												
1	3.29+0	4.33+1	3.18+3	1.12+3	2.7-7	2.7-7	9.8-7	1.0-6	24	29	55596	5826
3	3.83+0	3.00+1	2.06+3	2.57+2	1.3-7	3.4-7	3.8-6	1.0-6	22	36	100000	13031
4	2.97+1	2.42+3	1.19+3	5.86+2	5.2-7	9.6-9	8.6-7	7.4-7	13	27	4241	3401
5	1.70+0	1.29+2	3.30+2	9.27+1	8.0-7	1.7-8	8.9-7	6.5-7	9	20	18001	3959
6	2.51+1	1.39+3	3.60+3	3.60+3	1.3-8	1.4-7	5.8-5	2.6-7	11	22	20646	19075
7	1.22+0	1.88+1	5.99+2	2.69+1	5.3-8	2.1-8	6.9-7	9.9-7	15	23	41578	1685
8	5.75+0	1.47+2	1.14+2	1.94+2	2.5-7	3.7-7	4.4-7	9.8-7	15	25	4373	9974

ℓ_1 penality cont.

Table – Comparison of computation time : SMOP to solve CP(ϱ) vs. SSNAL and the smoothing Newton algorithm (SmthN) to solve reduced $\mathrm{P_{LS}}(\lambda^*)$ for some large scale instances. In this test, the stopping tolerance is 10^{-6}.

	idx	reduced n	SMOP	SSNAL	SmthN	SMOP/SSNAL	SMOP/SmthN
Test I	1	339	1.95	0.70	0.12	2.78	16.02
	2	110	2.25	0.98	0.03	2.29	72.58
	3	247	0.67	0.09	0.01	7.14	61.00
	4	405	1.75	0.78	0.08	2.23	21.81
Test II	1	796	2.03	2.36	0.14	0.86	14.50
	2	629	7.84	11.07	0.65	0.71	11.99
	3	517	0.77	0.14	0.03	5.50	24.84
	4	758	3.17	1.25	0.31	2.54	10.33

The reduced $\mathrm{P_{LS}}(\lambda^*)$:

1. Obtain the non-zero index set I of the solution generated by SSNAL for the original problem $\mathrm{P_{LS}}(\lambda^*)$.
2. Remove all the columns from matrix A that correspond to the complement of index set I.

Outline

Conclusion

- When $p(\cdot)$ is a gauge function, we prove that $\varphi(\cdot)$ is (strongly) semismooth for a wide class of instances of $p(\cdot)$.
- When $p(\cdot)$ is a polyhedral gauge function, we show that $\varphi(\cdot)$ is locally piecewise C^k on $(0, \lambda_\infty)$ for any integer $k \geq 1$; and for any $\bar{\lambda} \in (0, \lambda_\infty)$, $v > 0$ for any $v \in \partial\varphi(\bar{\lambda})$.
- Under the assumption that $p(\cdot)$ is a polyhedral gauge function, we show that the secant method converges at least 3-step Q-quadratically for solving (E_φ), and if $\partial_{\mathrm{B}}\varphi(\lambda^*)$ is a singleton, the secant method converges superlinearly with Q-order at least $(1+\sqrt{5})/2$.
- We target to address the computational challenges for solving (CP(ϱ)) : Level-set approach + Secant method + adaptive sieving ("nonlinear column generation").

Reference

Qian Li, Defeng Sun, and Yancheng Yuan. "An efficient sieving based secant method for sparse optimization problems with least-squares constraints." arXiv preprint arXiv :2308.07812 (2023). *SIAM Journal on Optimization* (2024).

Thank you for your attention !

[Van den Berg-Friedlander 2008] Ewout Van den Berg, and Michael P. Friedlander. "Probing the Pareto frontier for basis pursuit solutions." Siam journal on scientific computing 31.2 (2008) : 890-912.

[Van den Berg-Friedlander 2011] Ewout Van den Berg, and Michael P. Friedlander. "Sparse optimization with least-squares constraints." SIAM Journal on Optimization 21.4 (2011) : 1201-1229.

[Li-Sun-Toh 2018] Xudong Li, Defeng Sun, and Kim-Chuan Toh. "On efficiently solving the subproblems of a level-set method for fused lasso problems." SIAM Journal on Optimization 28.2 (2018) : 1842-1866.

[Traub 1964] Joseph Frederick Traub. Iterative methods for the solution of equations. Prentice-Hall, Englewood Cliffs, 1964.

[Potra-Qi-Sun 1998] Florian A. Potra, Liqun Qi, and Defeng Sun. "Secant methods for semismooth equations." Numerische Mathematik 80 (1998) : 305-324.

[Han-Sun 1997] Jiye Han, and Defeng Sun. "Newton and quasi-Newton methods for normal maps with polyhedral sets." Journal of optimization Theory and Applications 94.3 (1997) : 659-676.

[Yuan-Lin-Sun-Toh 2023] Yancheng Yuan, Meixia Lin, Defeng Sun, and Kim-Chuan Toh. "Adaptive sieving : A dimension reduction technique for sparse optimization problems." arXiv preprint arXiv :2306.17369 (2023).