SIAM J. OPTIM.
Vol. 27, No. 3, pp. 1637–1665

DIFFERENCE-OF-CONVEX LEARNING: DIRECTIONAL STATIONARITY, OPTIMALITY, AND SPARSITY*

MIJU AHN†, JONG-SHI PANG†, AND JACK XIN‡

Abstract. This paper studies a fundamental bicriteria optimization problem for variable selection in statistical learning; the two criteria are a loss/residual function and a model control (also called regularization, penalty). The former function measures the fitness of the learning model to data and the latter function is employed as a control of the complexity of the model. We focus on the case where the loss function is (strongly) convex and the model control function is a difference-of-convex (dc) sparsity measure. Our paper establishes some fundamental optimality and sparsity properties of directional stationary solutions to a nonconvex Lagrangian formulation of the bicriteria optimization problem, based on a specially structured dc representation of many well-known sparsity functions that can be profitably exploited in the analysis. We relate the Lagrangian optimization problem with the penalty constrained problem in terms of their respective d(irectional)-stationary solutions; this is in contrast to common analysis that pertains to the (global) minimizers of the problem which are not computable due to nonconvexity. Most importantly, we provide sufficient conditions under which the d(irectional)-stationary solutions of the nonconvex Lagrangian formulation are global minimizers (possibly restricted due to nondifferentiability), thereby filling the gap between previous minimizer-based analysis and practical computational considerations. The established relation allows us to readily apply the derived results for the Lagrangian formulation to the penalty constrained formulation. Specializations of the conditions to exact and surrogate sparsity functions are discussed, yielding optimality and sparsity results for existing nonconvex formulations of the statistical learning problem.

Key words. statistical learning, difference-of-convex programs, directional stationary points, optimality, sparsity

AMS subject classifications. 90C26, 65K10, 62B10

DOI. 10.1137/16M1084754

1. Introduction.

Sparse representation [27] is a fundamental methodology of data science in solving a broad range of problems from statistical and machine learning in artificial intelligence to physical sciences and engineering (e.g., imaging and sensing technologies), and to medical decision making (e.g., classification of healthy versus unhealthy patients, benign and cancerous tumors). Significant advances have been made in the last decade on constructing intrinsically low-dimensional solutions in high-dimensional problems via convex programming. In statistical learning, the Least Absolute Shrinkage and Selection Operator (LASSO) [49] is an efficient linear optimization method in regression and variable selection problems based on the minimization of the ℓ_1-norm $\|x\|_1 \triangleq \sum_{i=1}^n |x_i|$ of the n-dimensional model variable x, either subject to a certain prescribed residual constraint, leading to a constrained optimization problem, or employing such a residual as an additional criterion to be minimized, resulting in an unconstrained minimization problem. (Throughout this

*Received by the editors July 14, 2016; accepted for publication (in revised form) May 1, 2017; published electronically August 8, 2017.
http://www.siam.org/journals/siopt/27-3/M108475.html
Funding: The first and second author's work was supported by the National Science Foundation under grants CMMI-1333902 and IIS-1632971 and by the Air Force Office of Scientific Research under grant FA9550-15-1-0126. The third author's work was supported by the National Science Foundation grants DMS-1222507, DMS-1522383, and IIS-1632935.
†Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089 (mijuahn@usc.edu, jongship@usc.edu).
‡Department of Mathematics, University of California, Irvine, CA 92697 (jxin@math.uci.edu).

paper, we address the case of vector optimization and leave the analysis of matrix optimization problems to future work.) Such a convex norm is employed as a surrogate of the ℓ_0-function of x; i.e., $\|x\|_0 \triangleq \sum_{i=1}^n |x_i|_0$, where for a scalar t, $|t|_0 \triangleq \begin{cases} 1 & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases}$ counts the nonzero entries in the model variable x. Theoretical analysis of LASSO and its ability for recovery of the true support of the ground truth vector was pioneered by Candès and Tao [6, 7]. This operator has many good statistical properties such as model selection consistency [56], estimation consistency [3, 29], and persistence property for prediction [26]. There exist many modified versions and algorithms proposed to improve the computational efficiency of LASSO; including adaptive LASSO [58], Bregman Iterative Algorithms [52], the Dantzig Selector [10, 8], iteratively reweighted LASSO [9], elastic net [59], and LARS (Least Angle Regression) [15].

Until now, due to its favorable theoretical underpinnings and many efficient solution methods, convex optimization has been a principal venue for solving many statistical learning problems. Yet there is increasing evidence supporting the use of nonconvex formulations to enhance the realism of the models and improve their generalizations. For instance, in compressed sensing and image science [13], recent findings reported in [34, 51, 54] show that the difference of ℓ_1 and ℓ_2 norms (ℓ_{1-2} for short) and a (nonconvex) "transformed ℓ_1" surrogate outperform ℓ_1 and other known convex penalties in sparse signal recovery when the sensing matrix is highly coherent; such a regime occurs in superresolution imaging where one attempts to recover fine scale structure from coarse scale data information; see [12, 5] and references therein. More broadly, important applications in other areas such as computational statistics, machine learning, bio-informatics, and portfolio selection [28] offer promising sources of problems for the employment of nonconvex functionals to express model loss, promote sparsity, and enhance robustness.

The present paper is motivated by the recent flurry of activities pertaining to the use of nonconvex functionals for statistical learning problems. This idea dates back more than fifteen years ago in statistics when Fan and Li [17] pointed out that LASSO is a biased estimator and postulated several desirable statistics based properties for a good univariate surrogate function for ℓ_0 to have; the authors then introduced a univariate, butterfly shaped, *folded concave* [19], piecewise quadratic function called SCAD for *smoothly clipped absolute deviation* that is symmetric with respect to the origin and concave on $\mathbb{R}_+$; see also [18]. Besides SCAD and ℓ_{1-2}, there exist today many penalized regression methods using nonconvex penalties, including the ℓ_q (for $q \in (0,1)$) penalty [20, 23], the related $\ell_p - \ell_q$ penalty for $p \geq 1$ and $q \in [0,1)$ [11], the combination of ℓ_0 and ℓ_1 penalties [32], the smooth integration of counting and absolute deviation (SICA) [35], the minimax concave penalty (MCP) [53], the capped-ℓ_1 penalty [55], the logarithmic penalty [37], the truncated ℓ_1 penalty [48], the hard thresholding penalty [57], the $\ell_1 - \ell_2$ penalty [51], and the transformed ℓ_1 [40, 35, 54] mentioned above. While there are algorithms for solving several of these nonconvex optimization problems, such as [17, 19, 53, 37, 4], due to the nondifferentiability of the minimands in these nonconvex optimization problems, it is not clear what kind of stationary points these algorithms actually compute. Indeed, without a clear understanding of such points, it is not possible to rigorously ascertain the stationarity, let alone minimizing, properties of the limit points of the iterates produced by the algorithms. This will also help to close the gap between a minimizer-based statistical analysis and a practical computational procedure, relaxing the requirement of optimality to a more realistic stationarity property of the computational model in such an analysis.

Our work focuses on two types of sparsity measures: one type consists of those mentioned above that are surrogates of the ℓ_0-function; the other class of functions is motivated by the ℓ_{1-2}-function; this function has the interesting property that its zeros coincide with the 1-sparse vectors, i.e., $\|x\|_1 - \|x\|_2 = 0$ if and only if x has at most one nonzero component. Generalizing this function, we examine a function whose zeros are the K-sparse vectors for a given positive integer K; these are vectors with at most K nonzero components. While this function has been used to describe the rank of matrices [21, 25, 39], its use to express the sparsity of vectors has not received much attention in the literature. One exception is the reference [25] that has recognized the dc property [44, 45] of the K-sparsity function and employed it to describe cardinality constraints. All these functions are nonconvex and nondifferentiable, making the resulting optimization problems challenging to analyze and solve. A major goal of our study is to address these challenges, providing analysis, algorithms, and numerics to demonstrate the properties and benefits of such nonconvex sparsity functions. The overarching contention is that for the analysis of the resulting optimization problems, it is essential that we focus on solutions that are computable in practice, rather than on solutions that are not provably computable (such as global minimizers that cannot be guaranteed to be the outcome of an algorithm). Thus, our contribution is to provide the theoretical underpinning of what may be termed computable statistical learning, relying on the recent work [42] where numerical algorithms supported by convergence theories have been introduced for computing d-stationary solutions to general nonsmooth dc programs; see also [43]. An accompanying study [1] will report the details of computational experience of such algorithms applied to the problems described in this paper.

With the above background and review of the literature, we are ready to summarize the multifold contributions of our work whose setting is on vector problems. The starting point of these contributions is the introduction of a unified formulation with a piecewise-smooth dc objective function for all the sparsity measures mentioned above. Since the concave summand of several objective functions are of the piecewise smooth, thus nondifferentiable kind, care must be exerted in understanding the stationary solutions of the optimization problems. For this purpose, we will build on the theory in the recent paper [42] that has identified directional stationarity as the least restrictive concept for nonsmooth dc programs with the resulting d-stationary points being computable by practical algorithms. Most importantly, a major contribution of our work is the derivation of several theoretical results on the nonconvex Lagrangian formulation of the bicriteria statistical learning problem, giving conditions under which d-stationary solutions of the Lagrangian optimization problem are its global minima (perhaps restricted due to nondifferentiability), and in some cases, offering error bounds on the deviation for these solutions to an arbitrary vector, including the ground truth to be discovered. Unlike much analysis in the statistics literature, our approach is more in line with computational optimization by focusing on the optimization model that is employed as the principal computational vehicle for discovery. In this vein, our analysis is related to that in [11] which has studied the choice of the weighing factor and derived properties of a global minimizer of the particular "$\ell_2 - \ell_p$ problem"; in contrast, our general results in section 3 are applicable to an arbitrary (strongly) convex loss function and a dc penalty function that covers many known sparsity functions in the literature.

The analysis in the present paper is related to that in two recent articles [31, 33]. There are several differences, however. Perhaps the main one is the respective foci of the cited papers and ours. In [31], where there are results connecting the sparsity

minimization problems formulated using the original ℓ_0 objective and those using its surrogates, the emphasis therein is on the application of the dc algorithm to the latter surrogate formulations, linking them to existing machine learning schemes. In [33], the analysis is more from a statistical perspective while ours is more from a practical optimization perspective. Specifically, the goal in [33] is to gain a deep understanding of how the solution to a regularized optimization problem can "recover" a true minimizer of the expected loss function (the so-called *ground truth*), where the former problem is constructed by minimizing the sum of the loss function plus a "regularization term" and with the constraint that the support of the variables is contained in that of the ground truth. Under extensive assumptions on the parameter choices, the analysis therein is quite detailed and covers many loss and regularization functions known in the literature. Nevertheless, besides the support constraint imposed in the optimization problem, two restrictions were required of the regularizer, i.e., the penalty function: separability and differentiability (except at the origin); these restrictions rule out several important classes of sparsity functions, such as the exact K-sparse functions (see section 5) and the family of piecewise smooth functions such as the capped ℓ_1-function and the piecewise linear approximation of the ℓ_2-function in the ℓ_{1-2}-function. The exact K-sparsity function is nonseparable; both classes of functions have nontrivial manifolds of nondifferentiable points that are potentially the limit points of iterates computed by an optimization procedure. In contrast, our theory is applicable to all these functions. More importantly, we aim to address the following issue, which from a computational perspective is of more practical significance. Namely, given a loss function and a regularizing/penalty function, can one identify suitable weights combining these two criteria so that a d-stationary solution of the resulting nonconvex, nonsmooth optimization problem has some minimizing properties? Once this is done, the latter property then allows further analysis of the solution that includes error bounds from the ground truth in terms of some known statistical estimates. Our analysis is supported by numerical algorithms that can compute such a stationary solution, based on the recent work of [42]. Throughout, the dc property of the functions involved is the backbone of our results. Although this property is not explicitly stated in [33], one of the assumptions imposed therein on the univariate regularizer implies that it is a dc function.

In summary, while there exists previous analysis of the bicriteria optimization problem with sparsity functions, our work contributes to a deeper understanding of the problem, particularly for nondifferentiable sparsity functions. The principal goal of our analysis is to address the optimality and sparsity properties of a directional stationary solution of the problem that can be computed in practice using a dc-based optimization algorithm [42]. Another noteworthy point is that we attempt to deal with an analysis sufficiently broad that is applicable to a host of loss and penalty functions, rather than focusing on specialized problems whose generalizations to other formulations may not be immediate.

2. The unified DC program of statistical learning. We consider the following Lagrangian form of the bicriteria statistical learning optimization problem for selecting the model variable $w \in \mathbb{R}^m$:

$$\underset{w \in \mathcal{W}}{\text{minimize}} \; Z_\gamma(w) \triangleq \ell(w) + \gamma P(w), \tag{1}$$

where γ is a positive scalar parameter balancing the loss (or residual) function $\ell(w)$ and the model control (also called penalty) function $P(w)$ and $\mathcal{W}$ is a closed convex (typically polyhedral) set in $\mathbb{R}^m$ constraining the model variable w. The unconstrained

case where $\mathcal{W} = \mathbb{R}^m$ is of significant interest in many applications. Throughout the paper, we assume that

(A_0) $\ell(w)$ is a convex function on the closed convex set $\mathcal{W}$ and $P(w)$ is a difference-of-convex (dc) function given by

$$P(w) \triangleq g(w) - h(w), \quad \text{with } h(w) \triangleq \max_{1 \le i \le I} h_i(w), \quad \text{for some integer } I > 0, \tag{2}$$

where g is convex but not necessarily differentiable and each h_i is convex and differentiable. ∎

Thus the concave summand of $P(w)$, i.e., $-h(w)$, is the negative of a pointwise maximum of finitely many convex differentiable functions; as such $h(w)$ is piecewise differentiable (PC^1) according to the definition of such a piecewise smooth function [16, Definition 4.5.1]. (Specifically, a continuous function $f(w)$ is PC^1 on an open domain Ω if there exist finite many C^1 functions $\{f_i\}_{i=1}^M$ for some positive integer M such that $f(w) \in \{f_i(w)\}_{i=1}^M$ for all $w \in \Omega$. If each f_i is an affine function, we say that f is *piecewise affine* on Ω.) In many (inexact) surrogate sparsity functions, the function $P(w)$ is separable in its arguments; i.e., $P(w) = \sum_{i=1}^m p_i(w_i)$, where each $p_i(w_i)$ is a univariate dc function whose concave summand is the negative of a pointwise maximum of finitely many convex differentiable (univariate) functions; i.e., the univariate analog of (2).

It is important to note that while it has been recognized in the literature (e.g., [22, 24, 31, 50]) that several classes of sparsity functions can be formulated as dc functions, the particular form (2) of the function $h(w)$ has not been identified in the general dc approach to the sparsity optimization. Our work herein exploits this piecewise structure profitably. Since every convex function can be represented as the pointwise maximum of a family of affine functions, possibly a continuum of them, it is natural to ask if the results in our work can be extended to a general convex function h. A full treatment of this question is regrettably beyond the scope of this paper which is aimed at addressing problems in sparsity optimization and not for general dc programs.

2.1. A model-control constrained formulation. With two criteria, the loss $\ell(w)$ and model control $P(w)$, there are two alternative optimization formulations for the choice of the model variable w; one of these two alternative formulations constrains the latter function while minimizing the former; this is in contrast to balancing the two criteria using a weighing parameter γ into a single combined objective function to be minimized. Specifically, given a prescribed scalar $\beta > 0$, consider

$$\underset{w \in \mathcal{W}}{\text{minimize}}\ \ell(w) \quad \text{subject to } P(w) \le \beta, \tag{3}$$

which we call the model control-constrained version of (1). Since both (1) and (3) are nonconvex problems and involve nondifferentiable functions, the connections between them are not immediately clear. For one thing, from a computational point of view, it is not practical to relate the two problems via their global minima; the reason is simple: such minima are computationally intractable. Instead, one needs to relate these problems via their respective computable solutions. Toward this end, it seems reasonable to relate the d(irectional)-stationary solutions of (1) to the B(ouligand)-stationary solutions of (3) as both kinds of solutions can be computed by the majorization/linearization algorithms described in [42] and they are the "sharpest" kind of stationary solutions for directionally differentiable optimization problems. Before presenting the details of this theory relating the two problems (1) and (3), we note

that a third formulation can be introduced wherein one minimizes the penalty function $P(w)$ while constraining the loss function $\ell(w)$ not to exceed a prescribed tolerance. As the loss function $\ell(w)$ is assumed to be convex in our treatment, the latter formulation is a convex constrained dc program; this is unlike (3) that is a dc constrained dc program, a problem that is considerably more challenging than a convex constrained program. Thus we will omit this loss function constrained formulation and focus only on relating the Lagrangian formulation (1) and the penalty constrained formulation (3).

2.2. Stationary solutions. With $\mathcal{W}$ being a closed convex set, a vector $w^* \in \mathcal{W}$ is formally a d-stationary solution of (1) if the directional derivative

$$Z'_\gamma(w^*; w - w^*) \triangleq \lim_{\tau \downarrow 0} \frac{Z_\gamma(w^* + \tau(w - w^*)) - Z_\gamma(w^*)}{\tau} \geq 0 \quad \forall\, w \in \mathcal{W}.$$

Letting $\widehat{\mathcal{W}}_\beta \triangleq \{w \in \mathcal{W} \mid P(w) \leq \beta\}$ be the (nonconvex) feasible set of (3), we say that a vector $\bar{w} \in \widehat{\mathcal{W}}_\beta$ is a B-stationary solution of this problem if $\ell'(\bar{w}; dw) \geq 0$ for all $dw \in \mathcal{T}(\widehat{\mathcal{W}}_\beta; \bar{w})$, where $\mathcal{T}(\widehat{\mathcal{W}}_\beta; \bar{w})$ is the tangent cone of $\widehat{\mathcal{W}}_\beta$ at $\bar{w}$; the latter cone consists of all vector dw for which there exist a sequence of vectors $\{w^k\} \subset \widehat{\mathcal{W}}_\beta$ converging to $\bar{w}$ and a sequence of positive scalars $\{\tau_k\}$ converging to zero such that $dw = \lim_{k\to\infty} \frac{w^k - \bar{w}}{\tau_k}$. This paper focuses on the derivation of optimality and sparsity properties of directional stationary solutions to the problem (1). Through the connections with the constrained formulation (3) established below, the obtained results for (1) can be adopted to the former problem.

A natural question arises why we choose to focus on directional stationary solutions rather than stationary solutions of other kinds, such as that of a critical point for dc programs [24, 31, 44, 45]. For reasons given in [42, section 3.3] (see, in particular, the summary of implications therein), directional stationary solutions are the sharpest kind among stationary solutions of other kinds in the sense a directional stationary solution must be stationary according to other definitions of stationarity. Moreover, as shown in Proposition 2.1 below and Proposition 3.1 later, d-stationary solutions possess minimizing properties that are not in general satisfied by stationary solutions of other kinds. In essence, the reason for the sharpness of d-stationarity is because it captures all the active pieces as described by the index set $\mathcal{A}(w^*)$ at the point under consideration w^*; see (4). In contrast, a critical point of a dc program fails to capture all the active pieces. Thus any property that a critical point might have may not be applicable to all the active pieces.

Our first result is a characterization of a d-stationary solution of (1).

PROPOSITION 2.1. *Under assumption* (A_0)*, a vector* $w^* \in \mathcal{W}$ *is a d-stationary solution of* (1) *if and only if*

$$w^* \in \underset{w \in \mathcal{W}}{\mathit{argmin}} \left[\underbrace{\ell(w) + \gamma \left\{ g(w) - \nabla h_i(w^*)^T (w - w^*) \right\}}_{\text{convex function in } w} \right] \quad \forall\, i \in \mathcal{A}(w^*), \tag{4}$$

where $\mathcal{A}(w^*) \triangleq \{ i \mid h_i(w^*) = h(w^*) \}$*. Moreover, if the function* h *is piecewise affine, then any such d-stationary solution must be a local minimizer of* Z_γ *on* $\mathcal{W}$*.*

Proof. It suffices to note that $h'(w^*; dw) = \max_{i \in \mathcal{A}(w^*)} \nabla h_i(w^*)^T dw$ for all vectors $dw \in \mathbb{R}^m$. The last assertion about the local minimizing property follows readily

from the fact [16, Exercise 4.8.10] that we have $h(w) = h(w^*) + h'(w^*; w - w^*)$ for any piecewise affine function h, provided that w is sufficiently near w^*. ∎

Remark. The last assertion in the above proposition generalizes a result in [30] where an additional differentiability assumption was assumed. ∎

The characterization of a B-stationary point of (3) is less straightforward, the reason being the nonconvexity of the feasible set $\widehat{\mathcal{W}}_\beta$ that makes it difficult to unveil the structure of the tangent cone $\mathcal{T}(\widehat{\mathcal{W}}_\beta; \bar{w})$. If $P(\bar{w}) < \beta$, then $\mathcal{T}(\widehat{\mathcal{W}}_\beta; \bar{w}) = \mathcal{T}(\mathcal{W}; \bar{w})$. The following result shows that this case is not particularly interesting when it pertains to the stationarity analysis of $\bar{w}$.

PROPOSITION 2.2. *Under assumption* (A_0)*, if $\bar{w} \in \mathcal{W}_\beta$ is a B-stationary point of* (3) *such that $P(\bar{w}) < \beta$, then $\bar{w}$ is a global minimizer of* (3)*.*

Proof. As noted above, the fact that $P(\bar{w}) < \beta$ implies $\mathcal{T}(\widehat{\mathcal{W}}_\beta; \bar{w}) = \mathcal{T}(\mathcal{W}; \bar{w})$; hence $\bar{w}$ satisfies $\ell'(\bar{w}; dw) \geq 0$ for all $dw \in \mathcal{T}(\mathcal{W}; \bar{w})$. Since ℓ is a convex function and $\mathcal{W}$ is a convex set, it follows that $\bar{w}$ is a global minimizer of ℓ on $\mathcal{W}$. In turn, this implies that $\bar{w}$ is a global minimizer of ℓ on $\widehat{\mathcal{W}}_\beta$ because $\widehat{\mathcal{W}}_\beta$ is a subset of $\mathcal{W}$. ∎

To understand the cone $\mathcal{T}(\widehat{\mathcal{W}}_\beta; \bar{w})$ when $P(\bar{w}) = \beta$, we need certain constraint qualifications (CQs), under which it becomes possible to obtain a necessary and sufficient condition for a B-stationary point of (3) similar to that in Proposition 2.1 for a d-stationary point of (1). We list two such CQs as follows, one of which is a pointwise condition while the other one pertains to the entire set $\widehat{\mathcal{W}}_\beta$:

- The *pointwise Slater CQ* is said to hold at $\bar{w} \in \widehat{\mathcal{W}}_\beta$ satisfying $P(\bar{w}) = \beta$ if $\bar{d} \in \mathcal{T}(\mathcal{W}_\beta; \bar{w})$ exists such that $g'(\bar{w}; \bar{d}) < \nabla h_j(\bar{w})^T \bar{d}$ for all $j \in \mathcal{A}(\bar{w})$.
- The *piecewise affine CQ* (PACQ) is said to hold for $\widehat{\mathcal{W}}_\beta$ if $\mathcal{W}$ is a polyhedron and the function g is piecewise affine. (This CQ implies neither the convexity nor the piecewise polyhedrality of $\widehat{\mathcal{W}}_\beta$ because no linearity is assumed for the functions h_i.)

Under these CQs, we have the following characterization of a B-stationary point of the penalty constrained problem (3). See [42] for a proof.

PROPOSITION 2.3. *Under assumption* (A_0)*, if either the pointwise Slater CQ holds at $\bar{w}$ or the PACQ holds for the set $\widehat{\mathcal{W}}_\beta$, then $\bar{w}$ is a B-stationary point of* (3) *if and only if $\ell'(\bar{w}; dw) \geq 0$ for all $dw \in \widehat{C}^j(\bar{w}) \triangleq \{ d \in \mathcal{T}(\mathcal{W}; \bar{w}) \mid g'(\bar{w}; d) \leq \nabla h_j(\bar{w})^T d \}$ and every $j \in \mathcal{A}(w^*)$.* ∎

With the above two propositions, we can formally relate the two problems (1) and (3) as follows.

PROPOSITION 2.4. *Under assumption* (A_0)*, the following two statements hold.*

(a) *If w^* is a d-stationary solution of* (1) *for some $\gamma \geq 0$, and if the pointwise Slater CQ holds at w^* for the set $\widehat{\mathcal{W}}_\beta$ where $\beta \geq P(w^*)$, then w^* is a B-stationary solution of* (3)*.*

(b) *If $\bar{w}$ is a B-stationary solution of* (3) *for some $\beta > 0$, and if the pointwise Slater CQ holds at $\bar{w}$ for the set $\widehat{\mathcal{W}}_\beta$, then the following two statements hold:*

- *for each $j \in \mathcal{A}(\bar{w})$, a scalar $\gamma_j \geq 0$ exists such that $\bar{w}$ is a d-stationary solution of*

$$\underset{w \in \mathcal{W}}{\text{minimize}} \ [\, \ell(w) + \gamma_j \, (\, g(w) - h_j(w) \,) \,]\,; \tag{5}$$

- *nonnegative scalars $\bar{\gamma}$ and $\{\widehat{\gamma}_j\}_{j=1}^I$ exist with $\sum_{j=1}^I \widehat{\gamma}_j = 1$ and $\widehat{\gamma}_j \, (h(w) - h_j(w)) =$*

0 *for all* j *such that* $\bar{w}$ *is a d-stationary solution of*

$$\underset{w\in\mathcal{W}}{\text{minimize}} \left[\ell(w) + \bar{\gamma} \left(g(w) - \sum_{j=1}^{I} \widehat{\gamma}_j \, h_j(w) \right) \right].$$

Finally, statements (a) *and* (b) *remain valid if the PACQ holds instead of the pointwise Slater CQ.*

Proof. If w^* is a d-stationary solution of (1), then

$$\ell'(w^*; w - w^*) + \gamma \left[g'(w^*; w - w^*) - \nabla h_i(w^*)^T (w - w^*) \right] \geq 0$$

for all $w \in \mathcal{W}$ and all $i \in \mathcal{A}(w^*)$. Let $dw \in \widehat{C}^{\,j}(\bar{w})$ for some $j \in \mathcal{A}(w^*)$. We have $dw = \lim_{k\to\infty} \frac{w^k - w^*}{\tau_k}$ for some sequence $\{w^k\} \subset \mathcal{W}$ converging to w^* and sequence $\{\tau_k\} \subset \mathbb{R}_{++}$ converging to zero. For each k, we have

$$\ell'(w^*; w^k - w^*) + \gamma \left[g'(w^*; w^k - w^*) - \nabla h_i(w^*)^T (w^k - w^*) \right] \geq 0.$$

Dividing by τ_k and letting $k \to \infty$, we deduce

$$\ell'(w^*; dw) + \gamma \left[g'(w^*; dw) - \nabla h_i(w^*)^T dw \right] \geq 0,$$

which implies $\ell'(w^*; dw) \geq 0$ because the term in the square bracket is nonpositive. This proves parts (a) and (c).

To prove (b), let $\bar{w}$ be as given. By Proposition 2.3, it follows that for every $j \in \mathcal{A}(\bar{w})$, $\ell'(\bar{w}; dw) \geq 0$ for all $dw \in \widehat{C}^{\,j}(\bar{w}) \triangleq \left\{ d \in \mathcal{T}(\mathcal{W}; \bar{w}) \mid g'(\bar{w}; d) \leq \nabla h_j(\bar{w})^T d \right\}$. Thus $dw = 0$ is a minimizer of $\ell'(\bar{w}; dw)$ for $dw \in \widehat{C}^{\,j}(\bar{w})$. Since the latter set is convex and satisfies the Slater CQ, a standard result in nonlinear programming duality theory (see, e.g., [2]) yields the existence of a scalar γ_j such that $dw = 0$ is a minimizer of $\ell'(\bar{w}; dw) + \gamma_j \left[g'(\bar{w}; dw) - \nabla h_j(\bar{w})^T dw \right]$ on $\mathcal{T}(\mathcal{W}; \bar{w})$. This proves statement (i) in (b). Adding up the inequalities

$$\ell'(\bar{w}; dw) + \gamma_j \left[g'(\bar{w}; dw) - \nabla h_j(\bar{w})^T dw \right] \geq 0, \quad dw \in \mathcal{T}(\mathcal{W}; \bar{w}),$$

for $j \in \mathcal{A}(\bar{w})$, we deduce

$$\ell'(\bar{w}; dw) + \bar{\gamma} \left(g'(\bar{w}; dw) - \sum_{j \in \mathcal{A}(\bar{w})} \widehat{\gamma}_j \, \nabla h_j(\bar{w})^T dw \right) \geq 0, \quad dw \in \mathcal{T}(\mathcal{W}; \bar{w}),$$

where $\bar{\gamma} \triangleq \frac{1}{|\,\mathcal{A}(\bar{w})\,|} \sum_{j \in \mathcal{A}(\bar{w})} \gamma_j$ and $\widehat{\gamma}_j \triangleq \frac{\gamma_j}{\sum_{j' \in \mathcal{A}(\bar{w})} \gamma_{j'}}$. This establishes (b). ∎

Part (b) of Proposition 2.4 falls short in establishing the converse of part (a), i.e., in establishing the full equivalence of the two families of problems $\{(1)\}_{\gamma \geq 0}$ and $\{(3)\}_{\beta \geq 0}$ in terms of their d-stationary solutions and B-stationary solutions, respectively. This adds to the known challenges of the latter dc-constrained family of dc problems over the former convex constrained dc programs. In the next section, we focus on the Lagrangian formulation (1) only and rely on the connections established in this section to obtain parallel results for the constrained formulation (3).

Proposition 2.4 complements the penalty results in dc programming initially studied by [46] and recently expanded in [25, 42]. These previous penalty results address the question of when a fixed constrained problem (3) has an exact penalty formulation

as the problem (1) for sufficient large γ in terms of their global minima. Furthermore, in the case of a quadratic loss function, the reference [25] derives a lower bound for the scalar γ for the penalty formulation to be exact. In contrast, allowing for a general convex loss function, Proposition 2.4 deals with stationary solutions that from a computational perspective are more reasonable as the results pertaining to global minima lack pragmatism and should at best be regarded as providing only theoretical insights and conceptual evidence about the connection between the two formulations.

To close this subsection, we give a result pertaining to the case where the function h is piecewise affine; in this case, by Proposition 2.1, every d-stationary solution of (1) must be a local minimizer. Combining this fact with Proposition 2.4(b), we can establish a similar result for the problem (3).

PROPOSITION 2.5. *Let assumption* (A_0) *hold with each* h_i *being affine additionally. If* $\bar{w}$ *is a B-stationary solution of* (3) *for some* $\beta > 0$*, such that* $P(\bar{w}) = \beta$*, and if either the pointwise Slater CQ holds at* $\bar{w}$ *for the set* $\widehat{\mathcal{W}}_\beta$ *or the PACQ holds for the same set, then* $\bar{w}$ *is a local minimizer of* (3).

Proof. By Proposition 2.4(b) and (c), for each $j \in \mathcal{A}(\bar{w})$, a scalar $\gamma_j \geq 0$ exists such that $\bar{w}$ is a d-stationary solution, thus minimizer, of (5), which is a convex program because h_j is affine. To complete the proof, let $w \in \widehat{\mathcal{W}}_\beta$ be sufficiently close to $\bar{w}$ such that $\mathcal{A}(w) \subseteq \mathcal{A}(\bar{w})$. Let $j \in \mathcal{A}(w)$. We then have

$$\begin{aligned}
\ell(\bar{w}) &= \ell(\bar{w}) + \gamma_j\, P(\bar{w}) - \gamma_j\, \beta \\
&= \ell(\bar{w}) + \gamma_j\, (\, g(\bar{w}) - h_j(\bar{w})\,) - \gamma_j\, \beta \quad \text{because } j \in \mathcal{A}(\bar{w}) \\
&\leq \ell(w) + \gamma_j\, (\, g(w) - h_j(w)\,) - \gamma_j\, \beta \quad \text{by the minimizing property of } \bar{w} \\
&= \ell(w) + \gamma_j\, (\, P(w) - \beta\,) \;\leq\; \ell(w) \qquad \text{because } j \in \mathcal{A}(w) \text{ and } P(w) \leq \beta.
\end{aligned}$$

This establishes that $\bar{w}$ is a local minimizer of (3). ∎

3. Minimizing and sparsity properties. In general, a stationary solution of (1) is not guaranteed to possess any minimizing property. For smooth problems involving twice continuously differentiable functions, it follows from classical nonlinear programming theory that with an appropriate second-order sufficiency condition, a stationary solution can be shown to be strictly locally minimizing. Such a well-known result becomes highly complicated when the functions are nondifferentiable. Although one could, in principle, apply some generalized second derivatives, e.g., those based on Mordukhovich's no-smooth calculus [38], it would be preferable in our context to derive a simplified theory that is more readily applicable to particular sparsity functions, such as those to be discussed in sections 5 and 6.

3.1. Preliminary discussion of the main result. In a nutshell, our goal is to extend the characterization of a d-stationary solution of (1) in Proposition 2.1 by showing that under a set of assumptions, including a specific choice of the convex function g, for a range of values of γ to be identified in the analysis, any such nonzero d-stationary solution w^* either has $\| w^* \|_0$ bounded above by a scalar computable from certain model constants, or

$$w^* \in \operatorname*{argmin}_{w \in \mathcal{W}} \left[\underbrace{\ell(w) + \gamma\, (\, g(w) - h_i(w)\,)}_{\text{remains dc}} \right] \quad \forall\, i \in \mathcal{A}(w^*); \tag{6}$$

see Theorem 3.2 in subsection 3.4 for details. In terms of the original function Z_γ, the above condition implies a *restricted global optimality* property of w^*; namely,

$$w^* \in \operatorname*{argmin}_{w \in \mathcal{W}_*} Z_\gamma(w), \tag{7}$$

where $\mathcal{W}_* \triangleq \{ w \in \mathcal{W} \mid \mathcal{A}(w) \cap \mathcal{A}(w^*) \neq \emptyset \}$. To see this implication, it suffices to note that if $w \in \mathcal{W}_*$, then letting i be a common index in $\mathcal{A}(w) \cap \mathcal{A}(w^*)$, we have

$$Z_\gamma(w) - Z_\gamma(w^*) = [\, \ell(w) + \gamma\, (\, g(w) - h_i(w)\,)\,] - [\, \ell(w^*) + \gamma\, (\, g(w^*) - h_i(w^*)\,)\,] \geq 0.$$

Borrowing a terminology from piecewise programming, we say that the vectors w and w^* *share a common piece* if $\mathcal{A}(w) \cap \mathcal{A}(w^*) \neq \emptyset$. Thus the restricted global optimality property (7) of w^* says that w^* is a true global minimizer of $Z_\gamma(w)$ among those vectors $w \in \mathcal{W}$ that share a common piece with w^*. The subset $\mathcal{W}_*$ includes a neighborhood of w^* in $\mathcal{W}$; i.e., for a suitable scalar $\delta > 0$, $\mathbb{B}(w^*; \delta) \cap \mathcal{W} \subseteq \mathcal{W}_*$, where $\mathbb{B}(w^*; \delta)$ is an Euclidean ball centered at w^* and with radius $\delta > 0$. Thus, the optimality property (6) implies in particular that w^* is a local minimizer of Z_γ on $\mathcal{W}$. Another consequence of (6) is that if $I = 1$ (so that h is convex and differentiable), then w^* is a global minimizer of Z_γ on $\mathcal{W}$.

Besides the above special cases, our main result applies to many well-known sparsity functions in the statistical learning literature, plus the relatively new and largely unexplored class of exact K-sparsity functions to be discussed later.

3.2. The assumptions. The analysis requires two main sets of assumptions: one on the model functions $\ell(w)$ and $\{h_i(w)\}_{i=1}^I$ and the other one on the constraint set $\mathcal{W}$. We first state the former set of assumptions, which introduce the key model constants $\mathrm{Lip}_{\nabla \ell}$, $\lambda_{\min}^\ell$, $\mathrm{Lip}_{\nabla h}$, and ζ. In principle, we could localize these assumptions around a given d-stationary solution; instead, our intention of stating these assumptions globally on the set $\mathcal{W}$ is to use the model constants to identify the parameter γ that defines the optimization problem (1).

(A) In addition to (A_0),

$(\mathrm{A}_\ell^{\mathrm{L}})$ the loss function $\ell(w)$ is of class LC^1 (for continuously differentiable with a Lipschitz gradient) on $\mathcal{W}$; i.e., a positive scalar $\mathrm{Lip}_{\nabla \ell}$ exists such that for all w and w' in $\mathcal{W}$,

$$\| \nabla \ell(w) - \nabla \ell(w') \|_2 \leq \mathrm{Lip}_{\nabla \ell} \| w - w' \|_2 \quad \forall\, w, w' \in \mathcal{W}; \tag{8}$$

$(\mathrm{A}_\ell^{\mathrm{cvx}})$ a nonnegative constant $\lambda_{\min}^\ell$ exists such that

$$\ell(w) - \ell(w') - \nabla \ell(w')^T (\, w - w'\,) \geq \frac{\lambda_{\min}^\ell}{2} \| w - w' \|_2^2 \quad \forall\, w, w' \in \mathcal{W}; \tag{9}$$

$(\mathrm{A}_h^{\mathrm{L}})$ nonnegative constants $\mathrm{Lip}_{\nabla h}$ and β exist such that for each $i \in \{1, \ldots, I\}$ and all w and w' in $\mathcal{W}$, with $0/0$ defined to be zero,

$$0 \leq h_i(w) - h_i(w') - \nabla h_i(w')^T (\, w - w'\,) \leq \left(\frac{\mathrm{Lip}_{\nabla h}}{2} + \frac{\beta}{\| w' \|_2} \right) \| w - w' \|_2^2. \tag{10}$$

$(\mathrm{A}_h^{\mathrm{B}})$ $\max_{1 \leq i \leq I} [\, \sup_{w \in \mathcal{W}} \| \nabla h_i(w) \|_2 \,] \triangleq \zeta < \infty$. ∎

By the mean-value theorem for multivariate functions [41, 3.2.12], we derive from the inequality (8),

$$\ell(w) - \ell(w') - \nabla \ell(w')^T (\, w - w'\,) \leq \frac{\mathrm{Lip}_{\nabla \ell}}{2} \| w - w' \|_2^2 \quad \forall\, w, w' \in \mathcal{W}.$$

Thus $\mathrm{Lip}_{\nabla\ell} \geq \lambda^{\ell}_{\min}$. When $\ell(w) = \frac{1}{2}w^T Q w + q^T w$ is a strongly convex quadratic function with a symmetric positive definite matrix Q, the ratio $\frac{\lambda^{\ell}_{\min}}{\mathrm{Lip}_{\nabla\ell}} = \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}$, where the numerator and denominator of the right-side ratio are the smallest and largest eigenvalues of Q, respectively, is exactly the reciprocal of the condition number of the matrix Q.

We discuss a bit about the two assumptions $(\mathrm{A}^{\mathrm{cvx}}_{\ell})$ and $(\mathrm{A}^{\mathrm{L}}_{h})$. The constant $\lambda^{\ell}_{\min}$ expresses the convexity of the loss function ℓ, and the strong convexity when $\lambda^{\ell}_{\min} > 0$. For certain nonpolyhedral sparsity functions, this constant needs to be positive in order for interesting results to be obtained. One may argue that the latter positivity condition may be too restrictive to be satisfied in applications. For instance, in sparse linear regression where $\ell(w) = \frac{1}{2}\|Aw - b\|_2^2$, the matrix A can be expected to be column rank deficient so that ℓ is only convex but not strongly. For this problem, the results below are applicable to the regularized loss function $\ell_{\alpha}(w) = \ell(w) + \frac{\alpha}{2}\|w\|_2^2$ for any positive scalar α. More generally, any regularized convex function is strongly convex and thus satisfies condition $(\mathrm{A}^{\mathrm{cvx}}_{\ell})$ with a positive $\lambda^{\ell}_{\min}$. Another way to soften the positivity of $\lambda^{\ell}_{\min}$ satisfying (9) is to require that this inequality (with a positive $\lambda^{\ell}_{\min}$) holds only on a subset of $\mathcal{W}$, e.g., on the set $\mathcal{L}_* \cap \mathcal{W}$, where

$$\mathcal{L}_* \triangleq \{\, w \in \mathbb{R}^m \mid w_i = 0 \text{ whenever } w^*_i = 0 \,\}.$$

In this case, the minimizing property of w^* will be restricted to the set $\mathcal{L}_* \cap \mathcal{W}$. It turns out that if the functions h_i are affine so that the function $h(w) = \max_{1\leq i\leq I} h_i(w)$ is piecewise affine (and convex), the main result could hold for $\lambda^{\ell}_{\min} = 0$. Yet another localization of the global inequality (9) is to assume that the Hessian $\nabla^2\ell(w^*)$ exists and is positive definite (either on the entire space $\mathbb{R}^m$ or on the subspace $\mathcal{L}_*$). Under such a pointwise strong convexity condition (via the positive definiteness of the Hessian matrix), one obtains a locally strictly minimizing property of w^*. The upshot of this discussion is that, in general, if one desires a strong minimizing property to hold on the entire set $\mathcal{W}$, then the inequality (9) with a positive $\lambda^{\ell}_{\min}$ is essential to compensate for the nonconvexity of the sparsity function P when h_i is nonaffine; relaxation of the condition (9) with a positive $\lambda^{\ell}_{\min}$ is possible either with a piecewise linear function h or at the expense of weakening the minimizing conclusion when the functions h_i are not affine.

As we will see in the subsequent sessions, assumption $(\mathrm{A}^{\mathrm{L}}_{h})$ accommodates a host of convex functions h_i. Foremost is the case when the functions h_i are of class LC^1; in this case, we may take $\beta = 0$ and $\mathrm{Lip}_{\nabla h}$ to be the largest of the Lipschitz moduli of the gradients $\{\nabla h_i(w)\}_{i=1}^I$ on $\mathcal{W}$. In particular, when each function h_i is affine, we may further take $\mathrm{Lip}_{\nabla h}$ to be zero or any positive scalar. Assumption $(\mathrm{A}^{\mathrm{L}}_{h})$ also includes the case where $I = 1$ and $h(w) = \|w\|_2$. Indeed, this follows from the following inequality: provided $w^* \neq 0$,

$$\|w\|_2 - \|w^*\|_2 - \left(\frac{w^*}{\|w^*\|_2}\right)^T (w - w^*) \leq \frac{1}{\|w^*\|_2}\|w - w^*\|_2^2,$$

which is equivalent to

$$\|w\|_2\|w^*\|_2 - \|w^*\|_2^2 - (w^*)^T(w - w^*) \leq \|w - w^*\|_2^2.$$

The left-hand side is equal to $\|w\|_2\|w^*\|_2 - (w^*)^T w$. We have

$$\begin{aligned}
&\| w - w^* \|_2^2 - \left[\| w \|_2 \| w^* \|_2 - (w^*)^T w \right] \\
&\quad = \| w \|_2^2 - (w^*)^T w + \| w^* \|_2^2 - \| w \|_2 \| w^* \|_2 \\
&\quad \geq \| w \|_2^2 - 2 \| w^* \|_2 \| w \|_2 + \| w^* \|_2^2 = \left[\| w \|_2 - \| w^* \|_2 \right]^2 \geq 0,
\end{aligned}$$

which shows that we may take $\mathrm{Lip}_{\nabla h} = 0$ and $\beta = 1$ for the 2-norm function. It is important to remark that, in general, if (10) holds with a zero $\mathrm{Lip}_{\nabla h}$, it certainly holds with a positive $\mathrm{Lip}_{\nabla h}$. Nevertheless, we refrain from taking this constant as always positive because it affects the selection of the constant γ as it will become clear in the main theorem, Theorem 3.2. Finally, $(\mathrm{A}_h^{\mathrm{L}})$ also accommodates any linear combination of functions each satisfying this assumption. Such a combination may be relevant in problems of group sparsity minimization where one may consider the use of different sparsity functions for different groups.

The second assumption (B) below is on the constraint set $\mathcal{W}$. Consistent with the previous assumptions, we state assumption (B) with respect to an *arbitrary* vector $w^* \in \mathcal{W}$, making it a global-like condition; however, it will be clear from the analysis that the assumption is most specific to a stationary solution being analyzed. The assumption is satisfied for example if $\mathcal{W} = \prod_{i=1}^m \mathcal{W}_i$ where each $\mathcal{W}_i$ is one of three sets: $\mathbb{R}$ (for an unrestricted variable) or $\mathbb{R}_\pm$ (for problems with sign restrictions on the variables w_i). Thus our results are applicable in particular to sign constrained problems, a departure from the literature which typically either deals with unconstrained problems, or makes an interiority assumption on w^* which essentially converts the analysis to the unconstrained case.

(B) For a given vector $w^* \in \mathcal{W}$, there exists $\varepsilon_* > 0$ such that for all $\varepsilon \in (0, \varepsilon_*]$ and for every i with $w_i^* \neq 0$, the vectors $w^{\pm\varepsilon;i}$ defined below belong to $\mathcal{W}$:

$$w_j^{\pm\varepsilon;i} \triangleq \begin{cases} w_j^* \pm \varepsilon & \text{if } j = i, \\ w_j^* & \text{if } j \neq i, \end{cases} \qquad j = 1, \ldots, m.$$

While our analysis is applicable to a general convex function g, a more interesting case is when g is a weighted ℓ_1-function given by

$$g(w) \triangleq \sum_{i=1}^m \xi_i \, | \, w_i \, |, \quad w \in \mathbb{R}^m, \tag{11}$$

for some positive constants ξ_i. This choice of the function g was employed in the study [31] of DC approximations to sparsity minimization; Proposition 6 in this reference shows that a broad class of separable sparsity functions in the statistical learning literature has a dc decomposition with the above function g as the convex part. For this choice of g, if $\eta \in \partial g(w)$ with $\partial g(w)$ denoting the subdifferential of g at w, then

$$\| \eta_{\neq 0} \|_2 = \sqrt{\sum_{i \, : \, w_i \neq 0} \xi_i^2} \geq \left(\min_{1 \leq i \leq m} \xi_i \right) \sqrt{\| w \|_0},$$

where $\eta_{\neq 0}$ denotes the subvector of η with components corresponding to the nonzero components of w. Let $\xi_{\min} \triangleq \min_{1 \leq i \leq m} \xi_i$ so that $\| \eta_{\neq 0} \|_2 \geq \xi_{\min} \sqrt{\| w \|_0}$ for all $w \in \mathbb{R}^m$.

3.3. The case $\beta = 0$. We begin our analysis with a result pertaining to the case where condition $(\mathrm{A}_h^{\mathrm{L}})$ holds with $\beta = 0$. This is the case where each h_i is of class LC^1. The special property (B) on $\mathcal{W}$ is not needed in this case; also, no assumption is imposed on g except for its convexity.

PROPOSITION 3.1. *Suppose that* $(\mathrm{A}_\ell^{\mathrm{cvx}})$ *and* $(\mathrm{A}_h^{\mathrm{L}})$ *with* $\beta = 0$ *hold. For any scalar* $\gamma > 0$ *such that* $\delta \triangleq \lambda_{\min}^{\ell} - \gamma\, Lip_{\nabla h} \geq 0$*, the following two statements hold:*

(a) *If* w^* *is a d-stationary point of* Z_γ *on* $\mathcal{W}$*, then for all* $i \in \mathcal{A}(w^*)$*,*

$$[\,\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)\,] - [\,\ell(w^*) + \gamma\,(\,g(w^*) - h_i(w^*)\,)\,] \;\geq\; \frac{\delta}{2}\,\|\,w - w^*\,\|_2^2 \;\;\forall\, w \in \mathcal{W}.$$

Hence w^* *is a minimizer of* $Z_\gamma(w)$ *on* $\mathcal{W}_*$*. Moreover, if* $\delta > 0$*, then* w^* *is the unique minimizer of* $\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)$ *on* $\mathcal{W}$ *for all* $i \in \mathcal{A}(w^*)$*. Hence the unique minimizer of* $Z_\gamma(w)$ *on* $\mathcal{W}_*$*.*

(b) *If* $I = 1$*, then* Z_γ *is convex on* $\mathcal{W}$ *with modulus* δ *(thus is strongly convex if* $\delta > 0$*).*

Proof. To prove (a), let $w \in \mathcal{W}$ be arbitrary. For any $i \in \mathcal{A}(w^*)$, we have

$$\begin{aligned}
&[\,\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)\,] - [\,\ell(w^*) + \gamma\,(\,g(w^*) - h_i(w^*)\,)\,] \\
&\quad\geq\; [\,\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)\,] - [\,\ell(w^*) + \gamma\,(\,g(w^*) - h_i(w^*)\,)\,] \\
&\qquad\quad - \big[\,\nabla\ell(w^*)^T(\,w - w^*\,) + \gamma\,\big(\,g'(w^*; w - w^*) - \nabla h_i(w^*)^T(\,w - w^*\,)\,\big)\,\big] \\
&=\; [\,\ell(w) - \ell(w^*) - \nabla\ell(w^*)(\,w - w^*\,)\,] + \gamma\,[\,g(w) - g(w^*) - g'(w^*; w - w^*)\,] \\
&\qquad\quad - \gamma\,\big[\,h_i(w) - h_i(w^*) - \nabla h_i(w^*)^T(\,w - w^*\,)\,\big] \\
&\quad\geq\; \tfrac{1}{2}\,\big(\,\lambda_{\min}^{\ell} - \gamma\,\mathrm{Lip}_{\nabla h}\,\big)\,\|\,w - w^*\,\|_2^2.
\end{aligned}$$

Thus (a) follows. To prove (b), suppose $I = 1$. We have, for any w and w' in $\mathcal{W}$,

$$Z_\gamma(w) - Z_\gamma(w') - Z_\gamma'(w'; w - w') \;\geq\; \tfrac{1}{2}\,\big(\,\lambda_{\min}^{\ell} - \gamma\,\mathrm{Lip}_{\nabla h}\,\big)\,\|\,w - w'\,\|_2^2.$$

Statement (b) follows readily from the above inequality. ∎

Statement (b) of the above result is, in general, not true if $I > 1$ as illustrated by the univariate function $Z_\gamma(t) = t^2 - \gamma|t| = t^2 - \gamma\max(t, -t)$. (This univariate function fits our framework with $g = |t|$ and $h = 2|t|$.) It can be seen that for any $\gamma > 0$ this function is not convex because

$$Z_\gamma(0) \;=\; 0 \;>\; -\frac{\gamma^2}{4} \;=\; \tfrac{1}{2}\,[\,Z_\gamma(\gamma/2) + Z_\gamma(-\gamma/2)\,].$$

Nevertheless, this function has two stationary points at $\pm\gamma/2$ that are both global minima of Z_γ. This simple result illustrates a couple noteworthy points. First, the convexity of the function Z_γ cannot be expected for $I > 1$ even though the loss function ℓ may be strongly convex. Yet it is possible for a d-stationary solution of Z_γ to be a global minimizer. The latter possibility is encouraging and serves as the motivation for the subsequent extended analysis with $\beta > 0$, where an upper bound on γ will persist.

3.4. The case of g being a weighted ℓ_1 function. In this subsection, we refine the proof of Proposition 3.1 by taking g to be the weighted ℓ_1 function (11) and

including the constant β. We have, for all $w \in \mathcal{W}$ and all $i \in \mathcal{A}(w^*)$,

$$\begin{aligned}
&[\,\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)\,] - [\,\ell(w^*) + \gamma\,(\,g(w^*) - h_i(w^*)\,)\,] \\
&\quad \geq [\,\ell(w) - \ell(w^*) - \nabla\ell(w^*)(\,w - w^*\,)\,] + \gamma\,[\,g(w) - g(w^*) - g'(w^*; w - w^*)\,] \\
&\qquad - \gamma\,\left[\,h_i(w) - h_i(w^*) - \nabla h_i(w^*)^T(\,w - w^*\,)\,\right] \\
&\quad \geq \left[\,\frac{\lambda^\ell_{\min}}{2} - \gamma\,\left(\,\frac{\mathrm{Lip}_{\nabla h}}{2} + \frac{\beta}{\|\,w^*\|_2}\,\right)\right]\,\|\,w - w^*\,\|_2^2.
\end{aligned}$$

The remaining analysis uses the d-stationarity of w^* to establish the nonnegativity of the multiplicative factor $M_\gamma \triangleq \frac{\lambda^\ell_{\min}}{2} - \gamma\,\big(\,\frac{\mathrm{Lip}_{\nabla h}}{2} + \frac{\beta}{\|\,w^*\|_2}\,\big)$. By the characterization of d-stationarity in Proposition 2.1, there exists a subgradient $\eta^{\,i} \in \partial g(w^*)$ such that

$$\left[\,\nabla\ell(w^*) + \gamma\,\left(\,\eta^{\,i} - \nabla h_i(w^*)\,\right)\,\right]^T(\,w - w^*\,) \,\geq\, 0 \quad \forall\, w \,\in\, \mathcal{W}.$$

By assumption (B) on the constraint set $\mathcal{W}$, we may substitute $w = w^{\pm\varepsilon;k}$ for all k such that $w^*_k \neq 0$ and for arbitrary $\varepsilon \in (0, \varepsilon_*]$ and obtain

$$\left[\,\nabla\ell(w^*) + \gamma\,\left(\,\eta^{\,i} - \nabla h_i(w^*)\,\right)\,\right]_k \,=\, 0 \quad \forall\, k \text{ such that } w^*_k \,\neq\, 0. \tag{12}$$

Thus letting $\nabla_{\neq 0}\, h_i(w)$ be the vector $\big(\frac{\partial h_i(w)}{\partial w_k}\big)_{k:w_k\neq 0}$, we have

$$\begin{aligned}
\|\,w^*\,\|_2\,\mathrm{Lip}_{\nabla\ell} + \|\,\nabla\ell(0)\,\|_2 \,\geq\, \|\,\nabla\ell(w^*)\,\|_2 \quad &= \quad \gamma\,\|\,\eta_{\neq 0} - \nabla_{\neq 0} h_i(w^*)\,\|_2 \\
&\geq \quad \gamma\,[\,\|\,\eta_{\neq 0}\,\|_2 - \|\,\nabla_{\neq 0} h_i(w^*)\,\|_2\,],
\end{aligned}$$

where the first inequality is by the LC^1 property of ℓ and the triangle inequality which also yields the last inequality. Suppose $\gamma \geq c\,\|\,\nabla\ell(0)\,\|_2$ for a constant $c > 0$ to be determined momentarily. We deduce

$$\|\,w^*\,\|_2\,\mathrm{Lip}_{\nabla\ell} \,\geq\, \gamma\,\left[\,\|\,\eta_{\neq 0}\,\|_2 - \|\,\nabla_{\neq 0} h_i(w^*)\,\|_2 - c^{-1}\,\right].$$

With g being the weighted ℓ_1-function (11) and by recalling the scalar ζ in condition $(\mathrm{A}^{\mathrm{B}}_h)$, we deduce

$$\|\,w^*\,\|_2\,\mathrm{Lip}_{\nabla\ell} \,\geq\, \gamma\,\left[\,\xi_{\min}\,\sqrt{\|\,w^*\,\|_0} - \zeta - c^{-1}\,\right]. \tag{13}$$

Moreover, provided that $\xi_{\min}\,\sqrt{\|\,w^*\,\|_0} > \zeta + c^{-1}$, we obtain

$$\frac{\gamma}{\|\,w^*\,\|_2} \,\leq\, \frac{\mathrm{Lip}_{\nabla\ell}}{\xi_{\min}\,\sqrt{\|\,w^*\,\|_0} - \zeta - c^{-1}}.$$

Hence, if $\sqrt{\|\,w^*\,\|_0} > \frac{2}{\xi_{\min}}\,\left[\,\zeta + c^{-1}\,\right]$, (the factor 2 can be replaced by any constant greater than 1), then $\frac{\gamma}{\|\,w^*\,\|_2} < \frac{\mathrm{Lip}_{\nabla\ell}}{\zeta + c^{-1}}$. Thus $M_\gamma \geq \frac{1}{2}\,\big(\lambda^\ell_{\min} - \gamma\mathrm{Lip}_{\nabla h}\big) - \frac{\beta\,\mathrm{Lip}_{\nabla\ell}}{\zeta + c^{-1}}$. Consequently, if

$$\delta_\gamma(c) \,\triangleq\, \tfrac{1}{2}\,\left(\,\lambda^\ell_{\min} - \gamma\,\mathrm{Lip}_{\nabla h}\,\right) - \frac{\beta\,\mathrm{Lip}_{\nabla\ell}}{\zeta + c^{-1}} \,\geq\, 0, \tag{14}$$

it follows that $M_\gamma \geq 0$. It remains to determine the constant c. To ensure the validity of the above derivations, we need to have

$$\tfrac{1}{2}\,c\,\|\,\nabla\ell(0)\,\|_2\,\mathrm{Lip}_{\nabla h} \,\leq\, \frac{\lambda^\ell_{\min}}{2} - \frac{\beta\,\mathrm{Lip}_{\nabla\ell}}{\zeta + c^{-1}}, \tag{15}$$

which is equivalent to

$$q_*(c) \triangleq c^2\,\zeta\,\|\,\nabla\ell(0)\,\|_2\,\mathrm{Lip}_{\nabla h} + c\,\left[\,\|\,\nabla\ell(0)\,\|_2\,\mathrm{Lip}_{\nabla h} + 2\beta\,\mathrm{Lip}_{\nabla\ell} - \lambda^\ell_{\min}\,\zeta\,\right] - \lambda^\ell_{\min} \le 0.$$

Summarizing the above derivations, we have established the following main result which does not require further proof.

THEOREM 3.2. *Under the assumptions in subsection* 3.2 *with the constants* $\lambda^\ell_{\min}$, $Lip_{\nabla\ell}$, $Lip_{\nabla h}$, β, ζ, *and* $\xi_{\min}$ *as given therein, let* g *be given by* (11), *and let* $c > 0$ *be any constant satisfying* $q_*(c) \le 0$. *Let* $\gamma > 0$ *satisfy*

$$c\,\|\,\nabla\ell(0)\,\|_2\,Lip_{\nabla h} \le \gamma\,Lip_{\nabla h} \le \lambda^\ell_{\min} - \frac{2\,\beta\,Lip_{\nabla\ell}}{\zeta + c^{-1}}\,. \tag{16}$$

If w^* *is a d-stationary solution of* Z_γ *on* $\mathcal{W}$, *then either*

$$\sqrt{\|\,w^*\,\|_0} \le \frac{2}{\xi_{\min}}\,\left[\,\zeta + c^{-1}\,\right], \tag{17}$$

or for all $i \in \mathcal{A}(w^*)$ *and all* $w \in \mathcal{W}$,

$$[\,\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)\,] - [\,\ell(w^*) + \gamma\,(\,g(w^*) - h_i(w^*)\,)\,] \ge \delta_\gamma(c)\,\|\,w - w^*\,\|_2^2.$$

If the above inequality holds, then w^* *is a minimizer of* $Z_\gamma(w)$ *on* $\mathcal{W}_*$. *Moreover, if* $\delta_\gamma(c) > 0$, *then* w^* *is the unique minimizer of* $\ell(w) + \gamma\,(\,g(w) - h_i(w)\,)$ *on* $\mathcal{W}$ *for all* $i \in \mathcal{A}(w^*)$, *hence the unique minimizer of* $Z_\gamma(w)$ *on* $\mathcal{W}_*$. ∎

Remark. As shown above, $q_*(c) \le 0$ if and only if (15) holds. With the latter inequality, we can choose $\gamma > 0$ so that (16) holds. In turn, the inequalities in (16) implicitly impose a condition on the four model constants: $\lambda^\ell_{\min}$, $\mathrm{Lip}_{\nabla\ell}$, $\mathrm{Lip}_{\nabla h}$, and β and suggest a choice of γ in terms of them for the theorem to hold. ∎

As a (restricted) global minimizer of Z_γ, w^* has the property that for every $w \in \mathcal{W}_*$, either $\ell(w^*) \le \ell(w)$ or $P(w^*) < P(w)$. In particular, if w^* is not a global minimizer of the loss function ℓ on $\mathcal{W}_*$, then we must have $P(w^*) < P(\widetilde{w})$ where $\widetilde{w}$ is the latter minimizer. In the language of multicriteria optimization such a vector w^* is a *Pareto point* on $\mathcal{W}_*$ of the two criteria: loss and model control; specifically, there does not exist a $w \in \mathcal{W}_*$ such that the pair $(\ell(w), P(w)) \lneqq (\ell(w^*), P(w^*))$.

The implication of the (restricted) global minimizing property of a d-stationary point is reminiscent of the class of "pseudoconvex functions" [36] which, by their definition, are differentiable functions with the property that all its stationary points are global minimizers of the functions. There are two differences, however. One difference is that the function Z_γ is only directionally differentiable and its stationary solutions are defined in terms of the directional derivatives. Another difference is that when $\beta > 0$, a (non)sparsity stipulation (i.e., the failure of (17)) on the d-stationary solution w^* is needed for it to be a (restricted) global minimizer.

The following corollary of Theorem 3.2 is worth mentioning. For simplicity, we state the corollary in terms of Z_γ on the subset $\mathcal{W}_*$. No proof is required.

COROLLARY 3.3. *Under the assumptions of Theorem* 3.2, *suppose that* $Lip_{\nabla h} = 0$. *Let* g *be given by* (11). *Let* $c > 0$ *satisfy* $c\left[\,2\,\beta\,Lip_{\nabla\ell} - \lambda^\ell_{\min}\,\zeta\,\right] \le \lambda^\ell_{\min}$. *For any* $\gamma \ge c\,\|\,\nabla\ell(0)\,\|_2$, *if* w^* *is a d-stationary solution of* Z_γ *on* $\mathcal{W}$, *then either*

$$Z_\gamma(w) - Z_\gamma(w^*) \ge \left[\,\frac{\lambda^\ell_{\min}}{2} - \frac{\beta\,Lip_{\nabla\ell}}{\zeta + c^{-1}}\,\right]\,\|\,w - w^*\,\|_2^2 \quad \forall\, w \in \mathcal{W}_*, \tag{18}$$

or $\sqrt{\|\,w^*\,\|_0} \le \frac{2\,(\,\zeta + c^{-1})}{\xi_{\min}}$. ∎

We close this section by giving a bound on $\|w^*\|_0$ based on the inequality (13). The validity of this bound does not require the positivity of $\lambda^\ell_{\min}$; nevertheless, the sparsity bound preassumes a bound on $\|w^*\|_2$; this is in contrast to Theorem 3.2 which makes no assumption on w^* except for its d-stationarity.

PROPOSITION 3.4. *Suppose that assumptions* $(\mathrm{A}^{\mathrm{cvx}}_\ell)$, $(\mathrm{A}^{\mathrm{L}}_\ell)$, *and* (B) *hold. Let* g *be given by* (11). *For every* $\gamma > c\|\nabla\ell(0)\|_2$ *for some scalar* $c > 0$, *if* w^* *is a d-stationary solution of* Z_γ *on* $\mathcal{W}$ *such that* $\|w^*\|_2 \leq \frac{c\|\nabla\ell(0)\|_2}{Lip_{\nabla\ell}}\left[\xi_{\min}\sqrt{L+1} - \zeta - c^{-1}\right]$ *for some integer* $L > 0$ *then* $\|w^*\|_0 \leq L$.

Proof. Assume for contradiction that $\|w^*\|_0 \geq L+1$. Then $w^* \neq 0$; hence $\xi_{\min}\sqrt{L+1} > \zeta + c^{-1}$. We have

$$\begin{aligned}
& c\,\|\nabla\ell(0)\|_2\left[\xi_{\min}\sqrt{L+1} - \zeta - c^{-1}\right] \\
& < \gamma\left[\xi_{\min}\sqrt{L+1} - \zeta - c^{-1}\right] && \text{by the choice of } \gamma \\
& \leq \gamma\left[\xi_{\min}\|w^*\|_0 - \zeta - c^{-1}\right] && \text{by assumption on } \|w^*\|_0 \\
& \leq \|w^*\|_2\,\mathrm{Lip}_{\nabla\ell} && \text{by (13)} \\
& \leq c\,\|\nabla\ell(0)\|_2\left[\xi_{\min}\sqrt{L+1} - \zeta - c^{-1}\right] && \text{by assumption on } \|w^*\|_2.
\end{aligned}$$

This contradiction establishes the desired bound on $\|w^*\|_0$. ∎

4. Sparsity functions and their dc representations. In this and the next section, we investigate the application of the results in section 3 to a host of nonconvex sparsity functions that have been studied extensively in the literature. These nonconvex functions are deviations from the convex ℓ_1-norm that is a well-known convex surrogate for the ℓ_0-function. As mentioned in the introduction, we classify these sparsity functions into two categories: the exact ones and the surrogate ones. The next section discusses the class of exact sparsity functions; section 6 discusses the surrogate sparsity functions. In each case, we identify the constants $\mathrm{Lip}_{\nabla h}$, β, ζ, and $\xi_{\min}$ of the sparsity functions and present the specialization of the results in the last section to these functions. For convenience, we summarize in Table 1 the sparsity functions being analyzed. The results obtained in the next two sections are summarized in Tables 2 and 3 appended.

5. Exact sparsity functions. While the class of exact K-sparsity functions has been discussed exclusively in the context of matrix optimization, for the sake of completeness we formally define these functions for vectors and highlight their properties most relevant to our discussion. For an m-vector $w = (w_i)_{i=1}^m$, let $[|w|] \triangleq (w_{[i]})_{i=1}^m$ be derived from w by ranking the absolute values of its components in nonincreasing order: $\max_{1\leq i\leq m}|w_i| \triangleq |w_{[1]}| \geq |w_{[2]}| \geq \cdots \geq |w_{[m]}| \triangleq \min_{1\leq i\leq m}|w_i|$; thus $|w_{[k]}|$ is the kth largest of the m components of w in absolute values and $\{[1],\ldots,[m]\}$ is an arrangement of the index set $\{1,\ldots,m\}$. For a fixed positive integer K, let

$$P_{[K]}(w) \triangleq \underbrace{\|w\|_1}_{g_{[K]}(w)} - \underbrace{\sum_{k=1}^{K}|w_{[k]}|}_{h_{[K]}(w)} = \sum_{k=K+1}^{m}|w_{[k]}|.$$

Clearly, $P_{[K]}(w) = 0$ if and only if w has no more than K nonzero components. The convexity of $h_{[K]}$, thus the dc-property of $P_{[K]}$, follows from the value-function

TABLE 1
Exact and surrogate penalty functions and their properties.

Penalty	Difference of convex representation
K-sparsity	• $\|\| w \|\|_1 - \sum_{k=1}^{K} \| w_{[k]} \|$, • $\beta = \mathrm{Lip}_{\nabla h} = 0$
$\ell_1 - \ell_2$	• $\|\| w \|\|_1 - \|\| w \|\|_2$, • $\beta = 1$; $\mathrm{Lip}_{\nabla h} = 0$
SCAD	• $\sum_{i=1}^{m} \lambda_i \| w_i \| - \sum_{i=1}^{m} \begin{cases} 0 & \text{if } \| w_i \| \le \lambda, \\ \frac{(\| w_i \| - \lambda)^2}{2(a-1)} & \text{if } \lambda \le \| w_i \| \le a\lambda, \\ \lambda \| w_i \| - \frac{(a+1)\lambda^2}{2} & \text{if } \| w_i \| \ge a\lambda, \end{cases}$ • $\beta = 0$; $\mathrm{Lip}_{\nabla h} = 1$
Capped ℓ_1	• $\sum_{i=1}^{m} \frac{\|w_i\|}{a_i} - \sum_{i=1}^{m} \max\left\{0, \frac{w_i}{a_i} - 1, -\frac{w_i}{a_i} - 1\right\}$, • $\beta = \mathrm{Lip}_{\nabla h} = 0$
Transformed ℓ_1	• $\sum_{i=1}^{m} \frac{a_i+1}{a_i}\|w_i\| - \sum_{i=1}^{m} \left[\frac{a_i+1}{a_i}\|w_i\| - \frac{(a_i+1)\|w_i\|}{a_i+\|w_i\|}\right]$, • $\beta = 0$; $\mathrm{Lip}_{\nabla h} = 2 \max_{1\le i\le m} \frac{a_i+1}{a_i^2}$
Logarithmic	• $\sum_{i=1}^{m} \frac{\lambda_i}{\varepsilon_i}\|w_i\| - \sum_{i=1}^{m} \lambda_i \left[\frac{\|w_i\|}{\varepsilon_i} - \log(\|w_i\| + \varepsilon_i) + \log \varepsilon_i\right]$, • $\beta = 0$; $\mathrm{Lip}_{\nabla h} = \max_{1\le i\le m} \frac{\lambda_i}{\varepsilon_i}$

TABLE 2
Summary of general results. w^* is a d-stationary solution.

Result	Condition on constants	Conclusions
Thm. 3.2	• $c\|\|\nabla\ell(0)\|\|_2 \mathrm{Lip}_{\nabla h}$ $\le \gamma \mathrm{Lip}_{\nabla h} \le \lambda^{\ell}_{\min} - \frac{2\beta \mathrm{Lip}_{\nabla \ell}}{\zeta + c^{-1}}$, • $q_*(c) \le 0$	• Either $\sqrt{\|\|w^*\|\|_0} \le \frac{2}{\xi_{\min}} \left[\zeta + c^{-1}\right]$, • or w^* is a minimizer on $\mathcal{W}_*$, unique if $\gamma \mathrm{Lip}_{\nabla h} < \lambda^{\ell}_{\min} - \frac{2\beta \mathrm{Lip}_{\nabla \ell}}{\zeta + c^{-1}}$
Prop. 3.4	$c\|\|\nabla\ell(0)\|\|_2 < \gamma$	$\| w^* \|_2 \le \frac{c\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}} \left[\xi_{\min}\sqrt{L+1} - \zeta - c^{-1}\right]$ $\Rightarrow \ \|w^*\|_0 \le L.$

representation below:

$$
\begin{aligned}
h_{[K]}(w) \quad &\triangleq \quad \sum_{k=1}^{K} \| w_{[k]} \| \ = \ \max_{v \in \Delta_{K,m}} \sum_{i=1}^{m} v_i \| w_i \|, \\
\text{where} \quad \Delta_{K,m} \quad &\triangleq \quad \left\{ v \in [0,1]^m \ \mid \ \sum_{i=1}^{m} v_i \ = \ K \right\}.
\end{aligned} \tag{19}
$$

Since $h_{[K]}$ is piecewise linear, it is LC^1 with $\mathrm{Lip}_{\nabla h} = 0$; moreover, we have $\beta = 0$. In order to calculate the constant ζ associated with the function $h_{[K]}$, it would be useful to express $h_{[K]}(w)$ as the pointwise maximum of finitely many linear functions. For this purpose, let $\mathcal{E}(\Delta_{K,m})$ be the finite set of extreme points of the polytope $\Delta_{K,m}$.

TABLE 3

Summary of specialized results. $q_*(c) \triangleq c^2\,\zeta\,\|\,\nabla\ell(0)\,\|_2\,Lip_{\nabla h} + c\left[\,\|\,\nabla\ell(0)\,\|_2\,Lip_{\nabla h} + 2\beta\,Lip_{\nabla\ell} - \lambda^\ell_{\min}\,\zeta\,\right] - \lambda^\ell_{\min}$.

Result	Condition on constants	Conclusions
Thm. 5.1 (K-sparsity)		w^* is a minimizer on $\mathcal{W}_*$
Thm. 5.1 (K-sparsity)	• $c\|\|\nabla\ell(0)\|\|_2 < \gamma$, • $\sqrt{K+1} > \sqrt{K} + c^{-1}$	$\|w^*\|_2 \le \frac{c\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}}\left[\sqrt{K+1} - \sqrt{K} - c^{-1}\right]$ $\Rightarrow\ \|w^*\|_0 \le K$
Thm. 5.2 ($\ell_1 - \ell_2$)	• $c\|\|\nabla\ell(0)\|\|_2 < \gamma$, • $c \le \frac{\lambda^\ell_{\min}}{2\mathrm{Lip}_{\nabla\ell} - \lambda^\ell_{\min}}$	• Either $\sqrt{\|w^*\|_0} \le 2(1 + c^{-1})$, • or w^* is a global minimizer on $\mathcal{W}$
Thm. 6.1 (SCAD)	$\gamma \le \frac{1}{2}\lambda^\ell_{\min}$	w^* is a global minimizer on $\mathcal{W}$
Thm. 6.1 (SCAD)	$c\|\nabla\ell(0)\|_2 < \gamma \le \frac{1}{2}\lambda^\ell_{\min}$	$\|w^*\|_2 \le \frac{c\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}}\left[\min_{1\le j\le m}\lambda_j - c^{-1}\right]$ $\Rightarrow \sqrt{\|w^*\|_0} \le \frac{\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}}\left[c - \left(\min_{1\le j\le m}\lambda_j\right)^{-1}\right]$
Thm. 6.2 (Capped ℓ_1)		w^* is a global minimizer on $\mathcal{W}_*$
Thm. 6.2 (Capped ℓ_1)	$c\|\nabla\ell(0)\|_2 < \gamma$	$\|w^*\|_2 \le \frac{c\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}}\left[\min_{1\le j\le m}\frac{1}{a_j} - c^{-1}\right]$ $\Rightarrow \sqrt{\|w^*\|_0} \le \frac{c\|\nabla\ell(0)\|_2}{\left[\min_{1\le j\le m} a_j\right]\mathrm{Lip}_{\nabla\ell}}\left[\min_{1\le j\le m}\frac{1}{a_j} - c^{-1}\right]$
Thm. 6.3 (Transformed ℓ_1)	$c\|\|\nabla\ell(0)\|\|_2 \le \gamma \le \frac{\lambda^\ell_{\min}}{2\max_{1\le i\le m}\frac{a_i+1}{a_i^2}}$	w^* is a global minimizer on $\mathcal{W}$
Thm. 6.3 (Transformed ℓ_1)	$c\|\|\nabla\ell(0)\|\|_2 < \gamma \le \frac{\lambda^\ell_{\min}}{2\max_{1\le i\le m}\frac{a_i+1}{a_i^2}}$	$\|w^*\|_2 \le \frac{c\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}}\left[\frac{1}{4}\left(1 + \min_{1\le i\le m}\frac{1}{a_i}\right) - c^{-1}\right]$ $\Rightarrow \forall\, i = 1, \ldots, m$, either $w^*_i = 0$ or $\|w^*_i\| \ge a_i$
Thm. 6.4 (Logarithmic)	$c\|\|\nabla\ell(0)\|\|_2 \le \gamma \le \frac{\lambda^\ell_{\min}}{\max_{1\le i\le m}\frac{\lambda_i}{\varepsilon_i}}$	w^* is a global minimizer on $\mathcal{W}$
Thm. 6.4 (Logarithmic)	$c\|\nabla\ell(0)\|_2 < \gamma \le \frac{\lambda^\ell_{\min}}{\max_{1\le i\le m}\frac{\lambda_i}{\varepsilon_i}}$	$\|w^*\|_2 \le \frac{c\|\nabla\ell(0)\|_2}{\mathrm{Lip}_{\nabla\ell}}\left[\min_{1\le i\le m}\frac{\lambda_i}{2\varepsilon_i} - c^{-1}\right]$ $\Rightarrow \sqrt{\|w^*\|_0} \le \frac{c\|\nabla\ell(0)\|_2}{\left[\min_{1\le m}\varepsilon_i\right]\mathrm{Lip}_{\nabla\ell}}\left[\min_{1\le i\le m}\frac{\lambda_i}{2\varepsilon_i} - c^{-1}\right]$

We then have

$$h_{[K]}(w) \;=\; \max\left\{\sum_{i=1}^m v_i\,\sigma_i\,w_i \;\mid\; v \in \mathcal{E}(\Delta_{K,m});\, \sigma \in \{\pm 1\}^m\right\}.$$

Note that if $v \in \mathcal{E}(\Delta_{K,m})$ then each component of v is either 0 or 1 and there are exactly K ones. For a given pair $(v,\sigma) \in \mathcal{E}(\Delta_{K,m}) \times \{\pm 1\}^m$, the linear function $h_{v,\sigma} : w \mapsto \sum_{i=1}^m v_i\,\sigma_i\,w_i$ has gradient given by $v \circ \sigma$, where $\circ$ is the notation for the Hadamard product of two vectors. Hence,

$$\|\,\nabla_{\neq 0} h_{v,\sigma}(w)\,\|_2 \;=\; \sqrt{\sum_{i\,:\,w_i\neq 0} v_i^2\,\sigma_i^2} \;=\; \sqrt{\min(\,K,\,\|w\|_0\,)}.$$

Since for any nonzero vector w and for any $\eta \in \partial g_{[K]}(w)$, $\|\eta_{\neq 0}\|_2 = \sqrt{\|w\|_0}$, it follows

that for all $w \in \mathcal{W}$, all $\eta \in \partial g_{[K]}(w)$, and all pairs $(v, \sigma) \in \mathcal{E}(\Delta_{K,m}) \times \{\pm 1\}^m$,

$$\begin{aligned} \| \eta_{\neq 0} - \nabla_{\neq 0} h_{v,\sigma}(w) \|_2 &\geq \| \eta_{\neq 0} \|_2 - \| \nabla_{\neq 0} h_{v,\sigma}(w) \|_2 \\ &\begin{cases} = 0 & \text{if } \| w \|_0 \leq K, \\ \geq \sqrt{K+1} - \sqrt{K} & \text{if } \| w \|_0 \geq K+1. \end{cases} \end{aligned}$$

Since each function $h_{v,\sigma}$ is affine, Proposition 3.1 applies. In order to identify the vectors $w \in \mathcal{W}$ that share a common piece with w^*, i.e., $w \in \mathcal{W}_*$, we arrange the components of $|w^*|$ as follows:

$$\begin{aligned} |w^*_{[1]}| \geq \cdots \geq |w^*_{[s^*]}| &> \underbrace{|w^*_{[s^*+1]}| = \cdots = |w^*_{[K]}| = \cdots = |w^*_{[t^*-1]}|}_{\text{all equal to } |w^*_{[K]}|} \\ &> |w^*_{[t^*]}| \geq \cdots \geq |w^*_{[m]}|, \end{aligned}$$

and let

$$\mathcal{I}^*_{>} \triangleq \{[1], \ldots, [s^*]\}, \quad \mathcal{I}^*_{=} \triangleq \{[s^*+1], \ldots, [t^*-1]\}, \quad \text{and} \quad \mathcal{I}^*_{<} \triangleq \{[t^*], \ldots, [m]\}.$$

Vectors w that share a common piece with w^* are those such that $P_{[K]}(w) = \sum_{i \in \mathcal{I}^*_{<}} |w_i| + \sum_{i \in \mathcal{J}^*_{=}} |w_i|$, where $\mathcal{J}^*_{=}$ is any subset of $\mathcal{I}^*_{=}$ with $t^* - (K+1)$ elements. Let $Z_{[K];\gamma}(w) \triangleq \ell(w) + \gamma P_{[K]}(w)$.

THEOREM 5.1. *Assume that conditions* $(\mathrm{A}^{\mathrm{L}}_{\ell})$, $(\mathrm{A}^{\mathrm{cvx}}_{\ell})$, *and* (B) *hold. For a fixed integer* $K \geq 1$ *and scalar* $\gamma > 0$, *the following two statements hold for a d-stationary point* w^* *of the function* $Z_{[K];\gamma}(w)$ *on* $\mathcal{W}$:
(a) w^* *is a global minimizer of* $Z_{[K];\gamma}(w)$ *on the subset of vectors* $w \in \mathcal{W}$ *that share a common piece with* w^*.
(b) *Suppose that* $\gamma > c \| \nabla \ell(0) \|_2$ *for some* $c > 0$ *satisfying* $\sqrt{K+1} > \sqrt{K} + c^{-1}$. *If* $\| w^* \|_2 \leq \frac{c \| \nabla \ell(0) \|_2}{\mathrm{Lip}_{\nabla \ell}} \left[\sqrt{K+1} - \sqrt{K} - c^{-1} \right]$, *then* $\| w^* \|_0 \leq K$.

Proof. Statement (a) follows from Proposition 3.1 with $\mathrm{Lip}_{\nabla h} = 0$. Statement (b) follows from the inequality $\| w^* \|_2 \mathrm{Lip}_{\nabla \ell} \geq \gamma \left[\| \eta_{\neq 0} - \nabla_{\neq 0} h_{v,\sigma}(w) \|_2 - c^{-1} \right]$ by a contrapositive argument and the lower bound on $\| \eta_{\neq 0} - \nabla_{\neq 0} h_{v,\sigma}(w) \|_2$; cf. the proof of Proposition 3.4. ∎

5.1. The ℓ_1 ℓ_2 function. When $K = 1$, the zeros of the function $P_{[1]}(w) = \|w\|_1 - \|w\|_\infty$ are the 1-sparse vectors. It turns out that the function $P_{\ell_{1-2}}(w) \triangleq \|w\|_1 - \|w\|_2$ has the same property as $P_{[1]}(w)$ in this regard. Nevertheless, structurally, these two functions are different: specifically, $P_{[1]}$ is a piecewise linear function whereas the ℓ_2-function in $P_{\ell_{1-2}}$ is not piecewise smooth although it is differentiable everywhere except at the origin. As such, Corollary 3.3 is applicable. As shown previously, condition $(\mathrm{A}^{\mathrm{L}}_h)$ holds with $\mathrm{Lip}_{\nabla h} = 0$ and $\beta = 1$. Moreover, for any $w \neq 0$, $\|\nabla \|w\|_2\|_2 = 1$. Corresponding to $P_{\ell_{1-2}}$, we write $Z_{\ell_{1-2};\gamma}(w) \triangleq \ell(w) + \gamma P_{\ell_{1-2}}(w)$.

THEOREM 5.2. *Assume conditions* $(\mathrm{A}^{\mathrm{L}}_{\ell})$, $(\mathrm{A}^{\mathrm{cvx}}_{\ell})$, *and* (B). *Let* $c > 0$ *satisfy*

$$c \left[2 \, Lip_{\nabla \ell} - \lambda^{\ell}_{\min} \right] \leq \lambda^{\ell}_{\min}.$$

For any $\gamma > c \| \nabla \ell(0) \|_2$, *if* w^* *is a d-stationary solution of* $Z_{\ell_{1-2};\gamma}$ *on* $\mathcal{W}$, *then either* $\sqrt{\| w^* \|_0} \leq 2 (1 + c^{-1})$, *or*

$$Z_{\ell_{1-2};\gamma}(w) - Z_{\ell_{1-2};\gamma}(w^*) \geq \left[\frac{\lambda^{\ell}_{\min}}{2} - \frac{Lip_{\nabla \ell}}{1 + c^{-1}} \right] \| w - w^* \|_2^2 \quad \forall \, w \in \mathcal{W}. \tag{20}$$

Remark. The inequality (20) is global on the entire set $\mathcal{W}$. ∎

We should point out that there are other piecewise linear functions whose zeros are 1-sparse but are not necessarily exact sparsity functions. Indeed, for any finite subset V of the unit Euclidean ball, the function $P_V(w) = \|w\|_1 - \max_{v\in V} v^T w$ is one such function. This can be seen from the inequality $P_V(w) \geq \|w\|_1 - \|w\|_2$, from which it follows that the zeros of $P_V(w)$ must be 1-sparse. Nevertheless, the 1-sparse vectors are not necessarily the zeros of $P_V(w)$ if the subset V is not properly chosen. Yet a result similar to Theorem 5.1 can be derived for the function P_V.

6. Surrogate sparsity functions. We next examine several inexact sparsity functions. All such functions to be examined are of the *folded concave* type [19] and separable so that $P(w) = \sum_{i=1}^m p_i(w_i)$ with each $p_i(w_i)$ representable as the difference of a convex and a pointwise maximum of finitely many differentiable convex function; cf. 2); i.e., $p_i(w_i) = g_i(w_i) - \max_{1\leq j\leq I_i} h_{i,j}(w_i)$ where g_i and each $h_{i,j}$ are univariate convex functions with $h_{i,j}$ differentiable. For each sparsity function examined below, we discuss the applicability of the results in section 3.

6.1. The SCAD family. Foremost among the separable sparsity functions is the SCAD (*smoothly clipped absolute deviation*) family [17, 18, 35]. Parameterized by two scalars $a > 2$ and $\lambda > 0$ and with the origin as its unique zero, this univariate function is once continuously differentiable except at the origin and given by, for all $t \in \mathbb{R}$,

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

The representation of this function as a dc function $g_\lambda^{\mathrm{SCAD}}(t) - h_{a,\lambda}^{\mathrm{SCAD}}(t)$ with $h_{a,\lambda}^{\mathrm{SCAD}}(t)$ being differentiable and $g_\lambda^{\mathrm{SCAD}}(t)$ being a multiple of the absolute-value function is known [22]. (Nevertheless, Theorem 6.1 is new.) Specifically, we may take

$$g_\lambda^{\mathrm{SCAD}}(t) \triangleq \lambda\,|\,t\,| \quad \text{and} \quad h_{a,\lambda}^{\mathrm{SCAD}}(t) \triangleq \begin{cases} 0 & \text{if } |\,t\,| \leq \lambda, \\ \dfrac{(\,|\,t\,| - \lambda\,)^2}{2\,(\,a-1\,)} & \text{if } \lambda \leq |\,t\,| \leq a\,\lambda, \\ \lambda\,|\,t\,| - \dfrac{(\,a+1\,)\,\lambda^2}{2} & \text{if } |\,t\,| \geq a\,\lambda. \end{cases}$$

The function $h_{a,\lambda}^{\mathrm{SCAD}}(t)$ is continuously differentiable with derivative given by

$$\frac{dh_{a,\lambda}^{\mathrm{SCAD}}(t)}{dt} = \begin{cases} 0 & \text{if } |\,t\,| \leq \lambda, \\ \dfrac{|\,t\,| - \lambda}{a-1}\,\mathrm{sign}(t) & \text{if } \lambda \leq |\,t\,| \leq a\,\lambda, \\ \lambda\,\mathrm{sign}(t) & \text{if } |\,t\,| \geq a\,\lambda. \end{cases}$$

It is not difficult to see that the function $h_{a,\lambda}^{\mathrm{SCAD}}(t)$ is LC^1 on $\mathbb{R}$ with its derivative being Lipschitz continuous with a Lipschitz constant of 2 (recall that $a > 2$). To

verify the latter property, we need to show that

$$\left| \frac{dh^{\mathrm{SCAD}}_{a,\lambda}(t)}{dt} - \frac{dh^{\mathrm{SCAD}}_{a,\lambda}(t')}{dt} \right| \leq 2\,|\,t - t'\,|$$

for two arbitrary scalars t and t'. The derivation below establishes this for t and t' satisfying $-a\lambda \leq t \leq -\lambda$ and $\lambda \leq t' \leq a\lambda$:

$$\left| \frac{dh^{\mathrm{SCAD}}_{a,\lambda}(t)}{dt} - \frac{dh^{\mathrm{SCAD}}_{a,\lambda}(t')}{dt} \right| = \left| \frac{t+\lambda}{a-1} - \frac{t'-\lambda}{a-1} \right| = \left| \frac{t-t'}{a-1} - \frac{2\,\lambda}{a-1} \right| \leq 2\,|\,t - t'\,|$$

because $a > 2$. The same inequality can be derived for all other cases of the pair (t, t'). As a consequence of this LC^1 property, Proposition 3.1 is applicable to the SCAD family of surrogate sparsity functions $Z^{\mathrm{SCAD}}_{\gamma;a,\lambda}(w) \triangleq \ell(w) + \gamma\, P^{\mathrm{SCAD}}_{a,\lambda}(w)$, where for given positive constants $\{a_i, \lambda_i\}_{i=1}^m$ with $a_i > 2$ for all i,

$$P^{\mathrm{SCAD}}_{a,\lambda}(w) \triangleq \underbrace{\sum_{i=1}^m \lambda_i\,|\,w_i\,|}_{\text{weighted } \ell_1} - \underbrace{\sum_{i=1}^m h^{\mathrm{SCAD}}_{a_i,\lambda_i}(w_i)}_{h^{\mathrm{SCAD}}_{a,\lambda}(w)}.$$

The cited proposition yields the strong convexity of the objective $Z^{\mathrm{SCAD}}_{\gamma;a;\lambda}(w)$ on $\mathbb{R}^m$, provided that $\gamma \leq \lambda^{\ell}_{\min}/2$. To obtain a bound on $\|w^*\|_0$ from Theorem 3.2, we evaluate $\|\,\eta_{\neq 0} - \nabla_{\neq 0} h^{\mathrm{SCAD}}_{a;\lambda}(w)\,\|_2^2$ for a subgradient $\eta \in \partial g^{\mathrm{SCAD}}_{\lambda}(w)$. We have

$$\begin{aligned}
\|\,\eta_{\neq 0} - \nabla_{\neq 0} h^{\mathrm{SCAD}}_{a;\lambda}(w)\,\|_2^2 &= \sum_{k\,:\,w_k \neq 0} \left(\lambda_k \operatorname{sign}(w_k) - \frac{dh^{\mathrm{SCAD}}_{a_k,\lambda_k}(w_k)}{dw_k} \right)^2 \\
&= \sum_{k\,:\,0<|w_k|<\lambda_k} \lambda_k^2 + \sum_{k\,:\,\lambda_k \leq |w_k| \leq a_k\lambda_k} \left(\lambda_k - \frac{|\,w_k\,| - \lambda_k}{a_k - 1} \right)^2 \\
&= \sum_{k\,:\,0<|w_k|<\lambda_k} \lambda_k^2 + \sum_{k\,:\,\lambda_k \leq |w_k| \leq a_k\lambda_k} \left(\frac{a_k\,\lambda_k - |\,w_k\,|}{a_k - 1} \right)^2 \\
&\geq \min_{1\leq j\leq m} \lambda_j^2 \quad \text{provided that there is one } k \text{ such that } 0 < |w_k| < \lambda_k.
\end{aligned} \tag{21}$$

THEOREM 6.1. *Assume conditions* $(\mathrm{A}^{\mathrm{L}}_{\ell})$, $(\mathrm{A}^{\mathrm{cvx}}_{\ell})$, *and* (B). *Let* $\{a_k, \lambda_k\}_{k=1}^m$ *be positive scalars such that* $a_k > 2$ *for all* k. *For every positive scalar* $\gamma \leq \frac{1}{2}\,\lambda^{\ell}_{\min}$, *if* w^* *is a d-stationary point of* $Z^{\mathrm{SCAD}}_{\gamma;a;\lambda}(w)$ *on* $\mathcal{W}$, *then* w^* *is a minimizer of this function on* $\mathcal{W}$; *more precisely,*

$$Z^{\mathrm{SCAD}}_{\gamma;a;\lambda}(w) - Z^{\mathrm{SCAD}}_{\gamma;a;\lambda}(w^*) \geq \left[\frac{\lambda^{\ell}_{\min}}{2} - \gamma \right] \|\,w - w^*\,\|_2^2 \quad \forall\, w \in \mathcal{W}.$$

Moreover, if $\gamma > c\,\|\,\nabla\ell(0)\,\|_2$ *for some* $c > 0$ *and*

$$\|\,w^*\,\|_2 \leq \frac{c\,\|\,\nabla\ell(0)\,\|_2}{Lip_{\nabla\ell}} \left[\min_{1\leq j\leq m} \lambda_j - c^{-1} \right], \tag{22}$$

then for every $k = 1, \ldots, m$, *either* $w^*_k = 0$ *or* $|w^*_k| \geq \lambda_k$; *hence* $\sqrt{\|\,w^*\,\|_0} \leq \frac{\|\,\nabla\ell(0)\,\|_2}{\mathrm{Lip}_{\nabla\ell}} \left[c - \frac{1}{\min_{1\leq j\leq m}\,\lambda_j} \right]$.

Proof. It suffices to show the last two assertions of the theorem. Assume the choice of γ and the bound on $\|w^*\|_2$. There is nothing to prove if $w^* = 0$. Otherwise, we must have $\min_{1\le j\le m} \lambda_j - c^{-1} > 0$. Moreover, if there exists k such that $0 < |w_k^*| < \lambda_k$, then $\| \eta_{\neq 0} - \nabla_{\neq 0} h_{a;\lambda}^{\rm SCAD}(w) \|_2 \ge \min_{1\le j\le m} \lambda_j$. This contradicts the inequality $\| w^* \|_2 \, {\rm Lip}_{\nabla \ell} \ge \gamma \left[\| \eta_{\neq 0} - \nabla_{\neq 0} h_{a;\lambda}^{\rm SCAD}(w) \|_2 - c^{-1} \right]$. To complete the proof of the theorem, let $\| w^* \|_0 = K$. Then

$$\sqrt{K} \min_{1\le j\le m} \lambda_j \le \| w^* \|_2 \le \frac{c \, \| \nabla \ell(0) \|_2}{{\rm Lip}_{\nabla \ell}} \left[\min_{1\le j\le m} \lambda_j - c^{-1} \right],$$

from which the desired bound on K follows. ∎

Remarks. The two conditions $c\|\nabla \ell(0)\|_2 < \gamma \le \frac{1}{2} \lambda_{\min}^{\ell}$ and (22) together yield $\min_{1\le j\le m} \lambda_j \ge \frac{2 \| \nabla \ell(0) \|_2}{\lambda_{\min}^{\ell}}$, which offers a guide in choosing the parameters λ_j in the SCAD function so that Theorem 6.1 is applicable to the function $Z_{\gamma;a;\lambda}^{\rm SCAD}(w)$. The recipe of deriving a bound on $\|w^*\|_0$ from the individual components w_k^* persists in the later results; details will not be repeated. ∎

6.2. The MCP family. Next we discuss the MCP (*minimax concave penalty*) family of surrogate sparsity functions [53]. Also parameterized by two positive scalars $a > 2$ and λ, the building block of these functions is the univariate, piecewise quadratic function: for $t \in \mathbb{R}$,

$$p_{a,\lambda}^{\rm MCP}(t) \triangleq a \, \lambda^2 - \frac{[\, (a\lambda - |\, t \,|\,)_+ \,]^2}{a}.$$

Similar to the SCAD decomposition, we may take

$$g_{\lambda}^{\rm MCP}(t) \triangleq 2 \, \lambda \, |\, t \,| \quad \text{and} \quad h_{a,\lambda}^{\rm MCP}(t) \triangleq \begin{cases} \dfrac{t^2}{a} & \text{if } |\, t \,| \le a \, \lambda, \\ 2 \, \lambda \, |\, t \,| - a \, \lambda^2 & \text{if } |\, t \,| \ge a \, \lambda. \end{cases}$$

Moreover, the function $h_{a,\lambda}^{\rm MCP}(t)$ is convex and continuously differentiable with a derivative given by

$$\frac{dh_{a,\lambda}^{\rm MCP}(t)}{dt} = \begin{cases} \dfrac{2\,t}{a} & \text{if } |\, t \,| \le a \, \lambda, \\ 2 \, \lambda \, {\rm sign}(t) & \text{if } |\, t \,| \ge a \, \lambda. \end{cases}$$

Moreover, using the fact that $a > 2$, we can verify that, for any two scalars t and t',

$$\left| \frac{dh_{a,\lambda}^{\rm MCP}(t)}{dt} - \frac{dh_{a,\lambda}^{\rm MCP}(t')}{dt} \right| \le |\, t - t' \,|.$$

The MCP sparsity function is defined as follows: for some positive constants $\{a_i, \lambda_i\}_{i=1}^m$ with $a_i > 2$ for all i,

$$P_{a,\lambda}^{\rm MCP}(w) \triangleq \underbrace{2 \sum_{i=1}^m \lambda_i \, |\, w_i \,| - \sum_{i=1}^m h_{a_i,\lambda_i}^{\rm MCP}(w_i)}_{h_{a,\lambda}^{\rm MCP}(w)}.$$

At this point, we can proceed similarly to the SCAD function and obtain a result analogous to Theorem 6.1. In particular, since $h_{a,\lambda}^{\rm MCP}$ is differentiable, the globally minimizing property of a stationary solution is on the entire set $\mathcal{W}$. We omit the details.

6.3. The capped ℓ_1 family. One distinction of this family from the previous two families is that the capped ℓ_1 functions are piecewise linear; thus Propositions 3.1 and 3.4 are applicable. The building block of this family of functions is as follows: for a given scalar $a > 0$ and for all $t \in \mathbb{R}$,

$$p_a^{\mathrm{capL1}}(t) \triangleq \min\left(\frac{|\,t\,|}{a}, 1\right) = \frac{|\,t\,|}{a} - \max\left(0, \frac{|\,t\,|}{a} - 1\right).$$

This leads to the surrogate penalty function: given positive scalars $\{a_i\}_{i=1}^m$ and for all $t \in \mathbb{R}$,

$$P_a^{\mathrm{capL1}}(w) \triangleq \sum_{i=1}^m \min\left(\frac{|\,w_i\,|}{a_i}, 1\right) = \underbrace{\sum_{i=1}^m \frac{|\,w_i\,|}{a_i}}_{g_a^{\mathrm{capL1}}(w)} \underbrace{- \sum_{i=1}^m \max\left(0, \frac{w_i}{a_i} - 1, \frac{-w_i}{a_i} - 1\right)}_{h_a^{\mathrm{capL1}}(w)}.$$

We have the following result for the function $Z_{\gamma;a}^{\mathrm{capL1}}(w) \triangleq \ell(w) + \gamma P_a^{\mathrm{capL1}}(w)$. Due to the piecewise property of the function $h_a^{\mathrm{capL1}}(w)$, the minimizing property is of the restricted kind for vectors sharing a common piece with the given stationary solution on hand.

THEOREM 6.2. *Assume conditions* ($\mathrm{A}_\ell^{\mathrm{L}}$), ($\mathrm{A}_\ell^{\mathrm{cvx}}$), *and* (B). *Let* $\{a_k\}_{k=1}^m$ *be positive scalars. The following two statements hold for a d-stationary point* w^* *of* $Z_{\gamma;a}^{\mathrm{capL1}}(w)$ *on* $\mathcal{W}$ *for any* $\gamma > 0$.

(a) w^* *is a global minimizer of* $Z_{\gamma;a}^{\mathrm{capL1}}(w)$ *on the subset of vectors of* $\mathcal{W}$ *that share a common piece with* w^*.

(b) *If* $\gamma > c \,\|\, \nabla\ell(0) \,\|_2$ *for some* $c > 0$ *and* $\|\, w^* \,\|_2 \leq \frac{c \,\|\, \nabla\ell(0) \,\|_2}{\mathrm{Lip}_{\nabla\ell}} \left[\, \min_{1\leq j\leq m} \frac{1}{a_j} - c^{-1} \,\right]$, *then for every* $k = 1, \ldots, m$, *either* $w_k^* = 0$ *or* $|w_k^*| \geq a_k$; *thus* $\sqrt{\|\, w^* \,\|_0} \leq \frac{c \,\|\, \nabla\ell(0) \,\|_2}{\left[\min_{1\leq j\leq m} a_j\right] \mathrm{Lip}_{\nabla\ell}} \left[\, \min_{1\leq j\leq m} \frac{1}{a_j} - c^{-1} \,\right]$. ∎

6.4. The transformed ℓ_1-family. Employed recently by [54], this function is given as follows: for a given $a > 0$ and for $t \in \mathbb{R}$,

$$p_a^{\mathrm{TL1}}(t) \triangleq \frac{(\,a+1\,)\,|\,t\,|}{a + |\,t\,|}$$

which has the dc decomposition

$$p_a^{\mathrm{TL1}}(t) = \underbrace{\frac{a+1}{a}\,|\,t\,|}_{g_a^{\mathrm{TL1}}(t)} - \underbrace{\left[\frac{a+1}{a}\,|\,t\,| - \frac{(\,a+1\,)\,|\,t\,|}{a + |\,t\,|}\right]}_{h_a^{\mathrm{TL1}}(t)}.$$

The univariate function $h_a^{\mathrm{TL1}}(t)$ is strictly convex and (infinitely many times) differentiable on the real line as can be seen from its first and second derivatives:

$$\frac{dh_a^{\mathrm{TL1}}(t)}{dt} = \left[\frac{a+1}{a} - \frac{a\,(\,a+1\,)}{(\,a + |\,t\,|\,)^2}\right] \mathrm{sign}(t) \quad \text{and} \quad \frac{d^2 h_a^{\mathrm{TL1}}(t)}{dt^2} = \frac{2a\,(\,a+1\,)}{(\,a + |\,t\,|\,)^3}.$$

The second derivative also shows that $\frac{dh_a^{\mathrm{TL1}}(t)}{dt}$ is Lipchitz continuous with modulus $\frac{2\,(\,a+1\,)}{a^2}$. For given positive parameters $\{a_i\}_{i=1}^m$, and with

$$g_a^{\mathrm{TL1}}(w) \triangleq \sum_{i=1}^m \frac{a_i+1}{a_i}\,|\,w_i\,| \quad \text{and} \quad h_a^{\mathrm{TL1}}(w) \triangleq \sum_{i=1}^m \left[\frac{a_i+1}{a_i}\,|\,w_i\,| - \frac{(\,a_i+1\,)\,|\,w_i\,|}{a_i + |\,w_i\,|}\right],$$

we have for any $\eta \in \partial g_a^{\mathrm{TL1}}(w)$,

$$\begin{aligned} \left\| \, \eta_{\neq 0} - \nabla_{\neq 0} h_a^{\mathrm{TL1}}(w) \, \right\|_2 &= \sqrt{\sum_{i \, : \, w_i \neq 0} \left(\frac{a_i \, (\, a_i + 1 \,)}{(\, a_i + | \, w_i \, | \,)^2} \right)^2} \\ &\geq \frac{1}{4} \left(1 + \min_{1 \leq i \leq m} \frac{1}{a_i} \right) \quad \text{if } \exists \, k \;\; \text{such that} \;\; 0 < | \, w_k \, | < a_k. \end{aligned}$$

Taking $\mathrm{Lip}_{\nabla h} = 2 \max_{1 \leq i \leq m} \frac{a_i + 1}{a_i^2}$, we obtain a result for the function $Z_{\gamma;a}^{\mathrm{TL1}}(w) \triangleq \ell(w) + \gamma \, P_a^{\mathrm{TL1}}(w)$, where

$$P_a^{\mathrm{TL1}}(w) \triangleq \sum_{i=1}^{m} \frac{(\, a_i + 1 \,) \, | \, w_i \, |}{a_i + | \, w_i \, |}$$

that is similar to Theorem 6.1 for the function $Z_{\gamma;a;\lambda}^{\mathrm{SCAD}}$ and Theorem 6.2 for the function $Z_{\gamma;a}^{\mathrm{capL1}}(w)$.

THEOREM 6.3. *Assume conditions* $(\mathrm{A}_\ell^{\mathrm{L}})$, $(\mathrm{A}_\ell^{\mathrm{cvx}})$, *and* (B). *Let* $\{a_k\}_{k=1}^m$ *be positive scalars. For every positive scalars* c *and* γ *such that*

$$c \, \| \, \nabla \ell(0) \, \|_2 \leq \gamma \leq \frac{\lambda_{\min}^{\ell}}{2 \max\limits_{1 \leq i \leq m} \dfrac{a_i + 1}{a_i^2}},$$

if w^* *is a d-stationary point of* $Z_{\gamma;a}^{\mathrm{TL1}}(w)$ *on* $\mathcal{W}$, *then* w^* *is a minimizer of this function on* $\mathcal{W}$; *more precisely,*

$$Z_{\gamma;a}^{\mathrm{TL1}}(w) - Z_{\gamma;a}^{\mathrm{TL1}}(w^*) \geq \left[\frac{\lambda_{\min}^{\ell}}{2} - \gamma \max_{1 \leq i \leq m} \frac{a_i + 1}{a_i^2} \right] \| \, w - w^* \, \|_2^2 \quad \forall \, w \in \mathcal{W}.$$

Moreover, if $\gamma > c \, \| \, \nabla \ell(0) \, \|_2$ *and* $\| \, w^* \, \|_2 \leq \frac{c \, \| \, \nabla \ell(0) \, \|_2}{\mathrm{Lip}_{\nabla \ell}} \left[\, \frac{1}{4} \left(1 + \min_{1 \leq i \leq m} \frac{1}{a_i} \right) - c^{-1} \, \right]$, *then for every* $k = 1, \ldots, m$, *either* $w_k^* = 0$ *or* $|w_k^*| \geq a_k$. ∎

6.5. The logarithmic family. Introduced in [9] as an optimization formulation for the reweighted ℓ_1 procedure, and studied in particular in [22, 31], this family of functions is built from the univariate function: given positive scalars λ and ε,

$$p_{\lambda;\varepsilon}^{\log}(t) \triangleq \lambda \, \log(| \, t \, | + \varepsilon) - \lambda \, \log \varepsilon, \quad t \in \mathbb{R},$$

which has the dc decomposition

$$p_{\lambda;\varepsilon}^{\log}(t) = \frac{\lambda}{\varepsilon} \, | \, t \, | - \underbrace{\lambda \left[\frac{| \, t \, |}{\varepsilon} - \log(| \, t \, | + \varepsilon) + \log \varepsilon \right]}_{h_{\lambda;\varepsilon}^{\log}(t)}.$$

(Although this logarithmic function fails to satisfy the "continuity property" in the set of postulates of [17], we include it here to illustrate the breadth of our framework.) The univariate function $h_{\lambda;\varepsilon}^{\log}$ is strictly convex and (infinitely many times) differentiable on the real line as can be seen from its first and second derivatives:

$$\frac{dh_{\lambda;\varepsilon}^{\log}(t)}{dt} = \frac{\lambda \, t}{\varepsilon \, (\, \varepsilon + | \, t \, | \,)} \quad \text{and} \quad \frac{d^2 h_{\lambda;\varepsilon}^{\log}(t)}{dt^2} = \frac{\lambda}{(\, \varepsilon + | \, t \, | \,)^2} \quad \forall \, t \in \mathbb{R}.$$

The second derivative also shows that $\frac{dh^{\log}_{\lambda;\varepsilon}(t)}{dt}$ is Lipchitz continuous with modulus λ/ε^2. For given positive parameters $\{\lambda_i, \varepsilon_i\}_{i=1}^m$, and with

$$g^{\log}_{\lambda;\varepsilon}(w) \triangleq \sum_{i=1}^m \frac{\lambda_i}{\varepsilon_i} |\, w_i \,| \quad \text{and} \quad h^{\log}_{\lambda;\varepsilon}(w) \triangleq \sum_{i=1}^m \lambda_i \left[\frac{|\, w_i \,|}{\varepsilon_i} - \log(|\, w_i \,| + \varepsilon_i) + \log \varepsilon_i \right],$$

we have for any $\eta \in \partial g^{\log}_{\lambda;\varepsilon}(w)$,

$$\begin{aligned} \left\| \eta_{\neq 0} - \nabla_{\neq 0} h^{\log}_{\lambda;\varepsilon}(w) \right\|_2 &= \sqrt{\sum_{i\,:\,w_i \neq 0} \frac{\lambda_i^2}{(\varepsilon_i + |\, w_i \,|)^2}} \\ &\geq \min_{1 \leq i \leq m} \frac{\lambda_i}{2\,\varepsilon_i} \quad \text{if } \exists\, k \ \text{ such that } \ 0 < |\, w_k \,| < \varepsilon_k. \end{aligned}$$

Similar to previous results for the functions $Z^{\mathrm{SCAD}}_{\gamma;a;\lambda}(w)$ and $Z^{\mathrm{TL1}}_{\gamma;a}(w)$, the result below pertains to a d-stationary point of the function $Z^{\log}_{\gamma;\lambda;\varepsilon}(w) \triangleq \ell(w) + \gamma\, P^{\log}_{\lambda;\varepsilon}(w)$, where $P^{\log}_{\lambda;\varepsilon}(w) \triangleq \sum_{i=1}^m \lambda_i \log(|w_i| + \varepsilon_i)$.

THEOREM 6.4. *Assume conditions* $(\mathrm{A}^{\mathrm{L}}_{\ell})$, $(\mathrm{A}^{\mathrm{cvx}}_{\ell})$, *and* (B). *Let* $\{a_k, \varepsilon_k\}_{k=1}^m$ *be positive scalars. For every positive scalars* c *and* γ *such that*

$$c \,\|\, \nabla \ell(0) \,\|_2 \,\leq\, \gamma \,\leq\, \frac{\lambda^{\ell}_{\min}}{\max\limits_{1 \leq i \leq m} \dfrac{\lambda_i}{\varepsilon_i^2}},$$

if w^* *is a d-stationary point of* $Z^{\log}_{\gamma;\lambda;\varepsilon}(w)$ *on* $\mathcal{W}$, *then* w^* *is a minimizer of this function on* $\mathcal{W}$; *more precisely,*

$$Z^{\log}_{\gamma;\lambda;\varepsilon}(w) - Z^{\log}_{\gamma;\lambda;\varepsilon}(w^*) \,\geq\, \tfrac{1}{2} \left[\lambda^{\ell}_{\min} - \gamma \max_{1 \leq i \leq m} \frac{\lambda_i}{\varepsilon_i^2} \right] \|\, w - w^* \,\|_2^2 \quad \forall\, w \,\in\, \mathcal{W}.$$

Moreover, if $\gamma > c \,\|\, \nabla \ell(0) \,\|_2$ *and* $\|\, w^* \,\|_2 \,\leq\, \frac{c \,\|\, \nabla \ell(0) \,\|_2}{\mathrm{Lip}_{\nabla \ell}} \left[\, \min_{1 \leq i \leq m} \frac{\lambda_i}{2\,\varepsilon_i} - c^{-1} \,\right]$, *then for every* $k = 1, \ldots, m$, *either* $w^*_k = 0$ *or* $|w^*_k| \geq \varepsilon_k$. *Hence,*

$$\sqrt{\|\, w^* \,\|_0} \,\leq\, \frac{c \,\|\, \nabla \ell(0) \,\|_2}{\left[\min\limits_{1 \leq m} \varepsilon_i \right] \mathit{Lip}_{\nabla \ell}} \left[\min_{1 \leq i \leq m} \frac{\lambda_i}{2\,\varepsilon_i} - c^{-1} \right].$$

This completes the proof. ∎

6.6. Some comments on results. We note that in the statistical learning literature [57, 60] among others, the bounds on the estimators from minimizing a regularized cost function are probabilistic and typically in the limit of large sample size or dimension. In contrast, our sparsity bounds in Proposition 3.4 and Theorems 5.1, 5.2, 6.1, 6.2, 6.3, and 6.4 on a d-stationary point are deterministic and hold in any fixed sample size and dimension. Moreover, these results are explicit and helpful for analyzing numerical algorithms. In a set of regularized least-squares tests, we find that our sparsity bounds are robust and can still be valid when the theoretical sufficient conditions are not satisfied. In this regard, we should emphasize that these conditions are necessarily conservative and an advanced analysis could yield better sparsity bounds for targeted sparsity functions.

7. Concluding remarks and future work. Based on a general derivation, we have analyzed the minimizing and sparsity properties of a d-stationary solution of the minimization problem (1). The analysis makes it clear about the role of the parameter γ in obtaining these properties. In contrast to the statistical analysis, our approach is totally deterministic, taking as given the form of the optimization problem to be solved in practice and adopting a pragmatic perspective about the kind of solution being analyzed. In a companion paper [1] under preparation, we will examine the computational comparisons of the various problems studied in this paper, including the use of the algorithms described in [42] for computing the d-stationary solutions.

It is natural to ask what role the error bound

$$Z_\gamma(w) - Z_\gamma(w^*) \geq \left[\frac{\lambda^\ell_{\min}}{2} - \gamma \left(\frac{\mathrm{Lip}_{\nabla h}}{2} + \frac{\beta}{\| w^* \|_2} \right) \right] \| w - w^* \|_2^2 \quad \forall\, w \in \mathcal{W} \tag{23}$$

for a d-stationary solution w^* of $Z_\gamma(w)$ on the set $\mathcal{W}$ plays when this function is constructed with the goal of addressing the bicriteria optimization of an expected loss and the sparsity of an underlying statistical model; i.e., the problem

$$\text{bicriteria} \underset{w \in \mathcal{W}}{\text{minimization}} \left(\mathbb{E}\left[\ell(w, \widetilde{\omega}) \right], \| w \|_0 \right). \tag{24}$$

Here, the uncertainty $\widetilde{\omega}$ is a random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with Ω being the sample space, $\mathcal{F}$ being the σ-algebra generated by subsets of Ω, $\mathbb{P}$ being a probability measure defined on $\mathcal{F}$, and $\mathbb{E}$ being the associated expectation operator. We assume that $\ell(\bullet, \omega)$ is convex on $\mathcal{W}$ for every fixed but arbitrary $\omega \in \Omega$. We approximate the bicriteria stochastic program (24) by a regularized sample average approximation (SAA) wherein we take N independently and identically distributed samples $\{\omega^i\}_{i=1}^N$ of the random variable $\widetilde{\omega}$, regularize the loss function to ensure the strong convexity of the sampled loss function, introduce the sparsity function $P(w)$ to substitute for the ℓ_0-function, and convert this biobjective minimization problem (24) to a single-objective, sampled minimization problem:

$$\underset{w \in \mathcal{W}}{\operatorname{argmin}}\, Z_{\gamma;N}(w) \triangleq \underbrace{\frac{1}{N} \sum_{j=1}^N \ell(w, \omega^j) + \frac{\alpha_N}{2}\, w^T w}_{\ell_N(w) \text{ with convexity modulus } \lambda^\ell_{N;\min}} + \gamma_N\, P(w), \tag{25}$$

where $\alpha_N \geq 0$ and $\gamma_N > 0$ may depend on the sample size N. The asymptotic analysis of the standard SAA approach for solving a single-objective expected-value minimization problem: minimize$_{w \in \mathcal{W}}\, \mathbb{E}\left[\ell(w, \widetilde{\omega}) \right]$, is well understood in stochastic programming, particularly for the convex case; see [47, Chapter 5]. The existing analysis is complicated by the presence of the nonconvex function $P(w)$, which renders the resulting SAA subproblem (25) nonconvex. Such an analysis typically works with a global minimizer of (25) and/or its global optimum value, both of which are computationally not available. This is where an error bound such as (23) could be useful. Indeed, a minimizing property of a d-stationary solution provides the missing link between the statistical analysis and computational practice. It is our interest to pursue this line of research in a future work that will explore the role of (23) in such an analysis.

Acknowledgments. The authors gratefully acknowledge the constructive comments offered by two referees that have improved the presentation of this paper. The

efficient handling of the review by the Associate Editor, Professor Defeng Sun, is also greatly appreciated. The authors also thank Professor Xiaoming Huo and his graduate student Shanshan Cao for catching an incorrect expression of the $h_{a,\lambda}^{\mathrm{SCAD}}(t)$ function that has been corrected in this published version.

REFERENCES

[1] M. Ahn, J.S. Pang, and J. Xin, *Difference-of-convex learning* II: *Computations*, in preparation.
[2] D.P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.
[3] B.J. Bickel, Y. Ritov, and A.B. Tsybakov, *Simultaneous analysis of* LASSO *and Dantzig selector*, Ann. Statist., 37 (2009), pp. 1705–1732.
[4] P. Breheny and J. Huang, *Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection*, Ann. Appl. Stat., 5 (2011), pp. 232–253.
[5] E. Candès and C. Fernandez-Granda, *Towards a mathematical theory of super-resolution*, Comm. Pure Appl. Math., 67 (2014), pp. 906–956.
[6] E. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2005), pp. 4203–4215.
[7] E. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.
[8] E. Candès and T. Tao, *The Dantzig selector: Statistical stimation when p is much larger than n (with discussion)*, Ann. Statist., 35 (2007), pp. 2313–2404.
[9] E.J. Candès, M.B. Wakin, and S.P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.
[10] S.S. Chen, D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61, https://doi.org/10.1137/S1064827596304010.
[11] X. Chen, D. Ge, Z. Wang, and Y. Ye, *Complexity of unconstrained $\ell_2 - \ell_p$ minimization*, Math. Program., 143 (2014), pp. 371–383.
[12] D. Donoho, *Superresolution via sparsity constraints*, SIAM J. Math. Anal., 23 (1992), pp. 1309–1331, https://doi.org/10.1137/0523074.
[13] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[14] D. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 2197–2202.
[15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., 32 (2004), pp. 407–499.
[16] F. Facchinei and J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Volumes I and II, Springer-Verlag, New York, 2003.
[17] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360.
[18] J. Fan and J. Lv, *Nonconcave penalized likelihood with NP-dimensionality*, IEEE Trans. Inform. Theory, 57 (2011), pp. 5467–5484.
[19] J. Fan, L. Xue, and H. Zou, *Strong oracle optimality of folded concave penalized estimation*, Ann. Statist., 42 (2014), pp. 819–849.
[20] L.E. Frank and J.H. Friedman, *A statistical view of some chemometrics regression tools*, Technometrics, 35 (1993), pp. 109–135.
[21] Y. Gao and D. Sun, *A Majorized Penalty Approach for Calibrating Rank Constrained Correlation Matrix Problems*, manuscript, Department of Mathematics, National University of Singapore, Singapore, revised May 2010.
[22] G. Gasso, A. Rakotomamonjy, and S. Canu, *Recovering sparse signals with a certain family of nonconvex penalties and DC programming*, IEEE Trans. Signal Process., 57 (2009), pp. 4686–4698.
[23] D. Ge, X. Jiang, and Y. Ye, *A note on the complexity of L_p minimzation*, Math. Program., 129 (2011), pp. 285–299.
[24] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, *A general iterative shrinkage and thresholding algorithm for nonconvex regularized optimization problems*, in Proceedings of the 30th International Conference on Machine Learning, 28 (2013), pp. 37–45.
[25] J.Y. Gotoh, A. Takeda, and K. Tono, *DC Formulations and Algorithms for Sparse Optimization Problems*, manuscript, Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan, 2015.
[26] E. Greenshtein, *Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint*, Ann. Statist., 34 (2006), pp. 2367–2386.

[27] T. Hastie, R. Tibshirani, and M.J. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Series in Statistics and Applied Probability, Chapman and Hall/CRC Press, Boca Raton, FL, 2015.
[28] M. Ho, Z. Sun, and J. Xin, *Weighted elastic net penalized mean-variance portfolio design and computation*, SIAM J. Finan. Math., 6 (2015), pp. 1220–1244, https://doi.org/10.1137/15M1007872.
[29] K. Knight and W. Fu, *Asymptotics for lasso-type estimators*, Ann. Statist., 28 (2000), pp. 1356–1378.
[30] H.A. Le Thi and D.T. Pham, *The DC programming and DCA revised with DC models of real world nonconvex optimization problems*, Ann. Oper. Res., 133 (2005), pp. 25–46.
[31] H.A. Le Thi, D.T. Pham, and X.T. Vo, *DC approximation approaches for sparse optimization*, European J. Oper. Res., 244 (2015), pp. 26–46.
[32] Y. Liu and Y. Wu, *Variable selection via a combination of the L0 and L1 penalties*, J. Comput. Graph. Statist., 16 (2007), pp. 782–798.
[33] P.L. Loh and M.J. Wainwright, *Support Recovery without Incoherence: A Case for Nonconvex Regularization*, preprint, https://arxiv.org/abs/1412.5632, 2014.
[34] Y. Lou, P. Yin, and J. Xin, *Point source super-resolution via non-convex L1 based methods*, J. Sci. Comput., 68 (2016), pp. 1082–1100.
[35] J. Lv and Y. Fan, *A unified approach to model selection and sparse recovery using regularized least squares*, Ann. Statist., 37 (2009), pp. 3498–3528.
[36] O.L. Mangasarian, *Pseudo-convex functions*, J. Soc. Indust. Appl. Math. Ser. A Control, 3 (1965), pp. 281–290.
[37] R. Mazumder, J.H. Friedman, and T. Hastie, *SparseNet: Coordinate descent with nonconvex penalties*, J. Amer. Statist. Assoc., 106 (2011), pp. 1125–1138.
[38] B.S. Mordukhovich, *Variational Analysis and Generalized Differentiation.* I. *Basic theory*, Springer-Verlag, Berlin, 2006, http://www.springer.com/series/138.
[39] B.K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput., 24 (1995), pp. 227–234, https://doi.org/10.1137/S0097539792240406.
[40] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM J. Appl. Math., 61 (2000), pp. 633–658, https://doi.org/10.1137/S0036139997327794.
[41] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics Appl. Math. 30, SIAM, Philadelphia, 2000, https://doi.org/10.1137/1.9780898719468.
[42] J.S. Pang, M. Razaviyayn, and A. Alvarado, *Computing B-stationary points of nonsmooth DC programs*, Math. Oper. Res., 42 (2017), pp. 95–118, https://doi.org/10.1287/moor.2016.0795.
[43] J.S. Pang and M. Tao, *Decomposition Methods for Computing Directional Stationary Solutions of a Class of Non-smooth Non-convex Optimization Problems*, manuscript, 2017.
[44] D.T. Pham and H.A. Le Thi, *Recent advances in dc programming and dca*, Transactions on Computational Collective Intelligence XIII, 8342 (2014), pp. 1–37.
[45] D.T. Pham and H.A. Le Thi, *Convex analysis approach to D.C. programming: Theory, algorithms and applications*, ACTA Math. Vietnam., 22 (1997), pp. 289–355.
[46] D.T. Pham, H.A. Le Thi, and D.M. Le, *Exact penalty in d.c. programming*, Vietnam J. Math., 27 (1999), pp. 169–178.
[47] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718751.
[48] X. Shen, W. Pan, and Y. Zhu, *Likelihood-based selection and sharp parameter estimation*, J. Amer. Statist. Assoc., 107 (2012), pp. 223–232.
[49] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[50] Q. Yao and J.T. Kwok, *Efficient learning with a family of nonconvex regularizers by redistributing nonconvexity*, in Proceedings of the 33rd International Conference on Machine Learning, M.F. Balcan and K.Q. Weinberger, eds., 2016, pp. 2645–2654.
[51] P. Yin, Y. Lou, Q. He, and J. Xin, *Minimization of ℓ_{1-2} for compressed sensing*, SIAM J. Sci. Comput., 37 (2015), pp. A536–A563, https://doi.org/10.1137/140952363.
[52] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for L_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168, https://doi.org/10.1137/070703983.
[53] C. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Ann. Statist., 38 (2010), pp. 894–942.

[54] S. Zhang and J. Xin, *Minimization of transformed L_1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing*, ftp://ftp.math.ucla.edu/pub/camreport/cam14-86.pdf, submitted, 2016.
[55] T. Zhang, *Analysis of multi-stage convex relaxation for sparse regularization*, J. Mach. Learn. Res., 11 (2010), pp. 1081–1107.
[56] P. Zhao and B. Yu, *On model selection consistency of* LASSO, J. Mach. Learn. Res., 7 (2006), pp. 2541–2563.
[57] Z. Zheng, Y. Fan, and J. Lv, *High dimensional thresholded regression and shrinkage effect*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 76 (2014), pp. 627–649.
[58] H. Zou, *The adaptive Lasso and its oracle properties*, J. Amer. Statist. Assoc., 101 (2006), pp. 1418–1429.
[59] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B Stat. Methodol., 67 (2005), pp. 301–320.
[60] H. Zou and R. Li, *One-step sparse estimates in nonconcave penalized likelihood models*, Ann. Statist., 36 (2008), pp. 1509–1533.