Math. Program., Ser. B (2019) 176:95–135
https://doi.org/10.1007/s10107-018-1283-3

FULL LENGTH PAPER

Structural properties of affine sparsity constraints

Hongbo Dong[1] · **Miju Ahn**[2] · **Jong-Shi Pang**[2]

Received: 12 May 2017 / Accepted: 26 April 2018 / Published online: 4 May 2018

Abstract We introduce a new constraint system for sparse variable selection in statistical learning. Such a system arises when there are logical conditions on the sparsity of certain unknown model parameters that need to be incorporated into their selection process. Formally, extending a cardinality constraint, an affine sparsity constraint (ASC) is defined by a linear inequality with two sets of variables: one set of continuous variables and the other set represented by their nonzero patterns. This paper aims to study an ASC system consisting of finitely many affine sparsity constraints. We investigate a number of fundamental structural properties of the solution set of such a non-standard system of inequalities, including its closedness and the description of its closure, continuous approximations and their set convergence, and characterizations of its tangent cones for use in optimization. Based on the obtained structural properties of an ASC system, we investigate the convergence of B(ouligand) stationary solutions when the ASC is approximated by surrogates of the step ℓ_0-function commonly employed in sparsity representation. Our study lays a solid mathematical foundation

The research of the second and the third authors were partially supported by the U.S. National Science Foundation Grant IIS-1632971.

✉ Hongbo Dong
hongbo.dong@wsu.edu

Miju Ahn
mijuahn@usc.edu

Jong-Shi Pang
jongship@usc.edu

[1] Department of Mathematics and Statistics, Washington State University, Pullman, USA

[2] Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, USA

for solving optimization problems involving these affine sparsity constraints through their continuous approximations.

Keywords Sparsity systems · Set convergence · Nonconvex optimization · B(ouligand)stationary points

Mathematics Subject Classification 90C26 · 90C90 · 62J05

1 Introduction

Judicious variable selection is an important task in statistical modeling. Practitioners wish to explain the data in the simplest way, i.e., following the parsimony principle and omitting redundant/unnecessary predictor variables that may add noise in the prediction or estimation of the essential quantities. In addition, for highly ill-posed problems, or applications where model interpretability is a major concern, context-specific assumptions and domain knowledge may require regularization for stable estimation and development of an interpretable model.

A popular approach for simultaneous variable selection and parameter estimation is to solve an optimization problem with two criteria. One is a loss (or residual) function measuring "model accuracy", i.e., how well the model fits available data sets, while the other criterion is a "penalty function" aiming to reduce "model complexity", i.e., obeying the parsimony principle that simpler is better. One complexity measure is the number of nonzero variables in the model; in this case, the goal of the penalty function is to promote the sparsity of the unknown parameters to be estimated. Early examples in the statistics literature include the lasso [31], fused lasso [32] and covariance selection for Gaussian graphical models [13,17], all leading to convex optimization problems; see the monograph [19] for more variations of lasso. Nonconvex penalty functions have also been proposed and studied, among which are the *smoothly clipped absolute deviation* (SCAD) [15], the *minimax concave penalty* (MCP) [35], and *Capped-*ℓ_1 [36], to name a few. Such nonconvex methods have been shown to enjoy several theoretical and practical advantages over convex methods in increasingly more applied contexts. The recent paper [1] presents a unified formulation of many of the most commonly used nonconvex penalty functions employed in sparsity representation and provides references for their applications.

The present paper is motivated by the increasingly advanced modeling of the variable selection process. Namely, in many applications, logical conditions among the variables need to be considered, in order to ensure a meaningful/interpretable statistical model, or to exploit domain knowledge to increase model fidelity. Such logical conditions could be that certain variables are allowed to become active only if certain other variables are selected in the model [6], or that the selection must respect given groupings of the variables [20] for which there may be different application dependent stipulations such as "within group sparsity" or "groupwise sparsity". For a recent application of such a group selection problem for building an integrative model in genetic data analysis, see [27]. Current approaches in the statistics literature on the modeling of such logical conditions are mostly done in an ad hoc way and rely mainly

on convex formulations for computational ease. As it is well-known in the integer programming literature [25], such modeling simplifications are fundamentally flawed and could easily yield inaccurate descriptions of the true logical connections among the model variables. In this paper, we propose a rigorous framework as an attempt to address such problems in a faithful way, by first modeling the logical conditions exactly and then understanding what needs to be done in order to make the resulting formulations computationally tractable for subsequent optimization and analysis, and ultimately, for statistical inference. To be specific, we propose to model the logical conditions on sparsity with a new type of constraints which we call *affine sparsity constraints* (ASC). Such constraints are derived from a linear system of inequalities where some of the continuous variables are individually replaced by the ℓ_0-function of a scalar variable: $|t|_0 \triangleq \begin{cases} 1 \text{ if } t \neq 0 \\ 0 \text{ otherwise} \end{cases}$. A most significant departure of the resulting ASC system from a standard system of linear inequalities in linear or (mixed) integer programming is that the former system involves the discontinuous ℓ_0-function and thus the solution set may not be closed. This feature immediately challenges the optimization over such constraints and calls for suitable approximations by continuous functions. Inspired by the family of surrogate sparsity functions [1] employed to approximate the ℓ_0-function, we are led to investigate the approximation of an ASC by replacing the binary indicator function by a (continuous) surrogate sparsity function. Understanding the convergence of such an approximated continuous system to the given discontinuous ASC system is a primary goal of this paper. Prior to addressing this convergence issue, we obtain a necessary and sufficient condition for the solution set of an ASC system to be closed, and derive an explicit expression of the closure of such a set in general. Overall, this paper is devoted to the study of some fundamental structural properties of affine sparsity constraints in preparation for subsequent research of solving optimization problems subject to these constraints.

This paper is organized as follows. In Sect. 2, we formally define affine sparsity constraints, discuss some of their elementary features, and describe the applied problems that they model. We address the closure properties of the solutions of these constraints in Sect. 3, giving necessary and sufficient conditions for the closedness of the solution set and two representations of its closure: one in terms of some auxiliary integer variables and the other in terms of the continuous relaxations of the latter variables, under an appropriate assumption. Section 4 investigates the approximation of an ASC by replacing the ℓ_0-function by the family of univariate folded concave functions commonly employed in statistical learning and sparsity optimization, and presents set convergence results of the approximations. In Sect. 5, we derive representations of the tangent cones of the solution set of the ASC and its approximations, and present conditions under which the tangent cones of the approximate constraint sets can be characterized (algebraically) as a finite union of convex sets. These tangent results are then used in the last Sect. 6 to establish important convergence properties of the directional stationary solutions of approximations of an ASC-constrained optimization problem.

Notation We define some notations used in this paper. Based on the binary unvariate ℓ_0-function, we define the multivariate ℓ_0-function $\|\bullet\|_0 : \mathbb{R}^n \to \{0, 1\}^n$ by $\|x\|_0 \triangleq$

$(|x_i|_0)_{i=1}^n$ for any $x \in \mathbb{R}^n$. For a given subset $\mathscr{J}$ of $\{1, \cdots, n\}$ and a vector $x \in \mathbb{R}^n$, $x_{\mathscr{J}}$ denotes the sub-vector of x with components indexed by the elements of $\mathscr{J}$. For a matrix $A \in \mathbb{R}^{m \times n}$ and the same subset $\mathscr{J}$, $A_{\bullet \mathscr{J}}$ denotes the columns of A indexed by $\mathscr{J}$. A similar notation applies to the rows of A. The support of a vector x is the index set of nonzero components of x and is denoted by $\text{supp}(x)$. For any matrix $A \in \mathbb{R}^{m \times n}$, the matrices A^+ and A^- are the entry-wise nonnegative and nonpositive parts of A, such that for all i and j,

$$[A^+]_{ij} \triangleq \max(A_{ij}, 0) \quad \text{and} \quad [A^-]_{ij} \triangleq \max(-A_{ij}, 0). \tag{1}$$

Thus we have the decomposition: $A = A^+ - A^-$ where $A^{\pm}$ are nonnegative matrices. A similar definition applies to a vector x. For any $x \in \mathbb{R}^n$, $\mathbb{B}_n(x, r)$ is an open ball (in a suitable norm) centered at x with radius r. For any set $S \subseteq \mathbb{R}^n$, $\text{cl}(S)$ denotes its closure. A vector of all ones of a given dimension is written as $\mathbf{1}$ with the dimension omitted in the notation.

2 ASC systems: introduction and preliminary discussion

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \mathbb{R}^m$, the ASC system is defined as the problem of finding a vector $\beta \in \mathbb{R}^n$ such that

$$A \parallel \beta \parallel_0 \leq b \Leftrightarrow \left\{ \sum_{j=1}^{n} A_{ij} \mid \beta_j \mid_0 \leq b_i, \text{ for all } i = 1, \cdots, m \right\}. \tag{2}$$

The (possibly empty) solution set of this system is denoted by SOL-ASC(A, b). Clearly, the system (2) can be written as:

$$A^+ \parallel \beta \parallel_0 \leq A^- \parallel \beta \parallel_0 + b.$$

A particularly important special case is when A is 0–1 matrix, i.e., all its entries are either 0 or 1, and b is a positive integral vector. In this case, the constraint $\sum_{j=1}^n A_{ij} \mid \beta_j \mid_0 \leq b_i$ becomes

$$\sum_{j \in \mathscr{I}_i} \mid \beta_j \mid_0 \leq b_i, \text{ where } \mathscr{I}_i \triangleq \{ j \mid A_{ij} \neq 0 \},$$

which is an *upper cardinality constraint* stipulating that for each $i = 1, \cdots, m$, the number of nonzero components β_j for all $j \in \mathscr{I}_i$ is no more than the given cardinality b_i (assumed integer). Cardinality constraints of this type have been studied in [5, 7,9,16,38] using various reformulations; applications of such constraints to sparse portfolio selection can be found in [5,7,8,10].

Needless to say, the ASC(A, b) is closely related to the system of linear inequalities: $Aw \leq b$, $w \in \mathbb{R}^n$. Nevertheless, there are many important differences. One obvious

difference is that the solution set of the former system is a cone possibly with the origin omitted, i.e., the following clearly holds:

$$\beta \in \text{SOL-ASC}(A, b) \Rightarrow \tau\,\beta \in \text{SOL-ASC}(A, b), \forall \text{ scalars } \tau \neq 0;$$

thus, under no other restriction on β, a non-zero SOL-ASC(A, b) is always an unbounded set. In contrast, a polyhedron, which is the solution set of a system of linear inequalities, does not have the scaling property in general. Further, an important feature that distinguishes the general ASC(A, b) where A can have positive and negative entries from the special case of a cardinality constraint system where A is nonnegative is that the solution set of the latter system must be closed (due to the lower semi-continuity of the ℓ_0-function) while SOL-ASC(A, b) is not necessarily so in the general case, as illustrated by the simple Example 1 below.

Example 1 Consider the simple case where $A = [\,1\ -1\,]$ and $b = 0$, yielding the system $|\beta_1|_0 \leq |\beta_2|_0$. It is not difficult to see that SOL-ASC(A, b) is the entire plane $\mathbb{R}^2$ except for the two half β_1-axes; i.e., SOL-ASC$(A, b) = \left\{(\beta_1, \beta_2) \in \mathbb{R}^2 \mid \beta_2 \neq 0\right\} \cup \{(0, 0)\}$, which is obviously not closed. □

As we shall see later, the non-closedness of SOL-ASC(A, b) is due to the presence of some negative entries in the matrix A. An extreme case of this is when the entries of A are all either 0 or -1 and b is a negative integral vector. In this case, we obtain a *lower cardinality constraint* of the form:

$$\sum_{j \in \mathscr{I}_i} |\,\beta_j\,|_0 \geq |\,b_i\,|, \text{ where } \mathscr{I}_i \triangleq \{\, j \mid A_{ij} < 0 \,\},$$

that has minimally been studied in the literature to date. By imposing no sign restrictions on the matrix A, our treatment goes far beyond these special cases and accommodates recent interests in statistical variable selection subject to logical constraints. Another important difference between SOL-ASC and a polyhedron is their respective tangent cones; see Sect. 5 for details.

It is natural to consider an extension of the ASC by including continuous variables; specifically, let $b : \mathbb{R}^k \to \mathbb{R}^m$ be a given mapping and let Γ be a closed convex set in $\mathbb{R}^k$. Defined by the triplet (A, b, Γ), the extended ASC (xASC) system is the problem of finding a pair $(\beta, \gamma) \in \mathbb{R}^n \times \Gamma$ such that

$$A \,\|\, \beta \,\|_0 \leq b(\gamma).$$

The (possibly empty) solution set of this system is denoted by SOL-xASC(A, b, Γ). Subsequently, we will discuss how results of the ASC system can be extended to the xASC system, and show how these extended systems could arise in the approximation of the ASC (see Sect. 4.2).

2.1 Source problems

In general, if $\mathbf{M} \subseteq \{0, 1\}^n$ is a subset of binary vectors containing all admissible vectors of $\|\beta\|_0$, then $\|\beta\|_0 \in \mathbf{M}$ is equivalent to $\|\beta\|_0 \in \text{conv}(\mathbf{M})$, which in principle can be formulated as an ASC system as conv($\mathbf{M}$) is a polytope. More specifically, logical conditions on the sparsity of the model variables can be modeled by affine constraints using the binary ℓ_0-indicators of these variables. In what follows we present two models of statistical regression with logical conditions on the unknown parameters.

Hierarchical variable selection Consider the following regression model with interaction terms [6]:

$$y = \sum_{i=1}^{n} \beta_i^{(1)} x_i + \sum_{1 \le i < j \le n} \beta_{ij}^{(2)} x_i x_j + \varepsilon, \tag{3}$$

where $x \in \mathbb{R}^n$ is the vector of model inputs, $y \in \mathbb{R}$ is the (univariate) model output, the $\beta_i^{(1)}$ and $\beta_{ij}^{(2)}$ are the unknown model parameters to be estimated, and ε is the (random) error of the model. It is common practice in the variable selection process to maintain certain hierarchical conditions (also called "heredity constraints" or "marginality" in the literature [6,18,24]) between the coefficient of the linear terms, $\beta_i^{(1)}$, and those in the interaction terms, $\beta_{ij}^{(2)}$. There are two types of hierarchical conditions. The *strong* hierarchical condition means that an interaction term can be selected only if both of the linear terms are selected, i.e., $|\beta_{ij}^{(2)}|_0 \le \min\left\{|\beta_i^{(1)}|_0, |\beta_j^{(1)}|_0\right\}$ for any $i < j$, while the *weak* hierarchical condition means that an interaction term can be selected only if one of the corresponding linear terms is selected, i.e., $|\beta_{ij}^{(2)}|_0 \le |\beta_i^{(1)}|_0 + |\beta_j^{(1)}|_0$. Clearly, both conditions can be represented by linear inequalities in $\|\beta\|_0$, where β is the concatenated vector of $\beta^{(1)}$ and $\beta^{(2)}$.

The hierarchical variable selection problem has received much attention in the statistics literature; for its treatment, various convex relaxations of the hierarchical conditions have been proposed. For instance, in [6, Remark 1], it was suggested to employ the linear constraints: $|\beta_{ij}^{(2)}| \le \beta_{+,i}^{(1)} + \beta_{-,i}^{(1)}$ for all $i < j$, to model the constraint: $|\beta_{ij}^{(2)}|_0 \le |\beta_i^{(1)}|_0$ where $\beta_{\pm,i}^{(1)}$ are the nonnegative and nonpositive parts of $\beta_i^{(1)}$, as such, the latter two nonnegative variables ought to satisfy the important complementarity constraint: $\beta_{+,i}^{(1)} \beta_{-,i}^{(1)} = 0$. Nevertheless, the latter condition is dropped in the convex formulation. Note also the use of the absolute value $|\beta_{ij}^{(2)}|$ as a replacement for the ℓ_0-function $|\beta_{ij}^{(2)}|_0$. These two convexification maneuvers lead to a restriction of the hierarchical constraints in the sense that:

$$|\beta_{ij}^{(2)}| \le \beta_{+,i}^{(1)} + \beta_{-,i}^{(1)} \Rightarrow |\beta_{ij}^{(2)}|_0 \le |\beta_i^{(1)}|_0 \le |\beta_i^{(1)}|_0 + |\beta_j^{(1)}|_0.$$

Obviously, such simplistic convexifications can at best be considered a very crude treatment of the nonconvex hierarchial constraints; they were introduced merely to take advantage of the advances of convex programming for the optimization of an objective function subject to these constraints. In light of the recent surge in interest in using nonconvex surrogates to replace the ℓ_0-function, see e.g. [1] and the references

therein, one is led to investigate the use of these surrogates in approximating the hierarchical constraints within the broader framework of an ASC.

Several other works have developed specialized methods to attempt to induce the "hierarchical effects", based on convex penalty functions that are originally designed for grouped variable selection [34] (see discussion of the next topic). For example, the paper [37] proposed a general framework of constructing (convex) penalty functions to derive the grouping or hierarchical effects. Specialized to the strong hierarchical case, their proposed penalty function is: for $\lambda > 0$,

$$\lambda \sum_{j<k} \left[|\beta_{jk}^{(2)}| + \left\| \left(\beta_j^{(1)}, \beta_k^{(1)}, \beta_{jk}^{(2)} \right) \right\|_{\gamma_{jk}} \right] \text{where } 1 < \gamma_{jk}, \forall\, j < k$$

and $\| \bullet \|_{\gamma_{jk}}$ is the γ_{jk}-norm. Again, this is a very inaccurate way of expressing the strong hierarchical condition. Other related formulations can be found in [2], where combinations of convex functions are employed.

In summary the existing attempts are commendable as a first step towards the treatment of the hierarchical constraints; obviously there is much room for further studies and improvement in the treatment. Our work is a step in the latter direction.

Group variable selection In many applications, variables are naturally divided into groups. There are different versions of grouped variable selection, and a selective review is given in [20]; see also [21,34]. Here we consider two versions of this problem. One version is a variation of the group lasso discussed in [19, Section 4.3] where it is stated that: *It is desirable to have all coefficients within a group to become zero (or nonzero) simultaneously.* The formulation in the cited reference is as follows: instead of the basic linear regression model:

$$y = \sum_{i=1}^{n} \beta_i x_i + \varepsilon, \tag{4}$$

that expresses the model output directly as a combination of the core predictors $\{x_i\}_{i=1}^n$, an aggregate model:

$$y = \sum_{j=1}^{J} (z^j)^T \theta^j + \varepsilon'$$

in terms of the J group $\{\theta^j\}_{j=1}^J$ of (unknown) variates is postulated, where each $\theta^j \in \mathbb{R}^{p_j}$ represents a group of p_j regression coefficients among the β_i's, with the vectors z^j being the known covariates in group j. A convex least-squares objective function [19, expression (4.5)] that contains the penalty $\sum_{j=1}^J \|\theta^j\|_2$ is minimized whereby the unknown variables θ^j are obtained. While this provides a plausible approach to the group selection process, it does not exactly model the desired grouping as stipulated above with reference to the basic model (4) that is at the level of the individual predictors x_i's for $i = 1, \cdots, n$. Instead, we may minimize a combined (Lagrangian)

objective $f_\lambda(\beta) \triangleq \ell(\beta) + \lambda P(\beta)$, which comprises of a loss function weighed by a sparsity penalty function, both in terms of the original variable $\beta \triangleq (\beta_i)_{i=1}^n$, subject to the grouping conditions that can be formulated as follows. For each $j = 1, \cdots, J$, let $\mathscr{G}_j$ be the subset of $\{1, \cdots n\}$ containing indices i such that the variable β_i in group j; thus $\{1, \cdots, n\} = \bigcup_{j=1}^J \mathscr{G}_j$. Consider the system with some auxiliary group variables ζ_j:

$$| \beta_i |_0 = \zeta_j \forall\, i \in \mathscr{G}_j \text{ and } j = 1, \cdots, J,$$

which can easily be seen to model exactly the desired grouping requirement. Clearly, the above is an xASC system in the pair of variables (β, ζ). Note that no constraints are imposed on ζ_j. When the ℓ_0-function is subsequently approximated by a surrogate function, properties of such a function (e.g., nonnegativity and upper bounds) will naturally transfer to restrictions on ζ_j.

Consider the alternative stipulation of choosing the variables β_i so that the number of groups covering all nonzero components is minimal. Together with the Lagrangian function $f_\lambda(\beta) \triangleq \ell(\beta) + \lambda P(\beta)$, the optimization problem of variable selection is: given positive coefficients $\{c_j\}_{j=1}^J$,

$$\begin{aligned} &\underset{\beta,\zeta}{\text{minimize}} \;\; f_\lambda(\beta) + \sum_{j=1}^J c_j\, \zeta_j \\ &\text{subject to } | \beta_i |_0 \le \sum_{j\,:\, i \in \mathscr{G}_j} \zeta_j, \; \forall\, i = 1, \cdots, n \\ &\text{and} \qquad \zeta_j \in \{0, 1\}, \qquad j = 1, \cdots J, \end{aligned} \tag{5}$$

where each xASC models the coverage of the variate β_i in the groups that contain them; i.e., the variable β_i is selected only if at least one of the groups containing predictor i is selected. The minimization of the (weighed) sum of the binary variables ζ_j is a slight generalization of the goal of selecting the minimum number of groups in the coverage of all the nonzero β_i's which is the special case with equal weights. Clearly, the relaxation of the binary condition $\zeta_j \in \{0, 1\}$ to the continuous condition $\zeta_j \in [0, 1]$ leads to an xASC system in the pair of variables (β, ζ).

A special case: tree structure In what follows, we discuss a special case of the group selection problem that justifies the relaxation of the 0–1 restriction on the group indicator variables ζ_j in the problem (5). The key to this justification is the fact that for *fixed* β, the constraints of the problem are those of a well-known set covering problem [25, Part III]. As such the theory of the latter problem can be applied. In turn, this application is based on the theory of balanced matrices which we quickly review; see [11,12]. A 0–1 matrix is *balanced* matrix if and only if for any odd number $k \ge 3$, it does not contain a $k \times k$ submatrix whose row sums and column sums all equal 2 and which does not contain the 2×2 submatrix $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. A 0–1 matrix is called *totally balanced* (TB) if such condition holds for all $k \ge 3$ (not just odd k). A sufficient

condition for a matrix to be TB is that it does not contain a submatrix $F = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ [25, Part III, Proposition 4.5]. (In fact this condition is also necessary up to permutations of rows and columns [25, Part III, Proposition 4.8].) Clearly the TB property is invariant under permutations of rows and columns.

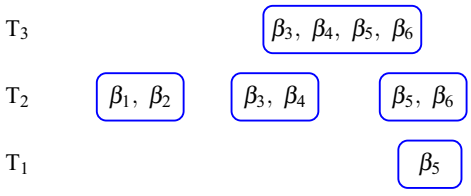
The model discussed below is related to the tree structured group lasso presented in [33]. Suppose that any two distinct members of the family $\mathbf{G} \triangleq \{\mathcal{G}_j\}_{j=1}^{J}$ are such that either they do not intersect or one is a proper subset of the other. In this case (which clearly includes the special case of non-overlapping groups; i.e., $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ for $i \neq j$), we can arrange the groups in a tree structure as follows. Let $d \triangleq \max_{1 \leq j \leq J} |\mathcal{G}_j|$ be the largest number of elements contained in the individual groups. For each integer $k = 1, \cdots, d$, let $T_k = \{\mathcal{G}_1^k, \cdots, \mathcal{G}_{q_k}^k\}$ be the sub-family of $\mathbf{G}$ consisting of all (distinct) groups each with exactly k variables; thus each $\mathcal{G}_r^k$ is one of the $\mathcal{G}_j$. For simplicity, assume that each T_k is a non-empty sub-family for $k = 1, \cdots, d$; thus $J = \sum_{k=1}^{d} q_k$. It then follows that any two different members of T_k do not overlap, i.e., for any $k = 1, \cdots, d$ and $r \neq r'$ both in $\{1, \cdots, q_k\}$, $\mathcal{G}_r^k \cap \mathcal{G}_{r'}^k = \emptyset$. Furthermore, for each pair (k, k') satisfying $1 \leq k' < k \leq d$, we call $\mathcal{G}_r^k$ a parent node of $\mathcal{G}_{r'}^{k'}$ if $\mathcal{G}_{r'}^{k'}$ is a proper subset of $\mathcal{G}_r^k$. Thus a parent node contains more elements than its descendent node(s). Note that it is not necessary for every element in the sub-family T_k to have a parent node in the sub-family $T_{k'}$, nor is it necessary for every element in $T_{k'}$ to have a parent node in T_k. We define the element-group incidence matrix $E \in \mathbb{R}^{n \times J}$ as follows. Arrange the columns of E in the order of ascendency of the groups in T_k for $k = 1, \cdots, d$; i.e.,

$$\underbrace{\mathcal{G}_1^1, \cdots, \mathcal{G}_{q_1}^1}_{T_1}; \underbrace{\mathcal{G}_1^2, \cdots, \mathcal{G}_{q_2}^2}_{T_2}; \cdots; \underbrace{\mathcal{G}_1^{d-1}, \cdots, \mathcal{G}_{q_{d-1}}^{d-1}}_{T_{d-1}}; \underbrace{\mathcal{G}_1^d, \cdots, \mathcal{G}_{q_d}^d}_{T_d};$$

then let $E_{ij} = \begin{cases} 1 \text{ if } i \text{ is contained in group } j \\ 0 \text{ otherwise} \end{cases}$. In terms of this matrix E, the constraints:

$$\sum_{j\,:\,i \in \mathcal{G}_j} \zeta_j \geq |\beta_i|_0 \forall\, i \in \text{supp}(\beta)$$

can be written simply as $E_{S\bullet}\zeta \geq \mathbf{1}_{|S|}$ where $S \triangleq \text{supp}(\beta)$ and $E_{S\bullet}$ denotes the rows of E indexed by S.



	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$
β_1	0	1	0	0	0
β_2	0	1	0	0	0
β_3	0	0	1	0	1
β_4	0	0	1	0	1
β_5	1	0	0	1	1
β_6	0	0	0	1	1

An illustration of tree structure and the element-group incidence matrix E
6 variables, 3 levels, and 5 groups:
$\mathscr{G}_1 = \{\beta_5\}$, $\mathscr{G}_2 = \{\beta_1, \beta_2\}$, $\mathscr{G}_3 = \{\beta_3, \beta_4\}$, $\mathscr{G}_4 = \{\beta_5, \beta_6\}$, and $\mathscr{G}_5 = \{\beta_3, \beta_4, \beta_5, \beta_6\}$

Proposition 1 *Let the family of groups $\{\mathscr{G}_j\}_{j=1}^J$ have the tree structure defined above. For any $\beta \in \mathbb{R}^n$ and any nonnegative coefficients $\{c_j\}_{j=1}^J$, it holds that*

$$\min\left\{\sum_{j=1}^J c_j\,\zeta_j \mid \zeta \in \{0,1\}^J,\ \sum_{j\,:\,i\in\mathscr{G}_j}\zeta_j \geq 1,\ \forall i \in \mathit{supp}(\beta)\right\}$$
$$= \min\left\{\sum_{j=1}^J c_j\,\zeta_j \mid \zeta \in [0,1]^J,\ \sum_{j\,:\,i\in\mathscr{G}_j}\zeta_j \geq 1,\ \forall i \in \mathit{supp}(\beta)\right\}.$$

Proof Let β be given. Suppose that the matrix $E_{S\bullet}$ has a submatrix $F \triangleq \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ corresponding to rows $i < i'$ and columns $j < j'$. Then i is contained in both groups corresponding to columns j and j'. By the tree structure assumption and the ordering of the columns of E, this can only happen if the group at column j is a descendent node of the group at column j'. However, since $E_{i'j} = 1$ and $E_{i'j'} = 0$, element i' is in the descendent group but not in the parent group, hence a contradiction. Thus the constraint matrix $E_{S\bullet}$ is TB. By [25, Part III, Thm 4.13], the polyhedron $\{\zeta \geq 0 \mid E_{S\bullet}\zeta \geq \mathbf{1}_{|S|}\}$ is integral; i.e., all its extreme points are binary. This is enough to establish the desired equality of the two minima. □

3 Closedness and closure

The closedness of a feasible region is very important in constrained optimization because the minimum value of a lower semi-continuous objective function may not be attained on a non-closed feasible region, even if the objective has bounded level sets in the region. This section addresses the closedness issue of an (x)ASC system and derives an expression of the closure of SOL-(x)ASC.

3.1 Necessary and sufficient conditions for closedness

It is convenient to define two related sets of binary vectors which capture the "combinatorial structures" of SOL-ASC(A, b):

$$\begin{aligned} &\mathscr{Z}(A,b) \triangleq \{\, z \in \{0,1\}^n \mid Az \leq b \,\} \\ \text{and } &\mathscr{Z}^{\leq}(A,b) \triangleq \{\, z \in \{0,1\}^n \mid \exists\, \bar{z} \in \{0,1\}^n \text{ such that } A\bar{z} \leq b \text{ and } z \leq \bar{z} \,\}. \end{aligned} \tag{6}$$

The latter set $\mathscr{Z}^{\leq}(A, b)$ contains binary vectors obtained by "zeroing out" some entries of the vectors in $\mathscr{Z}(A, b)$. These two sets provide the intermediate step to characterize

the closedness of SOL-ASC(A, b). The following proposition states the relationship between $\mathscr{Z}(A, b)$ and SOL-ASC(A, b). No proof is needed.

Proposition 2 *It holds that*

$$\begin{aligned}
&\{0, 1\}^n \cap \textit{SOL-ASC}(A, b) \\
&= \mathscr{Z}(A, b) \subseteq \underbrace{\textit{SOL-ASC}(A, b) = \{\beta \mid \|\beta\|_0 \in \mathscr{Z}(A, b)\}}_{\textit{to be extended to closure of SOL-ASC(A,b)}} \\
&= \bigcup_{d\,:\,d_i \neq 0,\,\forall\, i} \{d \circ z \mid z \in \mathscr{Z}(A, b)\}.
\end{aligned}$$

Moreover, the inclusion is strict if $\mathscr{Z}(A, b)$ is neither the empty set nor the singleton $\{0\}$. □

Clearly, $\mathscr{Z}(A, b)$ is always closed and bounded, while SOL-ASC(A, b) needs not be closed and must be unbounded unless it is the singleton $\{0\}$. The following result provides sufficient and necessary conditions for SOL-ASC(A, b) to be a closed set.

Proposition 3 *Suppose SOL-ASC$(A, b) \neq \emptyset$. The following three statements are equivalent.*

(a) *SOL-ASC(A, b) is closed.*
(b) $\mathscr{Z}(A, b) = \mathscr{Z}^{\leq}(A, b)$.
(c) *SOL-ASC*$(A, b) = \left\{\beta \mid A^+\|\beta\|_0 \leq b\right\}$ *and* $\mathscr{Z}(A, b) = \left\{z \in \{0, 1\}^n \mid A^+ z \leq b\right\}$.

Proof (a) $\Rightarrow$ (b). It suffices to prove $\mathscr{Z}^{\leq}(A, b) \subseteq \mathscr{Z}(A, b)$. By way of contradiction, suppose there exist binary vectors $z \leq \bar{z}$ with $\bar{z} \in \mathscr{Z}(A, b) \not\ni z$. Clearly $\mathrm{supp}(z) \subsetneqq \mathrm{supp}(\bar{z})$. Define the binary vector β^k with components

$$\beta_i^k \triangleq \begin{cases} 1/k \text{ if } i \in \mathrm{supp}(\bar{z}) \setminus \mathrm{supp}(z) \\ \bar{z}_i \quad \text{otherwise.} \end{cases}$$

It then follows that $\|\beta^k\|_0 = \bar{z} \in \mathscr{Z}(A, b) \subseteq$ SOL-ASC(A, b). Moreover, $\lim_{k\to\infty} \beta^k = z \notin \mathscr{Z}(A, b)$. This contradicts the closedness of SOL-ASC(A, b).

(b) $\Rightarrow$ (c). It suffices to show the inclusion: $\mathscr{Z}(A, b) \subseteq \left\{z \in \{0, 1\}^n \mid A^+ z \leq b\right\}$. Let $\bar{z} \in \mathscr{Z}(A, b)$. We claim that $\sum_{j=1}^n \max\left(A_{ij}, 0\right)\bar{z}_j \leq b_i$ for arbitrary $i = 1, \cdots, m$. For a given i, define a vector z such that $z_j \triangleq \begin{cases} \bar{z}_j \text{ if } A_{ij} \geq 0 \\ 0 \quad \text{if } A_{ij} < 0 \end{cases}$ which clearly satisfies $z \leq \bar{z}$. By (b), it follows that $z \in \mathscr{Z}(A, b)$. We have

$$b_i \geq \sum_{j=1}^n \max\left(A_{ij}, 0\right) z_j - \sum_{j=1}^n \max\left(-A_{ij}, 0\right) z_j = \sum_{j=1}^n \max\left(A_{ij}, 0\right)\bar{z}_j.$$

Thus the claim holds.

(c) $\Rightarrow$ (a). This is obvious because the mapping $\|\bullet\|_0$ is lower semi-continuous and A^+ contains only nonnegative entries. □

Remark 1 It is possible for SOL-ASC(A, b) to be a closed set when A contains negative entries. A trivial example is the set $\{\beta \mid -\|\beta\|_0 \leq 0\} = \mathbb{R}^n$ which corresponds to A being the negative identity matrix and $b = 0$. □

Example 2 We illustrate the two sets $\mathscr{Z}(A, b)$ and $\mathscr{Z}^{\leq}(A, b)$ using the regression model (3) with interaction terms that satisfy the strong hierarchical conditions: $|\beta_{ij}^{(2)}|_0 \leq \min\left(|\beta_i^{(1)}|_0,\ |\beta_j^{(1)}|_0\right)$ for all $i < j$ and also a cardinality constraint on $\beta^{(1)}$. In this case, we have, for some integer $K > 0$,

$$\mathscr{Z}(A, b) = \left\{ \begin{array}{c} (z^{(1)}, z^{(2)}) \mid z_{ij}^{(2)} \leq \min\left(z_i^{(1)},\ z_j^{(1)}\right), \forall\, i < j \\ \mid \sum_{i=1}^{n} z_i^{(1)} \leq K, z_{ij}^{(2)},\ z_i^{(1)} \text{ all binary} \end{array} \right\}.$$

We claim that

$$\begin{aligned} \mathscr{Z}^{\leq}(A, b) = \Bigg\{ (z^{(1)}, z^{(2)}) \mid & \sum_{i=1}^{n} \max_{j<i<\ell} \left(z_i^{(1)},\ z_{ji}^{(2)},\ z_{i\ell}^{(2)}\right) \\ & \leq K, z_{ij}^{(2)},\ z_i^{(1)} \text{ all binary} \Bigg\}. \end{aligned} \tag{7}$$

To see this, let $(z^{(1)}, z^{(2)})$ be a pair belonging to the right-hand set. Define

$$\bar{z}_i^{(1)} \triangleq \max_{j<i<\ell} \left(z_i^{(1)},\ z_{ji}^{(2)},\ z_{i\ell}^{(2)}\right).$$

It is not difficult to see that $(\bar{z}^{(1)}, z^{(2)}) \in \mathscr{Z}(A, b)$. Conversely, let $(z^{(1)}, z^{(2)})$ be a binary pair such that there exists $(\bar{z}^{(1)}, \bar{z}^{(2)}) \in \mathscr{Z}(A, b)$ such that $(z^{(1)}, z^{(2)}) \leq (\bar{z}^{(1)}, \bar{z}^{(2)})$. Then,

$$\sum_{i=1}^{n} \max_{j<i<\ell} \left(z_i^{(1)},\ z_{ji}^{(2)},\ z_{i\ell}^{(2)}\right) \leq \sum_{i=1}^{n} \max_{j<i<\ell} \left(\bar{z}_i^{(1)},\ z_{ji}^{(2)},\ z_{i\ell}^{(2)}\right) = \sum_{i=1}^{n} \bar{z}_i^{(1)} \leq K.$$

This proves that $(z^{(1)}, z^{(2)})$ belongs to the right-hand set in (7). Hence the equality in this expression holds. □

3.2 Closure of SOL-ASC

The expression (7) is interesting because the result below shows that the set $\mathscr{Z}^{\leq}(A, b)$ determines the closure of the solution set of the ASC system, not only for the regression model with interaction terms under the strong hierarchical relation among its variates, but also in general.

Proposition 4 *It holds that*

$$cl\,[\,SOL\text{-}ASC(A,b)\,]\;=\;\left\{\,\beta\;\mid\;\|\,\beta\,\|_0\;\in\;\mathscr{Z}^{\leq}(A,b)\,\right\}.$$

Proof Let $\{\beta^k\}$ be a sequence of vectors in SOL-ASC(A,b) converging to the limit β^∞. We then have $z^k \triangleq \|\,\beta^k\|_0 \in \mathscr{Z}(A,b)$. Since $\mathscr{Z}(A,b)$ is a compact set, we may assume without loss of generality that the binary sequence $\{z^k\}$ converges to binary vector z^∞ that must belong to $\mathscr{Z}(A,b)$. We then have,

$$|\,\beta_j^\infty\,|_0\;\leq\;\liminf_{k\to\infty}|\,\beta_j^k\,|_0\;=\;z_j^\infty,\forall\,j\;=\;1,\cdots,n.$$

By the definition of $\mathscr{Z}^{\leq}(A,b)$, it follows that $\|\beta^\infty\|_0 \in \mathscr{Z}^{\leq}(A,b)$. Thus

$$\mathrm{cl}\,[\,\mathrm{SOL\text{-}ASC}(A,b)\,]\subseteq\left\{\beta\;\mid\;\|\beta\|_0\in\mathscr{Z}^{\leq}(A,b)\right\}.$$

To show the reverse inclusion, let $\bar\beta$ be such that $\|\bar\beta\|_0 \leq \bar z$ for some $\bar z \in \mathscr{Z}(A,b)$. The sequence $\left\{\beta^k\right\}$, where

$$\beta_j^k\;\triangleq\;\left\{\begin{array}{ll}\bar\beta_j & \text{if } |\bar\beta_j|_0=\bar z_j\\ 1/k & \text{otherwise}\end{array}\right\}\;j\;=\;1,\cdots,n,$$

converges to $\bar\beta$. Moreover, since $\|\beta^k\|_0=\bar z$, it follows that $\beta^k\in$ SOL-ASC(A,b) for all k. So we have $\bar\beta\in\mathrm{cl}\,[\,\mathrm{SOL\text{-}ASC}(A,b)\,]$ as desired. □

Propositions 2 and 4 have established the fundamental role the two sets $\mathscr{Z}(A,b)$ and $\mathscr{Z}^{\leq}(A,b)$ play in the study of the ASC(A,b). The former set $\mathscr{Z}(A,b)$ is the intersection of the polyhedron $\{z\mid Az\leq b\}$ with the set $\{0,1\}^n$ of binary vectors, while a continuous relaxation of the latter set $\mathscr{Z}^{\leq}(A,b)$ is

$$\mathscr{H}^{\leq}(A,b)\;\triangleq\;\left\{\,z\;\in\;[\,0,1\,]^n\;\mid\;\exists\,\bar z\;\in\;[\,0,1\,]^n\text{ such that }A\bar z\;\leq\;b\text{ and }z\;\leq\;\bar z\,\right\}.$$

Clearly $\mathrm{cl}\,[\,\mathrm{SOL\text{-}ASC}(A,b)\,]\;\subseteq\;\left\{\,\beta\;\mid\;\|\,\beta\,\|_0\;\in\;\mathscr{H}^{\leq}(A,b)\right\}$. A natural question is whether equality holds. An affirmative answer to this question will simplify the task of verifying if a given vector β belongs to the left-hand closure, and facilitate the optimization over this closure. Indeed, according to Proposition 4, the former task can be accomplished by solving an integer program. If the equality in question holds, then this task amounts to solving a linear program. Furthermore, a representation in terms of the polytope $\mathscr{H}^{\leq}(A,b)$ is key to the convergence of the approximation of the ℓ_0-function by continuous surrogate functions.

Before addressing the above question, we give an example to show that the desired equality of the two sets in question does not always hold without conditions.

Example 3 Consider the ASC system in two variables (β_1,β_2): $|\beta_1|_0\;\leq\;0.8$ and $|\beta_2|_0\leq 2|\beta_1|_0$. Clearly, the only solution is $(0,0)$. The set $\mathscr{H}^{\leq}(A,b)$ for this system

is:

$$\Big\{ (z_1, z_2) \in [0,1]^2 \mid \exists\, (\bar{z}_1, \bar{z}_2) \in [0,1]^2 \text{ such that } \bar{z}_1 \le 0.8,\ \bar{z}_2 \le 2\bar{z}_1, \text{ and } (z_1, z_2) \le (\bar{z}_1, \bar{z}_2) \}$$

which clearly contains the point $(0, 1)$. Hence $\{\beta \mid \|\beta\|_0 \in \mathscr{H}^{\le}(A, b)\}$ is a superset of cl [SOL-ASC(A, b)]. □

Roughly speaking, the condition below has to do with the rounding of the elements on certain faces of $\mathscr{H}^{\le}(A, b)$ to integers without violating the linear system $Az \le b$.

Assumption A For any subset $\mathscr{I}$ of $\{1, \cdots, n\}$,

$$\big\{ z \in [0,1]^n \mid Az \le b \text{ and } z_{\mathscr{I}} = \mathbf{1}_{|\mathscr{I}|} \big\} \neq \emptyset \Rightarrow \mathscr{Z}(A, b) \cap \big\{ z \mid z_{\mathscr{I}} = \mathbf{1}_{|\mathscr{I}|} \big\} \neq \emptyset.$$

It is not difficult to see that Example 3 fails this assumption with $\mathscr{I} = \{2\}$ because the set

$$\Big\{ (z_1, z_2) \in [0,1]^2 \mid z_1 \le 0.8,\ z_2 \le 2z_1, \text{ and } z_2 = 1 \Big\}$$

is nonempty, but the constraints have no solutions with $z_1 \in \{0, 1\}$.

Proposition 5 *The following two statements hold:*

(a) *Under Assumption A, cl* [*SOL-ASC*(A, b)] $= \big\{\beta \mid \|\beta\|_0 \in \mathscr{H}^{\le}(A, b)\big\}$.
(b) *If Assumption A is violated by the index set* $\mathscr{I}$*, then any vector* β *with* $supp(\beta) = \mathscr{I}$ *belongs to the right-hand but not the left-hand set in (a).*

Proof It suffices to show that if β is such that $\|\beta\|_0 \le \bar{z}$ for some $\bar{z} \in [0, 1]^n$ satisfying $A\bar{z} \le b$, then β belongs to the closure of SOL-ASC(A, b). This follows readily by applying Assumption A to the index set $\mathscr{I} = \text{supp}(\bar{z})$. Thus (a) holds.

To prove (b), suppose $\mathscr{I}$ violates Assumption A. Let $\bar{z} \in [0, 1]^n$ be such that $A\bar{z} \le b$ and $\bar{z}_{\mathscr{I}} = \mathbf{1}_{|\mathscr{I}|}$ but there does not exist $z \in \mathscr{Z}(A, b)$ with $z_{\mathscr{I}} = \mathbf{1}_{|\mathscr{I}|}$. Clearly $\bar{z} \ge \|\beta\|_0$. Hence $\|\beta\|_0 \in \mathscr{H}^{\le}(A, b)$. But $\|\beta\|_0 \notin \mathscr{Z}^{\le}(A, b)$; for otherwise, there exists $\widehat{z} \in \mathscr{Z}(A, b)$ satisfying $\widehat{z} \ge \|\beta\|_0$, which implies $\widehat{z}_{\mathscr{I}} = \mathbf{1}_{|\mathscr{I}|}$, which is a contradiction. By Proposition 4, it follows that $\beta \notin$ cl [SOL-ASC(A, b)]. □

Assumption A holds for a matrix A satisfying the *column-wise uni-sign property*; i.e., when each column of A has either all nonpositive or all nonnegative entries (in particular, when A has only one row). In this case, fractional components (other than those in a given set $\mathscr{I}$) of a vector $z \in [0, 1]^n$ satisfying $Az \le b$ can be rounded either up or down, depending on the sign of the elements in the corresponding column, without violating the inequalities $Az \le b$. This special sign property of A turns out to be important in the optimization subject to ASC's; see Sect. 6.2. Another case where Assumption A holds is if

$$\text{convex hull of } \mathscr{Z}(A, b) = \big\{ z \in [0,1]^n \mid Az \le b \big\}. \tag{8}$$

Indeed, if the above equality holds, and if $z \in [0,1]^n$ satisfying $Az \le b$ is such that $z_{\mathscr{J}} = \mathbf{1}_{\mathscr{J}}$, then there exists $\{z^k\}_{k=1}^K \subseteq \mathscr{Z}(A,b)$ and positive scalars $\{\lambda_k\}_{k=1}^K$ summing to unity such that $z = \sum_{k=1}^K \lambda_k z^k$. Since $z_{\mathscr{J}} = \mathbf{1}_{\mathscr{J}}$ and each z^k is a binary vector, we must have $z^k_{\mathscr{J}} = \mathbf{1}_{\mathscr{J}}$ for all $k = 1, \cdots, K$. Thus Assumption A is valid under (8). Part (b) of Proposition 5 shows that Assumption A is sharp for part (a) to be valid.

Remark 2 Stand-alone lower cardinality constraints are not particularly interesting as the closure of such a constraint is the entire space. This may be one reason why lower cardinality constraints have not been found useful in an optimization context. However, when there are multiple ASCs, some of which may be of the lower cardinality kind, the closure operation must be performed for the whole system (instead of the individual constraints). The results in this section provide proper tools for this aim. Consider the ASC system in hierarchical variable section for example,

$$\text{SOL-ASC}(A,b) \;=\; \{\,\beta \mid \|\beta\|_0 \;\in\; \mathscr{Z}(A,b)\,\},$$

where $\mathscr{Z}(A,b)$ is defined as in Example 2. Each constraint in the form of $|\beta_{ij}|_0 \le \min\{|\beta_i|_0, |\beta_j|_0\}$ defines a set whose closure is the whole space. If one were to perform the closure operation to each constraint individually, and then take the intersection, one would end up with $\left\{ (\beta^{(1)}, \beta^{(2)}) \mid \sum_{j=1}^n |\beta_j^{(1)}|_0 \le K \right\}$. This set loses all hierarchical information and is a proper superset of the actual closure: $\text{cl}\,[\text{SOL-ASC}(A,b)] = \{(\|\beta^{(1)}\|_0, \|\beta^{(2)}\|_0) \in \mathscr{Z}^{\le}(A,b)\}$ with $\mathscr{Z}^{\le}(A,b)$ characterized in (7). □

3.3 Extension to SOL-xASC

In what follows, we extend two main results in the last subsection to an xASC system. In both extensions, we assume that $b(\gamma)$ is a continuous function and Γ is a closed set. We first extend Proposition 3.

Proposition 6 *The set SOL-xASC*(A,b,Γ) *is closed if and only if*

$$\textit{SOL-xASC}(A,b,\Gamma) \;=\; \left\{\,(\beta,\gamma) \;\in\; \mathbb{R}^n \times \Gamma \;\mid\; A^+\|\beta\|_0 \le b(\gamma)\,\right\}.$$

Proof It suffices to prove that if SOL-xASC(A,b,Γ) is closed, then it is contained in $\left\{(\beta,\gamma) \mid A^+\|\beta\|_0 \le b(\gamma)\right\}$. Let $(\bar\beta,\bar\gamma) \in$ SOL-xASC(A,b,Γ). Then $\bar\beta \in$ SOL-ASC$(A, b(\bar\gamma))$, which must be closed. Hence $A^+\|\bar\beta\|_0 \le b(\bar\gamma)$ by Proposition 3. □

Next is an extension of part (a) of Proposition 5. Part (b) is similar and omitted.

Proposition 7 *Suppose that for every* $\gamma \in \Gamma$, *Assumption A holds for the pair* $(A, b(\gamma))$. *Then*

$$cl\,[\,\textit{SOL-xASC}(A,b,\Gamma)\,] \;=\; \left\{\,(\beta,\gamma) \;\in\; \mathbb{R}^n \;\times\; \Gamma \;\mid\; \|\beta\|_0 \;\in\; \mathscr{H}^{\le}(A,b(\gamma))\,\right\}. \quad (9)$$

Proof Take any $(\bar{\beta}, \bar{\gamma})$ in the left-hand closure in (9). We claim that for every $\varepsilon > 0$, $\bar{\beta}$ belongs to the closure of the set $\{\beta \mid A\|\beta\|_0 \leq b(\bar{\gamma}) + \varepsilon \mathbf{1}\}$, which must be a subset of $\left\{ \beta \mid \| \beta \|_0 \in \mathscr{H}^{\leq}(A, \bar{b}^{\varepsilon}) \right\}$, where $\bar{b}^{\varepsilon} \triangleq b(\bar{\gamma}) + \varepsilon \mathbf{1}$. Let $\{(\beta^{\nu}, \gamma^{\nu})\}$ be a sequence in SOL-xASC(A, b, Γ) that converges to $(\bar{\beta}, \bar{\gamma})$ as $\nu \to \infty$. By definition, we have $A\|\beta^{\nu}\|_0 \leq b(\gamma^{\nu})$. Since $\{b(\gamma^{\nu})\}$ converges to $b(\bar{\gamma})$, for any $\varepsilon > 0$, there exists $\bar{\nu}$ such that for all $\nu \geq \bar{\nu}$, $A\|\beta^{\nu}\|_0 \leq b(\bar{\gamma}) + \varepsilon \mathbf{1}$. Since $\bar{\beta}$ is the limit of $\{\beta^{\nu}\}$, the claim holds. Hence, for every $\varepsilon > 0$ there exists $\bar{z}^{\varepsilon} \in [0, 1]^n$ such that

$$A\bar{z}^{\varepsilon} \leq b(\bar{\gamma}) + \varepsilon \mathbf{1} \text{ and } \|\beta\|_0 \leq \bar{z}^{\varepsilon}.$$

It then follows by a continuity property of polyhedra that there exists $\widehat{z} \in [0, 1]^n$ such that $A\widehat{z} \leq b(\bar{\gamma})$ and $\|\bar{\beta}\|_0 \leq \widehat{z}$. In other words, $(\bar{\beta}, \bar{\gamma})$ belong to the right-hand set in (9).

Conversely, suppose that $(\bar{\beta}, \bar{\gamma})$ belong to the right-hand set in (9). By Proposition 5, it follows that $\bar{\beta} \in \mathrm{cl}\left[\text{ SOL-ASC}(A, b(\bar{\gamma}))\right]$; so there exists $\{\beta^{\nu}\}$ converging to $\bar{\beta}$ such that $(\beta^{\nu}, \bar{\gamma})$ belongs to SOL-xASC(A, b, Γ) for every ν. Hence $(\bar{\beta}, \bar{\gamma})$ belong to the left-hand closure in (9). □

4 Continuous approximation and convergence

It is known that the ℓ_0-function can be formulated by a complementarity condition [16]. As such, regularizations and relaxations of the latter condition [14,22,28,30] can be employed as approximations of the former function. Another approach to deal with the discrete feature of the ℓ_0-function is by branch-and-bound for mixed-integer (non)linear programming; see e.g., [4,7]. When finite bounds of the variables are available, the indicator functions can be modeled with binary variables using such bounds followed by a suitable branch-and-bound method for optimization. However, an effective implementation of this approach is highly nontrivial and depends on many details, e.g., the way of choosing the bounds if they are known to exist only implicitly, the tradeoff between relaxation strength and computational complexity, primal heuristics and branching rules. In general, tools of building relatively tight lower bounds are rather limited [3], and can be computationally very expensive. In this section, we investigate the approximation of an (x)ASC system by another approach that has gained tremendous momentum in the field of statistical learning and sparsity representation where the nonzero count is a principal target to be controlled in a regression/estimation model.

Inspired by the family of difference-of-convex functions that were designed specifically in the statistical learning literature as surrogates of the ℓ_0-function (see [1,23] and the many references therein), we investigate the approximation of the set SOL-xASC(A, b, Γ) by replacing the univariate ℓ_0-function by a surrogate function parameterized by a scalar that control the tightness of the approximation. Convergence of such approximated functions to the ℓ_0-function has been ascertained in [23] in the context of sparsity optimization where the ℓ_0-function is part of the objective to be optimized. In contrast, our focus here is different from the latter reference in that we analyze the convergence of the approximated sets to the solution set of the xASC.

This analysis is complicated by the fact that SOL-xASC is generally not closed; so the convergence pertains to the closure of this solution set whose characterization in Proposition 7 is key. Before introducing the approximation functions, we first summarize two key notions of set convergence; see [29] for a comprehensive study of such convergence and the connection to optimization.

For any two closed sets C and D in $\mathbb{R}^N$ for some integer $N > 0$, the *Pompeiu–Hausdorff* (PH) distance is defined as:

$$\text{dist}_{\text{PH}}(C, D) \triangleq \max\left\{ \sup_{x\in C} \text{dist}(x, D),\ \sup_{x\in D} \text{dist}(x, C) \right\},$$

where the point-set distance $\text{dist}(x, D)$ is by definition equal to $\inf_{y\in D} \text{dist}(x, y)$ with $\text{dist}(x, y) \triangleq \|x - y\|$ and $\| \bullet \|$ being a given vector norm in $\mathbb{R}^N$. Let $C(\delta)$ be a closed set in $\mathbb{R}^N$ parameterized by $\delta \in \Delta$ where Δ is a closed set in some Euclidean space. The family $\{C(\delta)\}_{\delta\in\Delta}$ is said to converge to $C(\delta_0)$ in the PH sense if $\text{dist}_{\text{PH}}(C(\delta), C(\delta_0)) \to 0$ as $\delta \to \delta_0 \in \Delta$. Equivalently, $C(\delta)$ converges to $C(\delta_0)$ in the PH sense if for any $\varepsilon > 0$, there is an open neighborhood $\mathcal{N}$ of δ_0, such that for all $\delta \in \mathcal{N}$,

$$C(\delta) \subseteq C(\delta_0) + \text{cl}\,[\mathbb{B}_N(0, \varepsilon)] \text{ and} C(\delta_0) \subseteq C(\delta) + \text{cl}\,[\mathbb{B}_N(0, \varepsilon)]\,.$$

The other notion of set convergence is that of *Painlevé–Kuratowski* (PK) defined as follows. Again consider the case of $C(\delta)$ as $\delta \to \delta_0$; by definition, the outer and inner limits are, respectively,

$$\limsup_{\delta\to\delta_0} C(\delta) \triangleq \left\{ x \mid \liminf_{\delta\to\delta_0} \text{dist}(x, C(\delta)) = 0 \right\}$$
$$\liminf_{\delta\to\delta_0} C(\delta) \triangleq \left\{ x \mid \limsup_{\delta\to\delta_0} \text{dist}(x, C(\delta)) = 0 \right\}.$$

In other words, the outer limit $\limsup_{\delta\to\delta_0} C(\delta)$ contains all x such that *there exist* sequences $\{\delta^k\} \to \delta_0$ and $\{x^k\} \to x$ such that $x^k \in C(\delta^k)$ for all k . The inner limit $\liminf_{\delta\to\delta_0} C(\delta)$ contains all x such that *for any* sequence $\{\delta^k\} \to \delta_0$, there exists a sequence $\{x^k\} \to x$ such that $x^k \in C(\delta^k)$ for all k. It is easy to show that the inner limit is always a subset of the outer limit. The family $\{C(\delta)\}_{\delta\in\Delta}$ is said to converge to $C(\delta_0)$ as $\delta \to \delta_0$ in the PK sense if both of the outer and inner limits are equal to $C(\delta_0)$. It is proved in [29, Proposition 5.12] that PH convergence implies PK convergence; but the converse is not true.

In later applications, we will be speaking about the convergence of a sequence of sets $\{C(\delta^k)\}$ to a given set $C(\delta^\infty)$ as the sequence $\{\delta^k\}$ converges to a limit δ^∞. The definition of such sequential convergence is similar to the above that applies to *all* δ near the base value δ^∞.

In general, set convergence has an important role to play in the convergence of optimal solutions to an optimization problem when the constraints are being approximated, like in the case of SOL-ASC(A, b) to be discussed in the next section. This

is made precise in the result below; see [29, Section 7.E] where such convergence is discussed in the framework of functional epi-convergence.

Proposition 8 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous. Let $\{\delta^k\}$ be a sequence converging to δ^∞. If*

$$\lim_{k\to\infty} C(\delta^k) \;=\; C(\delta^\infty) \quad \textit{in the PK sense},$$

then $\limsup_{k\to\infty} \operatorname{argmin}_{x\in C(\delta^k)} f(x) \;\subseteq\; \operatorname{argmin}_{x\in C(\delta^\infty)} f(x)$.

Proof Let $\{\bar{x}^k\}$ be a sequence converging to x^∞ such that $\bar{x}^k \in \operatorname{argmin}_{x\in C(\delta^k)} f(x)$ for all k. It suffices to show two things: (i) $x^\infty \in C(\delta^\infty)$ and (ii) for every $x \in C(\delta^\infty)$, there exists $\{x^k\}$ converging to x such that $x^k \in C(\delta^k)$ for all k. Both follow easily from the assumed convergence of $\{C(\delta^k)\}$ to $C(\delta^\infty)$. □

While simple to prove and illustrative of the role of set convergence, Proposition 8 is only of conceptual importance because if the sets $C(\delta^k)$ are nonconvex and/or the objective $f(x)$ is nonconvex, then since global minima of nonconvex optimization problems are generally not computable in practice, the convergence of a sequence of minima cannot be used to deduce any property of the limit of a sequence of non-optimal solutions of the approximated problems. Instead, the convergence of computable stationary solutions should be investigated in order to eliminate the gap between practical computation and convergence of the computed solutions. This provides the motivation for Sect. 6 where we focus on a kind of stationary points that can be computed by methods of dc constrained optimization problems, such as those presented in the reference [26], and investigate the convergence of such stationary solutions. Before discussing this in detail, some preparatory work is needed that begins in the next subsection.

4.1 Approximating functions

We consider the approximation of the univariate ℓ_0-function $|\bullet|_0$ by various surrogate functions that are motivated by the families of surrogate sparsity functions summarized in [1,23]. Specifically, writing the constraint system $A\|\beta\|_0 \le b(\gamma)$ as

$$\sum_{j=1}^{n} A_{ij}^{+} \,|\, \beta_j \,|_0 \;\le\; \sum_{j=1}^{n} A_{ij}^{-} \,|\, \beta_j \,|_0 + b_i(\gamma), i \;=\; 1, \cdots, m,$$

we approximate each ASC constraint by

$$\sum_{j=1}^{n} A_{ij}^{+}\, p_j^{+}(\beta_j, \delta_j^{+}) \;\le\; \sum_{j=1}^{n} A_{ij}^{-}\, p_j^{-}(\beta_j, \delta_j^{-}) + b_i(\gamma), i \;=\; 1, \cdots, m, \tag{10}$$

where $\delta_j^{\pm}$ are positive scalars and each $p_j^{\pm}$ is a continuous bivariate function $\rho : \mathbb{R} \times (0, \infty) \to [0, 1]$ satisfying

(R1) $\lim_{\delta \downarrow 0} \rho(t, \delta) = |\, t \,|_0$ for any $t \in \mathbb{R}$ [this limit is not required to be uniform in t];
(R2) for every $\delta > 0$, $\rho(\bullet, \delta)$ is symmetric on $\mathbb{R}$ (i.e., $\rho(t, \delta) = \rho(-t, \delta)$ for all $t \geq 0$), and non-decreasing on $[0, \infty)$;
(R3) $\rho(t, \delta) = 1$ for all t such that $|t| \geq \delta$.

One special feature about the approximated system (10) is that we may use different approximation functions $p_j^{\pm}(\bullet, \delta_j^{\pm})$ corresponding to the individual entries of the matrix A. We let SOL-xASC$_{\delta^{\pm}}(A, b, \Gamma)$ denote the solution set of (10), emphasizing the control scalars $\delta_j^{\pm}$ of the approximating functions. We use Example 3 to show that SOL-xASC$_{\delta^{\pm}}(A, b, \Gamma)$ does not always converge to SOL-xASC(A, b, Γ) as $\delta^{\pm} \downarrow 0$. When this happens, a limiting solution of an optimization problem subject to the approximated constraints may not even be feasible to the xASC system.

Example 3 (cont.) We approximate the 2-variable system: $|\beta_1|_0 \leq 0.8$ and $|\beta_2|_0 \leq 2|\beta_1|_0$ using the capped ℓ_1-function $\rho(t, \delta) = \min\left(\frac{|t|}{\delta}, 1\right)$ for $\delta > 0$, obtaining the approximated system consisting of the inequalities below:

$$\min\left(\frac{|\beta_1|}{\delta}, 1\right) \leq 0.8 \text{and} \min\left(\frac{|\beta_2|}{\delta}, 1\right) \leq 2 \min\left(\frac{|\beta_1|}{\delta}, 1\right). \tag{11}$$

The solution set of the ASC, which is closed in this example, and that of the approximated system are depicted in Fig. 1 below. Algebraically, the solution set of (11), for fixed $\delta > 0$, is the union of two non-convex sets:

$$\{(\beta_1, \beta_2) \mid |\beta_1| \leq 0.8\,\delta,\ 2\,|\beta_2| \leq |\beta_1|\} \bigcup \underbrace{\{(\beta_1, \beta_2) \mid 0.5\,\delta \leq |\beta_1| \leq 0.8\,\delta\}}_{\text{the vertical stripes in the right-hand figure}}.$$

Since the two vertical stripes are always contained in the above union for any $\delta > 0$ and shrink to the vertical β_2-axis as $\delta \downarrow 0$, it is not difficult to see that this vertical axis is the limit of SOL-ASC$_\delta$ as $\delta \downarrow 0$. Clearly, this axis is a proper superset of the solution set of the ASC system which is the singleton $\{(0, 0)\}$. □

We present below various surrogate functions summarized in [1] and briefly discuss their satisfaction of the assumptions (R1)–(R3).

The SCAD **family.** Parameterized by two scalars $a > 2$ and $\lambda > 0$ and with the origin as its unique zero, this univariate *smoothly clipped absolute deviation* (SCAD) function is once continuously differentiable except at the origin and given by: for all $t \in \mathbb{R}$,

$$p_{a,\lambda}^{\text{SCAD}}(t) \triangleq \begin{cases} \lambda\,|\,t\,| & \text{if } |\,t\,| \leq \lambda \\ \dfrac{(a+1)\,\lambda^2}{2} - \dfrac{(a\,\lambda - |\,t\,|)^2}{2\,(a-1)} & \text{if } \lambda \leq |\,t\,| \leq a\,\lambda \\ \dfrac{(a+1)\,\lambda^2}{2} & \text{if } |\,t\,| \geq a\,\lambda\,. \end{cases}$$

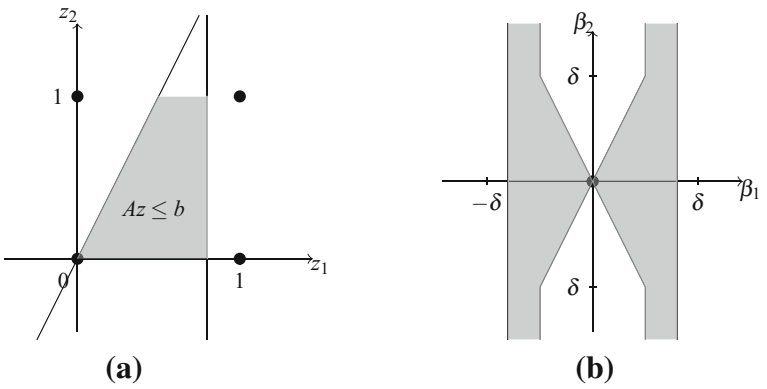


Fig. 1 **a** $\mathscr{Z}^{\leq} = \{z \in \{0,1\}^2 \mid Az \leq b\} = \{(0,0)\}$; **b** solution set of the approximated system (11)

To conform to the stated assumptions, we scale the function by the reciprocal of $\dfrac{(a+1)\,\lambda^2}{2}$ and identify $\delta = a\lambda$ and a being fixed, obtaining

$$\rho_a^{\mathrm{SCAD}}(t,\delta) \;=\; \begin{cases} \dfrac{2\,a}{(a+1)\,\delta}\,|\,t\,| & \text{if } |\,t\,| \leq \dfrac{\delta}{a} \\ 1 - \dfrac{(\,\delta - |\,t\,|\,)^2}{\left(1 - \dfrac{1}{a^2}\right)\delta^2} & \text{if } \dfrac{\delta}{a} \;\leq\; |\,t\,| \;\leq\; \delta \\ 1 & \text{if } |\,t\,| \;\geq\; \delta \end{cases}$$

Note that while $\lim\limits_{\delta \downarrow 0} \rho_a^{\mathrm{SCAD}}(t,\delta) = |t|_0$ for any $t \in \mathbb{R}$, this limit is not uniform in t near 0. The same remark applies to the following families of functions.

The MCP **family.** Also parameterized by two positive scalars $a > 2$ and λ, this is a univariate, piecewise quadratic function given by:

$$p_{a,\lambda}^{\mathrm{MCP}}(t) \;\triangleq\; \tfrac{1}{2}\left\{ a\,\lambda^2 - \frac{\left[\,(a\lambda - |\,t\,|\,)_+\,\right]^2}{a} \right\} \text{ for } t \in \mathbb{R}.$$

Again, to conform with the approximating conditions, we scale the function by $\dfrac{2}{a\,\lambda^2}$ and identify $\delta = a\lambda$ and a being fixed, obtaining

$$\rho^{\mathrm{MCP}}(t,\delta) \;=\; \begin{cases} \dfrac{2\,|\,t\,|}{\delta} - \dfrac{t^2}{\delta^2} & \text{if } |\,t\,| \;\leq\; \delta \\ 1 & \text{if } |\,t\,| \;\geq\; \delta \end{cases}$$

The capped ℓ_1 family. This has already been mentioned before; namely, for $\delta > 0$,

$$\rho^{\mathrm{C}\ell_1}(t, \delta) \;=\; \min\left(\frac{|\, t \,|}{\delta},\; 1 \right), \text{for } t \in \mathbb{R}.$$

The truncated transformed ℓ_1 family. Parameterized by a given scalar $a > 0$, this is a truncated modification of the transformed ℓ_1-function in [1] taking into account the scaling factor $\delta > 0$ that tends to zero:

$$\rho_a^{\mathrm{TTL}_1}(t, \delta) \;=\; \min\left(\frac{(a+\delta)\,|\, t \,|}{\delta\,(a+|\, t \,|)},\; 1 \right), \text{for } t \in \mathbb{R}.$$

It is easy to see that properties (R1) and (R2) hold for this function $\rho_a^{\mathrm{TTL}_1}(t, \delta)$ for $(t, \delta) \in \mathbb{R} \times (0, \infty)$. Noting that

$$\frac{(a+\delta)\,|\, t \,|}{\delta\,(a+|\, t \,|)} \;=\; 1 + \frac{a\,(|\, t \,| - \delta)}{\delta\,(a+|\, t \,|)},$$

we may conclude that condition (R3) is also satisfied.

The truncated logarithmic family. Derived similarly to the truncated transformed ℓ_1-functions and parameterized by a scalar $\varepsilon > 0$, a truncated logarithmic penalty function is defined as follows: for $\delta > 0$,

$$\rho_\varepsilon^{\mathrm{Tlog}}(t, \delta) \;\triangleq\; \min\left\{ \frac{1}{\log\left(1 + \dfrac{\delta}{\varepsilon}\right)} \left[\log(|\, t \,| + \varepsilon) - \log(\varepsilon) \right],\; 1 \right\}, \text{for } t \;\in\; \mathbb{R}.$$

It is not difficult to see that this function satisfies the desired conditions (R1), (R2), and (R3).

In summary, we have identified a number of well-known univariate surrogate sparsity functions, suitably truncated outside a δ-neighborhood of the origin, that satisfy the basic requirements of an approximating function of the ℓ_0-function.

4.2 Convergence

The main convergence result of the family of approximated solution sets $\{\text{SOL-xASC}_{\delta^\pm}(A, b, \Gamma)\}_{\delta^\pm > 0}$ as $\delta^\pm \to 0$ is the following.

Theorem 1 *Suppose that for every $\gamma \in \Gamma$, Assumption A holds for the pair $(A, b(\gamma))$. Assume further that for all $\delta_j^\pm > 0$ sufficiently small, the surrogate functions in the*

family $\{p_j^{\pm}(\bullet, \delta_j^{\pm})\}_{j=1}^n$ *satisfy in addition to (R1), (R2), and (R3),*

$$\begin{aligned} &\textit{either} \sum_{j=1}^{n} A_{ij}^{-}\, p_j^{-}(\beta_j, \delta_j^{-}) \;\le\; \sum_{j=1}^{n} A_{ij}^{-}\, p_j^{+}(\beta_j, \delta_j^{+}), \\ &\forall\, \beta \in \mathbb{R}^n \textit{ and } i = 1, \cdots, m. \end{aligned} \tag{12}$$

$$\begin{aligned} &\textit{or} \sum_{j=1}^{n} A_{ij}^{+}\, p_j^{-}(\beta_j, \delta_j^{-}) \;\le\; \sum_{j=1}^{n} A_{ij}^{+}\, p_j^{+}(\beta_j, \delta_j^{+}), \\ &\forall\, \beta \in \mathbb{R}^n \textit{ and } i = 1, \cdots, m. \end{aligned} \tag{13}$$

Then the family $\{SOL\text{-}xASC_{\delta^{\pm}}(A, b, \Gamma)\}_{\delta^{\pm} > 0}$ *converges to cl* [*SOL-xASC*(A, b, Γ)] *in the PH sense as* $\|\delta^{\pm}\|_{\infty} \triangleq \max\left\{\delta_j^{\pm} \mid j = 1, \cdots n\right\} \downarrow 0$.

Proof Let $C(\delta^{\pm}) \triangleq$ SOL-xASC$_{\delta^{\pm}}(A, b, \Gamma)$ and $C_0 \triangleq$ cl [SOL-xASC(A, b, Γ)]. We need to show that for every $\varepsilon > 0$, a scalar $\underline{\delta} > 0$ exists such that for all $\delta \in (0, \underline{\delta}]$,

$$C(\delta^{\pm}) \subseteq C_0 + \text{cl}\left[\mathbb{B}_{n+k}(0, \varepsilon)\right] \text{and} C_0 \subseteq C(\delta^{\pm}) + \text{cl}\left[\mathbb{B}_{n+k}(0, \varepsilon)\right]. \tag{14}$$

By Proposition 7, we have

$$C_0 = \left\{(\beta, \gamma) \in \mathbb{R}^n \times \Gamma \mid \exists\, \bar{z} \in [0, 1]^n \text{ such that } A\bar{z} \le b(\gamma) \text{ and } \|\beta\|_0 \le \bar{z}\right\}.$$

Let $(\bar{\beta}, \bar{\gamma}) \in C(\delta^{\pm})$. Then $\bar{\gamma} \in \Gamma$. For the rest of the proof, we assume that (12) holds. A similar proof can be applied when (13) holds. For all $i = 1, \cdots, m$

$$\sum_{j=1}^{n} A_{ij}^{+}\, p_j^{+}(\bar{\beta}_j, \delta_j^{+}) \le \sum_{j=1}^{n} A_{ij}^{-}\, p_j^{-}(\bar{\beta}_j, \delta_j^{-}) + b_i(\bar{\gamma}) \le \sum_{j=1}^{n} A_{ij}^{-}\, p_j^{+}(\bar{\beta}_j, \delta_j^{+}) + b_i(\bar{\gamma}),$$

By Assumption A, there exists $\bar{z} \in \{0, 1\}^n$ such that $A\bar{z} \le b(\bar{\gamma})$ and $\bar{z}_j = 1$ whenever $p_j^{+}(\bar{\beta}_j, \delta_j^{+}) = 1$. Define a vector $\beta(\delta)$ with components given by

$$\beta_j(\delta) \triangleq \begin{cases} \bar{\beta}_j & \text{if } |\bar{\beta}_j| \ge \delta_j^{+}; \text{ (thus } p_j^{+}(\bar{\beta}_j, \delta^{+}) = 1) \\ \text{sign}(\bar{\beta}_j)\, \delta_j^{+}\, \bar{z}_j & \text{if } |\bar{\beta}_j| < \delta_j^{+}, \end{cases}$$

where we define sign$(0) = 1$. It is easily seen that $\|\beta(\delta)\|_0 = \bar{z}$. Hence $A\|\beta(\delta)\|_0 \le b(\bar{\gamma})$; so $(\beta(\delta), \bar{\gamma}) \in C_0$. For an index j such that $|\bar{\beta}_j| < \delta_j^{+}$, we have

$$\beta_j(\delta) - \bar{\beta}_j = \text{sign}(\bar{\beta}_j)\left[\delta_j^{+}\, \bar{z}_j - |\,\bar{\beta}_j\,|\right],$$

implying that $|\,\beta_j(\delta) - \bar{\beta}_j\,| = \delta_j^{+}\, \bar{z}_j - |\,\bar{\beta}_j\,| \le \delta_j^{+}$. Hence the first inclusion in (14) holds. To prove the second inclusion, let $(\bar{\beta}, \bar{\gamma}) \in C_0$ be arbitrary. Define a vector

$\beta(\delta)$ with components given by:

$$\beta_j(\delta) \triangleq \begin{cases} \bar{\beta}_j & \text{if } |\bar{\beta}_j| \geq \delta_j \text{ or } \bar{\beta}_j = 0 \\ \text{sign}(\bar{\beta}_j)\, \delta_j & \text{otherwise,} \end{cases}$$

where $\delta_j \triangleq \max\left(\delta_j^+, \delta_j^-\right)$. It then follows that $p_j^{\pm}(\beta_j(\delta), \delta_j^{\pm}) = 1 = |\bar{\beta}_j|_0$ unless $\bar{\beta}_j = 0$, in which case, $p_j^{\pm}(\bar{\beta}_j, \delta_j^{\pm}) = 0 = |\bar{\beta}_j|_0$. Since $A\|\bar{\beta}\|_0 \leq b(\gamma)$, it follows that $\beta(\delta) \in C(\delta^{\pm})$. Since clearly $|\beta_j(\delta) - \bar{\beta}_j| \leq \delta_j$, the second inclusion in (14) also holds. □

Two special cases worth noting are when $A \geq 0$ or $A \leq 0$. It is easy to verify that Assumption A holds in both cases (see discussion after Proposition 5); moreover, (12) holds in the former case and (13) holds in the latter case. Hence Theorem 1 is valid under the basic properties (R1), (R2) and (R3). In general, the two requirements (12) and (13) can be enforced by choosing p_j^+ and p_j^- such that $p_j^-(\bullet, \delta_j^-) \leq p_j^+(\bullet, \delta_j^+)$ pointwise.

Our next convergence result pertains to the situation where we fix $\delta_j^- > 0$ for all $j = 1, \cdots, n$ and consider only a one-sided approximation $|\beta_j|_0 \approx p_j^+(\beta_j, \delta_j^+)$ with $\delta_j^+ \downarrow 0$. Specifically, we consider the approximation of the system

$$\sum_{j=1}^{n} A_{ij}^+ \,|\, \beta_j \,|_0 \;\leq\; \widehat{b}_i(\beta, \gamma) \;\triangleq\; \sum_{j=1}^{n} A_{ij}^- \, p_j^-(\beta_j, \delta_j^-) + b_i(\gamma), \;\forall\, i \;=\; 1, \cdots, m,$$

where we fix δ_j^- and approximate $|\bullet|_0$ on the left side as said. In this case, convergence in the PK sense can be established without Assumption A and with no restriction on the choice of the approximating functions $p_j^+(\bullet, \delta_j^+)$ except for the basic properties (R1), (R2), and (R3). Recognizing that the above system is a constraint system: $A\|\beta\|_0 \leq \widehat{b}(\beta, \gamma)$ with A being a nonnegative matrix and $\widehat{b}$ being a continuous function of the pair (β, γ) whose dependence on $\left\{\delta_j^-\right\}_{j=1}^n$ we have suppressed, we state and prove a version of Theorem 1 for such a system.

Proposition 9 *Let A be a nonnegative matrix, $\widehat{b}(\beta, \gamma)$ be a continuous function, and Γ be a closed set. For each $j = 1, \cdots, m$, let $p_j^+(\bullet, \delta_j^+)$ be an approximating function satisfying conditions (R1), (R2), and (R3). Define for any $\delta^+ \triangleq \left(\delta_j^+\right)_{j=1}^n > 0$, the sets*

$$C(\delta^+) \triangleq \left\{ (\beta, \gamma) \in \mathbb{R}^n \times \Gamma \;|\; \sum_{j=1}^{n} A_{\bullet j}\, p_j^+(\beta_j, \delta_j^{+,k}) \leq \widehat{b}(\beta, \gamma) \right\}$$

$$\text{and } C_0 \quad \triangleq \left\{ (\beta, \gamma) \in \mathbb{R}^n \times \Gamma \;|\; A\,\|\,\beta\,\|_0 \leq \widehat{b}(\beta, \gamma) \right\}.$$

Then the family $\left\{C(\delta^+)\right\}_{\delta^+ > 0}$ converges to C_0 in the PK sense as $\|\delta^+\|_\infty \triangleq \max\left\{\delta_j^+ \;|\; j = 1, \cdots n\right\} \downarrow 0$.

Proof Since C_0 is a subset of $C(\delta^+)$ for any $\delta^+ > 0$, it follows that the former set is a subset of $\liminf_{\delta^+ \downarrow 0} C(\delta^+)$. It remains to show that $\limsup_{\delta^+ \downarrow 0} C(\delta^+) \subseteq C_0$. Let $(\beta^\infty, \gamma^\infty)$ be the limit of a sequence $\{(\beta^k, \gamma^k)\}$ where for each k, the pair (β^k, γ^k) satisfies:

$$\sum_{j=1}^{n} A_{ij}\, p_j^+(\beta_j^k, \delta_j^{+,k}) \;\leq\; \widehat{b}_i(\beta^k, \gamma^k)\; \forall\, i \;=\; 1, \cdots, m,$$

corresponding to a sequence of positive scalars $\{\delta^{+,k}\} \downarrow 0$. By the nonnegativity of A, property (R3), and the nonnegativity of of the approximating functions $p_j^+(\bullet, \delta_j^+)$, we have for each $i = 1, \cdots, m$ and all k sufficiently large,

$$\sum_{j=1}^{n} A_{ij} \,|\, \beta_j^\infty \,|_0 \;=\; \sum_{j \in \mathrm{supp}(\beta^\infty)} A_{ij} \;\leq\; \sum_{j=1}^{n} A_{ij}\, p_j^+(\beta_j^k, \delta_j^{+,k}) \;\leq\; \widehat{b}_i(\beta^k, \gamma^k).$$

Passing to the limit $k \to \infty$ establishes the desired inclusion. □

5 Tangent properties of SOL-ASC and its approximation

For a given closed sets $C \subseteq \mathbb{R}^N$, the tangent cone of C at a vector $x \in C$, denoted $\mathscr{T}(x; C)$ consists of vectors v such that there exist sequences of vectors $\{x^k\} \subseteq C$ converging to x and positive scalars $\{\tau_k\}$ converging to zero such that $v = \lim_{k\to\infty} \dfrac{x^k - x}{\tau_k}$. Tangent vectors of closed sets play an important role in the stationarity conditions of optimization problems constrained by such sets. In this section, we characterize the tangent vectors of SOL-ASC(A, b) and those of its approximation SOL-ASC$_{\delta^\pm}(A, b)$. We omit the extension (A, b, Γ) in order to focus on the $\| \bullet \|_0$-function and its componentwise approximation by the penalty functions $p_j^\pm(\bullet, \delta_j^\pm)$. Recalling that the tangent cone must be a closed set, we have the following expression of the tangent cone of the SOL-ASC, which shows in particular that the latter cone is also defined by an ASC. We also obtain a superset of the tangent cone in the case where the matrix A is nonnegative which will be useful subsequently. For a given $\bar{\beta} \in$ SOL-ASC(A, b), we let $\mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$ be the set of indices $i \in \{1, \cdots, m\}$ such that $A_{i\bullet} \| \bar{\beta} \|_0 = b_i$.

Proposition 10 *Let $\bar{\beta} \in$ SOL-ASC(A, b). Let $\bar{S}^{\mathrm{c}}$ be the complement of $\bar{S} \triangleq$ supp$(\bar{\beta})$ in $\{1, \cdots, n\}$. It holds that*

$$\begin{aligned} \mathscr{T}\left(\bar{\beta}; \text{SOL-ASC}(A, b)\right) &= cl\Big[\Big\{ v \;\mid\; A_{\bullet \bar{S}^{\mathrm{c}}} \| \, v_{\bar{S}^{\mathrm{c}}} \|_0 \;\leq\; b - A \,\|\, \bar{\beta} \,\|_0 \Big\}\Big] \\ &= cl\left[\left\{ v \;\mid\; \begin{pmatrix} \bar{\beta}_{\bar{S}} \\ v_{\bar{S}^{\mathrm{c}}} \end{pmatrix} \in \text{SOL-ASC}(A, b) \right\}\right]. \end{aligned} \tag{15}$$

Thus if $A \geq 0$, then

$$\begin{aligned} \mathscr{T}\left(\bar{\beta}; \text{SOL-ASC}(A, b)\right) \;\subseteq\; \big\{ v \;\mid\; & v_j \\ = \; 0 \; & \text{for all } j \in \bar{S}^{\mathrm{c}} \text{ such that } \exists\, i \in \mathscr{A}_{\mathrm{ASC}}(\bar{\beta}) \text{ with } A_{ij} > 0 \big\}. \end{aligned}$$

Proof The equality of the two closures is easy to see. To prove the equality of the tangent cone and the first closure, write $C \triangleq \text{SOL-ASC}(A, b)$. We first show that

$$\left\{ v \mid A_{\bullet \bar{S}^c} \| v_{\bar{S}^c} \|_0 \leq b - A \| \bar{\beta} \|_0 \right\} \subseteq \mathscr{T}(\bar{\beta}; C).$$

This is enough to imply that the left-hand tangent cone in (15) contains the right-hand closures in the same expression. Let v be an arbitrary vector satisfying $A_{\bullet \bar{S}^c} \| v_{\bar{S}^c} \|_0 \leq b - A \| \bar{\beta} \|_0$. For all $\tau > 0$ sufficiently small, we have $\| \bar{\beta} + \tau v \|_0 = \max \left(\| \bar{\beta} \|_0, \| v \|_0 \right)$. So

$$A \| \bar{\beta} + \tau v \|_0 = A \| \bar{\beta} \|_0 + A_{\bullet \bar{S}^c} \| v_{\bar{S}^c} \|_0 \leq b,$$

which implies that v is a tangent vector of C at $\bar{\beta}$. Conversely, let $d \in \mathscr{T}(\bar{\beta}; C)$ be given. Let $\{\beta^k\} \subseteq C$ and $\{\tau_k\} \downarrow 0$ be such that

$$\lim_{k \to \infty} \beta^k = \bar{\beta} \text{ and } \lim_{k \to \infty} \frac{\beta^k - \bar{\beta}}{\tau_k} = d.$$

Clearly, $\left\| \dfrac{\beta^k_{\bar{S}^c} - \bar{\beta}_{\bar{S}^c}}{\tau_k} \right\|_0 = \| \beta^k_{\bar{S}^c} \|_0$; moreover, for all k sufficiently large,

$$b \geq A \| \beta^k \|_0 = A \| \bar{\beta} \|_0 + A_{\bullet \bar{S}^c} \| \beta^k_{\bar{S}^c} \|_0 = A \| \bar{\beta} \|_0 + A_{\bullet \bar{S}^c} \left\| \frac{\beta^k_{\bar{S}^c} - \bar{\beta}_{\bar{S}^c}}{\tau_k} \right\|_0.$$

Hence, it follows readily that d belongs to the right-hand closure in (15).

To prove the last assertion of the proposition, let $A \geq 0$. Let v satisfy $A_{\bullet \bar{S}^c} \| v_{\bar{S}^c} \|_0 \leq b - A \| \bar{\beta} \|_0$. For every $i \in \mathscr{A}_{\text{ASC}}(\bar{\beta})$, we have

$$\sum_{j \notin \bar{S} \text{ and } v_j \neq 0} A_{ij} \leq 0.$$

Therefore if $j \in \bar{S}^c$ and there is an $i \in \mathscr{A}_{\text{ASC}}(\bar{\beta})$ with $A_{ij} > 0$, then we must have $v_j = 0$. □

Proposition 10 yields two interesting properties of $\mathscr{T}\left(\bar{\beta}; \text{SOL-ASC}(A, b)\right)$ that can be contrasted with the tangent cone of the polyhedron $\mathscr{P}(A, b) \triangleq \{w \in \mathbb{R}^n \mid Aw \leq b\}$. First, it is known that for a given $\bar{w} \in \mathscr{P}(A, b)$, we have

$$\mathscr{T}(\bar{w}; \mathscr{P}(A, b)) = \left\{ v \in \mathbb{R}^n \mid A_{i \bullet} v \leq 0 \, \forall \, i \in \mathscr{A}(\bar{w}) \right\}$$

where $\mathscr{A}(\bar{w})$ is the index set of the active constraints at $\bar{w} \in \mathscr{P}(A, b)$. It follows from this representation of $\mathscr{T}(\bar{w}; \mathscr{P}(A, b))$ that v is a tangent vector of $\mathscr{P}(A, b)$ at $\bar{w}$ if and only if $\bar{w} + \tau v \in \mathscr{P}(A, b)$ for all $\tau > 0$ sufficiently small. In contrast, Proposition 10 shows that a vector v is a tangent of the set $\text{SOL-ASC}(A, b)$ at $\bar{\beta}$ if and only if it is the

limit of a sequence $\{v^k\}$ for which a scalar $\bar{\tau} > 0$ and an integer $\bar{k} > 0$ exist such that for all $\tau \in (0, \bar{\tau}]$ and all $k \geq \bar{k}$, $\bar{\beta} + \tau v^k \in$ SOL-ASC(A, b). From this, it follows that if v is a tangent of the set SOL-ASC(A, b), then $\bar{\beta} + \tau v \in$ cl[SOL-ASC(A,b)] for all $\tau > 0$ sufficiently small. Thus the tangents of SOL-ASC(A, b) have exactly the same feasibility property as those of $\mathscr{P}(A, b)$ provided that SOL-ASC(A, b) is closed. Another interesting consequence of Proposition 10 is that the entire set of constraints $A\|\beta\|_0 \leq b$ and the (in)activity of the sparsity constraints $|\beta_j| \geq 0$, $j = 1, \cdots, n$ are all involved in the definition of the tangent vectors of SOL-ASC(A, b). In contrast, $\mathscr{T}(\bar{w}; \mathscr{P}(A, b))$ involves only the active constraints at $\bar{w}$ of the system $Aw \leq b$.

Example 4 We illustrate Proposition 10 for a pair of upper and lower cardinality constraints: for two positive integers $\underline{k} < \bar{k}$,

$$\underline{k} \leq \sum_{j=1}^{n} |\beta_j|_0 \leq \bar{k}. \tag{16}$$

Let C be the solution set of this ASC. Let $\bar{\beta}$ be a given vector satisfying (16). There are 2 cases:

- $\sum_{j=1}^{n} |\bar{\beta}_j|_0 = \bar{k}$. In this case, it is not difficult to show that $\mathscr{T}(\bar{\beta}; C) = \left\{ v \mid v_j = 0 \text{ for all } j \in \bar{S}^{\mathrm{c}} \right\}$.
- $\sum_{j=1}^{n} |\bar{\beta}_j|_0 < \bar{k}$. In this case, it is not difficult to show that $\mathscr{T}(\bar{\beta}; C) = \left\{ v \mid \sum_{j \notin \bar{S}} |v_j|_0 \leq \bar{k} - \sum_{j=1}^{n} |\bar{\beta}_j|_0 \right\}$.

The noteworthy point of this illustrative example is that the lower cardinality constraint does not enter into the representation of the tangent cone, regardless of whether this constraint is binding or not. More generally, for the ASC(A, b), it is easy to see that if a row $A_{i\bullet}$ contains no positive entries, then the constraint $A_{i\bar{S}^{\mathrm{c}}} \| v_{\bar{S}^{\mathrm{c}}} \|_0 \leq b_i - A_{i\bullet} \| \bar{\beta} \|_0$ is trivially satisfied, thus redundant, in the representation of the tangent cone of the solution set of the ASC at $\bar{\beta}$. □

5.1 Tangent cone of approximated sets: fixed δ

Consider the approximated SOL-ASC$_{\delta^{\pm}}(A, b)$ with each approximating function $p_j^{\pm}(\bullet, \delta_j^{\pm})$ given by: for some positive integer K_j,

$$p_j^{\pm}(t, \delta_j^{\pm}) \triangleq \lambda_j^{\pm} |t| - \underbrace{\max_{1 \leq k \leq K_j} g_{jk}^{\pm}(t)}_{\triangleq\, g_j^{\pm}(t)}, t \in \mathbb{R}, \tag{17}$$

where each $\lambda_j^{\pm}$ is a positive scalar and each $g_{jk}^{\pm}$ is a univariate differentiable convex function, all dependent on $\left\{\delta_j^{\pm}\right\}_{j=1}^{n}$. As proved in [1], the surrogate sparsity functions in the SCAD, MCP, capped ℓ_1, the transformed ℓ_1, and the logarithmic families can all be expressed in the above difference-of-convex (dc) form. From the expression:

$$\min\left(\lambda\,|\,t\,|-g(t),\ 1\right)\ =\ \lambda\,|\,t\,|-\max\left(g(t),\ \lambda\,|\,t\,|-1\right),$$

we see that the truncation of a dc function of the above form can be represented in the same form; thus, dc functions given by (17) also include those in the truncated transformed ℓ_1 and truncated logarithmic families; thus all the functions discussed in Sect. 4.1 are covered by the form (17).

With $p_j^{\pm}(\bullet, \delta_j^{\pm})$ as given, the inequality

$$\sum_{j=1}^{n} A_{ij}^{+}\, p_j^{+}(\beta_j, \delta_j^{+})\ \le\ \sum_{j=1}^{n} A_{ij}^{-}\, p_j^{-}(\beta_j, \delta_j^{-}) + b_i$$

can be written very simply as

$$\underbrace{\phi_i(\beta)-\psi_i(\beta)}_{\zeta_i(\beta)} - b_i\ \le\ 0, \tag{18}$$

where

$$\phi_i(\beta)\ \triangleq\ \sum_{j=1}^{n}\left[\,A_{ij}^{+}\,\lambda_j^{+}\,|\,\beta_j\,|+A_{ij}^{-}\,g_j^{-}(\beta_j)\,\right]\quad\text{and}$$

$$\psi_i(\beta)\ \triangleq\ \sum_{j=1}^{n}\left[\,A_{ij}^{-}\,\lambda_j^{-}\,|\,\beta_j\,|+A_{ij}^{+}\,g_j^{+}(\beta_j)\,\right]$$

are convex functions. Thus, SOL-ASC$_{\delta^{\pm}}(A, b)$ is a "dc set"; i.e., it has the representation:

$$\text{SOL-ASC}_{\delta^{\pm}}(A, b)\ =\ \{\,\beta\ \mid\ \zeta_i(\beta)\ \le\ b_i,\ \forall\, i\ =1,\cdots,m\,\}.$$

We recall that the directional derivative of a function ζ at a given vector $\bar{\beta}$ in the direction v is given by:

$$\zeta'(\bar{\beta};\, v)\ \triangleq\ \lim_{\tau\downarrow 0}\frac{\zeta(\bar{\beta}+\tau v)-\zeta(\bar{\beta})}{\tau}$$

if the limit exists. The following representation of the tangent cone of SOL-ASC$_{\delta^{\pm}}$ (A, b) is directly adopted from Corollary 1 and Proposition 3 in [26] where a proof can be found.

Proposition 11 *Let $\bar{\beta} \in$ SOL-ASC$_{\delta^\pm}(A, b)$ be given. Let each $p_j^\pm(\bullet, \delta_j^\pm)$ be given by (17). Let*

$$\mathscr{A}_{\delta^\pm}(\bar{\beta}) \triangleq \left\{ i \mid \sum_{j=1}^{n} A_{ij}^+ p_j^+(\bar{\beta}_j, \delta_j^+) = \sum_{j=1}^{n} A_{ij}^- p_j^-(\bar{\beta}_j, \delta_j^-) + b_i \right\}.$$

If each function g_{jk}^- is linear for all $j = 1, \cdots, n$ and $k = 1, \cdots, K_j$, it holds that

$$\mathscr{T}\left(\bar{\beta}; \text{SOL-ASC}_{\delta^\pm}(A, b)\right) = \left\{ v \mid \zeta_i'(\bar{\beta}; v) \leq 0, \ \forall\, i \in \mathscr{A}_{\delta^\pm}(\bar{\beta}) \right\}; \tag{19}$$

hence $\mathscr{T}\left(\bar{\beta}; \text{SOL-ASC}_{\delta^\pm}(A, b)\right)$ is the union of finitely many closed convex cones. □

The linearity of the functions g_{jk}^- implies that each approximating function $p_j^-(\bullet, \delta_j^-)$ is piecewise linear although no such piecewise linearity is required on $p_j^+(\bullet, \delta_j^+)$. Among the five families: SCAD, MCP, capped ℓ_1, truncated transformed ℓ_1, and truncated logarithmic, only the capped ℓ_1 function is piecewise linear; the SCAD and MCP functions are differentiable on the real line except at the origin; the latter two truncated functions have two additional non-differentiable points at $t = \pm\delta$.

In terms of the functions $p_j^\pm(\bullet, \delta_j^\pm)$, the expression (19) yields the following:

$$\begin{aligned} &\mathscr{T}\left(\bar{\beta}; \text{SOL-ASC}_{\delta^\pm}(A, b)\right) \\ &= \left\{ v \mid \sum_{j=1}^{n} A_{ij}^+ p_j^+(\bullet, \delta_j^+)'(\bar{\beta}_j; v_j) \leq \sum_{j=1}^{n} A_{ij}^- p_j^-(\bullet, \delta_j^-)'(\bar{\beta}_j; v_j), \forall i \in \mathscr{A}_{\delta^\pm}(\bar{\beta}) \right\}. \end{aligned} \tag{20}$$

Unlike the tangent cone of SOL-ASC [cf. (15)], no closure is needed in the above right-hand set because this set is already closed. We provide an example showing that the equality (19) can fail without the linearity assumption on the functions g_{jk}^- in Proposition 11.

Example 1 cont. Consider the system $|\beta_1|_0 \leq |\beta_2|_0$ at the feasible point $\bar{\beta} = (2, 2)$. We approximate both ℓ_0 functions by the MCP functions with a fixed $\delta = 2$ as follows: for $i = 1, 2$,

$$\rho_i(\beta_i) = |\beta_i| - g_i(\beta_i), \text{ where } g_i(t) \triangleq \begin{cases} t^2/4 & \text{if } |t| \leq 2 \\ |t| - 1 & \text{if } |t| \geq 2. \end{cases}$$

It is easy to check that g_i is convex and continuously differentiable on $\mathbb{R}$. Moreover, $\rho_i(\beta_i)$ is differentiable everywhere except at $\beta_i = 0$. The dc set is

$$\begin{aligned} &\left\{ (\beta_1, \beta_2) \in \mathbb{R}^2 \mid \rho_1(\beta_1) \leq \rho_2(\beta_2) \right\} \\ &= \left\{ (\beta_1, \beta_2) \in \mathbb{R}^2 \mid |\beta_1| \leq |\beta_2| \text{ or } |\beta_2| \geq 2 \right\}. \end{aligned} \tag{21}$$

Since $\rho_i'(2) = 0$, it follows that

$$\left\{ (v_1, v_2) \in \mathbb{R}^2 \mid \rho_1'(2)v_1 \leq \rho_2'(2)\, v_2 \right\} = \mathbb{R}^2. \tag{22}$$

In contrast, the actual tangent cone of the set (21) at the given $\bar{\beta} = (2, 2)$ is:

$$\left\{ (d_1, d_2) \in \mathbb{R}^2 \mid d_1 \leq d_2 \text{ or } d_2 \geq 0 \right\}$$

which is clearly a subset of (22). □

6 Convergence of B-stationary solutions

In this section, we apply the results derived in the above sections to address the convergence of stationary solutions. Consider the following optimization problem:

$$\underset{\beta}{\text{minimize}}\ f(\beta) \triangleq h(\beta) - g(\beta) \ \text{ subject to } \beta \in C \triangleq \text{cl}\,[\text{SOL-ASC(A,b)}]\,, \tag{23}$$

where h is a convex function (not necessarily differentiable) and g is a continuously differentiable convex function, both defined on an open set Ω containing the feasible set C. Thus f is a difference-of-convex (dc) function. The non-convexity of cl [SOL-ASC(A,b)] adds complications to the problem. As a non-convex optimization problem, we cannot realistically hope to be able to compute a local minimizer easily, let alone a global minimizer. At best, a stationary solution of some kind is what an iterative algorithm can approximately compute. In general, stationarity is a necessary condition for the local minimizing property. For a non-convex non-differentiable optimization problem, there are many kinds of stationary solutions. When the problem is constrained by a non-convex set, one needs to consider a constrained notion of stationarity. As detailed in [26], to deal with a dc optimization problem with a non-convex non-differentiable feasible set, the sharpest kind of stationary solutions is that based on the elementary directional derivatives of the objective functions (which are well defined for a dc function) and the tangent cone of the constraint set, sharpest in the sense that the resulting stationary solution must satisfy all other definitions of stationarity. The cited reference contains more details about such a stationary point, which is termed a B(ouligand) stationary solution, and discussion about its computation.

By definition, a feasible vector $\bar{\beta} \in C$ is a *B(ouligand) stationary solution* of the problem (23) if

$$h'(\bar{\beta}; v) - \nabla g(\bar{\beta})^T v = f'(\bar{\beta}; v) \geq 0, \forall\, v \in \mathscr{T}(\bar{\beta}; C).$$

We attempt to approximate such a stationary solution $\bar{\beta}$ by solving the approximated problem:

$$\underset{\beta}{\text{minimize}}\ f(\beta) \text{ subject to } \beta \in C(\delta^{\pm}) \triangleq \text{SOL-ASC}_{\delta^{\pm}}(A, b), \tag{24}$$

where the feasible region is defined by the family of approximating functions $\left\{p_j^{\pm}(\bullet, \delta_j^{\pm})\right\}_{j=1}^{n}$ each satisfying assumptions (R1), (R2), and (R3) as well as condition (12) or (13).

Let $\{\delta^{\pm;k}\}$ be a sequence of positive scalars converging to zero. For each k, let $\bar{\beta}^k$ be a B-stationary solution of (24) corresponding to $\delta^{\pm;k}$; i.e.,

$$f'(\bar{\beta}^k; v) \geq 0, \forall\, v \in \mathscr{T}(\bar{\beta}^k; C(\delta^{\pm;k})).$$

Suppose that the sequence $\{\bar{\beta}^k\}$ converges to $\bar{\beta}$, which must necessarily belong to C by the convergence of $\{C(\delta^{\pm;k})\}$ to C. The question is whether $\bar{\beta}$ is a B-stationary solution of (23). For this question to have an affirmative answer, it suffices to identify for any $\widehat{v} \in \mathscr{T}(\bar{\beta}; C)$ an infinite index set $\kappa \subseteq \{1, 2, \cdots\}$ and a sequence of tangents $\{v^k\}_{k\in\kappa}$ with $v^k \in \mathscr{T}(\bar{\beta}^k; C(\delta^{\pm;k}))$ for each k such that $\{v^k\}_{k\in\kappa}$ converges to $\widehat{v}$. Indeed, if such v^k exist, then using the fact that $h'(\bar{\beta}; \widehat{v}) \geq \limsup_{k(\in\kappa)\to\infty} h'(\bar{\beta}^k; v^k)$, by the convexity of h, we deduce,

$$f'(\bar{\beta}; \widehat{v}) \geq \limsup_{k(\in\kappa)\to\infty} f'(\bar{\beta}^k; v^k),$$

from which the desired B-stationarity of $\bar{\beta}$ follows readily. If f is differentiable, then it suffices for $\{\nabla f(\bar{\beta}^k)^T v^k\} \to \nabla f(\bar{\beta})^T \widehat{v}$. Before constructing the desired sequence $\{v^k\}$, we provide an example to illustrate that this may not always be possible, thus establishing the failed convergence of such stationary solutions in general.

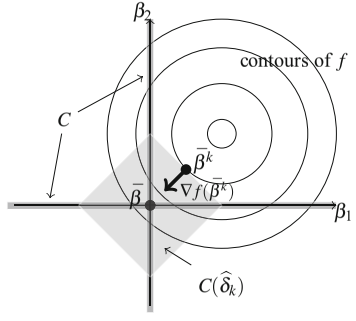
Example 5 Consider the 2-variable optimization problem:

$$\underset{\beta_1,\beta_2}{\text{minimize}}\ \tfrac{1}{2}\left[(\beta_1 - 1)^2 + (\beta_2 - 1)^2\right] \text{ subject to } |\beta_1|_0 + |\beta_2|_0 \leq 1. \tag{25}$$

We approximate the constraint by the capped ℓ_1 surrogate function, obtaining the approximated problem: for $\widehat{\delta}_k > 0$,

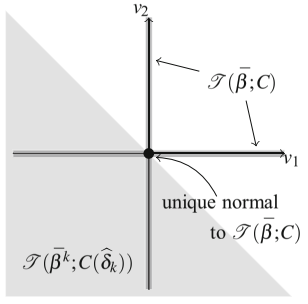
$$\begin{aligned} &\underset{\beta_1,\beta_2}{\text{minimize}}\ \tfrac{1}{2}\left[(\beta_1 - 1)^2 + (\beta_2 - 1)^2\right] \text{ subject to} \\ &\min\left(\frac{|\beta_1|}{\widehat{\delta}_k}, 1\right) + \min\left(\frac{|\beta_2|}{\widehat{\delta}_k}, 1\right) \leq 1. \end{aligned} \tag{26}$$

It is not difficult to verify that $\bar{\beta}(\widehat{\delta}_k) \triangleq \left(\frac{\widehat{\delta}_k}{2}, \frac{\widehat{\delta}_k}{2}\right)$ is a B-stationary solution of the latter problem. Yet the limit $(0, 0)$ is not a B-stationary solution of the original problem (25), by noting that the tangent cone of the feasible region at this limit point is equal to the feasible region itself. For this example, we note that the constraint in (26) is binding at the approximated pair $\bar{\beta}(\widehat{\delta}_k)$ but the constraint in (25) is not binding at the limit $(0, 0)$. Incidentally, this situation will not happen with a polyhedron under perturbation but happens here partly due to the discontinuity of the ℓ_0 function (Fig. 2). □



$$C = \{(\beta_1, \beta_2) \mid |\beta_1|_0 + |\beta_2|_0 \leq 1\}$$

$$C(\widehat{\delta}_k) = \left\{(\beta_1, \beta_2) \mid \min\left(\frac{|\beta_1|}{\widehat{\delta}_k}, 1\right) + \min\left(\frac{|\beta_2|}{\widehat{\delta}_k}, 1\right) \leq 1\right\}$$



$$\mathscr{T}(\bar{\beta}; C) = \{(v_1, v_2) \mid |v_1|_0 + |v_2|_0 \leq 1\}$$

$$\mathscr{T}(\bar{\beta}^k; C(\widehat{\delta}_k)) = \{(v_1, v_2) \mid v_1 + v_2 \leq 0\}$$

Fig. 2 Illustration of Example 5 using iterates

For operational purposes, we assume in the rest of the paper that the representation (20) of the tangent cone $\mathscr{T}(\bar{\beta}^k; C(\delta^{\pm;k}))$ of the approximated set $C(\delta^{\pm;k})$ is valid. To facilitate the identification of the desired approximating tangents, we write the two cones $\mathscr{T}(\bar{\beta}; C)$ and $\mathscr{T}(\bar{\beta}^k; C(\delta^{\pm;k}))$ as follows. First, let $\bar{S} \triangleq \text{supp}(\bar{\beta})$ with complement $\bar{S}^{\text{c}}$. We have

$$\mathscr{T}(\bar{\beta}; C) = \text{cl}\left[\left\{ v \mid \sum_{j \in \bar{S}^{\text{c}}} A_{ij} \mid v_j \mid_0 \leq b_i - \sum_{j=1}^{n} A_{ij} \mid \bar{\beta}_j \mid_0, \ \forall\, i = 1, \cdots, m \right\}\right]$$

while $\mathscr{T}\left(\bar{\beta}^k; C(\delta^{\pm;k})\right)$

$$= \left\{ v \mid \sum_{j=1}^{n} A_{ij}^{+} p_j^{+}(\bullet, \delta_j^{+;k})'(\bar{\beta}_j^k; v_j) \right.$$

$$\left. \leq \sum_{j=1}^{n} A_{ij}^{-} p_j^{-}(\bullet, \delta_j^{-;k})'(\bar{\beta}_j^k; v_j), \ \forall\, i \in \mathscr{A}_{\delta^{\pm;k}}(\bar{\beta}^k) \right\}.$$

For any vector $v \in \mathbb{R}^n$, let

$$\mathscr{V}_{=0} \triangleq \left\{ j \in \bar{S}^{\text{c}} \mid j \notin \text{supp}(v) \right\} \text{and} \quad \mathscr{V}_{\neq 0} \triangleq \left\{ j \in \bar{S}^{\text{c}} \mid j \in \text{supp}(v) \right\}$$

whose union is the complement $\bar{S}^{\text{c}}$ of the support of the vector $\bar{\beta}$. We divide the sum $\sum_{j=1}^{n} A_{ij}^{\pm} p_j^{\pm}(\bullet, \delta_j^{\pm;k})'(\bar{\beta}_j^k; v_j)$ according to a given vector $\widehat{v} \in \mathscr{T}(\bar{\beta}; C)$ and the two associated index sets $\widehat{\mathscr{V}}_{=0}$ and $\widehat{\mathscr{V}}_{\neq 0}$:

$$\sum_{j=1}^{n} A_{ij}^{\pm} p_j^{\pm}(\bullet, \delta_j^{\pm;k})'(\bar{\beta}_j^k; v_j) = \sum_{j \in \bar{S}} A_{ij}^{\pm} p_j^{\pm}(\bullet, \delta_j^{\pm;k})'(\bar{\beta}_j^k; v_j) + T_{k;0}^{\pm}(v) + T_{k;\neq 0}^{\pm}(v),$$

where

$$T_{k;0}^{\pm}(v) \triangleq \sum_{j \in \widehat{\mathscr{V}}_{=0}} A_{ij}^{\pm} p_j^{\pm}(\bullet, \delta_j^{\pm;k})'(\bar{\beta}_j^k; v_j) \text{and}$$

$$T_{k;\neq 0}^{\pm}(v) \triangleq \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{ij}^{\pm} p_j^{\pm}(\bullet, \delta_j^{\pm;k})'(\bar{\beta}_j^k; v_j).$$

Letting $\delta_j^{\pm;k} = \widehat{\delta}_k$ for all $j = 1, \cdots, n$, we can write

$$T_{k;0}^{\pm}(v) = \frac{1}{\widehat{\delta}_k} \sum_{j \in \widehat{\mathscr{V}}_{=0}} A_{ij}^{\pm} \left[\widehat{\delta}_k \, p_j^{\pm}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j) \right]$$

$$\text{and } T_{k;\neq 0}^{\pm}(v) = \frac{1}{\widehat{\delta}_k} \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{ij}^{\pm} \left[\widehat{\delta}_k \, p_j^{\pm}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j) \right].$$

Under assumption (R3) of the functions $p_j^{\pm}(\bullet, \delta)$, we deduce the following two one-sided derivatives for all k sufficiently large and all $j \in \bar{S}$:

$$p_j^{\pm}(\bullet, \widehat{\delta}_k)'(t; \pm 1) \;=\; 0, \forall\, t \text{ such that } |\, t\,| \;>\; \widehat{\delta}_k.$$

For an index $j \in \bar{S}$, since $\{\bar{\beta}_j^k\} \to \bar{\beta}_j \neq 0$ and $\{\widehat{\delta}_k\} \to 0$, it follows that for all but finitely many k, $\bar{\beta}_j^k > \widehat{\delta}_k$. Hence, under the stipulation that $\delta_j^{\pm;k} = \widehat{\delta}_k$ for all $j = 1, \cdots, n$, it follows that for all k sufficiently large, $v \in \mathscr{T}\left(\bar{\beta}^k; C(\delta^{\pm;k})\right)$ if and only if for all $i \in \mathscr{A}_{\delta^{\pm;k}}(\bar{\beta}^k)$,

$$\begin{array}{l} \displaystyle\sum_{j \in \widehat{\mathscr{V}}_{=0}} A_{ij}^{+} \left[\widehat{\delta}_k\, p_j^{+}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j) \right] + \displaystyle\sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{ij}^{+} \left[\widehat{\delta}_k\, p_j^{+}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j) \right] \\ \leq \displaystyle\sum_{j \in \widehat{\mathscr{V}}_{=0}} A_{ij}^{-} \left[\widehat{\delta}_k\, p_j^{-}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j) \right] + \displaystyle\sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{ij}^{-} \left[\widehat{\delta}_k\, p_j^{-}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j) \right] \end{array} \tag{27}$$

This can be contrasted with the necessary and sufficient condition for $\widehat{v} \in \mathscr{T}(\bar{\beta}; C)$, which is:

$$\sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{ij}^{+} \;\leq\; \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{ij}^{-} + \left[b_i - \sum_{j \in \bar{S}} A_{ij} \right], \; \forall\, i \;= 1, \cdots, m, \tag{28}$$

provided that the set of vectors $\widehat{v}$ satisfying the latter inequalities is closed. This is the case for instance when the matrix A is nonnegative as we will assume in Sect. 6.1. One obvious difference between (27) and (28) is that the components v_j, for $j \in \widehat{\mathscr{V}}_{=0}$, of the tangent v of the approximated set $C(\delta^{\pm;k})$ appear explicitly in the former, whereas the same components of $\widehat{v}$ do not in the latter.

At this point, it would be useful to provide the directional derivatives of the surrogate sparsity functions discussed in Sect. 4.1, in particular the derived expressions will verify the following properties of $\rho(\bullet, \delta)'(t; \pm 1)$ for all $\delta > 0$ and all nonzero $s \in \mathbb{R}$, namely,

(R4a) $\operatorname{sign}\left[\, \rho(\bullet, \delta)'(t; s) \,\right] \begin{cases} = \operatorname{sign}(s) \operatorname{sign}(t) & \text{if } 0 < |\, t\,| \;<\; \delta \\ = 1 & \text{if } t = 0; \end{cases}$

(R4b) $\rho(\bullet, \delta)'(t; s) \begin{cases} = 0 \text{ if } st > 0 \\ \leq 0 \text{ if } st < 0 \end{cases}$ for $|t| = \delta$; and

(R4c) $\rho(\bullet, \delta)'(t; s) = 0$ for $|t| > \delta$.

The function $\rho(\bullet, \delta)$ is not differentiable at the origin and possibly at $\pm\delta$ (see e.g. the truncated transformed ℓ_1-function below). By definition, we must have $\rho(\bullet, \delta)'(t; 0) = 0$ for all t. Moreover, for $st \neq 0$, $\rho(\bullet, \delta)'(t; s) > 0$ if and only if $|t| < \delta$ and $st > 0$.

In what follows, we give expressions of the directional derivatives of three such functions: SCAD, MCP, and the truncated transformed ℓ_1, and omit the other two: capped ℓ_1 and truncated logarithmic.

The SCAD **family.** We have

$$\rho_a^{\mathrm{SCAD}}(\bullet,\delta)'(t;\pm 1) = \begin{cases} \pm\,\mathrm{sign}(t)\,\dfrac{2\,a}{(a+1)\,\delta} & \text{if } 0 < |t| \le \dfrac{\delta}{a} \\ \dfrac{2\,a}{(a+1)\,\delta} & \text{if } t = 0 \\ \pm\,\mathrm{sign}(t)\,\dfrac{2\,(\delta - |t|)}{\left(1 - \dfrac{1}{a^2}\right)\delta^2} & \text{if } \dfrac{\delta}{a} \le |t| \le \delta \\ 0 & \text{if } |t| \ge \delta, \end{cases}$$

which is continuously differentiable at all nonzero $t \in \mathbb{R}$. Thus

$$\delta\,\rho_a^{\mathrm{SCAD}}(\bullet,\delta)'(t;s) = \begin{cases} s\,\mathrm{sign}(t)\,\dfrac{2\,a}{a+1} & \text{if } 0 < |t| \le \dfrac{\delta}{a} \\ |s|\,\dfrac{2\,a}{a+1} & \text{if } t = 0 \\ s\,\mathrm{sign}(t)\,\dfrac{2\left(1 - \dfrac{|t|}{\delta}\right)}{1 - \dfrac{1}{a^2}} & \text{if } \dfrac{\delta}{a} \le |t| \le \delta \\ 0 & \text{if } |t| \ge \delta. \end{cases}$$

The MCP **family.** We have

$$\rho^{\mathrm{MCP}}(\bullet,\delta)'(t;\pm 1) = \begin{cases} \pm\,\mathrm{sign}(t)\left(\dfrac{2}{\delta} - \dfrac{2\,|t|}{\delta^2}\right) & \text{if } 0 < |t| \le \delta \\ \dfrac{2}{\delta} & \text{if } t = 0 \\ 0 & \text{if } |t| \ge \delta, \end{cases}$$

which has the same differentiability properties as a SCAD function. Moreover,

$$\delta\,\rho^{\mathrm{MCP}}(\bullet,\delta)'(t;s) = \begin{cases} s\,\mathrm{sign}(t)\left(2 - \dfrac{2\,|t|}{\delta}\right) & \text{if } 0 < |t| \le \delta \\ 2\,|s| & \text{if } t = 0 \\ 0 & \text{if } |t| \ge \delta. \end{cases}$$

The truncated transformed ℓ_1 **family.** We have

$$\rho_a^{\mathrm{TTL_1}}(\bullet,\delta)'(t;\pm 1) = \begin{cases} \pm\,\mathrm{sign}(t)\,\dfrac{a\,(a+\delta)}{\delta\,(a+|t|)^2} & \text{if } 0 < |t| < \delta \\ \dfrac{a+\delta}{\delta\,a} & \text{if } t = 0 \\ \min\left(\pm\,\mathrm{sign}(t)\,\dfrac{a}{\delta\,(a+\delta)},\ 0\right) & \text{if } |t| = \delta \\ 0 & \text{if } |t| > \delta, \end{cases}$$

yielding

$$\delta\,\rho_a^{\mathrm{TTL}_1}(\bullet,\delta)'(t;s) \;=\; \begin{cases} s\,\mathrm{sign}(t)\,\dfrac{a\,(a+\delta\,)}{(a+|\,t\,|\,)^2} & \text{if } 0 \;<\; |\,t\,| \;<\; \delta \\ |\,s\,|\,\dfrac{a+\delta}{a} & \text{if } t = 0 \\ \min\left(s\,\mathrm{sign}(t)\,\dfrac{a}{a+\delta},\,0\right) & \text{if } |\,t\,| \;=\; \delta \\ 0 & \text{if } |\,t\,| \;>\; \delta. \end{cases}$$

6.1 The case $A \geq 0$

Using the last inclusion of the tangent cone in Proposition (10), this nonnegative case is relatively easy to deal with.

Proposition 12 *Let $A \geq 0$ and $f = h - g$ be a dc function with g and h both convex and g additionally continuously differentiable. Let $\delta_j^{\pm;k} = \widehat{\delta}_k$ for all $j = 1, \cdots, n$ and all k. Let $\lim_{k\downarrow 0} \widehat{\delta}_k = 0$. For each k, let $\bar{\beta}^k$ be a B-stationary solution of the problem:*

$$\underset{\beta}{\mathit{minimize}} \quad f(\beta)$$
$$\mathit{subject\ to}\ \beta \in \widehat{C}(\widehat{\delta}_k) \triangleq \left\{ \beta \;|\; \sum_{j=1}^{n} A_{ij}\, p_j^+(\beta_j, \widehat{\delta}_k) \;\leq\; b_i, \forall\, i \;=\; 1, \cdots, m \right\}.$$

where each surrogate function $p_j^+(\bullet, \widehat{\delta}_k)$ satisfies conditions (R1)–(R4). If $\bar{\beta}$ is the limit of $\{\bar{\beta}^k\}$ satisfying the property:

$\left(\mathrm{A}^{\geq 0}_{\bar{\beta}}\right)$*: if i is such that*

$$\sum_{j=1}^{n} A_{ij}\, p_j^+(\bar{\beta}_j^k, \widehat{\delta}_k) \;=\; b_i \tag{29}$$

for infinitely many k, then $i \in \mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$,

then $\bar{\beta}$ is a B-stationary solution of (23).

Proof Let $\widehat{v} \in \mathscr{T}(\bar{\beta}; C)$ with $C \triangleq \mathrm{SOL\text{-}ASC}(A, b)$. It suffices to construct a sequence $\{v^k\}$ and identify an infinite index set κ such that the subsequence $\{v^k\}_{k\in\kappa}$ converges to $\widehat{v}$ and for all $k \in \kappa$ sufficiently large,

$$\left.\begin{array}{l} 0 \geq \displaystyle\sum_{j\in\widehat{\mathscr{V}}_{=0}} A_{ij}\left[\widehat{\delta}_k\, p_j^+(\bullet,\widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k)\right] + \\ \quad\displaystyle\sum_{j\in\widehat{\mathscr{V}}_{\neq 0}} A_{ij}\left[\widehat{\delta}_k\, p_j^+(\bullet,\widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k)\right] \end{array}\right\} \forall\, i \text{ such that (29) holds.} \tag{30}$$

Define the components v_j^k as follows:

$$v_j^k \triangleq \begin{cases} \widehat{v}_j & \text{if either } j \in \bar{S} \text{ or } j \in \widehat{\mathscr{V}}_{\neq 0} \\ -\bar{\beta}_j^k & \text{if } j \in \widehat{\mathscr{V}}_{=0} \end{cases}$$

For every k, there is a (possibly empty) index set $\mathscr{A}_k$ of constraints i such that (29) holds corresponding to the pair (i, k). Since there are only finitely constraints, there exists an infinite subset κ of $\{1, 2, \cdots\}$ such that $\mathscr{A}_k$ is a constant set, say $\bar{\mathscr{A}}$, for all $k \in \kappa$. By assumption $\left(\mathrm{A}_{\bar{\beta}}^{\geq 0}\right)$, we have $\bar{\mathscr{A}} \subseteq \mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$. It follows from Proposition 10 that $\widehat{v}_j = 0$ provided that there exists an $i \in \mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$ such that $A_{ij} > 0$. Thus, for every index $j \in \widehat{\mathscr{V}}_{\neq 0}$, we must have $A_{ij} = 0$ for all $i \in \mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$. Hence, the requirement (30) for the sequence $\{v^k\}_{k\in\kappa}$ reduces to

$$0 \geq \sum_{j \in \widehat{\mathscr{V}}_{=0}} A_{ij} \left[\widehat{\delta}_k \, p_j^+(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k) \right] \forall\, i \in \bar{\mathscr{A}} \text{ and all } k \in \kappa.$$

By (R4), we have, for $j \in \widehat{\mathscr{V}}_{=0}$ with $\bar{\beta}_j^k \neq 0$, $\mathrm{sign}\left[p_j^+(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k) \right] = \mathrm{sign}(v_j^k)\,\mathrm{sign}(\bar{\beta}_j^k) = -1$, by the choice of v_j^k. Hence it follows that $v^k \in \mathscr{T}\left(\bar{\beta}^k; \text{SOL-ASC}_{\widehat{\delta}_k}(A, b)\right)$ for all $k \in \kappa$. It remains to show that $\{v^k\}_{k\in\kappa}$ converges to $\widehat{v}$. But this is clear from the definition of the components v_j^k and the fact that $\bar{\beta}_j^k \to \bar{\beta}_j = 0 = \widehat{v}_j$ for all $j \in \widehat{\mathscr{V}}_{=0}$. □

6.2 The case of column-wise uni-sign

In this subsection, we assume that the objective function f in (23) is continuously differentiable (C^1) so that we can focus on the choice of the sequence of approximate tangents. Let $W(\bar{\beta})$ be the set of vectors $\widehat{v}$ satisfying the inequalities in (28). Since the closure of $W(\bar{\beta})$ is equal to $\mathscr{T}(\bar{\beta}, \text{SOL-ASC}(A, b))$, it follows that a necessary and sufficient condition for $\bar{\beta}$ to be a B-stationary solution of (23) is that

$$0 \leq \nabla f(\bar{\beta})^T \widehat{v} = \sum_{j \in \bar{S}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \widehat{v}_j + \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \widehat{v}_j, \forall\, \widehat{v} \in W(\bar{\beta}). \tag{31}$$

In what follows, we obtain a necessary and sufficient condition for a given $\bar{\beta} \in \text{SOL-ASC}(A, b)$ to be a B-staionary solution of (23), based on which we will address the convergence of the B-stationary solutions of the approximated problems. For this purpose, define

$$\mathscr{J}(\bar{\beta}) \triangleq \left\{ J \subseteq \bar{S}^{\mathrm{c}} \mid \sum_{j \in J} A_{\bullet j} \leq b - A \parallel \bar{\beta} \parallel_0 \right\}.$$

This family $\mathscr{J}(\bar{\beta})$ collects all index sets $\widehat{\mathscr{V}}_{\neq 0}$ corresponding to the tangents $\widehat{v}$ in $\mathscr{T}(\bar{\beta}; \text{SOL-ASC}(A, b))$.

Lemma 1 *Let f be a C^1 function defined on an open set containing SOL-ASC(A, b). A vector $\bar{\beta} \in$ SOL-ASC(A, b) is a B-stationary solution of (23) if and only if*

$$\frac{\partial f(\bar{\beta})}{\partial \beta_j} = 0, \forall\, j \in \mathit{supp}(\bar{\beta}) \cup \bigcup_{J \in \mathscr{J}(\bar{\beta})} J. \tag{32}$$

Proof "If." Associated with every vector $\widehat{v} \in W(\bar{\beta})$ is an index set $\widehat{J}$ satisfying $\sum_{j \in \widehat{J}} A_{\bullet j} \leq b - A \parallel \bar{\beta} \parallel_0$. By assumption, $\dfrac{\partial f(\bar{\beta})}{\partial \beta_j} = 0$ for all $j \in \widehat{J}$. The equality in the expression for $\nabla f(\bar{\beta})^T \widehat{v}$ in (31) and the assumption (32) easily yields the B-stationarity of $\bar{\beta}$.

"Only if." Suppose that the inequality in (31) holds. Any vector $\widehat{v}$ with $\mathscr{V}_{\neq 0} = \emptyset$ belongs to the set $W(\bar{\beta})$. Thus, we must have

$$\sum_{j \in \bar{S}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \widehat{v}_j \geq 0, \forall \widehat{v}_{\bar{S}}.$$

This implies that $\dfrac{\partial f(\bar{\beta})}{\partial \beta_j} = 0$ for all $j \in \bar{S}$. Hence the B-stationarity of $\bar{\beta}$ reduces to

$$\sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \widehat{v}_j \geq 0, \forall \left(\widehat{v}_j\right)_{j \in \widehat{\mathscr{V}}_{\neq 0}} \text{ satisfying } \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} A_{\bullet j} \leq b - A \parallel \bar{\beta} \parallel_0.$$

Let $\widehat{j} \in \widehat{J}$ for some $\widehat{J} \in \mathscr{J}(\bar{\beta})$. We have $\sum_{j' \in \widehat{J}} A_{\bullet j'} \leq b - A \parallel \bar{\beta} \parallel_0$. For any scalar $\varepsilon > 0$, define the vectors $\widehat{v}^{\varepsilon;\pm}$ as follows:

$$\widehat{v}_j^{\varepsilon;\pm} \triangleq \begin{cases} 0 & \text{if } j \notin \widehat{J} \\ \varepsilon & \text{if } j \in \widehat{J} \text{ and } j \neq \widehat{j} \\ \pm 1 & \text{if } j = \widehat{j}. \end{cases}$$

It is not difficult to see that $\left\{ j \notin \bar{S} \mid \widehat{v}_j^{\varepsilon;\pm} \neq 0 \right\} = \widehat{J}$. Hence we have

$$0 \leq \varepsilon \sum_{j \in \widehat{J} \setminus \{\widehat{j}\}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \pm \frac{\partial f(\bar{\beta})}{\partial \beta_{\widehat{j}}}.$$

Since this holds for all $\varepsilon > 0$, it follows by passing to the limit $\varepsilon \downarrow 0$ that

$$\pm \frac{\partial f(\bar{\beta})}{\partial \beta_{\widehat{j}}} \geq 0, \text{ yielding} \frac{\partial f(\bar{\beta})}{\partial \beta_{\widehat{j}}} = 0,$$

establishing that (32) is necessary for B-stationarity. □

Based on the above lemma, we can establish the following result when the matrix A satisfies the column-wise uni-sign property. With this special structure on A, we may divide the columns of A into two groups: $\mathscr{C}_{\oplus}$ whose entries are all nonnegative, and $\mathscr{C}_{\ominus}$ whose entries are all nonpositive. The noteworthy part in the proof of the result is that not all components of the sequence $\{v^k\}$ converge to the corresponding components of $\widehat{v}$ (those with indices in the sets $\widehat{\mathscr{V}}_{\neq 0}^{\ominus;\oplus}$ do not; see notation below), but we must have $\nabla f(\bar{\beta}^k)^T v^k \to \nabla f(\bar{\beta})^T \widehat{v}$ restricted to an appropriate subsequence; this is sufficient to establish the desired B-stationarity of the limit $\bar{\beta}$.

Theorem 2 *Let A have the column-wise uni-sign property and f be a C^1 function defined on an open set containing SOL-ASC(A, b). Let $\delta_j^{\pm;k} = \widehat{\delta}_k$ for all $j = 1, \cdots, n$ and all k. Let $\lim_{k\downarrow 0} \widehat{\delta}_k = 0$. For each k, let $\bar{\beta}^k$ be a B-stationary solution of the problem:*

$$\begin{array}{l} \underset{\beta}{\text{minimize}} \quad f(\beta) \\ \text{subject to } \beta \in \widehat{C}(\widehat{\delta}_k) \triangleq \left\{ \beta \mid \sum_{j=1}^{n} A_{ij}^{+} p_j^{+}(\beta_j, \widehat{\delta}_k) \le b_i + \sum_{j=1}^{n} A_{ij}^{-} p_j^{-}(\beta_j, \widehat{\delta}_k), \right. \\ \qquad \left. \forall\, i = 1, \cdots, m \right\}, \end{array}$$

where each pair of surrogate functions $p_j^{\pm}(\bullet, \widehat{\delta}_k)$ satisfies conditions (R1)–(R4) and either (12) or (13). If $\bar{\beta}$ is the limit of $\{\bar{\beta}^k\}$ satisfying the property that

($A_{\bar{\beta}}$) if i is such that

$$\sum_{j=1}^{n} A_{ij}^{+} p_j^{+}(\bar{\beta}_j^k, \widehat{\delta}_k) = b_i + \sum_{j=1}^{n} A_{ij}^{-} p_j^{-}(\bar{\beta}_j^k, \widehat{\delta}_k) \tag{33}$$

for infinitely many k, then $i \in \mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$,

then $\bar{\beta}$ is a B-stationary solution of (23) if and only if

$$\frac{\partial f(\bar{\beta})}{\partial \beta_j} = 0, \forall\, j \in \bigcup_{J \in \mathscr{J}(\bar{\beta})} J \text{ such that } \exists\, i \in \mathscr{A}_{\mathrm{ASC}}(\bar{\beta}) \text{ with } A_{ij} \neq 0. \tag{34}$$

Proof It suffices to prove the "if" statement. We proceed as in the proof of Proposition 12. For every k, there is a (possibly empty) index set $\mathscr{A}_k$ of constraints i such that (33) holds corresponding to the pair (i, k). Since there are only finitely constraints, there exists an infinite subset κ of $\{1, 2, \cdots\}$ such that $\mathscr{A}_k$ is a constant set, say $\bar{\mathscr{A}}$, for

all $k \in \kappa$. By assumption $(\mathrm{A}_{\bar{\beta}})$, we have $\bar{\mathscr{A}} \subseteq \mathscr{A}_{\mathrm{ASC}}(\bar{\beta})$. Define the components v_j^k as follows:

$$v_j^k \triangleq \begin{cases} \widehat{v}_j & \text{if } j \in \bar{S} \\ \bar{\beta}_j^k & \text{if } j \in \widehat{\mathscr{V}}_{=0} \cap \mathscr{C}_{\ominus} \triangleq \widehat{\mathscr{V}}_{=0}^{\ominus} \\ -\bar{\beta}_j^k & \text{if } j \in \widehat{\mathscr{V}}_{=0} \cap \mathscr{C}_{\oplus} \triangleq \widehat{\mathscr{V}}_{=0}^{\oplus} \\ \widehat{v}_j & \text{if } j \in \widehat{\mathscr{V}}_{\neq 0} \text{ and } A_{ij} = 0 \,\forall\, i \in \mathscr{A}_{\mathrm{ASC}}(A,b); \quad \text{denoted } j \in \widehat{\mathscr{V}}_{\neq 0}^{0} \\ 0 & \text{if } j \in \widehat{\mathscr{V}}_{\neq 0} \cap \mathscr{C}_{\ominus} \text{ and } \exists\, i \in \mathscr{A}_{\mathrm{ASC}}(A,b) \text{ such that } A_{ij} \neq 0; \text{ denoted } j \in \widehat{\mathscr{V}}_{\neq 0}^{\ominus} \\ -\bar{\beta}_j^k & \text{if } j \in \widehat{\mathscr{V}}_{\neq 0} \cap \mathscr{C}_{\oplus} \text{ and } \exists\, i \in \mathscr{A}_{\mathrm{ASC}}(A,b) \text{ such that } A_{ij} \neq 0; \text{ denoted } j \in \widehat{\mathscr{V}}_{\neq 0}^{\oplus}. \end{cases}$$

Note that the components v_j^k for $j \in \widehat{\mathscr{V}}_{\neq 0}^{\ominus;\oplus}$ do not necessarily converge to $\widehat{v}_j$, whereas all other components do. Taking into account the column-wise uni-sign property of A, the inequality in (27) for $i \in \bar{\mathscr{A}}$, which is a subset of $\mathscr{A}_{\mathrm{ASC}}(A,b)$, can be written equivalently as

$$\begin{aligned} &\sum_{j \in \widehat{\mathscr{V}}_{=0}^{\oplus}} A_{ij}^{+} \left[\widehat{\delta}_k \, p_j^{+}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k) \right] + \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}^{\oplus}} A_{ij}^{+} \left[\widehat{\delta}_k \, p_j^{+}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k) \right] \\ &\leq \sum_{j \in \widehat{\mathscr{V}}_{=0}^{\ominus}} A_{ij}^{-} \left[\widehat{\delta}_k \, p_j^{-}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k) \right] + \underbrace{\sum_{j \in \widehat{\mathscr{V}}_{\neq 0}^{\ominus}} A_{ij}^{-} \left[\widehat{\delta}_k \, p_j^{-}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j^k; v_j^k) \right]}_{= 0 \text{ since } v_j^k = 0 \text{ for all such } j}. \end{aligned}$$

By the definition of v_j^k and property (R4) of the directional derivatives $p_j^{\pm}(\bullet, \widehat{\delta}_k)'(\bar{\beta}_j; \bullet)$ of the surrogate functions, it follows that for all $i \in \bar{\mathscr{A}}$, the left-hand sum is nonpositive while the right-hand sum is nonnegative. Thus, the above-defined vector $v^k \in \mathscr{T}(\bar{\beta}^k; \text{SOL-ASC}_{\widehat{\delta}_k}(A,b))$ for all $k \in \kappa$ sufficiently large. Hence, we have for all such k,

$$\begin{aligned} 0 \leq \nabla f(\bar{\beta}^k)^T v^k &= \sum_{j \in \bar{S}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k + \sum_{j \in \widehat{\mathscr{V}}_{=0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k + \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k \\ &= \underbrace{\sum_{j \in \bar{S}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} \widehat{v}_j}_{\substack{\text{converges to} \\ \sum_{j \in \bar{S}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \widehat{v}_j}} + \underbrace{\sum_{j \in \widehat{\mathscr{V}}_{=0}^{\ominus}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} \bar{\beta}_j^k - \sum_{j \in \widehat{\mathscr{V}}_{=0}^{\oplus}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} \bar{\beta}_j^k}_{\text{converges to } 0} + \sum_{j \in \widehat{\mathscr{V}}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k. \end{aligned}$$

We can write

$$\sum_{j\in\widehat{\mathscr{V}}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k = \sum_{j\in\widehat{\mathscr{V}}^{\ominus}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k + \sum_{j\in\widehat{\mathscr{V}}^{\oplus}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k + \sum_{j\in\widehat{\mathscr{V}}^{0}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} v_j^k$$

$$= -\sum_{j\in\widehat{\mathscr{V}}^{\oplus}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} \bar{\beta}_j^k + \sum_{j\in\widehat{\mathscr{V}}^{0}_{\neq 0}} \frac{\partial f(\bar{\beta}^k)}{\partial \beta_j} \widehat{v}_j$$

converges to $\sum_{j\in\widehat{\mathscr{V}}^{0}_{\neq 0}} \frac{\partial f(\bar{\beta})}{\partial \beta_j} \widehat{v}_j$, by (34).

Hence $\left\{\nabla f(\bar{\beta}^k)^T v^k\right\}_{k\in\kappa} \to \nabla f(\bar{\beta})^T \widehat{v}$ as desired. □

The condition (34) is void when A is a nonnegative matrix. At this time, we are not able to address the case where there are nonzero entries in some columns of A that have mixed signs.

Remark 3 The assumptions $(A^{\leqq}_{\bar{\beta}})$ and $(A_{\bar{\beta}})$ in Proposition 12 and Theorem 2, respectively, are not expected to be easily verifiable in practice. Nevertheless, they provide an explanation of the failure of convergence in Example 5; more importantly, these results show that the persistent holding of binding constraints in the limit is essential for the convergence of the B-stationary solutions of the approximated problems to a desired B-stationary solution of the ASC constrained problem (23). □

References

1. Ahn, M., Pang, J.S., Xin, J.: Difference-of-convex learning: directional stationarity, optimality, and sparsity. SIAM J. Optim. Revision under review (as of February 2017)
2. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Structured sparsity through convex optimization. Stat. Sci. **27**(4), 450–468 (2012)
3. Belotti, P., Kirches, C., Leyffer, S., Linderoth, J., Luedtke, J., Mahajan, A.: Mixed-integer nonlinear optimization. Acta Numer. **22**, 1–131 (2013)
4. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. Ann. Stat. **44**(2), 813–852 (2016)
5. Bertsimas, D., Shioda, R.: Algorithm for cardinality-constrained quadratic optimization. Comput. Optim. Appl. **43**(1), 1–22 (2009)
6. Bien, J., Taylor, J., Tibshirani, R.: A lasso for hierarchical interactions. Ann. Stat. **43**(3), 1111–1141 (2013)
7. Bienstock, D.: Computational study of a family of mixed-integer quadratic programming problems. Math. Program. Ser. A **74**(2), 121–140 (1996)
8. Brodie, J., Daubechies, I., De Mol, C., Giannone, D., Loris, I.: Sparse and stable Markowitz portfolios. Proc. Natl. Acad. Sci. **106**(30), 12267–12272 (2009)
9. Burdakov, O.P., Kanzow, C., Schwartz, A.: Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. SIAM J. Optim. **26**(1), 397–425 (2016)
10. Chen, C., Li, X., Tolman, C., Wang, S., Ye, Y.: Sparse portfolio selection via quasi-norm regularization. arXiv:1312.6350v1 (2013)
11. Conforti, M., Cornuejols, G.: A class of logic problems solvable by linear programming. J. ACM **42**(5), 1107–1112 (1995)

12. Conforti, M., Cornuejols, G.: Balanced matrices. In: Aardal, K., Nemhauser, G.L., Weismantel, R. (eds.) Discrete Optimization. Handbooks in Operations Research and Management Science, vol. 12, pp. 277–320. Elsevier, Amsterdam (2005)
13. d'Aspremont, A., Banerjee, O., El Ghaoui, L.: First-order methods for sparse covariance selection. SIAM J. Matrix Anal. Appl. **30**(1), 55–66 (2008)
14. de Miguel, A.-V., Friedlander, M., Nogales, F.J., Scholtes, S.: A two-sided relaxation scheme for mathematical programs with equilibrium constraints. SIAM J. Optim. **16**(2), 587–609 (2006)
15. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
16. Feng, M., Mitchell, J.E., Pang, J.S., Waechter, A., Shen, X.: Complementarity formulations of ℓ_0-norm optimization problems. Pac. J. Optim. Accepted Aug 2016
17. Friedman, J.H., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. Biostatistics **9**(3), 432–441 (2013)
18. Hamada, M., Wu, C.F.J.: Analysis of designed experiments with complex aliasing. J. Qual. Technol. **24**, 130–137 (1992)
19. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press Taylor & Francis Group, Boca Raton (2015)
20. Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high-dimensional models. Stat. Sci. **27**(4), 481–499 (2012)
21. Jacob, L., Obozinski, G., Vert, J.P.: Group lasso with overlap and graph lasso. In: Proceeding of the 26th Annual International Conference on Machine Learning, Montreal, Canada (ICML '09, ACM New York) pp. 433–440 (2009)
22. Kanzow, C., Schwartz, A.: A new regularization method for mathematical programs with complementarity constraints with strong convergence properties. SIAM J. Optim. **23**(2), 770–798 (2013)
23. Le Thi, H.A., Pham, D.T., Vo, X.T.: DC approximation approaches for sparse optimization. Eur. J. Oper. Res. **244**, 26–46 (2015)
24. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman & Hall, London (1983)
25. Nemhauser, G., Wolsey, L.: Integer and Combinatorial Optimization. Wiley, New York (1999)
26. Pang, J.S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. Math. Oper. Res. https://doi.org/10.1287/moor.2016.0795
27. Park, H., Niida, A., Miyano, S., Imoto, S.: Sparse overlapping group Lasso for integrative multi-Omics analysis. J. Comput. Biol. **22**(2), 73–84 (2015)
28. Ralph, D., Wright, S.J.: Some properties of regularization and penalization schemes for MPECs. Optim. Methods Softw. **19**(5), 527–556 (2004)
29. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
30. Scholtes, S.: Convergence properties of a regularisation scheme for mathematical programs with complementarity constraints. SIAM J. Optim. **11**(4), 918–936 (2001)
31. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
32. Tibshirani, R., Saunders, M.A., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**(1), 91–108 (2005)
33. Wang, J., Ye, J.: Multi-layer feature reduction for tree structured group lasso via hierarchical projection. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada pp. 1279–1287 (2015)
34. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methods **68**(1), 49–67 (2006)
35. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
36. Zhang, T.: Analysis of multi-stage convex relaxation for sparse regularization. J. Mach. Learn. Res. **11**, 1081–1107 (2010)
37. Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**(6A), 3468–3497 (2009)
38. Zheng, X., Sun, X., Li, D., Sun, J.: Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear DC approach. Comput. Optim. Appl. **59**(1–2), 379–397 (2014)