Mathematical Programming
https://doi.org/10.1007/s10107-021-01766-4

FULL LENGTH PAPER

Series B

Linear-step solvability of some folded concave and singly-parametric sparse optimization problems

Andrés Gómez[1] · Ziyu He[1] · Jong-Shi Pang[1]

Received: 18 April 2021 / Accepted: 22 December 2021

Abstract

This paper studies several versions of the sparse optimization problem in statistical estimation defined by a pairwise separation objective. The sparsity (i.e., ℓ_0) function is approximated by a folded concave function; the pairwise separation gives rise to an objective of the Z-type. After presenting several realistic estimation problems to illustrate the Z-structure, we introduce a linear-step inner-outer loop algorithm for computing a directional stationary solution of the nonconvex nondifferentiable folded concave sparsity problem. When specialized to a quadratic loss function with a Z-matrix and a piecewise quadratic folded concave sparsity function, the overall complexity of the algorithm is a low-order polynomial in the number of variables of the problem; thus the algorithm is strongly polynomial in this quadratic case. We also consider the parametric version of the problem that has a weighted ℓ_1-regularizer and a quadratic loss function with a (hidden) Z-matrix. We present a linear-step algorithm in two cases depending on whether the variables have prescribed signs or with unknown signs. In both cases, a parametric algorithm is presented and its strong polynomiality is established under suitable conditions on the weights. Such a parametric algorithm can be combined with an interval search scheme for choosing the parameter to optimize a secondary objective function in a bilevel setting. The analysis makes use of a least-element property of a Z-function, and, for the case of a quadratic loss function, the strongly polynomial solvability of a linear complementarity problem with a hidden Z-matrix. The origin of the latter class of matrices can be traced to an inspirational paper of Olvi Mangasarian to whom we dedicate our present work.

Keywords Sparse optimization · Folded concave functions · Strong polynomiality · Parametric

Mathematics Subject Classification 90C20 · 90C26 · 90C31 · 90C33 · 62J07

The work of the first author was based on research support by the National Science Foundation under grant CIF-2006762. The work of the third author was based on research supported by the U.S. Air Force Office of Scientific Research under Grant FA9550-18-1-0382. A tribute to Olvi L. Mangasarian.

Extended author information available on the last page of the article

Published online: 20 January 2022

1 Introduction

In 1976, Mangasarian published a paper [44] on the solution of linear complementarity problems (LCPs) as linear programs. The main result of the paper is the identification of a special class of matrices for which positive vectors can be constructed to serve as the objective coefficients of a single linear program whose optimal solutions yield complementary solutions of the LCP. The work has led to the Ph.D. thesis [49] of the third author which connects Mangasarian's findings with a previous theory of polyhedra having least elements by Cottle and Veinott [22].Kindly check and confirm that the article title is correctly identified. Coined "hidden Z" (see [21, Notes and Comments 3.13.26] for the background of this term), Mangasarian's class of matrices becomes the foundation for the strong polynomialilty of the parametric principle pivoting algorithm and Lemke's algorithm for solving LCPs with these matrices; see [1,50] and the related paper [48]. With the exception of Adler et al. [1], all these references are decades old. An important goal of the present paper is to highlight how Mangasarian's seminal work before the birth of the field of machine learning benefits the modern topic of sparse optimization with such structures.

Sparse optimization is an important topic in statistical estimation [32]. Typically, the objective function of the optimization problem consists of a weighted sum of a loss function and a sparsity term defined by the ℓ_0-function of the variables; i.e., $\ell_0(t) \triangleq \begin{cases} 1 \text{ if } t \neq 0 \\ 0 \text{ if } t = 0 \end{cases}$ for $t \in \mathbb{R}$. To handle the latter discontinuous function, various convex and nonconvex surrogate functions as well as integer and complementarity formulations have been proposed, the former as approximations and the latter for global optimization. Among the family of convex surrogate functions, the weighted ℓ_1-function is perhaps the most popular due to its simplicity although there are many variations [32]. As early as 2001, it has been recognized [27] in the statistics community that nonconvex surrogate sparsity functions have favorable statistical properties that are lacking in the family of convex sparsity functions. In particular, the class of folded concave functions [28] provides a unification of many nonconvex approximations of the ℓ_0-function; for the study of optimization problems using a folded concave sparsity function, see [2,40] where these problems are treated as difference-of-convex programs. In the machine learning literature, there is an early result [43] showing that for the least-squares regression problem, the Lasso regularization path has an exponential number (in the number of variables) of linear segments. More recently, the paper [16] presents some NP-hardness results pertaining to a variety of sparse optimization problems with folded concave penalty functions.

Motivated by some recent works [3,4] by the first author on the subject of sparse optimization with M-matrices (and a further related work [29]) and by the connection of these problems to strongly polynomially solvable LCPs with matrices of the same class, we are led to the question of whether some sparse optimization problems and their bilevel extensions of optimally selecting the sparsity penalty can be solved by *strongly polynomially bounded* algorithms. The main goal of this paper is to identify some such problems along with the solution algorithms that answer this question affirmatively. Thus this goal is in direct contrast to the negative results in [16,43]. Some comparative remarks with several most relevant references [9,10,41,47] are given in

Sect. 5.1. To clarify, by strongly polynomial complexity, we mean that the total number of arithmetic operations of the algorithm is a (low-degree) polynomial of the number of variables of the problem. In the presence of nonlinear functions (e.g., a non-quadratic loss function), we use the term "linear-step" to refer to the linear number (in the number of variables of the problem) of subproblems each solvable by an algorithm that involves solving a finite number of nonlinear equations; needless to say, each such equation step is in general an infinite process. When these equations are linear, then the linear-step algorithm becomes a strongly polynomial algorithm. The main factor contributing to our favorable complexity results is the Z-property of the loss functions. As a unification of the motivating works and other related problems, such as a sparse nearly isotonic regression problem [55], we frame our study based on a generalization of the deviation-separation problem defined in [34] parameterized by the sparsity of the variables. An important remark about the strong polynomiality property of an algorithm is that all computations are assumed to be exact. For sparse minimization problems with quadratic loss functions and piecewise quadratic regularizers, exactness of the computations and operation counts, both of which are finite, can be maintained throughout. Since the emphasis of our paper is on strong polynomiality derived from a "finite" algorithm in a more general setting, we do not concern with inexact versions of the algorithm, which typically will involve some inexactness measures that lead to a complexity analysis of a different kind.

We end this introduction by re-iterating the motivation for us to undertake this research. Namely, by way of the identification of important classes of sparse minimization problems that are strongly polynomially solvable, as opposed to much (if not all) of the existing literature that has not touched on this aspect of these problems, we are able to pay tribute to our beloved colleague Olvi Mangasarian for his pioneering contributions that allow us to deepen the understanding of some contemporary problems of significance in machine learning and statistical estimation. Overall, our study not only enriches the computational research of sparse optimization, it expands the domain of modern nonconvex nondifferentiable optimization [24] involving the minimization of a sum objective consisting of a convex function and a concave composite piecewise affine function.

2 Problem formulations

Generalizing the formulation in [34], consider the following parametric nonlinear program with sparsity control:

$$\underset{\ell \le x \le u}{\textbf{minimize}} \ \underbrace{\sum_{i=1}^{n} h_i(x_i) + \sum_{(i,j)\in V} g_{ij}(a_{ij}x_i - b_{ij}x_j)}_{\text{denoted } \theta(x)} + \gamma \sum_{i=1}^{n} |\, x_i \,|_0, \qquad \gamma \ge 0, \tag{1}$$

where each $h_i : \mathbb{R} \to \mathbb{R}_+$ is a continuously differentiable, strongly convex function; each $g_{ij} : \mathbb{R} \to \mathbb{R}_+$ is a continuously differentiable convex function; $V \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$; $\{a_{ij}, b_{ij}\}_{(i,j)\in V}$ are nonnegative scalars; and to avoid some tedious details,

$-\infty \leq \ell_i < 0 < u_i \leq \infty$ are given upper and lower bounds on the variable x_i. The "pairwise separation" summands $g_{ij}(a_{ij}x_i - b_{ij}x_j)$ generalize the deviation-separation objectives in [34] that have all the coefficients $\{a_{ij}, b_{ij}\}_{(i,j)\in V}$ equal to unity. We will denote the feasible set $\{x \in \mathbb{R}^n \mid \ell \leq x \leq u\} \triangleq [\,\ell, u\,]$. The latter interval notation allows some of the bounds to be $\pm\infty$; in this case, it is understood that the corresponding $[\,\ell_i, u_i\,]$ is meant to be an unbounded interval. In applications, there is often a multiplicative factor $\lambda > 0$ of the second sum which is a measure of the separation of the pairs of the variables (x_i, x_j) for $(i, j) \in V$; this factor is a weight of the deviation measure relative to the first sum. Throughout this paper, we take λ to be a constant and absorb it into the functions g_{ij}. Under the given setting, the function θ is continuously differentiable and strongly convex. For any positive integer k, we write $[k] \triangleq \{1, \ldots, k\}$.

The problem (1) encompasses two cases: fixed and parametric γ. While the fixed-γ case is of independent interest, the parametric-γ case is the cornerstone for solving the singly-parametric bilevel optimization problem:

$$\begin{array}{ll} \underset{(x,\gamma)\in\mathbb{R}^n\times\mathbb{R}_+}{\textbf{minimize}} & \psi(x,\gamma) \\ \textbf{subject to} & x \text{ is a "solution" of (1) corresponding to } \gamma, \end{array} \tag{2}$$

where $\psi : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ is a given "easy-to-handle" function; e.g., $\psi(x, \gamma)$ is a quadratic function in x alone; more generally, this outer objective function is such that for any "solution" path $x(\gamma)$ of the parametric problem (1), the univariate composite function $\psi(x(\gamma), \gamma)$ in γ alone is easy to minimize. The word "solution" is in quotes because we have yet to specify its meaning in view of the nonconvex and discontinuous objective function of (1). The use of bilevel programming as a systematic approach for the optimal selection of hyper-parameters in machine learning models has been investigated more than a decade ago; see [7,36,37] and more recently [38]. In these references, the learning models are support vector machines with 2 parameters, which are being optimized with respect to a cross-validation objective formed from some hidden data. As it is known, the global resolution of such a bi-parameter identification problem [39] is computationally very challenging. While the problem (2) is singly-parametric, the computation of its global solution remains elusive in general. Part of the contributions of this paper is the identification of some (surrogate) sparsity problems whose solution paths can be traced out in linear number of iterations in terms of the number of variables; thus the global solution of the bilevel parameter identification can be accomplished via linearly many successive interval searches.

2.1 Some source problems

We give several source problems that all lead to special cases of the problem (1).

Sparse and smooth signal estimation [4,5] Consider the problem of recovering an uncorrupted sparse signal represented by a vector $x \in \mathbb{R}^n$ from corrupted data $a \in \mathbb{R}^n$. In addition to sparsity of the signal which is at the core of compressed sensing, smoothness of the signal is often an important characteristics in denoising. Combining these two criteria, one simple version of the signal recovery problem can be formulated as:

$$\underset{x \in \mathbb{R}^n}{\textbf{minimize}} \ \frac{1}{2} \left[\sum_{i=1}^{n} (a_i - x_i)^2 + \lambda \sum_{(i,j) \in V} (x_i - x_j)^2 \right] + \gamma \sum_{i=1}^{n} | x_i |_0, \tag{3}$$

which is clearly of the form (1) for each fixed λ. The version of this problem considered in [4] has the data vector a as nonnegative; in this case, a nonnegative signal x is sought, rendering the above problem a nonnegatively constrained quadratic minimizaion problem with sparsity control.

Inference of time-varying Gaussian Markov Random Fields (MRFs) [29] The standard approach, the so-called *time-varying Graphical Lasso* [11,31], for addressing this problem of estimating the precision matrix of a time-varying Gaussian MRF given data calls for solving a semidefinite optimization problem (SDP). Similar to the previous problem, the objective of the MRF problem contains both an ℓ_1 regularization (capturing the sparsity of the precision matrix) and fused lasso terms (capturing a smooth evolution of the process over time). For the purpose of handling large-scale problems more effectively, the authors in [29] propose an alternative to the SDP approach to compute estimates $\{ \widehat{\Theta}_t \}_{t=1}^{T}$ of the precision matrices by solving an optimization problem of the form:

$$\underset{\Theta_t \in [\ell, u]^{n \times n}}{\textbf{minimize}} \sum_{t=1}^{T} \left[\frac{1}{2} \left\| \Theta_t - \widehat{\Sigma}_t^{-1} \right\|_{\mathrm{F}}^2 + \lambda \sum_{t=1}^{T-1} g(\Theta_{t+1} - \Theta_t) + \gamma \| \Theta_t \|_0 \right], \tag{4}$$

where $\| \cdot \|_{\mathrm{F}}$ is the Fröbenius norm of a matrix, $\widehat{\Sigma}_t^{-1}$ are estimates (computed a priori) of the precision matrix at time t that ignore the smooth evolution of the process, and $g : \mathbb{R}^{n \times n} \to \mathbb{R}$ is a separable function capturing the time dependence. The squared $\| \cdot \|_{\mathrm{F}}$-norm of the deviation from the estimates is an alternative to a constrained formulation used in [29].

Scaled nearly isotonic regression with sparsity control The isotonic (modeled by the constraints $x_i \leq x_{i+1}$ for $i \in [n-1]$) regression problem is classical in statistics [6]. The paper [55] introduces the "nearly isotonicity regression" problem without sparsity control wherein the order constraint is replaced by a penalty term $\max(x_i - x_{i+1}, 0)$ added to the objective function as a relaxation of the isotonic constraint. The near isotonicity can be generalized to scaled near isotonicity; this generalization with sparsity control leads to the problem:

$$\underset{x \in \mathbb{R}^n}{\textbf{minimize}} \ \frac{1}{2} \left[\sum_{i=1}^{n} (a_i - x_i)^2 + \lambda \sum_{i=1}^{n-1} \max(x_i - b_{i+1} x_{i+1}, 0)^2 \right] + \gamma \sum_{i=1}^{n} | x_i |_0, \tag{5}$$

where we have used a quadratic penalty function to model the violation of the scaled isotonicity constraint: $x_i \le b_{i+1} x_{i+1}$. Since the univariate function $t \mapsto \max(t, 0)^2$ is nondecreasing, (once but not twice) differentiable, and convex, we obtain yet another instance of the problem (1). An interesting application of the above problem pertains to spike detection in calcium imaging data [14,35,56]. In such data, moments in which a neuron spikes are characterized by an intense flood of calcium into the cell, followed by a smooth decay to a baseline level. The near isotonicity term $\sum_{i=1}^{n-1} \max(\, x_i - b_{i+1}\, x_{i+1},\, 0\,)^2$ is a variation of the smoothness term $\sum_{(i,j)\in V} (x_i - x_j)^2$ in (3) and may serve to more adequately model such sudden calcium spikes; moreover sparsity captures the fact that neurons are inactive most of the time.

Portfolio revision with transaction costs [3,8] Consider the problem of portfolio selection/revision with transaction costs. There are N risky assets with current holdings $w \in \mathbb{R}^N$, expected returns $\{\mu_i\}_{i=1}^N$, and variance-covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$. Let $a \in \mathbb{R}_{++}$ be the positive fixed transaction costs associated with the buying/selling any quantity, and $c \in \mathbb{R}^N$ be the unit variable costs of trading the assets. Let $u \in \mathbb{R}_{++}$ be the upper bounds of the transacted amounts; these bounds are used in lieu of the usual budget constraint. Then with the objective of maximizing returns less transaction costs and minimizing the variance of the transactions, with the two conflicting objectives being balanced by a parameter $\rho \ge 0$ in accordance with the classic Markowitz portfolio model, this problem with a suitable nonnegative ε-regularization for stability may be formulated as:

$$\begin{aligned} \underset{-u \le x \le u}{\textbf{minimize}} \ \frac{1}{2} \left[\, (\, w + x \,)^\top \Sigma (\, w + x \,) + \varepsilon \sum_{i=1}^n x_i^2 \, \right] \\ + \rho \sum_{i=1}^n \left[\, -\mu_i \, w_i - (\, \mu_i - c_i \,)\, x_i + a_i \,|\, x_i \,|_0 \,\right]. \end{aligned} \tag{6}$$

The quadratic term with the general variance-covariance matrix Σ may not correspond directly to the form (1); in particular, it does not have the required non-positive sign patterns in the off-diagonal elements as would be implied by the terms $g_{ij}(a_{ij}x_i - b_{ij}x_j)$. Nevertheless, with Σ being strictly quasi-diagonally dominant, i.e., if there exists a positive vector d such that

$$\Sigma_{ii}\, d_i \; > \; \sum_{j \ne i} |\, \Sigma_{ij} \,|\, d_j, \qquad \forall\, i \; = \; 1, \ldots, n, \tag{7}$$

or in general, with a large enough choice of the regularization scalar ε, the above parametric problem (in ρ) is still amenable to treatment by an easy extension of Algorithm III in Sect. 7.1. While diagonal dominance of Σ arises in a factor model if the idiosyncratic risk of a portfolio overshadows the systematic risk, the choice of a sufficiently large $\varepsilon > 0$ may help to induce less frequency of the portfolio transactions (thus more sparsity in x) due to the associated costs and also as a mechanism to robustify the problem against uncertainty as suggested in [8].

3 Folded concave approximations: fixed $\gamma > 0$

The complication of the problem (1) is due to the third sum, without which the problem is a standard differentiable, strongly convex program. Obviously, this problem as stated, not to mention its parametric extension (2), is computationally very challenging. Aside from the fact that (1) admits a mixed-integer nonlinear programming formulation, for any algorithm that attempts to relieve the global minimization of the problem, it is important to understand what property a computed solution can possibly have. For this purpose, we follow a common approach in statistics to deal with the ℓ_0-function, which is to approximate this discontinuous function by a scaled folded concave function. For each $i \in [n]$, let $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ satisfy:

- (*blanket assumption*): f_i is continuous, concave, and such that $f_i(0) = 0$ and the one-sided directional derivative $f_i{}'(0; 1) > 0$; moreover f_i is strictly increasing before it eventually becomes flat after a certain value, which we may assume without loss of generality is to the left of the upper bound u_i.

See the right-hand functions in Fig. 1 which we borrow from Cui and Pang [24]. We then approximate the problem (1) by

$$\underset{\ell \le x \le u}{\textbf{minimize}} \; \varphi_{\gamma,\delta}(x) \triangleq \theta(x) + \gamma \sum_{i=1}^{n} \underbrace{f_i\left(\frac{|\,x_i\,|}{\delta}\right)}_{\text{denoted}\,\rho(|x_i|,\delta)\,\text{in Fig. 1}}, \quad (\gamma, \delta) > 0. \tag{8}$$

The two parameters serve different roles: γ for sparsity control and δ for approximation accuracy. For a large family of concave functions f_i, we have $\lim_{\delta \downarrow 0} f_i\left(\frac{\xi}{\delta}\right) = |\,\xi\,|_0$ for all $\xi > 0$ although this limiting property is not essential throughout this paper. See [25,40] for studies on the use of the δ-parameter to control the approximation of the ℓ_0-function. In addition to the ℓ_1-function, there are several popular classes of folded concave functions, all being of the piecewise kind: the capped ℓ_1-function [40, Section 5] defined by $\xi(\ge 0) \mapsto \min(\xi, 1)$, the minimax concave penalty MCP function [26,58], and the smoothly clipped absolute deviation SCAD function [27]; as demonstrated in [2], all these functions are not differentiable at the origin as they are approximations of the ℓ_0 function that is discontinuous there; see Fig. 1. Omitting their explicit definitions, which can be found in the cited references, we note that both the SCAD and MCP functions are both once continuous differentiable and piecewise linear-quadratic; the latter means that for each of the associated f_i functions, there is a partition of the interval $[0, \infty)$ into a finite number of non-overlapping subintervals on each of which f_i is a quadratic function; see Sect. 6.2 for a more formal definition.

As a coupled nonconvex nondifferentiable optimization problem, a first question to ask about (8) is what kind of solution an algorithm can be expected to compute. Since the computation of a globally optimal solution is out of the question, one should settle for a stationary solution of the sharpest kind; see [24,51]. For (8), a (directional)-stationary solution is the best kind one can hope for. In general, for a function $\varphi : \mathcal{O} \subseteq$

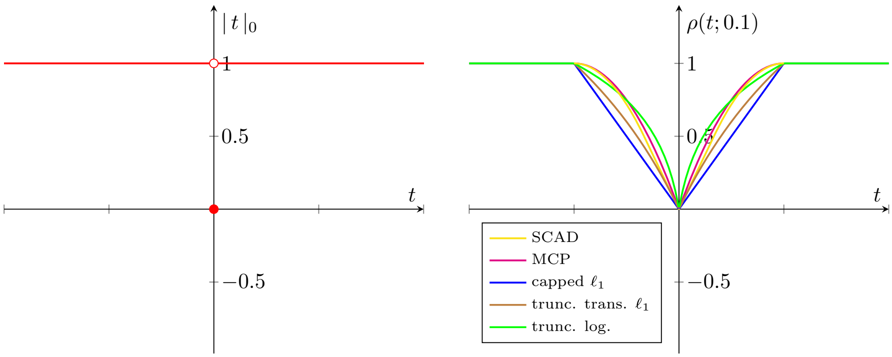


Fig. 1 The ℓ_0 function and surrogate sparsity functions $\rho(\cdot, 0.1)$

$\mathbb{R}^n \to \mathbb{R}$ defined on an open set $\mathcal{O}$, the directional derivative of φ at a vector $\bar{x} \in \mathcal{O}$ along a direction v, denoted $\varphi'(\bar{x}; v)$, is by definition

$$\varphi'(\bar{x}; v) \triangleq \lim_{\tau \downarrow 0} \frac{\varphi(\bar{x} + \tau v) - \varphi(\bar{x})}{\tau}$$

provided that the limit exists. A directional stationary solution of (8) is a feasible vector $\bar{x} \in [\ell, u]$ such that $\varphi'_{\gamma,\delta}(\bar{x}; x - \bar{x}) \geq 0$ for all $x \in [\ell, u]$. We have

$$\begin{aligned}
\varphi'_{\gamma,\delta}(\bar{x}; v) &= \nabla\theta(\bar{x})^\top v + \frac{\gamma}{\delta} \left[\sum_{i\,:\,\bar{x}_i > 0} f_i{}' \left(\frac{\bar{x}_i}{\delta};\ v_i \right) + \sum_{i\,:\,\bar{x}_i < 0} f_i{}' \left(\frac{-\bar{x}_i}{\delta};\ -v_i \right) \right. \\
&\quad \left. + \sum_{i\,:\,\bar{x}_i = 0} f_i{}'\,(0; |v_i|) \right] \\
&= \sum_{i\,:\,\bar{x}_i > 0} \left[\frac{\partial\theta(\bar{x})}{\partial x_i}\, v_i + \frac{\gamma}{\delta}\, f_i{}' \left(\frac{\bar{x}_i}{\delta};\ v_i \right) \right] \\
&\quad + \sum_{i\,:\,\bar{x}_i < 0} \left[\frac{\partial\theta(\bar{x})}{\partial x_i}\, v_i + \frac{\gamma}{\delta}\, f_i{}' \left(\frac{-\bar{x}_i}{\delta};\ -v_i \right) \right] \\
&\quad + \sum_{i\,:\,\bar{x}_i = 0} \left[\frac{\partial\theta(\bar{x})}{\partial x_i}\, v_i + \frac{\gamma}{\delta}\, f_i{}'(0; |v_i|) \right].
\end{aligned}$$

For a nonzero scalar t, we let $\operatorname{sgn}(t) \triangleq \pm 1$ depending on the sign of t. Define, for a given vector $\bar{x} \in \mathbb{R}^n$, the index sets:

$$\mathcal{A}_+(\bar{x}) \triangleq \{ i \mid \bar{x}_i > 0 \}; \quad \mathcal{A}_0(\bar{x}) \triangleq \{ i \mid \bar{x}_i = 0 \} \quad \text{and} \quad \mathcal{A}_-(\bar{x}) \triangleq \{ i \mid \bar{x}_i < 0 \}. \tag{9}$$

Based on the above expression for the directional derivatives, we obtain the following necessary and sufficient conditions for a vector $\bar{x} \in [\,\ell, u\,]$ to be a d-stationary point of the problem (8). The proposition also identifies a distinguished property of such a stationary point when each f_i is additionally piecewise affine. We recall that in our setting $-\infty \le \ell_i < 0 < u_i \le \infty$ for all $i \in [n]$.

Proposition 1 *For every pair $(\gamma, \delta) > 0$, problem (8) has a global minimizer, thus a d-stationary point. Moreover, a feasible vector $\bar{x}$ is a d-stationary point if and only if the following conditions hold:*

- *f_i is differentiable at $|\bar{x}_i|/\delta$ for all i such that $\bar{x}_i \neq 0$ and $\ell_i < \bar{x}_i < u_i$;*
- $\dfrac{\partial \theta(\bar{x})}{\partial x_i} + \mathrm{sgn}(\bar{x}_i)\, \dfrac{\gamma}{\delta}\, f_i{}'\left(\dfrac{|\bar{x}_i|}{\delta}\right) = 0$ *for all i such that $\bar{x}_i \neq 0$ and $\ell_i < \bar{x}_i < u_i$;*
- $\left|\dfrac{\partial \theta(\bar{x})}{\partial x_i}\right| \le \dfrac{\gamma}{\delta}\, f_i{}'(0;1)$ *for all $i \in \mathcal{A}_0(\bar{x})$;*
- $\dfrac{\partial \theta(\bar{x})}{\partial x_i} + \dfrac{\gamma}{\delta}\, f_i{}'\left(\dfrac{|\bar{x}_i|}{\delta}; -1\right) \ge 0$ *for all i such that $\bar{x}_i = \ell_i$;*
- $-\dfrac{\partial \theta(\bar{x})}{\partial x_i} + \dfrac{\gamma}{\delta}\, f_i{}'\left(\dfrac{|\bar{x}_i|}{\delta}; -1\right) \ge 0$ *for all i such that $\bar{x}_i = u_i$.*

If each f_i is additionally piecewise affine, then every d-stationary point of problem (8) is a local minimizer.

Proof Since θ is strongly convex and each f_i is nonnegative, the function $\varphi_{\gamma,\delta}$ is coercive. Therefore, problem (8) has a global minimizer which must necessarily be a d-stationary point. The sufficiency condition of such a stationary point is easy to show by using the directional derivative formula and verifying that $\varphi'_{\gamma,\delta}(\bar{x}; x - \bar{x}) \ge 0$ for all x satisfying $\ell \le x \le u$. To show necessity, let i be such that $0 < \bar{x}_i < u_i$. It then follows that

$$\frac{\partial \theta(\bar{x})}{\partial x_i}\, v_i + \frac{\gamma}{\delta}\, f_i{}'\left(\frac{\bar{x}_i}{\delta};\, v_i\right) \ge 0, \qquad \forall\, v_i \in \mathbb{R}.$$

Letting $v_i = \pm 1$ we obtain the two inequalities:

$$\frac{\partial \theta(\bar{x})}{\partial x_i} + \frac{\gamma}{\delta}\, f_i{}'\left(\frac{\bar{x}_i}{\delta}; 1\right) \ge 0 \quad \text{and} \quad -\frac{\partial \theta(\bar{x})}{\partial x_i} + \frac{\gamma}{\delta}\, f_i{}'\left(\frac{\bar{x}_i}{\delta}; -1\right) \ge 0.$$

Adding, we obtain

$$f_i{}'\left(\frac{\bar{x}_i}{\delta}; 1\right) + f_i{}'\left(\frac{\bar{x}_i}{\delta}; -1\right) \ge 0;$$

since f_i is concave, the left-hand sum is nonpositive. Hence we deduce

$$\frac{\partial \theta(\bar{x})}{\partial x_i} = -\frac{\gamma}{\delta}\, f_i{}'\left(\frac{\bar{x}_i}{\delta}; 1\right) = \frac{\gamma}{\delta}\, f_i{}'\left(\frac{\bar{x}_i}{\delta}; -1\right).$$

It follows that f_i is differentiable at $\bar{x}_i/\delta$ and we have

$$\frac{\partial \theta(\bar{x})}{\partial x_i} = -\frac{\gamma}{\delta}\, f_i{}'\left(\frac{\bar{x}_i}{\delta}\right).$$

By a similar argument, we can prove the case when $\ell_i < \bar{x}_i < 0$. This completes the proof of the case where i is such that $\bar{x}_i \neq 0$ and $\ell_i < \bar{x}_i < u_i$. The remaining three conditions can also be proved by similar arguments. Consider for instance the case where $\bar{x} = \ell_i$. We then have

$$\frac{\partial \theta(\bar{x})}{\partial x_i} v_i + \frac{\gamma}{\delta} f_i{}' \left(\frac{|\bar{x}_i|}{\delta}; -v_i \right) \geq 0, \qquad \forall\, v_i \in \mathbb{R}_+,$$

which is equivalent to $\dfrac{\partial \theta(\bar{x})}{\partial x_i} + \dfrac{\gamma}{\delta} f_i{}' \left(\dfrac{|\bar{x}_i|}{\delta}; -1 \right) \geq 0$. We omit the proof of the other two cases. The last assertion of proposition is an immediate consequence of Cui et al. [23, Proposition 4.1]. □

Remark 2 When each f_i is a piecewise linear-quadratic function, there is an equivalence of a local minimizer of problem (8) and a d-stationary point satisfying a "second-order directional stationarity condition". The latter enhanced stationarity condition is defined in terms of the "second-order directional derivatives" of f_i. For more details on the theory of piecewise linear-quadratic optimization, see [23] for details. The question of whether this theory can be sharpened when each f_i is additionally continuously differentiable, particularly for the MCP and SCAD regularizers, requires further investigation.

The following result establishes a necessary condition of a global minimizer of (1) that leads to a concept of a "pseudo-minimizer". This kind of "minimizers" turns out to be the limits of convergent sequences of d-stationary solutions to (8) as $\delta \downarrow 0$.

Proposition 3 *Let $X \subseteq \mathbb{R}^n$, $\theta : \mathbb{R}^n \to \mathbb{R}$, and $\gamma > 0$ be given. If $\bar{x}$ is a global minimizer of the problem:*

$$\underset{x \in X}{\textbf{minimize}}\ \theta(x) + \gamma \sum_{i=1}^{n} |\, x_i \,|_0, \tag{10}$$

then $\bar{x}$ is a global minimizer of the problem:

$$\underset{x \in X}{\textbf{minimize}}\ \theta(x) \quad \textbf{subject to}\ x_{\mathcal{A}_0(\bar{x})} = 0. \tag{11}$$

Proof Indeed, for any feasible solution x of (11), we have

$$\begin{aligned} \theta(x) &= \theta(x) + \gamma \sum_{i=1}^{n} |\, x_i \,|_0 - \gamma\, (\text{ number of nonzero components of } x) \\ &\geq \theta(\bar{x}) + \gamma \sum_{i=1}^{n} |\, \bar{x}_i \,|_0 - \gamma\, (\text{ number of nonzero components of } \bar{x}) \;=\; \theta(\bar{x}), \end{aligned}$$

where the inequality holds because $\bar{x}$ is a global solution of (10) and the number of nonzero components of x is no more than that of $\bar{x}$. □

We say that a vector $\bar{x} \in X$ is a *pseudo minimizer* of the problem (10) if $\bar{x}$ is a global minimizer of (11). The following result gives a necessary and sufficient condition for a feasible vector of (11) to be a pseudo minimizer when θ is convex and $X \triangleq [\,\ell, u\,]$.

Proposition 4 *Let $-\infty \leq \ell_i < 0 < u_i \leq \infty$ for all $i \in [n]$. If θ is convex, then a feasible vector $\bar{x}$ of (11) with $X \triangleq [\,\ell, u\,]$ is a pseudo minimizer of (10) if and only if the following three conditions hold for all $i \in [n]$,*

- *if $\bar{x}_i \neq 0$ and $\ell_i < \bar{x}_i < u_i$, then $\dfrac{\partial \theta(\bar{x})}{\partial x_i} = 0$;*
- *if $\bar{x}_i = \ell_i$, then $\dfrac{\partial \theta(\bar{x})}{\partial x_i} \geq 0$;*
- *if $\bar{x}_i = u_i$, then $\dfrac{\partial \theta(\bar{x})}{\partial x_i} \leq 0$.*

By our convention of the (positive) upper and (negative) lower bounds (to simplify the subsequent analysis), the origin is always a pseudo-minimizer. For problems where some such bounds are zero, the concept of a pseudo minimizer can be suitably modified. We omit such modifications. It turns out that the limits of d-stationary solutions of the δ-approximated problems (8) are pseudo minimizers of (1). This is the main reason for introducing this kind of a solution of (1). The convergence result below is in contrast to Le Thi et al. [40, Theorem 1] that pertains to global minimizers of the folded concave approximating problems. While such a convergence result about global minimizers is conceptually useful to justify the folded concave approximations, it is practically not very meaningful because such minimizers cannot be computed. This computational consideration motivates the following result that pertains to practically computable stationary solutions.

Proposition 5 *Let $-\infty \leq \ell_i < 0 < u_i \leq \infty$ for all $i \in [n]$. Let θ be strongly convex and let $\{\delta_k\}$ be a sequence of positive scalars converging to zero. For each k, let $\bar{x}^k$ be a d-stationary solution of (8) corresponding to δ_k. If for every $i \in [n]$, $f_i(t)$ is a constant for all $t > 0$ sufficiently large, then every accumulation point of the sequence $\{\bar{x}^k\}$ is a pseudo minimizer of (10).*

Proof Let $\bar{x}^\infty$ be the limit of a convergent subsequence, which without loss of generality we may take to be the entire sequence $\{\bar{x}^k\}$. We need to verify the three conditions in Proposition 4. For an index i such that $\bar{x}_i^\infty \neq 0$ and $\ell_i < \bar{x}_i^\infty < u_i$, we must have $\bar{x}_i^k \neq 0$ and $\ell_i < \bar{x}_i^k < u_i$ for all k sufficiently large; moreover, $\lim_{k\to\infty} \dfrac{|\bar{x}_i^k|}{\delta_k} = \infty$, which yields $f_i{}'\left(\dfrac{|\bar{x}_i^k|}{\delta_k}\right) = 0$ for all k sufficiently large, by the constancy assumption of $f_i(t)$ for $t > 0$ sufficiently large. Hence $\dfrac{\partial \theta(\bar{x}^k)}{\partial x_i} = 0$ for all such k, which yields $\dfrac{\partial \theta(\bar{x}^\infty)}{\partial x_i} = 0$. Consider next an index i such that $\bar{x}_i^\infty = \ell_i$. We then have $\bar{x}_i^k < 0 < u_i$ for all k sufficiently large. Moreover, either $\bar{x}_i^k = \ell_i$ for infinitely many k's, or $\bar{x}_i^k > \ell_i$

for all k sufficiently large. In either case, we can deduce that $\frac{\partial \theta(\bar{x}^\infty)}{\partial x_i} \geq 0$ by an argument similar to the previous case. Lastly, we can similarly show that for an index i such that $\bar{x}_i^\infty = u_i$, we must have $\dfrac{\partial \theta(\bar{x}^\infty)}{\partial x_i} \leq 0$. □

4 Z- and M-functions

In proving the linear (in n) number of steps of the main algorithm to be introduced in Sect. 6, the Z-property of the gradient $\nabla\theta$, which is motivated by the structure of the function θ in the applied pairwise separation problem (1), plays a prominent role. Specifically, the gradient $\nabla\theta : \mathbb{R}^n \to \mathbb{R}^n$ is a *Z-function*, also called an "off-diagonally antitone" function. As a generalization of a Z-matrix, which is a real square matrix with nonpositive off-diagonal entries, this class of vector functions was first introduced by Rheinboldt [52] and has been studied extensively in the context of complementarity problems [49,54]. We formally define a Z-function and related functions below. For two vectors a and b of the same dimension, the notation $a \leq b$ is meant to be componentwise.

Definition 6 [45] A function $F : \mathbb{R}^n \to \mathbb{R}^n$ is

- a *Z-function* if the univariate, scalar-valued function $t \mapsto F_i(x + te^j)$, where e^j is the j-th coordinate vector, is nonincreasing for all $i \neq j$ and all $x \in \mathbb{R}^n$;
- *inverse isotone* if $F(x) \leq F(y)$ implies $x \leq y$ for all x and y in $\mathbb{R}^n$;
- an *M-function* if it is an inverse isotone Z-function.
- *strongly monotone* if a constant $c > 0$ exists such that $(F(x) - F(y))^\top(x - y) \geq c\,(x - y)^\top(x - y)$ for all $x, y \in \mathbb{R}^n$.

For every index set $\beta \subseteq \{1, \ldots, n\}$ with complement α and for every vector $b_\beta \in \mathbb{R}^{|\beta|}$, the function $x_\alpha \mapsto F_\alpha(x_\alpha, b_\beta)$ is called a *principal subfunction* of F. It can be shown that every principal subfunction of an M-function is an M-function. It is well known that the gradient of a continuous differentiable, strongly convex function is strongly monotone. Moreover, by Moré and Rheinboldt [45, Theorem 3.8], a continuously, strongly monotone, Z-function must be inverse isotone, thus is an M-function. Moreover, it is well-known that a continuous, strongly monotone function must be surjective. A simple proof of this statement when F is the gradient of a continuously differentiable, strongly convex function is given in the proof below.

Proposition 7 *Let θ be defined in (1) and satisfy the stated assumptions. Then $F \triangleq \nabla\theta$ is a surjective M-function. Thus, for every vector $b \in \mathbb{R}^n$, a vector $a \in \mathbb{R}^n$ exists such that $F^{-1}[\,b, \infty\,) \subseteq [\,a, \infty\,)$.*

Proof For each i, observe that for all $(i, j) \in V$ and $(j, i) \in V$ with $j \neq i$,

$$\frac{\partial g_{ij}(a_{ij}x_i - b_{ij}x_j)}{\partial x_i} = a_{ij}\, g'_{ij}(a_{ij}x_i - b_{ij}x_j) \quad \text{and}$$
$$\frac{\partial g_{ji}(a_{ji}x_j - b_{ji}x_i)}{\partial x_i} = -b_{ji}\, g'_{ji}(a_{ji}x_j - b_{ji}x_i).$$

By the nonnegativity of the scalars $\{a_{ij}, b_{ij}\}_{(i,j)\in V}$ and the nondecreasing property of g'_{ij} and g'_{ji}, both due to the convexity of g_{ij} and g_{ji}, respectively, it follows that the above two partial derivatives are nonincreasing functions of x_j. Thus $\nabla\theta$ is a Z-function. By the aforementioned remark, it follows that $\nabla\theta$ is strongly monotone; the surjectivity of $\nabla\theta$ can be proved by noting that the equation $\nabla\theta(x) = b$ is equivalent to the problem $\textbf{minimize}_x\ \theta(x) - b^\top x$ which has a strongly convex objective, thus has a unique minimizer that satisfies the gradient equation. To prove the last assertion of the proposition, it suffices to note that, by surjectivity, for every vector $b \in \mathbb{R}^n$, there exists a vector $a \in \mathbb{R}^n$ such that $F(a) = b$; hence if $x \in F^{-1}[b, \infty)$, then $F(x) \geq b = F(a)$; The inverse isotonicity of F then yields $x \geq a$. □

5 Solving (8) with $\gamma = \delta = 1$: a synopsis and related literature

Having settled the convergence as $\delta \downarrow 0$, we proceed next to discuss the computation of a d-stationary solution of the problem (8) with a fixed δ which we take to be 1; we also take $\gamma = 1$ to simplify the notation. Thus, we consider:

$$\underset{\ell \leq x \leq u}{\textbf{minimize}}\ \varphi(x) \triangleq \theta(x) + \sum_{i=1}^{n} f_i(|\,x_i\,|). \tag{12}$$

Throughout θ is continuously differentiable and strongly convex such that $\nabla\theta$ is a Z-function; by the proof of Proposition 7, it follows that $\nabla\theta$ is a surjective M-function. Before proceeding, we should point out that as a stand-alone problem, there is to date no known algorithm that can compute a d-stationary solution of (12) when each f_i is a general concave function. Our main contribution is an algorithm with favorable computational complexity for the case where each f_i is piecewise smooth. This is a significant departure from the algorithms in [42,51] for structured difference-of-convex problems where only subsequential convergence is established.

The overall algorithm to be proposed for problem (12) consists of outer (Algorithm I) and inner (Algorithm II) loops which can be implemented (see Sect. 6.4) by a "greedy" procedure originated from the study of complementarity problems with Z-properties to preserve the favorable computational complexity. Each outer loop consists of solving a fixed-sign subproblem determined by a pair $(S, \bar{S})$ of complementary index sets whose union is $\{1, \ldots, n\}$:

$$\left.\begin{array}{l} \underset{\ell \leq x \leq u}{\textbf{minimize}}\ \ \theta(x) + \displaystyle\sum_{i=1}^{n} f_i(|\,x_i\,|) \\ \textbf{subject to } x_S \geq 0 \quad \text{and} \quad x_{\bar{S}} \leq 0 \end{array}\right\} \Leftrightarrow \left\{\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\textbf{minimize}} & \theta(x) + \displaystyle\sum_{i=1}^{n} f_i(|\,x_i\,|) \\ \textbf{subject to} & u_S \geq x_S \geq 0 \\ \textbf{and} & \ell_{\bar{S}} \leq x_{\bar{S}} \leq 0. \end{array}\right. \tag{13}$$

With each f_i being piecewise smooth (see Sect. 6.2), each inner loop solves a finite number of differentiable subproblems each defined by a "piece" of the sum function $\sum_{i=1}^n f_i$. Together, the two loops break down the nondifferentiability of the regularizer $\sum_{i=1}^n f_i(|\,x_i\,|)$ in two steps: (i) the outer loop deals with the positive and negative pieces of the absolute-value function, and (ii) the inner loop exploits the piecewise differentiability of f_i to define smooth subproblems that become the workhorse of the overall algorithm. Supported by the well-definedness results in Sects. 6.1 and 6.2, Theorem 11 shows that at most $n + 1$ subproblems of the kind (13) are solved. This explains the term "linear-step" in the title of this paper. When each f_i is differentiable, each subproblem (13) can be solved via its first-order Karush–Kuhn–Tucker conditions formulated as an upper-bounded nonlinear complementarity problem with a Z-function. This solution strategy is applicable to each differentiable subproblem in the piecewise smooth case. The literature for Z-structured complementarity problems starts with the work of Chandrasekaran [13] for the linear complementarity problem, extended to the nonlinear problem by Tamir [54], and further extended to problems with upper bounds by the last author in the paper [48]. A summary of this literature can be found in the Ph.D. thesis [49]. At the ground level, each of the cited algorithms solves a system of linear equations when θ is quadratic and each f_i is piecewise linear-quadratic (as in SCAD or MCP), and a system of nonlinear equations when θ is non-quadratic and/or some f_i are not piecewise linear-quadratic; in each case, there are at most $\mathrm{O}(n)$ many such equations to be solved (when the number of pieces of each f_i is independent of n). Putting together these complexity bounds for the two loops, the resulting algorithm for computing a directional stationary solution of (12) has a strongly polynomial complexity when θ is quadratic and each f_i is piecewise linear quadratic. In the general case, the proposed algorithm requires solving $\mathrm{O}(n^2)$ nonlinear equations each involving a subset of the n variables. More details of the computational efforts are described in Sect. 6.4. In addition to such favorable complexity, the iterations of the overall algorithm also yield descent in the objective function φ; see the last conclusion in Theorem 15 and the follow-up comment in the ensuing paragraph.

5.1 Related literature

A major distinction between the present work and the relevant literature on folded concave optimization is our emphasis of finiteness of the algorithms, which yields the strongly polynomial complexity for problems with quadratic loss and piecewise linear-quadratic regularizers. In what follows, we contrast our algorithmic complexity results with those in several most relevant papers. In [10], the authors study the ℓ_0-problem with a quadratic loss function; a specialized iterative thresholding algorithm that depends on the parameter γ was presented and analyzed. A most notable result therein is that under the condition that $\|\nabla^2\theta\|_2 < 1$ where $\nabla^2\theta$ is the Hessian of the quadratic function θ, the proposed algorithm produces a sequence of iterates that converge to a local minimizer of the problem. The cornerstone of the convergence is based on the key fact that under the norm condition, a minimizer of the ℓ_0-problem is a subset of a γ-dependent fixed-point mapping. Under a further assumption on the quadratic loss function θ, the rate of convergence is obtained. It is worth pointing out that it is not clear whether the algorithm in [10] is applicable to a weighted ℓ_0-problem with unequal weights on the separable sparsity terms. As noted in the paper [33], the equal-weighted and unequal-weighted versions of the ℓ_0-problem can be quite different. In contrast, with the use of the surrogate sparsity functions f_i, these two versions of the approximated problem can be treated in a unified way. In [41], the authors studied the problem (13) with differentiable functions f_i and used SCAD and MCP as the primary examples to illustrate the results. When it comes to algorithms, the authors mentioned on page 235 of the reference that they applied a potential reduction algorithm [57] for quadratic programming (QP) to compute a "second-order KKT solution" of the problem. Such an algorithm is of the iterative kind and computes only an approximate solution with complexity inversely proportional to the accuracy.

Capped folded concave approximations have been studied in two recent papers [9,47] as an approximation of the ℓ_0-function. The principal method for the solution of the resulting optimization problem (12) employed therein is smoothing [15]. Thus in terms of the original ℓ_0-minimization problem, there are two levels of approximation: using a folded concave approximation of the sparsity function, and a further approximation of the folded concave function by smoothing. In contrast, the method in this paper involves only one layer of approximation via the folded concave functions. In particular, for a piecewise affine (such as the capped ℓ_1-function) or a piecewise linear-quadratic (such as SCAD or MCP) regularizer appended to a quadratic loss function with a Stieltjes matrix, our overall algorithm is a finite method with a strongly polynomial complexity. This very favorable computational complexity is believed to be the first of its kind in the area of nonconvex sparsity minimization. As an initial extension of this study on a stand-alone problem (8) with a fixed pair (γ, δ), we briefly touch on the complete parametric ℓ_1-problem with δ fixed in the last Sect. 7. The goal there is to show that the entire solution path of certain instances of the problem (8) can be followed with strongly polynomial computational complexity also.

6 The linear-step two-loop algorithm

As a prerequisite, we present the following result similar to Proposition 1 that gives a necessary and sufficient condition for a directional stationary point of (13). No proof is needed. We recall the index set $\mathcal{A}_0(\bar{x})$ defined in (9).

Proposition 8 *A feasible vector $\bar{x}$ is a directional stationary point of (13) if and only if*

- f_i *is differentiable at* $|\bar{x}_i|$ *for all* i *such that* $\bar{x}_i \neq 0$ *and* $\ell_i < \bar{x}_i < u_i$*;*
- $\dfrac{\partial \theta(\bar{x})}{\partial x_i} + sgn(\bar{x}_i)\, f_i{}'(\,|\bar{x}_i|\,) = 0$ *for all* i *such that* $\bar{x}_i \neq 0$ *and* $\ell_i < \bar{x}_i < u_i$*;*
- $\dfrac{\partial \theta(\bar{x})}{\partial x_i} + f_i{}'(0; 1) \geq 0$ *for all* $i \in \mathcal{A}_0(\bar{x}) \cap S$*;*
- $-\dfrac{\partial \theta(\bar{x})}{\partial x_i} + f_i{}'(0; 1) \geq 0$ *for all* $i \in \mathcal{A}_0(\bar{x}) \cap \bar{S}$*;*
- $\dfrac{\partial \theta(\bar{x})}{\partial x_i} + f_i{}'(\,|\bar{x}_i|; -1\,) \geq 0$ *for all* i *such that* $\bar{x}_i = \ell_i$*;*
- $-\dfrac{\partial \theta(\bar{x})}{\partial x_i} + f_i{}'(\,|\bar{x}_i|; -1\,) \geq 0$ *for all* i *such that* $\bar{x}_i = u_i$*.*

Notice that a d-stationary solution of (13) can be expected to contain zero components where the absolute-value function is not differentiable. The following consequence of the proposition is immediate.

Corollary 9 *Let $\bar{x}$ be a directional stationary point of* (13) *such that $\bar{x}_i < 0$ for all $i \in \bar{S}$. Then $\bar{x}$ is a directional stationary point of* (12) *if and only if*

$$\frac{\partial \theta(\bar{x})}{\partial x_i} \leq f_i{}'(0;1) \quad \forall\, i \text{ such that } \bar{x}_i = 0.$$

Based on the above corollary, we present the outer loop of the method for computing a d-stationary solution of (12).

Algorithm I: Outer loop for computing a d-stationary point of (12)

Initialization. Let $S^0 = \{1, \cdots, n\}$ and $\nu = 0$.

General iteration. Let $x^{\nu+1}$ be a directional stationary solution of the fixed-signed subproblem (13) with the pair $(S^\nu, \bar{S}^\nu)$ such that $x^{\nu+1} \leq x^\nu$ (this condition is not needed when $\nu = 0$). Let

$$S^{\nu+1} \triangleq S^\nu \setminus \underbrace{\left\{ i \in S^\nu \mid x_i^{\nu+1} = 0 < \frac{\partial \theta(x^{\nu+1})}{\partial x_i} - f_i{}'(0;1) \right\}}_{\text{denoted} E^\nu}.$$

If $S^{\nu+1} = S^\nu$, then the iterate $x^{\nu+1}$ is a d-stationary solution of (12); if $S^{\nu+1} = \emptyset$, then the next iterate $x^{\nu+2}$ is d-stationary solution of (12). So the algorithm terminates successfully if either case occurs. Otherwise, let $\nu \leftarrow \nu + 1$ and return to the general iteration.

Provided that each iterate is well defined, Algorithm I will terminate in no more than $n+1$ visits to the general iteration. Postponing the well-definedness proof until later, we first establish this claim of linear-step termination of the algorithm. This is accomplished with the aid of a lemma that gives an immediate consequence of the isotonicity of the iterates; i.e, $x^{\nu+1} \leq x^\nu$. The lemma asserts that if at any point in the algorithm a coordinate is negative ($x_i^\nu < 0$) and $i \notin S^{\nu-1}$, then the coordinate will remain negative throughout the rest of the algorithm. Note that the condition $x_i^1 < 0$ for $i \notin S^0$ is vacuously true because S^0 is the full index set $[n]$.

Lemma 10 *Suppose $x_i^\nu < 0$ for all $i \notin S^{\nu-1}$ for a given $\nu \geq 1$. Then $x_i^{\nu+1} < 0$ for all $i \notin S^\nu$.*

Proof We note that

$$S^\nu = S^{\nu-1} \setminus E^{\nu-1} \quad \text{and} \quad \bar{S}^\nu = \bar{S}^{\nu-1} \cup E^{\nu-1}. \tag{14}$$

Since $x^{\nu+1} \leq x^\nu$, it follows that $x_i^{\nu+1} < 0$ for all $i \notin S^{\nu-1}$. It remains to show that $x_i^{\nu+1} < 0$ for all $i \in E^{\nu-1}$. All indices of the latter kind are in $\bar{S}^\nu$; thus $x_i^{\nu+1} \leq 0$ for all such indices. Suppose that $x_{\bar{i}}^{\nu+1} = 0$ for some $\bar{i} \in E^{\nu-1}$. By the d-stationarity of

$x^{\nu+1}$ for the fixed-signed subproblem (13) with the pair $(S^{\nu}, \bar{S}^{\nu})$, we have

$$f_i{}'(0;1) \geq \frac{\partial\theta(x^{\nu+1})}{\partial x_{\bar{i}}} \geq \frac{\partial\theta(x^{\nu})}{\partial x_{\bar{i}}} > f_i{}'(0;1),$$

where the second inequality holds by the Z-property of $\nabla\theta$ because $x^{\nu+1} \leq x^{\nu}$ and $x_{\bar{i}}^{\nu+1} = x_{\bar{i}}^{\nu}$. The above inequalities yield a contradiction. □

We state and prove the n-step convergence of Algorithm I in the result below.

Theorem 11 *Let θ be continuously differentiable and strongly convex such that $\nabla\theta$ is a Z-function. Suppose that the bounds satisfy: $-\infty \leq \ell_i < 0 < u_i \leq \infty$ for all $i \in [n]$. Provided that each $x^{\nu+1}$ is well-defined (see Sects. 6.1 and 6.2), then in no more than $n+1$ iterations, Algorithm I will terminate with an iterate that is a d-stationary point of the problem (12).*

Proof Since Algorithm I is initiated with $S^0 = [n]$, by Lemma 10, it follows inductively that $x_i^{\nu+1} < 0$ for all $i \notin S^{\nu}$ for all ν. By Corollary 9, a d-stationary solution $\bar{x}$ of (13) that has $\bar{x}_i < 0$ for all $i \in \bar{S}$ is a d-stationary point of (12) if and only if

$$\frac{\partial\theta(\bar{x})}{\partial x_i} \leq f_i{}'(0;1) \quad \forall\, i \in S \text{ such that } \bar{x}_i = 0.$$

So if $E^{\nu} = \emptyset$ at some iteration ν, then the above requirement clearly holds for the pair $(S^{\nu}, \bar{S}^{\nu})$ and the d-stationary solution $x^{\nu+1}$ of (13) corresponding to this pair. Thus $x^{\nu+1}$ is a d-stationary point of (12) in this case. On the other hand, if $E^{\nu} = S^{\nu}$ so that $S^{\nu+1} = \emptyset$, then the above condition holds vacuously at the next iteration $\nu+1$. Since the set S^{ν} always decreases by at least one element if $E^{\nu} \neq \emptyset$, the n-iteration termination of the algorithm follows readily. □

We prove the existence of the desired iterate $x^{\nu+1}$ in two cases of the concave functions f_i, for $i \in [n]$: (a) differentiable and Lipschitz continuous, and (b) piecewise smooth; (details of the assumptions will be clearly stated). In the former case, the proof relies on a least-element property of a special d-stationary solution of the fixed-signed subproblem (13) and the update (14) of the pair $(S^{\nu}, \bar{S}^{\nu})$ from the preceding pair; the latter case is similar. Subsequently, we will comment on the practical computation of each iterate. Note that these two classes of concave functions do not include the power function with an exponent less that unity; i.e., the function $t(>0) \mapsto t^p$ for $p \in (0,1)$ (due to the non-Lipschitz behavior near the origin); nevertheless, the ε-smoothed power function: $t(>0) \mapsto (t+\varepsilon)^p$ for $p \in (0,1)$ and $\varepsilon > 0$ belongs to the differentiable class. The power regularizer and its smoothing have been studied extensively; see e.g. [17–19]. One theoretical drawback of this family of power regularizers is the lack of favorable computational complexity, even with a quadratic loss function θ; see [17, Theorems 5 and 6]. This drawback can be contrasted with the n-step termination of Algorithm I and the subsequent discussion of the efficient computation of $x^{\nu+1}$.

6.1 Existence of $x^{\nu+1}$: the differentiable case

Throughout this subsection, we assume, in addition to the blanket assumption stated in Sect. 3, that

- **(diff + Lip)** each f_i is Lipschitz continuous on the closed interval $X_i \triangleq [\,0, \max(-\ell_i, u_i)\,]$ and differentiable therein except at the origin.

Classes of functions satisfying this assumption include SCAD, MCP, and suitable smoothings of nondifferentiable regularizers, such as that of the aforementioned power function with an exponent less than one. In fact, this assumption is consistent with the definition of a "folded concave penalty" employed in [41]. For a given pair $(S, \bar{S})$, we define the following set consisting of vectors in $\mathbb{R}^n$ satisfying three conditions:

$$\mathcal{Z}_S \triangleq \left\{ \begin{array}{l} x = (x_S, x_{\bar{S}}) \text{ satisfies} \\ \text{(a) } 0 \le x_S \le u_S \text{ and } 0 \ge x_{\bar{S}} \ge \ell_{\bar{S}} \\ \text{(b) } \forall\, i \in S : x_i < u_i \Rightarrow \dfrac{\partial \theta(x)}{\partial x_i} + f_i{}'(x_i; 1) \ge 0 \\ \text{(c) } \forall\, i \in \bar{S} : x_i < 0 \Rightarrow \dfrac{\partial \theta(x)}{\partial x_i} + f_i{}'(-x_i; -1) \ge 0 \end{array} \right\},$$

whose construction is motivated by the classical case of an upper-bounded linear complementarity problem with a Z-matrix; see [48]. It is not difficult to deduce from the necessary and sufficient conditions in Proposition 8 that the set $\mathcal{Z}_S$ contains all d-stationary solutions of the problem (13).

For ease of reference in the proof of Proposition 13, we cite a combined statement of Rockafellar [53, Theorems 24.1 and 25.4] pertaining to one-sided directional derivatives of univariate convex/concave functions. Under the differentiability assumption, the equalities (15) are for $\bar{t} = 0$ only while the second statement of the lemma applies to $\bar{t} \neq 0$.

Lemma 12 *Let f be a convex/concave function on an open interval containing $\bar{t}$. It holds that*

$$f'(\bar{t}; 1) = \pm \lim_{s \downarrow \bar{t}} f'(s; \pm 1) \quad \textit{and} \quad f'(\bar{t}; -1) = \mp \lim_{s \uparrow \bar{t}} f'(s; \pm 1). \tag{15}$$

Moreover, if f is differentiable at $\bar{t}$, then $f'(\cdot; \pm 1)$ is continuous at $\bar{t}$.

Under the afore-stated differentiability and Lipchitz continuity assumption, the following proposition asserts the well-definedness of the iterate $x^{\nu+1}$ as the least element of the set $\mathcal{Z}_{S^\nu}$.

Proposition 13 *Let θ be continuously differentiable and strongly convex such that $\nabla\theta$ is a Z-function. Suppose that Assumption (diff + Lip) holds and that the bounds satisfy: $-\infty \le \ell_i < 0 < u_i \le \infty$ for all $i \in [n]$. Let x^ν ($\nu \ge 1$) be an iterate in Algorithm* I. *Then x^ν belongs to the set $\mathcal{Z}_{S^\nu}$. Moreover, the set $\mathcal{Z}_{S^\nu}$ has a componentwise least element, say $\bar{x}$, which must necessarily satisfy $\bar{x} \le x^\nu$ and be a d-stationary solution of the fixed-signed subproblem* (13) *with the pair $(S^\nu, \bar{S}^\nu)$.*

Proof By definition, x^{ν} is a d-stationary point of (13) corresponding to the pair $(S^{\nu-1}, \bar{S}^{\nu-1})$. As such, $x^{\nu}_{S^{\nu-1}} \geq 0$; thus $x^{\nu}_{S^{\nu}} \geq 0$; similarly, $x^{\nu}_{\bar{S}^{\nu}} \leq 0$. We next show that x^{ν} satisfies the two properties (b) and (c) of elements in $\mathcal{Z}_{S^{\nu}}$. To show (b), let $i \in S^{\nu}$ be such that $x^{\nu}_i < u_i$. Then $i \in S^{\nu-1}$. It follows from the d-stationarity of x^{ν} that $\dfrac{\partial\theta(x^{\nu})}{\partial x_i} + f_i{}'(x^{\nu}_i) = 0$ if $x^{\nu}_i \neq 0$ (by the differentiability assumption of f_i) and $\dfrac{\partial\theta(x^{\nu})}{\partial x_i} + f_i{}'(0; 1) \geq 0$ if $x^{\nu}_i = 0$. Similarly, to show (c), let $i \in \bar{S}^{\nu}$ be such that $x^{\nu}_i < 0$. Then $i \in \bar{S}^{\nu-1}$. By d-stationarity, x^{ν} satisfies: $\dfrac{\partial\theta(x^{\nu})}{\partial x_i} - f_i{}'(|x^{\nu}_i|) = 0$ if $x^{\nu}_i > \ell_i$ (by the same differentiability assumption) and $\dfrac{\partial\theta(x^{\nu})}{\partial x_i} + f_i{}'(|x^{\nu}_i|; -1) \geq 0$ if $x^{\nu}_i = \ell_i$. This completes the proof that $x^{\nu} \in \mathcal{Z}_{S^{\nu}}$.

To show that the set $\mathcal{Z}_{S^{\nu}}$ has a least element, we need to show three things according to Pang [48, Theorem 2.2]:

(i) the set $\mathcal{Z}_{S^{\nu}}$ is closed;
(ii) if x and x' are two vectors in $\mathcal{Z}_{S^{\nu}}$, then their componentwise minimum $y \triangleq \min(x, x')$ is also an element in this set; and
(iii) vectors in $\mathcal{Z}_{S^{\nu}}$ are bounded below componentwise.

We start with the closedness of $\mathcal{Z}_{S^{\nu}}$. Let $\{x^k\} \subseteq \mathcal{Z}_{S^{\nu}}$ converge to x^{∞}. We need to show that x^{∞} satisfies the defining conditions (b) and (c) in the set $\mathcal{Z}_{S^{\nu}}$. For an index $i \in S^{\nu}$ such that $x^{\infty}_i < u_i$, we have $x^k_i < u_i$ for all k sufficiently large. Hence $\dfrac{\partial\theta(x^k)}{\partial x_i} + f_i{}'(|x^k_i|; 1) \geq 0$ for such k. There are two cases to consider: either $x^{\infty}_i \neq 0$ or $x^{\infty}_i = 0$. In the former case, f_i is differentiable at x^{∞}_i and $|x^k_i|$ for all k sufficiently large; thus by Lemma 12, we easily deduce $\dfrac{\partial\theta(x^{\infty})}{\partial x_i} + f_i{}'(|x^{\infty}_i|; 1) \geq 0$. In the latter case, we may assume without loss of generality that $x^k_i > 0$ for all k sufficiently large. It then follows that f_i is differentiable at $|x^k_i| = x^k_i$; moreover, by Lemma 12,

$$\lim_{k\to\infty} f_i{}'(x^k_i) = \lim_{k\to\infty} f_i{}'(|x^k_i|; 1) = f_i{}'(0; 1).$$

Consequently, we obtain $\dfrac{\partial\theta(x^{\infty})}{\partial x_i} + f_i{}'(|x^{\infty}_i|; 1) \geq 0$ also. This shows that x^{∞} satisfies condition (b) in $\mathcal{Z}_{S^{\nu}}$. Let $i \in \bar{S}^{\nu}$ be such that $x^{\infty}_i < 0$. We then have $\dfrac{\partial\theta(x^k)}{\partial x_i} + f_i{}'(|x^k_i|; -1) \geq 0$ for all k sufficiently large. Again, there are two cases to consider: $x^{\infty}_i = \ell_i$ or $x^{\infty}_i > \ell_i$. Similarly to the above, it suffices to consider the case where $x^k_i > \ell_i = x^{\infty}_i$ for all k sufficiently large. By the differentiability of f_i at $|x^k_i|$ for all k sufficiently large, we have

$$f_i{}'(|x^{\infty}_i|; -1) = \lim_{k\to\infty} f_i{}'(|x^k_i|; -1).$$

Hence it follows that $\dfrac{\partial\theta(x^\infty)}{\partial x_i} + f_i{}'(\,|x_i^\infty|; -1\,) \geq 0$. The proof of the closedness of $\mathcal{Z}_{S^\nu}$ is completed.

To prove (ii), we need to verify the three conditions (a)–(c) in the definition of an element in $\mathcal{Z}_{S^\nu}$ for the vector $y \triangleq \min(x, x')$. Since each component y_i is equal to either x_i or x_i', condition (a) is obvious. For the remaining two conditions (b) and (c), consider any index i and say $y_i = x_i$. We then have, by the Z-property of $\nabla\theta$, $\dfrac{\partial\theta(y)}{\partial x_i} \geq \dfrac{\partial\theta(x)}{\partial x_i}$; the desired implications in conditions (b) and (c) can easily be seen to hold for y.

To complete the proof of the existence of a least vector in $\mathcal{Z}_{S^\nu}$, it remains to show the bounded below property (iii) of elements in $\mathcal{Z}_{S^\nu}$. By the Lipschitz continuity of f_i on X_i, there exists a constant $L > 0$ such that

$$\max_{1 \leq i \leq n} \left\{ \max\left(|\, f_i{}'(0; 1)\,|;\ \sup_{\xi \in X_i} \left|\, f_i{}'(\,\xi; \pm 1\,)\,\right| \right) \right\} \leq L.$$

For any index set α equal to the disjoint union of two subsets α_1 and α_2, consider the subfunction:

$$F_\beta : x_\beta \mapsto \left(\frac{\partial\theta(x_\beta, u_{\alpha_1}, 0_{\alpha_2})}{\partial x_i} \right)_{i \in \beta},$$

where β is the complement of α in $\{1, \ldots, n\}$. By Proposition 7, it follows that there exists a vector $a \in \mathbb{R}^n$ such that for all such pairs (α, β) of index sets and all $x \in \mathbb{R}^n$,

$$F_\beta(x_\beta) \geq -L\,\mathbf{1}_\beta \;\Rightarrow\; x_\beta \geq a_\beta. \tag{16}$$

For any vector $x \in \mathcal{Z}_{S^\nu}$ let

$$\alpha_1 \triangleq \left\{ i \in S^\nu \mid x_i = u_i \right\} \quad \text{and} \quad \alpha_2 \triangleq \left\{ i \in \bar{S}^\nu \mid x_i = 0 \right\},$$

and let β be the complement of $\alpha = \alpha_1 \cup \alpha_2$. We can then deduce by conditions (b) and (c) in $\mathcal{Z}_{S^\nu}$ and the bound (16) that x is bounded below.

Finally, to show that the least vector $\bar{x}$ of the set $\mathcal{Z}_{S^\nu}$ is a d-stationary solution of the fixed-signed subproblem (13) with the pair $(S^\nu, \bar{S}^\nu)$, we need to show that $\bar{x}$ satisfies the conditions in Proposition 8. Consider an index i for which $\bar{x}_i \neq 0$ and $\ell_i < \bar{x}_i < u_i$. If such an index i belongs to S^ν, then $\bar{x}_i > 0$ and we need to show that $\dfrac{\partial\theta(\bar{x})}{\partial x_i} + f_i{}'(\,|\bar{x}_i|\,) = 0$. Indeed, if this is not true, then by condition (b) in $\mathcal{Z}_{S^\nu}$, we must have $\dfrac{\partial\theta(\bar{x})}{\partial x_i} + f_i{}'(\,|\bar{x}_i|\,) > 0$. Since f' is continuous at $|\bar{x}_i|$, it follows that for a sufficiently small $\delta > 0$, the vector $x^\delta \triangleq \bar{x} - \delta e^i$, where e^i is the ith coordinate vector, satisfies $\dfrac{\partial\theta(x^\delta)}{\partial x_i} + f_i{}'(\,|x_i^\delta|\,) > 0$; moreover, by the Z-property of $\nabla\theta$, x^δ can be shown to belong to $\mathcal{Z}_{S^\nu}$. This contradicts the least property of $\bar{x}$. If such an index i

belongs to $\bar{S}^\nu$, then $\bar{x}_i < 0$ and we need to show that $\dfrac{\partial\theta(\bar{x})}{\partial x_i} - f_i{}'(\,|\bar{x}_i|\,) = 0$. Again if this is not true, then by condition (c) in $\mathcal{Z}_{S^\nu}$, we must have $\dfrac{\partial\theta(\bar{x})}{\partial x_i} - f_i{}'(\,|\bar{x}_i|\,) > 0$. By a similar perturbation of $\bar{x}$, we can obtain a contradiction to the least property of $\bar{x}$. Next consider an index $i \in S^\nu$ for which $\bar{x}_i = 0$. Then $\dfrac{\partial\theta(\bar{x})}{\partial x_i} + f_i{}'(\,0;\,1\,) \geq 0$ by condition (b) in $\mathcal{Z}_{S^\nu}$. If $i \in \bar{S}^\nu$ and $\bar{x}_i = 0$ but $-\dfrac{\partial\theta(\bar{x})}{\partial x_i} + f_i{}'(\,0;\,1\,) < 0$, then since $f_i{}'(\,0;\,1\,) = \lim_{\delta\downarrow 0} f_i{}'(\,\delta;\,1\,)$ by Lemma 12, it follows that for a sufficiently small $\delta > 0$, the vector $x^\delta \triangleq \bar{x} - \delta e^i$ satisfies $-\dfrac{\partial\theta(x^\delta)}{\partial x_i} + f_i{}'(\,|x_i^\delta|;\,1\,) < 0$; or equivalently

$$0 < \frac{\partial\theta(x^\delta)}{\partial x_i} - f_i{}'(\,|x_i^\delta|;\,1\,) = \frac{\partial\theta(x^\delta)}{\partial x_i} + f_i{}'(\,|x_i^\delta|;\,-1\,),$$

where the latter equality is by the differentiability of f_i at $|x_i^\delta| = \delta > 0$. By the Z-property of $\nabla\theta$, we can again show that the vector $\bar{x}^\delta$ belongs to $\mathcal{Z}_{S^\nu}$, which is a contradiction by the least property of $\bar{x}$. For the two remaining cases: $\bar{x}_i \in \{\ell_i, u_i\}$, the case $\bar{x}_i = \ell_i$ can be argued as for the case with $i \in \bar{S}^\nu$ and $\bar{x}_i < 0$. For the case $\bar{x}_i = u_i$, we need to show that $\frac{\partial\theta(\bar{x})}{\partial x_i} + f_i{}'(\bar{x}_i) \leq 0$. If this is not true, then by decreasing $\bar{x}_i$ slightly and keeping the other components unchanged, the Z-property of $\nabla\theta$ yields a contradiction to the least property of $\bar{x}$. $\square$

6.2 Existence of $x^{\nu+1}$: the piecewise case

In this subsection, we consider the case where each f_i is a piecewise smooth function with a finite number of nondifferentiable points. To start, throughout this section, we assume for simplicity of notation that $-\ell_i = u_i$ for each $i \in [n]$. Moreover, there is a partition of the interval $[\,0, u_i\,]$: for some integer $K_i \geq 1$,

$$0 = \xi_{i0} < \xi_{i1} < \cdots < \xi_{i\,K_i} < \xi_{i\,K_i+1} \triangleq u_i \leq \infty \tag{17}$$

and an associated family of functions $\{\phi_{ik}(\xi)\}_{k=1}^{K_i+1}$ where each $\phi_{ik} : [\,0, u_i\,] \to \mathbb{R}$ for $k = 1, \ldots, K_i + 1$ is concave, Lipschitz continuous, and continuously differentiable such that $f_i(\xi) = \min\limits_{1\leq k\leq K_i+1} \phi_{ik}(\xi)$ for all $\xi \in [\,0, u_i\,]$; moreover, $\{\xi_{ik}\}_{i=1}^{K_i}$ are the breakpoints of f_i in the interval $(\,0, u_i\,)$ such that

- *Piece orderings:* $f_i(\xi) = \phi_{ik}(\xi)$ for all $\xi \in I_{ik} \triangleq [\,\xi_{ik-1}, \xi_{ik}\,]$; and the slopes satisfy:
- *Piecewise dominance:* $\phi'_{ik}(\xi) \geq \phi'_{ij}(\xi)$ for all $k = 1, \ldots, K_i$, $j = k+1, \ldots, K_i + 1$ and all $\xi \in [0, u_i\,]$; and
- *Strict dominance at breakpoints:* $\phi'_{ik}(\xi_{ik}) > \phi'_{ik+1}(\xi_{ik})$ for all $k = 1, \ldots, K_i$.

The latter two properties together yield $\phi'_{ik}(\xi_{ik}) > \phi'_{ij}(\xi_{ik})$ for all $k \in [K_i]$ and $j = k+1, \ldots, K_i + 1$. Noting that

$$\begin{aligned} -\phi'_{ik}(\xi_k) = f_i{}'(\xi_k; -1) \quad &\text{and} \quad \phi'_{ik+1}(\xi_k) = f_i{}'(\xi_k; 1), \qquad k \in [K_i] \\ \phi'_{i1}(0) = f_i{}'(0; 1) \quad &\text{and} \quad \phi'_{iK_1+1}(u_i) = -f_i{}'(u_i; -1), \end{aligned} \tag{18}$$

it follows that the only nondifferentiable points of f_i are at the end points of the intervals I_{ik} for all $k \in [K_i + 1]$; in particular, there are at most K_i of them in the open interval $(0, u_i)$. The strict dominance at the breakpoints is what distinguishes this nondifferentiable case from the previous differentiable case. For instance, both the MCP and SCAD belong to the latter case, whereas the capped ℓ_1 and any piecewise, non-C^1 approximation of a concave nondecreasing function belong to former case discussed here.

The existence proof of $x^{\nu+1}$ is constructive. The workhorse of the construction is the following differentiable optimization problem: for a given tuple $\kappa \triangleq \{k_i\}_{i=1}^n \subseteq \prod_{i=1}^n [K_i + 1]$,

$$\begin{aligned} &\underset{x \in \mathbb{R}^n}{\textbf{minimize}} \ \ \theta(x) + \sum_{i \in S} \phi_{ik_i}(x_i) + \sum_{i \in \bar{S}} \phi_{ik_i}(-x_i) \\ &\textbf{subject to} \ u_S \geq x_S \geq 0 \quad \text{and} \quad \ell_{\bar{S}} \leq x_{\bar{S}} \leq 0. \end{aligned} \tag{19}$$

With nonconvex functions $\{\phi_{ik}\}$, the above problem has a nonconvex objective. Part of the analysis is to show that (19) has a particular least-element solution that can be computed efficiently. Subsequently, we will provide some detailed discussion for the computation of such a solution. Associated with the problem (19), which is of the type (13) with each f_i being differentiable, we define the set

$$\mathcal{Z}_S^\kappa \triangleq \left\{ \begin{array}{l} x = (x_S, x_{\bar{S}}) \text{ satisfies} \\ \text{(a) } 0 \leq x_S \leq u_S \text{ and } 0 \geq x_{\bar{S}} \geq \ell_{\bar{S}} \\ \text{(b) } \forall\, i \in S: \ x_i < u_i \Rightarrow \dfrac{\partial \theta(x)}{\partial x_i} + \phi'_{ik_i}(x_i) \geq 0 \\ \text{(c) } \forall\, i \in \bar{S}: \ x_i < 0 \Rightarrow \dfrac{\partial \theta(x)}{\partial x_i} - \phi'_{ik_i}(-x_i) \geq 0 \end{array} \right\}.$$

By Proposition 13 applied to the differentiable functions $\{\phi_{ik_i}\}_{i=1}^n$, it follows that this set, if nonempty, has a least element that is a d-stationary solution of (19). We call such an element, which must be unique, the least d-stationary solution of the latter problem.

Given a vector $\bar{x} \in [\ell, u]$, we define a unique tuple $\bar{\kappa} \triangleq \{\bar{k}_i\}_{i=1}^n \subseteq \prod_{i=1}^n [K_i + 1]$ by:

$$\bar{k}_i = \begin{cases} k & \text{if } i \in S \text{ and } 0 < \bar{x}_i \in (\xi_{ik-1}, \xi_{ik}] \text{ for some } k \in [K_i + 1] \\ 1 & \text{if } i \in S \text{ and } \bar{x}_i = 0 \\ k & \text{if } i \in \bar{S} \text{ and } \bar{u}_i > |\bar{x}_i| \in [\xi_{ik-1}, \xi_{ik}) \text{ for some } k \in [K_i + 1] \\ K_i + 1 & \text{if } i \in \bar{S} \text{ and } |\bar{x}_i| = u_i. \end{cases}$$

This definition yields a necessary and sufficient condition for a vector to be a d-stationary point of (13) in terms of the differentiable optimization subproblem defined by this tuple of indices.

Proposition 14 *Let $\bar{x}$ be a feasible point of (19) with the zero-index set $\mathcal{A}_0(\bar{x})$ defined in (9). Then $\bar{x}$ is a d-stationary point of (13) if and only if the following conditions hold:*

- *f_i is differentiable at $|\bar{x}_i|$ for all i such that $\bar{x}_i \neq 0$ and $\ell_i < \bar{x}_i < u_i$;*
- *$\bar{x}$ is a d-stationary point of (19) defined by the unique tuple $\{\bar{k}_i\}_{i=1}^n$ of indices determined by $\bar{x}$.*

Thus, if $\bar{x}$ is a d-stationary point of (19) defined by the unique tuple $\{\bar{k}_i\}_{i=1}^n$ of indices determined by $\bar{x}$, then $\bar{x}$ is a stationary point of (13) if and only if $|\bar{x}_i|$ is not equal to any of the breakpoints of the partition (17) that lie within the interval $(0, u_i)$.

Proof Similar to Proposition 8, assuming the differentiability condition on f_i at $|\bar{x}_i|$, we may deduce that $\bar{x}$ is a d-stationary point of (19) defined by the unique tuple $\{\bar{k}_i\}_{i=1}^n$ if and only if

- $\dfrac{\partial\theta(\bar{x})}{\partial x_i} + \phi'_{i\bar{k}_i}(\bar{x}_i) = 0$ if $0 < \bar{x}_i < u_i$;
- $\dfrac{\partial\theta(\bar{x})}{\partial x_i} - \phi'_{i\bar{k}_i}(-\bar{x}_i) = 0$ if $\ell_i < \bar{x}_i < 0$;
- $\dfrac{\partial\theta(\bar{x})}{\partial x_i} + \phi'_{i\bar{k}_i}(0) \geq 0$ for all $i \in \mathcal{A}_0(\bar{x}) \cap S$;
- $-\dfrac{\partial\theta(\bar{x})}{\partial x_i} + \phi'_{i\bar{k}_i}(0) \geq 0$ for all $i \in \mathcal{A}_0(\bar{x}) \cap \bar{S}$;
- $\dfrac{\partial\theta(\bar{x})}{\partial x_i} - \phi'_{i\bar{k}_i}(-\bar{x}_i) \geq 0$ for all i such that $\bar{x}_i = \ell_i$;
- $-\dfrac{\partial\theta(\bar{x})}{\partial x_i} - \phi'_{i\bar{k}_i}(\bar{x}_i) \geq 0$ for all i such that $\bar{x}_i = u_i$.

The proof of the proposition can be completed by comparing the above conditions with those in Proposition 8 and by the expressions in (18). □

The algorithm below pertains to the pair $(S^\nu, \bar{S}^\nu)$ at the outer iteration $\nu + 1$ of Algorithm I. Associated with this pair is an iterate x^ν from the previous iteration which is a d-stationary solution of the problem (13) corresponding to the pair $(S^{\nu-1}, \bar{S}^{\nu-1})$. By the definition (14) of the pair $(S^\nu, \bar{S}^\nu)$ which involves the set

$$E^{\nu-1} \triangleq \left\{ i \in S^{\nu-1} \mid x_i^\nu = 0 < \frac{\partial\theta(x^\nu)}{\partial x_i} - f_i{}'(0; 1) \right\},$$

it follows that $x^\nu_{S^{\nu-1}} \geq 0$ and $x^\nu_{\bar{S}^{\nu-1}} < 0$. While the former property is obvious because $S^\nu \subseteq S^{\nu-1}$, the latter property holds by Lemma 10 and the initialization $S^0 = [n]$ of Algorithm I.

Theorem 15 *Let each $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ be a continuous, concave, function satisfying $f_i(0) = 0$ and $f_i{}'(0; 1) \in (0, \infty)$. Let $\{\phi_{ik}\}_{k=1}^{K_i+1}$ be the family of element functions*

Algorithm II: construction of $x^{\nu+1}$ for piecewise regularizers

Initialization. Let $(S, \bar{S}) = (S^{\nu}, \bar{S}^{\nu})$ be given and fixed. Let $x^{(\nu)0} = x^{\nu}$ and $t = 0$.

General iteration. Let $x^{(\nu)t+1}$ be the least d-stationary solution of the problem (19) corresponding to the unique tuple $\kappa_t^{(\nu)} \triangleq \left\{ \bar{k}_i^{(\nu)t} \right\}_{i=1}^n$ of indices determined by $x^{(\nu)t}$.

Terminate if $x^{(\nu)t+1} = x^{(\nu)t}$; otherwise, repeat the general iteration with $t \leftarrow t+1$.

associated with f_i as specified above. Then Algorithm II *is well defined and will terminate in no more than $2 + \sum_{i=1}^n K_i$ iterations with a d-stationary solution $x^{\nu+1}$ of* (13) *satisfying (A) $x^{\nu+1} \le x^{\nu}$, and (B) $\varphi(x^{\nu+1}) \le \varphi(x^{\nu})$, where φ is the objective function of* (12).

Proof The proof of assertion (A) consists of two main steps followed by a summary. The proof of assertion (B) is presented as Step 4.

Step 1: well-definedness We prove by induction on the inner iteration counter t that $x^{(\nu)t} \in \mathcal{Z}_{S^{\nu}}^{\kappa_t^{(\nu)}}$ for all t. Once this is proved, four things follow: (i) the well-definedness of the iterate $x^{(\nu)t+1}$ as the least d-stationary solution of the problem (19) corresponding to the tuple $\kappa_t^{(\nu)}$, (ii) $x^{(\nu)t+1} \le x^{(\nu)t}$, (iii) $\bar{k}_i^{(\nu)t+1} \le \bar{k}_i^{(\nu)t}$ for all $i \in S^{\nu}$, and (iv) $\bar{k}_i^{(\nu)t+1} \ge \bar{k}_i^{(\nu)t}$ for all $i \in \bar{S}^{\nu}$.

Consider $t = 0$. To verify condition (b) in the set $\mathcal{Z}_{S^{\nu}}^{\kappa_t^{(\nu)}}$ for $x^{(\nu)t}$, consider an index $i \in S^{\nu}$ such that $x_i^{\nu} < u_i$. Since x^{ν} is stationary for (13) corresponding to the pair $(S^{(\nu-1)}, \bar{S}^{(\nu-1)})$ and since $S^{\nu} \subseteq S^{\nu-1}$, it follows that $\dfrac{\partial \theta(x^{\nu})}{\partial x_i} + f_i{}'(0; 1) \ge 0$ if $x_i^{\nu} = 0$; and that $\dfrac{\partial \theta(x^{\nu})}{\partial x_i} + f_i{}'(x_i^{\nu}) = 0$ if $x_i^{\nu} > 0$. In either case, we have $\dfrac{\partial \theta(x^{\nu})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t}}(x_i^{\nu}) \ge 0$ by the definition of the index $\bar{k}_i^{(\nu)t}$. Next, to verify condition (c) in the set $\mathcal{Z}_{S^{\nu}}^{\kappa_t^{(\nu)}}$, consider an index $i \in \bar{S}^{\nu} = \bar{S}^{\nu-1} \cup E^{\nu-1}$ such that $x_i^{\nu} < 0$. Such an index must belong to $i \in \bar{S}^{\nu-1}$ and we have $\dfrac{\partial \theta(x^{\nu})}{\partial x_i} + f_i{}'(|x_i^{\nu}|; -1) \ge 0$ if $x_i^{\nu} = \ell_i$; and $\dfrac{\partial \theta(x^{\nu})}{\partial x_i} - f_i{}'(|x_i^{\nu}|) = 0$ if $x_i^{\nu} > \ell_i$. In either case, we deduce that condition (c) holds.

Next assume that $x^{(\nu)t} \in \mathcal{Z}_{S^{\nu}}^{\kappa_t^{(\nu)}}$ so that the iterate $x^{(\nu)t+1}$ is well defined and satisfies $x^{(\nu)t+1} \le x^{(\nu)t}$. To complete the induction, we need to show that $x^{(\nu)t+1} \in \mathcal{Z}_{S^{\nu}}^{\kappa_{t+1}^{(\nu)}}$. For this purpose, we compare the tuple $\kappa_{t+1}^{(\nu)} \triangleq \left\{ \bar{k}_i^{(\nu)t+1} \right\}_{i=1}^n$ of indices determined by $x^{(\nu)t+1}$ and the tuple $\kappa_t^{(\nu)} \triangleq \left\{ \bar{k}_i^{(\nu)t} \right\}_{i=1}^n$ of indices determined by $x^{(\nu)t}$. Since $x^{(\nu)t+1} \le x^{(\nu)t}$, it follows that $\bar{k}_i^{(\nu)t+1} \le \bar{k}_i^{(\nu)t}$ for all $i \in [n]$. To verify condition

(b) in the set $\mathcal{Z}_{S^\nu}^{\kappa_{t+1}^{(\nu)}}$ for $x^{(\nu)t+1}$, consider an index $i \in S^\nu$ such that $x_i^{(\nu)t+1} < u_i$. Since $x^{(\nu)t+1}$ is a d-stationary solution of the problem (19) corresponding to the tuple $\kappa_t^{(\nu)}$, we have $\dfrac{\partial\theta(x^{(\nu)t+1})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t}}(x_i^{(\nu)t+1}) \geq 0$ if $x_i^{(\nu)t+1} = 0$ and $\dfrac{\partial\theta(x^{(\nu)t+1})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t}}(x_i^{(\nu)t+1}) = 0$ if $x_i^{(\nu)t+1} > 0$. Since $\phi'_{i\bar{k}_i^{(\nu)t+1}}(x_i^{(\nu)t+1}) \geq \phi'_{i\bar{k}_i^{(\nu)t}}(x_i^{(\nu)t+1})$ by the piecewise dominance of the family $\{\phi_{ik}\}_{k=1}^{K_i+1}$, we easily deduce that condition (b) in the set $\mathcal{Z}_{S^\nu}^{\kappa_{t+1}^{(\nu)}}$ holds for $x^{(\nu)t+1}$. Similarly so does condition (c). The induction is thus completed.

Step 2: change of pieces For $i \in S^\nu$, the component $x_i^{(\nu)t+1}$ of the new iterate either transitions to a new interval to the left of the current interval $\left(\xi_{i\bar{k}_i^{(\nu)t}-1}, \xi_{i\bar{k}_i^{(\nu)t}}\right]$ that contains $x_i^{(\nu)t}$ or stays in the same interval. Similarly, for $i \in \bar{S}^\nu$, the component $\left|x_i^{(\nu)t+1}\right|$ of the new iterate either transitions to a new interval to the right of the current interval $\left[\xi_{i\bar{k}_i^{(\nu)t}-1}, \xi_{i\bar{k}_i^{(\nu)t}}\right)$ that contains $|x_i^{(\nu)t}|$ or stays in the same interval. We claim that in both cases, if $|x_i^{(\nu)t+1}|$ stays in the same interval as the previous iterate $|x_i^{(\nu)t}|$, then $x_i^{(\nu)t+1}$ must lie in the interior of this interval unless $x_i^{(\nu)t+1} \in \{0, \ell_i, u_i\}$. We prove this only for the case $i \in S^\nu$ as the proof for the case $i \in \bar{S}^\nu$ is similar. So let $i \in S^\nu$ be such that $0 < x_i^{(\nu)t+1} < u_i$ and assume that $x_i^{(\nu)t+1} = \xi_{i\bar{k}_i^{(\nu)t}}$. Then we have $x_i^{(\nu)t} = \xi_{i\bar{k}_i^{(\nu)t}} = x_i^{(\nu)t+1}$ and $x_i^{(\nu)t} \in (0, u_i)$. Since $x^{(\nu)t}$ is the d-stationary solution of the problem (19) corresponding to the tuple $\kappa_{t-1}^{(\nu)}$ associated with the previous iterate $x^{(\nu)t-1}$, and $x^{(\nu)t+1}$ is the d-stationary solution of the problem (19) corresponding to the tuple $\kappa_t^{(\nu)}$, we have

$$\begin{aligned}\frac{\partial\theta(x^{(\nu)t})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t-1}}(\xi_{i\bar{k}_i^{(\nu)t}}) &= \frac{\partial\theta(x^{(\nu)t})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t-1}}(x_i^{(\nu)t}) = 0 \\ &= \frac{\partial\theta(x^{(\nu)t+1})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t}}(x_i^{(\nu)t+1}) \\ &= \frac{\partial\theta(x^{(\nu)t+1})}{\partial x_i} + \phi'_{i\bar{k}_i^{(\nu)t}}(\xi_{i\bar{k}_i^{(\nu)t}}).\end{aligned}$$

We must have $\bar{k}_i^{(\nu)t-1} > \bar{k}_i^{(\nu)t}$; by the piecewise dominance of the gradients of the functions ϕ_{ik}, it follows that $\phi'_{i\bar{k}_i^{(\nu)t}}(\xi_{i\bar{k}_i^{(\nu)t}}) > \phi'_{i\bar{k}_i^{(\nu)t-1}}(\xi_{i\bar{k}_i^{(\nu)t}})$. This implies $\dfrac{\partial\theta(x^{(\nu)t})}{\partial x_i} > \dfrac{\partial\theta(x^{(\nu)t+1})}{\partial x_i}$. But this contradicts the antitonicity of the partial derivative $\dfrac{\partial\theta}{\partial x_i}$ because $x_j^{(\nu)t+1} \leq x_j^{(\nu)t}$ for all j with equality holding for $j = i$.

Step 3: conclusion As the iterations proceed, after no more than $2 + \sum_{i=1}^n K_i \leq 2 + n \max_{1\leq i\leq n} K_i$ iterations, we will arrive at an iterate $x^{(\nu)t+1}$ with the following property: (i) $x^{(\nu)t+1}$ is the least d-stationary solution of the problem (19) corresponding to the tuple $\kappa_t^{(\nu)}$; (ii) $\kappa_{t+1}^{(\nu)} = \kappa_t^{(\nu)}$; and (iii) for each $i \in [n]$, either $x_i^{(\nu)t+1} \in \{0, \ell_i, u_i\}$

or $x_i^{(\nu)t+1} \notin \{\xi_{ik}\}_{k=1}^{K_i}$. These 3 properties of $x^{(\nu)t+1}$ are enough to establish that it is the desired d-stationary solution asserted by the theorem. The count of $2 + \sum_{i=1}^n K_i$ is the result of piece transitions plus the initial subproblem and the last one confirming the d-stationarity of $x^{(\nu)t+1}$ for (13).

Step 4: assertion (B) We show inductively that $\varphi\left(x^{(\nu)t+1}\right) \leq \varphi\left(x^{(\nu)t}\right)$ for all iterations t throughout Algorithm II. From the piecewise structure of each f_i, it follows that for all iteration counts t,

$$\sum_{i=1}^n f_i\left(|x_i^{(\nu)t}|\right) = \sum_{i \in S^\nu} \phi_{i\bar{k}_i^{(\nu)t}}\left(x_i^{(\nu)t}\right) + \sum_{i \in \bar{S}^\nu} \phi_{i\bar{k}_i^{(\nu)t}}\left(-x_i^{(\nu)t}\right);$$

thus

$$\begin{aligned}
\varphi\left(x^{(\nu)t+1}\right) &= \theta\left(x^{(\nu)t+1}\right) + \sum_{i \in S^\nu} f_i\left(x_i^{(\nu)t+1}\right) + \sum_{i \in \bar{S}^\nu} f_i\left(-x_i^{(\nu)t+1}\right) \quad \text{by definition} \\
&\leq \theta\left(x^{(\nu)t+1}\right) + \sum_{i \in S^\nu} \phi_{i\bar{k}_i^{(\nu)t}}\left(x_i^{(\nu)t+1}\right) + \sum_{i \in \bar{S}^\nu} \phi_{i\bar{k}_i^{(\nu)t}}\left(-x_i^{(\nu)t+1}\right) \\
&\qquad \text{since } f_i(\xi) = \min_{1 \leq k \leq K_i+1} \phi_{ik}(\xi) \text{for all} \xi \in [0, u_i] \\
&\leq \theta\left(x^{(\nu)t}\right) + \sum_{i \in S^\nu} \phi_{i\bar{k}_i^{(\nu)t}}\left(x_i^{(\nu)t}\right) + \sum_{i \in \bar{S}^\nu} \phi_{i\bar{k}_i^{(\nu)t}}\left(-x_i^{(\nu)t}\right) \quad \text{by optimality of } x^{(\nu)t+1} \\
&= \varphi\left(x^{(\nu)t}\right) \quad \text{by the definition of the index } \bar{k}_i^{(\nu)t}
\end{aligned}$$

as desired. □

6.3 Linear-step complexity

At first sight, since Algorithm I requires solving $n + 1$ problems of the kind (13) and Algorithm II solves each such problem by solving (at most) $2+\sum_{i=1}^n K_i$ differentiable programs each of the kind (19), it seems that the overall procedure for computing a d-stationary point of (12) would need to solve $\mathrm{O}(n^2)$ differentiable programs, provided that each K_i is a constant independent of n. Nevertheless a careful look at the two algorithms when they are combined reveal that only $\mathrm{O}(n)$ differentiable programs need to be solved. We summarize this conclusion in the theorem below, which focuses on the nondifferentiable case where Algorithm II is needed. The differentiable case is directly covered by Theorem 11 with the supporting Proposition 13. With this remark, it is understood that both Theorem 16 and the subsequent one, Theorem 17, apply to all the differentiable and nondifferentiable piecewise regularizers.

Theorem 16 *Let θ be continuously differentiable and strongly convex such that $\nabla\theta$ is a Z-function. Suppose that the bounds satisfy: $-\infty \leq \ell_i < 0 < u_i \leq \infty$ for all $i \in [n]$ and that each integer K_i is a constant independent of n. Algorithm* I *with Algorithm II embedded in it computes a d-stationary point x^* of* (12) *by solving $2+2\sum_{i=1}^n (K_i+1)$ differentiable programs each of the kind* (19)*. Moreover, $\varphi(x^*) \leq \varphi(x^1)$, where x^1 is a d-stationary solution of* (13) *with $S^0 = [n]$.*

Proof Consider an outer iteration ν with the pair $(S^{\nu}, \bar{S}^{\nu})$ kept fixed within a run of Algorithm II that has x^{ν} as the initial iterate and ends with $x^{\nu+1}$ and a new pair $(S^{\nu+1}, \bar{S}^{\nu+1})$. Consider a variable x^i and say $i \in S^{\nu}$ so that x_i is nonnegative throughout Algorithm II. (The situation for an index $i \in \bar{S}^i$ is similar; thus the factor of 2 in the final count.) This index i will remain in $S^{\nu+1}$ unless $i \in E^{\nu}$. In the former case, the variable x_i will continue to be restricted nonnegative. In the latter case, the same variable switches from being nonnegative to being restricted nonpositive. In both cases, the variable x_i will only move to the left or stay at the same value. Consequently, each variable $|x_i|$ can traverse each of the $K_i + 1$ intervals $\left\{ \left[\xi_{ik}, \xi_{ik+1} \right] \right\}_{k=0}^{K_i}$ only once throughout the entire solution process. We further note that there must be one change in the index set S^{ν} at the end of each outer iteration. Putting these 3 properties together: (i) only left movement of each variable, (ii) no reverse visits, and (iii) at least one decrease in the S-set at each outer iteration, we deduce that the total count of piece transitions is at most $\sum_{i=1}^{n} (2K_i + 1)$; adding this count to the number $n + 1$ of d-stationarity confirmations in each inner loop per outer loop and the one initial subproblem of the outer loop corresponding to $S^0 = [n]$, we obtain the claimed total count of differentiable programs each of the kind (19) being solved. $\square$

Remark Although Algorithm I is initiated with a special pair of index sets, it actually can be initialized at any pair $(S, \bar{S})$ such that (13) has a d-stationary solution $\bar{x}$ such that $\bar{x}_i < 0$ for all $i \notin S$. The conclusions of Theorem 16 remain valid. The special initialization $([n], \emptyset)$ is a trivial choice satisfying the requirement.

6.4 Sketch of a finite-step algorithm for solving (19)

The workhorse in the update of the iterate $x^{\nu+1}$ is the computation of a directional stationary solution of the bounded-variable problem with a differentiable objective function:

$$
\begin{array}{ll}
\underset{x \in \mathbb{R}^n}{\textbf{minimize}} & \theta(x) + \sum_{i \in S} h_i(x_i) + \sum_{i \notin S} h_i(-x_i) \\
\textbf{subject to} & u_S \geq x_S \geq 0 \\
\text{and} & \ell_{\bar{S}} \leq x_{\bar{S}} \leq 0,
\end{array}
\tag{20}
$$

where θ is strongly convex and continuously differentiable on $\mathbb{R}^n$ and each h_i is a continuously differentiable function on $\mathbb{R}$; further, without loss of generality, we may take each lower bound $\ell_i > -\infty$ for all $i \notin S$. We make two remarks about this setting: (i) Even though in the differentiable case, the original functions f_i are differentiable only on $\mathbb{R}_{++}$, as far as the fixed-sign subproblem (13) is concerned, we can extend each f_i to the whole real line in a continuously differentiable way without affecting the solution of the subproblem, thus justifying the setting of (20). (ii) Even if some original lower bounds are equal to $-\infty$, as in an unconstrained problem, the strong convexity of θ will lead to some implied bounds on the solutions of the problem.

Since the objective of (20) is differentiable, any gradient method can be employed to compute a d-stationary point of the problem (without it being convex). In general, such a method converges only in the limit; many of them have complexity inversely proportional to the prescribed accuracy that is quite different from the strongly polynomial kind emphasized in this paper. In order to obtain a finite-step algorithm for solving (20), we again employ the Z-property of $\nabla\theta$. To derive such an algorithm, note that with the change of variables $y_{\bar{S}} \triangleq x_{\bar{S}} - \ell_{\bar{S}}$ (this is where the finiteness of the lower bound is needed) and by defining $\widetilde{\theta}(x_S, y_{\bar{S}}) \triangleq \theta(x_S, y_{\bar{S}} + \ell_{\bar{S}})$ and $\widetilde{h}_i(y_i) \triangleq h_i(-y_i - \ell_i)$, the problem (20) is equivalent to

$$\begin{array}{ll} \underset{(x_S, y_{\bar{S}}) \in \mathbb{R}^n}{\textbf{minimize}} & \widetilde{\theta}(x_S, y_{\bar{S}}) + \sum_{i \in S} h_i(x_i) + \sum_{i \notin S} \widetilde{h}_i(y_i) \\ \textbf{subject to} & u_S \geq x_S \geq 0 \\ \text{and} & -\ell_{\bar{S}} \geq y_{\bar{S}} \geq 0, \end{array} \tag{21}$$

where the pair $z \triangleq (x_S, y_{\bar{S}})$ is nonnegative and upper bounded. The gradient $\nabla\widetilde{\theta}$ remains a Z-function. We may write the Karush–Kuhn–Tucker conditions of (21) as the following upper-bounded nonlinear complementarity problem:

$$\begin{array}{l} 0 \leq z \perp F(z) + w \geq 0, \quad \text{where} \quad F(z) \triangleq \nabla\widetilde{\theta}(z) + D(z) \\ 0 \leq w \perp u - z \geq 0, \end{array} \tag{22}$$

with $D(z) = (d_i(z_i))_{i=1}^n$ being a separable vector function whose components are $h_i'(x_i)$ for $i \in S$ and $\tilde{h}_i'(y_i)$ for $i \notin S$, and w is the multiplier vector of the upper bound constraint on the primary variable z. At this point, by combining the algorithm in [54] for a nonlinear complementarity problem and that in [48] for an upper-bounded linear complementarity problem, we can obtain a linear-step algorithm of solving (22) in which the principal computational effort is the solution of a (square) system of nonlinear equations for various subsets $\alpha \subseteq [n]$ with complement $\bar{\alpha}$: $F_\alpha(z_\alpha, z_{\bar{\alpha}}^{\text{fixed}}) = 0$, where the components of $z_{\bar{\alpha}}^{\text{fixed}}$ are fixed at either their upper or lower bounds. Moreover, the index set α increases monotonically in size throughout the algorithm; thus, the solution of such an equation can be employed to initiate the computation of the next equation. In the case of an affine function F, which is derived when θ and all the element functions ϕ_{ik} are quadratic, efficient linear-algebraic updates can further speed up the numerical solution of the resulting linear equations and reduce the computational complexity by an order of magnitude. The solution produced by the overall algorithm is a pair $(\bar{z}, \bar{w})$ such that $\bar{z}$ is the least element of a set

$$\mathcal{Z} \triangleq \{ z \in [0, u] \mid \forall\, i \in [n],\ z_i < u_i \text{ implies } F_i(z) \geq 0 \}.$$

With a careful implementation that takes advantage of monotonic enlargement of the index sets α mentioned above, the total number of equations to be solved is at most $O(n)$.

6.5 A summarizing strong polynomiality result

Due to the importance of the special case defined by quadratic functions, we present the following summarizing strongly polynomial result. The $O(n^3)$ complexity of obtaining the least-element solution of the upper-bounded linear complementarity problem (23) below follows from a careful implementation of the algorithm in [48] as sketched above.

Theorem 17 *Let θ be a quadratic function with a Stieltjes Hessian matrix Q. Let each function $f_i \triangleq \min_{1 \le k \le K_i} \phi_{ik}$ with all the element functions $\{\phi_{ik}\}_{i=1}^{K_i+1}$ being quadratic. (For simplicity) assume that $-\ell = u_i < \infty$ for all $i \in [n]$. Then by solving for the least-element solution of at most $\bar{\nu} \triangleq 2 + 2\sum_{i=1}^{n}(K_i + 1)$ upper-bounded linear complementarity problems each of the kind:*

$$\begin{aligned} 0 \le z \perp q + \widehat{Q}z + w \ge 0, \quad &\text{where} \quad \widehat{Q} \triangleq Q + D, \\ 0 \le w \perp u - z \ge 0, \quad & \end{aligned} \tag{23}$$

where D is a diagonal matrix with nonpositive diagonal entries, Algorithm I generates a finite, monotonically nonincreasing sequence $\{x^{\nu}\}_{\nu=1}^{N}$ with $N \le \bar{\nu}$ such that $\varphi(x^{\nu+1}) \le \varphi(x^{\nu})$ for all $\nu = 1, \ldots, N-1$, where x^1 is a d-stationary solution of (13) with $S^0 = [n]$; moreover, x^N is a d-stationary solution of (12). In turn, each desired least-element solution of *(23) can be obtained in $O(n^3)$ complexity by solving linear equations each defined by a principal submatrix of $\widehat{Q}$. Finally, if the integers K_i are all independent of n, then the computational complexity of obtaining the d-stationary solution x^N is of order $O(n^4)$.*

7 The parametric problem: quadratic objective

In this section, we consider the problem (3) with three specifications: $\delta = 1$, θ is a strongly convex quadratic function, and each f_i is a positive multiple of the identity function. Thus, the quadratic optimization problem with the weighted-ℓ_1 surrogate sparsity control is as follows:

$$\underset{\ell \le x \le u}{\textbf{minimize}}\; q^\top x + \tfrac{1}{2}x^\top Q x + \gamma \sum_{i=1}^{n} p_i \,|\, x_i \,|, \qquad \gamma > 0, \tag{24}$$

where the matrix $Q \in \mathbb{R}^{n \times n}$ is symmetric positive definite, the vector $p \in \mathbb{R}^n$ is positive; and there is no sign restriction on the components of $q \in \mathbb{R}^n$. With a strongly convex objective function, the problem (24) has a unique solution, denoted by $x(\gamma)$, that is a piecewise affine function of the parameter γ. The main goal in the section is to identify conditions on the pair (Q, p) under which there is at most a linear number of pieces of linearity of the solution function $x(\gamma)$. As mentioned in Sect. 3, having an efficient algorithm for identifying all such pieces allows us to globally resolve the singly-parameter bilevel optimal parameter identification problem (2) via interval searches. The linear (in n) number of such pieces offers a significant advantage in that this bilevel problem can be solved with guaranteed effectiveness to global optimality with a friendly (such as quadratic) outer objective. Another benefit of solving the problem (24) for all values of γ is that one can determine a value γ such that the corresponding solution $x(\gamma)$ has the smallest ℓ_0-value.

Our treatment of the parametric problem (24) is divided into two cases: fixed sign pattern of the variables, and no a priori knowledge of the signs of the variables. These are the respective topics of discussion in the two subsections below. In the first subsection, the sign patterns of the off-diagonal entries of Q (in particular, the Z-structure) turn out not needed; but the Z-structure is re-instated in the second subsection.

7.1 Fixed sign partition

Let $(S, \bar{S})$ be a given pair of complementary index subsets of $\{1, \dots, n\}$. We are interested in solving the problem (cf. (13)):

$$\left\{ \begin{array}{l} \underset{x \in \mathbb{R}^n}{\textbf{minimize}} \;\; q^\top x + \frac{1}{2} x^\top Q x + \gamma \left[\sum_{i \in S} p_i \, x_i - \sum_{i \in \bar{S}} p_i \, x_i \right] \\ \textbf{subject to } u_S \, \geq \, x_S \, \geq \, 0 \quad \text{and} \quad \ell_{\bar{S}} \, \leq \, x_{\bar{S}} \, \leq \, 0, \end{array} \right.$$

parametrically for all values of $\gamma \geq 0$, with $-\infty \leq \ell_i < 0 < u_i \leq \infty$ for all $i \in [n]$. Letting $z_{\bar{S}} \triangleq -x_{\bar{S}}$, we may write the problem equivalently as one where all the variables are nonnegative and bounded above, the vectors q and p have the signs of their $\bar{S}$-components negated, and the matrix Q is changed accordingly to the following matrix:

$$Q^{\,\text{signed}} \triangleq \begin{bmatrix} Q_{SS} & -Q_{S\bar{S}} \\ -Q_{\bar{S}S} & Q_{\bar{S}\bar{S}} \end{bmatrix}.$$

which remains symmetric positive definite. Abusing the notation, we write the resulting problem as:

$$\left\{ \begin{array}{l} \underset{x \in \mathbb{R}^n}{\textbf{minimize}} \;\; q^\top x + \frac{1}{2} x^\top Q x + \gamma \, p^\top x \\ \textbf{subject to } u \, \geq \, x \, \geq \, 0 \end{array} \right\} \qquad p \, > \, 0. \tag{25}$$

Given a triplet of index subsets $(\beta_0, \alpha, \beta_{\text{ub}})$ that partitions the full index set $\{1, \dots, n\}$, we define a vector $x(\gamma)$ with $x_{\beta_0}(\gamma) = 0$, $x_{\beta_{\text{ub}}}(\gamma) = u_{\beta_{\text{ub}}}$, and $x_\alpha(\gamma) = -[Q_{\alpha\alpha}]^{-1}(q_\alpha + Q_{\alpha\beta_{\text{ub}}} u_{\beta_{\text{ub}}} + \gamma \, p_\alpha)$. By writing down the Karush–Kuhn–Tucker conditions of the above QP, it is easy to deduce that this vector is an optimal solution of (25) if and only if the following three conditions are satisfied:

- $\bar{q}_{\beta_0} + \gamma \, \bar{p}_{\beta_0} \geq 0$;
- $u_\alpha \geq -\bar{q}_\alpha - \gamma \, \bar{p}_\alpha \geq 0$; and
- $\bar{q}_{\beta_{\text{ub}}} + \gamma \, \bar{p}_{\beta_{\text{ub}}} \leq 0$; where

$$\begin{aligned} \begin{pmatrix} \bar{q}_\alpha \\ \bar{p}_\alpha \end{pmatrix} &\triangleq [\, Q_{\alpha\alpha} \,]^{-1} \begin{pmatrix} q_\alpha + Q_{\alpha\beta_{\text{ub}}} u_{\beta_{\text{ub}}} \\ p_\alpha \end{pmatrix} \\ \begin{bmatrix} \bar{q}_{\beta_0} & \bar{p}_{\beta_0} \\ \bar{q}_{\beta_{\text{ub}}} & \bar{p}_{\beta_{\text{ub}}} \end{bmatrix} &\triangleq \begin{bmatrix} q_{\beta_0} + Q_{\beta_0\beta_{\text{ub}}} u_{\beta_{\text{ub}}} & p_{\beta_0} \\ q_{\beta_{\text{ub}}} + Q_{\beta_{\text{ub}}\beta_{\text{ub}}} u_{\beta_{\text{ub}}} & p_{\beta_{\text{ub}}} \end{bmatrix} - \begin{bmatrix} Q_{\beta_0\alpha} \\ Q_{\beta_{\text{ub}}\alpha} \end{bmatrix} [\, \bar{q}_\alpha \;\, \bar{p}_\alpha \,]. \end{aligned} \tag{26}$$

The parametric algorithm is initiated at the triplet of index sets with $\alpha = \beta_{\rm ub} = \emptyset$ and the smallest value $\gamma_0 \geq 0$ so that $q + \gamma p \geq 0$ for all $\gamma \geq \gamma_0$; for such γ, we have $x(\gamma) = 0$. The value of γ is then gradually pivoted down; in the process, some x-variables will increase above zero and eventually reach their respective upper bounds. During the pivots, the parametric vector p will ensure that the values of the x-components will not decrease. This monotonic properties of the variables constitute the proof of the linear number of iterations that is at most $2n$ (from zero to positive to upper bound; thus the multiplicative factor 2). Phrased in terms of the triplet of index sets $(\beta_0, \alpha, \beta_{\rm ub})$, the algorithm below is in the spirit of the n-step parametric principal pivoting algorithm [21, Section 4.8] extended here to deal with the upper bounds.

Algorithm III: a $2n$-step parametric procedure for solving (25)

Initialization. Let $(\beta_0, \alpha, \beta_{\rm ub}) = (\{1, \cdots, n\}, \emptyset, \emptyset)$. Also let $\gamma_{\rm old} \triangleq \infty$.

General iteration. Solve the system of linear equations in (26) to obtain $\bar{q}_\alpha$ and $\bar{p}_\alpha$ and compute $\begin{bmatrix} \bar{q}_{\beta_0} & \bar{p}_{\beta_0} \\ \bar{q}_{\beta_{\rm ub}} & \bar{p}_{\beta_{\rm ub}} \end{bmatrix}$ by (26). Perform the ratio test (by convention, the maximum over an empty set is equal to $-\infty$):

$$\gamma_{\rm new} \triangleq \max \left\{ \max_{i \in \beta_0 \,:\, \bar{p}_i > 0} \left\{ -\frac{\bar{q}_i}{\bar{p}_i} \right\}, \max_{i \in \alpha \,:\, \bar{p}_i > 0} \left\{ -\frac{u_i + \bar{q}_i}{\bar{p}_i} \right\}, 0 \right\}$$

For $\gamma \in [\gamma_{\rm new}, \gamma_{\rm old}]$, let $x_{\beta_0}(\gamma) = 0$, $x_{\beta_{\rm ub}}(\gamma) = u_{\beta_{\rm ub}}$, and $x_\alpha(\gamma) = -\bar{q}_\alpha - \gamma\, \bar{p}_\alpha$. Stop if $\gamma_{\rm new} = 0$; in this case, the entire solution path $\{x(\gamma) \mid \gamma \geq 0\}$ is computed. Otherwise let $i \in \alpha \cup \beta_0$ be a maximizing index of $\gamma_{\rm new}$ (ties can be broken arbitrarily). Update the index sets as follows:

$$(\beta_0, \alpha, \beta_{\rm ub})_{\rm new} = \begin{cases} (\beta_0 \setminus \{i\}, \alpha \cup \{i\}, \beta_{\rm ub}) & \text{if} i \in \beta_0 \\ (\beta_0, \alpha \setminus \{i\}, \beta_{\rm ub} \cup \{i\}) & \text{if } i \in \alpha. \end{cases}$$

Set $\gamma_{\rm old} \leftarrow \gamma_{\rm new}$ and return to the beginning of the general iteration.

We recall [21, Section 4.8] that a vector $p > 0$ is an *n-step vector* for a P-matrix M (i.e., one whose principal minors are all positive) if $[M_{\alpha\alpha}]^{-1} p_\alpha > 0$ for all index subset α of $\{1, \ldots, n\}$. In the case of the symmetric positive definite matrix Q, it is known from Corollary 4.8.11 in the cited reference that Q has an n-step vector if and only if it is a hidden Z-matrix, this being the class of matrices discovered by Mangasarian [44]. There are many matrices known to belong to this class; a broad family is the class of positively-scaled diagonally dominant matrices. Specifically, for a symmetric strictly quasi-diagonally dominant matrix Q (not necessarily of class Z), i.e., one that satisfies (7) for a positive vector d, the vector p defined by:

$$p_i \triangleq Q_{ii}\, d_i + \sum_{j \,:\, Q_{ij} < 0} Q_{ij}\, d_j, \quad i = 1, \ldots, n, \tag{27}$$

is an n-vector of Q; see [20, Corollary 1(ii)]. Below, we establish the claimed $2n$-iteration of the above algorithm for computing the entire solution path of the parametric QP (25).

Theorem 18 *Let $p > 0$ be an n-step vector for the symmetric positive definite matrix Q. In at most $2n$ iterations, Algorithm III computes the entire solution path $\{x(\gamma) \mid \gamma \geq 0\}$ of the parametric QP (25).*

Proof It suffices to show that $\bar{p}_{\beta_{\mathrm{ub}}} \geq 0$ throughout the algorithm. Consider an arbitrary index $i \in \beta_{\mathrm{ub}}$ and the system of linear equations:

$$\begin{bmatrix} Q_{\alpha\alpha} & Q_{\alpha i} \\ Q_{i\alpha} & Q_{ii} \end{bmatrix} \begin{pmatrix} \widetilde{p}_\alpha \\ \widetilde{p}_i \end{pmatrix} = \begin{pmatrix} p_\alpha \\ p_i \end{pmatrix}$$

whose unique solution is positive. It is not difficult to show that

$$\bar{p}_i = \widetilde{p}_i \left[Q_{ii} - Q_{i\alpha}[Q_{\alpha\alpha}]^{-1} Q_{\alpha i} \right].$$

Being the Schur complement of $Q_{\alpha\alpha}$ in the matrix $\begin{bmatrix} Q_{\alpha\alpha} & Q_{\alpha i} \\ Q_{i\alpha} & Q_{ii} \end{bmatrix}$, which is positive definite, the positivity of $\bar{p}_i$ follows readily. □

In general, the choice of the weights p_i in the ℓ_1 function is dependent on the matrix Q. In particular, the use of equal weights may not yield such linear-iteration termination; see e.g., the vector p with components (27) in the case of a strictly quasi-diagonally dominant matrix Q. Thus our results supplement the experimental work of Candès et al. [12] which demonstrates that the use of weighted ℓ_1-functions can enhance sparsity in practice; here the advantage of a weighted ℓ_1 function is for the efficient identification of the sparsity parameter to optimize a secondary objective in a bilevel setting.

7.2 Unknown sign pattern

This subsection addresses the solution of the following parametric problem:

$$\operatorname*{\mathbf{minimize}}_{x \in \mathbb{R}^n} \; q^\top x + \tfrac{1}{2} x^\top Q x + \gamma \sum_{i=1}^{n} p_i \mid x_i \mid \tag{28}$$

for all values of $\gamma \geq 0$ by a linear-iteration algorithm, where for notational simplicity, we have dropped the bounds of the variables. Here we can no longer rely on a single parametric quadratic program to determine the correct signs of the variables as this is part of the solution requirement. Instead we may proceed similarly to the previous case by considering two disjoint subsets (α_+, α_-) of $\{1, \ldots, n\}$. This pair of index sets induces a vector x with $x_\beta(\gamma) = 0$, where $\beta \triangleq \{1, \ldots, n\} \backslash (\alpha_+ \cup \alpha_-)$, and the other components given by

$$\begin{pmatrix} x_{\alpha_+}(\gamma) \\ x_{\alpha_-}(\gamma) \end{pmatrix} = - \begin{bmatrix} Q_{\alpha_+\alpha_+} & Q_{\alpha_+\alpha_-} \\ Q_{\alpha_-\alpha_+} & Q_{\alpha_-\alpha_-} \end{bmatrix}^{-1} \begin{pmatrix} q_{\alpha_+} \\ q_{\alpha_-} \end{pmatrix} - \gamma \begin{bmatrix} Q_{\alpha_+\alpha_+} & Q_{\alpha_+\alpha_-} \\ Q_{\alpha_-\alpha_+} & Q_{\alpha_-\alpha_-} \end{bmatrix}^{-1} \begin{pmatrix} p_{\alpha_+} \\ -p_{\alpha_-} \end{pmatrix}$$
$$\triangleq - \begin{pmatrix} \bar{q}_{\alpha_+} \\ \bar{q}_{\alpha_-} \end{pmatrix} - \gamma \begin{pmatrix} \bar{p}_{\alpha_+} \\ \bar{p}_{\alpha_-} \end{pmatrix}.$$

[Remark: When explicit bounds are present, these index sets $\alpha_\pm$ need to be refined to account for the variables at bounds, thus complicating the notations without requiring new ideas.] It is easy to show that

$$\begin{pmatrix} \bar{p}_{\alpha_+} \\ \bar{p}_{\alpha_-} \end{pmatrix} = \begin{pmatrix} \left[Q_{\alpha_+\alpha_+} - Q_{\alpha_+\alpha_-} \left[Q_{\alpha_-\alpha_-} \right]^{-1} Q_{\alpha_-\alpha_+} \right]^{-1} \left(p_{\alpha_+} + Q_{\alpha_+\alpha_-} \left[Q_{\alpha_-\alpha_-} \right]^{-1} p_{\alpha_-} \right) \\ - \left[Q_{\alpha_-\alpha_-} - Q_{\alpha_-\alpha_+} \left[Q_{\alpha_+\alpha_+} \right]^{-1} Q_{\alpha_+\alpha_-} \right]^{-1} \left(p_{\alpha_-} + Q_{\alpha_-\alpha_+} \left[Q_{\alpha_+\alpha_+} \right]^{-1} p_{\alpha_+} \right) \end{pmatrix}.$$

This vector $x(\gamma)$ is a solution of (28) if and only if

- $x_{\alpha_+}(\gamma) \geq 0$ and $x_{\alpha_-}(\gamma) \leq 0$; and
- $0 \leq \bar{q}_\beta + \gamma\, \bar{p}_\beta \leq 2\gamma\, p_\beta$, where

$$\begin{pmatrix} \bar{q}_\beta & \bar{p}_\beta \end{pmatrix} \triangleq \begin{pmatrix} q_\beta & p_\beta \end{pmatrix} - \begin{bmatrix} Q_{\beta\alpha_+} & Q_{\beta\alpha_-} \end{bmatrix} \begin{bmatrix} \bar{q}_{\alpha_+} & \bar{p}_{\alpha_+} \\ \bar{q}_{\alpha_-} & \bar{p}_{\alpha_-} \end{bmatrix}. \tag{29}$$

The n-step algorithm for solving the problem (28) parametrically for all $\gamma \geq 0$ is similar to the previous case. Namely, we start with $\gamma > 0$ sufficiently large so that $\gamma\, p_i \geq |\, q_i \,|$ for all i which yields $x(\gamma) = 0$ to be the unique solution of (28). We then decrease γ by performing some ratio tests; during the algorithm, we are ensured that $\bar{p}_{\alpha_+} \geq 0$ and $\bar{p}_{\alpha_-} \leq 0$; these sign conditions in turn imply that a variable x_i once becomes positive will stay positive and similarly a variable x_i once becomes negative will stay negative. Before identifying conditions for these two sign conditions on $(\bar{p}_{\alpha_+}, \bar{p}_{\alpha_-})$ to hold, we stay the following algorithm and its n-step termination.

The n-step termination of the above algorithm assumes that Q is a Z-matrix, a reversal of the non-requirement of this property in the previous case where the signs of the variables are prescribed.

Theorem 19 *Let Q be a Z-matrix in addition to being symmetric positive definite. Provided that $p > 0$ is such that $p_i + Q_{i\alpha}[Q_{\alpha\alpha}]^{-1} p_\alpha > 0$ for all $\alpha \subseteq \{1, \ldots, n\}$ and $i \notin \alpha$, Algorithm* IV *will compute the entire solution path $\{x(\gamma) \mid \gamma \geq 0\}$ of problem* (28) *in no more than n iterations.*

Proof The assumptions on Q imply that the matrices $Q_{\alpha_\pm\alpha_\pm} - Q_{\alpha_\pm\alpha_\mp} \left[Q_{\alpha_\mp\alpha_\mp} \right]^{-1} Q_{\alpha_\mp\alpha_\pm}$, being Schur complements are M-matrices, are themselves M-matrices; thus they have nonnegative inverses. Thus $\bar{p}_{\alpha_+} > 0$ and $\bar{p}_{\alpha_-} < 0$ under the given assumptions on the vector p. This is enough to establish the claim of the theorem. □

Algorithm IV: an n-step parametric procedure for solving (28)

Initialization. Let $(\beta, \alpha_+, \alpha_-) = (\{1, \cdots, n\}, \emptyset, \emptyset)$. Also let $\gamma_{\text{old}} \triangleq \infty$.

General iteration. Solve the system of linear equations:

$$\begin{bmatrix} Q_{\alpha_+\alpha_+} & Q_{\alpha_+\alpha_-} \\ Q_{\alpha_-\alpha_+} & Q_{\alpha_-\alpha_-} \end{bmatrix} \begin{pmatrix} \bar{q}_{\alpha_+} & \bar{p}_{\alpha_+} \\ \bar{q}_{\alpha_-} & \bar{p}_{\alpha_-} \end{pmatrix} = \begin{pmatrix} q_{\alpha_+} & p_{\alpha_+} \\ q_{\alpha_-} & -p_{\alpha_-} \end{pmatrix},$$

and compute $(\bar{q}_\beta, \bar{p}_\beta)$ by (29). Perform the ratio test:

$$\gamma_{\text{new}} \triangleq \max \left\{ \max_{i \in \beta \,:\, \bar{p}_i > 0} \left\{ -\frac{\bar{q}_i}{\bar{p}_i} \right\}, \max_{i \in \beta \,:\, \bar{p}_i < 2p_i} \left\{ \frac{\bar{q}_i}{2p_i - \bar{p}_i} \right\}, 0 \right\}$$

For $\gamma \in [\gamma_{\text{new}}, \gamma_{\text{old}}]$, let $x_\beta(\gamma) = 0$, $x_{\alpha_+}(\gamma) = -\bar{q}_{\alpha_+} - \gamma\, \bar{p}_{\alpha_+}$ and $x_{\alpha_-}(\gamma) = -\bar{q}_{\alpha_-} - \gamma\, \bar{p}_{\alpha_-}$ Stop if $\gamma_{\text{new}} = 0$; in this case, the entire solution path $\{x(\gamma) \mid \gamma \geq 0\}$ is computed. Otherwise let $i \in \beta$ be a maximizing index of γ_{new} (ties can be broken arbitrarily). Update the index sets as follows:

$$(\beta, \alpha_+, \alpha_-)_{\text{new}} = \begin{cases} (\beta \setminus \{i\}, \alpha_+ \cup \{i\}, \alpha_-) & \text{if } i \in \beta \text{ and } \bar{p}_i > 0 \\ (\beta \setminus \{i\}, \alpha_+, \alpha_- \cup \{i\}) & \text{if } i \in \beta \text{ and } \bar{p}_i < 2p_i. \end{cases}$$

Set $\gamma_{\text{old}} \leftarrow \gamma_{\text{new}}$ and return to the beginning of the general iteration.

To give a class of Z-matrices Q for which such a vector p can be easily identified, we note that the scalar $p_i + Q_{i\alpha}[Q_{\alpha\alpha}]^{-1} p_\alpha$ is the Schur complement of the matrix $Q_{\alpha\alpha}$ in the bordered matrix $R \triangleq \begin{bmatrix} p_i & Q_{i\alpha} \\ -p_\alpha & Q_{\alpha\alpha} \end{bmatrix}$, which is of class Z. Thus,

$$p_i + Q_{i\alpha}[Q_{\alpha\alpha}]^{-1} p_\alpha = \frac{\det R}{\det Q_{\alpha\alpha}} = \frac{p_i}{\det Q_{\alpha\alpha}} \det \left(Q_{\alpha\alpha} + \frac{1}{p_i} p_\alpha Q_{i\alpha} \right).$$

Hence, $p_i + Q_{i\alpha}[Q_{\alpha\alpha}]^{-1} p_\alpha > 0$ if and only if $Q_{\alpha\alpha} + \dfrac{1}{p_i} p_\alpha Q_{i\alpha}$, which itself is a Z-matrix, has positive determinant. Based on this derivation, the following corollary of Theorem 19 is easy to prove.

Corollary 20 *Let Q be a symmetric positive definite Z-matrix. If there exists a positive vector d* such that

$$Q_{kk}\, d_k > \sum_{j \neq k} |\, Q_{kj} \,|\, d_j + \sum_{j \neq \ell} |\, Q_{\ell j} \,|\, d_j, \quad \forall\, k \neq \ell \in \{1, \ldots, n\}, \tag{30}$$

then the n-step algorithm will compute the entire solution path $\{x(\gamma) \mid \gamma \geq 0\}$ of problem (28) in no more than n iterations with p being the vector of all ones.

Proof It suffices to show that the matrix $Q_{\alpha\alpha} + \mathbf{1}_\alpha Q_{i\alpha}$ is an M-matrix. In turn, it is enough to show that $[\, Q_{\alpha\alpha} + \mathbf{1}_\alpha Q_{i\alpha} \,]\, d_\alpha > 0$. Indeed, the $k(\in \alpha)$-th element of the latter vector is equal to

$$Q_{kk}\,d_k + \sum_{j\in\alpha\,:\,j\neq k} Q_{kj}\,d_j + \sum_{j\in\alpha} Q_{ij}\,d_j \;\geq\; Q_{kk}\,d_k - \sum_{j\neq k} |\,Q_{kj}\,|\,d_j - \sum_{j\neq i} |\,Q_{ij}\,|\,d_j$$

which is positive by assumption. □

In general, for any matrix Q with positive diagonals, we may define a matrix $\widehat{Q}$ by:

$$\widehat{Q}_{ij} \;\triangleq\; \begin{cases} Q_{ii} & \text{if} i = j \\ -\left[\,|\,Q_{ij}\,| + \max\limits_{\ell\neq i,j} |\,Q_{\ell j}\,|\,\right] & \text{if } i \neq j. \end{cases}$$

It is then easy to see that (30) holds for some positive vector d if $\widehat{Q}$ is an M-matrix. In general, if there are permissible controls on the diagonal and/or off-diagonal entries, such as in the applied sparse-smooth signal problem (3) and portfolio revision problem (6), the resulting matrix $\widehat{Q}$ can be seen to have this M-property. The upshot is that for these two problems, with suitable choices on the smoothing parameter λ in the former, and the proximal parameter $\varepsilon > 0$ in the latter, the ℓ_1-version of the former problem can be solved parametrically for all values of the sparsity parameter $\gamma > 0$ in strongly polynomial time and similarly for the latter problem; thus so can their respective bilevel versions of optimally choosing the latter parameter with say, quadratic outer objectives.

8 Numerical results

In this section, we test out the practical efficiency of Algorithm I with Algorithm II embedded in it, which together we call the GHP Algorithm. We apply the GHP Algorithm to problem (8) with $\theta(x) = q^\top x + \frac{1}{2}x^\top Qx$ being quadratic with a Stieltjes matrix Q and with (a) $f_i(t) = \gamma\,p_i\,t$, which leads to the scaled ℓ_1-regularizer $\gamma\sum_{i=1}^n p_i\,|\,x_i\,|$, and (b) $f_i(t) = \gamma\,p_i\,\min\left(\frac{t}{\delta},\,1\right)$, which leads to the capped ℓ_1-regularizer $\gamma\sum_{i=1}^n p_i\,\min\left(\frac{|\,x_i\,|}{\delta},\,1\right)$. Problems with the ℓ_1-regularizer are (equivalent to) convex quadratic programs; we compare the GHP Algorithm with two well-established solvers GUROBI 9.1 [30] and MOSEK 9.3 [46]. For problems with the capped ℓ_1-regularizer, we compare the GHP Algorithm with GUROBI applied to a mixed integer program reformulation specified Sect. 8.2. We terminate Algorithms I and II as in their respective descriptions; for the latter algorithm, $x^{(\nu)t+1} = x^{(\nu)t}$ is deemed satisfied if $|\,x_i^{(\nu)t+1} - x_i^{(\nu)t}\,| \leq 10^{-9}$ for all $i \in [n]$.

We synthetically generate a sparse Stieltjes matrix Q with $n = 5{,}000$ (a meaningful size for experimental purposes) with the density of the off-diagonal entries equal to $\sigma \in \{0.01, 0.2\}$, i.e., the number of such entries is $\sigma n(n-1)$. The nonzero off-diagonal entries are uniformly generated within the interval $[-1, 0)$; then we set the

Table 1 Computational results for the ℓ_1-problems. In this case (convex), all methods solve the optimization problem to optimality

Settings	Unconstrained				Constrained			
	GHP		GRB time	MSK time	GHP		GRB time	MSK time
	Time	Steps I			Time	Steps I		
$(\gamma, \sigma) = (0.01, 0.01)$	1.32	5.2	176.20	59.71	1.49	5.4	197.50	77.12
$(\gamma, \sigma) = (0.01, 0.2)$	10.51	5.4	286.17	61.16	11.10	5.2	335.78	63.63
$(\gamma, \sigma) = (1, 0.01)$	1.35	5.2	174.93	52.49	1.32	5.0	198.54	81.95
$(\gamma, \sigma) = (1, 0.2)$	12.28	6.0	295.37	77.52	11.39	5.4	333.67	60.80
$(\gamma, \sigma) = (10, 0.01)$	1.13	5.0	196.06	56.13	1.16	5.0	185.70	88.44
$(\gamma, \sigma) = (10, 0.2)$	9.23	5.2	312.28	63.86	9.34	5.0	331.36	59.10

diagonal terms $Q_{ii} = 1.2 \sum_{j \in [n], j \neq i} |Q_{ij}|, \forall i \in [n]$ so that Q is positive definite. Additionally, we randomly generate the components of q independently and uniformly in $[-100, 100]$ and those of p in $(0, 1]$. When constraints $\ell \leq x \leq u$ are imposed, we set $-\ell_i = u_i = \frac{2}{3} \|Q^{-1}q\|_\infty$ for all $i \in [n]$, where $-Q^{-1}q$ is the unconstrained minimizer of θ. All the statistics reported in the tables below are averaged over 5 random instances. The experiments were carried out within MATLAB R2017b on a Mac OS X personal computer with 2.3 GHz Intel Core i7 and 8 GB RAM.

8.1 ℓ_1-Penalized problems

In Table 1, we summarize the computational time (in seconds) and the number of iterations in Algorithm I (labelled steps I) under the aforementioned settings with γ additionally fixed as 0.01, 1, or 10. (Note that Algorithm II is not needed for the ℓ_1-problems.) Each fixed-sign subproblem in Algorithm I is solved as a linear complementarity problem by the method described in [48]. The computational results of GUROBI and MOSEK are abbreviated as "GRB" (resp. "MSK") for short.

As we can observe from Table 1, the proposed method is many times faster than the benchmark commercial solvers (by a factor of at least five in almost all the settings), and can exploit sparsity in the data much more effectively as well. In particular, for the cases with $\sigma = 0.01$ (low off-diagonal density), GHP can be 160 (resp. 75) faster than GRB (resp. MSK), e.g., see setting with $(\gamma, \sigma) = (10, 0.01)$ and constrained problems. For similar cases with $\sigma = 0.2$, GHP is (only) 40 times faster than GRB, and 6 times faster than MSK. We also see that while GHP requires in the worst case $n+1$ iterations, in practice, it terminates in practice in five iterations on average. Thus, the practical performance of GHP may be substantially faster than the worst-case $O(n^4)$ complexity, see Sect. 8.3 for more details.

8.2 Capped ℓ_1-Penalized problems

Besides the ℓ_1-cases, we also test and highlight the advantages of the GHP Algorithm when applied to capped ℓ_1-problems. We fix $\gamma = 1$ and $\delta = \frac{1}{6} \|Q^{-1}q\|_\infty$ in both unconstrained and constrained settings. Table 2 summarize the computing time, total number of inner loops in Algorithm II (labelled steps II) and the final objective value (denoted as "obj.") when our method terminates with a d-stationary point x_{ghp}. As a comparison, we apply GUROBI to solve the capped ℓ_1-problem reformulated as the following mixed integer quadratic program:

$$\begin{array}{ll} \underset{x,t,s\in\mathbb{R}^n}{\textbf{minimize}} & \frac{1}{2}x^\top Qx + q^\top x + \gamma\, p^\top t \\ \textbf{subject to} & \dfrac{|x_i|}{\delta} \le t_i + Ms_i,\ \forall i \in [n] \\ & 1 \le t_i + M(1-s_i),\ \forall i \in [n] \\ & \ell \le x \le u,\ \textbf{and}\ \ s \in \{0,1\}^n. \end{array} \quad \text{where} \quad \left\{ \begin{array}{l} M \triangleq \left\{ \begin{array}{ll} \max\left\{ \dfrac{2\|\tilde{q}\|_2}{\delta\lambda_{\min}(Q)},\ 1 \right\} & \text{if uncons.} \\ \dfrac{u_1}{\delta}\ (=4) & \text{if constr.} \end{array} \right. \\ \tilde{q}_i \triangleq q_i + \text{sign}(q_i)\,\dfrac{\gamma}{\delta} \qquad \forall\, i \in [n] \\ \lambda_{\min}(Q) \triangleq \text{ minimal eigenvalue of } Q \end{array} \right.$$

The choice of M for the unconstrained cases is motivated by the fact that if x^* is an optimal solution of the capped ℓ_1-problem, then $\|x^*\|_\infty \le \dfrac{2\|q^*\|_2}{\lambda_{\min}(Q)} \le \dfrac{2\|\tilde{q}\|_2}{\lambda_{\min}(Q)}$, where $q_i^* \triangleq q_i + \frac{\gamma}{\delta}p_i$ if $0 \le x_i^* < \delta$, $q_i^* \triangleq q_i + \frac{\gamma}{\delta}p_i$ if $-\delta < x_i^* < 0$ and $q_i^* \triangleq q_i$ if $|x_i^*| > \delta$. In Table 2, we record the best objective values found by GUROBI when it is set to stop at the time limit of 200 and 600 s. Additionally, we also initialize GUROBI with x_{ghp} and rerun it for 600 seconds, with the results included in the column named "GRB cont.". Finally, we also present the gaps between the upper and lower bounds of global optimality when GUROBI is terminated.

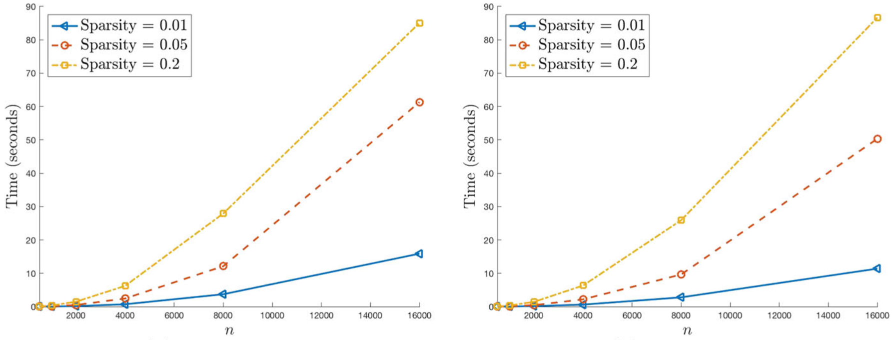
It is apparent that for all the cases except for the unconstrained, $\sigma = 0.01$ scenario,the d-stationary points x_{ghp} achieved by GHP (in 2 s for $\sigma = 0.01$, and 20 s for $\sigma = 0.2$) are better than incumbents found by GUROBI in 200 seconds. On the other hand, if the time limit is set as 600 seconds, GUROBI usually finds feasible solutions whose objective values are close to (but slightly worse) than the objective value of x_{ghp}, while requiring significantly more time. Finally, we note that the optimality gaps proven from GUROBI serve as a certificate that, for $\sigma = 0.01$, GHP delivers solutions whose objective is (at most) 1% worse than the global minimizer of the problem.

8.3 Verifying strongly polynomial complexity

To verify the strongly polynomial complexity of our method, we test it on the capped ℓ_1-problems, both constrained and unconstrained, with various sizes $n \in \{500, 1000, 2000, 4000, 8000, 16{,}000\}$ and with the off-diagonal sparsity of Q being $\sigma \in \{0.01, 0.05, 0.2\}$. Similar to previous experiments, γ is fixed to 1 and δ, u, ℓ are chosen identically. The computing time and the total number of subproblems (19) solved in Algorithm II (labelled as steps II) are summarized in Table 3. We also plot the computing time as a function of n under various settings in Fig. 2. The results indicate that doubling n results in an $4 - 6x$ increase in computational times, suggesting that the practical complexity of the GHP method is roughly of the order $n^{2.5}$ in the experiments, considerably less than the worst-case complexity of $O(n^4)$ established in Theorem 17; the total number of steps II is also significantly less than the worst-case bound of linear number of (inner) subproblems (19) stipulated in the same theorem.

Table 2 Computational results for the capped ℓ_1-problems where "Unc." ("Con.") stands for "unconstrained" ("constrained")

Settings	GHP			GRB (time lim. = 200)		GRB (time lim. = 600)		GRB cont. (time lim. = 600)	
	Obj.	Time	Steps II	Obj.	Gap (%)	Obj.	Gap (%)	Obj.	Gap (%)
Unc. $\sigma = 0.01$	−3.3732e5	1.58	13.4	−3.3732e5	> 100	−3.3732e5	0.58	−3.3732e5	0.53
Unc. $\sigma = 0.2$	−1.6360e4	21.18	23.0	0	> 100	−1.6306e4	> 100	−1.6360e4	> 100
Con. $\sigma = 0.01$	−3.7314e5	1.74	13.0	0	> 100	−3.3713e5	0.88	−3.7314e5	0.42
Con. $\sigma = 0.2$	−1.6290e4	18.53	21.4	0	> 100	−1.6214e5	20.93	−1.6291e4	18.61



(a) Unconstrained **(b)** Constrained

Fig. 2 Computing time vs. size n for capped ℓ_1-problems

Table 3 Computational results for the capped ℓ_1-problems

Settings		$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$	$n = 8000$	$n = 16{,}000$
Unconstrained							
$\sigma = 0.01$	Time	0.03	0.04	0.15	0.61	2.79	11.39
	Steps II	6.4	6.4	7.4	8.2	9.6	10.2
$\sigma = 0.05$	Time	0.04	0.10	0.43	2.18	9.64	50.36
	Steps II	7.0	8.4	8.6	10.0	11.4	14.6
$\sigma = 0.2$	Time	0.07	0.32	1.44	6.33	25.92	86.68
	Steps II	8.6	9.2	10.8	12.2	14.2	14.0
Constrained							
$\sigma = 0.01$	Time	0.04	0.04	0.14	0.71	3.72	15.89
	Steps II	6.4	6.4	7.4	8.6	9.8	10.4
$\sigma = 0.05$	Time	0.03	0.10	0.49	2.45	12.12	61.28
	Steps II	7.0	8.4	9.2	10.2	12.2	15.0
$\sigma = 0.2$	Time	0.08	0.32	1.49	6.23	27.92	85.06
	Steps II	8.6	8.8	10.4	11.4	14.2	13.6

In summary, we conclude that the strongly-polynomial worst-case complexity of the proposed method translates to a practically efficient algorithm: it significantly outperforms leading off-the-shelf solvers while delivering high-precision solutions without additional computational effort.

Acknowledgements The authors would like to thank Dr. Ying Cui at the University of Minnesota for her insightful comments that have helped to improve this manuscript, and for bringing to our attention the two references [16,43]. They are also grateful to two referees who have provided constructive comments and offered additional references that have helped to improve the presentation and quality of the paper.

References

1. Adler, I., Cottle, R.W., Pang, J.S.: Some LCPs solvable in strongly polynomial time with Lemke's algorithm. Math. Progr., Ser. A **160**(1), 477–493 (2016)
2. Ahn, M., Pang, J.S., Xin, J.: Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J. Optim. **27**(3), 1637–1665 (2017)
3. Atamtürk, A., Gómez, A.: Strong formulations for quadratic optimzation with M-matrices and indicator variables. Math. Progr. Seri. B **170**, 141–176 (2018)
4. Atamtürk, A., Gómez, A., Han, S.: Sparse and smooth signal estimation: convexification of L0 formulations. J. Mach. Learn. Res. **22**, 1–43 (2021)
5. Bach, F.: Submodular functions: from discrete to continuous domains. Math. Program. **175**(1), 419–459 (2019)
6. Barlow, R.E., Bartholomew, D., Bremmer, J.M., Brunk, H.D.: Statistical Inference Under Order Restrictions: The Theory and Application of Order Regression. Wiley, New York (1972)
7. Bennett, K.P., Kunapuli, G., Hu, J., Pang, J.S.: Bilevel optimization and machine learning. In: Computational Intelligence: Research Frontiers. Lecture Notes in Computer Science, vol. 5050, pp. 25–47 (2008)
8. Bertsimas, D., Cory-Wright, R.: A scalable algorithm for sparse portfolio selection. arXiv preprint (2018). arXiv:1811.00138
9. Bian, W., Chen, X.: A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. SIAM J. Numer. Anal. **58**(1), 858–883 (2020)
10. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. J. Four. Anal. Appl. **14**, 629–654 (2008)
11. Cai, B., Zhang, A., Stephen, J.M., Wilson, T.W., Calhoun, V.D., Wang, Y.P.: Capturing dynamic connectivity from resting state FMRI using time-varying graphical lasso. IEEE Trans. Biomed. Eng. **66**(7), 1852–1862 (2018)
12. Candès, E.J., Watkins, M.B., Boyd, S.P.: Enhancing sparsity by reweighted ℓ_1 minimization. J. Four. Anal. Appl. **14**, 877–905 (2008)
13. Chandrasekaran, R.: A special case of the complementary pivot problem. Opsearch **7**, 263–268 (1970)
14. Chen, T.W., Wardill, T., Sun, Y., Pulver, S., Renninger, S., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M., Jayaraman, V.: Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature **499**, 295–300 (2013)
15. Chen, X.: Smoothing methods for nonsmooth, novonvex minimization. Math. Progr. **134**, 71–99 (2012)
16. Chen, Y., Ge, D., Wang, M., Wang, Z., Ye, Y., Yin, H.: Strong NP-hardness for sparse optimization with concave penalty functions. In: Proceedings of the 34 the International Conference on Machine Learning, Sydney, Australia, PMLR 70 (2017)
17. Chen, X., Ge, D., Wang, Z., Ye, Y.: Complexity of unconstrained L2-Lp minimization. Math. Progr. **143**, 371–383 (2014)
18. Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. SIAM J. Sci. Comput. **32**, 2832–2852 (2010)
19. Chen, X., Zhou, W.: Convergence of the reweighted ℓ_1 minimization algorithm for ℓ_2-ℓ_p minimization. Comput. Optim. Appl. **59**, 47–61 (2014)
20. Cottle, R.W., Pang, J.S.: On solving linear complementarity problems as linear programs. Math. Progr. Study **7**, 88–107 (1978)
21. Cottle, R.W., Pang, J.S., Stone, R.E.: The linear complementarity problem, vol. 60. SIAM Classics in Applied Mathematics, Philadelphia (2009) [Originally published by Academic Press, Boston (1992)]
22. Cottle, R.W., Veinott, A.F., Jr.: Polyhedral sets having a least element. Math. Progr. **3**, 23–249 (1969)
23. Cui, Y., Chang, T.H., Hong, M., Pang, J.S.: A study of piecewise-linear quadratic programs. J. Optim. Theory Appl. **186**, 523–553 (2020)
24. Cui, Y., Pang, J.S.: Modern nonconvex and nondifferentiable optimization. In: Society for Industrial and Applied Mathematics. MOS-SIAM Series on Optimization, Philadelphia (2021)
25. Dong, H., Ahn, M., Pang, J.S.: Structural properties of affine sparsity constraints. Math. Progr., Ser. B **176**(1–2), 95–135 (2018)
26. Dong, H., Chen, K., Linderoth, J.: Regularization vs. relaxation: a conic optimization perspective of statistical variable selection (2015). arXiv:1510.06083
27. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)

28. Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**(3), 819–849 (2014)
29. Fattahi, S., Gómez A.: Scalable inference of sparsely-changing Markov random fields with strong statistical guarantees. Forthcoming in NeurIPS (2021). https://proceedings.neurips.cc/paper/2021/hash/33853141e0873909be88f5c3e6144cc6-Abstract.html
30. Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual (2021). https://www.gurobi.com
31. Hallac, D., Park, Y., Boyd, S., Leskovec, J.: Network inference via the time-varying graphical lasso. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 205–213 (2017)
32. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical learning with sparsity: the Lasso and generalizations. In: Monographs on Statistics and Applied Probability, vol. 143. CRC Press (2015)
33. He, Z., Han, S., Gómez, A., Cui, Y., Pang, J.S.: Comparing solution paths of sparse quadratic minimization with a Stieltjes matrix. Department of Industrial and Systems Engineering, University of Southern California (2021)
34. Hochbaum, D.S., Lu, Ch.: A faster algorithm for solving a generalization of isotonic median regression and a class of fused Lasso problems. SIAM J. Optim. **27**(4), 2563–2596 (2017)
35. Jewell, S., Witten, D.: Exact spike train inference via $\ell 0$ optimization. Ann. Appl. Stat. **12**(4), 2457–2482 (2018)
36. Kunapuli, G., Bennett, K., Hu, J., Pang, J.S.: Classification model selection via bilevel programming. Optim. Methods Softw. **23**(4), 475–489 (2008)
37. Kunapuli, G., Bennett, K., Hu, J., Pang, J.S.: Bilevel model selection for support vector machines. In: Hansen, P., Pardolos, P. (eds.) CRM Proceedings and Lecture Notes. American Mathematical Society, vol. 45, pp. 129–158 (2008)
38. Lee, Y.C., Mitchell, J.E., Pang, J.S.: Global resolution of the support vector machine regression parameters selection problem with LPCC. EURO J. Comput. Optim. **3**(3), 197–261 (2015)
39. Lee, Y.C., Mitchell, J.E., Pang, J.S.: An algorithm for global solution to bi-parametric linear complementarity constrained linear programs. J. Glob. Optim. **62**(2), 263–297 (2015)
40. Le Thi, H.A., Pham Dinh, T., Vo, X.T.: DC approximation approaches for sparse optimization. Eur. J. Oper. Res. **244**(1), 26–46 (2015)
41. Liu, H., Yao, T., Li, R., Ye, Y.: Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. Math. Progr. **166**, 207–240 (2017)
42. Lu, Z., Zhou, Z., Sun, Z.: Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization. Math. Progr. **176**(1–2), 369–401 (2019)
43. Mairal, J., Yu, B.: Complexity analysis of the Lasso regularization path. In: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK (2012)
44. Mangasarian, O.L.: Linear complementarity problems solvable by a single linear program. Math. Progr. **10**, 263–270 (1976)
45. Moré, J., Rheinboldt, W.C.: On P- and S-functions and related classes of nonlinear mappings. Linear Algebra Appl. **6**, 45–68 (1973)
46. Mosek ApS. The MOSEK optimization toolbox for MATLAB manual. Version 9.3 (2019). http://docs.mosek.com/9.3/toolbox/index.html
47. Pan, L., Chen, X.: Group sparse optimization for images recovery using capped folded concave functions. SIAM J. Image Sci. **14**(1), 1–25 (2021)
48. Pang, J.S.: On a class of least-element linear complementarity problems. Math. Progr. **16**, 111–126 (1979)
49. Pang, J.S.: Leaast-element complementarity theory. Ph.D. Thesis. Department of Operations Research, Stanford University (1976)
50. Pang, J.S., Chandrasekaran, R.: Linear complementarity problems solvable by a polynomially bounded pivoting algorithm. Math. Progr. Study **25**, 13–27 (1985)
51. Pang, J.S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth dc programs. Math. Oper. Res. **42**, 95–118 (2017)
52. Rheinboldt, W.C.: On M-functions and their applications to nonlinear Gauss-Seidel iterations and to network flows. J. Math. Anal. Appl. **32**, 274–307 (1970)
53. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970)
54. Tamir, A.: Minimality and complementarity properties associated with Z-functions and M-functions. Math. Progr. **7**, 17–31 (1974)

55. Tibshirani, R.J., Hoefling, H., Tibshirani, R.: Nearly-isotonic regression. Technometrics **53**(1), 54–61 (2011)
56. Vogelstein, J.C., Packer, A.M., Machado, T.A., Sippy, T., Babadi, B., Paninski, L.: Fast nonnegative deconvolution for spike train inference from population calcium imaging. J. Neurophysiol. **6**, 3691–3704 (2010)
57. Ye, Y.: On the complexity of approximating a KKT point of quadratic programming. Math. Progr. **80**, 195–211 (1998)
58. Zhang, C.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Andrés Gómez[1] · Ziyu He[1] · Jong-Shi Pang[1]

✉ Ziyu He
ziyuhe@usc.edu

Andrés Gómez
gomezand@usc.edu

Jong-Shi Pang
jongship@usc.edu

[1] Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA