Sparse Quadratically Constrained Quadratic Programming via Semismooth Newton Method

Shuai Li* Shenglong Zhou†† Ziyan Luo†

Abstract: Quadratically constrained quadratic programming (QCQP) has long been recognized as a computationally challenging problem, particularly in large-scale or high-dimensional settings where solving it directly becomes intractable. The complexity further escalates when a sparsity constraint is involved, giving rise to the problem of sparse QCQP (SQCQP), which makes conventional solution methods even less effective. Existing approaches for solving SQCQP typically rely on mixed-integer programming formulations, relaxation techniques, or greedy heuristics but often suffer from computational inefficiency and limited accuracy. In this work, we introduce a novel paradigm by designing an efficient algorithm that directly addresses SQCQP. To be more specific, we introduce P-stationarity to establish first- and second-order optimality conditions of the original problem, leading to a system of nonlinear equations whose generalized Jacobian is proven to be nonsingular under mild assumptions. Most importantly, these equations facilitate the development of a semismooth Newton-type method that exhibits significantly low computational complexity due to the sparsity constraint and achieves a locally quadratic convergence rate. Finally, extensive numerical experiments validate the accuracy and computational efficiency of the algorithm compared to several established solvers.

Keywords: Sparse quadratically constrained quadratic programming, P-stationarity, stationary equations, semismooth Newton method, locally quadratic convergence rate

1 Introduction

We aim to solve the following sparse quadratically constrained quadratic programming (SQCQP),

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_0\mathbf{x} + \mathbf{q}_0^\top \mathbf{x} + c_0 \\ \text{s.t.} \quad & \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_i\mathbf{x} + \mathbf{q}_i^\top \mathbf{x} + c_i \le 0,\ i = 1, 2, \ldots, k, \\ & \mathbf{A}\mathbf{x} - \mathbf{b} \le 0,\ \|\mathbf{x}\|_0 \le s,\ \mathbf{x} \in \mathbb{X}, \end{aligned} \tag{SQCQP}$$

where $\mathbf{Q}_i$ is an n-order symmetric matrix, $\mathbf{q}_i \in \mathbb{R}^n$, and $c_i \in \mathbb{R}$, $i = 0, 1, \ldots, k$, $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$, $\|\mathbf{x}\|_0$ is the ℓ_0-norm of $\mathbf{x}$, counting the number of its nonzero elements, $s \ll n$ is an integer, and $\mathbb{X}$ is a simple closed convex set. Throughout the paper, we assume $\mathbb{X} := X_1 \times X_2 \times \cdots \times X_n$, where each $X_i \ni 0$ is a closed interval in $\mathbb{R}$. The primary challenge of the SQCQP problem stems from sparsity constraint $\|\mathbf{x}\|_0 \le s$ since the ℓ_0-norm is non-convex and discontinuous, making the problem generally NP-hard. Importantly, this complexity remains even in the convex setting, where each $\mathbf{Q}_i$ is positive semi-definite. Another challenge is the involvement of quadratic constraints, especially in large-scale or high-dimensional settings. Nevertheless, the SQCQP problem finds

*School of Mathematics and Statistics, Beijing Jiaotong University, Beijing, 100044, China.
Emails: 24110488@bjtu.edu.cn, shlzhou@bjtu.edu.cn, zyluo@bjtu.edu.cn.
†Corresponding author.

extensive applications, including portfolio optimization with cardinality constraints on asset selection [4, 24], sparse generalized eigenvalue problems in high-dimensional data analysis and machine learning [8, 40, 49], and sparse array beamforming in signal processing [32, 35].

1.1 Related work

To overcome the computational challenges posed by the ℓ_0-norm, a variety of research efforts have focused on reformulating (SQCQP) as a mixed-integer programming (MIP) by introducing binary variables. This reformulation enables the application of various solution methods, including branch-and-bound techniques, evolutionary algorithms, and other heuristic-based approaches. A comprehensive overview of existing methods can be found in [9, 20, 26, 52] and the references therein. However, these methods often exhibit limited computational efficiency, particularly when applied to large-scale instances. Beyond the MIP-based reformulation, other existing approaches for solving the SQCQP problem can be broadly classified into two categories.

a) Relaxation methods. One strategy for addressing the challenges stemmed from the ℓ_0-norm is to relax it by a continuous or convex surrogate function [1, 21, 27, 29]. An alternative line of research transfers the original problem into nonlinear programming (NLP) with complementarity constraints [18, 30, 36, 37, 50, 51]. While this reformulation formally aligns with the standard NLP framework, it introduces significant theoretical and numerical challenges. Specifically, the resulting problems are inherently non-convex and degenerate, often violating classical constraint qualifications required for NLP [16]. Consequently, specialized techniques must be developed to ensure both theoretical rigor and computational tractability.

b) Greedy methods. In contrast, greedy approaches are capable of directly handling the ℓ_0-norm, with their development rooted in the field of compressed sensing. For problems involving only a single sparsity constraint set, a wide range of algorithms have developed, including first-order methods using gradient information [12, 11, 54, 44, 10]), as well as Newton-type algorithms that leverage second-order information [3, 58, 59, 62, 56, 61, 53]. However, compared to the case with a single sparsity constraint, research on problems involving additional constraints remains relatively limited. Existing studies mainly focus on scenarios with relatively simple constraint sets, such as the simplex set [38], non-negative constraints [45], box constraints [19], affine sets [5, 34], and symmetric sets [7, 41, 39], with algorithms typically based on gradient projection techniques. For problems involving more complex nonlinear equality or inequality constraints, only a few works have been reported. For example, the authors in [42] introduced a block coordinate descent algorithm using a penalty decomposition technique to handle such cases. Moreover, for problems involving semi-continuous variables and additional closed convex set constraints, [4] proposed an augmented Lagrangian algorithm again based on the penalty decomposition.

It is worth mentioning that most existing algorithms fall under the category of first-order methods. Although these methods are structurally simple and easy to implement, they often fail to achieve high solution accuracy and computational efficiency compared to Newton-type methods. In a recent study [60], an efficient Lagrange-Newton algorithm (LNA) was proposed for sparse NLP with additional equality constraints. The method can be viewed as a generalization of NHTP [62] using Newton's method to solve the stationary equations. It has been demonstrated that both algorithms achieve fast locally quadratic convergence rate under certain conditions. Despite the scarcity of

Newton-type algorithms in constrained sparse optimization, the impressive computational efficiency of LNA and the well-established semi-smooth Newton methods for solving the Karush-Kuhn-Tucker (KKT) system in traditional NLP [25, 28, 31, 47] suggest the potential to develop an efficient second-order algorithm for (SQCQP), which motivates the research of this paper.

1.2 Contribution

The primary contribution of this paper lies in both the theoretical analysis and the development of a Newton-type algorithm for solving problem (SQCQP). To the best of our knowledge, this is the first work to propose a direct numerical algorithm for this problem rather than relying on its reformulations. Our contributions can be summarized as follows.

A. Optimality analysis via P-stationarity. We introduce the concepts of KKT points and P-stationary points to establish the first- and second-order optimality conditions for problem (SQCQP), clarifying their connections to local minimizers. By leveraging the properties of the sparse projection operator and a nonlinear complementarity problem (NCP) function, we equivalently reformulate the P-stationarity condition as a system of stationary equations. This reformulation provides a crucial theoretical foundation for the development of a Newton-type method. However, besides the solution variables, these equations also include an unknown discrete index set that needs to be determined simultaneously. Consequently, solving these equations differs from the standard nonlinear equation by employing the semismooth Newton-type methods.

B. A semismooth Newton-type method with locally quadratic convergence. It is important to note that, in addition to the solution variables, the stationary equations also involve an unknown discrete index set that must be determined simultaneously. As a result, solving these equations differs from solving standard nonlinear equations using a Newton-type method. Nevertheless, we successfully address this challenge by developing a semismooth Newton-type method, SNSQP. The novel design ensures that each iteration produces an s-sparse solution $\mathbf{x}^\ell$, namely, $\|\mathbf{x}^\ell\|_0 \leq s$. Consequently, although a system of linear equations must be solved at each step, the algorithm maintains a remarkably low computational complexity, enabling efficient large-scale computation. Furthermore, we establish the nonsingularity of the generalized Jacobian of the stationary equations under mild assumptions, which guarantees a locally quadratic convergence rate of the proposed algorithm.

C. High numerical performance. We conduct extensive numerical experiments on various applications using both synthetic and real-world datasets. Comparative evaluations against state-of-the-art algorithms and commercial solvers, such as CPLEX and GUROBI, demonstrate that SNSQP achieves superior computational efficiency and solution accuracy, highlighting its strong potential for a variety of large-scale applications.

1.3 Organization

The paper is organized as follows. The next subsection introduces key notation used throughout the paper. Section 2 examines the optimality conditions of problem (SQCQP) by presenting KKT points and P-stationary points, then reformulates a P-stationary point as a system of stationary equations. Section 3 analyzes the nonsingularity of the generalized Jacobian matrix associated with these equations. In Section 4, we propose a semi-smooth Newton method for solving the stationary equations and establish its locally quadratic convergence. Section 5 presents extensive numerical

Notation	Description
$[n]$	$:= \{1, 2, \ldots, n\}$.
$\|t\|$	The absolute value of scalar t.
$\|T\|$	The cardinality of set $T \subseteq [n]$.
$\overline{T}$	The complementary set of T, namely, $\overline{T} = [n] \setminus T$.
$x_{(i)}$	The ith largest (in absolute value) entry of vector $\mathbf{x}$.
$\|\mathbf{x}\|$	The Euclidean norm of vector $\mathbf{x}$.
$\|\mathbf{x}\|_\infty$	The ℓ_∞-norm of vector $\mathbf{x}$.
$\mathrm{supp}(\mathbf{x})$	$:= \{i \in [n] : x_i \neq 0\}$, the support set of vector $\mathbf{x}$.
$\langle \mathbf{w}, \mathbf{x} \rangle$	The inner product of two vectors $\mathbf{w}$ and $\mathbf{x}$, i.e., $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^\top \mathbf{x} = \sum w_i x_i$.
$(\mathbf{w}; \mathbf{x})$	The vector formed by stacking $\mathbf{w}$ and $\mathbf{x}$, i.e., $(\mathbf{w}; \mathbf{x}) = (\mathbf{w}^\top \ \mathbf{x}^\top)^\top$.
$\mathbb{J}_s(\mathbf{x})$	$:= \{J \subseteq [n] : \|J\| = s, J \supseteq \mathrm{supp}(\mathbf{x})\}$.
$\mathbf{x}_T$	The subvector of $\mathbf{x}$ containing elements indexed by T.
$\mathbf{A}_{IJ}$	The submatrix of $\mathbf{A} \in \mathbb{R}^{m \times n}$ whose rows and columns are indexed by I and J. In particular, $\mathbf{A}_{:T} := \mathbf{A}_{[m]T}$ and $\mathbf{A}_{T:} := \mathbf{A}_{T[n]}$.
$\mathbb{X}_J$	$:= \times_{i \in J} X_i$, e.g., $\mathbb{X}_{\{1,3,4\}} = X_1 \times X_3 \times X_4$.
$\mathbf{e}_i$	The i-th column of identity matrix $\mathbf{I}$.
$\mathrm{bd}(\Omega)$	The boundary of set Ω.
$\mathrm{int}(\Omega)$	The interior of set Ω.

Table 1: A list of notation.

experiments, followed by conclusions in the final section.

1.4 Notation

We end this section by introducing some notation to be used throughout the paper, with most of them summarized in Table 1. In addition, we denote

$$\begin{aligned} f_i(\mathbf{x}) &:= \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_i \mathbf{x} + \mathbf{q}_i^\top \mathbf{x} + c_i, \ i = 0, 1, 2, \ldots, k, \\ \mathbb{F} &:= \{\mathbf{x} \in \mathbb{R}^n : \ f_i(\mathbf{x}) \leq 0, \ i = 1, 2, \ldots, k, \ \mathbf{A}\mathbf{x} - \mathbf{b} \leq 0\}, \\ \mathbb{S} &:= \{\mathbf{x} \in \mathbb{R}^n : \ \|\mathbf{x}\|_0 \leq s\}, \end{aligned} \tag{1.1}$$

where $\mathbb{S}$ is known as the sparse set. Let $\Pi_\Omega(\mathbf{x})$ be the projection of $\mathbf{x}$ onto set Ω, namely,

$$\Pi_\Omega(\mathbf{x}) = \mathrm{argmin}_{\mathbf{z} \in \Omega} \ \|\mathbf{z} - \mathbf{x}\|^2.$$

Therefore, $\Pi_{\mathbb{S}}(\mathbf{x})$ keeps the first s largest (in absolute value) entries of $\mathbf{x}$ and sets the remaining to be zeros. Let $T_\Omega(\mathbf{x})$, $N_\Omega(\mathbf{x})$, and $\widehat{N}_\Omega(\mathbf{x})$ be the Clarke tangent cone, Clarke/limiting normal cone, and Fréchet normal cone of $\mathbf{x}$ at Ω, respectively. One can refer to [48] for their definitions. It is easy to calculate that

$$T_{[a,b]}(x) = \begin{cases} [0, +\infty), & \text{if } \ x = a, \\ \mathbb{R}, & \text{if } \ x \in (a, b), \\ (-\infty, 0], & \text{if } \ x = b, \end{cases} \qquad N_{[a,b]}(x) \in \begin{cases} (-\infty, 0], & \text{if } \ x = a, \\ \{0\}, & \text{if } \ x \in (a, b), \\ [0, +\infty), & \text{if } \ x = b. \end{cases} \tag{1.2}$$

Moreover, by [46, Theorem 2.1], we have

$$\widehat{N}_{\mathbb{S}}(\mathbf{x}) = \begin{cases} \mathbb{R}^n_\Gamma := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_\Gamma = 0\}, & \text{if } \|\mathbf{x}\|_0 = s, \\ \{0\}, & \text{if } \|\mathbf{x}\|_0 < s. \end{cases} \tag{1.3}$$

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a locally Lipschitz continuous function. Then f is differentiable almost everywhere by Rademacher's Theorem. By denoting D_f as the set of points where f is differentiable, the Clarke generalized Jacobian [23] of f at $\mathbf{x} \in \mathbb{R}^n$ is defined as

$$\partial f(\mathbf{x}) = \text{co}\left\{\lim_{\mathbf{x}^\ell \in D_f, \mathbf{x}^\ell \to \mathbf{x}} \nabla f(\mathbf{x}^\ell)\right\},$$

where co(Ω) represents the convex hull of Ω. Let $\mathcal{U} \subseteq \mathbb{R}^n$ be an open set and $f : \mathcal{U} \to \mathbb{R}^m$ be a locally Lipschitz continuous function. We say that f is semismooth at $\mathbf{x} \in \mathcal{U}$ if it is directionally differentiable at $\mathbf{x}$, and for every $\Delta\mathbf{x} \to 0$ and $\mathbf{H} \in \partial f(\mathbf{x} + \Delta\mathbf{x})$ it satisfies

$$f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) - \mathbf{H}\Delta\mathbf{x} = o(\|\Delta\mathbf{x}\|).$$

Furthermore, if the above equation is replaced by

$$f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) - \mathbf{H}\triangle\mathbf{x} = O\left(\|\Delta\mathbf{x}\|^2\right),$$

then f is said to be strongly semismooth at $\mathbf{x}$. Finally, Fischer-Burmeister (FB) function [31] $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is defined by

$$\phi(a, b) := \sqrt{a^2 + b^2} - a - b.$$

It is well-known that the FB function is a nonlinear complementarity problem (NCP) function that satisfies $\phi(a, b) = 0$ if and only if $a \geq 0,\ b \geq 0$, and $ab = 0$.

2 Optimality Analysis

In this section, we analyze the optimality conditions of (SQCQP) by introducing P-stationarity, which is then equivalently formulated as stationary equations using the sparse projection operator and the FB function. This serves as a fundamental theoretical basis for algorithm design.

2.1 First-order optimality conditions

The Lagrangian function of (SQCQP) is,

$$L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) := f_0(\mathbf{x}) + \sum_{i=1}^{k} \mu_i f_i(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle, \tag{2.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{\lambda} \in \mathbb{R}^m$ are Lagrange multipliers. Given point $(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$, hereafter, we always denote

$$\boldsymbol{g}^* := \nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*), \quad \mathbf{H}^* := \nabla^2_{xx} L(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*), \quad \Gamma_* := \text{supp}(\mathbf{x}^*), \quad \mathbb{X}_* := \mathbb{X}_{\Gamma_*}.$$

Definition 2.1 (KKT Points). *We call $\mathbf{x}^*$ a KKT point of (SQCQP) if there is $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ such that,*

$$\begin{cases} -\boldsymbol{g}^* - \boldsymbol{\nu}^* \in \widehat{N}_{\mathbb{S}}(\mathbf{x}^*), \\ \boldsymbol{\nu}^*_{\Gamma_*} \in N_{\mathbb{X}_*}(\mathbf{x}^*_{\Gamma_*}),\ \boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0,\ \mathbf{x}^*_{\Gamma_*} \in \mathbb{X}_{\Gamma_*}, \\ f_i(\mathbf{x}^*) \leq 0,\ \mu_i^* \geq 0,\ \mu_i^* f_i(\mathbf{x}^*) = 0,\ i \in [k], \\ \mathbf{A}\mathbf{x}^* - \mathbf{b} \leq 0,\ \boldsymbol{\lambda}^* \geq 0,\ \langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x}^* - \mathbf{b} \rangle = 0. \end{cases} \tag{2.2}$$

To derive the relationship between a KKT point and a local minimizer of (SQCQP), we need the following restricted linear independent constraint qualification (LICQ) condition.

Assumption 2.1 (Restricted LICQ). *Let* $\mathbf{x}^* \in \mathbb{F} \cap \mathbb{S} \cap \mathbb{X}$ *and*

$$\mathcal{A}_1(\mathbf{x}^*) := \{i \in [k] : f_i(\mathbf{x}^*) = 0\}, \tag{2.3}$$

$$\mathcal{A}_2(\mathbf{x}^*) := \{i \in [m] : \langle \mathbf{a}_i, \mathbf{x}^* \rangle = \mathbf{b}_i\}, \tag{2.4}$$

$$\mathcal{A}_3(\mathbf{x}^*; T) := \{i \in T : x_i^* \in bd(X_i)\}. \tag{2.5}$$

Assume the following groups of vectors are linearly independent for any $T \in \mathbb{J}_s(\mathbf{x}^*)$,

$$\Big\{(\nabla f_i(\mathbf{x}^*))_{\Gamma_*} : i \in \mathcal{A}_1(\mathbf{x}^*)\Big\} \cup \Big\{(\mathbf{a}_i)_{\Gamma_*} : i \in \mathcal{A}_2(\mathbf{x}^*)\Big\} \cup \Big\{(\mathbf{e}_i)_{\Gamma_*} : i \in \mathcal{A}_3(\mathbf{x}^*; T)\Big\}.$$

Theorem 2.1. *Let* $\mathbf{x}^*$ *be a local minimizer of (SQCQP) and Assumption 2.1 hold at* $\mathbf{x}^*$. *Then there is a unique* $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ *such that* $\mathbf{x}^*$ *is a KKT point of (SQCQP).*

Proof. As $\mathbf{x}^*$ is a local minimizer of (SQCQP), it is also a local minimizer of the problem,

$$\min \ f_0(\mathbf{x}), \quad \text{s.t. } \mathbf{x} \in \mathbb{F}, \ \mathbf{x}_{\overline{T}} = 0, \ \mathbf{x}_T \in \mathbb{X}_T, \tag{2.6}$$

for any $T \in \mathbb{J}_s(\mathbf{x}^*)$. By $\Gamma_* \subseteq T$, Assumption 2.1 indicates that

$$\Big\{(\nabla f_i(\mathbf{x}^*))_T : i \in \mathcal{A}_1(\mathbf{x}^*)\Big\} \cup \Big\{(\mathbf{a}_i)_T : i \in \mathcal{A}_2(\mathbf{x}^*)\Big\} \cup \Big\{(\mathbf{e}_i)_T : i \in \mathcal{A}_3(\mathbf{x}^*; T)\Big\}$$

are linearly independent, namely, the LICQ holds at $\mathbf{x}^*$ for (2.6). Then by [55, Theorem 1], there exist a unique $(\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m \times \mathbb{R}^{n-s}$, such that

$$\begin{cases} \nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu} + \sum_{i \in \overline{T}} \gamma_i \mathbf{e}_i = 0, \\ \boldsymbol{\nu}_T \in N_{\mathbb{X}_T}(\mathbf{x}_T^*), \ \boldsymbol{\nu}_{\overline{T}} = 0, \ \mathbf{x}_T^* \in \mathbb{X}_T, \\ f_i(\mathbf{x}^*) \le 0, \ \mu_i \ge 0, \ \mu_i f_i(\mathbf{x}^*) = 0, \ i \in [k], \\ \mathbf{A}\mathbf{x}^* - \mathbf{b} \le 0, \ \boldsymbol{\lambda} \ge 0, \ \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{x}^* - \mathbf{b} \rangle = 0. \end{cases} \tag{2.7}$$

If $\|\mathbf{x}^*\|_0 = s$, then $\mathbb{J}_s(\mathbf{x}^*) = \{\Gamma_*\}$ and thus $T = \Gamma_*$. By letting $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda})$, we obtain $-\boldsymbol{g}^* - \boldsymbol{\nu}^* = \sum_{i \in \overline{\Gamma}_*} \gamma_i \mathbf{e}_i \in \mathbb{R}^n_{\overline{\Gamma}_*} = \widehat{N}_{\mathbb{S}}(\mathbf{x}^*)$ from (1.3) and $\boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0$.
If $\|\mathbf{x}^*\|_0 < s$, then for any $T \in \mathbb{J}_s(\mathbf{x}^*)$, rewrite the first row in (2.7) as follows

$$\begin{cases} [\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu}]_{\Gamma_*} = 0, \\ [\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu}]_{T \setminus \Gamma_*} = 0, \\ \big[\nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu} + \sum_{i \in \overline{T}} \gamma_i \mathbf{e}_i\big]_{\overline{T}} = 0. \end{cases} \tag{2.8}$$

Assumption 2.1 indicates the uniqueness of $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ from the first equation of (2.7). Therefore, we let $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = (\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda})$. Since $\cup_{T \in \mathbb{J}(x^*)}(T \setminus \Gamma_*) = \overline{\Gamma}_*$, it follows from the second row of (2.8) that $(\boldsymbol{g}^* + \boldsymbol{\nu})_{\overline{\Gamma}_*} = 0$. This together with the first row of (2.8) show $\boldsymbol{g}^* + \boldsymbol{\nu}^* = 0 \in \widehat{N}_{\mathbb{S}}(\mathbf{x}^*)$ from (1.3). Furthermore, $\boldsymbol{\nu}^*_{\overline{T}} = 0$ for any $T \in \mathbb{J}_s(\mathbf{x}^*)$ suffices to $\boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0$.
Therefore, both cases lead to the first condition in (2.2) and $\boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0$. Then the other conditions in (2.2) can be ensured by (2.7) with $(\boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = (\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$. □

According to Theorem 2.1, a KKT point is closely related to a local minimizer of (SQCQP). However, computing it directly is intractable. To address this, we introduce an alternative point that can be efficiently obtained through a well-designed numerical algorithm. This point is defined based on the projection onto sparse set $\mathbb{S}$, which is why we refer to it as a P-stationary point, where P represents the projection.

Definition 2.2. *We call $\mathbf{x}^* \in \mathbb{R}^n$ a P-stationary point of (SQCQP) associated with a constant $\tau > 0$ if there is $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ such that,*

$$\begin{cases} \mathbf{x}^* \ = \Pi_{\mathbb{S}}\left(\mathbf{x}^* - \tau(\boldsymbol{g}^* + \boldsymbol{\nu}^*)\right), \\ \mathbf{x}^*_{\Gamma_*} = \Pi_{\mathbb{X}_*}(\mathbf{x}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*}),\ \boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0, \\ f_i\left(\mathbf{x}^*\right) \le 0,\ \mu_i^* \ge 0,\ \mu_i^* f_i\left(\mathbf{x}^*\right) = 0,\ i \in [k], \\ \mathbf{A}\mathbf{x}^* - \mathbf{b} \le 0,\ \boldsymbol{\lambda}^* \ge 0,\ \langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x}^* - \mathbf{b}\rangle = 0. \end{cases} \tag{2.9}$$

One can easily observe that a P-stationary point must be a KKT-point because the fixed point equation based on the projection operator can lead to the condition characterized by the normal cones. Using the FB function, the two complementary conditions in (2.9) can be written as

$$\varphi\left(\mathbf{x}, \boldsymbol{\mu}\right) := \begin{bmatrix} \phi\left(-f_1\left(\mathbf{x}\right), \mu_1\right) \\ \vdots \\ \phi\left(-f_k\left(\mathbf{x}\right), \mu_k\right) \end{bmatrix} = 0, \quad \psi\left(\mathbf{x}, \boldsymbol{\lambda}\right) := \begin{bmatrix} \phi\left(b_1 - \langle \mathbf{a}_1, \mathbf{x}\rangle, \lambda_1\right) \\ \vdots \\ \phi\left(b_m - \langle \mathbf{a}_m, \mathbf{x}\rangle, \lambda_m\right) \end{bmatrix} = 0.$$

Based the above functions and the expression [6, Lemma 2.2] of $\Pi_{\mathbb{S}}(\cdot)$, $\mathbf{x}^*$ is P-stationary point if and only if $\mathbf{x}^* \in \mathbb{S}$ and

$$\begin{cases} \begin{cases} \boldsymbol{g}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*} = 0, \quad \tau\|\boldsymbol{g}^*_{\overline{\Gamma}_*}\|_\infty < x^*_{(s)}, & \text{if } \|\mathbf{x}^*\|_0 = s, \\ \boldsymbol{g}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*} = 0, \quad \boldsymbol{g}^*_{\overline{\Gamma}_*} = 0, & \text{if } \|\mathbf{x}^*\|_0 < s, \end{cases} \\ \mathbf{x}^*_{\Gamma_*} = \Pi_{\mathbb{X}_*}(\mathbf{x}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*}),\ \boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0, \\ \varphi\left(\mathbf{x}^*, \boldsymbol{\mu}^*\right) = 0, \quad \psi\left(\mathbf{x}^*, \boldsymbol{\lambda}^*\right) = 0, \end{cases} \tag{2.10}$$

which further results in the following optimality conditions for (SQCQP).

Theorem 2.2 (First-order necessary condition). *Let $\mathbf{x}^*$ be a local minimizer of (SQCQP) and Assumption 2.1 hold at $\mathbf{x}^*$. Then there is a unique $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ such that $\mathbf{x}^*$ is a P-stationary point of (SQCQP) for any $\tau \in (0, \overline{\tau})$, where*

$$\bar{\tau} := \begin{cases} \dfrac{x^*_{(s)}}{\left\|\boldsymbol{g}^*_{\overline{\Gamma}_*}\right\|_\infty}, & \text{if } \|\mathbf{x}^*\|_0 = s, \\ +\infty, & \text{if } \|\mathbf{x}^*\|_0 < s. \end{cases} \tag{2.11}$$

Proof. Theorem 2.1 states that there is a unique $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ such that $\mathbf{x}^*$ is a KKT point of (SQCQP), which by the first condition and $\boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0$ in (2.2) indicates

$$\begin{cases} \boldsymbol{g}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*} = 0, & \text{if } \|\mathbf{x}^*\|_0 = s, \\ \boldsymbol{g}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*} = 0, \quad \boldsymbol{g}^*_{\overline{\Gamma}_*} = 0, & \text{if } \|\mathbf{x}^*\|_0 < s, \end{cases} \tag{2.12}$$

The value of $\bar{\tau}$ means that for any $\tau \in (0, \bar{\tau})$, we have $\tau\,|g_i^*| < x^*_{(s)}$ for all $i \in \overline{\Gamma}_*$ when $\|\mathbf{x}^*\|_0 = s$. By (2.10), we can conclude that x^* is a P-stationary point of (SQCQP) for any $\tau \in (0, \bar{\tau})$. □

Theorem 2.3 (First-order sufficient condition). *Suppose that* $\mathbf{Q}_0, \mathbf{Q}_1 \ldots, \mathbf{Q}_k$ *are positive semi-definite. Then a P-stationary point of (SQCQP) is a local minimizer if* $\|\mathbf{x}^*\|_0 = s$ *and a global minimizer if* $\|\mathbf{x}^*\|_0 < s$.

Proof. Let $\mathbf{x}^*$ is a P-stationary point of (SQCQP). Then $\mathbf{x}^*$ is a KKT point and hence there is $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ such that (2.10) and (2.2). Note that $\boldsymbol{\nu}^*_{\Gamma_*} \in N_{\mathbb{X}_*}(\mathbf{x}^*_{\Gamma_*})$, it follows $\langle \boldsymbol{\nu}^*_{\Gamma_*}, \mathbf{x}_{\Gamma_*} - \mathbf{x}^*_{\Gamma_*} \rangle \leq 0$ for any $\mathbf{x} \in \mathbb{F} \cap \mathbb{S} \cap \mathbb{X}$. If $||\mathbf{x}^*||_0 < s$, then $\boldsymbol{g}^*_{\Gamma_*} = -\boldsymbol{\nu}^*_{\Gamma_*}$ and $\boldsymbol{g}^*_{\overline{\Gamma}_*} = 0$ by (2.10), yielding that

$$\langle \boldsymbol{g}^*, \mathbf{x} - \mathbf{x}^* \rangle = \langle \boldsymbol{g}^*_{\Gamma_*}, \mathbf{x}_{\Gamma_*} - \mathbf{x}^*_{\Gamma_*} \rangle = \langle -\boldsymbol{\nu}^*_{\Gamma_*}, \mathbf{x}_{\Gamma_*} - \mathbf{x}^*_{\Gamma_*} \rangle \geq 0, \tag{2.13}$$

Additionally, since $\{\mathbf{Q}_0, \mathbf{Q}_1 \ldots, \mathbf{Q}_k\}$ are positive semi-definite, $L(\cdot, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ is convex, thereby

$$f_0(\mathbf{x}) \geq L(\mathbf{x}, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \geq L(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) + \langle \boldsymbol{g}^*, \mathbf{x} - \mathbf{x}^* \rangle \geq f_0(\mathbf{x}^*), \tag{2.14}$$

for any $\mathbf{x} \in \mathbb{F} \cap \mathbb{S} \cap \mathbb{X}$, which means that $\mathbf{x}^*$ is a global minimizer of (SQCQP).
If $||\mathbf{x}^*||_0 = s$, then there is a sufficiently small neighborhood $\mathbb{N}^*$ of $\mathbf{x}^*$ such that $\mathrm{supp}(\mathbf{x}) = \Gamma_*$ for any $\mathbf{x} \in \mathbb{N}^* \cap (\mathbb{F} \cap \mathbb{S} \cap \mathbb{X})$, and hence $(\mathbf{x} - \mathbf{x}^*)_{\overline{\Gamma}_*} = 0$. This indicates condition (2.13) is also true, which by the convexity of $L(\cdot, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ implies that condition (2.14) holds for any $\mathbf{x} \in \mathbb{N}^* \cap (\mathbb{F} \cap \mathbb{S} \cap \mathbb{X})$. Hence, $\mathbf{x}^*$ is a local minimizer of (SQCQP). □

Based on Theorem 2.1, Theorem 2.2 and Theorem 2.3, we build the following relationships among the P-stationary points, KKT points and local minimizers,

$$P\text{-stationary points} \underset{\tau \in (0, \tilde{\tau})}{\Longleftrightarrow} \text{KKT points} \underset{\text{Assumption 2.1}}{\overset{\mathbf{Q}_i \succeq 0}{\Longleftrightarrow}} \text{Local minimizers} \tag{2.15}$$

where $\mathbf{Q}_i \succeq 0$ means $\mathbf{Q}_i$ is a symmetric positive semidefinite matrix.

2.2 Stationary equations

Hereafter, we denote some notation for simplicity. Given $\mathbf{x} \in \mathbb{S}$ and a constant $\tau > 0$, let

$$\begin{aligned} \mathbf{u} &:= \mathbf{x} - \tau \left(\nabla_x L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu} \right), \\ \mathbf{Y} &:= (\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m, \end{aligned}$$

which allows us to define a useful set by

$$\mathbb{T}_\tau(\mathbf{Y}) := \left\{ T \subseteq [n] : \ |T| = s, \ |u_i| \geq |u_j|, \ \forall i \in T, \ \forall j \in \overline{T} \right\}, \tag{2.16}$$

where u_i is the ith entry of $\mathbf{u}$. One can observe that any $T \in \mathbb{T}_\tau(\mathbf{Y})$ consists of s indices of the first s largest (in absolute value) entries of $\mathbf{u}$. Finally, given $\mathbf{Y}$ and an index set T, we define a system of equations as follows,

$$F(\mathbf{Y}; T) := \begin{bmatrix} (\nabla_x L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu})_T \\ \mathbf{x}_{\overline{T}} \\ \mathbf{x}_T - \Pi_{\mathbb{X}_T}(\mathbf{x}_T + \boldsymbol{\nu}_T) \\ \boldsymbol{\nu}_{\overline{T}} \\ \varphi(\mathbf{x}, \boldsymbol{\mu}) \\ \psi(\mathbf{x}, \boldsymbol{\lambda}) \end{bmatrix}. \tag{2.17}$$

This equation enables the further exploration of condition (2.9), as outlined below.

Theorem 2.4 (Stationary Equations). *Point $\mathbf{x}^*$ is a P-stationary point of (SQCQP) with $\tau > 0$ if and only if there is $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ such that*

$$F(\mathbf{Y}^*; T) = 0, \quad \forall\, T \in \mathbb{T}_\tau(\mathbf{Y}^*). \tag{2.18}$$

Furthermore,

$$\mathbb{T}_\tau(\mathbf{Y}^*) \equiv \mathbb{J}_s(\mathbf{x}^*). \tag{2.19}$$

Proof. By comparing (2.10) and (2.18), we need to show that

$$\mathbf{x}^* = \Pi_{\mathbb{S}}\left(\mathbf{x}^* - \tau(\boldsymbol{g}^* + \boldsymbol{\nu}^*)\right) \iff (\boldsymbol{g}^* + \boldsymbol{\nu}^*)_T = 0,\ \mathbf{x}^*_{\overline{T}} = 0,\ \forall\, T \in \mathbb{T}_\tau(\mathbf{Y}^*), \tag{2.20}$$

and the following relation,

$$\begin{aligned} &\begin{bmatrix} \mathbf{x}^*_{\Gamma_*} - \Pi_{\mathbb{X}_*}\left(\mathbf{x}^*_{\Gamma_*} + \boldsymbol{\nu}^*_{\Gamma_*}\right) \\ \boldsymbol{\nu}^*_{\overline{\Gamma}_*} \end{bmatrix} = 0, \quad \forall\, T \in \mathbb{T}_\tau(\mathbf{Y}^*) \\ \iff &\begin{bmatrix} \mathbf{x}^*_T - \Pi_{\mathbb{X}_T}\left(\mathbf{x}^*_T + \boldsymbol{\nu}^*_T\right) \\ \boldsymbol{\nu}^*_{\overline{T}} \end{bmatrix} = 0, \quad \forall\, T \in \mathbb{T}_\tau(\mathbf{Y}^*). \end{aligned} \tag{2.21}$$

Equivalence (2.20) follows from [62, Lemma 4] immediately. Thus, condition (2.19) follows from [60, Theorem 3] and (2.20). Based on this relation and $\mathbf{x}^*_{\overline{T}} = 0$, we can conclude that

$$\Gamma_* \subseteq T, \qquad \forall\, T \in \mathbb{T}_\tau(\mathbf{Y}^*). \tag{2.22}$$

We prove (2.21) by considering two cases. When $\|\mathbf{x}^*\|_0 = s$, it has $T = \Gamma_*$ due to (2.22) and $|T| = s$, which immediately shows (2.21). When $\|\mathbf{x}^*\|_0 < s$, it follows from (2.22) that $\overline{\Gamma}_* \supseteq \overline{T}$ for any $T \in \mathbb{T}_\tau(\mathbf{Y}^*)$, which means that $\boldsymbol{\nu}^*_{\overline{\Gamma}_*} = 0$ suffices to $\boldsymbol{\nu}^*_{\overline{T}} = 0$ and

$$\mathbf{x}^*_{T\setminus\Gamma_*} - \Pi_{\mathbb{X}_{T\setminus\Gamma_*}}\left(\mathbf{x}^*_{T\setminus\Gamma_*} + \boldsymbol{\nu}^*_{T\setminus\Gamma_*}\right) = 0 - \Pi_{\mathbb{X}_{T\setminus\Gamma_*}}(0+0) = 0,$$

showing inclusion ' $\Longrightarrow$'. From (2.21) and (2.19), $\boldsymbol{\nu}^*_{\overline{T}} = 0$ for any $T \in \mathbb{T}_\tau(\mathbf{Y}^*) = \mathbb{J}_s(\mathbf{x}^*)$, which contributes to $0 = \cup_{T\in\mathbb{J}_s(\mathbf{x}^*)}\boldsymbol{\nu}^*_{\overline{T}} = \boldsymbol{\nu}^*_{\overline{\Gamma}_*}$. Consequently, ' $\Longleftarrow$' follows due to $\Gamma_* \subseteq T$. □

The motivation for deriving the above theorem stems from numerical algorithm design. Although an unknown index set T is involved, the core of condition (2.18) is a system of equations, which makes a Newton-type method feasible.

2.3 Second-order optimality conditions

To end this section, we establish second-order optimality conditions for (SQCQP). For convenience, given $\mathbf{Y}^* := (\mathbf{x}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$, we define the following index sets,

$$\begin{aligned} \eta^1 &:= \{i \in [k] : \mu^*_i = 0, f_i(\mathbf{x}^*) < 0\}, & \eta^2 &:= \{i \in [m] : \lambda^*_i = 0, \langle \mathbf{a}_i, \mathbf{x}^*\rangle < b_i\}, \\ \theta^1 &:= \{i \in [k] : \mu^*_i = 0, f_i(\mathbf{x}^*) = 0\}, & \theta^2 &:= \{i \in [m] : \lambda^*_i = 0, \langle \mathbf{a}_i, \mathbf{x}^*\rangle = b_i\}, \\ \beta^1 &:= \{i \in [k] : \mu^*_i > 0, f_i(\mathbf{x}^*) = 0\}, & \beta^2 &:= \{i \in [m] : \lambda^*_i > 0, \langle \mathbf{a}_i, \mathbf{x}^*\rangle = b_i\}. \end{aligned} \tag{2.23}$$

In addition, for a given index set $T \in \mathbb{T}_\tau(\mathbf{Y}^*)$, let

$$\begin{aligned} \eta^3(T) &:= \{i \in T : \nu^*_i = 0, x^*_i \in \mathrm{int}(X_i)\}, & \eta^3 &:= \eta^3(\Gamma_*), \\ \theta^3(T) &:= \{i \in T : \nu^*_i = 0, x^*_i \in \mathrm{bd}(X_i)\}, & \theta^3 &:= \theta^3(\Gamma_*), \\ \beta^3(T) &:= \{i \in T : \nu^*_i \neq 0, x^*_i \in \mathrm{bd}(X_i)\}, & \beta^3 &:= \beta^3(\Gamma_*). \end{aligned} \tag{2.24}$$

Based on these sets, the critical cone for (SQCQP) is defined by

$$\mathbb{C}\left(\mathbf{Y}^*\right) := \left\{ \mathbf{d} \in \mathbb{R}^n : \begin{array}{rr} \langle \nabla f_i(\mathbf{x}^*), \mathbf{d}\rangle \le 0,\ i \in \theta^1, & \langle \nabla f_i(\mathbf{x}^*), \mathbf{d}\rangle = 0,\ i \in \beta^1 \\ \langle \mathbf{a}_i, \mathbf{d}\rangle \le 0,\ i \in \theta^2, & \langle \mathbf{a}_i, \mathbf{d}\rangle = 0,\ i \in \beta^2 \\ d_i \in T_{X_i}(x_i^*),\ i \in \theta^3, & d_i = 0,\ i \in \beta^3 \end{array} \right\}.$$

Theorem 2.5 (Second-order necessary condition). *Let $\mathbf{x}^*$ be a local minimizer of (SQCQP) and Assumption 2.1 hold at $\mathbf{x}^*$. Then there is unique $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ such that*

$$\langle \mathbf{d}, \mathbf{H}^*\mathbf{d}\rangle \ge 0, \quad \forall\ \mathbf{d} \in \mathbb{C}\left(\mathbf{Y}^*\right) \cap \mathbb{R}^n_{\Gamma_*}.$$

Proof. As $\mathbf{x}^*$ is a local minimizer of (SQCQP), it is also a local minimizer of

$$\min\ f_0(\mathbf{x}), \quad \text{s.t. } \mathbf{x} \in \mathbb{F},\ \mathbf{x}_{\overline{\Gamma}_*} = 0,\ \mathbf{x}_{\Gamma_*} \in \mathbb{X}_*.$$

Then the conclusion follows from Assumption 2.1 and [48, Example 13.25] immediately. □

Theorem 2.6 (Second-order sufficient condition). *Let $\mathbf{x}^*$ be a P-stationary point of (SQCQP), namely, there is $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ satisfying (2.9). Suppose that*

$$\langle \mathbf{d}, \mathbf{H}^*\mathbf{d}\rangle > 0, \quad \forall\ \mathbf{d}(\neq 0) \in \begin{cases} \mathbb{C}\left(\mathbf{Y}^*\right) \cap \mathbb{R}^n_{\Gamma_*}, & \text{if } \|\mathbf{x}^*\|_0 = s, \\ \mathbb{C}\left(\mathbf{Y}^*\right), & \text{if } \|\mathbf{x}^*\|_0 < s. \end{cases} \tag{2.25}$$

Then $\mathbf{x}^*$ is a strictly local minimizer of (SQCQP).

Proof. We argue by contradiction. Suppose that $\mathbf{x}^*$ is not a strictly local minimizer of (SQCQP). Then there is a sequence $\{\mathbf{x}^\ell\} \subset (\mathbb{F} \cap \mathbb{S} \cap \mathbb{X}) \backslash \{\mathbf{x}^*\}$ satisfying $\mathbf{x}^\ell \to \mathbf{x}^*$ and

$$f_0(\mathbf{x}^\ell) \le f_0\left(\mathbf{x}^*\right), \quad \ell = 1, 2, 3, \ldots \tag{2.26}$$

Let $\mathbf{d}^\ell := (\mathbf{x}^\ell - \mathbf{x}^*)/\|\mathbf{x}^\ell - \mathbf{x}^*\|$. Then $\|\mathbf{d}^\ell\| = 1$ and thus there as a convergent subsequence of $\{\mathbf{d}^\ell\}$ whose limit point $\mathbf{d}$ satisfies $\|\mathbf{d}\| = 1$. Without loss of generality, we assume that sequence $\{\mathbf{d}^\ell\}$ itself converges to $\mathbf{d}$. It is obvious that $\mathbf{d}_{\overline{\Gamma}_*} = 0$ when $\|\mathbf{x}^*\|_0 = s$ since $\mathrm{supp}(\mathbf{x}^\ell) = \Gamma_*$ when $\mathbf{x}^\ell$ is sufficiently close to $\mathbf{x}^*$. By (2.26),

$$0 \ge f_0(\mathbf{x}^\ell) - f_0(\mathbf{x}^*) = \langle \nabla f_0\left(x^*\right), \mathbf{x}^\ell - \mathbf{x}^*\rangle + o(\|\mathbf{x}^\ell - \mathbf{x}^*\|).$$

Dividing the both sides of the above inequality by $\|\mathbf{x}^\ell - \mathbf{x}^*\|$ and letting $\ell \to \infty$, we have

$$\langle \nabla f_0\left(\mathbf{x}^*\right), \mathbf{d}\rangle \le 0. \tag{2.27}$$

We note that a P-stationary point is also a KKT point, so $\boldsymbol{\nu}^*_{\Gamma_*} \in N_{\mathbb{X}_*}(\mathbf{x}^*_{\Gamma_*})$ by (2.2), leading to $\langle \boldsymbol{\nu}^*_{\Gamma_*}, \mathbf{x}^\ell_{\Gamma_*} - \mathbf{x}^*_{\Gamma_*}\rangle \le 0$ and hence $\langle \boldsymbol{\nu}^*_{\Gamma_*}, \mathbf{d}_{\Gamma_*}\rangle = \lim_{\ell\to\infty}\langle \boldsymbol{\nu}^*_{\Gamma_*}, \mathbf{d}^\ell_{\Gamma_*}\rangle \le 0$. This indicates $\mathbf{d}_{\Gamma_*} \in T_{\mathbb{X}_*}\left(\mathbf{x}^*_{\Gamma_*}\right)$ and thus $\langle \boldsymbol{\nu}^*_{\Gamma_*}, \mathbf{d}_{\Gamma_*}\rangle \le 0$. By the definitions of $(\theta^1, \beta^1, \theta^2, \beta^2)$ and $\{\mathbf{x}^\ell\} \subset \mathbb{F} \cap \mathbb{S} \cap \mathbb{X}$, it follows

$$\begin{aligned} f_i(\mathbf{x}^\ell) - f_i(\mathbf{x}^*) = f_i(\mathbf{x}^\ell) \le 0, &\quad \forall\ i \in \theta^1 \cup \beta^1, \\ \langle \mathbf{a}_i, \mathbf{x}^\ell - \mathbf{x}^*\rangle = \langle \mathbf{a}_i, \mathbf{x}^\ell\rangle - b_i - (\langle \mathbf{a}_i, \mathbf{x}^*\rangle - b_i) \le 0, &\quad \forall\ i \in \theta^2 \cup \beta^2. \end{aligned}$$

Based on these, the similar reasoning to show (2.27) enables us to derive that

$$\begin{cases} \langle \nabla f_i(\mathbf{x}^*), \mathbf{d} \rangle \leq 0, & i \in \theta^1 \cup \beta^1, \\ \langle \mathbf{a}_i, \mathbf{d} \rangle \leq 0, & i \in \theta^2 \cup \beta^2, \\ d_i \in T_{X_i}(x_i^*), & i \in \theta^3 \cup \beta^3. \end{cases} \tag{2.28}$$

Next, we further assert that

$$\begin{cases} \langle \nabla f_i(\mathbf{x}^*), \mathbf{d} \rangle = 0, & i \in \beta^1, \\ \langle \mathbf{a}_i, \mathbf{d} \rangle = 0, & i \in \beta^2, \\ d_i = 0, & i \in \beta^3. \end{cases} \tag{2.29}$$

Suppose there is an $i_0 \in \beta^1$ such that $\langle \nabla f_{i_0}(\mathbf{x}^*), \mathbf{d} \rangle < 0$. Then we have

$$\begin{aligned} 0 &= \langle \nabla_x L(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) + \boldsymbol{\nu}^*, \mathbf{d} \rangle \\ &= \langle \nabla f_0(\mathbf{x}^*), \mathbf{d} \rangle + \sum_{i \in \beta^1} \mu_i^* \langle \nabla f_i(\mathbf{x}^*), \mathbf{d} \rangle + \sum_{i \in \beta^2} \lambda_i^* \langle \mathbf{a}_i, \mathbf{d} \rangle + \sum_{i \in \beta^3} \nu_i^* d_i \\ &\leq \langle \nabla f_0(\mathbf{x}^*), \mathbf{d} \rangle + \mu_{i_0}^* \langle \nabla f_{i_0}(\mathbf{x}^*), \mathbf{d} \rangle \\ &< \langle \nabla f_0(\mathbf{x}^*), \mathbf{d} \rangle \leq 0, \end{aligned}$$

where the first equality is from (2.10) by the P-stationarity of $\mathbf{x}^*$ and $\mathbf{d}_{\overline{\Gamma}_*} = 0$ if $\|\mathbf{x}^*\|_0 = s$, the second equality and the first inequality are from (2.23), and the last inequality is from (2.27). This contradiction shows the first equation in (2.29). Then the same reasoning enables the last two equations in (2.29). Combining (2.28), (2.29), and $\mathbf{d}_{\overline{\Gamma}_*} = 0$ when $\|\mathbf{x}^*\|_0 = s$, we conclude that

$$\mathbf{d}(\neq 0) \in \begin{cases} \mathbb{C}(\mathbf{Y}^*) \cap \mathbb{R}^n_{\Gamma_*}, & \text{if } \|\mathbf{x}^*\|_0 = s, \\ \mathbb{C}(\mathbf{Y}^*), & \text{if } \|\mathbf{x}^*\|_0 < s. \end{cases} \tag{2.30}$$

Additionally, it follows from (2.23) and (2.26) that

$$L(\mathbf{x}^\ell, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = f_0(\mathbf{x}^\ell) + \sum_{i \in \beta^1} \mu_i^* f_i(\mathbf{x}^\ell) + \sum_{i \in \beta^2} \lambda_i^* (\langle \mathbf{a}_i, \mathbf{x}^* \rangle - b_i) \leq f_0(\mathbf{x}^*) = L(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*). \tag{2.31}$$

We note that $\mathrm{supp}(\mathbf{x}^\ell) = \Gamma_*$ when $\|\mathbf{x}^*\|_0 = s$ and $\boldsymbol{g}^*_{\overline{\Gamma}_*} = 0$ by (2.10) when $\|\mathbf{x}^*\|_0 < s$. Then the similar reasoning to show (2.13) can show $\langle \boldsymbol{g}^*, \mathbf{x}^\ell - \mathbf{x}^* \rangle \geq 0$. This together with (2.31) result in

$$\begin{aligned} 0 &\geq 2L(\mathbf{x}^\ell, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) - 2L(\mathbf{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \\ &= \langle \mathbf{x}^\ell - \mathbf{x}^*, \mathbf{H}^*(\mathbf{x}^\ell - \mathbf{x}^*) \rangle + 2\langle \boldsymbol{g}^*, \mathbf{x}^\ell - \mathbf{x}^* \rangle \\ &\geq \langle \mathbf{x}^\ell - \mathbf{x}^*, \mathbf{H}^*(\mathbf{x}^\ell - \mathbf{x}^*). \end{aligned}$$

Dividing the both sides of the above inequality by $\|\mathbf{x}^\ell - \mathbf{x}^*\|^2$ and letting $\ell \to \infty$, we have

$$\langle \mathbf{d}, \mathbf{H}^* \mathbf{d} \rangle \leq 0,$$

for any $\mathbf{d}$ satisfying (2.30), which contradicts to (2.25). The proof is completed. □

3 Nonsingularity of Generalized Jacobian

In this section, we aim to analyze the nonsingularity of the generalized Jacobian matrix ∂F of stationary equations (2.18). Given $\mathbf{Y}^* = (\mathbf{x}^*, \boldsymbol{\nu}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) \in \mathbb{R}^p$ with $p := 2n + k + m$, we select one index set $T \in \mathbb{T}_\tau(\mathbf{Y}^*)$. One can verify that $F(\cdot; T)$ is strongly semismooth at $\mathbf{Y}^*$ and any generalized Jacobian matrix $\mathbf{W} \in \partial F(\mathbf{Y}^*; T)$ takes the form of

$$\mathbf{W} := \begin{bmatrix} \mathbf{H}^*_{TT} & \mathbf{H}^*_{T\overline{T}} & \mathbf{B}^*_{T:} & \mathbf{A}^\top_{:T} & \mathbf{I} & 0 \\ 0 & \mathbf{I} & 0 & 0 & 0 & 0 \\ \mathbf{I} - \mathbf{C}^* & 0 & 0 & 0 & -\mathbf{C}^* & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{I} \\ \mathbf{U}^*_1(\mathbf{B}^*_{T:})^\top & \mathbf{U}^*_1(\mathbf{B}^*_{\overline{T}:})^\top & \mathbf{V}^*_1 & 0 & 0 & 0 \\ \mathbf{U}^*_2\mathbf{A}_{:T} & \mathbf{U}^*_2\mathbf{A}_{:\overline{T}} & 0 & \mathbf{V}^*_2 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{p\times p}. \tag{3.1}$$

The sub-matrices in (3.1) are defined by

$$\begin{aligned} \mathbf{B}^* &:= \begin{bmatrix} \nabla f_1(\mathbf{x}^*) & \cdots & \nabla f_k(\mathbf{x}^*) \end{bmatrix} \in \mathbb{R}^{n\times k}, \\ \mathbf{C}^* &:= \mathrm{Diag}\,(\mathbf{c}) \in \partial\Big(\Pi_{\mathbb{X}_T}\left(\mathbf{x}^*_T + \boldsymbol{\nu}^*_T\right)\Big) \in \mathbb{R}^{s\times s}, \\ \mathbf{U}^*_i &:= \mathrm{Diag}\left(\mathbf{u}^t\right) \in \mathbb{R}^{k\times k}, \\ \mathbf{V}^*_i &:= \mathrm{Diag}\left(\mathbf{v}^t\right) \in \mathbb{R}^{m\times m}, \quad i = 1, 2, \end{aligned} \tag{3.2}$$

where $\mathrm{Diag}(\mathbf{u})$ is a diagonal matrix with diagonal entries formed by $\mathbf{u}$, and for $t = 1, 2$,

$$\left(u^t_i, v^t_i\right) \begin{cases} = (0, -1), & \text{if } i \in \eta^t, \\ \in \left\{(\alpha, \beta) : \|(\alpha, \beta) - (1, -1)\|^2 \le 1\right\}, & \text{if } i \in \theta^t, \\ = (1, 0), & \text{if } i \in \beta^t, \end{cases} \qquad c_i \begin{cases} = 1, & \text{if } i \in \eta^3(T), \\ \in [0, 1], & \text{if } i \in \theta^3(T), \\ = 0, & \text{if } i \in \beta^3(T). \end{cases} \tag{3.3}$$

After simple elementary operations, we can get the following reduced matrix

$$\mathbf{G} := \begin{bmatrix} \mathbf{H}^*_{TT} & \mathbf{B}^*_{T:} & \mathbf{A}^\top_{:T} & \mathbf{I} \\ \mathbf{I} - \mathbf{C}^* & 0 & 0 & -\mathbf{C}^* \\ \mathbf{U}^*_1(\mathbf{B}^*_{T:})^\top & \mathbf{V}^*_1 & 0 & 0 \\ \mathbf{U}^*_2\mathbf{A}_{:T} & 0 & \mathbf{V}^*_2 & 0 \end{bmatrix} \in \mathbb{R}^{q\times q}, \tag{3.4}$$

where $q := 2s+k+m$, which has the same nonsingularity as $\mathbf{W}$ in (3.1). Based on these notation, we establish the CD-regularity of $F(\mathbf{Y}^*; T)$ for any given T as follows. We say $F(\mathbf{Y}^*; T)$ is CD-regular if any $\mathbf{W} \in \partial F(\mathbf{Y}^*; T)$ is nonsingular.

Theorem 3.1 (CD-regularity). *Let $\mathbf{x}^*$ be a P-stationary point of (SQCQP) with $\tau > 0$. Then for any given $T \in \mathbb{T}_\tau(\mathbf{Y}^*)$, $F(\mathbf{Y}^*; T)$ is CD-regular if Assumption 2.1 holds at $\mathbf{Y}^*$ and*

$$\langle \mathbf{d}, \mathbf{H}^*\mathbf{d}\rangle > 0, \forall 0 \neq \mathbf{d} \in \mathbb{Q}(\mathbf{Y}^*; T), \tag{3.5}$$

where $\mathbb{Q}$ is a cone defined by

$$\mathbb{Q}(\mathbf{Y}^*; T) := \left\{ \mathbf{d} \in \mathbb{R}^n : \begin{array}{r} \langle \nabla f_i(\mathbf{x}^*), \mathbf{d}\rangle = 0, \ i \in \beta^1 \\ \langle \mathbf{a}_i, \mathbf{d}\rangle = 0, \ i \in \beta^2 \\ d_i = 0, \ i \in \beta^3(T) \end{array} \right\}.$$

Proof. Given any $T \in \mathbb{T}_\tau(\mathbf{Y}^*)$, to show the nonsingularity of $\mathbf{W} \in \partial F\left(\mathbf{Y}^*; T\right)$ is equivalent to show that the following homogeneous system has a unique solution 0,

$$\mathbf{W}\left(\mathbf{d}^x;\ \mathbf{d}^\nu;\ \mathbf{d}^\mu;\ \mathbf{d}^\lambda\right) = 0.$$

This condition is equivalent to

$$\begin{cases} 0 = \mathbf{d}^x_{\overline{T}}, \\ 0 = \mathbf{d}^\nu_{\overline{T}}, \\ 0 = \mathbf{G}\left(\mathbf{d}^x_T;\ \mathbf{d}^\nu_T;\ \mathbf{d}^\mu;\ \mathbf{d}^\lambda\right), \end{cases} \tag{3.6}$$

which by (3.4), we get

$$\begin{cases} 0 = \mathbf{H}^*_{TT}\mathbf{d}^x_T + \mathbf{d}^\nu_T + \mathbf{B}^*_{T:}\mathbf{d}^\mu + \mathbf{A}^\top_{:T}\mathbf{d}^\lambda, \\ 0 = (\mathbf{I} - \mathbf{C}^*)\mathbf{d}^x_T - \mathbf{C}^*\mathbf{d}^\nu_T, \\ 0 = \mathbf{U}^*_1(\mathbf{B}^*_{T:})^\top \mathbf{d}^x_T + \mathbf{V}^*_1\mathbf{d}^\mu, \\ 0 = \mathbf{U}^*_2\mathbf{A}_{:T}\mathbf{d}^x_T + \mathbf{V}^*_2\mathbf{d}^\lambda. \end{cases} \tag{3.7}$$

Since $\mathbf{Y}^*$ satisfies (2.9) and (3.3), we have

$$\begin{aligned} \eta^t &= \{i : \left(u^t_i, v^t_i\right) = (0, -1)\}, & \eta^3(T) &= \{i \in T : c_i = 1\}, \\ \theta^t &= \{i : \left\|\left(u^t_i, v^t_i\right) - (1, -1)\right\|^2 \le 1\}, & \theta^3(T) &= \{i \in T : c_i \in [0, 1]\}, \\ \beta^t &= \{i : \left(u^t_i, v^t_i\right) = (1, 0)\}, & \beta^3(T) &= \{i \in T : c_i = 0\}, \end{aligned}$$

where $t = 1, 2$. Based on these indices, we further define

$$\begin{aligned} \widetilde{\eta}^t &:= \{i \in \theta^t : \left(u^t_i, v^t_i\right) = (0, -1)\} \cup \eta^t, & \widetilde{\eta}^3 &:= \{i \in \theta^3(T) : c_i = 1\} \cup \eta^3(T), \\ \widetilde{\theta}^t &:= \{i : \left\|\left(u^t_i, v^t_i\right) - (1, -1)\right\|^2 < 1\}, & \widetilde{\theta}^3 &:= \{i \in \theta^3(T) : c_i \in (0, 1)\}, \\ \widetilde{\beta}^t &:= \{i \in \theta^: \left(u^t_i, v^t_i\right) = (1, 0)\} \cup \beta^t, & \widetilde{\beta}^3 &:= \{i \in \theta^3(T) : c_i = 0\} \cup \beta^3(T). \end{aligned}$$

One can observe that $u^t_i > 0$ and $v^t_i < 0$ for any $i \in \widetilde{\theta}^t, t = 1, 2$. These and the last three equations in (3.7) lead to the facts in Table 2,

Table 2: Conditions on $(\widetilde{\eta}^t, \widetilde{\theta}^t, \widetilde{\beta}^t), t = 1, 2, 3$.

	$i \in \widetilde{\eta}^t$	$i \in \widetilde{\theta}^t$	$i \in \widetilde{\beta}^t$
$t = 1$	$d^\mu_i = 0$	$((\mathbf{B}^*_{T:})^\top \mathbf{d}^x_T)_i = -(v^1_i/u^1_i)d^\mu_i$	$\langle(\nabla f_i(\mathbf{x}^*))_T, \mathbf{d}^x_T\rangle = 0$
$t = 2$	$d^\lambda_i = 0$	$\langle(\mathbf{a}_i)_T, \mathbf{d}^x_T\rangle = -(v^2_i/u^2_i)d^\lambda_i$	$\langle(\mathbf{a}_i)_T,\ \mathbf{d}^x_T\rangle = 0$
$t = 3$	$i \in T,\ d^\nu_i = 0$	$i \in T, d^x_i = c_i d^\nu_j/(1 - c_i)$	$d^x_i = 0$

The conditions in the third column of Table 2, $\beta^t \subseteq \widetilde{\beta}^t, t = 1, 2$, and $\beta^3(T) \subseteq \widetilde{\beta}^3$ yield that

$$\mathbf{d}^x \in \mathbb{Q}\left(\mathbf{Y}^*; T\right). \tag{3.8}$$

Multiplying the both sides of the first equation in (3.7) by $\mathbf{d}^x_T$ derives

$$\langle \mathbf{d}^x_T, \mathbf{H}^*_{TT}\mathbf{d}^x_T\rangle + \sum_{i \in T \cap \widetilde{\theta}^3} \frac{c_i(d^\nu_j)^2}{1 - c_i} - \sum_{i \in \widetilde{\theta}^1} \frac{v^1_i(d^\mu_i)^2}{u^1_i} - \sum_{i \in \widetilde{\theta}^2} \frac{v^2_i(d^\lambda_i)^2}{u^2_i} = 0. \tag{3.9}$$

which together with (3.5) and (3.8) immediately results in

$$\left(\mathbf{d}_T^x,\ \mathbf{d}^{\nu}_{T\cap\widetilde{\theta}^3},\ \mathbf{d}^{\mu}_{\widetilde{\theta}^1},\ \mathbf{d}^{\lambda}_{\widetilde{\theta}^2}\right) = (0,0,0,0).$$

This recalling the first column of Table 2 further derives

$$\left(\mathbf{d}^{\nu}_{T\cap(\widetilde{\eta}^3\cup\widetilde{\theta}^3)},\ \mathbf{d}^{\mu}_{\widetilde{\eta}^1\cup\widetilde{\theta}^1},\ \mathbf{d}^{\lambda}_{\widetilde{\eta}^2\cup\widetilde{\theta}^2}\right) = (0,0,0).$$

By inserting these values into the first equation in (3.7), it holds that

$$\mathbf{I}_{T(T\cap\widetilde{\beta}^3)}\mathbf{d}^{\nu}_{(T\cap\widetilde{\beta}^3)} + \mathbf{B}^*_{T\widetilde{\beta}^1}\mathbf{d}^{\mu}_{\widetilde{\beta}^1} + \mathbf{A}^{\top}_{\widetilde{\beta}^2 T}\mathbf{d}^{\lambda}_{\widetilde{\beta}^2} = 0. \tag{3.10}$$

One can check that $\widetilde{\beta}^1 \subseteq \mathcal{A}_1(\mathbf{x}^*)$, $\widetilde{\beta}^2 \subseteq \mathcal{A}_2(\mathbf{x}^*)$, $T\cap\widetilde{\beta}^3 \subseteq \mathcal{A}_3(\mathbf{x}^*;T)$, and $\Gamma_* \subseteq T$ for any $T \in \mathbb{J}_s(\mathbf{x}^*)$. Hence, Assumption 2.1 indicates the following groups of vectors are linearly independent,

$$\left\{(\nabla f_i(\mathbf{x}^*))_T : i \in \widetilde{\beta}^1\right\} \cup \left\{(\mathbf{a}_i)_T : i \in \widetilde{\beta}^2\right\} \cup \left\{(\mathbf{e}_i)_T : i \in T\cap\widetilde{\beta}^3\right\},$$

which by (3.10) implies $(\mathbf{d}^{\nu}_{(T\cap\widetilde{\beta}^3)}, \mathbf{d}^{\mu}_{\widetilde{\beta}^1}, \mathbf{d}^{\lambda}_{\widetilde{\beta}^2}) = (0,0,0)$. Overall, $(\mathbf{d}^x_T, \mathbf{d}^{\nu}_T, \mathbf{d}^{\mu}, \mathbf{d}^{\lambda}) = 0$, which combining the first two conditions in (3.6) yields the conclusion. □

In the sequel, we pay attention to point $\mathbf{Y} = (\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ with $\mathbf{x} \in \mathbb{S}$ in a neighbourhood of $\mathbf{Y}^*$. Therefore, we define a set by

$$\mathbb{E} := \mathbb{S} \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m.$$

Moreover, we note that each $\nabla f_i(\mathbf{x})$ is Lipschitz continuous, so $\nabla_x L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu}$ is locally Lipschitz continuous. Leveraging this fact, we build the following property of any $\mathbf{Y}$ around $\mathbf{Y}^*$.

Lemma 3.1. *Let $\mathbf{x}^*$ be a P-stationary point of (SQCQP) with $\tau > 0$. Then there is a neighbourhood $\mathbb{N}^*$ of $\mathbf{Y}^*$ such that for any $\mathbf{Y} \in \mathbb{E}\cap\mathbb{N}^*$,*

$$\mathbb{T}_\tau(\mathbf{Y}) \subseteq \mathbb{T}_\tau(\mathbf{Y}^*), \qquad \Gamma_* \subseteq (\mathrm{supp}(\mathbf{x})\cap T),\ \forall\ T \in \mathbb{T}_\tau(\mathbf{Y}). \tag{3.11}$$

Consequently, $F(\mathbf{Y}^*;T) = 0$ for any $T \in \mathbb{T}_\tau(\mathbf{Y})$.

Proof. Since the second claim follows from (3.11) and (2.18), we only prove (3.11). Let neighbourhood $\mathbb{N}^*$ be a sufficiently small region. Then (3.11) is true when $\|\mathbf{x}^*\|_0 = s$ due to

$$\mathbb{T}_\tau(\mathbf{Y}) = \mathbb{T}_\tau(\mathbf{Y}^*) = \{\Gamma_*\} = \{\mathrm{supp}(\mathbf{x})\}.$$

Therefore, we only prove the case of $\|\mathbf{x}^*\|_0 < s$. One can easily check that for any $\mathbf{Y} \in \mathbb{E}\cap\mathbb{N}^*$,

$$\Gamma_* \subseteq \mathrm{supp}(\mathbf{x}), \qquad \mathbb{J}_s(\mathbf{x}) \subseteq \mathbb{J}_s(\mathbf{x}^*) = \mathbb{T}_\tau(\mathbf{Y}^*), \tag{3.12}$$

where the equality is from (2.19). If $\Gamma_* = \emptyset$, the conclusion holds evidently. Now we focus on $\Gamma_* \neq \emptyset$. This means that $\min_{i\in\Gamma_*}|x_i| > 0$ because $\mathbf{x}$ is close to $\mathbf{x}^*$. Moreover, it follows from (2.10) that $\boldsymbol{g}^* + \boldsymbol{\nu}^* = 0$. As a result of the locally Lipschitz continuity, $\tau\|\nabla_x L(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\nu}\|_\infty < \min_{i\in\Gamma_*}|x_i|$, which indicates that

$$\Gamma_* \in T, \quad \forall\ T \in \mathbb{T}_\tau(\mathbf{Y}).$$

This recalling the definition of $\mathbb{J}_s(\mathbf{x}^*)$ enables us to conclude that $\mathbb{T}_\tau(\mathbf{Y}) \subseteq \mathbb{J}_s(\mathbf{x}^*)$, which together with (3.12) shows the desired result. □

Theorem 3.2. *Let $\mathbf{x}^*$ be a P-stationary point of (SQCQP) with $\tau > 0$ and the assumptions in Theorem 3.1 hold at $\mathbf{Y}^*$. Then there is a neighbourhood $\mathbb{N}^*$ of $\mathbf{Y}^*$ such that the following statements are valid for any $\mathbf{Y} \in \mathbb{E} \cap \mathbb{N}^*$.*

1) Every $\mathbf{W} \in \partial F(\mathbf{Y}; T)$ is nonsingular for each given $T \in \mathbb{T}_\tau(\mathbf{Y})$.

2) There is a constant $C^* > 0$ such that

$$\left\|\mathbf{W}^{-1}\right\| \leq C^*, \quad \forall\ \mathbf{W} \in \partial F(\mathbf{Y}; T),\ \forall T \in \mathbb{T}_\tau(\mathbf{Y}). \tag{3.13}$$

where $\|\mathbf{W}\|$ represents the spectral norm of $\mathbf{W}$.

Proof. 1) By invoking Theorems 3.1, for each given $T_* \in \mathbb{T}_\tau(\mathbf{Y}^*)$, any $\mathbf{W}^* \in \partial F\left(\mathbf{Y}^*; T_*\right)$ is nonsingular. Consider any point $\mathbf{Y} \in \mathbb{E} \cap \mathbb{N}^*$. Lemma 3.1 means that $\mathbb{T}_\tau(\mathbf{Y}) \subseteq \mathbb{T}_\tau(\mathbf{Y}^*)$. As a result, for each given $T \in \mathbb{T}_\tau(\mathbf{Y})$, any $\mathbf{W}^* \in \partial F\left(\mathbf{Y}^*; T\right)$ is nonsingular. We first prove that for each given $T \in \mathbb{T}_\tau(\mathbf{Y})$, every $\mathbf{W} \in \partial F(\mathbf{Y}; T)$ is nonsingular and there is a constant $C_T > 0$ such that $\|\mathbf{W}^{-1}\| \leq C_T$. If this is not true, then given such T, there is a sequence $\mathbf{Y}^\ell (\in \mathbb{N}^*) \to \mathbf{Y}^*$ and $\mathbf{W}^\ell \in \partial F\left(\mathbf{Y}^\ell; T\right)$ satisfying either all $\mathbf{W}^\ell$ are singular or $\|(\mathbf{W}^\ell)^{-1}\| \to \infty$. The locally Lipschitz continuity of $F\left(\cdot; T\right)$ means $\partial F\left(\cdot; T\right)$ is bounded in $\mathbb{N}^*$. By passing to a subsequence, we may assume $\mathbf{W}^\ell \to \mathbf{W}^*$. Then $\mathbf{W}^*$ is singular, contradicting to the nonsingularity of $\mathbf{W}^*$
2) We note that there are finitely many $T \in \mathbb{T}_\tau(\mathbf{Y})$. By setting $C^* = \max_{T \in \mathbb{T}_\tau(\mathbf{Y})} C_T$, condition (3.13) follows immediately. □

4 A Semi-smooth Newton Algorithm

In this section, we leverage a semi-smooth Newton-type method to solve stationary equations (2.18) for problem (SQCQP), before which we define some notation for the ease of reading. Given a point $\mathbf{Y} = (\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^m$ and $T \in \mathbb{T}_s(\mathbf{Y})$, we define $\mathbf{y}$ and a direction $\mathbf{d}$ by

$$\begin{aligned} \mathbf{y} &:= \left(\mathbf{x}_T;\ \mathbf{x}_{\overline{T}};\ \boldsymbol{\nu}_T;\ \boldsymbol{\nu}_{\overline{T}};\ \boldsymbol{\mu};\ \boldsymbol{\lambda}\right) \quad \in \mathbb{R}^p, \\ \mathbf{d} &:= \left(\mathbf{d}^x_T;\ \mathbf{d}^x_{\overline{T}};\ \mathbf{d}^\nu_T;\ \mathbf{d}^\nu_{\overline{T}};\ \mathbf{d}^\mu;\ \mathbf{d}^\lambda\right) \in \mathbb{R}^p, \end{aligned} \tag{4.1}$$

namely, $\mathbf{y}$ is a vectorization of $\mathbf{Y}$. In addition, by letting

$$J := n + \overline{T}, \qquad K := [p] \setminus (\overline{T} \cup J),$$

one can write $\mathbf{d} = (\mathbf{d}_{\overline{T}};\ \mathbf{d}_J;\ \mathbf{d}_K)$, where

$$\mathbf{d}_{\overline{T}} = \mathbf{d}^x_{\overline{T}} \in \mathbb{R}^{n-s}, \qquad \mathbf{d}_J = \mathbf{d}^\nu_{\overline{T}} \in \mathbb{R}^{n-s}, \qquad \mathbf{d}_K = \left(\mathbf{d}^x_T;\ \mathbf{d}^\nu_T;\ \mathbf{d}^\mu;\ \mathbf{d}^\lambda\right) \in \mathbb{R}^q.$$

4.1 Algorithm design

To employ the Newton method, we need to find a solution to a system of linear equations. For (SQCQP), given current $\mathbf{Y}^\ell$ and $T_\ell \in \mathbb{T}_s(\mathbf{Y}^\ell)$ at step ℓ, we aim to find a direction $\mathbf{d}^\ell$ by solving stationary equations $\mathbf{W}^\ell \mathbf{d}^\ell = -F(\mathbf{Y}^\ell; T_\ell)$. In fact, it is easy to see from (3.1) and (3.4) that

$$\mathbf{W}^\ell \mathbf{d}^\ell = -F(\mathbf{Y}^\ell; T_\ell) \iff \begin{cases} \mathbf{d}^\ell_{\overline{T}_\ell} = -\mathbf{x}^\ell_{\overline{T}_\ell}, \\ \mathbf{d}^\ell_{J_\ell} = -\boldsymbol{\nu}^\ell_{\overline{T}_\ell}, \\ \mathbf{G}^\ell \mathbf{d}^\ell_{K_\ell} = \mathbf{D}^\ell \mathbf{x}^\ell_{\overline{T}_\ell} - \left(F(\mathbf{Y}^\ell; T_\ell)\right)_{K_\ell}, \end{cases} \tag{4.2}$$

Algorithm 1: SNSQP: Semi-smooth Newton algorithm for (SQCQP)

Initialize $\mathbf{Y}^0$, $\tau > 0$, $\varepsilon > 0$, $\rho \in (0,1)$, and $\sigma \in (0, 1/2)$. Select $T_0 \in \mathbb{T}_\tau(\mathbf{Y}^0)$, and set $\ell \Leftarrow 0$.

while $\|F(\mathbf{Y}^\ell; T_\ell)\| > \varepsilon$ **do**

Choose $\mathbf{W}^\ell \in \partial F(\mathbf{Y}^\ell; T_\ell)$.

Compute $\mathbf{d}^\ell_{\overline{T}_\ell} = -\mathbf{x}^\ell_{\overline{T}_\ell}$ and $\mathbf{d}^\ell_{J_\ell} = -\boldsymbol{\nu}^\ell_{\overline{T}_\ell}$.

Compute $\mathbf{d}^\ell_{K_\ell}$ by solving (4.5) if it is solvable and by solving (4.6) otherwise.

Find the smallest non-negative integer t_ℓ such that

$$\Psi\left(\mathbf{y}^\ell + \mathbf{d}^\ell(\rho^{t_\ell}); T_\ell\right) \le \Psi\left(\mathbf{y}^\ell;\ T_\ell\right) + \sigma\rho^{t_\ell}\left\langle F(\mathbf{Y}^\ell;\ T_\ell), \mathbf{W}^\ell\mathbf{d}^\ell\right\rangle \tag{4.4}$$

Update $\mathbf{y}^{\ell+1} = \mathbf{y}^\ell + \mathbf{d}^\ell(\rho^{t_\ell})$ and $T_{\ell+1} \in \mathbb{T}_\tau(\mathbf{Y}^\ell)$. Set $\ell \Leftarrow \ell + 1$.

where $\mathbf{D}^\ell$ is a sub-matrix of $\mathbf{W}^\ell$ defined by

$$\mathbf{D}^\ell := \begin{bmatrix} \mathbf{H}^\ell_{T_\ell\overline{T}_\ell} \\ 0 \\ \mathbf{U}^\ell_1(\mathbf{B}^\ell_{\overline{T}_\ell:})^\top \\ \mathbf{U}^\ell_2\mathbf{A}_{:\overline{T}_\ell} \end{bmatrix} \in \mathbb{R}^{q\times(n-s)}. \tag{4.3}$$

Hereafter, let $\mathbf{H}^\ell =: \nabla^2_{xx}L\left(\mathbf{x}^\ell, \boldsymbol{\mu}^\ell, \boldsymbol{\lambda}^\ell\right)$. Matrices $\mathbf{W}^\ell$ and $\mathbf{G}^\ell$ are calculated similarly to (3.1) and (3.4), where $(\mathbf{B}^\ell, \mathbf{C}^\ell, \mathbf{U}^\ell_1, \mathbf{U}^\ell_2, \mathbf{V}^\ell_1, \mathbf{V}^\ell_2)$ are computed using (3.2) by replacing $\mathbf{Y}^*$ by $\mathbf{Y}^\ell$. The equivalence in (4.2) enables a significant dimensional reduction, from $(p\times p)$ to $(q\times q)$, making large-scale computation possible. However, when matrix $\mathbf{G}^\ell$ is in a bad condition, solving the following equations may yield an unfavorable direction $\mathbf{d}^\ell_{K_\ell}$,

$$\mathbf{G}^\ell\mathbf{d}^\ell_{K_\ell} = \mathbf{D}^\ell\mathbf{x}^\ell_{\overline{T}_\ell} - \left(F(\mathbf{Y}^\ell; T_\ell)\right)_{K_\ell}. \tag{4.5}$$

To overcome such a drawback, we turn to solve the following equations as a compensation,

$$\left(\mathbf{G}^{\ell\top}\mathbf{G}^\ell + \kappa_\ell\mathbf{I}\right)\mathbf{d}^\ell_{K_\ell} = \mathbf{G}^{\ell\top}\left(\mathbf{D}^\ell\mathbf{x}^\ell_{\overline{T}_\ell} - \left(F(\mathbf{Y}^\ell; T_\ell)\right)_{K_\ell}\right), \tag{4.6}$$

where $\kappa_\ell > 0$ is a decreasing scalar along with ℓ rising and $\lim_{\ell\to\infty}\kappa_\ell = 0$. For example, $\kappa_\ell = 0.01/\ell$ used in the subsequent numerical experiment. Finally, given a step size $\alpha > 0$, we define

$$\mathbf{d}^\ell(\alpha) := \left(\mathbf{d}^\ell_{\overline{T}_\ell};\ \mathbf{d}^\ell_{J_\ell};\ \alpha\mathbf{d}^\ell_{K_\ell}\right) \tag{4.7}$$

to ensure point $\mathbf{x}^{\ell+1} \in \mathbb{S}$. This is because from (4.2), for any $\alpha > 0$, we have

$$\begin{aligned}\mathbf{y}^{\ell+1} = \mathbf{y}^\ell + \mathbf{d}^\ell(\alpha) &= \left(\mathbf{x}^\ell_{\overline{T}_\ell} + \mathbf{d}^\ell_{\overline{T}_\ell};\ \boldsymbol{\nu}^\ell_{J_\ell} + \mathbf{d}^\ell_{J_\ell};\ \mathbf{y}^{\ell+1}_{K^\ell_\ell} + \alpha\mathbf{d}^\ell_{K_\ell}\right) \\ &= \left(0;\ 0;\ \mathbf{y}^{\ell+1}_{K_\ell} + \alpha\mathbf{d}^\ell_{K_\ell}\right).\end{aligned} \tag{4.8}$$

We observe that $\mathbf{y}^{\ell+1}_{\overline{T}_\ell}$ corresponds to $\mathbf{x}^{\ell+1}_{\overline{T}_\ell}$, which by $|\overline{T}_\ell| = n - s$ results in $\mathbf{x}^{\ell+1} \in \mathbb{S}$. Consequently, $\mathrm{supp}(\mathbf{x}^\ell) \subseteq T_{\ell-1}$, leading to $|T| := |T_{\ell-1} \cap \overline{T}_\ell| \le s$. This suffices to $\mathbf{D}^\ell\mathbf{x}^\ell_{\overline{T}_\ell} = \mathbf{D}^\ell_{:T}\mathbf{x}^\ell_T$ in (4.2), which significantly reduces the computational complexity of evaluating $\mathbf{D}^\ell\mathbf{x}^\ell_{\overline{T}_\ell}$ from $O(q(n-s))$ to $O(qs)$, achieving a substantial efficiency gain. Overall, the complexity of solving (4.2) is about

$$O(s^3 + ks^2 + qs).$$

The developed algorithmic framework is presented in Algorithm 1, which is called SNSQP, an abbreviation for the Semi-smooth Newton for (SQCQP). To make the algorithm steady, we adopt an Armijo line search to determine the step size at every step based on a merit function,

$$\Psi\left(\mathbf{y};T\right) := \frac{1}{2}\left\|F\left(\mathbf{Y};T\right)\right\|^2. \tag{4.9}$$

Finally, it is important to point out that the stationary equations, (2.18), involve an unknown set T. If this set were known in advance, one could employ the semismooth Newton-type method to solve these equations with a fixed T and establish the local convergence rate in a way provided in [47]. However, set T may change from one iteration to the next. A different set T_ℓ leads to a distinct system of equations, $F(\mathbf{Y};T_\ell)=0$. Consequently, at each step, the algorithm computes a Newton direction for a different system of equations, rather than a fixed one. This is where the standard proof for the quadratic rate of convergence does not apply to our case. Nevertheless, we establish this property for Algorithm 1 under the assumptions in Theorem 3.1, as outlined below.

4.2 Local convergence rate

In the sequel, we use notation $\|\mathbf{Y}\|^2$ to denote

$$\|\mathbf{Y}\|^2 = \|\mathbf{x}\|^2 + \|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\lambda}\|^2 + \|\boldsymbol{\nu}\|^2.$$

Similar to $\mathbf{y}$ defined in (4.1), let $\mathbf{z}\in\mathbb{R}^p$ be the vectorization of $\mathbf{Z}\in\mathbb{E}$.

Corollary 4.1. *Let $\mathbf{x}^*$ be a P-stationary point of (SQCQP) with $\tau>0$ and the assumptions in Theorem 3.1 hold at $\mathbf{Y}^*$. Then for any given $\epsilon\in(0,1)$, there is a neighbourhood $\mathbb{N}_\epsilon^*$ of $\mathbf{Y}^*$ such that, for any point $\mathbf{Y}\in\mathbb{E}\cap\mathbb{N}_\epsilon^*$,*

$$\left\|F\left(\mathbf{Z};T\right)\right\| \le \epsilon\left\|F\left(\mathbf{Y};T\right)\right\|, \quad \forall\ \mathbf{W}\in\partial F(\mathbf{Y};T),\ \forall\ T\in\mathbb{T}_\tau(\mathbf{Y}). \tag{4.10}$$

where $\mathbf{z} := \mathbf{y} - \mathbf{W}^{-1}F\left(\mathbf{Y};T\right)$ is the vectorization of $\mathbf{Z}$.

Proof. Let $\mathbb{N}^*$ be a sufficiently small neighbourhood of $\mathbf{Y}^*$ and consider $\mathbf{Y}\in\mathbb{E}\cap\mathbb{N}^*$. For each given $T\in\mathbb{T}_\tau(\mathbf{Y})$, every $\mathbf{W}\in\partial F(\mathbf{Y};T)$ is nonsingular from Theorem 3.2. Then it follows,

$$\begin{aligned}
\|\mathbf{Z}-\mathbf{Y}^*\| &= \left\|\mathbf{y}-\mathbf{W}^{-1}F(\mathbf{Y};T)-\mathbf{y}^*\right\| \\
&= \left\|\mathbf{W}^{-1}\left(F(\mathbf{Y};T)-F\left(\mathbf{Y}^*;T\right)-\mathbf{W}(\mathbf{y}-\mathbf{y}^*)\right)\right\| \\
&\le \left\|\mathbf{W}^{-1}\right\|\left\|F(\mathbf{Y};T)-F\left(\mathbf{Y}^*;T\right)-\mathbf{W}(\mathbf{y}-\mathbf{y}^*)\right\| \\
&= O(\|\mathbf{Y}-\mathbf{Y}^*\|^2),
\end{aligned} \tag{4.11}$$

where the second equality is from Lemma 3.1 that $F(\mathbf{Y}^*;T)=0$ for any $T\in\mathbb{T}_\tau(\mathbf{Y})$ and the last equality holds due to the strongly semismoothness of $F(\cdot;T)$ at $\mathbf{Y}^*$ for given T. The above condition means that $\mathbf{Z}$ also lies in $\mathbb{N}^*$, which by semismooth implying B-differentiability, we have

$$\left\|F(\mathbf{Z};T)-F\left(\mathbf{Y}^*;T\right)-F'\left(\mathbf{Y}^*;T;\mathbf{Z}-\mathbf{Y}^*\right)\right\| = o(\|\mathbf{Z}-\mathbf{Y}^*\|) \tag{4.12}$$

where $F'\left(\mathbf{Y}^*;T;\mathbf{Z}-\mathbf{Y}^*\right)$ is the directional derivative of the function $F\left(\cdot;T\right)$ at point $\mathbf{Y}^*$ in direction $\mathbf{Z}-\mathbf{Y}^*$. Now we can claim that given $\epsilon\in(0,1)$, there is a smaller neighbourhood $\mathbb{N}_\epsilon^*\subseteq\mathbb{N}^*$ such that, for any $\mathbf{Y}\in\mathbb{N}_\epsilon^*\cap\mathbb{E}$,

$$\begin{aligned}
\left\|F(\mathbf{Z};T)-F'\left(\mathbf{Y}^*;T;\mathbf{Z}-\mathbf{Y}^*\right)\right\| &\le \|\mathbf{Z}-\mathbf{Y}^*\|, \\
\|\mathbf{Z}-\mathbf{Y}^*\| &\le \kappa_*\|\mathbf{Y}-\mathbf{Y}^*\|,
\end{aligned} \tag{4.13}$$

where $\kappa_* := \epsilon/(C^*(c^*+1)+\epsilon)$, the first inequality is from $F(\mathbf{Y}^*;T)=0$ and (4.12), the second one is due to (4.11), and $c^* := \sup_{\mathbf{Y}\in\mathbb{N}_\epsilon^*}\sup_{\mathbf{W}\in\partial F(\mathbf{Y};T)}\|\mathbf{W}\|$. The above conditions derive that

$$\begin{aligned}\|\mathbf{Y}-\mathbf{Y}^*\| &\le \|\mathbf{Z}-\mathbf{Y}\|+\|\mathbf{Z}-\mathbf{Y}^*\|\\ &= \|\mathbf{W}^{-1}F(\mathbf{Y};T)\|+\|\mathbf{Z}-\mathbf{Y}^*\|\\ &\le C^*\|F(\mathbf{Y};T)\|+\|\mathbf{Z}-\mathbf{Y}^*\|\\ &\le C^*\|F(\mathbf{Y};T)\|+\kappa_*\|\mathbf{Y}-\mathbf{Y}^*\|,\end{aligned}$$

where the first inequality is from (3.13). This leads to

$$\|\mathbf{Y}-\mathbf{Y}^*\| \le \frac{\epsilon}{(c^*+1)\kappa_*}\|F(\mathbf{Y};T)\|,$$

which by (4.13) further results in

$$\begin{aligned}\|F(\mathbf{Z};T)\| &\le \left\|F'\left(\mathbf{Y}^*;T;\mathbf{Z}-\mathbf{Y}^*\right)\right\|+\|\mathbf{Z}-\mathbf{Y}^*\|\\ &\le (c^*+1)\|\mathbf{Z}-\mathbf{Y}^*\|\\ &\le (c^*+1)\kappa_*\|\mathbf{Y}-\mathbf{Y}^*\|\\ &\le \epsilon\|F(\mathbf{Y};T)\|,\end{aligned}$$

showing the desired result. □

Theorem 4.1. *Let $\mathbf{x}^*$ be a P-stationary point of (SQCQP) with $\tau>0$ and the assumptions in Theorem 3.1 hold at $\mathbf{Y}^*$. Then in a neighborhood of $\mathbf{Y}^*$, Algorithm 1 admits full Newton steps and converges to $\mathbf{Y}^*$ quadratically.*

Proof. Let $\mathbb{N}^*$ be a sufficiently small neighbourhood of $\mathbf{Y}^*$ and $\{\mathbf{Y}^\ell\}$ be the sequence generated by Algorithm 1 in $\mathbb{N}^*$. According to (4.8), $\mathbf{x}^{\ell+1}\in\mathbb{S}$ and thus $\mathbf{Y}^\ell\in\mathbb{E}\cap\mathbb{N}^*$. By Lemma 3.1, we have $T_\ell\in\mathbb{T}_\tau(\mathbf{Y}^\ell)\subseteq\mathbb{T}_\tau(\mathbf{Y}^*)$. Recalling Theorem 3.2, any $\mathbf{W}^\ell\in\partial F(\mathbf{Y}^\ell;T_\ell)$ is nonsingular and $\|(\mathbf{W}^\ell)^{-1}\|$ is bounded. Consequently, equations (4.2) or (4.5) are solvable. Now, we prove that the full Newton steps always admit, namely $\rho^{t_\ell}=1$. Let $\mathbf{z}^\ell := \mathbf{y}^\ell+\mathbf{d}^\ell(1)=\mathbf{y}^\ell+\mathbf{d}^\ell$ be the vectorization of $\mathbf{Z}^\ell\in\mathbb{E}$. Given $\sigma\in(0,1/2)$, it follows from Corollary 4.1 that

$$\left\|F(\mathbf{Z}^\ell;T_\ell)\right\| \le \sqrt{1-2\sigma}\left\|F(\mathbf{Y}^\ell;T_\ell)\right\|,$$

which immediately delivers

$$\Psi(\mathbf{z}^\ell;T_\ell)=\frac{1}{2}\|F(\mathbf{Z}^\ell;T_\ell)\|^2 \le \frac{1-2\sigma}{2}\|F(\mathbf{Y}^\ell;T_\ell)\|^2=(1-2\sigma)\Psi(\mathbf{y}^\ell;T_\ell).$$

This condition indicates that

$$\Psi(\mathbf{y}^\ell+\mathbf{d}^\ell;T_\ell)-\Psi(\mathbf{y}^\ell;T_\ell)\le -2\sigma\Psi(\mathbf{y}^\ell;T_\ell)=-\sigma\langle F(\mathbf{Y}^\ell;T_\ell),\mathbf{W}^\ell\mathbf{d}^\ell\rangle,$$

which implies that condition (4.4) holds with a unit step length. Finally, the same reasoning to show (4.11) enables us to obtain $\|\mathbf{Y}^{\ell+1}-\mathbf{Y}^*\|=O(\|\mathbf{Y}^\ell-\mathbf{Y}^*\|^2)$, a quadratic convergence rate. □

5 Numerical Experiments

In this section, we conduct extensive numerical experiments on sparse recovery problems, sparse canonical correlation analysis, and sparse portfolio selection to showcase the performance of SNSQP. All experiments are implemented in MATLAB R2021a, running on a laptop computer of 16GB memory and Inter(R) Core(TM) i7 2.4Ghz CPU. In our numerical experiments, hyperparameters for SNSQP are set as follows: $\rho = 0.5$, $\sigma = 0.5$, $\nu_i^0 = 0, \mu_i^0 = 0.01$, and $\lambda_i^0 = 0.01$ for each i. The choices of τ and initial point $\mathbf{x}^0$ vary depending on the specific problems.

5.1 Recovery problems

We first demonstrate the performance of SNSQP for solving a recovery problem using synthetic datasets. The example is described as follows.

Example 5.1. *Let* $f_0(\mathbf{x}) = \frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{d}\|^2$ *and* $\mathbf{Q}_i = \mathbf{P}_i^\top \mathbf{P}_i + 0.01\mathbf{I}$ *for* $i \in [k]$, *where* $\mathbf{D} \in \mathbb{R}^{d\times n}$ *and* $\mathbf{P}_i \in \mathbb{R}^{n\times n}$. *Set* $\mathbb{X}$ *is chosen as* $\mathbb{R}^n$, $[-2,2]^n$, *or* $[0,\infty)^n$. *Let a ground truth* $\mathbf{x}^* \in \mathbb{S}$ *be given with* s *entries randomly selected to be* 1. *Then* $\mathbf{d} = \mathbf{D}\mathbf{x}^*$ *and each entry of* $\mathbf{D}$, $\mathbf{P}_i, \mathbf{q}_i$, *and* $\mathbf{A}$ *is generated from standard normal distribution* $\mathcal{N}(0,1)$. *To generate* c_i *and* $\mathbf{b}$, *we randomly select* $T_1 \subseteq [k]$ *with* $|T_1| = \lceil k/2 \rceil$ *and* $T_2 \subseteq [m]$ *with* $|T_2| = \lceil m/2 \rceil$, *where* $\lceil t \rceil$ *is the ceiling of* t. *Finally, define*

$$c_i = \begin{cases} -\frac{1}{2}\langle \mathbf{x}^*, \mathbf{Q}_i\mathbf{x}^*\rangle - \langle \mathbf{q}_i, \mathbf{x}^*\rangle - \zeta_i, & \text{if } i \in T_1, \\ -\frac{1}{2}\langle \mathbf{x}^*, \mathbf{Q}_i\mathbf{x}^*\rangle - \langle \mathbf{q}_i, \mathbf{x}^*\rangle, & \text{if } i \notin T_1, \end{cases} \qquad b_i = \begin{cases} \langle \mathbf{a}_i, \mathbf{x}^*\rangle - \xi_i, & \text{if } i \in T_2, \\ \langle \mathbf{a}_i, \mathbf{x}^*\rangle, & \text{if } i \notin T_2, \end{cases}$$

where both ζ_i *and* ξ_i *are uniform random variables from* $[0,1]$.

To solve this example, we set $\tau = 3$ and initialize $\mathbf{x}^0$ for SNSQP as follows: Randomly select s indices to form Γ and set $x_i^0 = 0.1$ if $i \in \Gamma$ and $x_i^0 = 0$ otherwise. The maximum number of iterations and the tolerance are set as 10^4 and $\varepsilon = 10^{-8}$. We report the relative error (Relerr $:= \|\mathbf{x}-\mathbf{x}^*\|/\|\mathbf{x}^*\|$), the objective function value (Fval), and the computational time in seconds to evaluate the performance of different benchmark algorithms, where $\mathbf{x}$ is the solution generated by one algorithm.

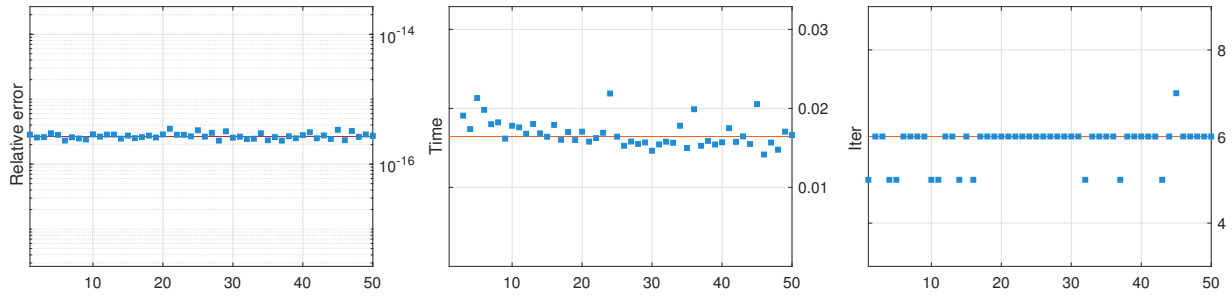


Figure 1: Effects of initial points for Example 5.1 ($X = \mathbb{R}^n$).

a) Effect of initial points. As stated in Theorem 4.1, SNSQP is a local method. To assess the effect of initial points $\mathbf{x}^0$, for Example 5.1 with $(n, d, k, m, s) = (1000, 1000, 1, 1, 10)$, we run SNSQP with 50 random initial points and visualize the results in Figure 1, where the x-axis represents the initial points. The stability of metrics such as relative errors, time, and iterations, suggests that it is insensitive to the choices of the initial points for solving this example.

Table 3: Numerical comparison with QCQP solvers for Example 5.1.

s	n	Relerr			Fval			Time(s)		
		SNSQP	CPLEX	CVX	SNSQP	CPLEX	CVX	SNSQP	CPLEX	CVX
					$\mathbb{X} = \mathbb{R}^n$					
	1000	7.97e-16	6.57e-02	1.97e-02	8.32e-17	6.30e-08	7.34e-09	0.010	1.893	4.574
	2000	3.07e-16	6.86e-02	2.61e-02	7.94e-16	1.15e-08	1.06e-08	0.017	36.51	52.96
$0.01n$	3000	3.75e-16	4.15e-02	1.57e-02	8.87e-16	7.63e-08	4.02e-08	0.047	442.84	200.54
	4000	2.25e-16	3.22e-02	1.08e-02	6.46e-16	5.21e-07	2.74e-08	0.079	1522.3	602.20
	5000	5.59e-16	–	1.41e-02	7.20e-16	–	1.99e-08	0.127	–	1249.6
	1000	4.31e-16	1.76e-02	1.46e-02	6.66e-16	5.50e-08	1.32e-08	0.019	2.046	4.695
	2000	4.45e-16	4.48e-02	2.44e-02	8.88e-16	1.25e-07	4.25e-08	0.036	35.502	51.72
$0.05n$	3000	5.12e-16	1.12e-02	2.36e-02	8.02e-16	5.21e-07	2.88e-08	0.049	469.40	215.55
	4000	5.00e-16	5.33e-02	1.52e-02	6.55e-16	7.23e-07	2.46e-08	0.198	1421.1	705.27
	5000	6.21e-16	–	1.52e-02	7.79e-16	–	2.33e-08	0.402	–	1201.1
					$\mathbb{X} = [-2, 2]^n$					
	1000	1.98e-16	8.67e-02	2.34e-04	8.21e-16	6.81e-06	3.85e-12	0.012	2.490	5.495
	2000	2.22e-16	1.69e-01	4.21e-04	7.15e-15	5.25e-05	1.71e-09	0.020	33.51	64.48
$0.01n$	3000	5.16e-16	1.36e-01	7.24e-04	8.46e-16	4.38e-06	8.66e-12	0.031	279.17	191.43
	4000	3.27e-16	1.02e-01	1.06e-03	8.57e-16	7.99e-06	1.03e-11	0.079	1554.5	614.24
	5000	4.47e-16	–	1.11e-03	6.35e-16	–	7.54e-12	0.131	–	1251.2
	1000	4.29e-16	8.03e-02	3.04e-04	2.22e-16	3.87e-06	5.16e-12	0.017	3.048	6.199
	2000	3.56e-15	1.43e-01	6.57e-04	6.33e-15	3.14e-05	7.69e-10	0.051	34.45	66.87
$0.05n$	3000	5.33e-16	1.13e-01	3.90e-03	7.02e-16	5.79e-06	5.03e-12	0.085	302.34	225.15
	4000	6.25e-16	1.21e-01	1.53e-03	6.73e-16	1.63e-05	4.12e-12	0.214	1513.4	722.78
	5000	4.11e-16	–	2.01e-03	6.25e-16	–	3.86e-12	0.464	–	1267.3
					$\mathbb{X} = [0, \infty)^n$					
	1000	5.27e-16	3.80e-02	2.00e-04	1.11e-16	2.83e-06	4.31e-09	0.014	2.347	8.527
	2000	1.31e-16	2.23e-02	1.69e-04	3.65e-16	1.03e-06	4.48e-09	0.023	34.68	54.51
$0.01n$	3000	3.34e-16	2.11e-02	2.06e-04	4.56e-16	6.81e-06	5.76e-09	0.051	217.73	258.91
	4000	3.92e-16	1.12e-02	1.86e-04	5.69e-16	7.63e-06	1.36e-09	0.091	1460.0	751.15
	5000	1.06e-15	–	1.79e-03	1.21e-16	–	5.90e-09	0.301	–	1631.5
	1000	4.48e-16	4.40e-02	1.81e-04	4.44e-16	7.40e-06	8.18e-09	0.023	3.464	8.414
	2000	5.53e-16	6.70e-02	1.49e-02	1.35e-16	6.81e-06	2.56e-09	0.063	32.85	55.28
$0.05n$	3000	5.79e-16	1.19e-02	1.59e-02	5.32e-16	2.36e-06	4.55e-09	0.248	239.25	275.00
	4000	6.75e-16	1.07e-02	1.38e-04	1.25e-15	7.69e-06	1.86e-09	1.201	1500.2	775.68
	5000	7.82e-16	–	1.82e-04	1.86e-16	–	1.33e-09	1.516	–	1816.8

b) Comparison with QCQP solvers. We compare SNSQP with two quadratically constrained quadratic programming solvers: CPLEX and CVX. For each $n \in \{1000, 2000, \ldots, 5000\}$, we set $d = n + 5$, $k = m = 0.001n$, $s = 0.01n$ or $0.05n$. Then, we run 20 independent trials for each (n, d, k, m, s). The median of the 20 trials is presented in Table 3, where $-$ indicates cases where the computational time exceeds one hour. The results show that SNSQP consistently achieves the lowest Relerr and Fval across all cases, with an accuracy on the order of at least 10^{-15}, indicating that it finds the optimal solution. Regarding computation time, SNSQP is the fastest. As n increases, the time required by CPLEX and CVX grows dramatically. For instance, when $\mathbb{X} = \mathbb{R}^n$, $n = 5000$, and $s = 0.01n$, CPLEX exceeds one hour, CVX requires more than 1200 seconds, whereas SNSQP consumes 0.127 seconds, a significant improvement.

c) Comparison with an MIP solver. We note that (SQCQP) is equivalent to the following MIP,

$$\min_{\mathbf{x},\mathbf{w}} f_0(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{x} \in \mathbb{F} \cap \mathbb{X}, \ \langle \mathbf{1}, \mathbf{w} \rangle \le s, \ |x_i| \le M w_i, \ w_i \in \{0, 1\}, \ i \in [n], \tag{5.1}$$

where $\mathbf{1} := (1, 1, \ldots, 1)^\top$ and $M > 0$ is a large enough constant, which is set as $M = 10$ in this example. The above problem can be directly solved by any MIP solver, such as GUROBI. Therefore, we compare it with our proposed algorithm. For each $n \in \{100, 200, \ldots, 1000\}$, other dimensions (d, k, m, s) are set similarly to the comparison with QCQP solvers. The median results of 20 independent trials are presented in Table 4. It is shown that SNSQP achieves the lowest Relerr, Fval, and the shortest computational time once again.

Table 4: Numerical comparison with an MIP solver for Example 5.1 with $\mathbb{X} = \mathbb{R}^n$.

s	n	Relerr		Fval		Time(s)	
		SNSQP	GUROBI	SNSQP	GUROBI	SNSQP	GUROBI
$0.01n$	100	8.63e-13	1.30e-03	2.08e-18	8.18e-08	0.002	0.617
	300	1.76e-14	2.73e-04	1.38e-17	2.93e-08	0.003	3.217
	500	6.64e-15	3.37e-04	5.55e-17	1.92e-08	0.006	8.490
	700	3.36e-16	6.45e-04	5.59e-18	1.89e-08	0.009	22.66
	1000	2.67e-16	9.63e-04	1.12e-17	2.17e-07	0.010	310.6
$0.05n$	100	2.78e-14	3.00e-04	5.55e-18	2.84e-08	0.003	0.886
	300	6.25e-14	2.86e-04	4.44e-17	7.91e-08	0.005	24.12
	500	2.84e-14	1.50e-04	3.33e-17	1.05e-08	0.006	699.8
	700	5.45e-16	1.30e-03	3.17e-17	1.87e-07	0.012	334.4
	1000	5.60e-16	–	4.10e-16	–	0.012	–

d) Effect of higher dimensions. We evaluate the performance of SNSQP in solving Example 5.1 in higher-dimensional settings. A comparison with other solvers is omitted, as Tables 3 and 4 indicate that they exhibit significantly higher computational costs for such cases. Table 5 presents the median results over 10 independent runs, where $d = n + 5$ and $k = m = 5$. The results demonstrate that the Relerr and Fval obtained by SNSQP remain stable, consistently ranging in magnitude from 10^{-13} to 10^{-15}, while maintaining efficient computational performance. Specifically, for $n = 30,000$, the computational time is 3.251 seconds and 33.59 seconds for $s = \lceil 0.01n \rceil$ and $s = \lceil 0.05n \rceil$.

Table 5: Results by SNSQP for Example 5.1 with $\mathbb{X} = [-2, 2]^n$ in higher dimensions.

n	$s = 0.01n$			$s = 0.05n$		
	Relerr	Fval	Time(s)	Relerr	Fval	Time(s)
10000	5.47e-15	4.44e-15	0.284	1.89e-14	7.10e-15	3.392
15000	1.03e-14	4.46e-15	0.895	1.42e-14	1.52e-15	4.581
20000	2.25e-14	7.10e-15	1.153	1.57e-15	1.05e-15	10.54
25000	1.82e-13	2.84e-14	2.741	2.13e-14	5.46e-15	15.82
30000	8.03e-14	7.10e-15	3.251	5.68e-14	1.99e-15	33.59

5.2 Sparse canonical correlation analysis

The sparse canonical correlation analysis (SCCA) problem takes the form of:

$$\begin{aligned} \max\ & \langle \mathbf{w}^x, \mathbf{\Sigma}^{xy}\mathbf{w}^y\rangle \\ \text{s.t. } & \langle \mathbf{w}^x, \mathbf{\Sigma}^{xx}\mathbf{w}^x\rangle \le 1,\ \|\mathbf{w}^x\|_0 \le s_x, \\ & \langle \mathbf{w}^y, \mathbf{\Sigma}^{yy}\mathbf{w}^y\rangle \le 1,\ \|\mathbf{w}^y\|_0 \le s_y, \end{aligned} \tag{SCCA}$$

where $\mathbf{\Sigma}^{xy} \neq 0$ is the cross-covariance matrix between two given data matrices $\mathbf{X}$ and $\mathbf{Y}$, namely $\mathbf{\Sigma}^{xy} = \mathbf{X}\mathbf{Y}^\top$, and $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}^{yy}$ represent the covariance matrices for $\mathbf{X}$ and $\mathbf{Y}$, namely $\mathbf{\Sigma}^{xx} = \mathbf{X}\mathbf{X}^\top$ and $\mathbf{\Sigma}^{yy} = \mathbf{Y}\mathbf{Y}^\top$. By letting $s = s_x + s_y$ and

$$\mathbf{Q}_0 = \begin{bmatrix} 0 & \mathbf{\Sigma}^{xy} \\ \mathbf{\Sigma}^{xy\top} & 0 \end{bmatrix}, \quad \mathbf{Q}_1 = \begin{bmatrix} \mathbf{\Sigma}^{xx} & 0 \\ 0 & \mathbf{\Sigma}^{yy} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \mathbf{w}^x \\ \mathbf{w}^y \end{bmatrix},$$

problem (5.2) can be relaxed in the form,

$$\max\ \langle \mathbf{x}, \mathbf{Q}_0\mathbf{x}\rangle, \quad \text{s.t.}\ \ \langle \mathbf{x}, \mathbf{Q}_1\mathbf{x}\rangle \le 2,\ \|\mathbf{x}\|_0 \le s. \tag{5.2}$$

This is a relaxation of problem (SCCA) because its feasible region includes the feasible region of (SCCA). However, the following result shows that the optimal solutions to them have a close relationship. Therefore, we employ SNSQP to solve (5.2), a special case of (SQCQP).

Proposition 5.1. *Any optimal solution* $(\mathbf{x}^x; \mathbf{x}^y)$ *of (5.2) satisfies* $\|\mathbf{x}^x\|_0 > 0$ *and* $\|\mathbf{x}^y\|_0 > 0$.

Proof. Since $\mathbf{\Sigma}^{xy} \neq 0$, there is $\Sigma^{xy}_{i_0 j_0} \neq 0$. Consider a point $\mathbf{x} := (\mathbf{u}; \mathbf{v})$ such that

$$u_i = \begin{cases} \frac{1}{\sqrt{\Sigma^{xx}_{i_0 i_0}}} & \text{if } i = i_0,\ \Sigma^{xx}_{i_0 i_0} \neq 0, \\ 1 & \text{if } i = i_0,\ \Sigma^{xx}_{i_0 i_0} = 0, \\ 0 & \text{if } i \neq i_0, \end{cases} \qquad v_j = \begin{cases} \frac{\operatorname{sign}\left(\Sigma^{xy}_{i_0 j_0}\right)}{\sqrt{\Sigma^{yy}_{j_0 j_0}}} & \text{if } j = j_0,\ \Sigma^{yy}_{j_0 j_0} \neq 0, \\ 1 & \text{if } j = j_0,\ \Sigma^{yy}_{j_0 j_0} = 0, \\ 0 & \text{if } j \neq j_0, \end{cases}$$

where $\operatorname{sign}(t)$ is the sign of t. Therefore, $\mathbf{x} = (\mathbf{u}; \mathbf{v})$ is feasible to problem (5.2), and meanwhile it satisfies that $\langle \mathbf{x}, \mathbf{Q}_0\mathbf{x}\rangle > 0$. Note that any point $\mathbf{x}' = (\mathbf{u}'; \mathbf{v}')$ satisfying $\|\mathbf{u}'\|_0 = 0$ or $\|\mathbf{v}'\|_0 = 0$ leads to $\langle \mathbf{x}', \mathbf{Q}_0\mathbf{x}'\rangle = 0$. Hence, $\mathbf{x}'$ is not the optimal solution, showing the desired result. □

5.2.1 Benchmark methods

Many algorithms have been proposed to address SCCA problems. The majority of these methods are based on relaxations of the ℓ_0 norm, while a subset employs greedy strategies. We select three representative relaxation-based approaches: SGEM [49], SCCA [22], and SCCAPD [33], in addition to two greedy algorithms: SpanCCA [2] and SWCCA [43]. It is important to note that the greedy methods are applicable only in scenarios where the covariance matrices are identity matrices. Consequently, in the subsequent numerical experiments, we specifically set $\mathbf{\Sigma}^{xx} = \mathbf{I}$ and $\mathbf{\Sigma}^{yy} = \mathbf{I}$ when evaluating SpanCCA and SWCCA.

The hyperparameters for all algorithms are configured as follows. For SNSQP, initial point $\mathbf{x}^0$ is generated using MATLAB built-in function `canoncorr`, while τ is selected via a grid search over $\{0.001, 0.002, \ldots, 0.01\}$. For SGEM, the same initial point as in SNSQP is used, with parameters set as $\epsilon = 10^{-6}$ and $\tau = \max\{0, -\lambda_{\min}(\mathbf{Q}_0)\}$, where $\lambda_{\min}(\mathbf{Q}_0)$ represents the smallest eigenvalue of $\mathbf{Q}_0$. For SCCA, maximum number $\ell_{\max}$ of iterations is fixed at 5000, and δ is selected from a predefined grid $\{0.1, 0.2, \ldots, 1\}$. For SCCAPD, we use its default settings. For SpanCCA, we set $\ell_{\max} = 5000$ and $r = 5$. For SWCCA, the initial point is the same as in SNSQP, and the sparsity level of $\mathbf{w}$ is set to be $0.6N$. It terminates when either $\ell > 5000$ or $\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 < 10^{-6}$ is satisfied. To evaluate the performance, we report the correlation, sparsity levels ρ_x and ρ_y of $\mathbf{w}^x$ and $\mathbf{w}^y$, violation VOC_x and VOC_y of the unit variance constraint, and the computational time, where

$$\begin{aligned}
&\text{Correlation} := \frac{\langle \mathbf{w}^x, \Sigma^{xy}\mathbf{w}^y\rangle}{\sqrt{\langle \mathbf{w}^x, \Sigma^{xx}\mathbf{w}^x\rangle\langle \mathbf{w}^y, \Sigma^{yy}\mathbf{w}^y\rangle}},\\
&\rho_x := \frac{n_x - \|\mathbf{w}^x\|_0}{n_x}, \qquad \mathrm{VOC}_x := \|\langle \mathbf{w}^x, \mathbf{\Sigma}^{xx}\mathbf{w}^x\rangle - 1\|,\\
&\rho_y := \frac{n_y - \|\mathbf{w}^y\|_0}{n_y}, \qquad \mathrm{VOC}_y := \|\langle \mathbf{w}^y, \mathbf{\Sigma}^{yy}\mathbf{w}^y\rangle - 1\|.
\end{aligned}$$

5.2.2 Testing examples

Example 5.2 (**Synthetic data [22]**). *Let* $\mathbf{X} \in \mathbb{R}^{n_x \times N}$ *and* $\mathbf{Y} \in \mathbb{R}^{n_y \times N}$ *be generated by*

$$\mathbf{X} = ((\mathbf{1}; -\mathbf{1}; 0) + \boldsymbol{\epsilon})\,\mathbf{u}^\top, \quad \mathbf{Y} = ((0; \mathbf{1}; -\mathbf{1}) + \boldsymbol{\varepsilon})\,\mathbf{u}^\top, \tag{5.3}$$

where $\mathbf{1} \in \mathbb{R}^{n_x/8}$, $\boldsymbol{\epsilon} \in \mathbb{R}^{n_x}$ *and* $\boldsymbol{\varepsilon} \in \mathbb{R}^{n_y}$ *are two noise vectors with* $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$ *and* $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$, *and* $\mathbf{u} \in \mathbb{R}^N$ *is a random vector with* $u_i \sim \mathcal{N}(0, 1)$.

Example 5.3 (**Real data**). *Four real datasets are selected to generate* $\mathbf{X}$ *and* $\mathbf{Y}$. *They are* `SRBCT`, `lymphoma`, `breast cancer`, *and* `glioma`[1] *[14, 15] with details summarized in in Table 6. The sample-wise normalization is conducted on them so that each sample has mean zero and variance one. In addition,* `SRBCT` *and* `lymphoma` *only contain one set of samples, so we manually divide the variables into two equal parts to obtain* $\mathbf{X}$ *and* $\mathbf{Y}$.

5.2.3 Numerical comparison

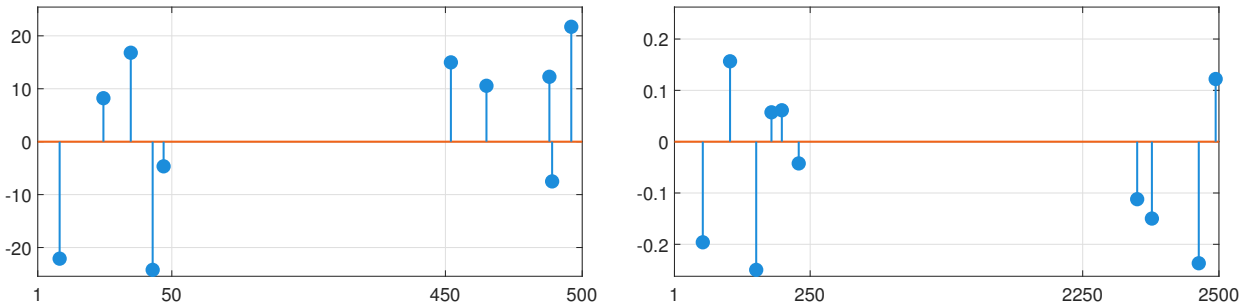
From the construction of $\mathbf{X}$ and $\mathbf{Y}$ in Example 5.2, it is evident that the first $n_x/4$ variables of $\mathbf{X}$ exhibit correlation with the last $n_x/4$ variables of $\mathbf{Y}$. An effective sparse CCA algorithm should be

[1] Available at www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2223.

Table 6: Descriptions of four real datasets.

Datasets	Source	N	n	n_x	n_y
`SRBCT`	R package: plsgenomics [13]	82	2308	1154	1154
`lymphoma`	R package: KODAMA [17]	61	4026	2013	2013
`breast cancer`	R package: PMA [57]	89	21821	2149	19672
`glioma`	NCBI: Gene Expression Omnibus	55	62127	22962	39165

capable of computing weight vectors $\mathbf{w}_x$ and $\mathbf{w}_y$ that accurately identify these correlated variables. Specifically, the nonzero elements of $\mathbf{w}_x$ should be restricted to the first $n_x/4$ components, while those of $\mathbf{w}_y$ should be confined to the last $n_x/4$ components. As illustrated in Figure 2, the SNSQP algorithm successfully achieves this objective.



(a) $n_x = 200, n_y = 300$ (b) $n_x = 1000, n_y = 1500$

Figure 2: Solutions obtained by SNSQP for Example 5.2 with $s = 10$.

Table 7: Comparison with relaxation methods for Example 5.2.

Algs.	Para.	Correlation	ρ_x	ρ_y	VOC$_x$	VOC$_y$	Time(s)
			$n_x = 200$, $n_y = 300$				
SNSQP	$s = 5$	1.0000	99.0%	99.0%	6.43e-15	5.99e-15	0.006
	$s = 10$	1.0000	99.0%	97.3%	1.34e-10	1.78e-10	0.009
	$\rho = 0.01$	0.9980	99.5%	99.7%	1.94e-02	1.94e-02	18.51
	$\rho = 0.005$	1.0000	76.0%	84.3%	1.69e-04	1.69e-04	6.388
SCCA	$\mu = 5$	1.0000	99.0%	99.0%	2.35e-05	8.64e-04	0.120
	$\mu = 10$	1.0000	99.5%	99.3%	4.65e-05	9.17e-04	0.207
SCCAPD		1.0000	75.0%	61.6%	6.43e-16	5.99e-16	0.024
			$n_x = 1000$, $n_y = 1500$				
SNSQP	$s = 5$	1.0000	99.8%	99.8%	1.61e-15	1.62e-15	0.005
	$s = 10$	1.0000	99.7%	99.5%	1.64e-10	1.64e-10	0.008
SGEM	$\rho = 0.01$	0.9990	96.5%	97.5%	2.86e-05	2.86e-05	166.52
	$\rho = 0.005$	1.0000	95.6%	96.8%	6.71e-05	6.71e-05	138.08
SCCA	$\mu = 5$	1.0000	99.5%	99.2%	1.01e-10	9.90e-04	0.419
	$\mu = 10$	1.0000	99.8%	99.3%	5.09e-13	9.90e-04	0.700
SCCAPD		1.0000	92.9%	64.8%	2.22e-16	5.99e-16	0.035

A comprehensive numerical comparison of SNSQP with three relaxation algorithms for solving Example 5.2 and Example 5.3 is presented in Table 7 and Table 8. One can observe that SNSQP outperforms the others in terms of overall performance, achieving the highest correlation and sparsity levels while maintaining the fastest computational efficiency.

Table 8: Comparison with relaxation methods for Example 5.3.

Algs.	Para.	Correlation	ρ_x	ρ_y	VOC_x	VOC_y	Time(s)
			`SRBCT`				
SNSQP	$s = 40$	0.9968	97.83%	98.70%	1.26e-14	1.26e-14	0.014
	$s = 80$	1.0000	96.62%	96.45%	4.68e-11	4.68e-11	0.035
SGEM	$\rho = 0.01$	0.6317	1.822%	1.215%	3.10e-03	3.10e-03	1587.2
	$\rho = 0.005$	0.5879	0.435%	0.782%	2.61e-03	2.61e-03	1177.9
SCCA	$\mu = 5$	1.0000	86.40%	86.74%	8.11e-04	3.51e-04	0.842
	$\mu = 10$	0.9887	91.16%	90.81%	7.23e-02	6.45e-03	0.826
SCCAPD		0.9316	98.09%	66.20%	3.32e-16	2.22e-16	0.063
			`lymphoma`				
SNSQP	$s = 20$	0.9904	99.35%	99.65%	5.57e-11	5.57e-11	0.014
	$s = 50$	0.9999	98.56%	98.96%	6.94e-11	6.94e-11	0.034
SCCA	$\mu = 5$	0.9014	98.61%	98.56%	1.82e-02	2.80e-02	2.759
	$\mu = 10$	0.4016	99.65%	99.60%	1.32e-02	2.80e-02	3.865
SCCAPD		0.9210	99.06%	56.38%	4.44e-16	4.44e-16	0.162
			`breast cancer`				
SNSQP	$s = 40$	0.9941	99.58%	99.84%	9.28e-09	9.28e-09	0.110
	$s = 80$	1.0000	99.44%	99.76%	3.87e-09	3.87e-09	0.160
SCCA	$\mu = 0.5$	0.9835	95.86%	99.72%	1.34e-03	1.55e-02	36.27
	$\mu = 1$	0.7723	96.32%	99.94%	8.04e-03	3.95e-01	36.15
SCCAPD		0.9193	99.67%	99.74%	3.33e-16	5.55e-16	0.164
			`glioma`				
SNSQP	$s = 50$	1.0000	99.92%	99.92%	1.07e-11	1.07e-11	0.476
	$s = 100$	1.0000	99.80%	99.86%	3.31e-08	3.31e-08	0.341
SCCA	$\mu = 0.2$	0.9955	99.31%	99.56%	1.48e-02	9.83e-03	355.49
	$\mu = 0.5$	0.8825	99.53%	99.78%	9.29e-02	2.64e-01	353.38
SCCAPD		0.9223	99.98%	99.84%	2.22e-16	1.11e-16	3.117

We further compare SNSQP with two greedy methods, SpanCCA and SWCCA, to examine the effect of the sparsity level s. As shown in Figure 3, the correlation computed by all three methods exhibits an increasing trend as s grows. SNSQP consistently achieves the highest correlation level. It is noteworthy that SWCCA is very fast for this example, as it solves the problem with identity covariance matrices, i.e., $\mathbf{\Sigma}xx = \mathbf{I}$ and $\mathbf{\Sigma}yy = \mathbf{I}$, that enable closed-form solutions to subproblems. However, due to this simplification, SWCCA fails to yield desirable correlation.

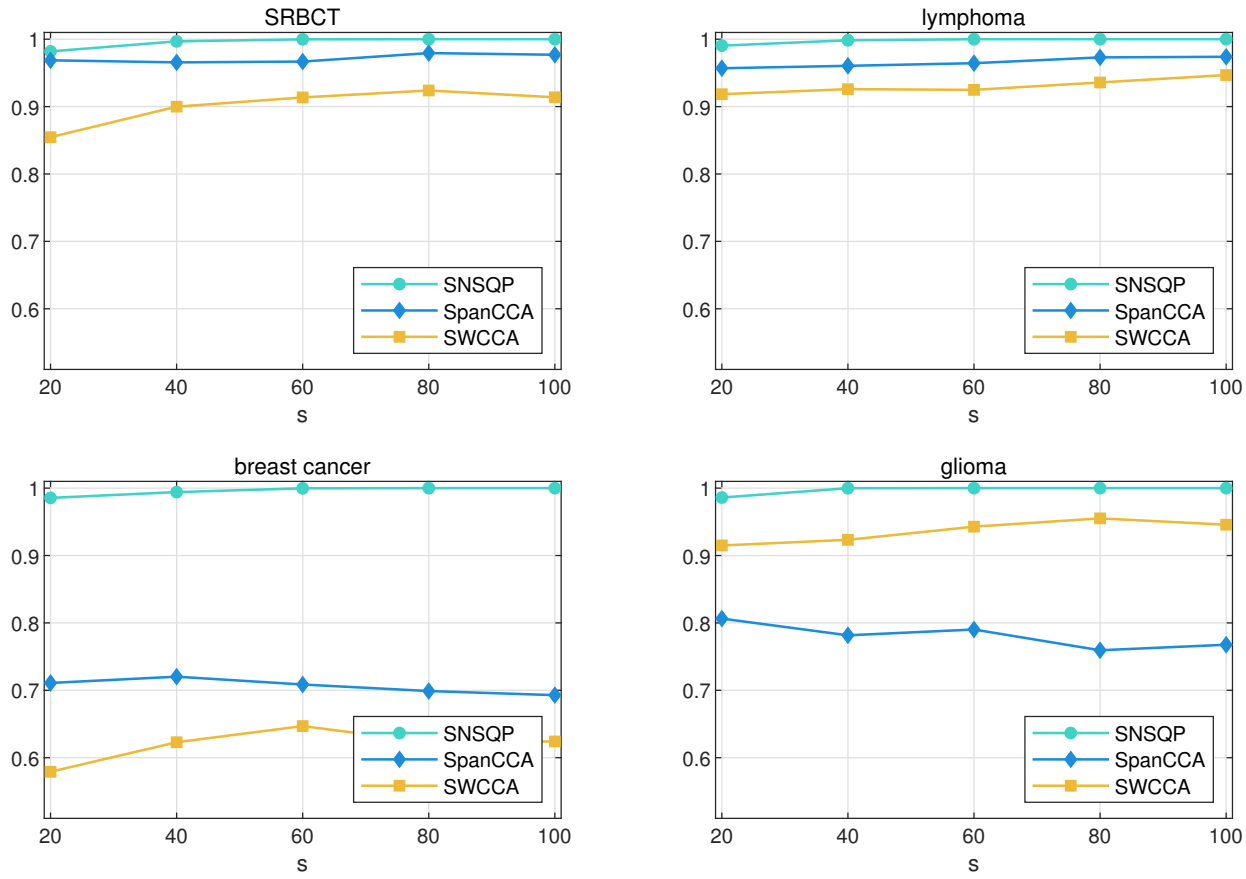


Figure 3: Effect of sparsity level s for Example 5.2.

5.3 Sparse portfolio selection problem

The sparse portfolio selection (SPS) problem proposed in [24] is formulated as follows,

$$\begin{aligned} \min\ & \langle \mathbf{x}, (\mathbf{Q}+\mathbf{Q}_1)\mathbf{x}\rangle \\ \text{s.t.}\ & \langle \mathbf{x}, \mathbf{Q}_1\mathbf{x}\rangle \le \sigma_0,\ \langle \mathbf{x}, \mathbf{a_1}\rangle \ge r_0,\ \langle \mathbf{x}, \mathbf{1}\rangle = 1, \\ & \|\mathbf{x}\|_0 \le s,\ \mathbf{x} \in \{0\} \cup [\mathbf{a}, \mathbf{b}]\,, \end{aligned} \tag{SPS}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is a symmetric positive semi-definite matrix, $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a non-negative diagonal matrix, $\mathbf{a}_1 \in \mathbb{R}^n$ is the return vector, σ_0 is the prescribed nonsystematic risk level and r_0 is the prescribed weekly return level, and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are the lower and upper bounds of $\mathbf{x}$. In the objection function, first part $\langle \mathbf{x}, \mathbf{Q}\mathbf{x}\rangle$ is called *systematic risk* and second part $\langle \mathbf{x}, \mathbf{Q}_1\mathbf{x}\rangle$ is called *nonsystematic/specific risk*. In the sequel, we set set $(\sigma_0, r_0) = (0.001, 0.002)$, $\mathbf{a} = \mathbf{0}$, and $\mathbf{b} = (0.3, 0.3, \ldots, 0.3)^\top$. We note that the formulation in (SPS) includes an equality constraint. Following the approach in LNA [60], we handle this constraint by directly incorporating it into the stationary equations for Newton's method iterations, rather than using the NCP function.

5.3.1 Benchmark methods

We compare SNSQP with SALM [4] and GUROBI. In this comparison, GUROBI is used to solve a reformulation of (SPS), where the sparsity constraint is replaced by the ones in (5.1). Its maximal runtime is set to one hour. For SALM, we configure the step size as $\omega = 0.3$ for Example 5.4 and $\omega = 1$ for Example 5.5. The initial Lagrange multiplier and penalty parameter are set to

$\boldsymbol{\lambda}^0 = 0$ and $\rho = 1$. For SNSQP, we set $\tau = 1$. Both algorithms terminate when either $\ell > 1000$ or the tolerance condition $|\cdot| \leq 10^{-6}$ is met. Additionally, we use CVX to solve (SPS) without the sparsity constraint to provide an initial point for all methods. Since GUROBI enables high-quality approximation of the global solution, we use the relative error defined by Relerr $= |f - f_{\text{MIP}}|/f_{\text{MIP}}$ to measure the accuracy of SNSQP and SALM, where f is their objective function values and f_{MIP} is the one obtained by GUROBI.

5.3.2 Testing examples

Example 5.4 (**Portfolio datasets**). *The dataset are drawn from the Standard and Poor's 500 (S&P 500) and Russell 2000 indices. Specifically, we utilize weekly return data for 468 stocks from the S&P 500 and 873 stocks from the Russell 2000 over the period from 2015 to 2020. The matrices* $\mathbf{Q}$ *and* $\mathbf{Q}_1$ *are derived from factor models constructed for these datasets. In addition, we randomly select* n *stocks from the entire stock pool.*

Table 9: Comparison with two methods for Example 5.4

n	sn	Relerr		Fval			Time(s)		
		SNSQP	SALM	SNSQP	SALM	GUROBI	SNSQP	SALM	GUROBI
					S&P500				
50	5	0.032	0.048	2.88e-04	2.90e-04	2.79e-04	0.011	0.069	0.336
	10	0.004	0.008	2.49e-04	2.51e-04	2.48e-04	0.030	0.190	0.455
100	5	0.085	0.114	2.89e-04	2.97e-04	2.66e-04	0.013	0.527	1.092
	10	0.013	0.024	2.45e-04	2.48e-04	2.42e-04	0.051	0.996	2.526
200	5	0.127	0.135	2.52e-04	2.54e-04	2.24e-04	0.072	1.156	5.279
	10	0.003	0.013	1.96e-04	1.98e-04	1.95e-04	0.170	1.789	12.42
300	5	0.044	0.086	2.38e-04	2.48e-04	2.28e-04	0.180	2.880	29.60
	10	0.060	0.091	1.90e-04	1.95e-04	1.79e-04	0.199	3.724	126.88
400	5	0.063	0.075	2.52e-04	2.55e-04	2.37e-04	0.070	2.696	116.16
	10	0.037	0.092	1.86e-04	1.96e-04	1.79e-04	0.091	5.069	1379.6
					Russell2000				
50	5	0.016	0.027	5.22e-04	5.28e-04	5.16e-04	0.020	0.081	0.233
	10	0.008	0.150	4.13e-04	4.22e-04	4.09e-04	0.014	0.043	0.320
200	5	0.016	0.039	2.70e-04	2.77e-04	2.66e-04	0.061	1.421	11.66
	10	0.002	0.010	2.71e-04	2.73e-04	2.70e-04	0.035	0.836	7.501
400	5	0.006	0.035	2.09e-04	2.16e-04	2.08e-04	0.031	4.023	804.93
	10	0.019	0.034	2.08e-04	2.12e-04	2.04e-04	0.088	5.688	830.44
600	5	0.002	0.016	2.08e-04	2.10e-04	2.07e-04	0.038	8.602	510.25
	10	0.012	0.045	2.43e-04	2.50e-04	2.40e-04	0.041	9.542	3593.1
800	5	0.000	0.014	2.04e-04	2.07e-04	2.04e-04	0.052	15.82	2951.6
	10	0.007	0.038	1.61e-04	1.66e-04	1.59e-04	0.061	12.08	3600.0

Example 5.5 (**Synthetic data**). *Slightly higher-dimensional datasets are generated as follows. Let* $\mathbf{Q}_0 = \mathbf{D}^\top \mathbf{D}$ *with* $\mathbf{D} \in \mathbb{R}^{n/4 \times n}$. *Each entry of* $\mathbf{D}$ *and each diagonal entry of* $\mathbf{Q}_1$ *are independently sampled from a uniform distribution over* $[0, 0.01]$, *while each entry of* $\mathbf{a}_1$ *is drawn from a normal distribution* $\mathcal{N}(0, 0.5^2)$. *We vary* n *over set* $\{1000, 1500, \ldots, 3000\}$.

5.3.3 Numerical comparison

As shown in Table 9, SNSQP consistently achieves lower Relerr and Fval than SALM across all cases for both datasets, indicating superior solution quality. Additionally, SNSQP demonstrates a significant speed advantage over its counterparts in every scenario, particularly in high-dimensional cases. For instance, on the Russell2000 dataset with $(n, S) = (800, 5)$, the computational time of SNSQP, SALM, and GUROBI is 0.052, 15.82, and 2951.6 seconds, respectively. Finally, we examine the numerical comparison for slightly higher-dimensional scenarios based on Example 5.5. Since GUROBI requires a significantly longer time to solve the problem, it is excluded from the comparison. The results in Table 10 show that SNSQP not only achieves lower Fval but also runs much faster than SALM across all test instances.

Table 10: Comparison for Example 5.5 in higher dimensions.

n	$s = 5$				$s = 10$			
	Fval		Time(s)		Fval		Time(s)	
	SNSQP	SALM	SNSQP	SALM	SNSQP	SALM	SNSQP	SALM
1000	5.83e-03	5.84e-03	0.012	13.18	5.69e-03	5.71e-03	0.029	8.280
1500	8.43e-03	8.46e-03	0.042	46.31	8.47e-03	8.49e-03	0.047	29.92
2000	1.16e-02	1.17e-02	0.030	104.67	1.14e-02	1.15e-02	0.032	91.58
2500	1.50e-02	1.51e-02	0.032	219.62	1.46e-02	1.48e-02	0.035	180.06
3000	1.80e-02	1.82e-02	0.039	443.45	1.75e-02	1.76e-02	0.042	406.82

6 Conclusion

It is known that SQCQP is a computationally challenging problem, particularly in large-scale or high-dimensional settings. In contrast to existing methods, which primarily focus on solving its mixed-integer programming reformulations or relaxations, we introduce a novel paradigm by designing an efficient semismooth Newton-type algorithm, SNSQP, that directly tackles SQCQP. The key innovation of our approach lies in the formulation of a newly introduced system of stationary equations, which characterizes the optimality conditions of the original problem. The algorithm is classified as a second-order method and thus exhibits a locally quadratic convergence rate while maintaining relatively low computational complexity due to the sparse structure of the solution. Extensive numerical experiments demonstrate that the algorithm consistently produces high-accuracy solutions with fast computational speed. However, given the challenges in establishing global convergence, the method remains local. Considering its strong numerical performance and extensive applications, ensuring its global convergence is worthy of further research.

Acknowledgments

This work is supported by the National Key R&D Program of China (2023YFA1011100)

References

[1] M. Ahn, J.-S. Pang, and J. Xin. Difference-of-convex learning: directional stationarity, optimality, and sparsity. *SIAM Journal on Optimization*, 27(3):1637–1665, 2017.

[2] M. Asteris, A. Kyrillidis, O. Koyejo, and R. Poldrack. A simple and provable algorithm for sparse diagonal cca. In *International Conference on Machine Learning*, pages 1148–1157. PMLR, 2016.

[3] S. Bahmani, B. Raj, and P. T. Boufounos. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 14(1):807–841, 2013.

[4] Y. Bai, R. Liang, and Z. Yang. Splitting augmented lagrangian method for optimization problems with a cardinality constraint and semicontinuous variables. *Optimization Methods and Software*, 31(5):1089–1109, 2016.

[5] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang. Restricted normal cones and sparsity optimization with affine constraints. *Foundations of Computational Mathematics*, 14:63–83, 2014.

[6] A. Beck and Y. C. Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

[7] A. Beck and N. Hallak. On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. *Mathematics of Operations Research*, 41(1):196–223, 2016.

[8] A. Beck and Y. Vaisbourd. The sparse principal component analysis problem: Optimality conditions and algorithms. *Journal of Optimization Theory and Applications*, 170:119–143, 2016.

[9] D. Bertsimas and R. Shioda. Algorithm for cardinality-constrained quadratic optimization. *Computational Optimization and Applications*, 43(1):1–22, 2009.

[10] J. D. Blanchard, J. Tanner, and K. Wei. Cgiht: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion. *Information and Inference: A Journal of the IMA*, 4(4):289–327, 2015.

[11] T. Blumensath and M. E. Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008.

[12] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[13] A.-L. Boulesteix, G. Durif, S. Lambert-Lacroix, J. Peyre, and K. Strimmer. *plsgenomics: PLS Analyses for Genomics*, 2024. R package version 1.5-3.

[14] M. Bredel, C. Bredel, D. Juric, G. R. Harsh, H. Vogel, L. D. Recht, and B. I. Sikic. Functional network analysis reveals extended gliomagenesis pathway maps and three novel myc-interacting genes in human gliomas. *Cancer research*, 65(19):8679–8689, 2005.

[15] M. Bredel, C. Bredel, D. Juric, G. R. Harsh, H. Vogel, L. D. Recht, and B. I. Sikic. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer research*, 65(10):4088–4096, 2005.

[16] O. P. Burdakov, C. Kanzow, and A. Schwartz. Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. *SIAM Journal on Optimization*, 26(1):397–425, 2016.

[17] S. Cacciatore and L. Tenori. *KODAMA: Knowledge Discovery by Accuracy Maximization*, 2023. R package version 2.4.

[18] M. Červinka, C. Kanzow, and A. Schwartz. Constraint qualifications and optimality conditions for optimization problems with cardinality constraints. *Mathematical Programming*, 160:353–377, 2016.

[19] F. Cesarone, A. Scozzari, and F. Tardella. A new method for mean-variance portfolio optimization with cardinality constraints. *Annals of Operations Research*, 205:213–234, 2013.

[20] T.-J. Chang, N. Meade, J. E. Beasley, and Y. M. Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, 27(13):1271–1302, 2000.

[21] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

[22] D. Chu, L.-Z. Liao, M. K. Ng, and X. Zhang. Sparse canonical correlation analysis: New formulation and algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3050–3065, 2013.

[23] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[24] X. Cui, X. Zheng, S. Zhu, and X. Sun. Convex relaxations and miqcqp reformulations for a class of cardinality-constrained portfolio selection problems. *Journal of Global Optimization*, 56(4):1409–1423, 2013.

[25] T. De Luca, F. Facchinei, and C. Kanzow. A semismooth equation approach to the solution of nonlinear complementarity problems. *Mathematical Programming*, 75:407–439, 1996.

[26] D. Di Lorenzo, G. Liuzzi, F. Rinaldi, F. Schoen, and M. Sciandrone. A concave optimization-based approach for sparse portfolio selection. *Optimization Methods and Software*, 27(6):983–1000, 2012.

[27] H. Dong, M. Ahn, and J.-S. Pang. Structural properties of affine sparsity constraints. *Mathematical Programming*, 176:95–135, 2019.

[28] F. Facchinei and J. Soares. A new merit function for nonlinear complementarity problems and a related algorithm. *SIAM Journal on Optimization*, 7(1):225–247, 1997.

[29] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[30] M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter. Complementarity formulations of l0-norm optimization problems. *Industrial Engineering and Management Sciences. Technical Report. Northwestern University, Evanston, IL, USA*, 5, 2013.

[31] A. Fischer. A special newton-type optimization method. *Optimization*, 24(3-4):269–284, 1992.

[32] S. A. Hamza and M. G. Amin. Hybrid sparse array beamforming design for general rank signal models. *IEEE Transactions on Signal Processing*, 67(24):6215–6226, 2019.

[33] D. R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83:331–353, 2011.

[34] R. Hesse, D. R. Luke, and P. Neumann. Alternating projections and douglas-rachford for sparse affine feasibility. *IEEE Transactions on Signal Processing*, 62(18):4868–4881, 2014.

[35] H. Huang, H. C. So, and A. M. Zoubir. Sparse array beamformer design via admm. *IEEE Transactions on Signal Processing*, 2023.

[36] C. Kanzow, A. B. Raharja, and A. Schwartz. An augmented lagrangian method for cardinality-constrained optimization problems. *Journal of Optimization Theory and Applications*, 189(3):793–813, 2021.

[37] C. Kanzow, A. Schwartz, and F. Weiß. The sparse(st) optimization problem: reformulations, optimality, stationarity, and numerical results. *Computational Optimization and Applications*, 90:77–112, 2025.

[38] A. Kyrillidis, S. Becker, V. Cevher, and C. Koch. Sparse projections onto the simplex. In *International Conference on Machine Learning*, pages 235–243. PMLR, 2013.

[39] X. Li, N. Xiu, and S. Zhou. Matrix optimization over low-rank spectral sets: stationary points and local and global minimizers. *Journal of Optimization Theory and Applications*, 184:895–930, 2020.

[40] O. Lindenbaum, M. Salhov, A. Averbuch, and Y. Kluger. L0-sparse canonical correlation analysis. In *International Conference on Learning Representations*, 2021.

[41] Z. Lu. Optimization over sparse symmetric sets via a nonmonotone projected gradient method. *arXiv preprint arXiv:1509.08581*, 2015.

[42] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.

[43] W. Min, J. Liu, and S. Zhang. Sparse weighted canonical correlation analysis. *Chinese Journal of Electronics*, 27(3):459–466, 2018.

[44] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[45] L. Pan, S. Zhou, N. Xiu, and H.-D. Qi. A convergent iterative hard thresholding for nonnegative sparsity optimization. *Pacific Journal of Optimization*, 13(2):325–353, 2017.

[46] L.-L. Pan, N.-H. Xiu, and S.-L. Zhou. On solutions of sparsity constrained optimization. *Journal of the Operations Research Society of China*, 3:421–439, 2015.

[47] L. Qi and H. Jiang. Semismooth karush-kuhn-tucker equations and convergence analysis of newton and quasi-newton methods for solving these equations. *Mathematics of Operations Research*, 22(2):301–325, 1997.

[48] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[49] B. K. Sriperumbudur, D. A. Torres, and G. R. Lanckriet. A majorization-minimization approach to the sparse generalized eigenvalue problem. *Machine Learning*, 85:3–39, 2011.

[50] S. Steffensen. Relaxation approaches for nonlinear sparse optimization problems. *Optimization*, 73(10):3237–3258, 2024.

[51] S. Steffensen and M. Ulbrich. A new relaxation scheme for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization*, 20(5):2504–2539, 2010.

[52] F. Streichert, H. Ulmer, and A. Zell. Evolutionary algorithms and the cardinality constrained portfolio optimization problem. In *Operations research proceedings 2003: Selected papers of the international conference on operations research (OR 2003) Heidelberg, September 3–5, 2003*, pages 253–260. Springer, 2004.

[53] J. Sun, L. Kong, and S. Zhou. Gradient projection newton algorithm for sparse collaborative learning using synthetic and real datasets of applications. *Journal of Computational and Applied Mathematics*, 422:114872, 2023.

[54] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[55] G. Wachsmuth. On LICQ and the uniqueness of lagrange multipliers. *Operations Research Letters*, 41(1):78–80, 2013.

[56] R. Wang, N. Xiu, and S. Zhou. An extended newton-type algorithm for ℓ_2-regularized sparse logistic regression and its efficiency for classifying large-scale datasets. *Journal of Computational and Applied Mathematics*, 397:113656, 2021.

[57] D. Witten, R. Tibshirani, S. Gross, and B. Narasimhan. *PMA: Penalized Multivariate Analysis*, 2024. R package version 1.2-3.

[58] X.-T. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.

[59] X.-T. Yuan and Q. Liu. Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4122–4129, 2014.

[60] C. Zhao, N. Xiu, H. Qi, and Z. Luo. A Lagrange–Newton algorithm for sparse nonlinear programming. *Mathematical Programming*, 195(1-2):903–928, 2022.

[61] S. Zhou. Gradient projection newton pursuit for sparsity constrained optimization. *Applied and Computational Harmonic Analysis*, 61:75–100, 2022.

[62] S. Zhou, N. Xiu, and H.-D. Qi. Global and quadratic convergence of newton hard-thresholding pursuit. *Journal of Machine Learning Research*, 22(12):1–45, 2021.