Sequential Convex Programming Revisited

Mohsen Kheirandishfard, Fariba Zohrizadeh, Shahrouz Ryan Alimo, Farhad Kamangar, and Ramtin Madani

***Abstract*—This paper revisits the well-known family of sequential convex programming methods. We adopt the difference of convex programming technique to relax a wide variety of non-convex optimization problems into convex programs. We extend this approach to a sequential convex programming algorithm that can generate a convergent sequence of feasible points whose objective values monotonically improve. As an improvement upon the existing sequential methods, we prove that under certain assumptions, the proposed algorithm reaches feasibility within a finite number of rounds, as opposed to asymptotic feasibility. The effectiveness of the proposed approach is corroborated through experiments on the problem of robust linear regression.**

I. Introduction

This paper is concerned with the class of difference of convex programming (DCP) optimization problems of the form

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad u_0(\boldsymbol{x}) - v_0(\boldsymbol{x}) \tag{1a}$$

$$\text{subject to} \quad u_k(\boldsymbol{x}) - v_k(\boldsymbol{x}) \leq 0 \quad k \in \{1,\dots,m\} \tag{1b}$$

where $u_k, v_k : \mathbb{R}^n \to \mathbb{R}$ are convex functions, for every $k \in \{0,1,\dots,m\}$. The formulation (1a) – (1b) covers a wide variety of computationally challenging problems with applications in control [1], [2], [3], machine learning and statistics [4], [5], [6], [7], [8], [9], among other areas [10]. The popularity of DCP is due to the fact that a broad class of functions can be decomposed into difference of convex, including function with continuous Hessian, as well as functions with bounded Hessian [11], [12], [10].

Several sequential algorithms have been studied in the literature for solving problems of the form (1a) – (1b) which rely on local approximations at each step [13], [14], [15], [16], [1], [2], [17], [18]. For instance, the paper [2] proposes an exact penalty method with trust region that integrates constraint violations into the objective. The papers [15], [16] generate a sequence of convex problems by approximating the constraints (1b) with the following convex inequalities:

$$u_k(\boldsymbol{x}) - v_k(\check{\boldsymbol{x}}) - \nabla v_k(\check{\boldsymbol{x}})^\top(\boldsymbol{x}-\check{\boldsymbol{x}}) \leq 0 \quad k\in\{1,\dots,m\}, \tag{2}$$

where $\check{\boldsymbol{x}} \in \mathbb{R}^n$ is a guess for the solution. Due to convexity of the functions $\{v_k\}_{k\in\mathcal{M}}$, the constraints (2) result in an inner-approximation of the feasible set, which can be empty if $\check{\boldsymbol{x}}$ is not feasible for the problem (1a) – (1b). Hence, this method may result in an infeasible approximation. To circumvent this issue, the papers [19], [20], [21] improve this approach by penalizing the violation of the approximate inequalities (2), instead of imposing them as hard constraints.

In this paper, we relax DCP problems of the form (1a) – (1b) and incorporate a novel penalty term into the objective that enforces feasibility. We introduce certain conditions under which the penalized relaxation is guaranteed to produce feasible points for the original problem (1a) – (1b). Moreover, we extend this approach to an algorithm that generates a convergent sequence of feasible points whose objective values monotonically improve. While the exiting sequential methods are proven to offer asymptotic feasibility, our method achieves feasibility within a finite number of rounds, under the standard assumption that the initial point is sufficiently close to the feasible set of (1a) – (1b). To demonstrate the potential of the proposed approach, we perform experiments on the problems of robust linear regression.

A. Notations

Throughout this paper, scalars, vectors, and matrices are respectively shown by italic, bold lower-case and bold upper-case letters. The symbols $\mathbb{R}^n$, $\mathbb{R}^n_+$, $\mathbb{R}^{m\times n}$, $\mathbb{S}_n$, and $\mathbb{S}^+_n$ denote the sets of n-dimensional real vectors, n-dimensional real positive vectors, $m \times n$ real matrices, $n \times n$ real symmetric matrices, and $n \times n$ symmetric positive semidefinite matrices, respectively. For a given vector $\boldsymbol{a}$ and matrix $\boldsymbol{A}$, the symbols a_i and A_{ij}, respectively, indicate the i^{th} element of $\boldsymbol{a}$ and $(i,j)^{th}$ element of $\boldsymbol{A}$. The symbol $(.)^\top$ represents the transpose operator. The notation $\boldsymbol{A} \succ 0$ means that $\boldsymbol{A}$ is positive definite. The notation $\|.\|_p$ refers to either matrix norm or vector norm depending on the context, $\|.\|_{\mathrm{F}}$ represents the Frobenius norm, and $|.|$ indicates the cardinality of a set or absolute value depending on the context. ∇ and H represent the gradient and Hessian operators, respectively.

II. Problem Formulation and Theoretical Results

With no loss of generality, we consider the following reformulation of (1a) - (1b):

$$\underset{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \boldsymbol{a}\in\mathbb{R}^{m+1}}}{\text{minimize}} \quad u_0(\boldsymbol{x}) - a_0 \tag{3a}$$

$$\text{subject to} \quad u_k(\boldsymbol{x}) - a_k \leq 0 \qquad k \in \mathcal{M} \tag{3b}$$

$$a_k = v_k(\boldsymbol{x}) \qquad k \in \{0\} \cup \mathcal{M} \tag{3c}$$

where $\mathcal{M} = \{1,\dots,m\}$ represents the set of constraints. Throughout the paper, we assume that this problem is feasible.

In order to solve problems of the form (3a) – (3c), we propose to relax the set of equality constrains (3c) to inequalities and incorporate the following penalty term into the objective:

$$g_{\check{\boldsymbol{x}}}(\boldsymbol{x},\boldsymbol{a}) \triangleq \sqrt{\|\boldsymbol{x}-\check{\boldsymbol{x}}\|_2^2 + \boldsymbol{\mu}^\top[\boldsymbol{a} - \boldsymbol{v}(\check{\boldsymbol{x}}) - \boldsymbol{V}(\check{\boldsymbol{x}})(\boldsymbol{x}-\check{\boldsymbol{x}})]} \tag{4}$$

where $\check{\boldsymbol{x}} \in \mathbb{R}^n$ is an initial guess for the solution, $\boldsymbol{\mu} \in \mathbb{R}^{m+1}_+$ is constant, the function $\boldsymbol{v} : \mathbb{R}^n \to \mathbb{R}^{m+1}$ is defined as

$$\boldsymbol{v}(\boldsymbol{x}) \triangleq [v_0(\boldsymbol{x}),\, v_1(\boldsymbol{x}),\, v_2(\boldsymbol{x}),\, \dots,\, v_m(\boldsymbol{x})]^\top \tag{5}$$

and $\boldsymbol{V}:\mathbb{R}^n\to\mathbb{R}^{(m+1)\times n}$ represents the Jacobian matrix of $\boldsymbol{v}$. Given $\check{\boldsymbol{x}}\in\mathbb{R}^n$ and $\boldsymbol{\mu}\in\mathbb{R}^m_+$, the penalized convex relaxation problem can be formulated as

$$\begin{aligned}
\underset{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \boldsymbol{a}\in\mathbb{R}^{m+1}}}{\text{minimize}} \quad & u_0(\boldsymbol{x})-a_0+\eta\times g_{\check{\boldsymbol{x}}}(\boldsymbol{x},\boldsymbol{a})^2 && && \text{(6a)}\\
\text{subject to} \quad & u_k(\boldsymbol{x})-a_k\le 0 && k\in\mathcal{M} && \text{(6b)}\\
& a_k\ge v_k(\boldsymbol{x}) && k\in\{0\}\cup\mathcal{M} && \text{(6c)}
\end{aligned}$$

where $\eta>0$ is a constant regularization parameter that controls the trade-off between the original objective function and the penalty term. Unlike the original problem (3a) – (3c), the penalized relaxation problem (6a) – (6c) is convex. We make two basic assumptions throughout the paper that are the basis of our theoretical results.

Assumption 1. *There exist $\alpha,\beta\ge 0$ such that*

$$|f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)|\le\alpha\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2+\beta\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2 \quad (7a)$$
$$\|\nabla_{\boldsymbol{x}}(f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2))\|_2\le\beta\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2 \quad (7b)$$

for every pair $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbb{R}^n$, where $f(\boldsymbol{x})=u_0(\boldsymbol{x})-v_0(\boldsymbol{x})$ denotes the objective function (1a)*.*

Assumption 2. *There exist $\delta,\varepsilon\ge 0$ such that*

$$\|\boldsymbol{J}(\boldsymbol{x}_1)-\boldsymbol{J}(\boldsymbol{x}_2)\|_2\le\delta\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2 \quad (8a)$$
$$\|\boldsymbol{V}(\boldsymbol{x}_1)-\boldsymbol{V}(\boldsymbol{x}_2)\|_2\le\varepsilon\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2 \quad (8b)$$

for every pair $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathbb{R}^n$, where $\boldsymbol{J}:\mathbb{R}^n\to\mathbb{R}^{m\times n}$ denotes the Jacobian matrix of constraints (1b)*.*

The following two definitions introduce the notions of feasibility distance and singularity that will be used later to state our theoretical results.

Definition 1. *Let $\mathcal{F}\subseteq\mathbb{R}^n$ denote the feasible set of problem* (1a) – (1b)*. For every $\check{\boldsymbol{x}}\in\mathbb{R}^n$ define*

$$g^{\min}_{\check{\boldsymbol{x}}}\triangleq\min\{g_{\check{\boldsymbol{x}}}(\boldsymbol{x},\boldsymbol{v}(\boldsymbol{x}))\,|\,\boldsymbol{x}\in\mathcal{F}\}. \quad (9)$$

as the feasibility distance of $\check{\boldsymbol{x}}$.

Definition 2. *Define the singularity function $s:\mathbb{R}^n\to\mathbb{R}$ as*

$$s(\boldsymbol{x})\triangleq\begin{cases}\sigma_{\min}(\boldsymbol{J}(\boldsymbol{x})) & \text{if } \boldsymbol{J}(\boldsymbol{x}) \text{ is full row rank}\\ 0 & \text{otherwise}\end{cases} \quad (10)$$

where $\sigma_{\min}(.)$ is the smallest singular value operator.

The following theorem guarantees that if the initial guess $\check{\boldsymbol{x}}$ is feasible for (3a) – (3c), then its feasibility is preserved by solving the penalized convex relaxation problem (6a) – (6c).

Theorem 1. *Let $\check{\boldsymbol{x}}\in\mathcal{F}$ be a feasible point for* (3a) – (3c) *such that $s(\check{\boldsymbol{x}})>0$. If $\eta>\max\{\beta,\mu_1^{-1}\}$ and*

$$\eta>\alpha+\frac{\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2+(3+\varepsilon\|\boldsymbol{\mu}\|+\delta\mu_{\min})\beta}{\mu_{\min}s(\check{\boldsymbol{x}})} \quad (11)$$

then every optimal solution $(\overset{*}{\boldsymbol{x}},\overset{*}{\boldsymbol{a}})$ for the penalized relaxation problem (6a) – (6c) *satisfies* (3c) *and therefore $\overset{*}{\boldsymbol{x}}\in\mathcal{F}$ as well. Moreover $f(\overset{*}{\boldsymbol{x}})\le f(\check{\boldsymbol{x}})$.*

Proof. See Section IV for the proof. □

Algorithm 1 Sequential Penalized Relaxation

Input: $\boldsymbol{x}^0\in\mathbb{R}^n$, $\boldsymbol{\mu}\in\mathbb{R}^{m+1}_+$, and $\eta>0$
1: $k\leftarrow 0$
2: **repeat**
3: $\quad k\leftarrow k+1$
4: $\quad\boldsymbol{x}^k\leftarrow$ use $\boldsymbol{x}^{k-1}$ as initial point and solve (6a) – (6c)
5: **until** stopping criteria is satisfied
Output: $\boldsymbol{x}^k$

There exist numerous optimization problems for which finding a feasible point is a challenging task. Assuming no feasible initial point is available, the next theorem investigates the required conditions that ensure (6a) – (6c) gives a feasible point for (3a) – (3c).

Theorem 2. *Let $\check{\boldsymbol{x}}\in\mathbb{R}^n$ satisfy*

$$s(\check{\boldsymbol{x}})>(3+\delta+\varepsilon\|\boldsymbol{\mu}\|)(\mu_{\min})^{-1}g^{\min}_{\check{\boldsymbol{x}}} \quad (12)$$

if $\eta>\max\{\beta,\mu_1^{-1}\}$ and

$$\eta>\alpha+\frac{\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2+(3+\varepsilon\|\boldsymbol{\mu}\|+\delta\mu_{\min})(\alpha g^{\min}_{\check{\boldsymbol{x}}}+\beta)}{\mu_{\min}s(\check{\boldsymbol{x}})-(3+\varepsilon\|\boldsymbol{\mu}\|+\delta\mu_{\min})g^{\min}_{\check{\boldsymbol{x}}}} \quad (13)$$

then every optimal solution $(\overset{*}{\boldsymbol{x}},\overset{*}{\boldsymbol{a}})$ for the penalized relaxation problem (6a) – (6c) *satisfies* (3c) *and therefore $\overset{*}{\boldsymbol{x}}\in\mathcal{F}$.*

Proof. See Section IV for the proof. □

Motivated by Theorems 1 and 2, a question arises as to whether we can solve a sequence of convex relaxations (6a) – (6c) to recover a high quality solution to (3a) – (3c). In response to this question, we propose Algorithm (1) which starts from an initial point and solves a sequence of penalized relaxations. If feasibility is attained, we theoretically prove that the algorithm can provide a convergent sequence of feasible points whose objective values monotonically improve. Theorem 3 investigates conditions under which Algorithm 1 converges to at least a locally optimal solution to (3a) – (3c).

Theorem 3. *Let $\check{\mathcal{F}}\triangleq\{\boldsymbol{x}\in\mathcal{F}\,|\,f(\boldsymbol{x})\le\tilde{c}\}$ denote an epigraph of the problem* (3a) – (3c) *through which, the functions $\{u_k\}_{k=0}^m$ and $\{v_k\}_{k=0}^m$ are twice continuously differentiable, and also $s(\boldsymbol{x})\ge\tilde{s}>0$ and $\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2<\tilde{d}$ for every $\boldsymbol{x}\in\check{\mathcal{F}}$. If $\eta>\max\{\beta,\mu_1^{-1}\}$ and*

$$\eta>\alpha+\frac{\tilde{d}+(3+\varepsilon\|\boldsymbol{\mu}\|+\delta\mu_{\min})\beta}{\mu_{\min}\tilde{s}} \quad (14)$$

then the sequence generated by Algorithm 1 converges to a local minimizer of the problem (3a) – (3c)*.*

Proof. See Section IV for the proof. □

III. Applications

In this section, we experimentally evaluate the efficacy of our approach, referred to as SCR, in solving an example of the form (1a) – (1b).

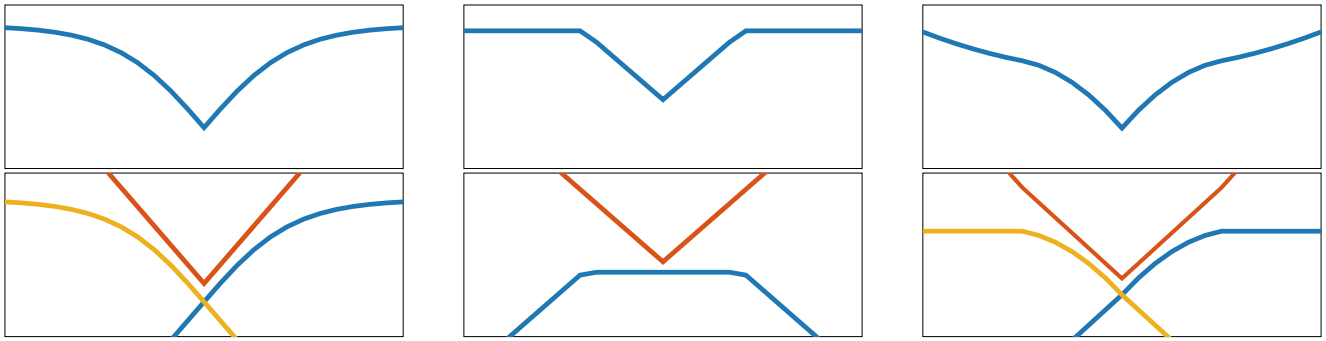

Fig. 1: Decomposition of non-convex regularization functions. The top row shows (**left**) SCAD, (**middle**) Capped ℓ_1-norm, (**right**) Hard-ridge regularization functions and the bottom row shows the corresponding convex/concave decompositions.

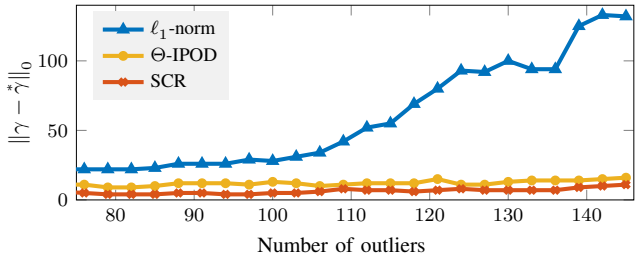

Fig. 2: Outlier identification result. Both SCR and Θ-IPOD [5] use SCAD regularization function.

A. Nonconvex Penalized Regression

Let $\{y_i\}_{i=1}^n$ be a set of samples randomly drawn from a low-dimensional subspace which is spanned by the columns of a matrix $\boldsymbol{A}\in\mathbb{R}^{n\times q}$, and contaminated with unknown outlying entries $\{\gamma_i\}_{i=1}^n$ and unknown noise values $\{\zeta_i\}_{i=1}^n$. Given that, we model observation the vector $\boldsymbol{y}\triangleq[y_1,\ldots,y_n]^\top$ as

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{z}+\boldsymbol{\gamma}+\zeta, \tag{15}$$

where $\boldsymbol{z}\in\mathbb{R}^q$ is an unknown regression vector, $\boldsymbol{\gamma}=[\gamma_1,\ldots,\gamma_n]^\top$ is a sparse outlier vector with arbitrary (possibly large) entries, and vector $\zeta=[\zeta_1,\ldots,\zeta_n]^\top$ corresponds to noise values sampled independently from a Gaussian distribution. The linear model (15) is well-studied in the literature [22], [23], [5], [24], where the main focus has been on estimating the model parameters while limiting the influence of outliers. To develop such regression model, consider the following optimization problem

$$\underset{\boldsymbol{\gamma},\zeta\in\mathbb{R}^n,\boldsymbol{z}\in\mathbb{R}^q}{\text{minimize}} \quad \|\zeta\|_2^2+\bar{\eta}\,l(\boldsymbol{\gamma}) \tag{16a}$$

$$\text{subject to} \quad \boldsymbol{y}=\boldsymbol{A}\boldsymbol{z}+\boldsymbol{\gamma}+\zeta, \tag{16b}$$

where ℓ_2-norm enforces the noise components to be small, function $l:\mathbb{R}^n\to\mathbb{R}$, defined as $l(\boldsymbol{\gamma})=\sum_{i=1}^n r(\gamma_i)$, imposes sparsity on vector $\boldsymbol{\gamma}$, $r:\mathbb{R}\to\mathbb{R}$ is a Hessian bounded element-wise regularization function, and parameter $\bar{\eta}>0$ sets a trade-off between the objective function and the regularization term. A natural choice for sparsity-promoting function l is the ℓ_0-norm which simply counts the number of outliers. However, this renders (16a)–(16b) computationally intractable [25]. To circumvent this drawback, ℓ_0-norm is often substituted by its convex surrogates (e.g. ℓ_1-norm). Despite the computational advantages offered by these functions, they mostly exhibits poor performance in the presence of multiple outliers [5]. To promote the robustness, one can adopt nonconvex regularization functions such as smoothly clipped absolute deviation (SCAD) [26], Capped ℓ_1-norm, Hard-Ridge function, etc. Next, we review some well-known nonconvex regularization functions and their convex decompositions.

a) SCAD: is a nonconvex quadratic spline function with application in outlier detection [5], [27], [28]. The SCAD regularization function can be decomposed as $r_{v\mathrm{S}}(\gamma)=b|\gamma|-g_{v\mathrm{S}}(\gamma)-g_{v\mathrm{S}}(-\gamma)$, where functions $r_{v\mathrm{S}}:\mathbb{R}\to\mathbb{R}$ and $g_{v\mathrm{S}}:\mathbb{R}\to\mathbb{R}$ are given by

$$r_{v\mathrm{S}}(\gamma)\triangleq\begin{cases}b|\gamma| & |\gamma|<b\\ \frac{|\gamma|^2-2ab|\gamma|+b^2}{2(1-a)} & b\le|\gamma|\le ab\\ \frac{(a+1)b^2}{2} & |\gamma|>ab\end{cases},$$

$$g_{v\mathrm{S}}(\gamma)\triangleq\begin{cases}-b\gamma & \gamma\le b\\ \frac{\gamma^2-2ab\gamma+b^2}{2(a-1)} & b<\gamma\le ab,\\ \frac{-(a+1)b^2}{2} & \gamma>ab\end{cases}$$

and parameters $a>2$ and b determine the shape of the function.

b) Capped ℓ_1-norm: is a sparsity-inducing regularization function which is widely used in various applications such as outlier detection, dictionary learning [29], Multi-task sparse feature learning [30], [31], [32], etc. This function is defined as $r_{v\mathrm{C}}(\gamma)=\min(|\gamma|,b)$ where parameter $b>0$ determines the maximum loss generated by the Capped ℓ_1-norm. Observed that this regularization function can be decomposed as $r_{v\mathrm{C}}(\gamma)=|\gamma|-\max(|\gamma|-b,0)$ where both functions $|\gamma|$ and $\max(|\gamma|-b,0)$ are convex.

c) Hard-ridge: function [33], [5], [34] enforces sparsity by hybridizing the hard-penalty [33] and ℓ_2-norm penalty. The function is given by

$$r_{v\mathrm{H}}(\gamma)\triangleq\begin{cases}-\frac{1}{2}\gamma^2+b|\gamma| & |\gamma|\le\frac{b}{1+a}\\ \frac{a}{2}\gamma^2+\frac{b^2}{2(1+a)} & |\gamma|>\frac{b}{1+a}\end{cases}.$$

The DC decomposition of the Hard-ridge function is $r_{v\mathrm{H}}(\gamma)=g_{v\mathrm{H}}(\gamma)-h_{v\mathrm{H}}(\gamma)-h_{v\mathrm{H}}(-\gamma)$, where convex func-

tions $g_{v\mathrm{H}}:\mathbb{R}\to\mathbb{R}$ and $h_{v\mathrm{H}}:\mathbb{R}\to\mathbb{R}$ are defined as

$$g_{v\mathrm{H}}(\gamma)\triangleq\begin{cases}\frac{b}{1+a}|\gamma| & |\gamma|\leq\frac{b}{1+a}\\ \frac{a}{2}\gamma^2+\frac{b}{1+a}|\gamma|-\frac{ab^2}{2(1+a)^2} & |\gamma|>\frac{b}{1+a}\end{cases},$$

$$h_{v\mathrm{H}}(\gamma)\triangleq\begin{cases}-\frac{b}{1+a}\gamma & \gamma\leq 0\\ \frac{1}{2}\gamma^2-b\gamma & 0\leq\gamma\leq\frac{b}{1+a}\\ \frac{b^2}{2(1+a)^2}-\frac{b^2}{1+a} & \gamma\geq\frac{b}{1+a}\end{cases}.$$

Figure 1 depicts the convex/concave decomposition of the above-mentioned regularization functions. Observe that optimization problem (16a)–(16b) with either of the above mentioned regularization functions, can be cast as a special case of (1a) – (1b). We reuse the experimental design proposed in [5] to verify the effectiveness of Algorithm 1 in solving (16a)–(16b). We generate the observations according to linear model (15) for $n=300$ and $m=70$; matrix $\boldsymbol{A}$ is set to $\boldsymbol{A}=\boldsymbol{V}\boldsymbol{S}^{\frac{1}{2}}$ where $V_{ij}\overset{\text{iid}}{\sim}U(-15,15)$ and $S_{ij}=\rho^{1_{i\neq j}}$ with $\rho=0.5$; the nonzero outlying entries of vector $\boldsymbol{\gamma}$ are sampled independently from $U(-2\tau,-\tau)+U(\tau,2\tau)$ with $\tau=5$, and $\zeta\overset{\text{iid}}{\sim}\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. We compare the outlier identification result of SCR with that of the ℓ_1-norm penalized problem and Θ-IPOD [5]. Both SCR and Θ-IPOD use SCAD regularization function. Figure 2 plots the results across various number of outliers averaged from 10 independent runs with random initialization. It can be observed that the proposed approach with minimal tuning performs on par or better than Θ-IPOD.

IV. Proofs

This section presents the proofs for the theoretical results in the paper. We first provide a number of prerequisite lemmas and definitions. Consider the following auxiliary optimization problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\quad u_0(\boldsymbol{x})-v_0(\boldsymbol{x})+\eta\times g_{\check{\boldsymbol{x}}}(\boldsymbol{x},\boldsymbol{v}(\boldsymbol{x}))^2 \tag{17a}$$

$$\text{subject to}\quad u_k(\boldsymbol{x})-v_k(\boldsymbol{x})\leq 0\qquad k\in\mathcal{M} \tag{17b}$$

The subsequent lemma states that for appropriate choices of η, the distance between the optimal solution of (17a) – (17b) and the initial point $\check{\boldsymbol{x}}$ can be bounded.

Lemma 1. *If $\eta>\alpha$, then the problem* (17a) – (17b) *has an attainable optimal value and every optimal solution $\overset{*}{\boldsymbol{x}}$ satisfies,*

$$0\leq g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))-g_{\check{\boldsymbol{x}}}^{\min}\leq\frac{\alpha g_{\check{\boldsymbol{x}}}^{\min}+\beta}{\eta-\alpha} \tag{18}$$

where $g_{\check{\boldsymbol{x}}}^{\min}$ is defined by the equation (9)*.*

Proof. According to Assumption 1, the objective function (17a) is bounded. Hence, attainability of the optimal solution is an immediate consequence of the fact that $\mathcal{F}$ is closed and nonempty.

The left side equality is followed by the definition of $g_{\check{\boldsymbol{x}}}^{\min}$. To prove the upper bound, let $\bar{\boldsymbol{x}}\in\mathbb{R}^n$ be an arbitrary member of $\{\boldsymbol{x}\in\mathcal{F}\,|\,g_{\check{\boldsymbol{x}}}(\boldsymbol{x},\boldsymbol{v}(\boldsymbol{x}))=g_{\check{\boldsymbol{x}}}^{\min}\}$. Since $\overset{*}{\boldsymbol{x}}$ is a solution to the problem (17a) – (17b) and $\bar{\boldsymbol{x}}$ is feasible, one can write:

$$\begin{aligned}&(\eta-\alpha)g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))^2-\beta g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))\\ &\leq -\boldsymbol{u}_0(\check{\boldsymbol{x}})+\boldsymbol{v}_0(\check{\boldsymbol{x}})+\boldsymbol{u}_0(\overset{*}{\boldsymbol{x}})-\boldsymbol{v}_0(\overset{*}{\boldsymbol{x}})+\eta g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))^2 &\text{(19a)}\\ &\leq -\boldsymbol{u}_0(\check{\boldsymbol{x}})+\boldsymbol{v}_0(\check{\boldsymbol{x}})+\boldsymbol{u}_0(\bar{\boldsymbol{x}})-\boldsymbol{v}_0(\bar{\boldsymbol{x}})+\eta g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}}))^2 &\text{(19b)}\\ &\leq(\eta+\alpha)g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}}))^2+\beta g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}})) &\text{(19c)}\end{aligned}$$

where (19a) and (19c) are concluded from (7a) and (19b) is concluded from the optimality of $\overset{*}{\boldsymbol{x}}$. Hence:

$$\begin{aligned}&\Rightarrow(\eta-\alpha)(g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))^2-g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}}))^2)\\ &\quad\leq 2\alpha g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}}))^2+\beta g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))+\beta g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}})) &\text{(20)}\\ &\Rightarrow(\eta-\alpha)(g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))-g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}})))\\ &\quad\leq\frac{2\alpha g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}}))^2}{g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))+g_{\check{\boldsymbol{x}}}(\bar{\boldsymbol{x}},\boldsymbol{v}(\bar{\boldsymbol{x}}))}+\beta &\text{(21)}\end{aligned}$$

which concludes the upper bound. □

The next lemma shows that if $\check{\boldsymbol{x}}$ is nonsingular and η is sufficiently large, then $\overset{*}{\boldsymbol{x}}$ is nonsingular as well.

Lemma 2. *Every optimal solution $\overset{*}{\boldsymbol{x}}$ to the problem* (17a) – (17b) *satisfies:*

$$s(\overset{*}{\boldsymbol{x}})\geq s(\check{\boldsymbol{x}})-\delta\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2 \tag{22}$$

Proof. The proof follows directly from the definition of s:

$$\begin{aligned}s(\overset{*}{\boldsymbol{x}})&=\min\{\|\boldsymbol{J}(\overset{*}{\boldsymbol{x}})^\top\boldsymbol{\nu}\|_2\,|\,\|\boldsymbol{\nu}\|_2=1\}\\ &\geq\min\{\|\boldsymbol{J}(\check{\boldsymbol{x}})^\top\boldsymbol{\nu}\|_2-\|[\boldsymbol{J}(\check{\boldsymbol{x}})-\boldsymbol{J}(\overset{*}{\boldsymbol{x}})]^\top\boldsymbol{\nu}\|_2\,|\,\|\boldsymbol{\nu}\|_2=1\}\\ &\geq\min\{\|\boldsymbol{J}(\check{\boldsymbol{x}})^\top\boldsymbol{\nu}\|_2\,|\,\|\boldsymbol{\nu}\|_2=1\}-\|\boldsymbol{J}(\check{\boldsymbol{x}})-\boldsymbol{J}(\overset{*}{\boldsymbol{x}})\|_2\\ &\geq s(\check{\boldsymbol{x}})-\delta\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2.\end{aligned} \tag{23}$$

As a consequence, if $\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2$ is small, and $\check{\boldsymbol{x}}$ is nonsingular, then $\overset{*}{\boldsymbol{x}}$ is nonsingular as well. □

Lemma 3. *Let $\check{\boldsymbol{x}}\in\mathbb{R}^n$ satisfy* (12)*. If $\eta>\max\{\beta\}$ and* (13) *is satisfied, then for every optimal solution $\overset{*}{\boldsymbol{x}}$ of the problem* (17a) – (17b)*, there exists a vector of Lagrange multipliers $\overset{*}{\boldsymbol{\lambda}}\in\mathbb{R}^m$ associated with the constraint* (17b)*, which satisfies:*

$$\|\eta^{-1}\overset{*}{\boldsymbol{\lambda}}\|_2<\mu_{\min}. \tag{24}$$

Proof. Let $\overset{*}{\boldsymbol{x}}$ denote an optimal point of the auxiliary problem (17a) – (17b). According to the assumption (12) and Lemma 2, we have:

$$\begin{aligned}s(\overset{*}{\boldsymbol{x}})&\geq s(\check{\boldsymbol{x}})-\delta\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2\\ &\geq s(\check{\boldsymbol{x}})-\delta\left(g_{\check{\boldsymbol{x}}}^{\min}+\frac{\alpha g_{\check{\boldsymbol{x}}}^{\min}+\beta}{\eta-\alpha}\right)>0.\end{aligned} \tag{25}$$

Hence, $\overset{*}{\boldsymbol{x}}$ is nonsingular and satisfies the linear independence constraint qualification condition. Therefore, there exists a vector of Lagrange multipliers $\overset{*}{\boldsymbol{\lambda}}\in\mathbb{R}^m$ that satisfies the following stationarity condition:

$$\nabla_{\boldsymbol{x}}f(\overset{*}{\boldsymbol{x}})+2\eta(\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}})+\eta(\boldsymbol{V}(\overset{*}{\boldsymbol{x}})-\boldsymbol{V}(\check{\boldsymbol{x}}))^\top\boldsymbol{\mu}+\boldsymbol{J}(\overset{*}{\boldsymbol{x}})^\top\overset{*}{\boldsymbol{\lambda}}=0 \tag{26}$$

Now one can upper bound $\|\eta^{-1}\overset{*}{\boldsymbol{\lambda}}\|_2$ as follows:

$$\begin{aligned}\|\eta^{-1}\overset{*}{\boldsymbol{\lambda}}\|_2&\leq\frac{\|\eta^{-1}\nabla_{\boldsymbol{x}}f(\overset{*}{\boldsymbol{x}})+2(\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}})+(\boldsymbol{V}(\overset{*}{\boldsymbol{x}})-\boldsymbol{V}(\check{\boldsymbol{x}}))^\top\boldsymbol{\mu}\|_2}{s(\overset{*}{\boldsymbol{x}})}\\ &\leq\frac{\eta^{-1}\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2+(\eta^{-1}\beta+2+\varepsilon\|\boldsymbol{\mu}\|)\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2}{s(\overset{*}{\boldsymbol{x}})}\\ &\leq\frac{\eta^{-1}\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2+(\eta^{-1}\beta+2+\varepsilon\|\boldsymbol{\mu}\|)\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2}{s(\check{\boldsymbol{x}})-\delta\|\overset{*}{\boldsymbol{x}}-\check{\boldsymbol{x}}\|_2}\\ &\leq\frac{\eta^{-1}\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2+(\eta^{-1}\beta+2+\varepsilon\|\boldsymbol{\mu}\|)g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))}{s(\check{\boldsymbol{x}})-\delta g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))}\\ &\leq\frac{(\eta-\alpha)^{-1}\|\nabla_{\boldsymbol{x}}f(\check{\boldsymbol{x}})\|_2+(3+\varepsilon\|\boldsymbol{\mu}\|)g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))}{s(\check{\boldsymbol{x}})-\delta g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}},\boldsymbol{v}(\overset{*}{\boldsymbol{x}}))}\end{aligned}$$

where the second and third inequalities are concluded from (8b) and Lemma 2, respectively. Now, according to Lemma 1 and the assumption (13) we have $\|\eta^{-1}\overset{*}{\boldsymbol{\lambda}}\|_2 < \mu_{\min}$. □

Proof of Theorem 2. Let $\overset{*}{\boldsymbol{x}}$ denote an optimal point of the auxiliary problem (17a) – (17b). According to Lemma 3, there exists a vector of Lagrange multipliers $\overset{*}{\boldsymbol{\lambda}} \in \mathbb{R}^m$ that satisfies (24). Now, the following serves as a primal and dual solution for the problem (6a) – (6c):

$$\boldsymbol{x} = \overset{*}{\boldsymbol{x}} \qquad \boldsymbol{a} = \boldsymbol{v}(\overset{*}{\boldsymbol{x}}) \tag{28a}$$

$$\boldsymbol{\lambda} = \overset{*}{\boldsymbol{\lambda}} \qquad \boldsymbol{\nu} = [1, \overset{*}{\boldsymbol{\lambda}}^\top]^\top - \eta\boldsymbol{\mu} < 0 \tag{28b}$$

where $\boldsymbol{\nu}$ represents the vector of Lagrange multipliers associated with (6c). Since $\boldsymbol{\nu} < 0$, the relaxation is lossless, which completes the proof. □

Proof of Theorem 1. The first part is an immediate consequence of Theorem 2 and the fact that $g^{\min}_{\check{\boldsymbol{x}}} = 0$ for every $\check{\boldsymbol{x}} \in \mathcal{F}$. Now, according to optimality of $\overset{*}{\boldsymbol{x}}$ and feasibility of $\check{\boldsymbol{x}}$, we have:

$$f(\overset{*}{\boldsymbol{x}}) + \eta \times g_{\check{\boldsymbol{x}}}(\overset{*}{\boldsymbol{x}}, \boldsymbol{v}(\overset{*}{\boldsymbol{x}}))^2 \leq f(\check{\boldsymbol{x}}) \tag{29}$$

which concludes that $f(\overset{*}{\boldsymbol{x}}) \leq f(\check{\boldsymbol{x}})$. □

Lemma 4. *Define $h_\eta : \mathcal{F} \to \mathcal{F}$ as a mapping that associates an initial point $\check{\boldsymbol{x}}$ of the problem* (17a) – (17b) *to its primal solutions. Consider an arbitrary $\check{\boldsymbol{x}} \in \mathcal{F}$ that satisfies* (12) *and assume that the functions $\{u_k\}_{k=0}^m$ and $\{v_k\}_{k=0}^m$ are twice continuously differentiable at every optimal point of the problem* (17a) – (17b)*. If $\eta > \max\{\beta, \mu_1^{-1}\}$ and* (13) *is satisfied then h_η is a continuous function at a vicinity of $\check{\boldsymbol{x}}$.*

Proof. Let $\overset{*}{\boldsymbol{x}} \in \mathcal{F}$ denote an optimal solution for the problem (17a) – (17b). According to Lemma 3 there exists a vector of Lagrange multipliers $\overset{*}{\boldsymbol{\lambda}} \in \mathbb{R}^m$ that satisfies the following KKT conditions:

$$\nabla_{\boldsymbol{x}} f(\overset{*}{\boldsymbol{x}}) + 2\eta\overset{*}{\boldsymbol{x}} + \eta\boldsymbol{V}(\overset{*}{\boldsymbol{x}})^\top\boldsymbol{\mu} + \boldsymbol{J}_{\mathcal{B}}(\overset{*}{\boldsymbol{x}})^\top\overset{*}{\boldsymbol{\lambda}}_{\mathcal{B}} = \eta(2\check{\boldsymbol{x}} + \boldsymbol{V}(\check{\boldsymbol{x}})^\top\mu)$$
$$u_k(\boldsymbol{x}) - v_k(\boldsymbol{x}) = 0 \qquad k \in \mathcal{B}$$

where $\mathcal{B} \subseteq \mathcal{M}$ denotes the set of binding constraints for $\overset{*}{\boldsymbol{x}}$. The Jacobian matrix of the above equations is

$$\begin{bmatrix} H^f_{\boldsymbol{x}}(\overset{*}{\boldsymbol{x}}) + 2\eta\boldsymbol{I} + \eta\sum_{k=0}^m \mu_{k+1} H^{v_k}_{\boldsymbol{x}}(\overset{*}{\boldsymbol{x}}) + \sum_{k\in\mathcal{B}} \lambda_k H^{f_k}_{\boldsymbol{x}}(\overset{*}{\boldsymbol{x}}) & \boldsymbol{J}_{\mathcal{B}}(\overset{*}{\boldsymbol{x}})^\top \\ \boldsymbol{J}_{\mathcal{B}}(\overset{*}{\boldsymbol{x}}) & \boldsymbol{0} \end{bmatrix}$$

which is nonsingular due to nonsingularity of $\boldsymbol{J}_{\mathcal{B}}(\overset{*}{\boldsymbol{x}})$ and the inequality (24). Therefore, h_η is a differentiable function. □

Proof of Theorem 3. Let $\{\boldsymbol{x}^k\}_{k=0}^\infty$ denote the sequence generated by Algorithm 1, where $\boldsymbol{x}^0 = \check{\boldsymbol{x}}$. Since $\boldsymbol{x}^0 \in \check{\mathcal{F}}$, then according to Theorem 1 and the assumption (14), every member of $\{\boldsymbol{x}^k\}_{k=0}^\infty$ belongs to $\check{\mathcal{F}}$ as well and the sequence $\{f(\boldsymbol{x}^k)\}_{k=0}^\infty$ is non-increasing and therefore convergent. On the other hand, for every nonnegative integer k, due to optimality of $\boldsymbol{x}^k$ and feasibility of $\boldsymbol{x}^{k-1}$, we have

$$f(\boldsymbol{x}^k) + \eta \times g_{\boldsymbol{x}^{k-1}}(\boldsymbol{x}^k, \boldsymbol{v}(\boldsymbol{x}^k))^2 \leq f(\boldsymbol{x}^{k-1}) \tag{30}$$

which implies that the sequence $\{g_{\boldsymbol{x}^{k-1}}(\boldsymbol{x}^k, \boldsymbol{v}(\boldsymbol{x}^k))\}_{k=1}^\infty$ converges to zero. Hence, the sequence $\{\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\|_2\}_{k=1}^\infty$ converges to zero and therefore, due to closeness of $\check{\mathcal{F}}$, the sequence $\{\boldsymbol{x}^k\}_{k=0}^\infty$ is convergent to

$$\boldsymbol{x}^\infty \triangleq \lim_{k\to\infty} \boldsymbol{x}^k \in \check{\mathcal{F}}. \tag{31}$$

According to Lemma 4, we have $h_\eta(\boldsymbol{x}^\infty) = \boldsymbol{x}^\infty$ and since $s(\boldsymbol{x}^\infty) > 0$, there exists a vector of Lagrange multipliers $\boldsymbol{\lambda}^\infty \in \mathbb{R}^m$ associated with the constraint (17b), which satisfies the following stationarity condition:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^\infty) + 2\eta(\boldsymbol{x}^\infty - \boldsymbol{x}^\infty) +$$
$$\eta(\boldsymbol{V}(\boldsymbol{x}^\infty) - \boldsymbol{V}(\boldsymbol{x}^\infty))^\top\boldsymbol{\mu} + \boldsymbol{J}(\boldsymbol{x}^\infty)^\top\boldsymbol{\lambda}^\infty = 0 \tag{32}$$

Hence, we have

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^\infty) + \boldsymbol{J}(\boldsymbol{x}^\infty)^\top\boldsymbol{\lambda}^\infty = 0 \tag{33}$$

which is the stationarity condition for the problem (17a) – (17b). As a result, the pair $(\boldsymbol{x}^\infty, \boldsymbol{\lambda}^\infty)$ is primal and dual optimal for the problem (17a) – (17b) and the proof is complete. □

V. Conclusion

This paper presented a sequential convex programming approach which leverages the difference of convex programming (DCP) technique to convexify a wide range of nonconvex optimization problems. We theoretically proved that under certain assumptions, solving a sequence of the convex programs provides a convergent sequence of points which are feasible for the original problem. The proposed method improves upon the existing sequential methods by reaching feasibility within a finite number of steps rather than asymptotic feasibility. We conducted experiments on a machine learning application to demonstrate the effectiveness of our approach.

References

[1] Q. T. Dinh, S. Gumussoy, W. Michiels, and M. Diehl, "Combining convex–concave decompositions and linearization approaches for solving BMIs, with application to static output feedback," *IEEE Transactions on Automatic Control*, vol. 57, no. 6, pp. 1377–1390, 2012.

[2] J. Schulman, Y. Duan, J. Ho, A. Lee, I. Awwal, H. Bradlow, J. Pan, S. Patil, K. Goldberg, and P. Abbeel, "Motion planning with sequential convex optimization and convex collision checking," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1251–1270, 2014.

[3] R. Doelman and M. Verhaegen, "Sequential convex relaxation for convex optimization with bilinear matrix equalities," in *2016 European Control Conference (ECC)*. IEEE, 2016, pp. 1946–1951.

[4] I. Gannaz, "Robust estimation and wavelet thresholding in partially linear models," *Statistics and Computing*, vol. 17, no. 4, pp. 293–310, Dec 2007.

[5] Y. She and A. B. Owen, "Outlier detection using nonconvex penalized regression," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 626–639, 2011.

[6] Y. She, "An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors," *Computational Statistics & Data Analysis*, vol. 56, no. 10, pp. 2976–2990, 2012.

[7] Y. Sun, N. R. Zhang, A. B. Owen *et al.*, "Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data," *The Annals of Applied Statistics*, vol. 6, no. 4, pp. 1664–1688, 2012.

[8] N. Wang and D.-Y. Yeung, "Bayesian robust matrix factorization for image and video processing," in *ICCV*, 2013.

[9] Y. Fu, T. M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao, "Robust subjective visual property prediction from crowdsourced pairwise labels," *IEEE TPAMI*, vol. 38, no. 3, pp. 563–577, 2016.

[10] A. A. Ahmadi and G. Hall, "DC decomposition of nonconvex polynomials with algebraic techniques," *Mathematical Programming*, pp. 1–26, 2018.

[11] P. Hartman *et al.*, "On functions representable as a difference of convex functions." *Pacific Journal of Mathematics*, vol. 9, no. 3, pp. 707–713, 1959.

[12] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," in *Advances in neural information processing systems*, 2002, pp. 1033–1040.

[13] C. T. Lawrence and A. L. Tits, "A computationally efficient feasible sequential quadratic programming algorithm," *SIAM Journal on Optimization*, vol. 11, no. 4, pp. 1092–1118, 2001.

[14] J. Nocedal and S. J. Wright, *Nonlinear Equations*. Springer, 2006.

[15] Q. T. Dinh and M. Diehl, "Local convergence of sequential convex programming for nonconvex optimization," in *Recent Advances in Optimization and its Applications in Engineering*. Springer, 2010, pp. 93–102.

[16] T. D. Quoc and M. Diehl, "Sequential convex programming methods for solving nonlinear optimization problems with DC constraints," *arXiv preprint arXiv:1107.5841*, 2011.

[17] O. Mehanna, K. Huang, B. Gopalakrishnan, A. Konar, and N. D. Sidiropoulos, "Feasible point pursuit and successive approximation of non-convex QCQPs," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 804–808, 2015.

[18] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–Part I: Theory," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1929–1944, 2017.

[19] H. A. Le Thi, T. P. Dinh *et al.*, "DC programming and DCA for general DC programs," in *Advanced Computational Methods for Knowledge Engineering*. Springer, 2014, pp. 15–35.

[20] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, 2016.

[21] S. Diamond, R. Takapoui, and S. Boyd, "A general system for heuristic minimization of convex functions over non–convex sets," *Optimization Methods and Software*, vol. 33, no. 1, pp. 165–193, 2018.

[22] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[23] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John wiley & sons, 2005.

[24] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE TPAMI*, vol. 35, no. 11, pp. 2765–2781, 2013.

[25] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1-2, pp. 237–260, 1998.

[26] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[27] D. Kong, H. Bondell, and W. Shen, "Outlier detection and robust estimation in nonparametric regression," in *AISTATS*, 2018.

[28] A. Georgogiannis, "The generalization error of dictionary learning with Moreau envelopes," in *ICML*, 2018.

[29] T. Zhang, "Analysis of multi–stage convex relaxation for sparse regularization," *Journal of Machine Learning Research*, vol. 11, no. Mar, pp. 1081–1107, 2010.

[30] T. Zhang *et al.*, "Multi–stage convex relaxation for feature selection," *Bernoulli*, vol. 19, no. 5B, pp. 2277–2293, 2013.

[31] P. Gong, J. Ye, and C. Zhang, "Multi–stage multi–task feature learning," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2979–3010, 2013.

[32] W. Jiang, F. Nie, and H. Huang, "Robust dictionary learning with capped ℓ_1–norm," in *IJCAI*, 2015.

[33] Y. She *et al.*, "Thresholding–based iterative selection procedures for model selection and shrinkage," *Electronic Journal of statistics*, vol. 3, pp. 384–415, 2009.

[34] Y. She, J. Wang, H. Li, and D. Wu, "Group iterative spectrum thresholding for super–resolution sparse spectral selection," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6371–6386, 2013.