Lagrangian Duality and Saddle Points for Sparse Linear Programming

Chen Zhao* Ziyan Luo† Weiyue Li ‡ Houduo Qi § Naihua Xiu ¶

February 18, 2019

Abstract

The sparse linear programming (SLP) is a linear programming problem equipped with a sparsity constraint, which is nonconvex, discontinuous and generally NP-hard to solve due to the combinatorial property involved. In this paper, by rewriting the sparsity constraint into a disjunctive form, we present an explicit formula of Lagrangian dual problem for the SLP, in terms of an unconstrained piecewise-linear convex programming problem which admits a strong duality. Furthermore, we show a saddle point theorem based on the strong duality and give two first-order necessary and sufficient optimality conditions for the saddle point problem without any constraint qualification for SLP. Additionally, as an extension, we show that these main results are also valid for the case when the nonnegative box constraint $[0, u]$ in the SLP is replaced by $[-u, u]$ for any positive vector u.

Key words. Sparse linear programming, Lagrangian dual problem, Strong duality, Saddle point theorem, Optimality condition.

AMS subject classifications. 190C26 90C30 90C46

1 Introduction

In recent decades, sparsity has attracted a great deal of attention in a wide range of applications including compressed sensing in signal processing [1], the variable selection [2] and the best subset selection [3] in high-dimensional statistical regression and machine learning, the sparse portfolio in finance [4], the network design [5], the data and knowledge intelligence [6], etc. Sparsity constrained optimization problems (SCOP) aim to minimize a general continuously differentiable function subject to sparsity, equality and inequality constraints. It is generally NP-hard due to the intrinsic combinatorial property possessed by the sparsity constraint in terms of the zero-norm. Thus, classical continuous optimization theory is unable to cope with such a problem and new theory tailored for SCOP is in need. Most of the existing theoretical analysis of SCOP is concentrated on the optimality conditions. For example, Beck et al. [7] presented three kinds of first-order necessary optimality conditions for

*Corresponding author; Department of Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China; 14118409@bjtu.edu.cn.

†State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, P. R. China; starkey-nature@hotmail.com.

‡Department of Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China; 13121542@bjtu.edu.cn.

§School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK; hdqi@soton.ac.uk.

¶Department of Mathematics, Beijing Jiaotong University, Beijing 100044, P. R. China; nhxiu@bjtu.edu.cn.

problem of SCOP merely with the sparsity constraint. Later, by exploring projection properties on the so-called symmetric sets, Beck et al. [8] extended the above results to the problem of minimizing a general continuously differentiable function over sparse and symmetric sets. Based on the Robinson constraint qualification from nonlinear programming, Lu et al. [9] introduced a first-order necessary optimality condition for general SCOP. For the SCOP problem equipped with sparse and polyhedral constraints, Li et al. [10] presented the first-order necessary conditions based on the expressions of the Fréchet and Mordukhovich normal cones to such a special feasible set. Decomposition properties of different types of normal cones and tangent cones to the intersection of the sparse set and some special structured sets were further characterized to establish first- and second-order optimality conditions for SCOP with nonnegativity constraints in [11], general nonlinear equality and inequality constraints in [12], and even conic constraints in [13], besides the sparsity constraints. More recently, by reformulating the sparsity constraint into a set of relatively complicated but smooth constraints, Bucher and Schwartz [14] gave some new optimality conditions with the help of some well-known constraint qualifications from the traditional nonlinear programming.

In this paper, we mainly focus on a special case of SCOP, the sparse linear programming (SLP) problem which is to minimize a linear function subject to linear equality, box and sparsity constraints. Its mathematical model takes the form of

$$(\mathcal{P}) \qquad \min\ c^T x, \quad \text{s.t. } Ax = b,\ 0 \le x \le u,\ \|x\|_0 \le r,$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ $c \in \mathbb{R}^n$, u is a positive vector in $\mathbb{R}^n$, $0 < r \le n$ is a positive integer and $\|x\|_0$ is the l_0 norm of x which counts the number of nonzero components in x. Assume throughout the paper that the feasible region of $(\mathcal{P})$ is nonempty, which leads to the solvability of $(\mathcal{P})$ (i.e., $(\mathcal{P})$ has at least one optimal solution) due to the box constraint involved. SLP has been widely employed to many fields, such as, the reformulation of linear compressed sensing [1], transportation problems [16], Sudoku games [17], and so on. It also resembles the problems of finding the sparsest solution of a linear programming problem [18, 19] or a linear complementarity problem [20, 21], but not exactly so. In contrast to latter cases, SLP is more applicable to the circumstances when people demand a vector with a certain degree of sparsity rather than the sparsest one. It is worth mentioning that the nonnegativity involved in the box constraint of $(\mathcal{P})$ is motivated from the prior information from the aforementioned application problems, along with the consideration that the standard linear programming problems are always equipped with nonnegativity.

With the special structure in $(\mathcal{P})$, it is hopeful that more properties can be explored for SLP than those for general SCOP. In this paper, we aim to give an explicit form of the Lagrangian dual problem for SLP and then emphasize on the study of the corresponding strong duality theorem. Distinct from the traditional duality theory for linear programming (LP), the nonconvexity of SLP embedded in the additional sparsity constraint makes things not straightforward any longer. A natural question is, can we write out an explicit form of the Lagrangian dual problem and further achieve the strong duality for SLP without any constraint qualification (CQ), just like the case for LP [22]. Normally, even for smooth nonlinear programming, the Lagrangian dual problem cannot be formulated explicitly, and the desired strong duality fails if no CQ is required. Interestingly, by rewriting the sparsity constraint into a disjunctive form, we can derive an explicit expression of the Lagrangian dual problem of SLP. By taking the advantage of such an explicit form of the dual problem, we can show the strong duality without any CQ for the nonconvex discontinuous problem SLP. To our best knowledge, this is the first time that the strong duality theorem with no CQ is achieved for SCOP.

We summarize our main results below and explain why the strong duality is not end of the story. The strong duality result (i.e., zero duality gap between the primal and the Lagrange dual problems)

without any CQs is stated in Thm. 3.1. Two more nice features of the dual problem is that the objective has an explicit formulation and it is piecewise linear over the whole dual space. Furthermore, the strong duality allows us to recover the primal optimal solution for some important cases, see Thm. 3.5. In other words, the strong duality only provides a partial answer to the original (NP-hard) problem. In order to provide a complete characterization of the primal and dual optimal solutions, we develop a saddle point theorem (Thm. 4.3). This further motivates us to study first-order optimality conditions for the saddle point problem. In particular, we propose the β- and $\widehat{N}$-stationary points and investigate their relationship to the primal and dual optimal solutions of the original problem (see Thm. 4.6). As an extension, the box constraint of the form $-u \leq x \leq u$ for any given positive vector u is also considered in the setting of SLP, which is shown to inherit all aforementioned nice properties following a similar routine with subtle modifications.

This paper is organized as follows. The model of SLP and the explicit formula of its Lagrangian dual are stated in Section 2. The strong duality theorem is then established in Section 3. We present saddle point theorem and two first-order necessary optimality conditions for the SLP in Section 4. An extension with a different box constraint in the SLP is proposed in Section 5. Conclusions are drawn in Section 6.

Notation For convenience, the following notations will be used throughout the paper. Denote the set of all nonnegative (positive) vectors in $\mathbb{R}^n$ by $\mathbb{R}^n_+$ ($\mathbb{R}^n_{++}$). For any given closed set $\Omega \subseteq \mathbb{R}^n$ and any $x \in \mathbb{R}^n$, define $\Pi_\Omega(x) := \arg\min_{y\in\Omega} \|x-y\|^2$. The l_1 norm, the l_2 norm and the projection onto the nonnegative orthant $\mathbb{R}^n_+$ of x are defined as $\|x\|_1 := \sum_{i=1}^n |x_i|$, $\|x\|_2 := \left(\sum_{i=1}^n x_i^2\right)^{1/2}$ and $[x]_+ := \max\{x, 0\}$, respectively. For any given positive integer n, denote $[n] := \{1, \ldots, n\}$. For any index set $I \subseteq [n]$, x_I represents the subvector of x indexed by index set I. For any given vector $d \in \mathbb{R}^n$, $\mathrm{Diag}(d)$ stands for the diagonal matrix whose diagonal elements is generated by d. Let $x^\downarrow$ be the vector of entries of x being arranged in the non-increasing order $x_1^\downarrow \geq \cdots \geq x_n^\downarrow$. The symbol $\circ$ stands for the Hadamard product, i.e., for vectors $u, v \in \mathbb{R}^n$, $(u \circ v)_i = u_i v_i$, $i \in [n]$. For any given optimization problem labeled by (P), we denote its optimal value by Val(P) and its optimal solution set by S^*_P. Finally, we denote the convex hull of Ω by $\mathrm{conv}(\Omega)$.

2 The Lagrangian Dual Problem

Given $u \in \mathbb{R}^n_{++}$ and any positive integer $r \leq n$, denote

$$S(r) := \{x \in \mathbb{R}^n | \|x\|_0 \leq r\},\ C(u; r) := \{x \in \mathbb{R}^n | 0 \leq x \leq u, x \in S(r)\}.$$

Problem $(\mathcal{P})$ can be rewritten as

$$\min_{x\in C(u;r)} \{c^T x \mid Ax = b\}.$$

By relaxing the linear equality constraint, we define the Lagrangian function of $(\mathcal{P})$ as

$$L(x, y) = c^T x - y^T(Ax - b), \quad \forall x \in C(u; r), y \in \mathbb{R}^m.$$

Then the corresponding Lagrangian dual problem takes the form of

$$\sup_{y\in\mathbb{R}^m} \left\{\theta(y) := \min_{x\in C(u,r)} L(x, y)\right\}. \tag{2.1}$$

It is worth pointing out that with different ways of relaxation for the constraints we can get different formulas of dual problems for the same primal problem. For example, in [15], the sparsity constraint is

relaxed to get a discontinuous Lagrangian function involving ℓ_0-norm and then an implicit formula for the corresponding Lagrange dual problem. Here, the constraints embedded in $C(u;r)$ are treated as an indicator function and the linear equality constraint $Ax = b$ is relaxed to construct the Lagrangian function which leads to an explicit formula of the corresponding Lagrange dual. The remainder of this section is devoted to the explicit form of $\theta(y)$ as defined in (2.1). Before proceeding, several essential lemmas are stated for preparation.

Note that $(\mathcal{P})$ is actually a disjunctive programming since the feasible region can be rewritten as

$$\bigcup_{I\in\mathcal{J}(r)} \{x \in \mathbb{R}^n_I \mid Ax = b, 0 \le x \le u\} := \bigcup_{I\in\mathcal{J}(r)} F_I,$$

where $\mathcal{J}(r) = \{J \subseteq \{1,\ldots,n\} \mid |J| = r\}$, $\mathbb{R}^n_I$ is the subspace of $\mathbb{R}^n$ spanned by $\{e_i \mid i \in I\}$ with $e_i \in \mathbb{R}^n$ being the i-th column of the identity matrix. Thus, $(\mathcal{P})$ can be cast as

$$\min_{I\in\mathcal{J}(r)} \ \min_{x\in F_I} c^T x. \tag{2.2}$$

For any given $I \in \mathcal{J}(r)$, the corresponding inner minimization problem

$$(\mathcal{P}_I) \qquad \min_{x\in F_I} c^T x$$

is a linear program and hence it admits the strong duality. By utilizing the following lemma, we can get an explicit form of the dual problem of $(\mathcal{P}_I)$.

Lemma 2.1. *Given $I \subseteq [n]$, $p \in \mathbb{R}^n$ and $u \in \mathbb{R}^n_{++}$, we have*

$$\min\left\{p^T x : 0 \le x \le u, x \in \mathbb{R}^n_I\right\} = -\|\Pi_{\mathbb{R}^n_I\cap\mathbb{R}^n_+}(-u\circ p)\|_1. \tag{2.3}$$

Proof. Set $x^*_i = u_i$, if $i \in I$ and $p_i < 0$, and $x^*_i = 0$ otherwise. Then

$$p^T x^* = \sum_{i\in I,\ p_i<0} u_i p_i = -\sum_{i\in I}[-u_i p_i]_+ = -\|\Pi_{\mathbb{R}^n_I\cap\mathbb{R}^n_+}(-u\circ p)\|_1. \tag{2.4}$$

Thus, the desired equality can be obtained from that for any $x \in \mathbb{R}^n_I \cap [0,u]$,

$$p^T x = \sum_{i\in I} x_i([p_i]_+ - [-p_i]_+) \ge -\sum_{i\in I} u_i[-p_i]_+ = p^T x^*.$$

This completes the proof. □

Lemma 2.2. *The dual problem of $(\mathcal{P}_I)$ is*

$$(\mathcal{D}_I) \qquad \max_{y\in\mathbb{R}^m} b^T y - \|\Pi_{\mathbb{R}^n_I\cap\mathbb{R}^n_+}[u\circ(A^T y - c)]\|_1. \tag{2.5}$$

Proof. The desired formula follows directly from Lemma 2.1 by taking $p = c - A^T y$. □

For later analysis, we review the explicit expression of projection on $S(r)\cap\mathbb{R}^n_+$ in Lemma 2.3 below which is a well known conclusion and has been mentioned in several papers [11].

Lemma 2.3. *Let $z \in \mathbb{R}^n$ and $\pi \in \Pi_{S(r)\cap\mathbb{R}^n_+}(z)$. Then $\pi_{t_i} = [z_{t_i}]_+$ for all $i \in [r]$, and $\pi_{t_i} = 0$ otherwise, where $\{t_1, t_2, \ldots, t_n\}$ satisfies $z_{t_1} \ge z_{t_2} \ge \ldots \ge z_{t_n}$. Furthermore, $\|\pi\|_1 = \sum_{i=1}^{r}[z_{t_i}]_+ = \sum_{i=1}^{r}([z]_+)^{\downarrow}_i$, for any $\pi \in \Pi_{S(r)\cap\mathbb{R}^n_+}(z)$.*

As indicated in Lemma 2.3, all projection vectors of z onto $S(r) \cap \mathbb{R}^n_+$ share the same ℓ_1 norm, which allows us to simply use the notation $\|\Pi_{S(r)\cap\mathbb{R}^n_+}(z)\|_1$ for the involved identical ℓ_1 norm without ambiguity. Combining with Lemma 2.1, this notation can further help characterize the optimal value for a relevant nonconvex optimization problem of $(\mathcal{P})$ as shown by the following theorem.

Theorem 2.4. *Given $p \in \mathbb{R}^n$, $u \in \mathbb{R}^n_{++}$ and an integer r $(0 < r \leq n)$, we have*

$$\min\left\{p^T x \mid x \in C(u;r)\right\} = -\|\Pi_{S(r)\cap\mathbb{R}^n_+}(-u \circ p)\|_1. \tag{2.6}$$

Furthermore, x^* is an optimal solution of Problem (2.6) if and only if for any $\lambda > 0$, x^* satisfies $x^* \in \Pi_{C(u;r)}(x^* - \lambda p)$, i.e.

$$S^*(p) := \arg\min\left\{p^T x \mid x \in C(u;r)\right\} = \left\{x^* \in \mathbb{R}^n \mid x^* \in \Pi_{C(u;r)}(x^* - \lambda p), \forall \lambda > 0\right\}. \tag{2.7}$$

Proof. The desired assertion can be derived by

$$\begin{aligned}
\min_{x\in C(u;r)} p^T x &= \min_{I\in\mathcal{J}(r)} \min_{x\in\mathbb{R}^n_I, 0\leq x\leq u} p^T x = \min_{I\in\mathcal{J}(r)} -\|\Pi_{\mathbb{R}^n_I\cap\mathbb{R}^n_+}(-u\circ p)\|_1 \\
&= -\max_{I\in\mathcal{J}(r)} \sum_{i\in I} [-u_i p_i]_+ = -\sum_{i=1}^{r} ([-u\circ p]_+)^{\downarrow}_i \\
&= -\left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(-u\circ p)\right\|_1,
\end{aligned} \tag{2.8}$$

where the first equality is from $S(r) = \bigcup_{I\in\mathcal{J}(r)} \mathbb{R}^n_I$, the second and the last equalities from Lemmas 2.1 and 2.3, and the fourth equality from (2.4). The following part is devoted to prove the conclusion of optimal solution.

Suppose that x^* is an optimal solution of (2.6), then $p^T x^* \leq p^T x$ is true for any $x \in C(u;r)$. Furthermore, for any $\lambda > 0$,

$$\begin{aligned}
\|x - (x^* - \lambda p)\|^2 &= \|x - x^*\|^2 + \|\lambda p\|^2 + 2\langle x - x^*, \lambda p\rangle \\
&\geq 2\lambda\langle x - x^*, p\rangle + \|\lambda p\|^2 \\
&\geq \|\lambda p\|^2 = \|x^* - (x^* - \lambda p)\|^2.
\end{aligned}$$

According to the definition of projection, we have $x^* \in \Pi_{C(u;r)}(x^* - \lambda p)$.

Now let us consider another side of the conclusion. Suppose that for any $\lambda > 0$, x^* satisfy $x^* \in \Pi_{C(u;r)}(x^* - \lambda p)$.

$$\begin{aligned}
\|x^* - (x^* - \lambda p)\|^2 &\leq \|x - (x^* - \lambda p)\|^2 \\
\|\lambda p\|^2 &\leq \|x - x^*\|^2 + \|\lambda p\|^2 + 2\langle x - x^*, \lambda p\rangle \\
2\lambda\langle x^* - x, p\rangle &\leq \|x - x^*\|^2
\end{aligned} \tag{2.9}$$

Assume that x^* is not the minimum of (2.6), then

$$\langle x^* - \widehat{x}, p\rangle > 0, \tag{2.10}$$

where $\widehat{x}$ is the optimal solution of (2.6). Taking $x = \widehat{x}$ in (2.9), we render

$$2\lambda\langle x^* - \widehat{x}, p\rangle \leq \|\widehat{x} - x^*\|^2 \tag{2.11}$$

Due to the arbitrariness of λ, (2.10) contradicts (2.11). □

Remark 2.5. *Theorem 2.4 shows that, for the special sparse linear programming problem (2.6), the optimal solution x^* can be obtained from its stablility condition (i.e. first-order optimality condition): $x^* \in \Pi_{C(u;r)}(x^* - \lambda p)$ and the arbitrariness of $\lambda > 0$ is essential. If there exists some $\lambda > 0$ such that $x^* \in \Pi_{C(u;r)}(x^* - \lambda p)$, x^* may not be the minimizer. For example, let $r = 2$, $p = (-2, -1, -1, 0)^T$, $u = (1, 1, 1, 1)^T$, $\lambda_1 = 0.5$, $\lambda_2 = 1$. Then $x^* = (0, 1, 1, 0)^T$ satisfies $x^* \in \Pi_{C(u;r)}(x^* - \lambda_1 q^*)$, $x^* \notin \Pi_{C(u;r)}(x^* - \lambda_2 q^*)$, and obviouly $x^* \neq \arg\min_{x \in C(u;r)} p^T x$.*

Now we are in a position to explicitly formulate the dual problem (2.1).

Theorem 2.6. *The Lagrangian dual problem* (2.1) *can be explicitly formulated as the following unconstrained convex program*

$$(\mathcal{D}) \qquad \max_{y \in \mathbb{R}^m} \theta(y) = b^T y - \|\Pi_{S(r) \cap \mathbb{R}^n_+}(u \circ (A^T y - c))\|_1.$$

Moreover, $\theta(\cdot)$ is piecewise linear over $\Re^m$. Consequently, y^* is an optimal solution of $(\mathcal{D})$ if and only if $b \in A\mathrm{conv}(S^*(c - A^T y^*))$.

Proof. It follows from the definition of $\theta(y)$ in (2.1) that

$$\theta(y) = \min_{x \in C(u,r)} \left\{ c^T x - y^T (Ax - b) \right\} = b^T y + \min_{x \in C(u,r)} (c - A^T y)^T x.$$

Then the formula of $(\mathcal{D})$ follows readily from Theorem 2.4. To verify the convexity of $-\theta(\cdot)$, one can regard $-\theta(\cdot)$ as the pointwise supremum of a collection of affine functions (and hence convex) in the way of

$$-\theta(y) = \max_{x \in C(u;r)} \left\{ \left(c - A^T y\right)^T x + b^T y \right\}.$$

and the convexity follows directly from [23, Theorem 5.5]. Since $\min_{y \in \mathbb{R}^m} \{-\theta(y)\}$ is an unconstrained convex program, the first-order optimality condition then indicates that y^* is an optimal solution of $(\mathcal{D})$ if and only if $0 \in \partial\left(-\theta(y^*)\right)$. To explicitly characterize such a sub-differential, we first claim that $-\theta(\cdot)$ is piecewise linear over $\mathbb{R}^m$. Denote by $\mathbf{\Pi}$ the set of all permutation matrices in $\mathbb{R}^{n \times n}$ and by

$$\mathbf{\Pi}_r = \left\{ B \circ P \mid B \in \{0,1\}^{n \times n}, P \in \mathbf{\Pi}, \|B \circ P\|_0 \leq r \right\}.$$

The cardinality (the number of distinct members) of $\mathbf{\Pi}_r$ is less than $\binom{n}{r} \cdot 2^r \cdot n!$ (finitely many). In view of

$$g(p) := \|\Pi_{S(r) \cap \mathbb{R}^n_+}(-u \circ p)\|_1 = \max_{M \in \mathbf{\Pi}_r} \left\{ -e^T M U p \right\},$$

where $e \in \mathbb{R}^n$ is the all-one vector and U is the diagonal matrix generated from u, the piecewise linearity of g over $\mathbb{R}^n$ is then achieved (see, e.g., [26, Chapter 2] or [27, Section 2]). Armed with the function g, we can rewrite $-\theta$ as

$$-\theta(y) = -b^T y + g\left(c - A^T y\right).$$

Thus, by applying [25, Exercise 10.22 (b)], we can show that $-\theta(\cdot)$ is convex and piecewise linear over $\mathbb{R}^m$ and

$$\partial\left(-\theta(y)\right) = -b - A\partial g\left(c - A^T y\right). \tag{2.12}$$

On the other hand, according to Theorem 2.4, $g(p) = \max\left\{-p^T x \mid x \in C(u;r)\right\}$. By employing the sub-differential of supremum functions addressed in [24, Theorem 2.93], one has

$$\partial g(p) = \mathrm{conv} \bigcup_{\bar{x} \in \{x \mid g(p) = -p^T x, x \in C(u;r)\}} (-\bar{x}) = \mathrm{conv} \bigcup_{\bar{x} \in S^*(p)} (-\bar{x}) = \mathrm{conv}\left(-S^*(p)\right), \tag{2.13}$$

where the second equality is due to (2.7). Invoking the display in (2.12), we can conclude that

$$\partial\left(-\theta(y)\right) = -b + A\mathrm{conv}(S^*(c - A^T y)).$$

Thus, the optimality condition $0 \in \partial\left(-\theta(y^*)\right)$ is equivalent to $b \in A\mathrm{conv}\left(S^*(c - A^T y^*)\right)$. This completes the proof. □

3 Strong Duality

This section is devoted to the strong duality between $(\mathcal{P})$ and $(\mathcal{D})$.

Theorem 3.1. (Strong Duality Theorem) *If $(\mathcal{P})$ (or $(\mathcal{D})$) is solvable, then so is $(\mathcal{D})$ (or $(\mathcal{P})$) and the duality gap is zero.*

Proof. "$S^*_{\mathcal{P}} \neq \emptyset \Rightarrow S^*_{\mathcal{D}} \neq \emptyset$": Assume that x^* is an optimal solution of $(\mathcal{P})$. The weak duality indicates that $\mathrm{Val}(\mathcal{D}) \leq c^T x^*$. Note that $\theta(\cdot)$ is proper and hence $\theta(y) > -\infty$ for all $y \in \mathbb{R}^m$. Thus $\mathrm{Val}(\mathcal{D})$ is finite. Now we claim that this finite supremum is attainable. Assume on the contrary that $\theta(y)$ can not reach $\mathrm{Val}(\mathcal{D})$ at any finite $y \in \mathbb{R}^m$. Then there exist $y_0 \in \mathbb{R}^m$ and $d \in \mathbb{R}^m \backslash \{0\}$ such that

$$\mathrm{Val}(\mathcal{D}) = \lim_{\lambda \to +\infty} \theta(y_0 + \lambda d). \tag{3.14}$$

Denote the following three index sets with respect to the signs of components in the vector $u \circ A^T d$:

$$I_+ = \left\{i \in [n] \mid \left(u \circ A^T d\right)_i > 0\right\},\ I_0 = \left\{i \in [n] \mid \left(u \circ A^T d\right)_i = 0\right\},\ \text{and } I_- = \left\{i \in [n] \mid \left(u \circ A^T d\right)_i < 0\right\}.$$

Denote by Σ the set of all permutations of $[n]$ such that for any $\{t_1, \ldots, t_n\} \in \Sigma$,

$$\left(u \circ A^T d\right)_{t_1} \geq \cdots \geq \left(u \circ A^T d\right)_{t_n}. \tag{3.15}$$

Such an order can be certainly preserved by multiplying $u \circ A^T d$ with any positive $\alpha \in \mathbb{R}$. Set $v := u \circ (c - A^T y_0)$. Then we can always find a permutation $\{t'_1, \ldots, t'_n\} \in \Sigma$ such that $v_{t'_i} \leq v_{t'_{i+1}}$ provided that $\left(u \circ A^T d\right)_{t'_i} = \left(u \circ A^T d\right)_{t'_{i+1}}$. This can be achieved since if $v_{t'_i} > v_{t'_{i+1}}$, then by exchanging the positions of t'_i and t'_{i+1}, the resulting new permutation still lies in Σ. Furthermore, define

$$\gamma_1 := \max_{\substack{\left(u \circ A^T d\right)_{t'_i} > \left(u \circ A^T d\right)_{t'_{i+1}} \\ i = 1, \cdots, n-1}} \left\{\frac{v_{t'_i} - v_{t'_{i+1}}}{(u \circ A^T d)_{t'_i} - (u \circ A^T d)_{t'_{i+1}}}\right\},\ \text{and } \gamma_2 := \max_{k \in I_+ \cup I_-} \left\{\frac{v_k}{(u \circ A^T d)_k}\right\}.$$

Combining with (3.15), we have that for any given $\lambda_0 > \max\{0, \gamma_1, \gamma_2\}$ and any given $\alpha > 0$,

$$\begin{cases} (u \circ (A^T(y_0 + \lambda_0 d) - c))_i > 0, \quad \alpha\left(u \circ A^T d\right)_i > 0, \quad \forall i \in I_+, \\ (u \circ (A^T(y_0 + \lambda_0 d) - c))_k < 0, \quad \alpha\left(u \circ A^T d\right)_k < 0, \quad \forall k \in I_-, \\ \left(u \circ \left(A^T(y_0 + \lambda_0 d) - c\right)\right)_{t_1} \geq \cdots \geq \left(u \circ \left(A^T(y_0 + \lambda_0 d) - c\right)\right)_{t_n}, \alpha\left(u \circ A^T d\right)_{t_1} \geq \cdots \geq \alpha\left(u \circ A^T d\right)_{t_n}. \end{cases} \tag{3.16}$$

The properties in (3.16) further lead to the following decomposition equality for any $\alpha > 0$:

$$\begin{aligned} &\left\|\Pi_{S(r) \cap \mathbb{R}^n_+}\left(u \circ \left(A^T\left(y_0 + \lambda_0 d\right) - c\right) + \alpha\left(u \circ A^T d\right)\right)\right\|_1 \\ =& \left\|\Pi_{S(r) \cap \mathbb{R}^n_+}\left(u \circ (A^T(y_0 + \lambda_0 d) - c)\right)\right\|_1 + \alpha\left\|\Pi_{S(r) \cap \mathbb{R}^n_+}\left(u \circ A^T d\right)\right\|_1. \end{aligned} \tag{3.17}$$

Note that for any $\lambda > \lambda_0$,

$$\begin{aligned}
&\theta(y_0 + \lambda d) \\
&= b^T(y_0 + \lambda d) - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ (A^T(y_0 + \lambda d) - c))\right\|_1 \\
&= b^T(y_0 + \lambda_0 d) + (\lambda - \lambda_0) b^T d - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}\left((\lambda - \lambda_0)\left(u \circ A^T d\right)\right)\right\|_1 \\
&\quad - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ (A^T(y_0 + \lambda_0 d) - c))\right\|_1 \\
&= \theta(y_0 + \lambda_0 d) + (\lambda - \lambda_0)\left(b^T d - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ A^T d)\right\|_1\right), \qquad (3.18)
\end{aligned}$$

where the second equality follows from (3.17). Taking the limit of $\theta(y_0 + \lambda d)$ as λ tends to $+\infty$, (3.18) and (3.14) tell us that

$$\mathrm{Val}(\mathcal{D}) = \lim_{\lambda\to+\infty} \theta(y_0 + \lambda d) = \theta(y_0 + \lambda_0 d) + \lim_{\lambda\to+\infty} (\lambda - \lambda_0)\left(b^T d - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ A^T d)\right\|_1\right).$$

Since $\theta(y_0+\lambda_0 d)$ and $b^T d - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ A^T d)\right\|_1$ are constant in λ, $\lim_{\lambda\to+\infty}(\lambda-\lambda_0) = +\infty$, and $\mathrm{Val}(\mathcal{D})$ is finite, we can conclude that $b^T d - \left\|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ A^T d)\right\|_1 = 0$ and hence $\theta(y_0 + \lambda_0 d) = \mathrm{Val}(\mathcal{D})$. This contradicts to the hypothesis that $\mathrm{Val}(D)$ is not attained at any finite $y \in \mathbb{R}^m$. Thus, $(\mathcal{D})$ is solvable.

"$S^*_{\mathcal{D}} \neq \emptyset \Rightarrow S^*_{\mathcal{P}} \neq \emptyset$ & $\mathrm{Val}(\mathcal{P}) = \mathrm{Val}(\mathcal{D})$": Taking $p = -\left(A^T y - c\right)$ in (2.8) yields that the dual problem $(\mathcal{D})$ can be equivalently rewritten as

$$\max_{y\in\mathbb{R}^m} \min_{I\in\mathcal{J}(r)} \left\{b^T y - \left\|\Pi_{\mathbb{R}^n_I\cap\mathbb{R}^n_+}(u \circ (A^T y - c))\right\|_1\right\}. \qquad (3.19)$$

Let y^* be an optimal solution of $(\mathcal{D})$. According to (3.19), there exists some index set $\hat{I} \in \mathcal{J}(r)$ such that

$$y^* \in \arg\max_{y\in\mathbb{R}^m} \left\{b^T y - \left\|\Pi_{\mathbb{R}^n_{\hat{I}}\cap\mathbb{R}^n_+}\left(u \circ \left(A^T y - c\right)\right)\right\|_1\right\}. \qquad (3.20)$$

As one can see, the involved maximization problem in (3.20) is exactly problem $\left(\mathcal{D}_{\hat{I}}\right)$, and it shares the same optimal value of problem $(\mathcal{D})$, i.e., $\mathrm{Val}\left(\mathcal{D}_{\hat{I}}\right) = \mathrm{Val}(\mathcal{D})$. Additionally, as stated in Lemma 2.2, $\left(\mathcal{D}_{\hat{I}}\right)$ is indeed the dual problem of the linear programming problem $\left(\mathcal{P}_{\hat{I}}\right)$. Then the strong duality admitted by linear programming supports us with an optimal solution $\hat{x}$ of the problem $\left(\mathcal{P}_{\hat{I}}\right)$ such that

$$c^T\hat{x} = \mathrm{Val}\left(\mathcal{P}_{\hat{I}}\right) = \mathrm{Val}\left(\mathcal{D}_{\hat{I}}\right) = b^T y^* - \left\|\Pi_{\mathbb{R}^n_{\hat{I}}\cap\mathbb{R}^n_+}(u \circ (A^T y^* - c))\right\|_1. \qquad (3.21)$$

Notice that $\hat{x} \in F_{\hat{I}}$ is certainly a feasible solution of $(\mathcal{P})$. Along with $\mathrm{Val}\left(\mathcal{D}_{\hat{I}}\right) = \mathrm{Val}(\mathcal{D})$, we have

$$\mathrm{Val}(\mathcal{P}) \leq \mathrm{Val}\left(\mathcal{P}_{\hat{I}}\right) = c^T\hat{x} = \mathrm{Val}(\mathcal{D}).$$

The weak duality forces the above inequality to be an equality, which actually implies that $\hat{x}$ is also an optimal solution of $(\mathcal{P})$ and the duality gap vanishes at $(\hat{x}, y^*)$ for problems $(\mathcal{P})$ and $(\mathcal{D})$. This completes the proof. □

We would like to take this opportunity to emphasize the essential difference between our strong duality result and that from mixed integer linear programming (MILP). The reason is that the sparse LP considered in this paper can be reformulated as a MILP and there already exists rich duality theory for MILP. Therefore, in principle, we could apply the duality framework of MILP to the sparse LP and

establish the corresponding strong duality results. We will see that it would result in a significantly different duality formulation from ours. The difference is a bit surprising and is indeed in favour of ours.

We first recall the duality framework for MILP from [31]. Consider

$$\text{(MILP)} \qquad \nu^{\text{IP}} := \min\ c^T x, \qquad \text{s.t.} \quad Ax = b, \quad x \in X,$$

where the data c, A, b are as in the sparse LP and X is a mixed integer linear set. The augmented Lagrangian dual is defined as follows [31, Eq. 7, Sect. 3.1]:

$$\begin{aligned} \nu_\rho^{\text{LD}+} &:= \sup_{\lambda\in\Re^m} \inf_{x\in X} \left\{c^T x + \lambda^T(b - Ax) + \rho\psi(b - Ax)\right\} \\ &= \sup_{\lambda\in\Re^m} \inf_{x\in X,\ \psi(b-Ax)\le\omega} \left\{c^T x + \rho\omega + \lambda^T(b - Ax)\right\} \\ &= \sup_{\lambda\in\Re^m} \inf_{x,\ \omega} \left\{c^T x + \lambda^T(b - Ax) + \rho\omega \ \mid\ (x,\omega) \in \text{conv}(S_\psi)\right\}, \end{aligned} \tag{3.22}$$

where $\rho \ge 0$, $\psi : \Re^m \mapsto \Re_+$ is an augmenting function satisfying certain properties, and

$$S_\psi := \{(x,\omega) :\ x \in X \ \text{ and } \ \psi(b - Ax) \le \omega\}.$$

To simplify, we take ψ to be the quadratic augmenting function: $\psi(z) = \|z\|^2$, $z \in \Re^m$.

A key message from [31] is that, for the quadratic augmenting function, there exists MILP examples ([31, Example 1]) such that $\nu_\rho^{\text{LD}+} < \nu^{\text{IP}}$ for any finite value $\rho > 0$. In other words, there exists non-zero duality gap between the primal and its augmented Lagrangian dual for any choice $\rho > 0$. If we let $\rho = 0$ resulting in the Lagrangian dual, the duality gap gets even bigger. Therefore, developing strong duality results for MILP is not trivial. We also note that the example constructed in [31] is not from the sparse LP. We wonder if the (disappointing) non-zero duality gap would happen when the sparse LP is reformulated as a special MILP. If it does not happen, then our strong duality would not show any advantage over that of the MILP reformulation. As the example shows below, it can happen.

Example 3.2. *Consider the sparse LP*

$$\nu^{SLP} := \min\ -x_1 - x_2, \qquad s.t. \qquad -x_1 + x_2 = 0, \quad 0 \le x \le 1, \quad \textit{and} \quad \|x\|_0 \le 1.$$

Obviously, the only feasible point is $(0, 0)$ *and the optimal objective is* $\nu^{SLP} = 0$. *According to [30, Lemma 1], the constraints* $0 \le x \le 1$ *and* $\|x\|_0 \le 1$ *is equivalent to* $x \in X$ *with*

$$X := \left\{(x_1, x_2, x_3, x_4) \ \middle|\ \begin{array}{l} x_3 + x_4 \le 1, \quad x_3 \in \{0, 1\}, \quad x_4 \in \{0, 1\} \\ 0 \le x_1 \le x_3, \quad 0 \le x_2 \le x_4. \end{array} \right\}.$$

Hence, the sparse LP has a MILP reformulation. The augmented Lagrangian dual ν_ρ^{LD+} *in (3.22) satisfies (see also [31, Eq. 15]):*

$$\begin{aligned} \nu_\rho^{LD+} &\le \inf_{(x,\omega)\in\text{conv}(\text{S}_\psi)} \sup_{\lambda\in\Re^m} \left\{c^T x + \lambda^T(b - Ax) + \rho\omega\right\} \\ &= \inf_{x,\ \omega} \left\{c^T x + \lambda^T(b - Ax) + \rho\omega \ \mid\ (x,\omega) \in \text{conv}(\text{S}_\psi)\right\}, \end{aligned}$$

where, with the choice being the quadratic augmenting function,

$$S_\psi = \left\{(x_1, x_2, x_3, x_4, \omega) \mid (x_1 - x_2)^2 \le \omega, \quad (x_1, x_2, x_3, x_4) \in X\right\}.$$

The following choice of two points are motivated by [31, Example 1]. Let $0 < \gamma \le \min(1, 1/(2\rho))$ *and*

$$\left(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{\omega}\right) = (0,\ \gamma,\ 0,\ 1,\ \gamma^2) \in S_\psi, \quad \left(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4, \tilde{\omega}\right) = (\gamma,\ 0,\ 1,\ 0,\ \gamma^2) \in S_\psi.$$

Then the new point

$$\left(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{\omega}\right) = \frac{1}{2}\left(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{\omega}\right) + \frac{1}{2}\left(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4, \tilde{\omega}\right) = \left(\gamma/2, \gamma/2, 1/2, 1/2, \gamma^2.\right)$$

For this choice, we have

$$\nu_\rho^{LD+} \le -\bar{x}_1 - \bar{x}_2 + \rho\bar{\omega} = -\gamma + \rho\gamma^2 \le \max\{-1/2,\ -1/(4\rho)\} < 0 = \nu^{SLP}.$$

Hence, the duality gap exists for any finite value of $\rho \ge 0$.

Since SLP admits a strong duality, one question is that whether the optimal solution to the dual problem of $(\mathcal{D})$ is the solution of $(\mathcal{P})$. This is true for convex problems, but we are not sure for the nonconvex cases. In order to explore this question, we work out the explicit formulation of the dual problem of $(\mathcal{D})$. For the convenience in analyzing the following result, we rewrite $(\mathcal{D})$ into its equivalent form :

$$\begin{array}{lll} (\hat{\mathcal{D}}) & \max & b^T y - \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ z)\|_1 \\ & \text{s.t.} & z = A^T y - c. \end{array}$$

Theorem 3.3. *The dual problem of* $(\hat{\mathcal{D}})$ *can be formulated as*

$$(\hat{\mathcal{P}}) \qquad \min \quad c^T w \quad s.t. \ \ Aw = b,\ w^T u^{-1} \le r,\ w \ge 0.$$

Furthermore, $S^*_{\mathcal{P}} = S^*_{\hat{\mathcal{P}}} \cap S(r)$.

Proof. Problem $(\hat{\mathcal{D}})$ is equivalent to

$$\min_{y\in\mathbb{R}^m, z\in\mathbb{R}^n} \left\{-b^T y + \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ z)\|_1 \mid z = A^T y - c\right\}, \tag{3.23}$$

in the sense that they share the same optimal solution set and have the same optimal value in magnitude with opposite signs. Define the Lagrangian function of problem (3.23) by

$$\hat{L}(y, z, w) := -b^T y + h(z) + w^T(A^T y - z - c), \quad \forall (y, z, w) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n,$$

where $h(z) = \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u\circ z)\|_1$. The dual problem of (3.23) takes the form of $\sup\limits_{w\in\mathbb{R}^n} \inf\limits_{y\in\mathbb{R}^m, z\in\mathbb{R}^n} \hat{L}(y, z, w)$. Note that

$$\begin{aligned} &\inf_{y\in\mathbb{R}^m, z\in\mathbb{R}^n} \hat{L}(y, z, w) \\ =& - c^T w + \inf_{y\in\mathbb{R}^m} \{-b^T y + (Aw)^T y\} + \inf_{z\in\mathbb{R}^n} \{h(z) - w^T z\} \\ =& - c^T w - \delta_{\{w|Aw=b\}}(w) - h^*(w) \end{aligned} \tag{3.24}$$

where $h^*(w) = \sup\limits_{z\in\mathbb{R}^n} \{z^T w - h(z)\}$ is the conjugate function of $h(z)$. According to Theorem 2.4,

$$h(z) = \max_{x\in C(u;r)} z^T x,$$

which is exactly the supporting function with respect to the set $C(u;r)$. Applying the conjugacy between the indicator function and the supporting function (see, e.g., [25, Example 11.4]), we can rewrite h as $h(z) = \delta^*_{C(u;r)}(z)$. By the virtue of [25, Theorem 11.1], together with the observation

$$\mathrm{cl}(\mathrm{conv}(C(u;r))) = \left\{ w \in \mathbb{R}^n \,\middle|\, w^T u^{-1} \le r, w \ge 0 \right\},$$

we have

$$h^*(w) = \delta^{**}_{C(u;r)}(w) = \mathrm{cl}\left(\mathrm{conv}\left(\delta_{C(u;r)}\right)\right)(w) = \delta_{\left\{w \in \mathbb{R}^n_+ \,|\, w^T u^{-1} \le r\right\}}(w). \tag{3.25}$$

Thus, the dual problem of (3.23) takes the form of

$$\max_{w \in \mathbb{R}^n} \left\{ -c^T w \mid Aw = b, w^T u^{-1} \le r, w \ge 0 \right\},$$

which implies the desired dual problem for $(\hat{\mathcal{D}})$. Since the dual problem of $(\hat{\mathcal{D}})$ optimizes over the convex hull of the feasible set of $(\mathcal{P})$, by utilizing the strong duality as shown in Theorem 3.1, we can verify that $S^*_{\mathcal{P}} = S^*_{\hat{\mathcal{P}}} \cap S(r)$. □

Figure 1 shows the relationship between $(\mathcal{P})$, $(\mathcal{D})$, $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{D}})$. As we have mentioned in Theorem 3.3, due to the sparsity constraint set, the solution of $(\mathcal{P})$ cannot be obtained through $(\mathcal{D})$ directly. But when the l_0 norm of $(\hat{\mathcal{P}})$'s optimal solution $\hat{x}$ is not larger than r, $\hat{x}$ is also the minimizer of $(\mathcal{P})$.

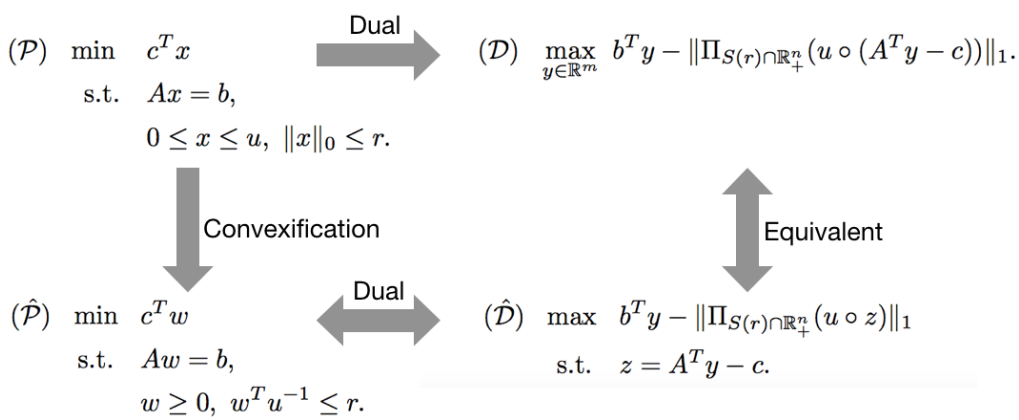

Figure 1: Relationship between $(\mathcal{P})$, $(\mathcal{D})$, $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{D}})$

Definition 3.4. *An index set $I^* \in \mathcal{J}(r)$ is said to be an optimal index set of $(\mathcal{P})$ if there exists an optimal solution x^* such that $x^* \in \mathbb{R}^n_{I^*}$.*

The following theorem provides several sufficient conditions for getting an optimal index set of $(\mathcal{P})$ by utilizing the optimal solution of $(\hat{\mathcal{D}})$.

Theorem 3.5. *Suppose that (y^*, z^*) is the optimal solution of $(\hat{\mathcal{D}})$. Let $\{t_1, \ldots, t_n\}$ be a permutation of $[n]$ such that $u_{t_1} z^*_{t_1} \ge u_{t_2} z^*_{t_2} \ge \ldots \ge u_{t_n} z^*_{t_n}$. Then,*

(i) 0 is an optimal solution of $(\mathcal{P})$ if $z^* < 0$;

(ii) $\{t_1, \ldots, t_r\}$ is an optimal index set of $(\mathcal{P})$ if one of the following holds

(a) $\|[z^*]_+\|_0 = r$;

(b) $\|[z^*]_+\|_0 > r$ *and* $u_{t_r} z^*_{t_r} > u_{t_{r+1}} z^*_{t_{r+1}}$;

(c) $0 < \|[z^*]_+\|_0 < r$ *and* $z^*_{t_{r+1}} < 0$.

Proof. Note that y^* is also the optimal solution to $(\mathcal{D})$. Thus, there exists an index set $I^* \in \mathcal{J}(r)$ such that

$$b^T y^* - \|\Pi_{\mathbb{R}^n_{I^*} \cap \mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1 = b^T y^* - \|\Pi_{S(r) \cap \mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1.$$

It then follows from Lemma 2.3 that for any index set $I \in \mathcal{J}(r)$,

$$\|\Pi_{\mathbb{R}^n_{I^*} \cap \mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1 \quad \geq \quad \|\Pi_{\mathbb{R}^n_{I} \cap \mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1. \tag{3.26}$$

Let x^* be an optimal solution of $(\mathcal{P}_{I^*})$. It follows from the proof of Theorem 3.1 that x^* is exactly an optimal solution of $(\mathcal{P})$. By invoking (3.26), we can get the assertions in (a) and (b) immediately. From the first-order optimality for $(\mathcal{P}_{I^*})$, we know that (x^*, y^*) will satisfy

$$\begin{cases} (A^T y^* - c)_{I^*} = -w + v, \\ w^T x^*_{I^*} = 0, \ x^*_{I^*} \geq 0, \ w \geq 0, \\ v^T (u_{I^*} - x^*_{I^*}) = 0, \ u - x^*_{I^*} \geq 0, \ v \geq 0. \end{cases}$$

Note that $z^* = A^T y^* - c$. By direct calculation, we can verify that for any $i \in I^*$, if $z^*_i < 0$, then $x^*_i = 0$. Thus, it is straightforward to verify that the results in (i) and (c) are valid. □

We can apply the conditions introduced in the above theorem to determine whether the solution of $(\mathcal{P})$ is detectable by $(\mathcal{D})$, see the following numerical illustration.

Example 3.6. *Set* $A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$, $b = (0,0)^T$, $c = (0,0,-1,-1)^T$, $u = (1,1,1,1)^T$ *and* $r = 2$ *in problem* $(\mathcal{P})$. *One can find that the optimal solution of* $(\mathcal{P})$ *is* $(0,0,1,1)^T$ *with the optimal value* -2. *By simple calculation, we know that the optimal solution of related* $(\hat{D})$ *is* (y^*, z^*) *with* $y^* = (0,0)^T$ *and* $z^* = (0,0,1,1)^T$. *As can be seen that* (y^*, z^*) *meets condition (a) presented in Theorem 3.5, and hence the optimal index set of* $(\mathcal{P})$ *is* $I^* = \{3,4\}$. *Then by solving* $(\mathcal{P}_{I^*})$, *we can obtain the optimal solution of* $(\mathcal{P})$.

4 Saddle Point Theorem

In this section, we study the saddle point theorem of SLP.

Definition 4.1. *Let* $x^* \in C(u;r)$ *and* $y^* \in \mathbb{R}^m$. *If*

$$L(x^*, y) \leq L(x^*, y^*) \leq L(x, y^*), \ \forall \ x \in C(u;r), \ y \in \mathbb{R}^m,$$

i.e.,

$$x^* = \arg \min_{x \in C(u;r)} L(x, y^*), \quad \textit{and} \quad y^* = \arg \max_{y \in \mathbb{R}^m} L(x^*, y), \tag{4.27}$$

then (x^*, y^*) *is called a saddle point of the Lagrange function* $L(x,y)$ *of* $(\mathcal{P})$.

Before presenting the saddle point theorem, we propose a necessary and sufficient optimality condition for $(\mathcal{P})$ and $(\mathcal{D})$ which follows immediately from strong duality theorem.

Theorem 4.2. (Optimality Condition) $(x^*, y^*) \in \mathbb{R}^n \times \mathbb{R}^m$ *is a solution of*

$$\begin{cases} c^T x - b^T y + \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ (A^T y - c))\|_1 = 0, \\ Ax = b, \\ 0 \le x \le u, \\ x \in S(r), \end{cases} \tag{4.28}$$

if and only if x^* and y^* are optimal solutions of ($\mathcal{P}$) and ($\mathcal{D}$) respectively.

By utilizing the first-order optimality condition of convex program, together with Lemma 2.4, we can easily get the equivalence between the system (4.27) and the system (4.28) with $x = x^*$ and $y = y^*$. Combining with Theorem 4.2, the following theorem comes directly.

Theorem 4.3. (Saddle Point Theorem) (x^*, y^*) *is the saddle point of the Lagrange function of* ($\mathcal{P}$), *if and only if x^* and y^* are the global optimal solutions of ($\mathcal{P}$) and ($\mathcal{D}$) respectively.*

Proof. ($\Rightarrow$) From the definition of saddle point, we have $y^* = \arg\max_{y\in\mathbb{R}^m} L(x^*, y)$. Then

$$\nabla_y L(x^*, y) = Ax^* - b = 0.$$

This is the second formula of (4.28). From the definition of saddle point, we also have $x^* = \arg\min_{x\in C(u;r)} L(x, y^*)$. According to Lemma 2.6,

$$L(x^*, y^*) = b^T y^* - \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1. \tag{4.29}$$

Substituting (4.29) and $Ax^* = b$ to

$$L(x^*, y) \le L(x^*, y^*) \le L(x, y^*)$$

and let $x = x^*$ in the second inequality, we can derive

$$c^T x^* \le b^T y^* - \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1 \le c^T x^*.$$

Then $c^T x^* = b^T y^* - \|\Pi_{S(r)\cap\mathbb{R}^n_+}(u \circ (A^T y^* - c))\|_1$, which is the first formula of (4.28). Above all, if (x^*, y^*) is the saddle point of the Lagrange function of (SLP), (x^*, y^*) is the solution of (4.28). Applying Theorem 4.2, we show that x^* and y^* are the global optimal solutions of ($\mathcal{P}$) and ($\mathcal{D}$), respectively.

($\Leftarrow$) Denote $\theta(y) = \min_{x\in C(u;r)} L(x, y)$. Since x^* is the optimal solution of ($\mathcal{P}$), we have $x^* \in C(u; r)$ and $Ax^* = b$. Then

$$\begin{aligned} \theta(y^*) &= \min_{x\in C(u;r)} c^T x - {y^*}^T (Ax - b) \\ &\le c^T x^* - {y^*}^T (Ax^* - b) \\ &= L(x^*, y^*) = c^T x^* \end{aligned}$$

Since the duality gap is zero, we can get $\theta(y^*) = L(x^*, y^*) = c^T x^*$. Furthermore, according to the definition of $\theta(y^*)$ and $Ax^* = b$, we can conclude that

$$L(x^*, y) \le L(x^*, y^*) = \theta(y^*) \le L(x, y^*),\ x \in C(u; r).$$

□

The remaining part of this section is devoted to studying the first-order necessary optimality conditions of saddle point problem (4.27). We first recall the expression of Fréchet normal cone of $S(r)$ [11, Theorem 2.1] and show the decomposition property of $\widehat{N}_{C(u;r)}(x)$. For any $x \in S(r)$ and letting $\Gamma = supp(x)$, the Fréchet normal cone to $S(r)$ at x is

$$\widehat{N}_{S(r)}(x) = \begin{cases} \mathbb{R}^n_\Gamma, & |\Gamma| = r; \\ \{0\}, & |\Gamma| < r, \end{cases} \tag{4.30}$$

where $supp(x) = \{i \in \mathcal{I} \mid x_i \neq 0\}$.

Lemma 4.4. *For any $x \in C(u;r)$, denote $\Gamma = supp(x)$ and $T = \{x \in \mathbb{R}^n \mid x \leq u\}$. The Fréchet normal cone to $C(u;r)$ at x has the following decomposition:*

$$\widehat{N}_{C(u;r)}(x) = \widehat{N}_{S(r)}(x) + \widehat{N}_{\mathbb{R}^n_+}(x) + \widehat{N}_T(x) \tag{4.31}$$

i.e.

$$\widehat{N}_{C(u;r)}(x) = \begin{cases} \left\{ \xi \in \mathbb{R}^n \middle| \xi_i \in \begin{cases} \mathbb{R}, & x_i = 0; \\ \{0\}, & 0 < x_i < u_i; \\ \mathbb{R}_+, & x_i = u_i. \end{cases} \right\}, & \text{if } |\Gamma| = r; \\ \left\{ \xi \in \mathbb{R}^n \middle| \xi_i \in \begin{cases} \mathbb{R}_-, & x_i = 0; \\ \{0\}, & 0 < x_i < u_i; \\ \mathbb{R}_+, & x_i = u_i. \end{cases} \right\}, & \text{if } |\Gamma| < r. \end{cases} \tag{4.32}$$

Proof. (4.31) can be obtained directly from [12, Corollary 2.9]

$$\widehat{N}_{\mathbb{R}^n_+}(x) = \left\{ \xi \in \mathbb{R}^n : \xi_i \in \begin{cases} \mathbb{R}_-, & x_i = 0; \\ \{0\}, & x_i > 0. \end{cases} \right\} \tag{4.33}$$

$$\widehat{N}_T(x) = \left\{ \xi \in \mathbb{R}^n : \xi_i \in \begin{cases} \mathbb{R}_+, & x_i = u_i; \\ \{0\}, & x_i < u_i. \end{cases} \right\} \tag{4.34}$$

After simple calculation, we can have the explicit formula. □

Now we introduce the definitions of $\beta-$stationary point and $\widehat{N}-$stationary point of saddle point problem (4.27) based on the expression of projection and Fréchet normal cone of $C(u;r)$.

Definition 4.5. *Given $x^* \in C(u;r)$, $y^* \in \mathbb{R}^m$, and $\beta > 0$, set $q^* := c - A^T y^*$.*

(i) (x^*, y^*) is called a $\beta-$stationary point of the saddle point problem (4.27) if it satisfies

$$\begin{cases} Ax^* = b, \\ x^* \in \Pi_{C(u;r)}(x^* - \beta q^*). \end{cases} \tag{4.35}$$

(ii) (x^*, y^*) is called an $\widehat{N}-$stationary point of the saddle point problem (4.27) if it satisfies

$$\begin{cases} Ax^* = b, \\ 0 \in q^* + \widehat{N}_{C(u;r)}(x^*). \end{cases} \tag{4.36}$$

It is noteworthy that the $\widehat{N}-$stationarity coincides with the β-stationarity for any given $\beta > 0$ with $C(u;r)$ replaced by any closed convex set C, due to the fact that

$$\widehat{N}_C(x^*) = N_C(x^*) = \left\{ v \middle| \langle v, x - x^* \rangle \leq 0, \ \forall x \in C \right\}$$

(See, [25, Theorem 6.9]). However, it is not the case for the nonconvex set $C(u;r)$. The relationship among these two types of stationarity and the optimality of problems $(\mathcal{P})$ and $(\mathcal{D})$ are elaborated in the following theorem.

Theorem 4.6. *Given $x^* \in C(u;r)$ and $y^* \in \mathbb{R}^m$, denote $q^* = c - A^T y^*$ and*

$$\mathrm{B}(x^*, y^*) = \left\{ \beta \in \mathbb{R}_{++} \middle| \beta \leq \min_{x_i^*=0, q_i^*<0} \left\{ -\frac{x^*_{(r)}}{q_i^*} \right\}, \text{ if } \|x^*\|_0 = r \text{ and } \{i \in [n] \mid x_i^* = 0, q_i^* < 0\} \neq \emptyset \right\},$$

where $x^*_{(r)}$ is the rth largest element of x^*. Consider the following statements

(a) *x^* is an optimal solution of ($\mathcal{P}$) and y^* is an optimal solution of ($\mathcal{D}$);*

(b) *(x^*, y^*) is a $\beta-$stationary point of the saddle point problem (4.27) for any $\beta > 0$;*

(c) *(x^*, y^*) is a $\beta-$stationary point of the saddle point problem (4.27) for any $\beta \in \mathrm{B}(x^*, y^*)$;*

(d) *(x^*, y^*) is an $\widehat{N}-$stationary point of the saddle point problem (4.27).*

We have

(i) *(a) $\Leftrightarrow$ (b);*

(ii) *(c) $\Leftrightarrow$ (d);*

(iii) *if $\mathrm{B}(x^*, y^*) = \mathbb{R}_{++}$ (e.g., $\|x^*\|_0 < r$), then (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ (c) $\Leftrightarrow$ (d).*

Proof. (i) "(a) $\Leftrightarrow$ (b)": Taking $p = q^*$ in Theorem 2.4, the following equivalence

$$x^* \in \arg\min_{x \in C(u,r)} L(x, y^*) \Longleftrightarrow x^* \in \Pi_{C(u;r)}(x^* - \beta q^*), \quad \forall \beta > 0,$$

is due to (2.7). Along with the observation

$$y^* \in \arg\max_{y \in \mathbb{R}^m} L(x^*, y) \Longleftrightarrow Ax^* = b,$$

we can conclude the equivalence between assertions in (a) and (b).
(ii)&(iii): For any given $\beta \in \mathrm{B}(x^*, y^*)$, Definition 4.5 (i) allows us to get an equivalent characterization for (x^*, y^*) being a $\beta-$stationary point of the saddle point problem (4.27) by

$$\left\{ \begin{array}{l} \left\{ \begin{array}{ll} q_i^* \geq -\frac{x^*_{(r)}}{\beta}, & \text{if } x_i^* = 0; \\ q_i^* = 0, & \text{if } 0 < x_i^* < u_i; \\ q_i^* \leq 0, & \text{if } x_i^* = u_i. \end{array} \right. , \quad \text{if } \|x^*\|_0 = r; \\ \left\{ \begin{array}{ll} q_i^* \geq 0, & \text{if } x_i^* = 0; \\ q_i^* = 0, & \text{if } 0 < x_i^* < u_i; \\ q_i^* \leq 0, & \text{if } x_i^* = u_i. \end{array} \right. , \quad \text{if } \|x^*\|_0 < r. \end{array} \right. \tag{4.37}$$

Combining with the expression of $\widehat{N}_{C(u;r)}$ as presented in (4.32), we can easily get the implication "(c) $\Rightarrow$ (d)". On the other hand, from the definition of the set $\mathrm{B}(x^*, y^*)$, we can verify that if the assertion in (d) holds, then for any choice of β from $\mathrm{B}(x^*, y^*)$, the system (4.37) is valid, which indeed shows the $\beta-$stationarity of (x^*, y^*). Thus, the equivalence in (ii) is established. Due to the two equivalences in (i) and (ii), the assertion in (iii) follows readily. This completes the proof. □

We demonstrate the subtle difference between the two concepts of stationary point below.

Example 4.7. *Let A, b, c, u, r be defined in Example 3.6. It is easy to verify that $x^* = (0, 0, 1, 1)^T$ is the unique optimal solution to ($\mathcal{P}$), and $y^* = (0, 0)^T$ is an optimal solution to ($\mathcal{D}$), with the common optimal value -2. Note that*

$$\Pi_{C(u,r)}\left(x^* - \beta(c - A^T y^*)\right) = \Pi_{C(u,r)}\left((0, 0, 1+\beta, 1+\beta)^T\right) = \{x^*\}, \quad \forall \beta > 0$$

$c - A^T y^* \in \widehat{N}_{C(u;r)}(x^*)$, and $\mathrm{B}(x^*, y^*) = \mathbb{R}_{++}$. Thus, (x^*, y^*) is an $\widehat{N}-$stationary point and a $\beta-$stationary point of the saddle point problem (4.27) for any $\beta > 0$. This confirms all conclusions in Theorem 4.6. Now, if we take $\bar{y} = (10, 1)^T$. By direct calculation, we have

$$\mathrm{B}(x^*, \bar{y}) = (0, 1/10]\,, \quad A^T\bar{y} - c \in \widehat{N}_{C(u;r)}(x^*),$$

while

$$\Pi_{C(u;r)}\left(x^* - \beta(c - A^T\bar{y})\right) = \left\{(1, 0, 1, 0)^T\right\} \not\ni x^*, \quad \forall \beta > 1/10.$$

Thus, $(x^*, \bar{y})$ is an $\widehat{N}-$stationary point but not a $\beta-$stationary point for any $\beta > 1/10$.

Recall from [12] that a point x^* is called a B-KKT point of $(\mathcal{P})$ if there exist $y^* \in \mathbb{R}^m$, $\mu \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^n$ such that

$$\begin{cases} -(c - A^T y^*) + \mu^* - \lambda^* \in \widehat{N}_{S(r)}(x^*), \\ \mu^* \geq 0, x^* \geq 0, {\mu^*}^T x^* = 0, \\ \lambda^* \geq 0, u - x^* \geq 0, {\lambda^*}^T(u - x^*) = 0, \\ Ax^* = b, \\ x^* \in S(r). \end{cases} \tag{4.38}$$

Note that

$$\mu^* \geq 0, x^* \geq 0, {\mu^*}^T x^* = 0 \quad \Longleftrightarrow \quad -\mu^* \in \widehat{N}_{\mathbb{R}^n_+}(x^*)$$

and

$$\lambda^* \geq 0, u - x^* \geq 0, {\lambda^*}^T(u - x^*) = 0 \quad \Longleftrightarrow \quad \lambda^* \in \widehat{N}_T(x^*),$$

where the involved two normal cones are established in (5.45) and (5.46). The decomposition property as stated in Lemma 4.4 then leads to the following characterization for a B-KKT point x^*:

$$\begin{cases} -(c - A^T y^*) \in \widehat{N}_{S(r)}(x^*) + \widehat{N}_{\mathbb{R}^n_+}(x^*) + \widehat{N}_T(x^*) = \widehat{N}_{C(u;r)}(x^*), \\ Ax^* = b, \end{cases} .$$

which is exactly the defining formula for (x^*, y^*) being an $\widehat{N}-$stationary point.

It is well-known in the theory of constrained nonlinear programming that a local minimizer is not always a KKT point, even for nonlinear convex programming problems, unless some regularity conditions, also termed as constraint qualifications (CQs for short), are enforced to the constraint system (see, e.g., [28]). For sparse optimization problems, the nonconvexity and discontinuity resulting from the involved ℓ_0-norm function will unsurprisingly aggravate the need of such CQs in general. For example, the popular CQs include the Robinson CQ in [29] and the restricted linear independence constraint qualification (R-LICQ for short, see [12, Definition 2.4]) in [12] for general sparsity constrained nonlinear programming. More specifically, as discussed in [12, Theorem 3.2 (ii)], if x^* is a minimizer of $(\mathcal{P})$, then x^* is a $B-$KKT point under the restricted linear independence constraint qualification (R-LICQ). The linear independency as required in R-LICQ may fail for general SLP due to the free choice of A and the constraint functions corresponding to the box constraint in problem $(\mathcal{P})$. See, for example, in the instance of $(\mathcal{P})$ in Example 3.6, the gradients of active constraints at x^* form the following set of linearly dependent vectors:

$$\left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\},$$

and hence the R-LICQ is not valid anymore. In this regard, to get the strong duality and the equivalences that are addressed in Theorems 3.1 and 4.6 without any CQs comes to be the main contribution in this paper, which in turn enriches the theory of nonlinear programming.

5 An Extension

The box constraint with 0 as the lower bound in $(\mathcal{P})$ can be extended to the case of $-u \le x \le u$ for any given $u \in \mathbb{R}^n_{++}$. The corresponding sparse linear programming problem takes the form of

$$(\tilde{\mathcal{P}}) \qquad \min\ c^T x, \ \text{ s.t. } Ax = b, \ -u \le x \le u, \ \|x\|_0 \le r.$$

This section is devoted to the strong duality and optimality conditions for $(\tilde{\mathcal{P}})$ by utilizing a similar scheme as that for $(\mathcal{P})$. Before proceeding, some essential technical properties are stated as follows.

Lemma 5.1. *Given $I \subseteq [n]$, $p \in \mathbb{R}^n$ and $u \in \mathbb{R}^n_{++}$, we have*

$$\min\left\{p^T x \mid -u \le x \le u, x \in \mathbb{R}^n_I\right\} = -\|\Pi_{\mathbb{R}^n_I}(u \circ p)\|_1. \tag{5.39}$$

Proof. Set $x^*_i = -\text{sign}(p_i)u_i$, if $i \in I$ and $p_i \neq 0$, and $x^*_i = 0$ otherwise. Such an x^* is obviously an optimal solution to (5.39), and

$$p^T x^* = \sum_{i \in I,\ p_i \neq 0} (-|u_i p_i|) = -\|\Pi_{\mathbb{R}^n_I}(u \circ p)\|_1.$$

This completes the proof. □

Corollary 5.2. *Given $p \in \mathbb{R}^n$, $u \in \mathbb{R}^n_{++}$ and an integer r $(0 < r \le n)$, we have*

$$\min\left\{p^T x \mid x \in \tilde{C}(u;r)\right\} = -\|\Pi_{S(r)}(u \circ p)\|_1, \tag{5.40}$$

where $\tilde{C}(u;r) = \{x \in \mathbb{R}^n \mid -u \le x \le u, \|x\|_0 \le r\}$. Furthermore, x^* is an optimal solution of problem (5.40) if and only if $x^* \in \Pi_{\tilde{C}(u;r)}(x^* - \lambda p)$ for any $\lambda > 0$, i.e.,

$$\tilde{S}^*(p) := \arg\min_{x \in \mathbb{R}^n}\left\{p^T x \ \middle|\ x \in \tilde{C}(u;r)\right\} = \left\{x^* \in \mathbb{R}^n \ \middle|\ x^* \in \Pi_{\tilde{C}(u;r)}(x^* - \lambda p), \forall \lambda > 0\right\}. \tag{5.41}$$

Proof. By employing Lemma 5.1, we have

$$\begin{aligned}
\min_{x \in \tilde{C}(u;r)} p^T x &= \min_{I \in \mathcal{J}(r)} \min_{x \in \mathbb{R}^n_I, -u \le x \le u} p^T x = \min_{I \in \mathcal{J}(r)} -\|\Pi_{\mathbb{R}^n_I}(u \circ p)\|_1 \\
&= -\max_{I \in \mathcal{J}(r)} \sum_{i \in I,\ p_i \neq 0} (|u_i p_i|) = -\sum_{i=1}^{r} (|u \circ p|)^{\downarrow}_i \\
&= -\left\|\Pi_{S(r)}(u \circ p)\right\|_1.
\end{aligned}$$

Mimicking the proof of Theorem 2.4, we can derive the desired assertions. □

Define the Lagrangian function of $(\tilde{\mathcal{P}})$ as

$$\tilde{L}(x, y) = c^T x - y^T(Ax - b), \quad \forall x \in \tilde{C}(u;r), y \in \mathbb{R}^m,$$

and denote by $\mathbf{\Pi}^s$ the set of all signed permutation matrices in $\mathbb{R}^{n \times n}$ and by

$$\mathbf{\Pi}^s_r = \left\{B \circ P \mid B \in \{0,1\}^{n \times n}, P \in \mathbf{\Pi}^s, \|B \circ P\|_0 \le r\right\}.$$

The cardinality of $\mathbf{\Pi}^s_r$ is less than $\binom{n}{r} \cdot 3^r \cdot n!$ (finitely many). In this case, we can verify that

$$\tilde{g}(p) := \|\Pi_{S(r)}(u \circ p)\|_1 = \max_{M \in \mathbf{\Pi}^s_r}\left\{e^T M U p\right\},$$

where $e \in \mathbb{R}^n$ is the all-one vector and U is the diagonal matrix generated from u. Thus, we can also get the piecewise linearity of $\tilde{g}$ over $\mathbb{R}^n$. With a similar proof as used for Theorem 2.6, we can write the explicit form for the dual problem of $(\tilde{\mathcal{P}})$.

Proposition 5.3. *The Lagrangian dual problem of ($\tilde{\mathcal{P}}$) can be explicitly formulated as the following unconstrained convex program*

$$(\tilde{\mathcal{D}}) \qquad \max_{y\in\mathbb{R}^m} \tilde{\theta}(y) = b^T y - \|\Pi_{S(r)}(u \circ (c - A^T y))\|_1.$$

Furthermore, y^* is an optimal solution of ($\mathcal{D}$) if and only if $b \in A\mathrm{conv}(\tilde{S}^*(c - A^T y^*))$.

Utilizing almost the same routine of the proof for Theorem 3.1, we can also establish the strong duality between ($\tilde{\mathcal{P}}$) and its dual problem ($\tilde{\mathcal{D}}$) that addressed in Proposition 5.3. We present this duality theorem as below without proof.

Theorem 5.4. *If ($\tilde{\mathcal{P}}$) (or ($\tilde{\mathcal{D}}$)) is solvable, then so is ($\tilde{\mathcal{D}}$) (or ($\tilde{\mathcal{P}}$)) and the duality gap is zero.*

Now, let us discuss the corresponding saddle point of ($\tilde{\mathcal{P}}$).

Definition 5.5. *Let $x^* \in \tilde{C}(u;r)$ and $y^* \in \mathbb{R}^m$. If*

$$\tilde{L}(x^*, y) \le \tilde{L}(x^*, y^*) \le \tilde{L}(x, y^*), \ \forall\ x \in \tilde{C}(u;r),\ y \in \mathbb{R}^m,$$

i.e.,

$$x^* = \arg\min_{x\in\tilde{C}(u;r)} \tilde{L}(x, y^*), \quad \textit{and} \quad y^* = \arg\max_{y\in\mathbb{R}^m} \tilde{L}(x^*, y), \tag{5.42}$$

then (x^*, y^*) is called a saddle point of the Lagrange function $\tilde{L}(x, y)$ of ($\tilde{\mathcal{P}}$).

The results reported in Theorems 4.2 and 4.3 can be extended to the case of ($\tilde{\mathcal{P}}$) by replacing $\|\Pi_{S(r)\cap\mathbb{R}^n_+}(\cdot)\|_1$ and $C(u;r)$ by $\|\Pi_{S(r)(\cdot)}\|_1$ and $\tilde{C}(u;r)$, respectively, facilitated with the following two lemmas.

Lemma 5.6. *For any $x \in \mathbb{R}^n$, we have $\Pi_{\tilde{C}(u;r)}(x) = \mathrm{sign}(x) \circ \Pi_{[0,u]}\left(\Pi_{S(r)}(|x|)\right)$, where $\mathrm{sign}(x)$ and $|x|$ are the component-wise sign function and the absolute value of x, respectively.*

Lemma 5.7. *For any $x \in \tilde{C}(u;r)$, denote $\Gamma = supp(x)$ and $T_1 = \{x \in \mathbb{R}^n \mid x \le u\}$, $T_2 = \{x \in \mathbb{R}^n \mid x \ge -u\}$. The Fréchet normal cone to $\tilde{C}(u;r)$ at x is*

$$\widehat{N}_{\tilde{C}(u;r)}(x) = \widehat{N}_{S(r)}(x) + \widehat{N}_{T_1}(x) + \widehat{N}_{T_2}(x) \tag{5.43}$$

i.e.

$$\widehat{N}_{\tilde{C}(u;r)}(x) = \begin{cases} \left\{ \xi \in \mathbb{R}^n \middle| \xi_i \in \left\{ \begin{array}{ll} \mathbb{R}_-, & x_i = -u_i; \\ \{0\}, & -u_i < x_i < u_i \ \textit{and}\ x_i \neq 0; \\ \mathbb{R}, & x_i = 0; \\ \mathbb{R}_+, & x_i = u_i. \end{array} \right. \right\}, & \text{if } |\Gamma| = r; \\ \left\{ \xi \in \mathbb{R}^n \middle| \xi_i \in \left\{ \begin{array}{ll} \mathbb{R}_-, & x_i = -u_i; \\ \{0\}, & -u_i < x_i < u_i; \\ \mathbb{R}_+, & x_i = u_i. \end{array} \right. \right\}, & \text{if } |\Gamma| < r. \end{cases} \tag{5.44}$$

Proof. (5.43) can be obtained directly from [12, Corollary 2.9]. Note that

$$\widehat{N}_{T_1}(x) = \left\{ \xi \in \mathbb{R}^n \middle| \xi_i \in \left\{ \begin{array}{ll} \mathbb{R}_+, & x_i = u_i; \\ \{0\}, & x_i < u_i. \end{array} \right. \right\} \tag{5.45}$$

and

$$\widehat{N}_{T_2}(x) = \left\{ \xi \in \mathbb{R}^n \middle| \xi_i \in \left\{ \begin{array}{ll} \mathbb{R}_-, & x_i = -u_i; \\ \{0\}, & x_i > -u_i. \end{array} \right. \right\} \tag{5.46}$$

Together with (4.30), we can get the explicit formula (5.44) by direct calculation. □

Additionally, all those concepts and properties in Lemma 4.4, Definition 4.5 and Theorem 4.6 can also been extended for ($\tilde{\mathcal{P}}$) and ($\tilde{\mathcal{D}}$) with subtle modifications.

Definition 5.8. *Given $x^* \in \tilde{C}(u;r)$, $y^* \in \mathbb{R}^m$, and $\beta > 0$, set $q^* := c - A^T y^*$.*

(i) (x^*, y^*) is called a $\beta-$stationary point of the saddle point problem (5.42) if it satisfies

$$\begin{cases} Ax^* = b, \\ x^* \in \Pi_{\tilde{C}(u;r)}(x^* - \beta q^*). \end{cases} \tag{5.47}$$

(ii) (x^*, y^*) is called an $\widehat{N}-$stationary point of the saddle point problem (5.42) if it satisfies

$$\begin{cases} Ax^* = b, \\ 0 \in q^* + \widehat{N}_{\tilde{C}(u;r)}(x^*). \end{cases} \tag{5.48}$$

Theorem 5.9. *Given $x^* \in \tilde{C}(u;r)$ and $y^* \in \mathbb{R}^m$, denote $q^* = c - A^T y^*$ and*

$$\tilde{\mathrm{B}}(x^*, y^*) = \left\{ \beta \in \mathbb{R}_{++} \middle| \beta \leq \min_{x_i^*=0, q_i^* \neq 0} \left\{ \frac{|x^*|_{(r)}}{|q_i^*|} \right\}, \text{ if } \|x^*\|_0 = r \text{ and } \{i \in [n] \mid x_i^* = 0, q_i^* \neq 0\} \neq \emptyset \right\},$$

where $|x^*|_{(r)}$ is the rth largest element of $|x^*|$. Consider the following statements

(a) *x^* is an optimal solution of $(\tilde{\mathcal{P}})$ and y^* is an optimal solution of $(\tilde{\mathcal{D}})$;*

(b) *(x^*, y^*) is a $\beta-$stationary point of the saddle point problem (5.42) for any $\beta > 0$;*

(c) *(x^*, y^*) is a $\beta-$stationary point of the saddle point problem (5.42) for any $\beta \in \tilde{\mathrm{B}}(x^*, y^*)$;*

(d) *(x^*, y^*) is an $\widehat{N}-$stationary point of the saddle point problem (5.42).*

We have

(i) *(a) $\Leftrightarrow$ (b);*

(ii) *(c) $\Leftrightarrow$ (d);*

(iii) *if $\tilde{\mathrm{B}}(x^*, y^*) = \mathbb{R}_{++}$ (e.g., $\|x^*\|_0 < r$), then (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ (c) $\Leftrightarrow$ (d).*

Proof. (i) "(a) $\Leftrightarrow$ (b)": By utilizing Corollary 5.2, the equivalence between assertions in (a) and (b) can be obtained by mimicking the first part of proof of Theorem 4.6.
(ii)&(iii): For any given $\beta \in \tilde{\mathrm{B}}(x^*, y^*)$, we can rewrite the second condition in (5.47) by

$$\begin{cases} \begin{cases} q_i^* \geq 0, & \text{if } x_i^* = -u_i; \\ q_i^* = 0, & \text{if } -u_i < x_i^* < u_i \text{ and } x_i^* \neq 0; \\ |q_i^*| \leq \dfrac{|x^*|_{(r)}}{\beta}, & \text{if } x_i^* = 0; \\ q_i^* \leq 0, & \text{if } x_i^* = u_i. \end{cases}, & \text{if } \|x^*\|_0 = r; \\ \begin{cases} q_i^* \geq 0, & \text{if } x_i^* = -u_i; \\ q_i^* = 0, & \text{if } -u_i < x_i^* < u_i; \\ q_i^* \leq 0, & \text{if } x_i^* = u_i. \end{cases}, & \text{if } \|x^*\|_0 < r. \end{cases} \tag{5.49}$$

Combining with the expression of $\widehat{N}_{\tilde{C}(u;r)}$ as presented in (5.44), we can easily get the implication "(c) $\Rightarrow$ (d)". On the other hand, by the definition of $\tilde{\mathrm{B}}(x^*, y^*)$, we can verify that if the assertion in (d) holds, then for any $\beta \in \tilde{\mathrm{B}}(x^*, y^*)$, all conditions in (5.49) are satisfied. This shows the $\beta-$stationarity of (x^*, y^*). Thus, the equivalence in (ii) is established. By virtue of those two equivalences in (i) and (ii), the assertion in (iii) follows readily. This completes the proof. □

An interesting question arises: Can we make the box constraint more general, e.g., $l \leq x \leq u$ for any given vectors $l \leq 0$ and $u \geq 0$? This will be one of our research topics in the future.

6 Conclusions

In this paper, we have given an explicit formula for the Lagrangian dual of SLP by employing the disjunctive structure of the sparsity constraint involved. The resulting dual problem has been shown to admit the strong duality theorem without any constraint qualification. Based on the strong duality, we have studied when the optimal index sets of the original problem can be obtained and established a saddle point theorem. Relations among the global optimality, the β-stationarity and the $\widehat{N}$-stationarity of the saddle point problem have been analyzed as well. These results have been extended to the case with $x \in [-u, u]$ instead of $x \in [0, u]$ for any given $u \in \mathbb{R}^n_{++}$ in the setting of SLP, by exploiting the properties of the projection and the normal cone with respect to the resulting new constraint system. More work needs to be done when working with more general box constraints (e.g., $l \leq x \leq u$ for any given $l \leq 0$ and $u > 0$). Additionally, from the numerical perspectively, our future research will be focused on designing efficient numerical algorithms for the SLP based on the equivalences addressed in Theorem 4.6, especially for large-scale problems with special structures of the inputs (A, b, r) arising from real application problems.

Acknowledgements

We sincerely appreciate the suggestions and comments from two anonymous referees for the improvement of the paper. We would like to acknowledge the financial support for the National Natural Science Foundation of China (11431002, 11771038, 11728101), the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University (RCS2017ZJ001) and China Scholarship Council (Grant No. 201707090019).

References

[1] Donoho D L. Compressed sensing. IEEE Trans Inform Theory, 2006, 52: 1289-1306

[2] Liu J, Chen J, Ye J. Large-scale sparse logistic regression. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009, 547-556

[3] Bertsimas D, King A, Mazumder R. Best subset selection via a modern optimization lens. Ann Stat, 2016, 44(2): 813-852

[4] Li D, Sun X L, Wang J. Optimal lot solution to cardinality constrained mean-variance formulation for portfolio selection. Math Finance, 2006, 16(1): 83-101

[5] Sun C C, Dai R, Mesbahi M. Weighted Network Design with Cardinality Constraints via Alternating Direction Method of Multipliers. IEEE Trans Control Netw Syst, 2018

[6] Wei Z H, Link S. Embedded Cardinality Constraints. In: Proceedings of the International Conference on Advanced Information Systems Engineering. Cham: Springer, 2018, 523-538

[7] Beck A, Yonina C E. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. SIAM J Optim, 2013, 23(3): 1480-1509

[8] Beck A, Hallak N. On the minimization over sparse symmetric sets: Projections, optimality conditions and algorithms. Math Oper Res, 2015, 41: 196-223

[9] Lu Z S, Yong Z. Sparse approximation via penalty decomposition methods. SIAM J Optim, 2013, 23(4): 2448-2478

[10] Li X, Song W. The first-order necessary conditions for sparsity constrained optimization. J Oper Res Soc China, 2015, 3(4): 521-535

[11] Pan L L, Xiu N H, Zhou S L. On solutions of sparsity constrained optimization. J Oper Res Soc China, 2015, 3: 421-439

[12] Pan L L, Xiu N H, Fan J. Optimality conditions for sparse nonlinear programming. Sci China Math, 2017, 60(5): 759-776

[13] Pan L L, Luo Z Y, Xiu N Y. Restricted Robinson constraint qualification and optimality for cardinality-constrained cone programming. J Optim Theory Appl, 2017, 175(1): 104-118

[14] Bucher M, Schwartz A. Second-order optimality conditions and improved convergence results for regularization methods for cardinality-constrained optimization problems. J Optim Theory Appl, 2018, 178(2): 383-410

[15] Zhu W X, Dong Z S, Yu Y L, et al. Lagrange dual method for sparsity constrained optimization. IEEE Access, 2018, 6: 28404-28416

[16] Chen A I, Graves S C. Sparsity-constrained transportation problem. arXiv preprint arXiv:1402.2309, 2014.

[17] Babu P, Pelckmans K, Stoica P, Li J. Linear systems, sparse solutions, and sudoku. IEEE Signal Process Lett, 2010, 17(1): 40-42

[18] Friedlander M P, Tseng P. Exact regularization of convex programs. SIAM J Optim, 2007,18(4): 1326-1350

[19] Donoho D L, Tanner J. Sparse nonnegative solution of underdetermined linear equations by linear programming. Proc Natl Acad Sci, 2005, 102(27): 9446-9451

[20] Shang M J, Zhang C, Xiu N H. Minimal zero norm solutions of linear complementarity problems. J Optim Theory Appl, 2014, 163(3): 795-814

[21] Chen X J, Xiang S H. Sparse solutions of linear complementarity problems. Math Program, 2016, 159(1-2): 539-556

[22] Diewert W E. Applications of Duality Theory. Frontiers of Quantitative Economics, Volume II, North-Holland Publishing Company, 1974

[23] Rockafellar R T. Convex analysis. New Jersey: Princeton university press, 2015

[24] Mordukhovich B S, Nam N M. An easy path to convex analysis and applications. Synthesis Lectures on Mathematics and Statistics, 2013, 6(2): 1-218.

[25] Rockafellar R T, Wets R J. Variational Analysis. Berlin: Springer, 1998

[26] Scholtes S. Introduction to Piecewise Differentiable Equations. Springer Science & Business Media, 2012

[27] Cui Y, Pang J S. On the finite number of directional stationary values of piecewise programs. arXiv preprint arXiv:1803.00190, 2018

[28] Bonnans J F, Shapiro A. Perturbation Analysis of Optimization Problems. Springer Science & Business Media, New York, 2000

[29] Lu Z, Zhang Y. Sparse approximation via penalty decomposition methods[J]. SIAM J Optim, 2013, 23(4): 2448-2478

[30] Bourguignon S, Ninin J, Carfantan H, et al. Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance[J]. IEEE Trans Signal Process, 2016, 64(6): 1405-1419

[31] Feizollahi M J, Ahmed S, Sun A. Exact augmented Lagrangian duality for mixed integer linear programming[J]. Math Program, 2017, 161(1-2): 365-387