MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

DC Formulations and Algorithms for Sparse Optimization Problems

Jun-ya GOTOH, Akiko TAKEDA,
and Katsuya TONO

METR 2015–27 August 2015

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

DC Formulations and Algorithms for Sparse Optimization Problems

Jun-ya GOTOH

Department of Industrial and Systems Engineering,
Chuo University
jgoto@indsys.chuo-u.ac.jp

Akiko TAKEDA and Katsuya TONO

Department of Mathematical Informatics,
The University of Tokyo
takeda@mist.i.u-tokyo.ac.jp
katsuya_tono@mist.i.u-tokyo.ac.jp

August 2015

Abstract

In this paper a DC (Difference of two Convex functions) formulation approach for sparse optimization problems is proposed. First we provide an exact DC representation of the cardinality constraint by using the largest-k norm. Next we show exact penalties for quadratic minimization problems which often appear in practice. A DC Algorithm (DCA) is presented, where the dual step at each iteration can be efficiently carried out due to the accessible subgradient of the largest-k norm. Furthermore, we can solve each DCA subproblem in linear time via a soft thresholding operation if there are no additional constraints. The framework is extended to the rank-constrained problem as well as the cardinality- and the rank-minimization problems. Numerical experiments demonstrate the efficiency of the proposed DCA in comparison with existing methods which have other penalty terms.

1 Introduction

Optimization problems which seek sparsity of solutions have recently received broad attention. For example, it is often required to select a subset of informative variables in regression analysis [2, 6] and principal component analysis [54]; compressed sensing aims at a sparse representation of signal

or image data (e.g., [12, 17]); fund management requires a relatively small number of invested assets in a portfolio (e.g., [11, 44]); finding meaningful fragments in genes is vital in bioinformatics (e.g., [41]).

Such types of problems can be cast as an optimization problem having a nonconvex constraint called the *cardinality constraint*:

$$\|\boldsymbol{w}\|_0 := |\{i \in \{1, \ldots, n\} : w_i \neq 0\}| \leq K,$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $K \in \{1, \ldots, n\}$. By convention, we call $\|\boldsymbol{w}\|_0$ the ℓ_0-*norm* of $\boldsymbol{w}$. Due to the nonconvexity and discontinuity of the ℓ_0-norm, the resulting optimization problem is known to be intractable (see, e.g., [31] for the NP-hardness of the ℓ_0-minimization over a linear system). While deterministic global optimization algorithms (see, e.g., [23]) can be employed (or developed) to solve such a non-convex optimization problem, it is impractical to rigorously apply them because guaranteeing its global optimality is often unacceptably time-consuming except for small-scale instances; accordingly, local search algorithms based on relaxation and/or approximation via tractable convex optimization are more popular.

A common approach is to replace the ℓ_0-norm with the ℓ_1-norm, $\|\boldsymbol{w}\|_1 := \sum_{i=1}^n |w_i|$, as ℓ_1-norm is a tight convex relaxation of ℓ_0-norm. Especially in compressed sensing, popularity of the ℓ_1-relaxation is enormous due to the exact recovery property under some conditions [15]. A host of nonconvex approximation approaches based on the so-called *DC (Difference of two Convex functions)* formulations have also been studied (e.g., [21, 26, 53, 33]). Most of the existing approaches based on DC representations regard the ℓ_0-norm as a sum of indicator functions and replace each indicator function with a difference of two convex functions:

$$\|\boldsymbol{w}\|_0 = \sum_{i=1}^n 1_{\{w_i \neq 0\}} \approx \sum_{i=1}^n (d(w_i) - c(w_i)),$$

where $1_{\{\text{cond}\}}$ is the indicator function, i.e., it returns 1 if "cond" is true, and 0 otherwise and d and c are some continuous convex functions on $\mathbb{R}$ with their difference approximating the indicator function (see, e.g., Table 1 of [26] for examples of the pairs). Approaches in this direction have been applied to many applications, such as the sparse Fisher discriminant analysis [34, 26], feature selection in Support Vector Machines (SVMs) [10, 21], portfolio selection [53], and compressed sensing [33].

In this paper, a different approach to the cardinality-constrained problem is developed. More precisely, we propose an exact DC representation of the cardinality constraint, and employ the DC Algorithm (DCA) [38, 25, 37].

An exact DC representation has been proposed by [37], where the cardinality constraint is represented with 0-1 variables, as

$$\mathbf{1}^\top \boldsymbol{u} \leq K,\ \boldsymbol{u} \in \{0,1\}^n,\ |w_i| \leq M_j u_j,\ j = 1, \ldots, n, \tag{1}$$

by assuming that the so-called big-M constants, M_j, are available.

Our formulation is, however, different and advantageous in that while tightness of the resulting DC formulation based on (1) can be affected by the magnitude of the big-M constants,[1] our formulation and the behavior of DCA does not rely on such a big-M parameter.

In addition, our DC formulation is quite simple and easier to interpret. Actually, we rewrite the cardinality constraint as

$$\text{“the } (K+1)\text{-st largest absolute value component of } \boldsymbol{w}\text{”} = 0,$$

and represent this by using the difference of two norms, i.e.,

$$\|\boldsymbol{w}\|_{K+1} - \|\boldsymbol{w}\|_K = 0,$$

where $\|\boldsymbol{w}\|_K$ is the largest-k norm (or vector k-norm [52]), which is defined as the sum of the k largest absolute value elements of $\boldsymbol{w}$.[2] One advantage of the use of the largest-k norm representation is that its subgradient can be efficiently computed, which can make DCAs efficient.

More interestingly, we can replace the difference with that of the ℓ_1-norm and the largest-K norm, i.e., $\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K = 0$. This fact motivates us to develop a soft thresholding technique, which is popular in the context of proximal methods, and thus allows us to use a closed-form solution of the DCA subproblem.

Our DC approach can be also applied to matrix optimization problems with rank-constraint. Considering that the nuclear norm, the Ky Fan k norm, and the number of non-zero singular values of a matrix correspond to the ℓ_1-norm, the largest-k norm, and the ℓ_0-norm, respectively, of the singular values of the matrix, our DC approach can be straightforward applied to the rank-constrained problems, which have attracted huge attentions in the context of compressed sensing. Our DC-penalty formulation of the problem can be associated with the existing nuclear norm minimization (e.g., [48]), and provides an insight and interpretation.

On the other hand, our approach has some commonality with a couple of existing papers proposing non-convex penalties (or regularizers) in the context of the rank minimization. Hu et al. [24] addresses the matrix completion problem, which minimizes the rank of a matrix, by reformulating it to the minimization of the difference of two Ky Fan k norms (or the truncated nuclear norm in their term), and proposes to apply the Alternating Direction Method of Multipliers (ADMM), which requires to introduce additional

[1] When some state-of-the-art solver is employed, (1) should be represented in the so-called Specially Ordered Sets of Type 1 (SOS-1) and it does not have to bother about the big-M constants.

[2] While the largest-k norm has been anonymously known (see, e.g., exercise problems in [7, 9, 16]), it became popular in operations research community after [4] implicitly introduced it in the context of robust optimization.

variable updates. Likewise, using the Ky Fan k norm, [22] provides equivalent bilinear (matrix) inequality representations of the rank and cardinality constraints, but does not present algorithm or computational results. It is contrastive that our main concern is to associate those sparsity constraints with DC optimizations and to provide DCAs.

Contributions of this paper are summarized as follows:

- Exact DC reformulations of the cardinality constraint and the rank constraint are posed;
- Lower bounds of the exact penalty for the cardinality-constrained quadratic optimization problems which appear in regression and matrix completion are shown. Those results can be helpful in relating the sparsity-constrained formulations to the existing penalty methods;
- A proximal gradient technique is proposed so that the DCA can be efficient and treat large-scale problems with the help of a soft thresholding operation.

The remainder of this paper is structured as follows. The next section presents equivalent DC expressions for the cardinality constraint. In Section 3, we further reformulate the DC-constrained problem to a DC minimization problem and show exact penalty parameters, above which the two DC formulations become equivalent. A DCA framework and a soft thresholding operation are proposed in the second part of the section. Section 4 extends the results for the cardinality-constrained problems to the rank-constrained ones. An exact penalty is proved for the matrix completion problem. Section 5 extends to the cardinality and the rank minimizations and a matrix norm-constrained problem. Section 6 reports numerical experiments, demonstrating the behavior of the proposed DCAs in comparison with other existing methods which employ different penalty terms.

Notation. Lower-case bold type is reserved for representing vectors in $\mathbb{R}^n$, while upper-case bold type is for matrices. Especially, $\boldsymbol{I}$ and $\boldsymbol{O}$ denote the identity and zero matrices, respectively; $\mathbf{1}$ and $\mathbf{0}$ denote the vectors with all the elements being ones and zeros, respectively, while $\boldsymbol{e}_i$ denotes the column vector whose i-th element is 1 and 0 otherwise. The inner product of vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ are denoted by $\boldsymbol{a}^\top \boldsymbol{b}$ or $\boldsymbol{b}^\top \boldsymbol{a}$, while that of matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$ is denoted by $\boldsymbol{A} \bullet \boldsymbol{B} = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$ where $\mathrm{Tr}(\boldsymbol{C})$ denotes the trace of a square matrix $\boldsymbol{C}$. $\lambda_{\max}(\boldsymbol{Q})$ and $\lambda_{\min}(\boldsymbol{Q})$ denote the largest and smallest eigenvalues of $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, respectively. Frobenius norm of a matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ is defined by $\|\boldsymbol{W}\|_{\mathrm{F}} := \sqrt{\boldsymbol{W} \bullet \boldsymbol{W}}$, while the nuclear norm (also known as trace norm) is by $\|\boldsymbol{W}\|_* := \sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{W})$, where $\sigma_i(\boldsymbol{W})$ is the i-th largest singular value of $\boldsymbol{W}$, and the spectral norm is denoted by $\|\boldsymbol{W}\|_2 := \sqrt{\lambda_{\max}(\boldsymbol{W}^\top \boldsymbol{W})}$. By $\|\cdot\|_p$, we denote the ℓ_p-norm, i.e., $\|\boldsymbol{w}\|_p := (\sum_{i=1}^n |w_i|^p)^{1/p}$ for $p \in [1, \infty)$, and $\max_i\{|w_i|\}$ for $p = \infty$. $[x]^+ := \max\{x, 0\}$.

2 DC Formulations of Cardinality-constrained Problems

We consider a cardinality-constrained optimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{w}}{\text{minimize}} & f(\boldsymbol{w}) \\ \text{subject to} & \|\boldsymbol{w}\|_0 \leq K,\ \boldsymbol{w} \in S, \end{array} \tag{2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $S \subset \mathbb{R}^n$ is a closed convex set, and $K \in \{1, 2, \ldots, n\}$. Besides, we allow for nonconvexity of the objective function.

Assumption 1. *f is represented as a difference of two convex functions: $f(\boldsymbol{w}) = g(\boldsymbol{w}) - h(\boldsymbol{w})$, where $g, h : \mathbb{R}^n \to \mathbb{R}$ are closed and convex.*

While our aim is to show DC formulations and algorithms for tackling with (2), special attention will be paid to cases where the objective and constraints are given by quadratic (or linear) functions since they include many important applications listed below.

Example 1 (Subset selection in regression)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{b} \in \mathbb{R}^m$. When $f(\boldsymbol{w}) = g(\boldsymbol{w}) = \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}\|_2^2$ (i.e., $h(\boldsymbol{w}) = 0$), and $S = \mathbb{R}^n$, the problem (2) is least square estimation of a linear model equipped with variable selection [12, 6]. Some variations are found in [2, 28, 29].*

Example 2 (SVM with feature selection)**.** *Let $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, \ell\}$ be a given data set, where $\boldsymbol{x}_i \in \mathbb{R}^n$ denotes attributes of sample i and $y_i \in \{+1, -1\}$ is its label. An SVM equipped with a feature selection is formulated by setting $f(\boldsymbol{w}, b, \boldsymbol{z}) = g(\boldsymbol{w}, b, \boldsymbol{z}) = \mathbf{1}^\top \boldsymbol{z}$, $S = \{(\boldsymbol{w}, b, \boldsymbol{z}) : z_i \geq 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w} - b), \boldsymbol{z} \geq \mathbf{0}\}$. An ℓ_0 minimization version is in [10, 21].*

Example 3 (Sparse eigenvalue problems)**.** *Let $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ be a symmetric matrix such that $\lambda_{\min}(\boldsymbol{Q}) < 0$. When $f(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w}$ and $S = \{\boldsymbol{w} \in \mathbb{R}^n : \|\boldsymbol{w}\|_2^2 \leq 1\}$, the problem (2) is a problem seeking a sparse vector which approximates the smallest eigenvector. Note that f is DC-decomposable as $g(\boldsymbol{w}) = \boldsymbol{w}^\top(\boldsymbol{Q} - \lambda_{\min}(\boldsymbol{Q})\boldsymbol{I})\boldsymbol{w}$ and $h(\boldsymbol{w}) = -\lambda_{\min}(\boldsymbol{Q})\boldsymbol{w}^\top \boldsymbol{w}$. Especially when $\boldsymbol{Q}$ is negative semidefinite, it can be regarded as (a variant of) the sparse principal component analysis (PCA) [47]. Besides, for the classification problem as in Example 2, the sparse Fisher Linear Discriminant Analysis [30] can be formulated by setting $\boldsymbol{Q} = -\boldsymbol{V_1}$ and $S = \{\boldsymbol{w} : \boldsymbol{w}^\top \boldsymbol{V_2}\boldsymbol{w} \leq 1\}$ where $\boldsymbol{V_1}, \boldsymbol{V_2}$ are the between-class and within-class covariance matrices, respectively.*

Example 4 (Sparse portfolio selection)**.** *Let $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ be a covariance matrix, $\boldsymbol{r} \in \mathbb{R}^n$ a mean return vector, $\boldsymbol{l}, \boldsymbol{u} \in \mathbb{R}^n$, and $\tau \in \mathbb{R}$. When $f(\boldsymbol{w}) = g(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{V}\boldsymbol{w}$ (i.e., $h(\boldsymbol{w}) = 0$), and $S = \{\boldsymbol{w} \in \mathbb{R}^n : \boldsymbol{r}^\top \boldsymbol{w} \geq \tau, \mathbf{1}^\top \boldsymbol{w} = 1, \boldsymbol{l} \leq \boldsymbol{w} \leq \boldsymbol{u}\}$, the problem (2) is a sparse portfolio selection (see, e.g., [20, 44, 53] for its variations).*

While we allow for a nonconvex objective in the problem (2), our main concern in this paper is the nonconvexity stemming from the combinatorial nature of the cardinality constraint. To address that, we propose to rewrite the constraint with its exact DC representation by employing the largest-k norm.

Let us denote by $w_{(i)}$ the element whose absolute value is the i-th largest among the n elements of a vector $\boldsymbol{w} \in \mathbb{R}^n$, i.e., $|w_{(1)}| \geq |w_{(2)}| \geq \cdots \geq |w_{(n)}|$, while w_i indicates the i-th element of $\boldsymbol{w}$.

Definition 1. *For an integer $k \in \{1, \ldots, n\}$, the largest-k norm of $\boldsymbol{w} \in \mathbb{R}^n$, denoted by $\|\boldsymbol{w}\|_k$, is defined as the sum of the largest k components in absolute value. Namely,*

$$\|\boldsymbol{w}\|_k := |w_{(1)}| + |w_{(2)}| + \ldots + |w_{(k)}|.$$

Following [5, 36, 19], $\|\boldsymbol{w}\|_k$ can be extensively defined for non-integer $k \in [1, n]$,[3] but for simplicity, we consider only integers for k in the remainder of this paper.

With the norm, we can obtain simple, but key representations of the cardinality constraint of (2).

Theorem 1. *For any integers K, h such that $1 \leq K < h \leq n$, and $\boldsymbol{w} \in \mathbb{R}^n$, the following three conditions are equivalent:*

1. $\|\boldsymbol{w}\|_0 \leq K$,
2. $\|\boldsymbol{w}\|_h - \|\boldsymbol{w}\|_K = 0$, *and*
3. $\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K = 0$.

Furthermore, the following three conditions are equivalent:

4. $\|\boldsymbol{w}\|_0 = K$,
5. $K = \min\{k : \|\boldsymbol{w}\|_h - \|\boldsymbol{w}\|_k = 0\}$, *and*
6. $K = \min\{k : \|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_k = 0\}$.

Proof. By definition, $\|\boldsymbol{w}\|_h - \|\boldsymbol{w}\|_K$ is nonnegative and equals the sum of the $(K+1)$-st to h-th largest absolute value elements, i.e., $|w_{(K+1)}| + \cdots + |w_{(h)}|$, and for any $h \in \{K+1, \ldots, n\}$, the condition 2. ensures that at least $n - K$ elements are zero. This condition is equivalent to that the number of non-zero elements is no greater than K, i.e., $\|\boldsymbol{w}\|_0 \leq K$. The equivalence of the conditions 2. and 3. comes from the fact $\|\boldsymbol{w}\|_1 = \|\boldsymbol{w}\|_n$. The second part is shown by noting that $\|\boldsymbol{w}\|_0 = K$ if and only if $|w_{(K)}| > 0$ and $|w_{(K+1)}| = 0$. □

[3] In general, for $k \in [1, n]$, it is valid that $\|\boldsymbol{w}\|_k = \min_c \left\{kc + \sum_{i=1}^n [|w_i| - c]^+\right\}$, which can be further rewritten as a linear program and solved in time of order n. For further properties of the norm, see [5, 36, 19, 52].

Note that $\|\boldsymbol{w}\|_h - \|\boldsymbol{w}\|_K \geq 0$ holds for $K < h$, and that the conditions 2. and 5. are equivalent to $\|\boldsymbol{w}\|_h - \|\boldsymbol{w}\|_K \leq 0$ and $K = \min\{k : \|\boldsymbol{w}\|_h - \|\boldsymbol{w}\|_k \leq 0\}$, respectively. Likewise the equality signs in the conditions 3. and 6. can be replaced with inequality signs, $\leq$.

While Theorem 1 shares a base with that of [22], which also presents an equivalent expression of the cardinality constraint, the above equivalent relations are beneficial both for DC optimization approaches and connections to existing ℓ_1-relaxation methods, as will be shown.

By Theorem 1, (2) is rewritten in a DC form.

Corollary 1. *The cardinality-constrained problem* (2) *can be rewritten as a DC-constrained problem:*

$$\begin{array}{ll} \underset{\boldsymbol{w}\in S}{\textit{minimize}} & f(\boldsymbol{w}) \\ \textit{subject to} & \|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K = 0. \end{array} \tag{3}$$

Problem (3) is straightforward from the equivalence between the conditions 1. and 3. of Theorem 1. Although another DC-constrained formulation can be obtained by using the condition 2. of Theorem 1 with $h = K+1, \ldots, n-1$, we omit to present it in the remainder of this paper for simplicity.

One of the advantages of the formulation (3) over the existing DC approaches, such as those in [21, 53], is that (3) is an exact DC reformulation of the cardinality constraint (and, thus, more interpretable), while the others approximate the ℓ_0-norm with some functions, leading to pass over the discontinuity of the ℓ_0-norm. Moreover, (3) is advantageous over another exact DC formulation of [26] in that it does not need a big-M constant. On the other hand, (3) may look to have a disadvantage that it has a DC constraint, which we will address in the next section.

3 DC Algorithms for Cardinality-Constrained Problems

Associated with (3), we solve a penalized formulation:

$$f^* := \underset{\boldsymbol{w}\in S}{\text{minimize}} \quad f(\boldsymbol{w}) + \rho\left(\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K\right), \tag{4}$$

where ρ is a positive constant.

Since $\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K \geq 0$ for any $\boldsymbol{w} \in \mathbb{R}^n$, we may view that the added term plays a role of a penalty function of the cardinality constraint, $\|\boldsymbol{w}\|_0 \leq K$.

Conditions for exact penalty, under which a penalized DC problem, such as (4), is equivalent to a DC-constrained problem, such as (3), have been studied for a general class of problems (see, e.g., [39] for exact penalty for

DC optimization problems). Those conditions tend to be restrictive and inappropriate for general problems in practice.

Now we here analyze a case where the objective function is quadratic:

$$f(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} + \boldsymbol{q}^\top \boldsymbol{w}, \tag{5}$$

where $\boldsymbol{Q} = (q_{ij}) \in \mathbb{R}^{n\times n}$ is symmetric and $\boldsymbol{q} = (q_i) \in \mathbb{R}^n$.

Lemma 1. *Suppose that f is given by (5) and there exists a constant $C > 0$ such that $\|\boldsymbol{w}^\star\|_2 \le C$ for any optimal solution $\boldsymbol{w}^\star$ of (4). Then, the problems (3) and (4) are equivalent if*

$$\rho > \max_i \left\{ |q_i| + (\|\boldsymbol{Q}\boldsymbol{e}_i\|_2 + |q_{ii}|/2)C \right\}.$$

See Appendix A.1 for the proof of Lemma 1.

Based on Lemma 1, we can show the equivalence for a case where the sparse regression problem in Example 1 is included.

Theorem 2. *Suppose that $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is positive semi-definite and*

$$K < R := \max\{r : \textit{any } r \times r \textit{ principal minor of } \boldsymbol{Q} \textit{ is positive definite}\}.$$

Then the problems (3) and (4) are equivalent if

$$\rho > \max_i \left\{ |q_i| + \frac{(2\|\boldsymbol{Q}\boldsymbol{e}_i\|_2 + |q_{ii}|)\|\boldsymbol{q}\|_2}{\hat{\lambda}_R} \right\},$$

where $\hat{\lambda}_R := \min\{\lambda_{\min}([\boldsymbol{Q}]_I) : I \subset \{1,\dots,n\}, |I| = R\}$, and $[\boldsymbol{Q}]_I$ denotes the principal minor of $\boldsymbol{Q}$, corresponding to the index set I.

Note that $\hat{\lambda}_R$ equals $\lambda_{\min}(\boldsymbol{Q})$ if $\boldsymbol{Q}$ is positive definite. See Appendix A.2 for the proof of Theorem 2.

Lemma 1 directly shows an exact penalty for the sparse eigenvalue problems of Example 3 since $C = 1$ holds.

Corollary 2. *For the sparse eigenvalue problem, i.e., f is given in (5) and $S = \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \le 1\}$, the problems (3) and (4) are equivalent if*

$$\rho > \max_i \left\{ |q_i| + \|\boldsymbol{Q}\boldsymbol{e}_i\|_2 + |q_{ii}|/2 \right\}.$$

These results motivate us to solve (4), which is more tractable to apply DCA than (3). [4]

On the other hand, the bounds for exact penalty are often too conservative. In addition, a large value of the parameter ρ can induce computational

[4] Needless to say, if $\|\boldsymbol{w}^*\|_1 - \|\boldsymbol{w}^*\|_K > 0$ at an optimal solution $\boldsymbol{w}^*$ for such a sufficiently large ρ, the constrained problem (3) is proved to be infeasible.

instability during the DCA. To address this issue, updating the value of ρ is proposed by [37], where a sequence $\{\rho_t : \rho_0 < \rho_1 < \cdots\}$ is applied instead of a fixed value of ρ. Appropriate updating rules for the sequence $\{\rho_t\}$ ensure global convergence to a critical point of the problem (3).

Here let us consider the meaning of the penalized reformulation (4) again. An essential difference in (4) from the ℓ_1-approximation methods is the largest-K norm term "$-\rho\|\boldsymbol{w}\|_K$." Namely, our DC approach to the cardinality-constrained problem can be viewed as a modification of the ℓ_1-relaxation. By definition, the parameter K denotes the upper bound of the density of a vector, i.e., the number of the nonzero elements. The effect of the parameter K on the resulting sparseness will be numerically examined in Section 6.

The penalized formulation (4) can also be viewed as one of the non-convex regularization methods, which are recently attracting attentions (e.g., [18]). The equivalence between (3) and (4) provides a new clear-cut perspective on the cardinality constraint and non-convex penalty methods.

3.1 DC Algorithms for Problem with DC Objective

Next we consider a DCA to solve (4).

To that aim, first let us observe that the objective function of (4) is expressed by a difference of two convex functions, u and v, as follows:

$$\begin{aligned} f(\boldsymbol{w}) + \rho\left(\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K\right) &= \underbrace{g(\boldsymbol{w}) + \rho\|\boldsymbol{w}\|_1}_{\text{convex: } u(\boldsymbol{w})} - \underbrace{\left(h(\boldsymbol{w}) + \rho\|\boldsymbol{w}\|_K\right)}_{\text{convex: } v(\boldsymbol{w})}, \\ &= u(\boldsymbol{w}) - v(\boldsymbol{w}). \end{aligned}$$

At each iteration, DCA solves a convex subproblem, which is defined by linearizing the concave term $-v(\boldsymbol{w}) = -(h(\boldsymbol{w}) + \rho\|\boldsymbol{w}\|_K)$, and repeats until a convergence condition is fulfilled. More specifically, at the t-th iteration, DCA solves the following convex optimization problem:

$$f^t := \underset{\boldsymbol{w}\in S}{\text{minimize}} \quad g(\boldsymbol{w}) + \rho\|\boldsymbol{w}\|_1 - \boldsymbol{w}^\top \boldsymbol{g}_{\boldsymbol{w}}^{t-1}, \tag{6}$$

where $\boldsymbol{g}_{\boldsymbol{w}}^{t-1}$ is a subgradient of $v(\boldsymbol{w})$ at $\boldsymbol{w}^{t-1}$, i.e.,

$$\boldsymbol{g}_{\boldsymbol{w}}^{t-1} \in \partial v(\boldsymbol{w}^{t-1}) = \partial h(\boldsymbol{w}^{t-1}) + \rho\partial\|\boldsymbol{w}^{t-1}\|_K,$$

and provides an optimal solution of (6), $\boldsymbol{w}^t$. The entire picture of a generic DCA is described below as Algorithm 1.

Especially, if either the function u or v is polyhedral, the DCA is said to be *polyhedral* and guaranteed to terminate in finite iterations [38]. Note that the above DCA is *polyhedral* if h is polyhedral since the largest-K norm term

Algorithm 1 DC Algorithm (DCA) for DC objective problem

Require: $\boldsymbol{w}^0$, and a small value $\epsilon > 0$.
  $t = 1$.
  repeat
    Select a $\boldsymbol{g}_w^{t-1} \in \partial v(\boldsymbol{w}^{t-1})$. [dual step]
    Solve the convex subproblem (6). [primal step]
    Increment t.
  until $\left|f^{t-1} - f^t\right| < \epsilon$ holds.

$-\rho\|\boldsymbol{w}\|_K$ is expressed as a pointwise maximum of $2^K\binom{n}{K}$ linear functions, i.e.,

$$-\rho\|\boldsymbol{w}\|_K = -\max_{\boldsymbol{z}}\{\rho\boldsymbol{z}^\top\boldsymbol{w} : \boldsymbol{z} \in \{-1, 0, 1\}^n, \boldsymbol{z}^\top\boldsymbol{z} = K\}.$$

For the details of DCA convergence properties, see [38, 37].

Remark 1. *There is another approach for DC-constrained problems, which iteratively linearizes the nonconvex part of the DC constraint in the problem* (3) *until convergence (e.g., [42, 43, 53]). [37] discussed, however, a drawback of the approach. Indeed, applying the approach to the cardinality-constrained problem (3), the convex constraint,* $\|\boldsymbol{w}\|_1 - \boldsymbol{w}^\top\boldsymbol{q}^{t-1} \leq 0$*, can be overly conservative because* $0 \leq \|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K \leq \|\boldsymbol{w}\|_1 - \boldsymbol{w}^\top\boldsymbol{q}^{t-1}$*. Even if there exists a feasible solution for the original problem* (3) *satisfying* $\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K \leq 0$*, there is not necessarily a solution satisfying* $\|\boldsymbol{w}\|_1 - \boldsymbol{w}^\top\boldsymbol{q}^{t-1} \leq 0$*.*

3.2 Solving DCA Subproblems

The subdifferential of $\|\boldsymbol{w}\|_K$ at a point $\boldsymbol{w}^t$ is given in, e.g., [51, 49], as

$$\partial\|\boldsymbol{w}^t\|_K = \operatorname*{argmax}_{\boldsymbol{g}}\left\{\sum_{i=1}^n |w_i^t| g_i : \sum_{i=1}^n g_i = K,\ 0 \leq g_i \leq 1, i = 1, \ldots, n\right\} \quad (7)$$
$$= \{(g_1, \ldots, g_n) : g_{(1)} = \cdots = g_{(K)} = 1,\ g_{(K+1)} = \cdots = g_{(n)} = 0,\ \text{for } \boldsymbol{w}^t\},$$

where $g_{(i)}$ denotes the element of $\boldsymbol{g}$, corresponding to $w_{(i)}$ in the linear program (7). Note that a subgradient $\boldsymbol{g} \in \partial\|\boldsymbol{w}\|_K$ can be computed efficiently as follows: (i) sort the elements of $|\boldsymbol{w}|$ in decreasing order, i.e., $|w_{(1)}| \geq |w_{(2)}| \geq \cdots \geq |w_{(n)}|$; (ii) assign 1 to g_i which corresponds to $w_{(1)}, \ldots, w_{(K)}$. [5]

For example, let us consider the case where f is a convex quadratic function given by (5) and $S = \mathbb{R}^n$. At the $(t-1)$-st dual step of the DCA,

[5] Theoretically, the subgradient of the largest-k norm of $\boldsymbol{w} \in \mathbb{R}^n$ can be obtained in time of the order $O(n)$ by using a selection algorithm.

we pick a subgradient $\boldsymbol{g}_{\boldsymbol{w}}^{t-1}$ of $v(\boldsymbol{w}) = -\boldsymbol{q}^\top \boldsymbol{w} + \rho \|\boldsymbol{w}\|_K$[6] of the penalized objective, i.e.,

$$\boldsymbol{g}_{\boldsymbol{w}}^{t-1} \in \rho \cdot \{\boldsymbol{g} : g_{(1)} = \cdots = g_{(K)} = 1,\ g_{(K+1)} = \cdots = g_{(n)} = 0,\ \text{for } \boldsymbol{w}^{t-1}\} - \{\boldsymbol{q}\}.$$

Then at the $(t-1)$-st primal step, we solve the convex subproblem:

$$\boldsymbol{w}^t \in \operatorname*{argmin}_{\boldsymbol{w}} \left\{ \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{Q} \boldsymbol{w} + \rho \|\boldsymbol{w}\|_1 - \boldsymbol{w}^\top \boldsymbol{g}_{\boldsymbol{w}}^{t-1} \right\}. \tag{8}$$

More generally, as long as the subgradient of h at $\boldsymbol{w}^{t-1}$ is readily available, so is $\boldsymbol{g}_{\boldsymbol{w}}^{t-1}$. For example, if the subgradient of h at $\boldsymbol{w}$ can be given as $\boldsymbol{A}\boldsymbol{w}$ using a matrix $\boldsymbol{A}$,[7] $\boldsymbol{g}_{\boldsymbol{w}}^{t-1}$ is computed by

$$\boldsymbol{g}_{\boldsymbol{w}}^{t-1} \in \{\boldsymbol{A}\boldsymbol{w}^{t-1}\} + \rho \cdot \{\boldsymbol{g} : g_{(1)} = \cdots = g_{(K)} = 1,\ g_{(K+1)} = \cdots = g_{(n)} = 0,\ \text{for } \boldsymbol{w}^{t-1}\}.$$

3.3 Closed-form Solution via a Soft Thresholding

The proposed DCA needs to solve subproblems (6) at each iteration. Each subproblem differs only in the linear term, $-\boldsymbol{w}^\top \boldsymbol{g}_{\boldsymbol{w}}^{t-1}$, of its objective function. Therefore, we can solve them efficiently by using the incumbent (i.e., $(t-1)$-st) solution $\boldsymbol{w}^{t-1}$ as an initial solution of the t-th subproblem, or even using the solution information of $\boldsymbol{w}^{t-1}$ (e.g., basis information).

Recently some research papers (e.g., [18, 27]) criticize that DCA requires some other iterative algorithm to solve its subproblems, which can be computationally expensive for large-scale problems. For special types of nonconvex problems, however, closed-form solutions are readily available at each iteration. For example, for large-scale nonsmooth convex optimization problems having the ℓ_1-term, [3, 32] report that proximal gradient methods and accelerated proximal-gradient methods efficiently solve the problems by using proximal mappings to handle the nonsmooth part, which leads to a closed-form solution in each iteration; for the trust region subproblem, [45] derived a simple closed-form solution of its subproblems.

Proximal DC Decomposition Algorithm. We here derive a closed-form solution of DCA subproblems for the case where a quadratic function given by (5) is minimized only with the cardinality constraint, i.e., $S = \mathbb{R}^n$. Following [45], the function $f(\boldsymbol{w})$ is decomposed as $g(\boldsymbol{w}) - h(\boldsymbol{w})$, where

$$g(\boldsymbol{w}) = \frac{\bar{\lambda}}{2} \|\boldsymbol{w}\|_2^2, \quad h(\boldsymbol{w}) = \frac{1}{2} \boldsymbol{w}^\top (\bar{\lambda} \boldsymbol{I} - \boldsymbol{Q}) \boldsymbol{w} - \boldsymbol{q}^\top \boldsymbol{w},$$

[6] For notational convenience in the latter part, the linear term is included in the concave part $-v(\boldsymbol{w})$.

[7] Note that many applications, such as Examples 1 to 4, fulfill this.

where $\bar{\lambda}$ is a positive value defined by $\bar{\lambda} := [\lambda_{\max}(\boldsymbol{Q})]^+ + \epsilon$ with $\epsilon > 0$, so that it makes $\bar{\lambda}\boldsymbol{I} - \boldsymbol{Q}$ positive definite.[8] Then the problem (3) is equivalently rewritten as

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \left\{ \frac{\bar{\lambda}}{2}\|\boldsymbol{w}\|_2^2 - \frac{1}{2}\boldsymbol{w}^\top(\bar{\lambda}\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{w} + \boldsymbol{q}^\top\boldsymbol{w} + \rho(\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K) \right\}. \quad (9)$$

The DCA for (9) amounts to repeating the following procedure until convergence:

$$\boldsymbol{g}_{\boldsymbol{w}}^{t-1} \in \{(\bar{\lambda}\boldsymbol{I} - \boldsymbol{Q})\boldsymbol{w}^{t-1} - \boldsymbol{q}\} + \rho \cdot \operatorname*{argmax}_{\boldsymbol{g}} \left\{ \sum_{i=1}^n |w_i^{t-1}| g_i : \begin{array}{l} \mathbf{1}^\top\boldsymbol{g} = K, \\ \mathbf{0} \le \boldsymbol{g} \le \mathbf{1} \end{array} \right\},$$

$$\boldsymbol{w}^t \in \operatorname*{argmin}_{\boldsymbol{w}} \left\{ \frac{\bar{\lambda}}{2}\|\boldsymbol{w}\|_2^2 - \boldsymbol{w}^\top\boldsymbol{g}_{\boldsymbol{w}}^{t-1} + \rho\|\boldsymbol{w}\|_1 \right\}. \quad (10)$$

Note that the subproblem (10) is equivalent to

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{2}\left\| \boldsymbol{w} - \frac{\boldsymbol{g}_{\boldsymbol{w}}^{t-1}}{\bar{\lambda}} \right\|_2^2 + \frac{\rho}{\bar{\lambda}}\|\boldsymbol{w}\|_1 \right\},$$

and its optimal solution is explicitly given by

$$w_i^t = \begin{cases} \dfrac{g_{w,i}^{t-1} - \rho}{\bar{\lambda}} & (g_{w,i}^{t-1} \ge \rho), \\ 0 & (-\rho \le g_{w,i}^{t-1} \le \rho), \\ \dfrac{g_{w,i}^{t-1} + \rho}{\bar{\lambda}} & (g_{w,i}^{t-1} \le -\rho), \end{cases}$$

where $g_{w,i}^{t-1}$ denotes the i-th element of $\boldsymbol{g}_{\boldsymbol{w}}^{t-1}$. This type of operation is called *soft thresholding* in the context of proximal methods[9]. This gives a sparse solution $\boldsymbol{w}^t$ close to

$$\frac{\boldsymbol{g}_{\boldsymbol{w}}^{t-1}}{\bar{\lambda}} = \boldsymbol{w}^{t-1} - \frac{1}{\bar{\lambda}}(\boldsymbol{Q}\boldsymbol{w}^{t-1} + \boldsymbol{q}) + \frac{\rho}{\bar{\lambda}}\boldsymbol{g}^*,$$

[8] Note that we can assume a nonlinear twice continuously differentiable function with a Lipschitz constant L of $\nabla f(\boldsymbol{w})$ for feasible points $\boldsymbol{w}$ (i.e., upper bound on the maximum eigenvalue of the Hessian of $f(\boldsymbol{w})$ for feasible $\boldsymbol{w}$) and similarly solve the DCA subproblems using the decomposition with the convex function $\frac{L}{2}\|\boldsymbol{w}\|_2^2 - f(\boldsymbol{w})$ as $\frac{L}{2}\|\boldsymbol{w}\|_2^2 - (\frac{L}{2}\|\boldsymbol{w}\|_2^2 - f(\boldsymbol{w}))$, but to make the paper simple, we assume a quadratic function for $f(\boldsymbol{w})$.

[9] In general, the proximal mapping of a function $h : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{u} \in \mathbb{R}^n$ is defined as

$$\mathrm{prox}_h(\boldsymbol{u}) := \operatorname*{argmin}_{\boldsymbol{w}} \ \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{u}\|_2^2 + h(\boldsymbol{w}).$$

For $h(\boldsymbol{w}) = \beta\|\boldsymbol{w}\|_1$, each element of $\mathrm{prox}_h(\boldsymbol{u})$ is explicitly given by

$$\mathrm{prox}_h(\boldsymbol{u})_i = \begin{cases} u_i - \beta, & (u_i \ge \beta), \\ 0, & (-\beta \le u_i \le \beta), \\ u_i + \beta, & (u_i \le -\beta). \end{cases}$$

where $\boldsymbol{g}^* \in \operatorname{argmax}_{\boldsymbol{g}} \left\{ \sum_{i=1}^n |w_i^{t-1}| g_i : \mathbf{1}^\top \boldsymbol{g} = K,\ \mathbf{0} \le \boldsymbol{g} \le \mathbf{1} \right\}$. Here $\frac{1}{\lambda}$ can be regarded as a step size for the direction $-(\boldsymbol{Q}\boldsymbol{w}^{t-1} + \boldsymbol{q} - \rho \boldsymbol{g}^*)$.

Thus we can solve each subproblem (10) more efficiently, while the forced DC-decomposition, $f = h - g$, may destroy the polyhedrality of DCA.

When $\boldsymbol{Q}$ is positive semidefinite, i.e., f is convex, the DC-decomposition seems to be a redundant procedure. However, the decomposition generates a proximal term for the subproblem (6). Seemingly there must be a trade-off between the light computation and the number of iterations, which will be numerically examined in Section 6.

Note also that the above technique can be used for any differentiable objective function f, by replacing $\boldsymbol{Q}\boldsymbol{w}^{t-1} + \boldsymbol{q}$ with $\nabla f(\boldsymbol{w}^{t-1})$.

Remark 2. *[46] investigated this type of DC decomposition and called the resulting DCA the projection DC decomposition algorithm. While their concern is in a quadratic optimization, ours derives the soft thresholding with the help of the ℓ_1-norm term of the subproblem (10).*

4 DC Formulations and Algorithms for Rank-constrained Problems

In this section we extend the DC approach developed in the preceding sections to matrix optimization problems.

Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, and denote by $\operatorname{rank}(\boldsymbol{W})$ the rank of a matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$. A rank-constrained minimization of f is then formulated as

$$\begin{array}{ll} \underset{\boldsymbol{W}}{\text{minimize}} & f(\boldsymbol{W}) \\ \text{subject to} & \operatorname{rank}(\boldsymbol{W}) \le K,\ \boldsymbol{W} \in S, \end{array} \tag{11}$$

where $K \in \{1, \ldots, \min\{m, n\}\}$, and S is a closed convex set of matrices of size $m\times n$.

Example 5 (Matrix Completion)**.** *The matrix completion problem is to recover a data matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ from a sampling of its entries and often formulated as*

$$\begin{array}{ll} \textit{minimize}_{\boldsymbol{W}} & \operatorname{rank}(\boldsymbol{W}) \\ \textit{subject to} & W_{ij} = M_{ij}, \quad (i,j) \in \Omega, \end{array}$$

where Ω is the index set of known entries of $\boldsymbol{M}$. Replacing $\operatorname{rank}(\boldsymbol{W})$ *by the nuclear norm $\|\boldsymbol{W}\|_*$ brings a tight convex relaxation [14]. By using a linear mapping $\mathcal{A}$ from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^p$, defined by $\mathcal{A}(\boldsymbol{W}) = (\boldsymbol{A}_1 \bullet \boldsymbol{W}, \boldsymbol{A}_2 \bullet \boldsymbol{W}, \ldots, \boldsymbol{A}_p \bullet \boldsymbol{W})^\top$ with $p = |\Omega|$ and $\boldsymbol{b} \in \mathbb{R}^p$, it can be described with a general notation*

$$\begin{array}{ll} \textit{minimize}_{\boldsymbol{W}} & \|\boldsymbol{W}\|_* \\ \textit{subject to} & \mathcal{A}(\boldsymbol{W}) = \boldsymbol{b}. \end{array}$$

With a given parameter $\mu > 0$, the nuclear norm regularized linear least squares problem

$$\underset{\boldsymbol{W}}{\text{minimize}} \quad \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}) - \boldsymbol{b}\|_2^2 + \mu\|\boldsymbol{W}\|_* \tag{12}$$

is proposed as a variation [48]. Assuming that $f(\boldsymbol{W}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}) - \boldsymbol{b}\|_2^2$, $S = \mathbb{R}^{m\times n}$, and an integer $K \in \{1, \ldots, p\}$, we will relate (11) to (12) in Section 4.3.

A connection between the ℓ_0-norm on $\mathbb{R}^n$ and the rank function for a matrix is discussed in [40, 22]. On the basis of that, our approach to a cardinality-constrained problem (2) can be extended to (11).

4.1 DC Formulations and Algorithms for Rank-constrained Problems

Definition 2. *For an integer $k \in \{1, \ldots, n\}$, the Ky Fan k norm of a matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, denoted by $|||\boldsymbol{W}|||_k$, is defined as the sum of k largest singular values of $\boldsymbol{W}$. Namely,*

$$|||\boldsymbol{W}|||_k := \sum_{i=1}^{k} \sigma_i(\boldsymbol{W}).$$

Unless confusion occurs, we use the same notation, $|||\cdot|||_k$, for the Ky Fan k norm as that for the largest-k norm. [10]

We can immediately have equivalent representations of the rank constraint via the Ky Fan k norm from Theorem 1.

Corollary 3. *For any integers K, h such that $1 \le K < h \le \min\{m, n\}$, and $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, the following three conditions are equivalent:*

1. $\mathrm{rank}(\boldsymbol{W}) \le K$,
2. $|||\boldsymbol{W}|||_h - |||\boldsymbol{W}|||_K = 0$, *and*
3. $\|\boldsymbol{W}\|_* - |||\boldsymbol{W}|||_K = 0$.

Furthermore, the following three conditions are equivalent:

4. $\mathrm{rank}(\boldsymbol{W}) = K$,
5. $K = \min\{k : |||\boldsymbol{W}|||_h - |||\boldsymbol{W}|||_k = 0\}$, *and*

[10] Following [35, 1], the Ky Fan k norm of a matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ can be computed by solving the following semidefinite programming (SDP) problem:

$$|||\boldsymbol{W}|||_k = \min_{\boldsymbol{W},\boldsymbol{Z},c} \left\{ kc + \mathrm{Tr}(\boldsymbol{Z}) : \boldsymbol{Z} \succeq \begin{pmatrix} \boldsymbol{O} & \boldsymbol{W}^\top \\ \boldsymbol{W} & \boldsymbol{O} \end{pmatrix} - c\boldsymbol{I},\ \boldsymbol{Z} \succeq \boldsymbol{O} \right\},$$

where $\boldsymbol{Z} \succeq \boldsymbol{Y}$ denotes that $\boldsymbol{Z} - \boldsymbol{Y}$ is positive semi-definite.

6. $K = \min\{k : \|\boldsymbol{W}\|_1 - \|\boldsymbol{W}\|_k = 0\}$.

Noting that the rank of a matrix equals the number of nonzero singular values of the matrix, i.e., $\mathrm{rank}(\boldsymbol{W}) = \|\boldsymbol{\sigma}(\boldsymbol{W})\|_0$, and the nuclear norm of $\boldsymbol{W}$ is given as the sum of all its singular values, i.e., $\|\boldsymbol{W}\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(\boldsymbol{W}) = \|\boldsymbol{W}\|_{\min\{m,n\}}$, all the statements are straightforward from Theorem 1.

Assuming that f is DC-representable with two convex functions, i.e., $f = g - h$, as in Assumption 1, the subproblem at the t-th iteration of the DCA for (11) is given as

$$\underset{\boldsymbol{W} \in S}{\text{minimize}} \quad g(\boldsymbol{W}) + \rho\|\boldsymbol{W}\|_* - \boldsymbol{G}_{\boldsymbol{W}}^{t-1} \bullet \boldsymbol{W}, \tag{13}$$

where $\boldsymbol{G}_{\boldsymbol{W}}^{t-1} \in \partial h(\boldsymbol{W}^{t-1}) + \rho\partial\|\boldsymbol{W}^{t-1}\|_K$. The subdifferential of the Ky Fan k norm at $\boldsymbol{W}^t$ is given in [50] as

$$\partial\|\boldsymbol{W}^t\|_K := \Bigg\{ \boldsymbol{U}\mathrm{diag}(\boldsymbol{q}^*)\boldsymbol{V}^\top : \boldsymbol{q}^* \in \underset{\boldsymbol{q} \in \mathbb{R}^d}{\mathrm{argmax}} \left\{ \sum_{i=1}^{d} \sigma_i(\boldsymbol{W}^t) q_i : \mathbf{1}^\top \boldsymbol{q} = K, \mathbf{0} \le \boldsymbol{q} \le \mathbf{1} \right\} \Bigg\},$$

where $d = \min\{m, n\}$ and $\boldsymbol{U}\mathrm{diag}(\boldsymbol{\sigma}(\boldsymbol{W}^t))\boldsymbol{V}^\top$ is a singular value decomposition (SVD) of $\boldsymbol{W}^t$. Note that a component of $\partial\|\boldsymbol{W}^t\|_K$ is efficiently obtained by computing the SVD and picking up the SVD vectors corresponding to the K largest singular values.

4.2 Closed-form Solution via a Soft Thresholding

While (13) results in an SDP, [11] solving SDP at each iteration can be costly. Therefore, as shown in Section 3.3 for the subproblems (6), let us show a closed-form solution of (13) for the case of $S = \mathbb{R}^{m \times n}$.

Suppose that $f(\boldsymbol{W})$ is twice differentiable and there exists $\bar{\lambda}$ which makes the function $\frac{\bar{\lambda}}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2 - f(\boldsymbol{W})$ convex. Then $f(\boldsymbol{W})$ can be expressed as the difference of two convex functions as

$$f(\boldsymbol{W}) = \frac{\bar{\lambda}}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2 - \left(\frac{\bar{\lambda}}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2 - f(\boldsymbol{W}) \right).$$

[11] Problem (13) can be recast as the following problem:

$$\begin{array}{ll} \underset{\boldsymbol{W}, \boldsymbol{Z}_1, \boldsymbol{Z}_2}{\text{minimize}} & g(\boldsymbol{W}) + \frac{\rho}{2}(\mathrm{Tr}(\boldsymbol{Z}_1) + \mathrm{Tr}(\boldsymbol{Z}_2)) - \boldsymbol{G}_{\boldsymbol{W}}^{t-1} \bullet \boldsymbol{W} \\ \text{subject to} & \boldsymbol{W} \in S, \ \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{W} \\ \boldsymbol{W}^\top & \boldsymbol{Z}_2 \end{pmatrix} \succeq \boldsymbol{O}. \end{array}$$

If g is a linear function and S is given by a system of linear functions on $\mathbb{R}^{m \times n}$, the above problem can be solved by a standard (linear) SDP solver.

The DCA subproblem (13) then becomes

$$\underset{\boldsymbol{W}}{\text{minimize}} \ \ \frac{\bar{\lambda}}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2 + \rho\|\boldsymbol{W}\|_* - \boldsymbol{G}_{\boldsymbol{W}}^{t-1} \bullet \boldsymbol{W},$$

where $\boldsymbol{G}_{\boldsymbol{W}}^{t-1} \in \{\bar{\lambda}\boldsymbol{W}^{t-1} - \nabla f(\boldsymbol{W}^{t-1})\} + \rho\partial\|\!|\boldsymbol{W}^{t-1}|\!\|_K$. It is equivalent to

$$\underset{\boldsymbol{W}}{\text{minimize}} \ \ \frac{1}{2}\left\|\boldsymbol{W} - \frac{1}{\bar{\lambda}}\boldsymbol{G}_{\boldsymbol{W}}^{t-1}\right\|_{\mathrm{F}}^2 + \frac{\rho}{\bar{\lambda}}\|\boldsymbol{W}\|_*. \tag{14}$$

An optimal solution of (14) is then given by the proximal mapping:

$$\begin{aligned}\mathrm{prox}_{(\rho/\bar{\lambda})\|\cdot\|_*}\left(\frac{\boldsymbol{G}_{\boldsymbol{W}}^{t-1}}{\bar{\lambda}}\right) &:= \underset{\boldsymbol{W}}{\mathrm{argmin}} \ \frac{1}{2}\left\|\boldsymbol{W} - \frac{1}{\bar{\lambda}}\boldsymbol{G}_{\boldsymbol{W}}^{t-1}\right\|_{\mathrm{F}}^2 + \frac{\rho}{\bar{\lambda}}\|\boldsymbol{W}\|_* \\ &= \boldsymbol{P}\mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)\boldsymbol{Q}^\top,\end{aligned}$$

where the SVD of $\boldsymbol{G}_{\boldsymbol{W}}^{t-1}/\bar{\lambda}$ is supposed to be given as $\boldsymbol{P}\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\boldsymbol{Q}^\top$ and

$$\hat{\sigma}_i = \begin{cases} \sigma_i - \rho/\bar{\lambda}, & (\sigma_i \geq \rho/\bar{\lambda}), \\ 0, & (0 \leq \sigma_i \leq \rho/\bar{\lambda}). \end{cases}$$

4.3 Application to Matrix Completion Problem

For the matrix completion problem in Example 5, we consider a rank-constraint formulation, which is equivalently recast as a DC-constrained problem:

$$\begin{array}{ll} \underset{\boldsymbol{W}}{\text{minimize}} & \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}) - \boldsymbol{b}\|_2^2 \\ \text{subject to} & \|\boldsymbol{W}\|_* - \|\!|\boldsymbol{W}|\!\|_K = 0. \end{array} \tag{15}$$

Theorem 3. *Suppose that K satisfies*

$$K(m+n-K) < R := \max\left\{r : \textit{any } r \times r \textit{ principal minor of } \sum_{i=1}^{p} \mathrm{vec}(\boldsymbol{A}_i)\mathrm{vec}(\boldsymbol{A}_i)^\top \textit{ is positive definite}\right\},$$

where $\mathrm{vec}(\boldsymbol{A})$ *denotes the* $mn \times 1$ *column vector obtained by stacking each column of* $\boldsymbol{A}$ *on the top of another. The DC-constrained problem (15) is equivalent to the DC-penalty problem*

$$\underset{\boldsymbol{W}}{\textit{minimize}} \ \ \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}) - \boldsymbol{b}\|_2^2 + \rho(\|\boldsymbol{W}\|_* - \|\!|\boldsymbol{W}|\!\|_K), \tag{16}$$

if

$$\rho > \sum_{i=1}^{p} \|\boldsymbol{A}_i\|_2 \left(C\|\boldsymbol{A}_i\|_{\mathrm{F}} + \frac{C}{2}\|\boldsymbol{A}_i\|_2 + |b_i|\right),$$

where $C = \sum_{i=1}^{p} |b_i| \|\boldsymbol{A}_i\|_{\mathrm{F}} / \hat{\lambda}_R$ *and*

$$\hat{\lambda}_R = \min_{I:|I|=R} \lambda_{\min}([\sum_{i=1}^{p} \mathrm{vec}(\boldsymbol{A}_i)\mathrm{vec}(\boldsymbol{A}_i)^\top]_I).$$

Note that $K(m+n-K)$ is the degree of freedom of an $m \times n$ matrix of rank K [14]. See Appendix A.3 for the proof of Theorem 3.

As in the vector-case, the difference between the DCA subproblem (16) and (12) is noteworthy. Identifying the parameters ρ and μ, the difference of (16) from (12) is in the concave term, $-\rho\|\boldsymbol{W}\|_K$, and we may regard (16) as a non-convex modification of the penalty in (12). A numerical comparison of the two methods will be reported in Section 6.2.

The DCA subproblem (14) for the matrix completion problem amounts to

$$\min_{\boldsymbol{W}} \frac{\bar{\lambda}}{2} \left\| \boldsymbol{W} - \left(\boldsymbol{W}^{t-1} - \frac{1}{\bar{\lambda}} (\mathcal{A}^*(\mathcal{A}(\boldsymbol{W}^{t-1}) - \boldsymbol{b}) - \rho \boldsymbol{B}^{t-1}) \right) \right\|_F^2 + \rho \|\boldsymbol{W}\|_*, \tag{17}$$

where $\boldsymbol{B}^{t-1} \in \partial\|\boldsymbol{W}^{t-1}\|_K$ and $\mathcal{A}^*$ is the adjoint of $\mathcal{A}$. It is solved by the soft thresholding, while proximal gradient methods solve the nuclear norm minimization (12) by applying the soft thresholding to

$$\min_{\boldsymbol{W}} \frac{\tau}{2} \left\| \boldsymbol{W} - \left(\boldsymbol{W}^{t-1} - \frac{1}{\tau} \mathcal{A}^*(\mathcal{A}(\boldsymbol{W}^{t-1}) - \boldsymbol{b}) \right) \right\|_F^2 + \mu \|\boldsymbol{W}\|_*, \tag{18}$$

where τ is a given parameter larger than the spectral norm of the linear mapping $\mathcal{A}$ (i.e.,$\|\mathcal{A}\|_2 := \max\{\|\mathcal{A}(\boldsymbol{W})\|_2 : \|\boldsymbol{W}\|_{\mathrm{F}} = 1\}$), as is the case with $\bar{\lambda}$.

5 Extensions

In this section we consider two extensions of the DC reformulation technique developed so far.

5.1 Cardinality and Rank Minimization Problems

Let us consider the cardinality minimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{w}}{\text{minimize}} & \|\boldsymbol{w}\|_0, \\ \text{subject to} & \boldsymbol{w} \in S, \end{array} \tag{19}$$

on the basis of the method developed for the cardinality-constrained problem.

Problem (19) can be reduced to a decision problem:

Find the smallest $k \in \{1, \ldots, n\}$ such that the following system is feasible:

$$\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_k = 0,\ \boldsymbol{w} \in S.$$

To tackle this, consider to minimize DC objectives for $k = 1, \ldots, n$:

$$\phi_k := \min\{\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_k : \boldsymbol{w} \in S\}. \tag{20}$$

The smallest k such that $\phi_k = 0$ is the optimal value of (19) and the obtained solution is that of (19). Since the term $\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_k$ is non-increasing in k, we may employ binary search to find the smallest k.

Note that if the DCA is applied to (20) with some k and finds a feasible solution whose objective value is 0, then we can conclude that the true optimal value of (19) is at most k.

The above approach can also be applied to a rank minimization problem:

$$\begin{array}{ll} \underset{\boldsymbol{w}}{\text{minimize}} & \text{rank}(\boldsymbol{W}), \\ \text{subject to} & \boldsymbol{w} \in S, \end{array} \tag{21}$$

which is equivalent to finding the smallest $k \in \{1, \ldots, \min\{m, n\}\}$ such that

$$\min\{\|\boldsymbol{W}\|_* - \|\boldsymbol{W}\|_k : \boldsymbol{W} \in S\} = 0.$$

5.2 Sparse Matrix Norm Problems

We consider the following problem

$$\begin{array}{ll} \underset{\boldsymbol{W}}{\text{minimize}} & f(\boldsymbol{W}) \\ \text{subject to} & \boldsymbol{W} \in S \subset \mathbb{R}^{m \times n},\ \|\boldsymbol{W}\|_{2,0} \leq K, \end{array} \tag{22}$$

where $\|\boldsymbol{W}\|_{2,0} := \left\|(\|\boldsymbol{w}_1\|_2^2, \ldots, \|\boldsymbol{w}_d\|_2^2)^\top\right\|_0$ is called the $\ell_{2,0}$-norm of a matrix $\boldsymbol{W} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{w}_i$ denotes the i-th row of $\boldsymbol{W}$. This type of problem is recently discussed in [13] for a multi-class feature selection problem, where for input data $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ (d training data) and response data $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$ (n classes), the residual $f(\boldsymbol{W}, \boldsymbol{b}) = \|\boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{W} - \boldsymbol{1}\boldsymbol{b}^\top\|_{2,1}$ is minimized over $(\boldsymbol{W}, \boldsymbol{b}) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{n \times 1}$. Here, $\|\boldsymbol{W}\|_{2,1}$ is defined as

$$\|\boldsymbol{W}\|_{2,1} := \sqrt{\sum_{i=1}^{m} \|\boldsymbol{w}_i\|_2^2} \equiv \|\boldsymbol{W}\|_{\mathrm{F}}.$$

Denoting by $\boldsymbol{w}_{(i)}$ the row of $\boldsymbol{W}$ whose ℓ_2-norm is the i-th largest among all the m rows, let

$$\|\boldsymbol{W}\|_{2,k} := \sqrt{\|\boldsymbol{w}_{(1)}\|_2^2 + \cdots + \|\boldsymbol{w}_{(k)}\|_2^2},$$

where $k \in \{1, \ldots, m\}$. The cardinality constraint $\|\boldsymbol{W}\|_{2,0} \leq K$ can then be rewritten as $\|\!|\boldsymbol{W}|\!\|_{2,h}^2 - \|\!|\boldsymbol{W}|\!\|_{2,K}^2 = 0$ for any $1 \leq K < h \leq n$, or as $\|\boldsymbol{W}\|_{\mathrm{F}}^2 - \|\!|\boldsymbol{W}|\!\|_{2,K}^2 = 0$. It is straightforward to apply a DCA by noting that

$$\partial \|\!|\boldsymbol{W}^t|\!\|_{2,K} = \left\{ 2\mathrm{diag}(\boldsymbol{q}^*)\boldsymbol{W}^t : \boldsymbol{q}^* \in \operatorname*{argmax}_{\boldsymbol{q} \in \mathbb{R}^d} \left\{ \sum_{i=1}^{d} \|\boldsymbol{w}_i^t\|_2^2 q_i : \begin{array}{l} \mathbf{1}^\top \boldsymbol{q} = K, \\ \mathbf{0} \leq \boldsymbol{q} \leq \mathbf{1} \end{array} \right\} \right\}.$$

6 Numerical Experiments

This section reports numerical results of the DCA-based approaches, examining the capability of the framework by focusing on the behavior of them.[12] In the first subsection, we solve the cardinality-constrained linear regression (Example 1) and compare with another exact DC reformulation based on [37]. In the subsequent subsection, we solve a matrix completion problem (Example 5) and compare with a convex relaxation formulation which uses the nuclear norm.

6.1 Subset Selection in Regression

The aim of this section is to compare two different exact DC formulations of the cardinality-constrained problem.

MIP-based DCA. As mentioned in Introduction, [37] rewrites the cardinality constraint by (1), which is further transformed into

$$\mathbf{1}^\top \boldsymbol{u} \leq K,\ \boldsymbol{u} \in [0,1]^n, \boldsymbol{u}^\top(\mathbf{1} - \boldsymbol{u}) \leq 0,\ |w_i| \leq M_j u_j,\ j = 1, \ldots, n.$$

Moving the nonconvex constraint $\boldsymbol{u}^\top(\mathbf{1} - \boldsymbol{u}) \leq 0$ to the objective part brings an exact penalty formulation as

$$\begin{array}{ll} \underset{\boldsymbol{w},\boldsymbol{u}}{\text{minimize}} & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} + \boldsymbol{q}^\top \boldsymbol{w} + \rho(\mathbf{1}^\top \boldsymbol{u} - \boldsymbol{u}^\top \boldsymbol{u}) \\ \text{subject to} & -M_j u_j \leq w_j \leq M_j u_j, j = 1, \ldots, n,\ \mathbf{1}^\top \boldsymbol{u} \leq K,\ \mathbf{0} \leq \boldsymbol{u} \leq \mathbf{1}, \end{array} \tag{23}$$

[12] All the numerical experiments in this section were performed on an Intel Core i7 2.9 GHz personal computer with 8GB of physical memory using Matlab (R2013a) with IBM ILOG CPLEX 12.

where $\rho > 0$ is a sufficiently large constant.[13] Finally, using the fact that an optimal solution of (23) satisfies $\mathbf{1}^\top \boldsymbol{u} = K$, we rewrite (23) as

$$\begin{array}{ll} \underset{\boldsymbol{w},\boldsymbol{u}}{\text{minimize}} & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} + \boldsymbol{q}^\top \boldsymbol{w} - \rho \boldsymbol{u}^\top \boldsymbol{u} \\ \text{subject to} & -M_j u_j \le w_j \le M_j u_j, j = 1, \ldots, n,\ \mathbf{1}^\top \boldsymbol{u} = K,\ \mathbf{0} \le \boldsymbol{u} \le \mathbf{1}. \end{array} \tag{24}$$

We call the DCA applied to (24) the *MIP-based DCA*.[14]

We compare the three DCA algorithms: the MIP-based DCA, the polyhedral DCA shown in (8), and the proximal DCA in (10), which are abbreviated as MIP-DCA, Poly-DCA, and Prox-DCA, respectively. The diabetes data set [8] and synthetic data sets are used for the comparison.

Initialization and Termination. All the three algorithms start from (i) $\boldsymbol{w}^0 = \boldsymbol{w}_{\mathrm{OLS}}$ (the ordinary least squares (OLS) solution) or (ii) $\boldsymbol{w}^0 = \mathbf{0}$ (plus $\boldsymbol{u}_0 = (K/n)\mathbf{1}$ for the MIP-DCA), and terminate if $\|\boldsymbol{w}^t\|_0$ becomes no greater than a target cardinality, K^*, and $|\mathrm{obj}(\boldsymbol{w}^t) - \mathrm{obj}(\boldsymbol{w}^{t-1})|/\max\{1, \mathrm{obj}(\boldsymbol{w}^t)\} < 10^{-4}$ is fulfilled.

Behavior for Diabetes Data. Figure 1 shows the behaviors of the three algorithms for each fixed ρ, $K = 5$, and $\boldsymbol{w}^0 = \boldsymbol{w}_{\mathrm{OLS}}$. It reports the attained cardinality, the sum of squared residuals (SSR), defined by $\|\boldsymbol{A}\boldsymbol{w}^* - \boldsymbol{b}\|_2^2$ for an output solution $\boldsymbol{w}^*$, the numbers of iterations and computation time in [sec.] for finding a solution $\boldsymbol{w}^*$. We see from the figure that (a) Poly- and Prox-DCAs stably attained smaller SSR values than MIP-DCA at large ρs; (b) attained cardinality was non-increasing with respect to ρ whereas the SSR was non-decreasing; however, (c) the three methods often overshot the target cardinality $K^* = 5$. Although neither of obtained solutions are globally optimal, Poly- and Prox-DCAs attained the target cardinality at

[13] $M_j = 4\|\boldsymbol{q}\|_2/\hat{\lambda}_R$ is a sufficiently large constant for any j so that an optimal solution $\boldsymbol{w}^\star$ of the cardinality constrained problem satisfies $M_j \ge \|\boldsymbol{w}^\star\|_\infty \ge |w_j^\star|$. In a similar manner to Theorem 2, we can provide a lower bound of ρ, above which (23) becomes an exact penalty formulation of the cardinality constraint. It is given as

$$\rho > \max_i \left\{ \frac{8\|\boldsymbol{q}\|_2}{\hat{\lambda}_R} \left(|q_i| + \frac{2\|\boldsymbol{Q}\boldsymbol{e}_i\|_2 \|\boldsymbol{q}\|_2 + \|\boldsymbol{q}\|_2 Q_{ii}}{\hat{\lambda}_R} \right) \right\}.$$

Note that the lower bound of ρ for (23) is $8\|\boldsymbol{q}\|_2/\hat{\lambda}_R$ times as large as that for (4). See Appendix A.4 for the lower bound of ρ.

[14] DCA for (24) repeats the following procedure:

$$\begin{aligned} &(\boldsymbol{g}_{\boldsymbol{w}}^{t-1}, \boldsymbol{v}_u^{t-1}) = (-\boldsymbol{q}, 2\rho\boldsymbol{u}), \\ &\quad (\boldsymbol{w}^t, \boldsymbol{u}^t) \in \underset{\boldsymbol{w},\boldsymbol{u}}{\mathrm{argmin}} \Big\{ \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} - \boldsymbol{w}^\top \boldsymbol{g}_{\boldsymbol{w}}^{t-1} - \boldsymbol{u}^\top \boldsymbol{v}_u^{t-1} : \\ &\qquad\qquad - M_j u_j \le w_j \le M_j u_j, j = 1, \ldots, n,\ \mathbf{1}^\top \boldsymbol{u} = K,\ \mathbf{0} \le \boldsymbol{u} \le \mathbf{1} \Big\}. \end{aligned}$$

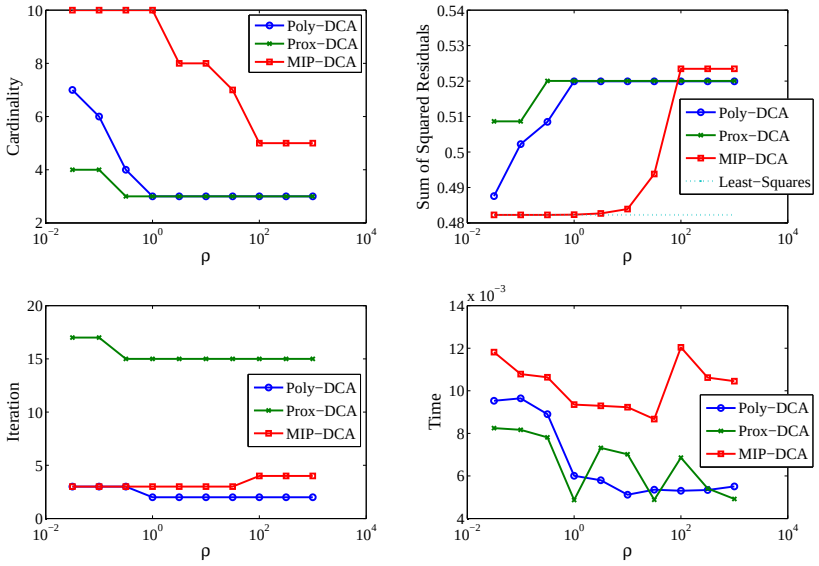

Figure 1: Comparison of algorithms with fixed $K = 5$ and ρ starting from the least-squares initial solution for diabetes data set

smaller ρs, which seems consistent with the discussion of the footnote 13 saying that the lower bound of ρ for our formulation is smaller than that for MIP-DCA.

For the diabetes data set ($m = 442, n = 10$), the theoretical lower bounds of the exact penalty are $\rho_{\text{theo}} = 3.92 \times 10^3$ for Poly- and Prox-DCAs and $\rho_{\text{theo}} = 4.42 \times 10^6$ for MIP-DCA.[15] We see from the figure that ρ can be set to a smaller value than ρ_{theo} in practice.

Figure 2 reports the attained cardinality, SSR, and the accumulated computational time as functions of iteration for the case where the algorithms started from $\boldsymbol{w}^0 = \mathbf{0}$ or $\boldsymbol{w}_{\text{OLS}}$ and ρ is fixed at $\rho = \rho_{\text{theo}}/100$. The cardinality parameter K was updated in each iteration by $K^t = \max\{\lfloor 0.9K^{t-1} \rfloor, K^*\}$ from $K^0 = n$ to the target cardinality K^* set to $n/2$. Note that the update rule aims at finding a sparse solution satisfying the target, i.e., $\|\boldsymbol{w}\|_0 = n/2$, and Prox- and Poly-DCAs succeeded in finding such sparse solutions having smaller SSR.

Behavior for Synthetic Data. To see results with larger-scale data, the three algorithms were compared over synthetic data. Each column $\boldsymbol{a_i}$ of the matrix $\boldsymbol{A}^\top = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$ was drawn independently from the normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\sigma_{ij}) = (0.5^{|i-j|})$, and each column of $\boldsymbol{A}$

[15] These values are computed by replacing $\hat{\lambda}_R$ with $\lambda_{\min}(\boldsymbol{Q})$.

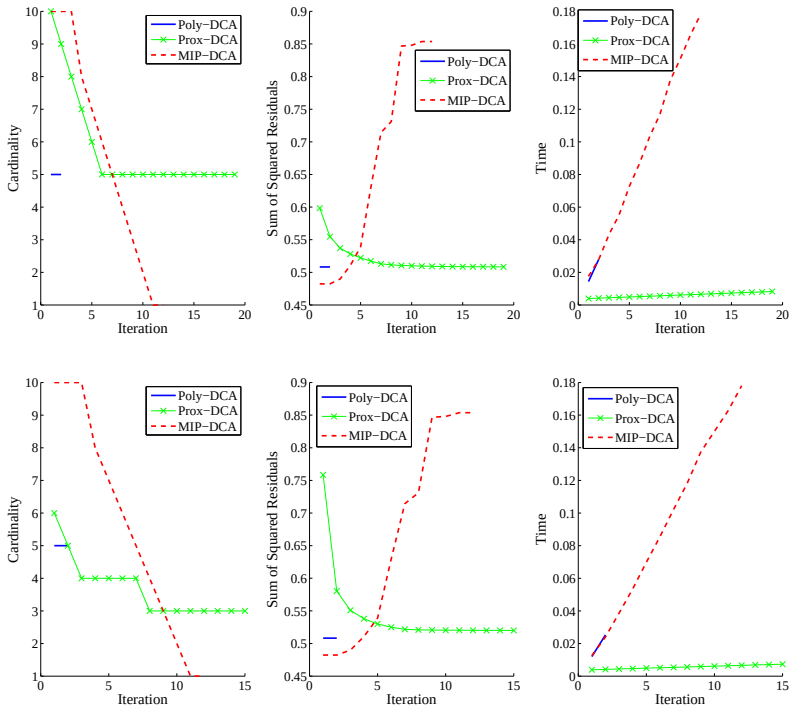

Figure 2: Iterative performance of algorithms with fixed $K = 5$ and $\rho = \rho_{\text{theo}}/100$ for diabetes data set. The upper panel shows the results for the zero initial solution and the lower one shows the results for the least-squares initial solution.

was then standardized, i.e, $\|\boldsymbol{a}_i\|_2 = 1$; $\boldsymbol{b}$ was generated by $\boldsymbol{b} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{\epsilon}$, where $\bar{x}_i \sim U(0,1)$ and $\epsilon_i \sim N(0,1)$.

Average performance among 10 runs are reported in Tables 1 and 2 where $\boldsymbol{w}^0 = \boldsymbol{0}$ and $\boldsymbol{w}^0 = \boldsymbol{w}_{\text{OLS}}$ were used, respectively. The penalties ρ of all the algorithms are fixed to $\rho_{\text{theo}}/100$ and K is updated by $K^t = \max\{\lfloor 0.9K^{t-1} \rfloor, K^*\}$ from $K^0 = n$ or $n/2$ until $K^* = n/10$.

All the three algorithms found solutions satisfying the target cardinality $\|\boldsymbol{w}\|_0 = n/10$ by using the penalty of $\rho = \rho_{\text{theo}}/100$. For each data set, Poly-DCA and Prox-DCA achieved smaller residuals than MIP-DCA and Prox-DCA ran quite faster than the others. We see that all the algorithms run faster and tend to have larger residuals by starting with smaller rank requirement $K^0 = n/2$. It seems to be reasonable to consider that choosing the smaller K^0 restricts the search for better solutions at an early stage of the algorithm.

We also compared two different initial solutions for DCA algorithms in these two tables. Our approaches, Poly- and Prox-DCAs, tend to have better performance in terms of computational time and residual by starting from

Table 1: Performance among 10 runs using the zero solution as an initial solution. The numbers in parenthesis are the standard deviations.

n	m	K^0	method	iter.	time	card.	residual
100	1000	n	MIP-DCA	22.0	0.719 (0.045)	10	8.304e-01 (2.807e-02)
			Poly-DCA	21.0	0.509 (0.025)	10	**6.827e-01** (1.802e-02)
			Prox-DCA	26.6	**0.024** (0.002)	10	6.986e-01 (2.263e-02)
500	1000	n	MIP-DCA	24.1	31.310 (3.109)	50	7.950e-01 (1.031e-01)
			Poly-DCA	24.0	23.604 (1.939)	50	**5.282e-01** (1.358e-02)
			Prox-DCA	32.9	**0.139** (0.024)	50	5.374e-01 (1.333e-02)
500	5000	n	MIP-DCA	24.5	32.125 (1.511)	50	9.137e-01 (1.881e-02)
			Poly-DCA	24.0	21.410 (0.755)	50	**6.794e-01** (8.984e-03)
			Prox-DCA	28.4	**0.118** (0.016)	50	6.874e-01 (1.169e-02)
1000	5000	n	MIP-DCA	24.6	297.556 (19.745)	100	8.615e-01 (4.740e-02)
			Poly-DCA	24.0	233.123 (12.390)	100	**6.332e-01** (7.270e-03)
			Prox-DCA	30.3	**0.502** (0.026)	100	6.386e-01 (7.842e-03)
100	1000	$n/2$	MIP-DCA	15.8	0.394 (0.021)	10	8.059e-01 (4.684e-02)
			Poly-DCA	15.0	0.279 (0.006)	10	**7.426e-01** (1.415e-02)
			Prox-DCA	21.7	**0.017** (0.001)	10	7.482e-01 (1.671e-02)
500	1000	$n/2$	MIP-DCA	17.4	21.388 (1.274)	50	8.640e-01 (1.285e-02)
			Poly-DCA	17.0	15.842 (0.425)	50	**5.780e-01** (1.984e-02)
			Prox-DCA	28.6	**0.119** (0.019)	50	5.909e-01 (1.924e-02)
500	5000	$n/2$	MIP-DCA	17.9	20.455 (0.801)	50	8.445e-01 (1.581e-02)
			Poly-DCA	17.0	15.085 (0.499)	50	**7.424e-01** (1.044e-02)
			Prox-DCA	23.3	**0.100** (0.010)	50	7.492e-01 (9.009e-03)
1000	5000	$n/2$	MIP-DCA	17.7	193.367 (7.910)	100	8.296e-01 (1.391e-02)
			Poly-DCA	17.0	162.284 (1.436)	100	**7.040e-01** (1.311e-02)
			Prox-DCA	24.7	**0.431** (0.009)	100	7.109e-01 (1.244e-02)

$\boldsymbol{w}^0 = \boldsymbol{w}_{\mathrm{OLS}}$.

6.2 Matrix Completion Problem

The main aim of this subsection is to compare the DC penalty formulation (16) with the nuclear norm penalty formulation (12). To that aim, we solve the matrix completion problem, which is described in Example 5.

Data Generation. Data sets are generated by following [48, 14]. Let r be the rank of a target matrix $\boldsymbol{M}$. We generated $\boldsymbol{M}$ as $\boldsymbol{M} = \boldsymbol{M}_L \boldsymbol{M}_R^\top$, where each element of $\boldsymbol{M}_L \in \mathbb{R}^{m\times r}$ and $\boldsymbol{M}_R \in \mathbb{R}^{n\times r}$ were drawn i.i.d. from $N(0,1)$. Note that an $m \times n$ matrix whose rank is r has $d_r = r(m+n-r)$ degrees of freedom. In this experiment, we change p/d_r which determines the number of p, i.e., the number of the observed entries.

Parameter Updating. Following [48], the initial solution $\boldsymbol{W}^0$ is set to the zero matrix, $\tau = \bar{\lambda} = \|\mathcal{A}\|_2 = 1$ is used, and the μ (ρ in our model) is

Table 2: Performance among 10 runs using the least-squares solution as an initial solution. The numbers in parenthesis are the standard deviations.

n	m	K^0	method	iter.	time	card.	residual
100	1000	n	MIP-DCA	22.0	0.536 (0.016)	10	8.304e-01 (2.807e-02)
			Poly-DCA	21.0	0.379 (0.015)	10	**6.827e-01** (1.802e-02)
			Prox-DCA	26.4	**0.018** (0.002)	10	7.016e-01 (1.976e-02)
500	1000	n	MIP-DCA	24.1	29.605 (1.419)	50	7.950e-01 (1.031e-01)
			Poly-DCA	24.0	21.766 (0.752)	50	**5.282e-01** (1.358e-02)
			Prox-DCA	30.2	**0.116** (0.012)	50	5.438e-01 (1.846e-02)
500	5000	n	MIP-DCA	24.1	29.605 (1.419)	50	7.950e-01 (1.031e-01)
			Poly-DCA	24.0	21.766 (0.752)	50	**5.282e-01** (1.358e-02)
			Prox-DCA	30.2	**0.116** (0.012)	50	5.438e-01 (1.846e-02)
1000	5000	n	MIP-DCA	24.6	310.133 (26.087)	100	8.615e-01 (4.740e-02)
			Poly-DCA	24.0	244.985 (19.279)	100	**6.332e-01** (7.270e-03)
			Prox-DCA	29.9	**0.514** (0.029)	100	6.426e-01 (6.032e-03)
100	1000	$n/2$	MIP-DCA	15.8	0.380 (0.014)	10	8.059e-01 (4.684e-02)
			Poly-DCA	15.0	0.270 (0.015)	10	**6.838e-01** (1.943e-02)
			Prox-DCA	20.5	**0.016** (0.002)	10	6.937e-01 (1.903e-02)
500	1000	$n/2$	MIP-DCA	17.4	21.549 (1.494)	50	8.640e-01 (1.285e-02)
			Poly-DCA	17.0	15.047 (0.459)	50	**5.288e-01** (1.544e-02)
			Prox-DCA	25.3	**0.106** (0.015)	50	5.424e-01 (1.617e-02)
500	5000	$n/2$	MIP-DCA	17.9	21.084 (1.348)	50	8.445e-01 (1.581e-02)
			Poly-DCA	17.0	14.823 (0.241)	50	**6.812e-01** (8.670e-03)
			Prox-DCA	22.1	**0.098** (0.009)	50	6.908e-01 (9.320e-03)
1000	5000	$n/2$	MIP-DCA	17.7	212.709 (18.116)	100	8.296e-01 (1.391e-02)
			Poly-DCA	17.0	173.532 (9.518)	100	**6.340e-01** (5.971e-03)
			Prox-DCA	23.3	**0.447** (0.024)	100	6.435e-01 (3.236e-03)

updated from $\mu^0 = \|\mathcal{A}^*(\boldsymbol{b})\|_2$ by the formula

$$\mu^t = \max\{0.9\mu^{t-1}, 10^{-4}\|\mathcal{A}^*(\boldsymbol{b})\|_2\},$$

implying that it would gradually decrease from $\|\mathcal{A}^*(\boldsymbol{b})\|_2$ to $10^{-4}\|\mathcal{A}^*(\boldsymbol{b})\|_2$[16]. The update rule for K from a designated K^0 is given by

$$K^t = \max\{\text{round}(0.8K^{t-1}), 1\} \tag{25}$$

when the number of positive singular values of $\boldsymbol{W}^t$ is no greater than $K^{t-1}+1$, and otherwise, $K^t = K^{t-1}$. The termination criterion is either of the maximum iteration number 500 or

$$|\text{obj}(\boldsymbol{W}^t) - \text{obj}(\text{obj}(\boldsymbol{W}^{t-1}))|/\max\{1, \text{obj}(\boldsymbol{W}^t)\} < 10^{-6}.$$

Here $\text{obj}(\boldsymbol{W})$ is given as the objective function of (18) for the nuclear norm minimization, and that of (17) for the DCA.

[16]In the original program code of [48], the coefficient was not 0.9, but 0.7. We slowed down the speed of decrease of μ because the proximal gradient (PG) without acceleration techniques needs much more iterations than the original code.

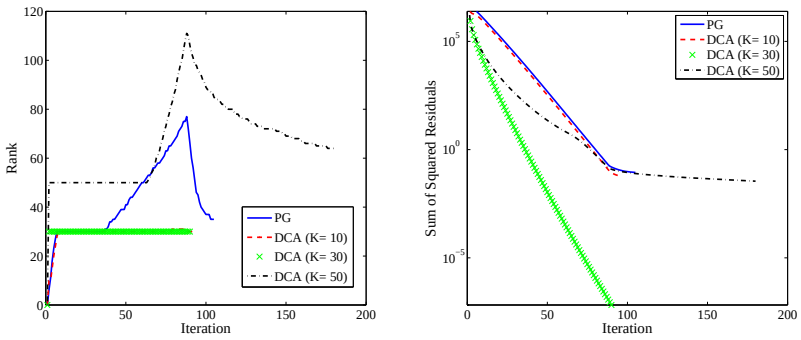

Figure 3: Comparison of PG and DCA with fixed K for a synthetic data set

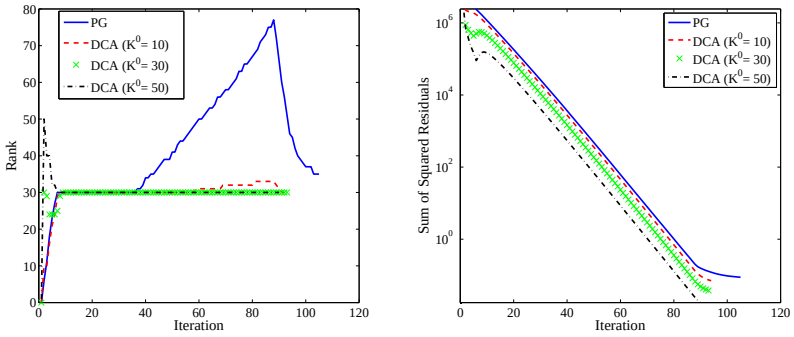

Figure 4: Comparison of PG and DCA with updated K starting from K^0 for a synthetic data set

We compare the performance of Prox-DCA (DCA in short) and the proximal gradient method for (12) (PG in short), i.e., iterative soft-thresholding for (18) in terms of error $:= \|\boldsymbol{W}^* - \boldsymbol{M}\|_F / \|\boldsymbol{M}\|_F$, the rank of the obtained matrix $\boldsymbol{W}^*$, the squared residual (SSR) $\|\mathcal{A}(\boldsymbol{W}^*) - \boldsymbol{b}\|_2^2$ and accumulated computational time. Figure 3 shows the rank and SSR for the PG and DCAs with fixed K for a synthetic data set generated with $m = n = 500$, $r = 30$, $p = 116322$ by setting $p/d_r = 4$. The error is 2.497×10^{-4} (PG), 1.725×10^{-4} (DCA with $K = 10$), 2.360×10^{-7} (DCA with $K = 30$), 2.441×10^{-2} (DCA with $K = 50$). This implies that the parameter K is sensitive to the performance of DCA.

The PG in Figure 4 is the same to the one in Figure 3 and the parameter K of DCAs is updated by (25) from K^0. Figure 4 indicates that the updating rule of K works and the performance of DCA is not so sensitive to the value of K^0. Indeed, the errors of DCA are also stable for any K^0: 1.811×10^{-4} (DCA with $K^0 = 10$), 1.346×10^{-4} (DCA with $K^0 = 30$), 9.477×10^{-5} (DCA with $K^0 = 50$).

Table 3: PG and DCA starting from $K^0 = n/10$

$n = m$	p	r	p/d_r	method	iter.	time	rank	residual	error
100	1192.8	1	6	PG	500.0	1.775	**28.7**	**1.053e-02**	4.934e-01
				DCA	440.1	**1.540**	30.8	1.064e-02	**4.617e-01**
100	3895.4	5	4	PG	500.0	**1.949**	44.4	3.978e-02	5.702e-02
				DCA	500.0	1.970	**16.0**	**2.373e-02**	**1.831e-02**
100	5717.6	10	3	PG	498.7	1.831	16.5	4.542e-02	9.794e-03
				DCA	186.1	**0.795**	**11.3**	**2.878e-02**	**1.038e-03**
500	59364.3	10	6	PG	500.0	55.894	165.0	2.351e-01	3.396e-02
				DCA	153.7	**21.896**	**14.2**	**6.069e-02**	**1.122e-03**
500	116392.9	30	4	PG	106.2	12.934	33.6	3.026e-01	5.021e-04
				DCA	90.3	**10.925**	**30.0**	**1.328e-01**	**9.570e-05**
500	142475.7	50	3	PG	107.7	13.280	54.7	4.695e-01	2.822e-04
				DCA	92.2	**11.990**	**50.0**	**3.116e-01**	**1.589e-04**
1000	237656.3	20	6	PG	500.0	474.140	317.5	6.576e-01	1.796e-02
				DCA	105.0	**83.977**	**22.6**	**1.656e-01**	**5.330e-04**
1000	465617.4	60	4	PG	95.3	71.957	60.5	8.691e-01	2.075e-04
				DCA	90.0	**69.254**	**60.0**	**3.938e-01**	**9.840e-05**
1000	569897.3	100	3	PG	97.1	91.993	101.2	1.337e+00	2.353e-04
				DCA	92.0	**85.198**	**100.0**	**8.448e-01**	**1.509e-04**

In Table 3, the average numerical results of PG and DCA with $K^0 = n/10$ are shown among 10 runs. PG hardly converged for data sets with small rank r and terminated with the maximum iteration 500. The DCA tends to find solutions with smaller rank, residual and error with smaller computation time than PG.

Finally, let us mention the Nesterov's acceleration technique [32] which FISTA [3] is also adopted. The technique accelerates the PG drastically[17] Indeed, Figure 5 implies that the performance of the accelerated PG (APG in short) drastically improved for the same synthetic data to Figures 3 and 4. The error also becomes smaller to 1.626×10^{-4}. There is no theoretical guarantee for the acceleration technique for nonconvex optimization problems but the DCA with the acceleration technique (ADCA in short) achieved the smallest error, 5.921×10^{-5}, among these four methods.

7 Conclusion

In this paper we propose an exact DC representation of the cardinality constraint by employing the largest-k norm, and rewrite the optimization problem as a DC optimization problem. We extend the reformulation technique to related optimization problems such as those with the rank constraint, the

[17] For the convex objective function $f(\boldsymbol{w})$, the rate of convergence to the optimal solution $\boldsymbol{w}^*$ changes from $f(\boldsymbol{w}^t) - f(\boldsymbol{w}^*) = O(1/t)$ to $O(1/t^2)$.

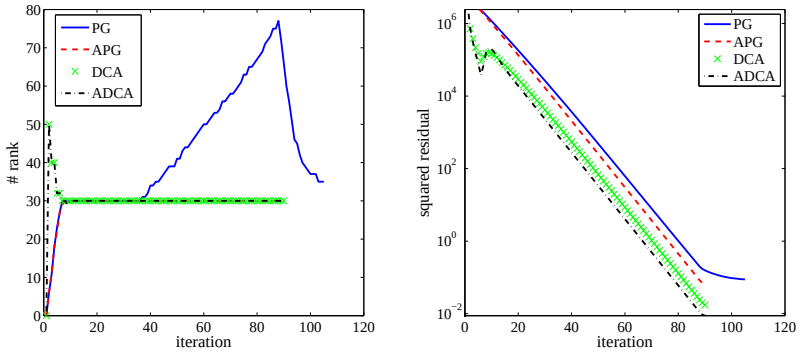

Figure 5: Accelerated PG and DCA for a synthetic data set

ℓ_0-objective, and some matrix norm constraint. For the penalized reformulation of the problem expression, we apply the DCA. An advantage of the use of the largest-k norm is that in each iteration of DCA, the subgradient computation is efficiently carried out.

While our reformulation is also advantageous over the existing exact reformulation in that we obtain smaller lower bounds above which the exact penalty is achieved, the numerical experiment supports its advantage, showing more stability.

Furthermore, since the penalty term contains the ℓ_1-norm in addition to the largest-k norm, a proximal method is constructed and each subproblems of DCA can be solved efficiently.

In general, the performance of first-order methods, which are practical and popular, can be tuned by incorporating various small improvements in program codes. Such a tuning so as to make the proposed algorithms truly practical is left for the future research. Also, the so-called acceleration techniques seem to be worth developing.

Acknowledgment. The research of the first author is supported by JSPS KAKENHI Grant Number 15K01204, 22510138, and 26242027.

A Proofs of Propositions

A.1 Proof of Lemma 1

Let $f_1(\boldsymbol{w}) := \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} + \boldsymbol{q}^\top \boldsymbol{w} + \rho(\|\boldsymbol{w}\|_1 - |||\boldsymbol{w}|||_K)$. We assume that $\|\boldsymbol{w}^\star\|_1 - |||\boldsymbol{w}^\star|||_K > 0$. We consider a feasible solution $\tilde{\boldsymbol{w}} := \boldsymbol{w}^\star - w_j^\star \boldsymbol{e}_j$, where j is the index of the $(K+1)$-st largest element of $(|w_1^\star|, \ldots, |w_n^\star|)^\top$. Then, from the

Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
& f_1(\boldsymbol{w}^\star) - f_1(\tilde{\boldsymbol{w}}) \\
&= \frac{1}{2}(\boldsymbol{w}^\star)^\top \boldsymbol{Q}\boldsymbol{w}^\star + \boldsymbol{q}^\top \boldsymbol{w}^\star - \frac{1}{2}(\boldsymbol{w}^\star - w_j^\star \boldsymbol{e}_j)^\top \boldsymbol{Q}(\boldsymbol{w}^\star - w_j^\star \boldsymbol{e}_j) \\
&\quad - \boldsymbol{q}^\top(\boldsymbol{w}^\star - w_j^\star \boldsymbol{e}_j) + \rho|w_j^\star| \\
&= w_j^\star \boldsymbol{e}_j^\top \boldsymbol{Q}\boldsymbol{w}^\star - \frac{1}{2} w_j^{\star 2} Q_{jj} + w_j^\star q_j + \rho|w_j^\star| \\
&\geq -|w_j^\star| \|\boldsymbol{Q}\boldsymbol{e}_j\|_2 \|\boldsymbol{w}^\star\|_2 - \frac{1}{2}|w_j^\star| \|\boldsymbol{w}^\star\|_2 |Q_{jj}| \\
&\quad - |q_j||w_j^\star| + \rho|w_j^\star| \quad (\because |w_j^\star| \leq \|\boldsymbol{w}^\star\|_2) \\
&\geq |w_j^\star| \left(\rho - C\|\boldsymbol{Q}\boldsymbol{e}_j\|_2 - C|Q_{jj}|/2 - |q_j|\right) > 0,
\end{aligned}
$$

which contradicts the optimality of $\boldsymbol{w}^\star$. (End of Proof of Lemma 1)

A.2 Proof of Theorem 2

Let $f_2(\boldsymbol{w}) := \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} + \boldsymbol{q}^\top \boldsymbol{w} + \rho(\|\boldsymbol{w}\|_1 - \|\boldsymbol{w}\|_K)$ and $\boldsymbol{w}^\star$ an optimal solution of (4). Assume by contradiction that $\|\boldsymbol{w}^\star\|_2 > 2\|\boldsymbol{q}\|_2/\hat{\lambda}_R$. We have

$$
\begin{aligned}
f_2(\boldsymbol{w}^\star) &= \frac{1}{2}(\boldsymbol{w}^\star)^\top \boldsymbol{Q}\boldsymbol{w}^\star + \boldsymbol{q}^\top \boldsymbol{w}^\star + \rho(\|\boldsymbol{w}^\star\|_1 - \|\boldsymbol{w}^\star\|_K) \\
&\geq \frac{1}{2}\hat{\lambda}_R \|[\boldsymbol{w}^\star]_{I^*}\|_2^2 - \|\boldsymbol{q}\|_2 \|\boldsymbol{w}^\star\|_2,
\end{aligned}
$$

where $I^* = \underset{I:|I|=R}{\operatorname{argmin}} \lambda_{\min}([\boldsymbol{Q}]_I)$. Since I^* is a maximal index set such that $[\boldsymbol{Q}]_I$ is positive definite, we can assign 0 to $w_j^\star$ for every $j \notin I_*$, without loss of generality. Then we have

$$
f_2(\boldsymbol{w}^\star) \geq \frac{1}{2}\hat{\lambda}_R \|\boldsymbol{w}^\star\|_2^2 - \|\boldsymbol{q}\|_2 \|\boldsymbol{w}^\star\|_2 > 0.
$$

On the other hand, since $f_2(\boldsymbol{0}) = 0$, the optimal value of (4) must be non-positive. Hence it suffices to consider the case $\|\boldsymbol{w}^\star\|_2 \leq 2\|\boldsymbol{q}\|_2/\hat{\lambda}_R$. Applying $C = \hat{\lambda}_R$ to Lemma 1, we have the desired result. (End of Proof of Theorem 2)

A.3 Proof of Theorem 3

Proof. Suppose that an optimal solution $\boldsymbol{W}^\star$ of (16) satisfies $\|\boldsymbol{W}^\star\|_{\mathrm{F}} > C$. We have

$$\begin{aligned} f_3(\boldsymbol{W}^\star) &:= \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}^\star) - \boldsymbol{b}\|_2^2 + \rho(\|\boldsymbol{W}^\star\|_* - \|\!|\boldsymbol{W}^\star|\!\|_K) \\ &\geq \sum_{i=1}^{p}\left(\frac{1}{2}(\boldsymbol{A}_i \bullet \boldsymbol{W}^\star)^2 - b_i(\boldsymbol{A}_i \bullet \boldsymbol{W}^\star)\right) + \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{b} \\ &\geq \frac{1}{2}\mathrm{vec}(\boldsymbol{W}^\star)^\top \sum_{i=1}^{p}(\mathrm{vec}(\boldsymbol{A}_i)\mathrm{vec}(\boldsymbol{A}_i)^\top)\mathrm{vec}(\boldsymbol{W}^\star) \\ &\quad - \sum_{i=1}^{p}(|b_i|\|\boldsymbol{A}_i\|_{\mathrm{F}})\|\boldsymbol{W}^\star\|_{\mathrm{F}} + \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{b} \\ &\geq \hat{\lambda}_R\|[\boldsymbol{W}^\star]_{I^*}\|_{\mathrm{F}}^2 - \sum_{i=1}^{p}(|b_i|\|\boldsymbol{A}_i\|_{\mathrm{F}})\|\boldsymbol{W}^\star\|_{\mathrm{F}} + \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{b}, \end{aligned}$$

where $I^* = \underset{I:|I|=R}{\mathrm{argmin}}\, \lambda_{\min}([\sum_{i=1}^{p}\mathrm{vec}(\boldsymbol{A}_i)\mathrm{vec}(\boldsymbol{A}_i)^\top]_I)$. Since I^* is maximal, we can assign 0 to $[\mathrm{vec}(\boldsymbol{W}^\star)]_j$ for every $j \notin I^*$, without loss of generality. Then we have

$$f_3(\boldsymbol{W}^\star) \geq \hat{\lambda}_R\|\boldsymbol{W}^\star\|_{\mathrm{F}}^2 - \sum_{i=1}^{p}(|b_i|\|\boldsymbol{A}_i\|_{\mathrm{F}})\|\boldsymbol{W}^\star\|_{\mathrm{F}} + \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{b} > \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{b}.$$

On the other hand, since $f_3(\boldsymbol{O}) = \boldsymbol{b}^\top\boldsymbol{b}/2$, it suffices to consider the case $\|\boldsymbol{W}^\star\| \leq C$.

Now we show that $\|\boldsymbol{W}^\star\|_* - \|\!|\boldsymbol{W}^\star|\!\|_K = 0$. Assume by contradiction that $\|\boldsymbol{W}^\star\|_* - \|\!|\boldsymbol{W}^\star|\!\|_K > 0$. Consider a feasible solution $\tilde{\boldsymbol{W}} := \boldsymbol{W}^\star - \boldsymbol{W}^\star\boldsymbol{v}_{K+1}\boldsymbol{v}_{K+1}^\top$, where $\boldsymbol{v}_{K+1}$ is the $(K+1)$-st leading eigenvector of $\boldsymbol{W}^{\star\top}\boldsymbol{W}^\star$

with $\|\boldsymbol{v}_{K+1}\|_2 = 1$. Then we have

$$
\begin{aligned}
&f_3(\boldsymbol{W}^\star) - f_3(\tilde{\boldsymbol{W}}) \\
&= \rho\sigma_{K+1}(\boldsymbol{W}^\star) + \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}^\star) - \boldsymbol{b}\|_2^2 - \frac{1}{2}\|\mathcal{A}(\boldsymbol{W}^\star - \boldsymbol{W}^\star\boldsymbol{v}_{K+1}\boldsymbol{v}_{K+1}^\top) - \boldsymbol{b}\|_2^2 \\
&= \rho\sigma_{K+1}(\boldsymbol{W}^\star) + \sum_{i=1}^{p}\Bigg\{\mathrm{Tr}(\boldsymbol{A}_i^\top\boldsymbol{W}^\star\boldsymbol{v}_{K+1}\boldsymbol{v}_{K+1}^\top) \\
&\quad\cdot\left(\mathrm{Tr}(\boldsymbol{A}_i^\top\boldsymbol{W}^\star) - \frac{1}{2}\mathrm{Tr}(\boldsymbol{A}_i^\top\boldsymbol{W}^\star\boldsymbol{v}_{K+1}\boldsymbol{v}_{K+1}^\top) - b_i\right)\Bigg\} \\
&\geq \rho\sigma_{K+1}(\boldsymbol{W}^\star) - \sigma_{K+1}\sum_{i=1}^{p}\|\boldsymbol{A}_i\|_2\cdot\left(\|\boldsymbol{A}_i\|_{\mathrm{F}}\|\boldsymbol{W}^\star\|_{\mathrm{F}} + \frac{1}{2}\|\boldsymbol{A}_i\|_2\|\boldsymbol{W}^\star\|_2 + |b_i|\right) \\
&\geq \sigma_{K+1}(\boldsymbol{W}^\star)\left(\rho - \sum_{i=1}^{p}\|\boldsymbol{A}_i\|_2\left(C\|\boldsymbol{A}_i\|_{\mathrm{F}} + \frac{C}{2}\|\boldsymbol{A}_i\|_2 + |b_i|\right)\right) > 0.
\end{aligned}
$$

□

A.4 Exact penalty for the MIP-based formulation

Theorem 4. *Problem* (23) *is equivalent to the cardinality-constrained problem if*

$$
\rho > \max_i\left\{\frac{8\|\boldsymbol{q}\|_2}{\hat{\lambda}_R}\left(|q_i| + \frac{2\|\boldsymbol{Q}\boldsymbol{e}_i\|_2\|\boldsymbol{q}\|_2 + \|\boldsymbol{q}\|_2 Q_{ii}}{\hat{\lambda}_R}\right)\right\}.
$$

Proof. Let $f_4(\boldsymbol{w},\boldsymbol{u}) := \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{Q}\boldsymbol{w} + \boldsymbol{q}^\top\boldsymbol{w} + \rho(\mathbf{1}^\top\boldsymbol{u} - \boldsymbol{u}^\top\boldsymbol{u})$. In the same manner as the proof of Theorem 2, an optimal solution $\boldsymbol{w}^\star$ of (23) is bounded by $\|\boldsymbol{w}^\star\|_2 \leq 2\|\boldsymbol{q}\|_2/\hat{\lambda}_R$. As for M_j, we have a naive bound $M_j \geq 2\|\boldsymbol{q}\|_2/\hat{\lambda}_R$ by $M_j \geq \|\boldsymbol{w}^\star\|_\infty$, but to derive the concerned bound of ρ, we set $M_j = 4\|\boldsymbol{q}\|_2/\hat{\lambda}_R$.

It suffices to show that $\mathbf{1}^\top\boldsymbol{u}^\star - (\boldsymbol{u}^\star)^\top\boldsymbol{u}^\star = 0$ for any solution $(\boldsymbol{w}^\star,\boldsymbol{u}^\star)$. Assume by contradiction that $\mathbf{1}^\top\boldsymbol{u}^\star - (\boldsymbol{u}^\star)^\top\boldsymbol{u}^\star > 0$.

- If there exists j such that $0 < u_j^\star \leq 1/2$, by constructing a feasible

solution $(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{u}}) = (\boldsymbol{w}^\star - w_j^\star \boldsymbol{e}_j, \boldsymbol{u}^\star - u_j^\star \boldsymbol{e}_j)$, we have

$$
\begin{aligned}
& f_4(\boldsymbol{w}^\star, \boldsymbol{u}^\star) - f_4(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{u}}) \\
&= \frac{1}{2}(\boldsymbol{w}^\star)^\top \boldsymbol{Q}\boldsymbol{w}^\star + \boldsymbol{q}^\top \boldsymbol{w}^\star - \frac{1}{2}(\boldsymbol{w}^\star - w_j^\star \boldsymbol{b}e_j)^\top \boldsymbol{Q}(\boldsymbol{w}^\star - w_j^\star \boldsymbol{b}e_j) \\
&\quad - \boldsymbol{q}^\top(\boldsymbol{w}^\star - w_j^\star \boldsymbol{e}_j) + \rho(u_j^\star - (u_j^\star)^2) \\
&\geq w_j^\star \boldsymbol{e}_j^\top \boldsymbol{Q}\boldsymbol{w}^\star - \frac{1}{2} w_j^{\star 2} Q_{jj} + w_j^\star q_j + \frac{1}{2}\rho u_j^\star \\
&\quad (\because x - x^2 \geq x/2 \text{ for } 0 < x \leq 1/2) \\
&\geq \frac{1}{2}\rho u_j^\star - |w_j^\star| \|\boldsymbol{Q}\boldsymbol{e}_j\|_2 \|\boldsymbol{w}^\star\|_2 - \frac{1}{2}|w_j^\star| \|\boldsymbol{w}^\star\|_2 Q_{jj} - |w_j^\star||q_j| \\
&\geq \frac{\rho}{2M_j}|w_j^\star| - \frac{2|w_j^\star| \|\boldsymbol{Q}\boldsymbol{e}_j\|_2 \|\boldsymbol{q}\|_2}{\hat{\lambda}_R} - \frac{|w_j^\star| \|\boldsymbol{q}\|_2 Q_{jj}}{\hat{\lambda}_R} - |w_j^\star||q_j| \\
&\geq \frac{|w_j^\star|}{2M_j}\left[\rho - \frac{8\|\boldsymbol{q}\|_2}{\hat{\lambda}_R}\left(|q_j| + \frac{2\|\boldsymbol{Q}\boldsymbol{e}_j\|_2 \|\boldsymbol{q}\|_2 + \|\boldsymbol{q}\|_2 Q_{jj}}{\hat{\lambda}_R}\right)\right] > 0.
\end{aligned}
$$

- Otherwise, if $u_j^\star \notin \{0, 1\}$, then $1/2 < u_j^\star < 1$.
  - Case: $\mathbf{1}^\top \boldsymbol{u}^\star < K$. We can take $\epsilon > 0$ such that $\mathbf{1}^\top \boldsymbol{u}^\star + \epsilon \leq K$ and $u_j^\star + \epsilon \leq 1$. Then by putting $(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{u}}) = (\boldsymbol{w}^\star, \boldsymbol{u}^\star + \epsilon \boldsymbol{e}_j)$, we have

    $$
    \begin{aligned}
    & f_4(\boldsymbol{w}^\star, \boldsymbol{u}^\star) - f_4(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{u}}) \\
    &= \rho[\mathbf{1}^\top \boldsymbol{u}^\star - (\boldsymbol{u}^\star)^\top \boldsymbol{u}^\star - \mathbf{1}^\top(\boldsymbol{u}^\star + \epsilon \boldsymbol{e}_j) + (\boldsymbol{u}^\star + \epsilon \boldsymbol{e}_j)^\top(\boldsymbol{u}^\star + \epsilon \boldsymbol{e}_j)] \\
    &= \rho(-\epsilon + 2\epsilon u_j^\star + \epsilon^2) > 0.
    \end{aligned}
    $$

  - Case: $\mathbf{1}^\top \boldsymbol{u}^\star = K$. There exists i $(\neq j)$ such that $1/2 < u_i^\star < 1$. Assume $u_i^\star \geq u_j^\star$ without loss of generality. If we choose $\epsilon > 0$ such that $u_i^\star + \epsilon \leq 1$ and $u_j^\star - \epsilon \geq 1/2$, a solution $(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{u}}) = (\boldsymbol{w}^\star, \boldsymbol{u}^\star + \epsilon \boldsymbol{e}_i - \epsilon \boldsymbol{e}_j)$ is feasible, since $M_j(u_j^\star - \epsilon) \geq M_j/2 = 2\|\boldsymbol{q}\|_2/\hat{\lambda}_R \geq |w_j^\star|$. Then we have

    $$
    \begin{aligned}
    & f_4(\boldsymbol{w}^\star, \boldsymbol{u}^\star) - f_4(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{u}}) \\
    &= \rho(-(\boldsymbol{u}^\star)^\top \boldsymbol{u}^\star + (\boldsymbol{u}^\star + \epsilon \boldsymbol{e}_i - \epsilon \boldsymbol{e}_j)^\top(\boldsymbol{u}^\star + \epsilon \boldsymbol{e}_i - \epsilon \boldsymbol{e}_j)) \\
    &= \rho(2\epsilon u_i^\star - 2\epsilon u_j^\star + 2\epsilon^2) > 0.
    \end{aligned}
    $$

□

References

[1] F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5(1):13–51, 1995.

[2] T.S. Arthanari and Y. Dodge. *Mathematical programming in statistics*, volume 341. Wiley New York, 1981.

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems.

[4] D. Bertsimas and M. Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

[5] D. Bertsimas, D. Pachamanova, and M. Sim. Robust linear optimization under general norms. *Operations Research Letters*, 32(6):510–516, 2004.

[6] D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *arXiv preprint arXiv:1507.03133*, 2015.

[7] R. Bhatia. Matrix analysis, volume 169 of graduate texts in mathematics, 1997.

[8] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/\~mlearn/MLRepository.html`.

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

[10] P.S. Bradley and O.L. Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[11] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[12] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[13] X. Cai, F. Nie, and H. Huang. Exact top-k feature selection via l2,0-norm constraint. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[14] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[15] E.J. Candès and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[16] J. Dattorro. *Convex optimization & Euclidean distance geometry*. Lulu. com, 2010.

[17] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[18] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of International Conference on Machine Learning*, volume 28, pages 37–45, 2013.

[19] J. Gotoh and S. Uryasev. Two pairs of families of polyhedral norms versus ℓ_p-norms: proximity and applications in optimization. *Mathematical Programming*, 2015, OnlineFirst.

[20] N. Gulpinar, L.T.H. An, and M. Moeini. Robust investment strategies with discrete asset choice constraints using dc programming. *Optimization*, 59(1):45–62, 2010.

[21] V.V. Nguyen H. A. Le Thi, M. Le Hoai and T. Pham Dinh. A dc programming approach for feature selection in support vector machines learning. *Advances in Data Analysis and Classification*, 2(3):259–278, 2008.

[22] A.B. Hempel and P.J. Goulart. A novel method for modelling cardinality and rank constraints. In *IEEE Conference on Decision and Control*, pages 4322–4327, Los Angeles, USA, December 2014. URL `http://control.ee.ethz.ch/index.cgi?page=publications;action=details;id=4712`.

[23] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches*. Springer-Verlag, Berlin, 3rd edition, 1996.

[24] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(9):2117–2130, 2013.

[25] H. A. Le Thi and T. Pham Dinh. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1-4):23–46, 2005.

[26] H. A. Le Thi and T. Pham Dinh. Dc approximation approaches for sparse optimization. *European Journal of Operational Research*, 244 (1):26–46, 2015.

[27] C. Lu, J. Tang, S. Yan, and Z. Lin. Generalized nonconvex nonsmooth low-rank minimization. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 4130–4137. IEEE, 2014.

[28] R. Miyashiro and Y. Takano. Mixed integer second-order cone programming formulations for variable selection in linear regression. *European Journal of Operational Research*, 2015.

[29] R. Miyashiro and Y. Takano. Subset selection by mallows 'cp: A mixed integer programming approach. *Expert Systems with Applications*, 42 (1):325–331, 2015.

[30] B. Moghaddam, Y. Weiss, and S. Avidan. Generalized spectral bounds for sparse lda. In *Proceedings of the 23rd international conference on Machine learning*, pages 641–648. ACM, 2006.

[31] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[32] Y. Nesterov. Gradient methods for minimizing composite objective function. core discussion papers.

[33] T.B.T. Nguyen, H.A. Le Thi, H.M. Le, and X.T. Vo. Dc approximation approach for ℓ_0-minimization in compressed sensing. In *Advanced Computational Methods for Knowledge Engineering*, pages 37–48. Springer, 2015.

[34] P.D. Nhat, M.C. Nguyen, and H.A. Le Thi. A dc programming approach for sparse linear discriminant analysis. In *Advanced Computational Methods for Knowledge Engineering*, pages 65–74. Springer, 2014.

[35] M.L. Overton and R.S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(1-3):321–357, 1993.

[36] K. Pavlikov and S. Uryasev. Cvar norm and applications in optimization. *Optimization Letters*, 8(7):1999–2020, 2014.

[37] T. Pham Dinh and H. A. Le Thi. Recent advances in dc programming and dca. *Transactions on Computational Collective Intelligence*, 8342: 1–37, 2014.

[38] T. Pham Dinh and H.A. Le Thi. Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

[39] T. Pham Dinh, H.A. Le Thi, and Le D. Muu. Exact penalty in d.c. programming. *Vietnam Journal of Mathematics*, 27(2):169–178, 1999.

[40] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[41] S.K. Shevade and S.S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17): 2246–2253, 2003.

[42] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 325–332, 2005.

[43] B. K. Sriperumbudur and G. R. G. Lanckriet. A proof of convergence of the concave-convex procedure using zangwill's theory. *Neural Computation*, 24(6):1391–1407, 2012.

[44] A. Takeda, M. Niranjan, J. Gotoh, and Y. Kawahara. Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Computational Management Science*, 10(1):21–49, 2013.

[45] P.D. Tao and An L.T.H. A d.c. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

[46] H.A. Le Thi, T. Pham Dinh, and N. Dong Yen. Properties of two dc algorithms in quadratic programming. *Journal of Global Optimization*, 49(3):481–495, 2011.

[47] M. Thiao, P.D. Tao, and L.T. An. A dc programming approach for sparse eigenvalue problem. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1063–1070, 2010.

[48] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[49] G. A. Watson. Linear best approximation using a class of polyhedral norms. *Numerical Algorithms*, 2(3):321–335, 1992.

[50] G. A. Watson. On matrix approximation problems with ky fan k norms. *Numerical Algorithms*, 5(5):263–272, 1993.

[51] B. Wu, C. Ding, D.F. Sun, and K.C. Toh. On the moreau-yoshida regularization of the vector k-norm related functions. *SIAM Journal on Optimization*, 24:766–794, 2014.

[52] Ding C. Sun D. Toh K.-C. Wu, B. On the moreau–yosida regularization of the vector k-norm related functions. *SIAM Journal on Optimization*, 24(2):766–794, 2014.

[53] X. Zheng, X. Sun, D. Li, and J. Sun. Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear dc approach. *Computational Optimization and Applications*, 59(1-2):379–397, 2014.

[54] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.