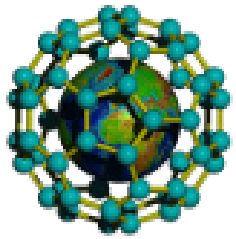


GOL

UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

A Unifying Framework for Sparsity Constrained Optimization

ODS 2021, Rome, 14th September 2021

M. Lapucci, T. Levato, F. Rinaldi, M. Sciandrone

DINFO, Univerisità di Firenze
Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova

Sparsity Constrained Optimization Problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & \|x\|_0 \leq s, \\ & x \in X, \end{aligned} \tag{SCOP}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable;
- $X \subseteq \mathbb{R}^n$ closed and convex;
- $s < n$;
- $\mathcal{X} = X \cap \{x \in \mathbb{R}^n \mid \|x\|_0 \leq s\}$.

Optimality

- **Global optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in \mathcal{X}$;

Optimality

- **Global optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in \mathcal{X}$;
- **Local optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in B(x^*, \epsilon) \cap \mathcal{X}$;

Optimality

- **Global optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in \mathcal{X}$;
- **Local optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in B(x^*, \epsilon) \cap \mathcal{X}$;
  - no combinatorial element;

Optimality

- **Global optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in \mathcal{X}$;
- **Local optimizer:** x^* s.t. $f(x^*) \leq f(x) \; \forall \, x \in B(x^*, \epsilon) \cap \mathcal{X}$;
  - no combinatorial element;
- **Necessary optimality conditions:**
  - *Component-Wise (CW) Minimum* (Beck and Eldar 2013)
    - $X = \mathbb{R}^n$, argmin required, single change in support;
  - *L-stationarity* (Beck and Eldar 2013; Beck and Hallak 2016)
    - Requires projection over $\mathcal{X}$;
  - *Lu and Zhang 2013*
    - KKTs w.r.t. super support;
  - *Basic Feasibility* (Beck and Eldar 2013; Beck and Hallak 2016)
    - stationarity w.r.t. super support;
  - *S-stationarity and M-stationarity* (Burdakov, Kanzow, and Schwartz 2016);
    - KKTs w.r.t. support.

Mixed-Integer Reformulation (Burdakov, Kanzow, and Schwartz 2016)

$$
\begin{aligned}
\min_{x,y}\ & f(x) \\
\text{s.t. }\ & e^\top y \geq n - s, \\
& x_i y_i = 0, \quad \forall\, i = 1, \ldots, n, \\
& x \in X, \\
& y \in \{0,1\}^n.
\end{aligned}
\tag{CCMIP}
$$

$$
\begin{aligned}
\mathcal{Y} &= \{y \mid y \in \{0,1\}^n, e^\top y \geq n - s\}, \\
\mathcal{X}(y) &= \{x \in X \mid x_i y_i = 0\ \forall\, i = 1, \ldots, n\}.
\end{aligned}
$$

Discrete Neighborhoods

- **Discrete Neighborhood:** $\mathcal{N}(x, y) \subset \mathcal{X} \times \mathcal{Y}$:

Discrete Neighborhoods

- **Discrete Neighborhood:** $\mathcal{N}(x, y) \subset \mathcal{X} \times \mathcal{Y}$:
  - $(x, y) \in \mathcal{N}(x, y)$,
  - $\hat{y} \in \mathcal{Y}$, $\hat{x} \in \mathcal{X}(\hat{y})$ for all $(\hat{x}, \hat{y}) \in \mathcal{N}(x, y)$,
  - $|\mathcal{N}(x, y)| < \infty$;

Discrete Neighborhoods

- **Discrete Neighborhood:** $\mathcal{N}(x, y) \subset \mathcal{X} \times \mathcal{Y}$:
  - $(x, y) \in \mathcal{N}(x, y)$,
  - $\hat{y} \in \mathcal{Y}$, $\hat{x} \in \mathcal{X}(\hat{y})$ for all $(\hat{x}, \hat{y}) \in \mathcal{N}(x, y)$,
  - $|\mathcal{N}(x, y)| < \infty$;
- **Tailored neighborhood:**

Discrete Neighborhoods

- **Discrete Neighborhood:** $\mathcal{N}(x, y) \subset \mathcal{X} \times \mathcal{Y}$:
  - $(x, y) \in \mathcal{N}(x, y)$,
  - $\hat{y} \in \mathcal{Y}$, $\hat{x} \in \mathcal{X}(\hat{y})$ for all $(\hat{x}, \hat{y}) \in \mathcal{N}(x, y)$,
  - $|\mathcal{N}(x, y)| < \infty$;
- **Tailored neighborhood:** $\mathcal{N}_\rho$
  - $(\hat{x}, \hat{y}) \in \mathcal{N}_\rho(x, y)$
  - $d_H(y, \hat{y}) \leq \rho$,
  - $\hat{x}_i = \begin{cases} x_i & \text{if } y_i = \hat{y}_i \\ 0 & \text{otherwise.} \end{cases}$

Discrete Neighborhoods

- **Discrete Neighborhood:** $\mathcal{N}(x, y) \subset \mathcal{X} \times \mathcal{Y}$:
  - $(x, y) \in \mathcal{N}(x, y)$,
  - $\hat{y} \in \mathcal{Y}$, $\hat{x} \in \mathcal{X}(\hat{y})$ for all $(\hat{x}, \hat{y}) \in \mathcal{N}(x, y)$,
  - $|\mathcal{N}(x, y)| < \infty$;
- **Tailored neighborhood:** $\mathcal{N}_\rho$
  - $(\hat{x}, \hat{y}) \in \mathcal{N}_\rho(x, y)$
  - $d_H(y, \hat{y}) \leq \rho$,
  - $\hat{x}_i = \begin{cases} x_i & \text{if } y_i = \hat{y}_i \\ 0 & \text{otherwise.} \end{cases}$
- Example: $n = 3$, $s = 2$, $\rho = 2$, $(x, y) \in \mathcal{X} \times \{0, 1\}^n$:

$$\mathcal{N}_2\left(\begin{pmatrix}1\\2\\0\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}\right) = \left\{\begin{pmatrix}1\\2\\0\end{pmatrix}\begin{pmatrix}0\\0\\1\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}\begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\2\\0\end{pmatrix}\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}\begin{pmatrix}0\\1\\1\end{pmatrix}, \begin{pmatrix}0\\2\\0\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix}, \begin{pmatrix}0\\0\\0\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix}\right\}$$

$\mathcal{N}$-stationarity

- **$\mathcal{N}$-stationary point of (SCOP):** x^* s.t.
  - $\exists y^* \in \mathcal{Y} : x^* \in \mathcal{X}(y^*)$;

$\mathcal{N}$-stationarity

- **$\mathcal{N}$-stationary point of (SCOP):** x^* s.t.
  - $\exists y^* \in \mathcal{Y} : x^* \in \mathcal{X}(y^*)$;
  - x^* is stationary for

$$\min_{x \in \mathcal{X}(y^*)} f(x);$$

$\mathcal{N}$-stationarity

- **$\mathcal{N}$-stationary point of (SCOP):** x^* s.t.
  - $\exists y^* \in \mathcal{Y} : x^* \in \mathcal{X}(y^*)$;
  - x^* is stationary for

$$\min_{x \in \mathcal{X}(y^*)} f(x);$$

  - $f(\hat{x}) \geq f(x^*)$ for all $(\hat{x}, \hat{y}) \in \mathcal{N}(x^*, y^*)$ and if $f(\hat{x}) = f(x^*)$ then $\hat{x}$ is stationary for

$$\min_{x \in \mathcal{X}(\hat{y})} f(\hat{x}).$$

$\mathcal{N}$-stationarity

- **$\mathcal{N}$-stationary point of (SCOP):** x^* s.t.
  - $\exists y^* \in \mathcal{Y} : x^* \in \mathcal{X}(y^*)$;
  - x^* is stationary for

$$\min_{x \in \mathcal{X}(y^*)} f(x);$$

  - $f(\hat{x}) \geq f(x^*)$ for all $(\hat{x}, \hat{y}) \in \mathcal{N}(x^*, y^*)$ and if $f(\hat{x}) = f(x^*)$ then $\hat{x}$ is stationary for

$$\min_{x \in \mathcal{X}(\hat{y})} f(\hat{x}).$$

- x^* **optimizer of (SCOP)** $\implies$ x^* **$\mathcal{N}$-stationary for (SCOP).**

Sparse Neighborhood Search

Algorithm 1 Sparse Neighborhood Search (SNS)

input: $y^0 \in \mathcal{Y}, x^0 \in \mathcal{X}(y^0), \xi \geq 0, \theta \in (0,1), \eta_0 > 0, \mu_0 > 0, \delta \in (0,1)$.

Step 0: Set $k = 0$.

Step 1: Compute $\tilde{x}^k$ by ProjectedGradientLineSearch(x^k, y^k).

Step 2: Define $W_k = \{(x,y) \in \mathcal{N}(\tilde{x}^k, y^k) \mid f(x) \leq f(\tilde{x}^k) + \xi\}$.

2.1: If $W_k \neq \emptyset$, choose $(x', y') \in W_k$, set $j = 1, x^j = x'$. Otherwise, go to Step 3.

2.2: Compute x^{j+1} by ProjectedGradientLineSearch(x^j, y').

2.3: If $f(x^{j+1}) \leq f(\tilde{x}^k) - \eta_k$, set $x^{k+1} = x^{j+1}, y^{k+1} = y', \eta_{k+1} = \eta_k$ and go to Step 4.

2.4: If $\|x^j - \Pi_{\mathcal{X}(y')}\left[x^j - \nabla f(x^j)\right]\| > \|x^k - \Pi_{\mathcal{X}(y^k)}\left[x^k - \nabla f(x^k)\right]\| + \mu_k$, set $j = j + 1$ and go to 2.2. Otherwise, set $W_k = W_k \setminus \{(x', y')\}$ and go to 2.1.

Step 3: Set $x^{k+1} = \tilde{x}^k, y^{k+1} = y^k$. If $f(x^{k+1}) \leq f(x^k) - \eta_k$, set $\eta_{k+1} = \eta_k$. Otherwise set $\eta_{k+1} = \theta\eta_k$.

Step 4: Set $\mu_{k+1} = \delta\mu_k$, $k = k + 1$ and go to Step 1.

Convergence Properties I

Assumption

The gradient $\nabla f(x)$ is Lipschitz-continuous over $\mathcal{X}$.

Assumption

Given $y^0 \in \mathcal{Y}$, $x^0 \in \mathcal{X}(y^0)$ and a scalar $\xi > 0$, the level set $\mathcal{L}(x^0, y^0) = \{(x, y) \in \mathcal{X}(y) \times \mathcal{Y} \mid f(x) \leq f(x^0) + \xi\}$ is compact.

Assumption

Let $\{(x^k, y^k)\}$ be converging to $(\bar{x}, \bar{y})$. Then, for any $(\hat{x}, \hat{y}) \in \mathcal{N}(\bar{x}, \bar{y})$, a sequence $\{(\hat{x}^k, \hat{y}^k)\}$ exists, converging to $(\hat{x}, \hat{y})$ such that $(\hat{x}^k, \hat{y}^k) \in \mathcal{N}(x^k, y^k)$.

Convergence Properties II

Proposition

For each iteration k of the SNS algorithm, the loop on the points in W_k terminates in a finite number of steps.

Theorem

Let $\{(x^k, y^k)\}$ be the sequence of iterates generated by SNS, and let $K_u = \{k \mid \eta_k < \eta_{k-1}\}$. Then:

1. *$\{(x^k, y^k)\}_{K_u}$ admits accumulation points;*
2. *every accumulation point (x^*, y^*) of $\{(x^k, y^k)\}_{K_u}$ is such that x^* is an $\mathcal{N}$-stationary point of problem (SCOP).*

Convergence Properties III

Theorem

Let $\{(x^k, y^k)\}$ *be the sequence of iterates generated by SNS equipped with* $\mathcal{N}_\rho$ *as neighborhood and* $\mathcal{A}^*$ *the set of the accumulation points of the sequence* $\{(x^k, y^k)\}_{K_u}$ *of unsuccessful iterates. If* $\rho \geq 2(s - \delta^*)$, *in the definition of the set* $\mathcal{N}_\rho(x, y)$, *and* $\delta^* = \min\{\|x^*\|_0 \mid (x^*, y^*) \in \mathcal{A}^*\}$, *then given a point* $(x^*, y^*) \in \mathcal{A}^*$, x^* *is basic feasible for problem (SCOP).*

Employing KKT stationarity I

Assumption

Given $\bar{y} \in \mathcal{Y}$ and $\bar{x} \in \mathcal{X}(\bar{y})$, we have that $\bar{x}$ is a stationary point for

$$\min_{x \in \mathcal{X}(\bar{y})} f(x)$$

if and only if there exist multipliers $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^n$ such that

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) + \sum_{i=1}^{p} \mu_i \nabla h_i(\bar{x}) + \sum_{i=1}^{n} \gamma_i e_i = 0,$$
$$\lambda_i \geq 0, \ \lambda_i g_i(\bar{x}) = 0, \ \forall i = 1, \ldots, m,$$
$$\gamma_i = 0, \ \forall \ i \text{ such that } \bar{y}_i = 0.$$

Employing KKT stationarity II

Theorem

Let $\{(x^k, y^k)\}$ be the sequence generated by SNS. Every accumulation point (x^*, y^*) of the sequence of unsuccessful iterates $\{(x^k, y^k)\}_{K_u}$ is S-stationary for (CCMIP) and x^* is M-stationary for (SCOP).

Theorem

Let $\{(x^k, y^k)\}$ be generated by SNS equipped with $\mathcal{N}_\rho$ as neighborhood and $\mathcal{A}^*$ the set of the cluster points of the sequence $\{(x^k, y^k)\}_{K_u}$ of unsuccessful iterates. If $\rho \geq 2(s - \delta^*)$, in the definition of the set $\mathcal{N}_\rho(x, y)$, and $\delta^* = \min\{\|x^*\|_0 \mid (x^*, y^*) \in \mathcal{A}^*\}$, then the optimality conditions from (Lu and Zhang 2013) hold.

Computational Experiments - Setup

- **Sparse Logistic Regression Problem**:

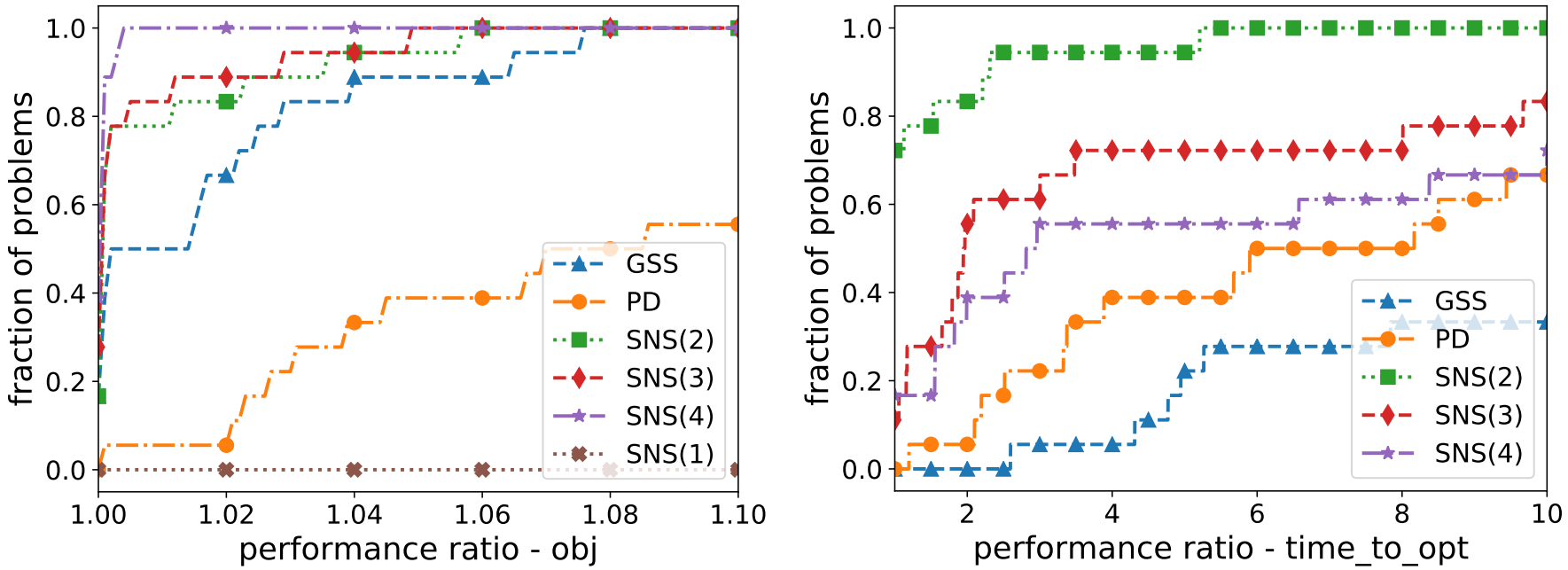
$$\min_w L(w) = \sum_{i=1}^{N} \log\left(1 + \exp\left(-t_i(w^\top z^i)\right)\right)$$
$$\text{s.t. } \|w\|_0 \leq s.$$

- **Datasets:**

Dataset	***N***	***n***	**Abbreviation**
Heart (Statlog)	270	25	`heart`
Breast Cancer Wisconsin (Prognostic)	194	33	`breast`
QSAR Biodegradation	1055	41	`biodeg`
SPECTF Heart	267	44	`spectf`
Spambase	4601	57	`spam`
Adult a2a	2265	123	`a2a`

- **Sparsity:** $s \in \{3, 5, 8\}$

Computational Experiments - Results I

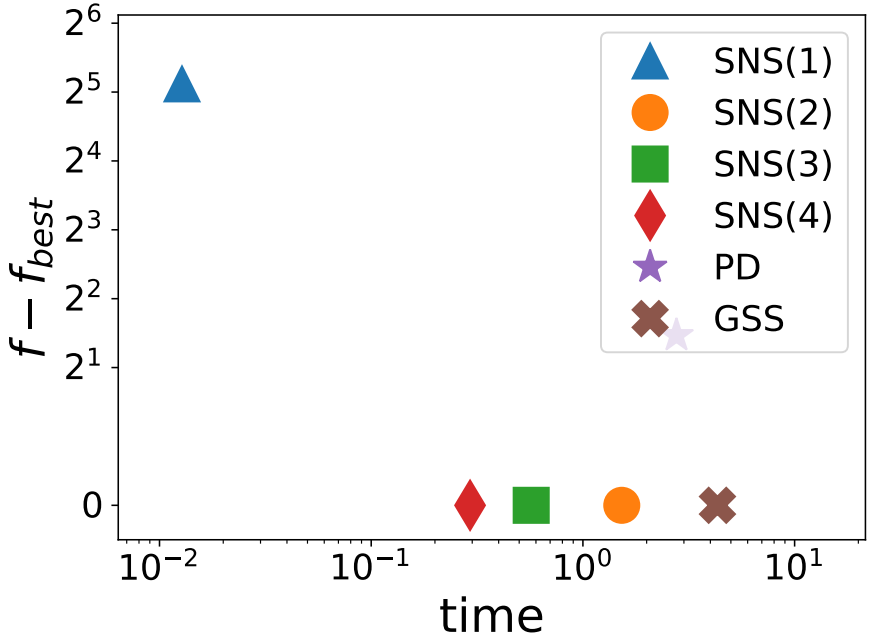


(a) objective value

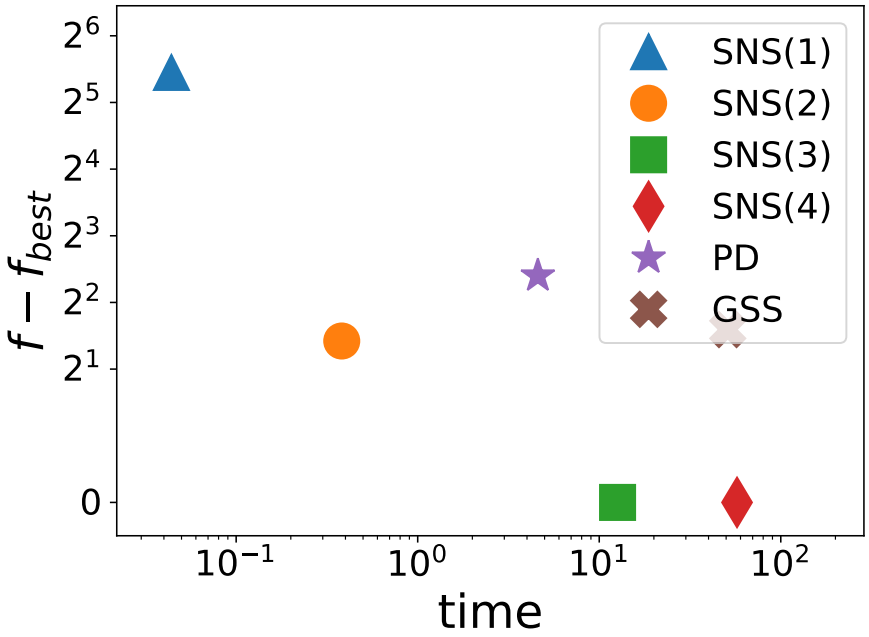
(b) time

Figure: Performance profiles for the considered algorithms on 18 sparse logistic regression problems.

Computational Experiments - Results II

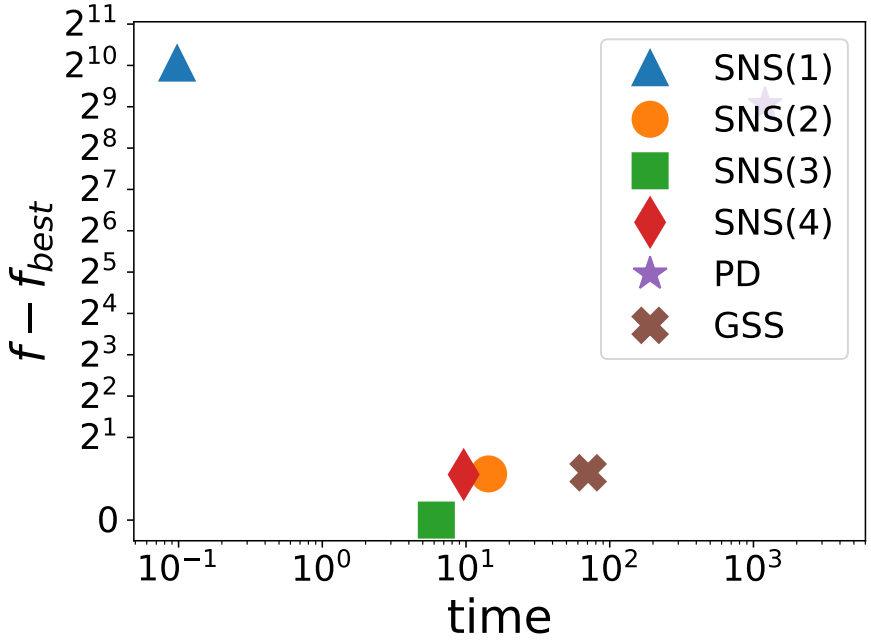


(a) `breast` - $s = 3$

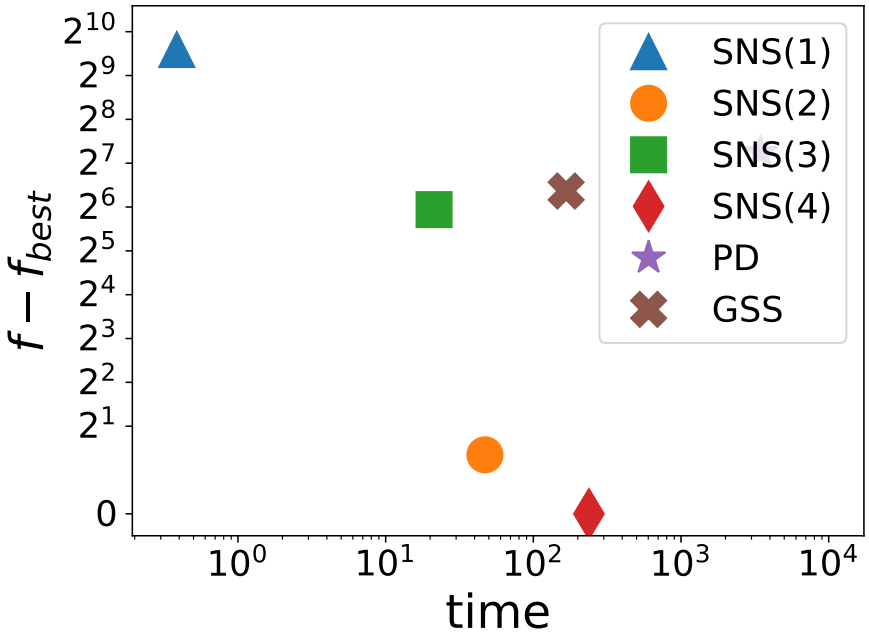


(b) `breast` - $s = 8$

Computational Experiments - Results III

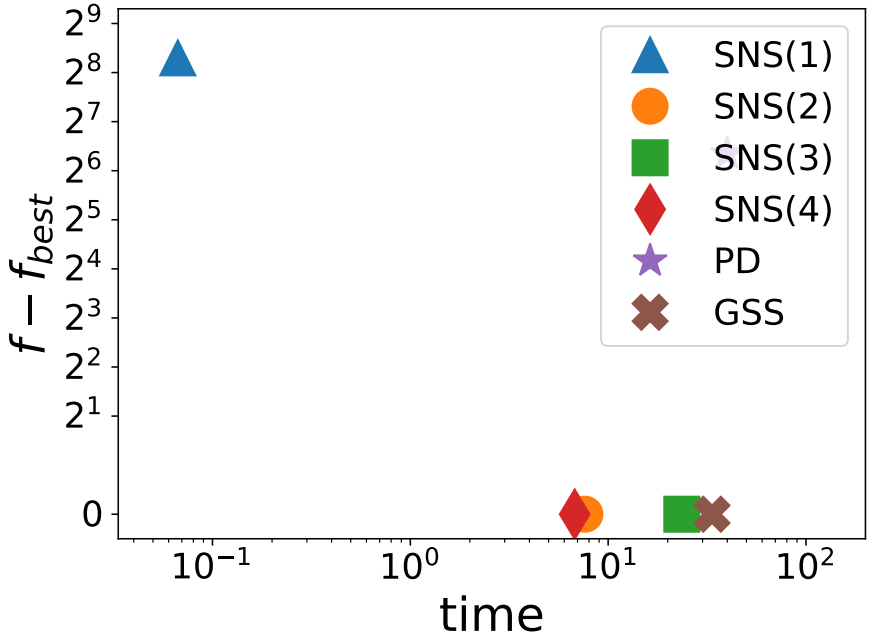


(c) spam - $s = 3$

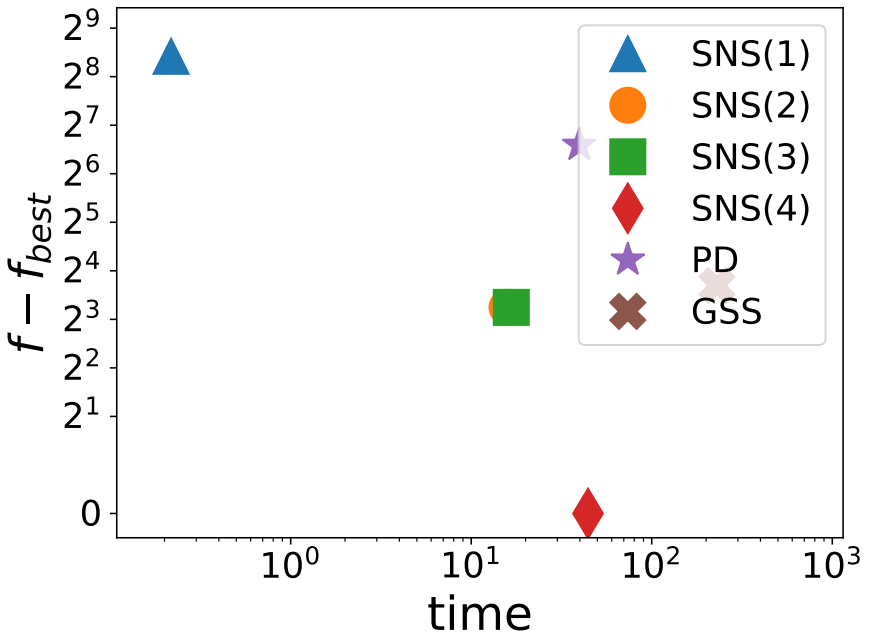


(d) spam - $s = 8$

Computational Experiments - Results IV



(e) a2a - $s = 3$



(f) a2a - $s = 8$

Figure: Quality/cost trade-off for the algorithms on sparse logistic regression problems from datasets `breast`, `spam` and a2a.

Conclusions

- New necessary optimality condition for (SCOP)
  - Takes into account changes in the support;
- New algorithmic framework
  - General convergence guarantees;
  - Convergence to well-known optimality conditions with tailored neighborhood;
  - Strong computational performance.
- Details in (Lapucci et al. 2021)

References I

A. Beck and Y. Eldar. "Sparsity Constrained Nonlinear Optimization: Optimality Conditions and Algorithms." In: *SIAM Journal on Optimization* 23.3 (2013), pp. 1480–1509.

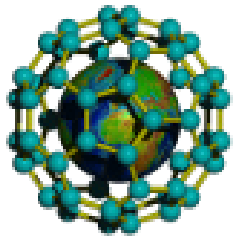
Amir Beck and Nadav Hallak. "On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms." In: *Mathematics of Operations Research* 41.1 (2016), pp. 196–223.

References II

O. Burdakov, C. Kanzow, and A. Schwartz. "Mathematical Programs with Cardinality Constraints: Reformulation by Complementarity-Type Conditions and a Regularization Method." In: *SIAM Journal on Optimization* 26.1 (2016), pp. 397–425.

M Lapucci et al. "A Unifying Framework for Sparsity Constrained Optimization." In: *arXiv preprint arXiv:2104.13244* (2021).

Z. Lu and Y. Zhang. "Sparse Approximation via Penalty Decomposition Methods." In: *SIAM Journal on Optimization* 23.4 (2013), pp. 2448–2478.



GOL

UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

A Unifying Framework for Sparsity Constrained Optimization

ODS 2021, Rome, 14th September 2021

M. Lapucci, T. Levato, F. Rinaldi, M. Sciandrone

DINFO, Univerisità di Firenze
Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova