Noname manuscript No.
(will be inserted by the editor)

A Lagrange-Newton Algorithm for Sparse Nonlinear Programming

Chen Zhao · Naihua Xiu · Houduo Qi · Ziyan Luo

October 5, 2021

Abstract The sparse nonlinear programming (SNP) problem has wide applications in signal and image processing, machine learning and finance, etc. However, the computational challenge posed by SNP has not yet been well resolved due to the nonconvex and discontinuous ℓ_0-norm involved. In this paper, we resolve this numerical challenge by developing a fast Newton-type algorithm. As a theoretical cornerstone, we establish a first-order optimality condition for SNP based on the concept of strong β-Lagrangian stationarity via the Lagrangian function, and reformulate it as a system of nonlinear equations called the Lagrangian equations. The nonsingularity of the corresponding Jacobian is discussed, based on which the Lagrange-Newton algorithm (LNA) is then proposed. Under mild conditions, we establish the locally quadratic convergence and its iterative complexity estimation. To further demonstrate the efficiency and superiority of our proposed algorithm, we apply LNA to two specific problems arising from compressed sensing and sparse high-order portfolio selection, in which significant benefits accrue from the restricted Newton step.

This research was partially supported by the National Natural Science Foundation of China (11771038, 11971052, 12011530155) and Beijing Natural Science Foundation (Z190002).

Chen Zhao
Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China
E-mail: 14118409@bjtu.edu.cn

Naihua Xiu
Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China
E-mail: nhxiu@bjtu.edu.cn

Houduo Qi
School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK
E-mail: hdqi@soton.ac.uk

Ziyan Luo
Corresponding author
Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China
E-mail: zyluo@bjtu.edu.cn

Keywords Sparse nonlinear programming · Lagrangian equation · The Newton method · Locally quadratic convergence · Application

Mathematics Subject Classification (2010) 90C30 · 49M15 · 90C46

1 Introduction

In this paper, we are mainly concerned with the following sparse nonlinear programming (SNP) problem:

$$\min_{x\in\mathbb{R}^n} f(x), \quad \text{s.t. } h(x)=0,\ x\in\mathbb{S}, \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ and $h:=(h_1,\ldots,h_m)^\top:\mathbb{R}^n\to\mathbb{R}^m$ are twice continuously differentiable functions, $\mathbb{S}:=\{x\in\mathbb{R}^n:\|x\|_0\le s\}$ is the sparse constraint set with integer $s<n$ and $\|x\|_0$ is the ℓ_0-norm of x that counts the number of nonzero components of x. By denoting $\Omega:=\{x\in\mathbb{R}^n:h(x)=0\}$, the feasible set of problem (1) is abbreviated as $\Omega\cap\mathbb{S}$ and the optimal solution set can be written as $\arg\min_{x\in\Omega\cap\mathbb{S}} f(x)$. The SNP problem has wide applications ranging from linear and nonlinear compressed sensing [12,14] in signal processing, the sparse portfolio selection[17,39] in finance, to variable selection [26,20] and sparse principle component analysis [47,3] in high-dimensional statistical analysis and machine learning, etc. Unfortunately, due to the intrinsic combinatorial property in $\mathbb{S}$, the SNP problem is generally NP-hard, even for the simple convex quadratic objective function [27].

To well resolve the computational challenge resulting from $\mathbb{S}$, efforts have been made in two mainstreams in the literature. The first mainstream is "*relaxation*" approach, with a rich variety of relaxation schemes developed in [10, 15,18] (to just name a few). The second one is the "*greedy*" approach that tackles the involved ℓ_0-norm directly, with a large number of algorithms tailored for SNP with the feasible set $\mathbb{S}$ only (i.e., $\Omega=\mathbb{R}^n$), see, e.g., the first-order algorithms [36,28,5], and second-order algorithms with the Newton-type steps interpolated [42,43,41,9]. As the first-order information such as gradients are used in first-order greedy algorithms, linear rate convergence results are established as expected. While benefiting from the second-order information such as Hessian matrices, the aforementioned second-order greedy algorithms are witnessed in the numerical experiments with superior performance in terms of fast computation speed and high solution accuracy. For example, in a very recent work [44] which reports notable computational advantage, Zhou et al. proposed a new algorithm called Newton Hard-Thresholding Pursuit (NHTP) with cheap Newton steps in a restricted fashion and rigorously establish the quadratic convergence rate.

In sharp contrast to the fruitful computational algorithms for nonlinear programming with the single sparse constraint set $\mathbb{S}$, only a small portion of research on greedy algorithms is devoted to general SNP over $\mathbb{S}$ intersecting with some additional constraint set. The limited works are reported in [21, 33,2,22,23]. It is noteworthy that these algorithms are mostly gradient-based,

and no quadratic convergence rate can be expected. To fill this gap and also being inspired by the appealing theoretical and computational properties of NHTP [44], we develop a quadratically convergent Newton-type method for SNP when Ω is characterized by nonlinear equality constraints as presented in problem (1). The main contributions of this paper are summarized as below:

(i) The strong β-Lagrangian stationarity is introduced to characterize the optimality condition for SNP. Moreover, an equivalent characterization of such a stationary point as a system of equations is built and it makes possible to develop a Newton-type method.
(ii) The crucial Jacobian nonsingularity of the underlying system is well addressed under some mild assumptions, and the essential linear system in each iteration is reduced to be of size $(s+m)\times(s+m)$, a significant dimension reduction benefiting from the intrinsic sparsity.
(iii) Theoretically, the resulting Lagrange-Newton algorithm (LNA) is shown to possess the locally quadratic convergence. Numerically, LNA leads to outstanding performance for some specific applications including compressed sensing and sparse portfolio selection.

The remainder of this paper is organized as follows. In Section 2, the optimality condition in terms of Lagrangian stationarity is established. In Section 3, an equivalent Lagrangian equation system is proposed and the nonsingularity of its Jacobian is discussed. The framework of LNA and its locally quadratic convergence are elaborated in Section 4. Two well-known applications are analyzed in Section 5. Extensive numerical experiments are conducted in Section 6. Conclusions are made in Section 7.

For convenience, the following notations will be used throughout the paper. For any given positive integer n, denote $[n] := \{1,\dots,n\}$. For an index set $J \subseteq [n]$, let $|J|$ be the cardinality of J that counts the number of elements in J, and $J^c := [n] \setminus J$ be its complementary set. The collection of all index sets with cardinality s in $[n]$ is defined by $\mathcal{J}_s := \{J \subseteq [n] : |J| = s\}$. Given $x \in \mathbb{R}^n$, denote $\mathrm{supp}(x) := \{i \in [n] : x_i \neq 0\}$ and $\mathcal{J}_s(x) := \{J \in \mathcal{J}_s : \mathrm{supp}(x) \subseteq J\}$. We define $x_T \in \mathbb{R}^{|T|}$ as the subvector of $x \in \mathbb{R}^n$ indexed by T. For a matrix $A \in \mathbb{R}^{m\times n}$, define $A_{I,J}$ as a submatrix whose rows and columns are respectively indexed by I and J. In particular, we write A_T as its submatrix consisting of columns indexed by T and $A_{T,\cdot}$ as its submatrix consisting of rows indexed by T. Given a twice continuously differentiable function g with its gradient $\nabla g(x)$ and Hessian $\nabla^2 g(x)$ at x, denote $\nabla_T g(x) := (\nabla g(x))_T$ and $\nabla^2_{I,J} g(x) := (\nabla^2 g(x))_{I,J}$. The Euclidean norm of a vector x is denoted by $\|x\|$, and the spectral norm of a matrix A is denoted by $\|A\|$.

2 Lagrangian Stationarity

This section is devoted to the optimality conditions for (1) in terms of the Lagrangian stationarity, which will build up the theoretical fundamentals to our new proposed algorithm in the sequel.

Firstly, we consider the projection on the sparse set $\mathbb{S}$. For any given nonempty closed set $Q \subseteq \mathbb{R}^n$ and any $x \in \mathbb{R}^n$, define the projection operator $\Pi_Q(x) := \arg\min_{y\in Q} \|x-y\|^2$. Recall from [2,32] that the projection operator $\Pi_{\mathbb{S}}$ admits an explicit formula as follows: for any $z \in \mathbb{R}^n$, and for any $\pi \in \Pi_{\mathbb{S}}(z)$, we have

$$\pi_{t_i} = \begin{cases} z_{t_i}, \, i \in [s], \\ 0, \quad \text{otherwise}, \end{cases} \tag{2}$$

where $\{t_1, \ldots, t_n\}$ satisfies $|z_{t_1}| \geq \ldots \geq |z_{t_n}|$. Utilizing the sparse projection $\Pi_{\mathbb{S}}$, Beck and Eldar [1, Theorem 2.2] introduced and characterized an optimality condition for SNP with the single sparse constraint set $\mathbb{S}$: if x^* is a global minimizer, then for any $L > L_f$, $x^* = \Pi_{\mathbb{S}}(x^* - \nabla f(x^*)/L)$, where L_f is the Lipschitz constant of ∇f. Later, Lu [22] extended such a result to the case of $\Omega' \cap \mathbb{S}$ by means of $x^* = \Pi_{\Omega'\cap\mathbb{S}}(x^* - \nabla f(x^*)/L)$, where Ω' is a closed convex set and enjoys certain symmetric property. Limitation follows when $\Pi_{\Omega'\cap\mathbb{S}}(\cdot)$ has no explicit expression. This motivates us to introduce the following Lagrangian stationarity.

Definition 1 Given $x^* \in \mathbb{R}^n$ and $\beta > 0$, x^* is called a strong β-Lagrangian stationary point of problem (1) if there exists a Lagrangian multiplier $y^* \in \mathbb{R}^m$ such that

$$\begin{cases} x^* = \Pi_{\mathbb{S}}(x^* - \beta \nabla_x L(x^*, y^*)), \\ h(x^*) = 0, \end{cases} \tag{3}$$

where $L(x,y) := f(x) - \langle y, h(x)\rangle$ is the Lagrangian function associated with (1) for any $x \in \mathbb{S}$ and $y \in \mathbb{R}^m$.

By employing (2) of $\Pi_{\mathbb{S}}$, one can easily rewrite system (3) as follows.

Lemma 1 *Given $x^* \in \mathbb{R}^n$ and $y^* \in \mathbb{R}^m$, denote $\Gamma^* := supp(x^*)$ and $q^* := \nabla_x L(x^*, y^*)$. Then x^* is a strong β-Lagrangian stationary point of problem* (1) *with y^* if and only if*

$$x^* \in \mathbb{S}, \; h(x^*) = 0, \; \begin{cases} \beta \|q^*_{(\Gamma^*)^c}\|_\infty < |x^*|_{(s)} \;\&\; q^*_{\Gamma^*} = 0, & \text{if } \|x^*\|_0 = s, \quad (4) \\ q^* = 0, & \text{if } \|x^*\|_0 < s, \quad (5) \end{cases}$$

where $supp(x) := \{i \in [n] : x_i \neq 0\}$ and $x_{(s)}$ is the sth largest component of x.

The following assumption is introduced, followed by optimality analysis.

Assumption 1 Given $x \in \Omega \cap \mathbb{S}$, $\operatorname{rank}(\nabla_\Gamma h(x)) = m$, where $\nabla h(x) := (\nabla h_1(x), \ldots, \nabla h_m(x))^\top$, $\Gamma = \operatorname{supp}(x)$.

Theorem 1 (First-order necessary optimality condition) *Suppose that x^* is a local minimizer of* (1) *and Assumption 1 holds at x^*. Then there exists a unique $y^* \in \mathbb{R}^m$ such that x^* is a strong β-Lagrangian stationary point of problem* (1) *for any $\beta \in (0, \hat{\beta})$, where*

$$\hat{\beta} := \begin{cases} \dfrac{|x^*|_{(s)}}{\left\|q^*_{(\Gamma^*)^c}\right\|_\infty}, & \text{if } \|x^*\|_0 = s \text{ and } q^*_{(\Gamma^*)^c} \neq 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{6}$$

Proof. Since x^* is a local minimizer of (1.1), for any $J \in \mathcal{J}_s(x^*)$, x^* is also a local minimizer of

$$(P_J) \quad \min_{x\in\mathbb{R}^n} f(x), \ \text{ s.t. } h(x) = 0, \ x_{J^c} = 0.$$

Assumption 1 implies that $\nabla h_i(x^*),\ i \in [m],\ e_j,\ j \in J^c$ are linearly independent, where $e_j \in \mathbb{R}^n$ is the jth column in the identity matrix. It means that linear independent constraint qualification (LICQ) holds at x^* for (P_J). Thus for any given $J \in \mathcal{J}_s(x^*)$, there exists a unique $y^J \in \mathbb{R}^m$ and a unique $z^J \in \mathbb{R}^{n-s}$ such that

$$\nabla f(x^*) = (\nabla h(x^*))^\top y^J + \sum_{j\in J^c} z_j^J e_j. \tag{7}$$

Case I: When $\|x^*\|_0 = s$, we have $\mathcal{J}_s(x^*) = \{\Gamma^*\}$. Set $J = \Gamma^*$ in (7), and let $y^* = y^{\Gamma^*}$. Direct calculations yield

$$q^*_{\Gamma^*} = \left(\nabla f(x^*) - (\nabla h(x^*))^\top y^*\right)_{\Gamma^*} = \left(\sum_{i\in(\Gamma^*)^c} z_i^{\Gamma^*} e_i\right)_{\Gamma^*} = 0 \ \text{ and } \ q^*_{(\Gamma^*)^c} = z^{\Gamma^*},$$

from which (4) holds for all $\beta \in (0, \hat{\beta})$.
Case II: When $\|x^*\|_0 < s$, $|\mathcal{J}_s(x^*)| = \mathcal{C}_{n-\Gamma^*}^{s-\Gamma^*}$ is finite. For any $J \in \mathcal{J}_s(x^*)$, rewrite (7) into the block form

$$\begin{cases} \nabla_{\Gamma^*} f(x^*) = (\nabla_{\Gamma^*} h(x^*))^\top y^J, \\ \nabla_{J\setminus\Gamma^*} f(x^*) = (\nabla_{J\setminus\Gamma^*} h(x^*))^\top y^J, \\ \nabla_{J^c} f(x^*) = (\nabla_{J^c} h(x^*))^\top y^J + z^J. \end{cases} \tag{8}$$

By virtue of Assumption 1, the first equation in (8) indicates that y^J's coincide for all $J \in \mathcal{J}_s(x^*)$, which we assign to y^*. It then leads to $q^*_{\Gamma^*} = 0$. As $\bigcup_{J\in\mathcal{J}_s(x^*)}(J\setminus\Gamma^*) = (\Gamma^*)^c$, it follows from the second equation in (8) that $q^*_{(\Gamma^*)^c} = 0$. Thus, $q^* = 0$. This completes the proof by utilizing Lemma 1. □

Theorem 2 (First-order sufficient optimality condition) *Let f be a convex function and h be an affine function. Given $\beta > 0$, suppose that x^* is a strong β-Lagrangian stationary point of* (1) *with the Lagrangian multiplier $y^* \in \mathbb{R}^m$. If $\|x^*\|_0 = s$, then x^* is a local minimizer of* (1)*; If $\|x^*\|_0 < s$, then x^* is a global minimizer of* (1).

Proof. Under the hypotheses on f and h, the Lagrangian function $L(x, y)$ is convex with respect to x. It follows that

$$L(x, y^*) \geq L(x^*, y^*) + \langle q^*, x - x^*\rangle, \ \forall\ x \in \Omega \cap \mathbb{S}. \tag{9}$$

Using the facts $x, x^* \in \Omega \cap \mathbb{S}$, we have $L(x, y^*) = f(x),\ L(x^*, y^*) = f(x^*)$. Since x^* is a strong β-Lagrangian stationary point with y^*, if $\|x^*\|_0 < s$, $\langle q^*, x - x^*\rangle = 0$ from (5). Then for any $x \in \Omega \cap \mathbb{S}$, $f(x) \geq f(x^*)$. Thus x^* is a global minimizer. If $\|x^*\|_0 = s$, there exists a sufficiently small $\delta > 0$ such

that for any $x \in \mathcal{N}(x^*, \delta) \cap (\Omega \cap \mathbb{S})$, $x_{(\Gamma^*)^c} = 0$ and hence $(x - x^*)_{(\Gamma^*)^c} = 0$. By invoking (4), it yields that

$$\langle q^*, x - x^* \rangle = \langle q^*_{\Gamma^*}, (x - x^*)_{\Gamma^*} \rangle + \langle q^*_{(\Gamma^*)^c}, (x - x^*)_{(\Gamma^*)^c} \rangle = 0.$$

Thus for any $x \in \mathcal{N}(x^*, \delta) \cap (\Omega \cap \mathbb{S})$, $f(x) \geq f(x^*)$, which implies that x^* is a local minimizer of (1). □

Remark 1 We make a few comments on problem (1). (i) By virtue of Theorem 1 and Lemma 1, we summarize the relations among the strong β-Lagrangian stationarity (Strong-β-LS), B-KKT point and C-KKT point in [31, Definition 3.1], S-stationarity (S-stat) and M-stationarity (M-stat) in [8, Definition 4.1] as below.

$$\begin{array}{ccccccc} & & \text{Local minimizer} & \overset{\text{Assumption 1}}{\Longrightarrow} & \text{Strong-}\beta\text{-LS } \forall \beta \in (0, \hat{\beta}) & & \\ & & & & \Updownarrow & & \\ \text{M-stat} & \Longleftarrow & \text{S-stat} & \Longleftrightarrow & \text{C-KKT point} & \Longleftarrow & \text{B-KKT point} \end{array}$$

(ii) Given a feasible solution x, we can show that Assumption 1, the restricted Robinson constraint qualification (R-RCQ) in [30, Definition 3.1], and the cardinality constraints linear independence constraint qualification (CC-LICQ) in [8, Definition 3.11] are equivalent. They are stronger than the restricted linear independent constraint qualification (R-LICQ) in [31, Definition 2.4] when $\|x\|_0 = s$. Learning from [31], together with (i) of this remark, R-LICQ also ensures the existence of strong β-Lagrangian stationary point. However, it is not sufficient to guarantee the uniqueness of the corresponding Lagrangian multiplier as stated in Theorem 1.

(iii) As stated in Theorem 1, when the ℓ_0-norm of a local minimizer x^* is smaller than s, then x^* is a strong β-Lagrangian stationary point for any $\beta > 0$ under Assumption 1. In other words, $\hat{\beta}$ can be taken as ∞ in such a case. This, to some extent, refines the unified bound $1/L_f$ (L_f is the Lipschitz constant of ∇f) as proposed in [1, Theorem 2.2] where $\Omega = \mathbb{R}^n$. However, when $\|x^*\|_0 = s$, it is difficult to estimate an easily checkable and unified lower bound of $\hat{\beta}$ for all local minimizers in general, due to the additional equality constraint. Nevertheless, if f is strongly convex, then it leads to finitely many local minimizers of problem (1) (attributing to unique minimizers of subproblems (P_J)). For those finitely many local minimizers x^*'s with nonzero multipliers y^*'s (but bounded under Assumption 1), we can then take the minimum among those ratios defined in (6) to obtain a unified lower bound of $\hat{\beta}$.

3 Lagrangian Equations and Jacobian Nonsingularity

In this section, we will present an equivalent reformulation for the Lagrangian stationarity in terms of nonlinear equations, and discuss the Jacobian nonsingularity of the resulting equation system.

3.1 Lagrangian Equations

The optimality conditions in terms of the strong β-Lagrangian stationary point, as established in Theorems 1 and 2, provide a way of solving (1). As one knows that $\Pi_{\mathbb{S}}(\cdot)$ is not differentiable, the main challenge is how to tackle the non-differentiability. By exploiting the special structure possessed by the projection operator $\Pi_{\mathbb{S}}$, we propose a differentiable reformulation of the strong β-Lagrangian stationary point (3), using a finite sequence of Lagrangian equations.

Definition 2 Given $x \in \mathbb{S}$, $y \in \mathbb{R}^m$ and $\beta > 0$, denote $u := x - \beta\nabla_x L(x, y)$. Define the collection of sparse projection index sets of u by

$$\mathbb{T}(x, y; \beta) := \{T \in \mathcal{J}_s : |u_i| \geq |u_j|,\ \forall i \in T, \forall j \in T^c\}. \tag{10}$$

For any given $T \in \mathbb{T}(x, y; \beta)$, define the corresponding Lagrangian equation as

$$F(x, y; T) := \begin{bmatrix} (\nabla_x L(x, y))_T \\ x_{T^c} \\ -h(x) \end{bmatrix} = 0. \tag{11}$$

As one can see, the function $F(x, y; T)$ in (11) is differentiable with respect to x and y once T is selected. Moreover, we have the following equivalent relationship between (11) and (3).

Theorem 3 *Given $x^* \in \mathbb{S}$, $y^* \in \mathbb{R}^m$ and $\beta > 0$, x^* is a strong β-Lagrangian stationary point of* (1) *with the Lagrangian multiplier y^* if and only if for any $T \in \mathbb{T}(x^*, y^*; \beta)$, $F(x^*, y^*; T) = 0$. Meanwhile, $\mathbb{T}(x^*, y^*; \beta) = \mathcal{J}_s(x^*)$.*

Proof. By invoking the proof of Lemma 4 in [44], we can obtain the equivalent relationship directly. Now, we prove the rest part of the theorem. As just proved, if x^* is a strong β-Lagrangian stationary point of (1) with the Lagrangian multiplier y^*, then for any $T \in \mathbb{T}(x^*, y^*; \beta)$, $F(x^*, y^*; T) = 0$. Thus, $x^*_{T^c} = 0$ and hence $\mathrm{supp}(x^*) \subseteq T$. It then yields that $T \in \mathcal{J}_s(x^*)$. The arbitrariness of T leads to the inclusion $\mathbb{T}(x^*, y^*; \beta) \subseteq \mathcal{J}_s(x^*)$. It now suffices to show $\mathcal{J}_s(x^*) \subseteq \mathbb{T}(x^*, y^*; \beta)$. If $\|x^*\|_0 < s$, then $\nabla_x L(x^*, y^*) = 0$ by invoking (5), and hence $u^* = x^* - \beta\nabla_x L(x^*, y^*) = x^*$. For any $T \in \mathcal{J}_s(x^*)$, it is easy to verify that for any $i \in T$ and any $j \in T^c$, $|u^*_i| = |x^*_i| \geq |x^*_j| = |u^*_j|$, which indicates that $T \in \mathbb{T}(x^*, y^*; \beta)$. If $\|x^*\|_0 = s$, then $|x^*|_{(s)} > 0$ and $\mathcal{J}_s(x^*) = \{\Gamma^*\}$. By virtue of (4), we have that for any $i \in \Gamma^*$ and any $j \notin \Gamma^*$, $|u^*_i| = |x^*_i| \geq |x^*|_{(s)} > \beta|(\nabla_x L(x^*, y^*))_j| = |u^*_j|$. Thus, $\Gamma^* \in \mathbb{T}(x^*, y^*; \beta)$. Therefore, for both cases, we can conclude $\mathcal{J}_s(x^*) \subseteq \mathbb{T}(x^*, y^*; \beta)$. This completes the proof. □

3.2 Jacobian Nonsingularity

In this subsection, let x^* be a strong β-Lagrangian stationary point with y^*. To handle the Lagrangian equation (11) for a given index set $T \in \mathbb{T}(x^*, y^*; \beta)$,

it is crucial to investigate the nonsingularity of the Jacobian of $F(x,y;T)$ with respect to (x,y) in a neighborhood of (x^*,y^*), namely,

$$\nabla_{(x,y)}F(x,y;T)=\begin{bmatrix}(\nabla^2_{xx}L(x,y))_{T,\cdot} & -(\nabla_T h(x))^\top\\ I_{T^c,\cdot} & 0\\ -\nabla h(x) & 0\end{bmatrix}\in\mathbb{R}^{(n+m)\times(n+m)}. \quad (12)$$

Here, $\nabla^2_{xx}L(x,y)=\nabla^2 f(x)-\sum_{i=1}^{m} y_i\nabla^2 h_i(x)$ is Hessian matrix of $L(x,y)$ with respect to x, and I is the identity matrix. It is worth mentioning that since T is related to (x,y), the conventional Jacobian of $F(x,y;T)$ may differ with (12) if we treat T as a function of (x,y). To deal with the changing nature of T we will update T in our proposed iterative algorithm adaptively. Two more assumptions are stated below.

Assumption 1′ $\mathrm{rank}(\nabla_T h(x^*))=m$, for any $T\in\mathcal{J}_s(x^*)$.

Assumption 2 (Second-order optimality condition) For any $T\in\mathcal{J}_s(x^*)$, $(\nabla^2_{xx}L(x^*,y^*))_{T,T}$ is positive definite restricted to the null space of $\nabla_T h(x^*)$:

$$d^\top\left(\nabla^2_{xx}L(x^*,y^*)\right)_{T,T}d>0,\forall\, 0\neq d\in\mathbb{N}(\nabla_T h(x^*)):=\{d\in\mathbb{R}^s:\nabla_T h(x^*)d=0\}.$$

Given any $T\in\mathcal{J}_s$ and $(x,y)\in\mathbb{S}\times\mathbb{R}^m$, elementary row operations yield the equivalence between the nonsingularity of $\nabla_{(x,y)}F(x,y;T)$ and that of

$$G(x,y;T):=\begin{bmatrix}(\nabla^2_{xx}L(x,y))_{T,T} & -(\nabla_T h(x))^\top\\ -\nabla_T h(x) & 0\end{bmatrix}\in\mathbb{R}^{(s+m)\times(s+m)}. \quad (13)$$

Thus, we call $G(x,y;T)$ the *reduced Jacobian* of $F(x,y;T)$. Furthermore, when Assumptions 1′ and 2 hold, we can directly obtain the desired nonsingularity at (x^*,y^*) as stated in the following theorem.

Theorem 4 *Let x^* be a strong β-Lagrangian stationary point with y^*. If Assumptions 1′ and 2 hold, then $\nabla_{(x,y)}F(x^*,y^*;T)$ is nonsingular for each index set $T\in\mathbb{T}(x^*,y^*;\beta)$.*

The rest of this subsection is devoted to the nonsingularity of $G(x,y;T)$ when (x,y) are sufficiently close to (x^*,y^*), by employing the achieved nonsingularity of $G(x^*,y^*;T)$ and the following assumption.

Assumption 3 $\nabla^2 f$ and $\nabla^2 h_i$ $(i\in[m])$ are Lipschitz continuous near x^*.

The locally Lipschitz continuity in Assumption 3 allows us to find positive constants δ_0^*, L_1 and L_2 such that for any $z:=(x;y)$, $\hat{z}:=(\hat{x};\hat{y})\in\mathcal{N}(z^*,\delta_0^*)$ with $z^*:=(x^*;y^*)$, we have

$$\|\nabla_x L(x,y)-\nabla_x L(\hat{x},\hat{y})\|\le L_1\|z-\hat{z}\|, \quad (14)$$
$$\|\nabla^2 L(x,y)-\nabla^2 L(\hat{x},\hat{y})\|\le L_2\|z-\hat{z}\|. \quad (15)$$

Let $x^* \neq 0$ (since the trivial case $x^* = 0$ is not desired in practice) be a strong β-Lagrangian stationary point with y^*. We can define

$$\delta_1^* := \frac{\min\limits_{i\in\Gamma^*} |x_i^*| - \beta \max\limits_{i\in(\Gamma^*)^c} |q_i^*|}{\sqrt{2}(1+\beta L_1)},$$

where $\Gamma^* \neq \emptyset$ (because $x^* \neq 0$) and q^* are defined in Lemma 1. By employing Lemma 1, one can easily verify that $\delta_1^* > 0$ since $x^* \neq 0$. Denote

$$\delta^* := \min\{\delta_0^*, \delta_1^*\}, \tag{16}$$

$$\mathcal{N}_{\mathbb{S}}(z^*;\delta^*) := \{z \in \mathbb{R}^{n+m} : x \in \mathbb{S},\ \|z - z^*\| < \delta^*\}. \tag{17}$$

Lemma 2 *Let x^* be a strong β-Lagrangian stationary point with y^*. Denote $z^* = (x^*; y^*)$. If Assumption 3 holds, then for any $z = (x; y) \in \mathcal{N}_{\mathbb{S}}(z^*;\delta^*)$, we have*

$$\mathbb{T}(x,y;\beta) \subseteq \mathbb{T}(x^*,y^*;\beta) \quad \textit{and} \quad \Gamma^* \subseteq \textit{supp}(x) \cap T, \forall\ T \in \mathbb{T}(x,y;\beta). \tag{18}$$

Furthermore, if $\|x^*\|_0 = s$, then $\{\textit{supp}(x)\} = \mathbb{T}(x,y;\beta) = \mathbb{T}(x^*,y^*;\beta) = \{\Gamma^*\}$.

Proof. Since x^* is a strong β-Lagrangian stationary point with y^*, we have $F(x^*,y^*;\bar{T}) = 0$, $\forall\ \bar{T} \in \mathbb{T}(x^*,y^*;\beta) = \mathcal{J}_s(x^*)$ from Theorem 3. Consider any given $z = (x;y) \in \mathcal{N}_{\mathbb{S}}(z^*;\delta^*)$, denote $\Gamma = \mathrm{supp}(x)$ and $q = \nabla_x L(x,y)$. For any $i \in \Gamma^*$ and any $j \in (\Gamma^*)^c$, we have

$$\begin{aligned}
& |x_i - \beta q_i| - |x_j - \beta q_j| \\
\geq\ & |x_i^*| - |x_i - x_i^*| - |x_j - x_j^*| - \beta|q_i - q_i^*| - \beta|q_j - q_j^*| - \beta|q_j^*| \\
\overset{(14)}{\geq}\ & \min_{t\in\Gamma^*} |x_t^*| - \sqrt{2}\|z - z^*\| - \sqrt{2}\beta L_1\|z - z^*\| - \beta \max_{t\in(\Gamma^*)^c} |q_t^*| \\
\geq\ & \min_{t\in\Gamma^*} |x_t^*| - \beta \max_{t\in(\Gamma^*)^c} |q_t^*| - \sqrt{2}(1+\beta L_1)\delta^* \\
\geq\ & 0.
\end{aligned}$$

This indicates that $i \in T$ and hence

$$\Gamma^* \subseteq T, \quad \forall\ T \in \mathbb{T}(x,y;\beta). \tag{19}$$

Furthermore, we have

$$\mathbb{T}(x,y;\beta) \subseteq \mathcal{J}_s(x^*) = \mathbb{T}(x^*,y^*;\beta). \tag{20}$$

Next we claim that Γ^* is also a subset of Γ. If not, there exists an index $i_0 \in \Gamma^* \setminus \Gamma$. Then

$$\|z - z^*\| \geq \|x - x^*\| \geq |(x - x^*)_{i_0}| = |x_{i_0}^*| \geq \min_{t\in\Gamma^*} |x_t^*| \geq \delta^*,$$

which is a contradiction to $z \in \mathcal{N}_{\mathbb{S}}(z^*;\delta^*)$. Thus,

$$\Gamma^* \subseteq \Gamma. \tag{21}$$

Summarizing (19), (20) and (21), we get (18). Particularly, if $\|x^*\|_0 = s$, then $|\Gamma^*| = s$ and $\mathcal{J}_s(x^*) = \{\Gamma^*\}$. Utilizing (19), (20) and (21) again, together with the fact $|\Gamma| \leq s$, we immediately get the rest of the desired assertion. □

Note that the technique used in the proof of Lemma 2 is similar to that in [24, Appendix A], which is to construct a lower bound of the radius of a neighborhood to ensure the required inclusion property for all reference points in the corresponding neighborhood.

Finally, we state the result of the desired nonsingularity.

Theorem 5 *Let x^* be a strong β-Lagrangian stationary point with y^*. If Assumptions 1′, 2 and 3 hold, then there exist constants $\tilde{\delta}^* \in (0, \delta^*]$ and $M^* \in (0, +\infty)$ such that for any $z = (x; y) \in \mathcal{N}_{\mathbb{S}}(z^*; \tilde{\delta}^*)$ with $z^* = (x^*; y^*)$, the reduced Jacobian matrix $G(x, y; T)$ is nonsingular and*

$$\|G^{-1}(x, y; T)\| \leq M^*, \quad \forall T \in \mathbb{T}(x, y; \beta). \tag{22}$$

Proof. By invoking Theorems 3 and 4, we can get the nonsingularity of $G(x^*, y^*; T)$ for all $T \in \mathbb{T}(x^*, y^*; \beta) = \mathcal{J}_s(x^*)$. For any $z \in \mathcal{N}_{\mathbb{S}}(z^*; \delta^*)$, the inclusion $\mathbb{T}(x, y; \beta) \subseteq \mathbb{T}(x^*, y^*; \beta)$ from Lemma 2 immediately yields the nonsingularity of $G(x^*, y^*; T)$ for each $T \in \mathbb{T}(x, y; \beta)$. Furthermore, it follows from (15) that for any $T \in \mathbb{T}(x, y; \beta)$,

$$\|G(x, y; T) - G(x^*, y^*; T)\| \leq \|\nabla^2 L(x, y) - \nabla^2 L(x^*, y^*)\| \leq L_2 \|z - z^*\|. \tag{23}$$

This indicates that $G(\cdot, \cdot\ ; T)$ is Lipschitz continuous near z^* for any given $T \in \mathbb{T}(x, y; \beta)$. Thus, there exists $\delta_T > 0$ and $M_T > 0$ such that for any $z \in \mathcal{N}(z^*; \delta_T)$, $G(x, y; T)$ is nonsingular and $\|G^{-1}(x, y; T)\| \leq M_T$. Set

$$\tilde{\delta}^* := \min\{\delta^*, \{\delta_T\}_{T \in \mathcal{J}_s(x^*)}\} \quad \text{and} \quad M^* := \max_{T \in \mathcal{J}_s(x^*)} \{M_T\}. \tag{24}$$

It follows readily that for any $z \in \mathcal{N}(z^*; \tilde{\delta}^*)$, $G(x, y; T)$ is nonsingular and $\|G^{-1}(x, y; T)\| \leq M^*$, for all $T \in \mathbb{T}(x, y; \beta)$. □

4 The Lagrange-Newton Algorithm

In this section, we propose a Newton algorithm for solving Lagrangian equation (11) of problem (1), named as Lagrange-Newton Algorithm (LNA), and analyze the convergence rate of the algorithm.

4.1 LNA Framework

By employing the relationship between the strong β-Lagrangian stationary point and the Lagrangian equations as stated in Theorem 3, we give the basic idea behind our algorithm as follows: solve the Lagrangian equation $F(x, y; T) = 0$ iteratively by using the Newton method, and update the involved index

set T accordingly from $\mathbb{T}(x, y; \beta)$ in each iteration. This leads to Algorithm 1. Detailed explanation of the algorithm is given below.

Given $\beta > 0$, let $(x^k, y^k) \in \mathbb{S} \times \mathbb{R}^m$ be the current iteration.

`Index Set Selection:` Choose one index set T_k from $\mathbb{T}(x^k, y^k; \beta)$ defined as in (10). This can be safely accomplished by picking the indices of the first s largest elements (in magnitude) in $x^k - \beta \nabla_x L(x^k, y^k)$.

`The Newton Step:` The classical Newton equation is

$$\nabla_{(x,y)} F(x^k, y^k; T_k)(x^{k+1} - x^k; y^{k+1} - y^k) = -F(x^k, y^k; T_k). \tag{25}$$

After simple calculations, (25) can be rewritten as

$$\begin{cases} x^{k+1}_{T_k^c} = 0; \\ G(x^k, y^k; T_k) \begin{bmatrix} x^{k+1}_{T_k} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} -\nabla_{T_k} f(x^k) + (\nabla^2_{xx} L(x^k, y^k))_{T_k,.} x^k \\ h(x^k) - \nabla h(x^k) x^k \end{bmatrix}, \end{cases} \tag{26}$$

from which a significant dimension reduction is attained, from $(n+m)\times(n+m)$ to $(s+m)\times(s+m)$. Under the conditions presented in Theorem 5, $G(x^k, y^k; T_k)$ is nonsingular, and hence the next iteration can be obtained from the unique solution of (26), which can be solved in a direct way if $s + m$ is small or by employing the conjugate gradient (CG) method when $s + m$ is relatively large. As indicated, the low computational cost of the Newton step is greatly beneficial from the intrinsic sparsity, especially when $s \ll n$.

`Stopping Criterion:` Given the current iteration triplet (x^k, y^k, T_k), to measure how far x^k is from being a strong β-Lagrangian stationary point, the following quantity is adopted

$$\eta_\beta(x^k, y^k; T_k) := \|F(x^k, y^k; T_k)\| + \max_{i \in T_k^c} \Big\{ \max\Big(|(\nabla_x L(x^k, y^k))_i| - |x^k|_{(s)}/\beta,\ 0 \Big) \Big\}. \tag{27}$$

The first term on the right-hand side of (27) is to measure the residual of the Lagrangian equation system, and the second term is to testify

$$\Big\| \big(\nabla_x L(x^k, y^k) \big)_{T_k^c} \Big\|_\infty \le \frac{1}{\beta} |x^k|_{(s)},$$

which comes from (10) and (11) after simple manipulations. The stopping criterion is then designed in terms of $\eta_\beta(x^k, y^k; T_k)$.

Algorithm 1 Lagrange-Newton Algorithm (LNA) for (1)

Step 0. `(Initialization)` Give $\beta > 0$ and $\epsilon > 0$, choose $(x^0, y^0) \in \mathbb{S} \times \mathbb{R}^m$ and set $k = 0$.
Step 1. `(Index Set Selection)` Choose $T_k \in \mathbb{T}(x^k, y^k; \beta)$ by (10).
Step 2. `(Stopping Criterion)` If $\eta_\beta(x^k, y^k;\ T_k) \le \epsilon$, then stop. Otherwise, go to **Step 3.**
Step 3. `(The Newton Step)` Update (x^{k+1}, y^{k+1}) by (26), set $k = k + 1$ and go to **Step 1.**

Remark 2 We note that LNA shares a similar framework with the piecewise sequential quadratic programming(SQP) method [25] for nonlinear programming problems with equilibrium constraints, which is based on the classical SQP [38,19,35]. One may regard the Newton step of LNA as one step of the SQP method for solving the original problem in a local piece generated by restricting $x_{T_k^c} = 0$, where T_k is chosen based on the current iteration point (x^k, y^k). The selection of this piece is due to the Lagrangian equation defined by (11), which is a reformulation of the first-order optimality condition. It is worth pointing out that the idea of selecting a local piece at each iteration in LNA is not trivial, comparing to the traditional piecewise SQP. As can be seen from the construction of the Lagrangian equation in Definition 2, the pieces are not generated by using the disjunctive property of $\mathbb{S} = \cup_{J\in\mathcal{J}_s}\mathbb{R}^n_J$ or simply the support set of x^k, but by the special property of sparse projection. More importantly, such a selection scheme can significantly reduce the dimension of the subproblems.

4.2 Locally Quadratic Convergence

The locally quadratic convergence of LNA is shown to be inherited from the classic Newton method, armed with the essential invariance property of index sets as stated in Lemma 2. Specifically, we have

Theorem 6 *Given $\beta > 0$, suppose x^* is a strong β-Lagrangian stationary point of (1) with y^*. If Assumptions 1′, 2 and 3 hold. Let $\tilde{\delta}^*$ and M^* be defined as in (24), and $\mathbb{T}(x^*, y^*; \beta)$ be defined as in (10), respectively. Denote $z^* = (x^*; y^*)$. Suppose that the initial point $z^0 := (x^0; y^0)$ of LNA satisfies $z^0 \in \mathcal{N}_{\mathbb{S}}(z^*, \delta)$ with $\delta = \min\{\tilde{\delta}^*, \frac{1}{M^* L_2}\}$. Then the sequence $\{z^k := (x^k; y^k)\}$ generated by LNA is well-defined and for any $k \geq 0$,*

(i) $\lim\limits_{k\to\infty} z^k = z^*$ *with quadratic convergence rate, namely*

$$\|z^{k+1} - z^*\| \leq \frac{M^* L_2}{2} \left\|z^k - z^*\right\|^2.$$

(ii) $\lim\limits_{k\to\infty} F(z^k; T_k) = 0$ *with quadratic convergence rate, namely*

$$\|F(z^{k+1}; T_{k+1})\| \leq \frac{M^* L_2 \sqrt{L_1^2 + 1}}{\lambda_H} \|F(z^k; T_k)\|^2,$$

where $\lambda_H := \min\limits_{T_k \in \mathcal{J}_s(x^*)} \lambda_{\min}\left(\nabla_z F(z^*; T_k)^\top \nabla_z F(z^*; T_k)\right)$.

(iii) $\lim\limits_{k\to\infty} \eta_\beta(z^k; T_k) = 0$ *with* $\eta_\beta(z^{k+1}; T_{k+1}) \leq M^* L_2 \sqrt{L_1^2 + 1} \|z^k - z^*\|^2$ *and LNA will terminate when*

$$k \geq \left\lceil \frac{\log_2\left(4\delta^2 M^* L_2 \sqrt{L_1^2 + 1}/\epsilon\right)}{2} \right\rceil,$$

where $\lceil t \rceil$ denotes the smallest integer no less than t.

Proof. By employing Theorem 5, we know that the sequence $\{z^k\}$ generated by LNA is well-defined from the nonsingularity of $G(x^k, y^k; T_k)$ for all $k \geq 0$.

(i) Choose $T_0 \in \mathbb{T}(x^0, y^0; \beta)$. Lemma 2, together with Theorem 3, yields $x^*_{T_0^c} = 0$. Meanwhile, following from (26) in Algorithm 1, we also have $x^1_{T_0^c} = 0$. With some routine work, one can further obtain

$$\begin{aligned} & \|z^1 - z^*\| \\ \overset{(26)}{=} & \left\| \begin{bmatrix} x^1_{T_0} \\ y^1 \end{bmatrix} - \begin{bmatrix} x^*_{T_0} \\ y^* \end{bmatrix} \right\| \overset{(22)}{\leq} M^* \left\| G(x^0, y^0; T_0) \left(\begin{bmatrix} x^1_{T_0} \\ y^1 \end{bmatrix} - \begin{bmatrix} x^*_{T_0} \\ y^* \end{bmatrix} \right) \right\| \\ \leq & \frac{M^* L_2}{2} \left\| z^0 - z^* \right\|^2 \leq \frac{\left\| z^0 - z^* \right\|}{2} < \frac{\delta}{2}. \end{aligned} \tag{28}$$

Similar reasoning allows us to sequentially get

$$\|z^{k+1} - z^*\| \leq \frac{M^* L_2}{2} \left\| z^k - z^* \right\|^2 \text{ and } z^k \in \mathcal{N}_{\mathbb{S}} \left(z^*, \frac{\delta}{2^k} \right). \tag{29}$$

Hence $\lim\limits_{k \to \infty} z^k = z^*$ with quadratic convergence rate.

(ii) Similar to the case $k = 1$, Lemma 2 and Theorem 3 also yield $x^*_{T^c_{k+1}} = 0$ for any $k \geq 1$ from (29). After basic manipulations, we have

$$\|F_\beta(z^{k+1}; T_{k+1})\|^2 \overset{(28)}{\leq} (L_1^2 + 1) \left(\frac{M^* L_2}{2} \right)^2 \|z^k - z^*\|^4. \tag{30}$$

Additionally, the index set property in Lemma 2 ensures

$$T_k \in \mathbb{T}(z^k; \beta) \subseteq \mathbb{T}(z^*; \beta) = \mathcal{J}_s(x^*), \ \forall \ k \geq 0$$

which further leads to $F(z^*; T_k) = 0$ from Theorem 3, and the nonsingularity of $\nabla_z F(z^*; T_k)$ from Theorem 4. Thus, $\lambda_H > 0$ and

$$\begin{aligned} \|z^k - z^*\|^2 &\leq \frac{1}{\lambda_H} \|\nabla_z F(z^*; T_k)(z^k - z^*)\|^2 \\ &\leq \frac{2}{\lambda_H} \|\nabla_z F(z^*; T_k)(z^k - z^*) + o(\|z^k - z^*\|)\|^2 \\ &= \frac{2}{\lambda_H} \|F(z^k; T_k)\|^2, \end{aligned} \tag{31}$$

where the last equality is from $F(z^*; T_k) = 0$. Combining with (30), we have

$$\|F(z^{k+1}; T_{k+1})\| \leq \frac{M^* L_2 \sqrt{L_1^2 + 1}}{2} \|z^k - z^*\|^2 \leq \frac{M^* L_2 \sqrt{L_1^2 + 1}}{\lambda_H} \|F(z^k; T_k)\|^2.$$

(iii) Denote $\zeta_k := \max\limits_{j \in T_k^c} \{\max\{|(\nabla_x L(x^k, y^k))_j| - |x^k|_{(s)}/\beta, 0\}$, for any given $k \geq 1$. We claim that

$$\zeta_k \leq L_1 \|z^k - z^*\|, \quad \forall k \geq 1. \tag{32}$$

For each given $k \geq 1$, we consider the following two cases.

Case I: If $\|x^*\|_0 < s$, then $\nabla_x L(x^*, y^*) = 0$. It follows from the definition of ζ_k that

$$\zeta_k \le \|(\nabla_x L(x^k, y^k))_{T_k^c}\| \le \|\nabla_x L(x^k, y^k) - \nabla_x L(x^*, y^*)\| \le L_1 \|z^k - z^*\|.$$

Case II: If $\|x^*\|_0 = s$, we have $\mathbb{T}(z^k; \beta) = \{\Gamma_k\} = \{\Gamma^*\}$ from Lemma 2. Thus, for any $k \ge 1$, $T_k = \Gamma^*$ and $(\nabla_x L(x^*, y^*))_{T_k} = 0$. Besides, since $|x^k|_{(s)} > 0$, there exists $i_k \in T_k$ such that $|x^k_{i_k}| = |x^k|_{(s)}$. For any $k \ge 1$, it follows from the definition of $\mathbb{T}(x^k, y^k; \beta)$ that for any $j \in T_k^c = (\Gamma^*)^c$,

$$\begin{aligned}
\beta|(\nabla_x L(x^k, y^k))_j| &= |x^k_j - \beta(\nabla_x L(x^k, y^k))_j| \\
&\le \min_{i \in T_k} \{|x^k_i - \beta(\nabla_x L(x^k, y^k))_i|\} \\
&\le |x^k_{i_k} - \beta(\nabla_x L(x^k, y^k))_{i_k}| \\
&\le |x^k|_{(s)} + \beta \|(\nabla_x L(x^k, y^k))_{T_k}\|. \qquad (33)
\end{aligned}$$

This implies that $\max\{|(\nabla_x L(x^k, y^k))_j| - |x^k|_{(s)}/\beta, 0\} \le \|(\nabla_x L(x^k, y^k))_{T_k}\|$, $\forall j \in T_k^c$, which means $\zeta_k \le \|(\nabla_x L(x^k, y^k))_{T_k}\|$. Together with $(\nabla_x L(x^*, y^*))_{T_k} = 0$, we have

$$\zeta_k \le \|(\nabla_x L(x^k, y^k))_{T_k} - (\nabla_x L(x^*, y^*))_{T_k}\| \le L_1 \|z^k - z^*\|, \quad \forall k \ge 1.$$

This shows the claim in (32). Combining with (30) and (28), we further get that for $k \ge 1$,

$$\begin{aligned}
\eta_\beta(x^k, y^k; T_k) &= \|F(z^k; T_k)\| + \zeta_k \\
&\le \frac{M^* L_2 \sqrt{L_1^2 + 1}}{2} \|z^{k-1} - z^*\|^2 + \frac{M^* L_2 L_1}{2} \|z^{k-1} - z^*\|^2 \\
&\le M^* L_2 \sqrt{L_1^2 + 1} \|z^{k-1} - z^*\|^2. \qquad (34)
\end{aligned}$$

In addition, by virtue of (29), we obtain $\eta_\beta(x^k, y^k; T_k) \le \frac{\delta^2 M^* L_2 \sqrt{L_1^2+1}}{2^{2k-2}}$. To meet the stopping criterion $\eta_\beta(x^k, y^k; T_k) \le \epsilon$ in LNA, it suffices to have $\frac{\delta^2 M^* L_2 \sqrt{L_1^2+1}}{2^{2k-2}} \le \epsilon$, which leads to the bound of k as desired. This completes the proof. □

Remark 3 The local convergence result is very much in the form of the classical Newton method. However, we would like to point out one more restriction than the common requirement of the initial point being close enough to the interested stationary point. This hidden restriction is that we have to choose a proper β a priori, which defines the stationary point. For general cases of (1), it is difficult to give an estimate of β. However, if $f(x)$ is strongly convex and the Lagrangian multipliers are all bounded, then we may estimate a bound for β, which has been detailed in Remark 1 (iii). Additionally, for the common requirement of the initial point, a very useful consequence from the above proof is that $T_k \in \mathbb{T}(x^*, y^*; \beta)$ holds automatically for all $k \ge 0$ if the initial point (x^0, y^0) is chosen to be close enough to the strong β-Lagrangian stationary

point. This means that changing T_k becomes less often when the iterate gets close to the stationary point and it consequently explains strong performance of the algorithm for the two classes of problems tested below.

5 Applications

Two selected SNP problems arising from some important applications are considered to demonstrate the effectiveness of our proposed Lagrange-Newton algorithm.

5.1 Compressed Sensing

Compressed sensing (CS) [13] has been widely applied in signal and image processing [14], machine learning [46], and statistics [29]. A more general framework is considered with some noise-free observations being allowed and added as hard constraints into the standard CS model, taking the form of

$$\min \ \frac{1}{2}\|Ax-b\|^2, \ \ \text{s.t. } Cx=d, \ x\in\mathbb{S}, \tag{35}$$

where $A\in\mathbb{R}^{(p-m)\times n}$, $C\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^{p-m}$ and $d\in\mathbb{R}^m$. Set

$$f(x)=\frac{1}{2}\|Ax-b\|^2 \ \text{ and } \ h(x)=Cx-d.$$

The Lagrangian function of (35) is

$$L(x,y)=\frac{1}{2}\|Ax-b\|^2-y^\top(Cx-d),$$

for any $x\in\mathbb{S}$ and $y\in\mathbb{R}^m$. Direct calculations lead to

$$\begin{cases} \nabla h(x)=C, \ \nabla_x L(x,y)=A^\top(Ax-b)-C^\top y, \\ \nabla^2 f(x)=\nabla^2_{xx}L(x,y)=A^\top A, \ \nabla^2 h(x)=0, \ \nabla^2 L(x,y)=\begin{bmatrix} A^\top A & -C^\top \\ -C & 0 \end{bmatrix}. \end{cases} \tag{36}$$

Since $\nabla^2 f(\cdot)$ and $\nabla^2 h(\cdot)$ are constant, Assumption 3 holds automatically everywhere. To ensure Assumptions $1'$ and 2 hold, we introduce the following assumption on the input matrices A and C.

Assumption 4 For any index set $T\in\mathcal{J}_s$, A_T is full column rank and C_T is full row rank.

Suppose that Assumption 4 holds. Note that for any $(x;y)\in\mathbb{R}^{n+m}$ and $\beta>0$, $\mathbb{T}(x,y;\beta)\subseteq\mathcal{J}_s$. Together with $\mathrm{rank}(\nabla_T h(x))=\mathrm{rank}(C_T)$, we can conclude that Assumption $1'$ holds everywhere once C_T is full row rank for all $T\in\mathcal{J}_s$. Similarly, since $\big(\nabla^2_{xx}L(x,y)\big)_{T,T}=A_T^\top A_T$, it is positive definite in the entire space $\mathbb{R}^s$ once A_T is full column rank. Thus, Assumption 2 follows.

It is worth mentioning that Assumption 4 is actually a mild condition for problem (35). Indeed, the full column rankness of A_T is the so-called s-regularity introduced by Beck and Eldar [1] which has been widely used in the CS community, and limiting the number of hard constraints will make the full row rankness of C_T accessible (here m is no more than s and hence $s+m \leq 2s$). Under Assumption 4, we have the following optimality conditions for problem (35).

Proposition 1 *Assume that the feasible set of problem* (35) *is nonempty and Assumption 4 holds. Then the optimal solution set S^*_{cs} of* (35) *is nonempty. Furthermore, for any strong β-Lagrangian stationary point x^* of* (35)*, it is either a strictly local minimizer if $\|x^*\|_0 = s$ or a global minimizer otherwise.*

Proof. The nonemptiness of S^*_{cs} follows from the Frank-Wolfe Theorem and the observation $\mathbb{S} = \cup_{J\in\mathcal{J}_s}\mathbb{R}^n_J$. The rest of the assertion follows from Theorem 2 and [31, Theorem 4.2]. □

With the above optimality result, we can apply LNA to solve (35) efficiently, since the linear system in each iteration is of size no more than $2s \times 2s$ and the algorithm will have a fast quadratic convergence rate, as stated in the following proposition.

Proposition 2 *Suppose that Assumption 4 holds. For given $\beta > 0$, let x^* be a strong β-Lagrangian stationary point of* (35) *with y^*. Suppose that the initial point z^0 of sequence $\{z^k\}$ generated by LNA satisfies $z^0 \in \mathcal{N}_S(z^*, \delta)$, where $\delta = \min\{\tilde{\delta}^*, 1\}$ and $\tilde{\delta}^*$ is defined as in* (24)*. Then for any $k \geq 0$,*

(i) $\lim\limits_{k\to\infty} z^k = z^*$ *with quadratic convergence rate, i.e.,* $\|z^{k+1} - z^*\| \leq \dfrac{1}{2}\|z^k - z^*\|^2$.

(ii) LNA terminates with accuracy ϵ when $k \geq \left\lceil \dfrac{\log_2\left(4\delta^2\sqrt{\|[A^\top A, -C^\top]\|^2+1}/\epsilon\right)}{2} \right\rceil$.

Proof. Since $\nabla^2 L(x, y)$ is constant and hence (15) holds everywhere for any $L_2 > 0$. Thus, take $L_2 = 1/M^*$ with M^* defined as in (24). Similarly, from (36), we also have (14) at any $z \in \mathbb{R}^{n+m}$ with $L_1 = \|[A^\top A, -C^\top]\|$. By employing Theorem 6, we can obtain the desired assertions. □

5.2 Sparse High-Order Portfolio Selection

In financial markets, returns have often been found to be skewed and extreme events frequently observed can be measured by skewness and kurtosis. Thus, based on Markowitz's mean-variance (MV) portfolio model, several studies consider the high-order portfolio selection with only a limited number of assets, i.e., the mean-variance-skewness-kurtosis model with cardinality constraint (MVSKC). Suppose that $\tilde{r} \in \mathbb{R}^n$ is the return vector of n assets and

$x \in \mathbb{R}^n$ is the vector of portfolio weights. The MVSKC model, which is first introduced in [7], takes the form of

$$\begin{array}{ll} \min & -\lambda_1 x^\top \mu + \lambda_2 x^\top \Sigma x - \lambda_3 x^\top \Phi(x \otimes x) + \lambda_4 x^\top \Psi(x \otimes x \otimes x), \\ \text{s.t.} & e^\top x = 1,\ \|x\|_0 \le s. \end{array} \tag{37}$$

Here $\mu = E(\tilde{r})$ is the mean return vector with $E(\cdot)$ being the expectation operator, $\Sigma = E(rr^\top)$ is the covariance matrix, $\Phi = E(r(r^\top \otimes r^\top))$ is the co-skewness, $\Psi = E(r(r^\top \otimes r^\top \otimes r^\top))$ is the co-kurtosis matrix with $r := \tilde{r} - \mu$ being the centred return vector, $\otimes$ the Kronecker product, $\lambda_i > 0$, $i = 1, \ldots, 4$ are parameters to balance the four moments of the portfolio return, and e is the vector of all ones. Set

$$f(x) = -\lambda_1 x^\top \mu + \lambda_2 x^\top \Sigma x - \lambda_3 x^\top \Phi(x \otimes x) + \lambda_4 x^\top \Psi(x \otimes x \otimes x),\ h(x) = e^\top x - 1.$$

It is noted that the objective function f is nonconvex and twice continuously differentiable. The corresponding Lagrangian function associated with problem (37) is

$$L(x, y) = f(x) - y(e^\top x - 1), \quad \forall x \in \mathbb{S}, \quad y \in \mathbb{R}.$$

Utilizing Lemma 1 in [37], we have

$$\begin{cases} \nabla h(x) = e^\top,\ \nabla^2 h(x) = 0 \\ \nabla_x L(x, y) = -\lambda_1 \mu + 2\lambda_2 \Sigma x - 3\lambda_3 \Phi(x \otimes x) + 4\lambda_4 \Psi(x \otimes x \otimes x) - ye, \\ \nabla^2 f(x) = \nabla^2_{xx} L(x, y) = 2\lambda_2 \Sigma - 6\lambda_3 \Phi(I \otimes x) + 12\lambda_4 \Psi(I \otimes x \otimes x), \\ \nabla^2 L(x, y) = \begin{bmatrix} \nabla^2 f(x) & -e \\ -e^\top & 0 \end{bmatrix}. \end{cases} \tag{38}$$

It is easy to verify that Assumption 1′ holds for any $x \in \mathbb{R}^n$. Since $\nabla^2 f(\cdot)$ is continuously differentiable and $\nabla^2 h(\cdot)$ is constant, Assumption 3 automatically holds near x^*. To make the required Assumption 2 hold for problem (37), we introduce the following assumption.

Assumption 5 $\lambda_i > 0$, $i = 1, \ldots, 4$, satisfy $4\lambda_4(2\lambda_2 - 1) > 3\lambda_3^2$, and Σ is positive definite restricted to the set $\{d \in \mathbb{R}^n : e^\top d = 0\}$.

Learning from (38), for any $d \in \{d \in \mathbb{R}^n : e^\top d = 0\}$, we have

$$\begin{aligned} d^\top \nabla^2_{xx} L(x, y) d &= d^\top (2\lambda_2 \Sigma - 6\lambda_3 \Phi(I \otimes x) + 12\lambda_4 \Psi(I \otimes x \otimes x)) d \\ &= d^\top E(2\lambda_2 - 6\lambda_3 r^\top x + 12\lambda_4 (r^\top x)^2) rr^\top) d \\ &\ge d^T E(rr^\top) d > 0, \end{aligned}$$

where the second equality is from the definitions of Σ, Φ and Ψ, and the first inequality is from Assumption 5. Thus, Assumption 2 is valid.

Note that the condition in Assumption 5 is mild in real-world instances of sparse portfolio, since the covariance matrix is always positive definite. Under Assumption 5, we have the following optimality conditions for problem (37) by Theorem 1 and [31, Theorem 4.2].

Proposition 3 *(i) Suppose that x^* is a local minimizer of (5), then there exists a unique $y^* \in \mathbb{R}^m$ such that for any $\beta \in (0, \hat{\beta})$, x^* is a strong β-Lagrangian stationary point, where $\hat{\beta}$ is defined as in (6).*

(ii) Assume that Assumption 5 holds and x^* is a strong β-Lagrangian stationary point of (37), then x^* is a strictly local minimizer of (37).

With the above optimality result, we can apply LNA to solve (37) efficiently. It is noteworthy that the computational cost per iteration is extremely low since we just need to handle $s+1$ linear equations in each iteration, and the algorithm will have a fast quadratic convergence rate as stated below.

Proposition 4 *Suppose that Assumption 5 holds. For given $\beta > 0$, let x^* be a strong β-Lagrangian stationary point of (37) with y^*. Then LNA for (37) has locally quadratic convergence rate as stated in Theorem 6.*

6 Numerical Experiments

This section reports numerical results of LNA on the compressed sensing problem and the sparse high-order portfolio selection with both synthetic and real data. All experiments are conducted by using MATLAB (R2018a) on a laptop of 8GB memory and Inter(R) Core(TM) i5 1.8Ghz CPU. We terminate our method at kth step if $\eta_\beta(x^k, y^k; T_k) \leq 10^{-6}$ where $\eta_\beta(x^k, y^k; T)$ is defined as (27) or k reaches 1000.

6.1 Compressed Sensing

The aim of this subsection is to compare LNA with six state-of-the-art methods for compressed sensing problem (35), including HTP[16][1], NIHT[6][2], GP[4][2], OMP [34,36][2], CoSaMP [28][3] and SP[11][3].

6.1.1 Testing examples

We generate the sensing matrix $\mathcal{A}$ in the same way as [40,45]. Each column of $\mathcal{A}$ is normalized to $\|\mathcal{A}_j\| = 1$ to make it consistent with the algorithms used in [16,6,4]. The true signal x^* and the measurement $\mathcal{B}$ are generated by the following pseudo MATLAB codes:

$$x^* = \texttt{zeros}(n,1),\ \Gamma = \texttt{randperm}(n),\ x^*(\Gamma(1:s)) = \texttt{randn}(s,1),\ \mathcal{B} = \mathcal{A}x^*.$$

[1] HTP is available at: *https://github.com/foucart/HTP.*

[2] NIHT, GP and OMP are available at *https://www.southampton.ac.uk/engineering/about/staff/tb1m08.page#software*. We use the version sparsity_0_5 in which NIHT, GP and OMP are called hard_l0_Mterm, greed_gp and greed_omp.

[3] CoSaMP and SP are available at: *http://media.aau.dk/null_space_pursuits/2011/07/a-few-corrections-to-cosamp-and-sp-matlab.html.*

We then randomly choose $m = \lceil 0.1s \rceil$ rows of $\mathcal{A}$ as C in (35). The rest part of $\mathcal{A}$ composes A in the objective function. See the following pseudo MATLAB code for details:

$$J = \texttt{randperm}(p),\ J_1 = J(1:m),\ J_2 = J(m+1:end);$$
$$A = \mathcal{A}(J_1),\ b = \mathcal{B}(J_1),\ C = \mathcal{A}(J_2),\ d = \mathcal{B}(J_2).$$

Example 1 (Gaussian matrix) Let $\mathcal{A} \in \mathbb{R}^{p\times n}$ be a random Gaussian matrix with each column being identically and independently generated from the standard normal distribution.

Example 2 (Partial DCT matrix) Let $\mathcal{A} \in \mathbb{R}^{p\times n}$ be a random partial discrete cosine transform (DCT) matrix generated by

$$\mathcal{A}_{ij} = \cos(2\pi(j-1)\psi_i), \quad i = 1,\ldots,p, \quad j = 1,\ldots,n$$

where ψ_i $(i = 1,\ldots,p)$ is uniformly and independently sampled from $[0,1]$.

6.1.2 Numerical comparisons

We set the maximum number of iterations and the tolerance as 1000 and 10^{-6}, respectively, in all of the six comparison methods mentioned above. The initializations are set to be $x^0 = 0$, $y^0 = 0$ and $\beta = 5/n$ for LNA. For comparison purpose, HTP, NIHT, GP, OMP, CoSaMP and SP as tested in this subsection are all initialized with the origin in their default setups.

We say a recovery of a method is successful if $\|x - x^*\| < 0.01\|x^*\|$, where x is the solution produced by the method. The corresponding success rate is defined as the percentage of the number of successful recovery instances over all trials.

Firstly, we run 500 independent trials with fixed $n = 256, p = \lceil n/4 \rceil$ at different sparsity levels s from 6 to 36. The corresponding success rates are illustrated in Fig. 1. One can observe that LNA always yields the highest success rate for each s under both Examples 1 and 2, while a lowest success rate is generated by GP. For instance, when $s = 20$ for Gaussian matrix, 85% successful recoveries are guaranteed by our method, much better than any other method, whose success rates are all less than 60%.

Next, we implement 500 independent trials by varying $p = \lceil rn \rceil$ in $r \in \{0.1, 0.12, \ldots, 0.3\}$ when $n = 256, s = \lceil 0.05n \rceil$. We report the results in Fig. 2, which indicates that the larger m is, the easier the problem becomes to be solved. Again, LNA outperforms the others in the success rate for each s, and GP still comes the last.

We now examine these algorithms with higher dimensions n between 5000 and 25000 with 50 trials when $p = \lceil n/4 \rceil, s = \lceil 0.01n \rceil, \lceil 0.05n \rceil$ in the framework of Example 1. The average absolute error $\|x - x^*\|$ and CPU time are presented in Table 1 and Table 2, respectively. One can see that the highest accurate recovery is obtained by LNA with the least CPU time for most cases. Although OMP and HTP render solutions as accurate as those by LNA when

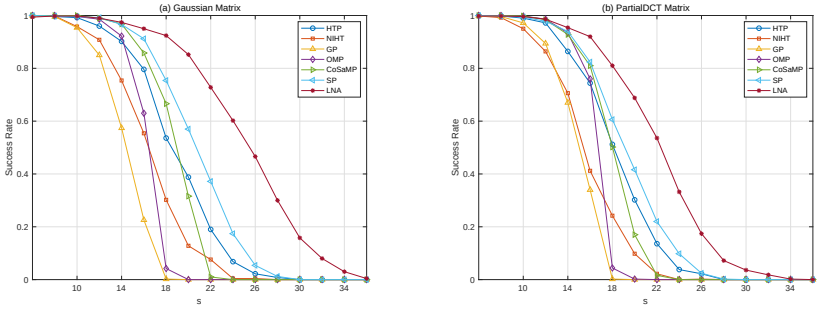

Fig. 1 Success rates. $n = 256, p = \lceil n/4 \rceil, s \in \{6, 8, \ldots, 36\}$.

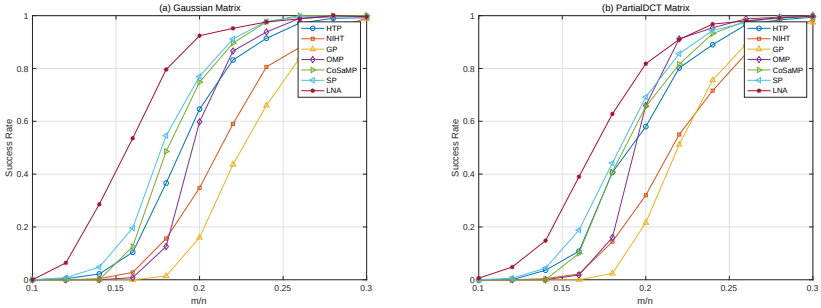

Fig. 2 Success rates. $n = 256, s = \lceil 0.05n \rceil, p = \lceil rn \rceil$ with $r \in \{0.1, 0.12, \ldots, 0.3\}$.

n is small, they present some shortcomings by comparing to LNA. In OMP, the accuracy cannot be guaranteed when n is large. Some inaccurate ones are observed particularly for $s = \lceil 0.05n \rceil$ and $n \geq 20000$, and this implies that OMP only works well when the solution is very sparse. On the other hand, the CPU time consumed by HTP increases drastically as n increases. For example, when $n = 25000$ and $s = \lceil 0.05n \rceil$, CPU is 5.99 seconds by LNA against 159.38 seconds by HTP. For NIHT, it is the fastest one among the six methods. However, its accuracy is much worse than others as it struggles at achieving the solutions with accuracy beyond the order 10^{-7}, while others can get as high as 10^{-14}. That is to say, the superiority of LNA becomes more obvious in the trade off of high accuracy and convergence speed with high dimensional data.

As stated in Theorem 6, LNA is a local method. Therefore, we conduct numerical experiments with randomly generated initial points for CS problems to see how the initial points would affect LNA. To proceed, we apply LNA to Examples 1 and 2 with $n = 10000, p = \lceil n/4 \rceil, s = \lceil 0.05n \rceil$. We run LNA with 50 different initial points, which are randomly generated from the uniform distribution, namely, $(x^0, y^0) \sim \mathrm{U}[0, 1]$. The absolute error $\|x - x^*\|$, the number of iterations and CPU time are plotted in Fig. 3, where the x-axis stands for the 50 initial points. One can see that all the results stabilize at a

Table 1 Average absolute error $\|x - x^*\|$ for Example 1.

s	n	LNA	HTP	NIHT	GP	OMP	CoSaMP	SP
	5000	2.71e-15	3.13e-15	2.03e-8	4.01e-15	2.78e-15	1.41e-14	1.41e-14
	10000	4.86e-15	5.70e-15	2.27e-8	7.04e-15	4.80e-15	2.15e-14	2.15e-14
$\lceil 0.01n \rceil$	15000	6.52e-15	7.39e-15	2.92e-8	1.06e-14	6.82e-15	2.98e-14	2.98e-14
	20000	8.87e-15	9.97e-15	4.37e-8	1.37e-14	9.34e-15	4.08e-14	4.08e-14
	25000	1.04e-14	1.21e-14	3.95e-8	1.72e-14	1.16e-14	4.44e-14	4.44e-14
	5000	1.14e-14	1.11e-14	1.63e-7	1.49e-14	1.08e-14	4.10e-14	4.10e-14
	10000	2.80e-14	2.28e-14	3.30e-7	2.97e-14	2.42e-14	7.75e-14	7.75e-14
$\lceil 0.05n \rceil$	15000	3.91e-14	3.70e-14	3.06e-7	5.02e-14	4.31e-14	1.10e-13	1.10e-13
	20000	5.22e-14	4.77e-14	4.03e-7	5.83e-14	5.15e-04	1.34e-13	1.34e-13
	25000	6.30e-14	6.12e-14	3.75e-7	7.74e-14	6.30e-04	1.82e-13	1.82e-13

Table 2 Average CPU time (in seconds) for Example 1.

s	n	LNA	HTP	NIHT	GP	OMP	CoSaMP	SP
	5000	0.06	0.62	0.21	1.92	0.30	0.57	0.04
	10000	0.25	4.02	0.83	14.07	2.28	0.28	0.18
$\lceil 0.01n \rceil$	15000	0.60	13.07	1.93	46.88	7.70	1.01	0.78
	20000	1.08	32.08	3.50	110.07	18.16	2.13	1.34
	25000	1.77	111.94	6.11	230.27	37.43	4.11	2.57
	5000	0.14	0.80	0.62	2.18	1.73	1.44	0.93
	10000	0.59	5.71	2.42	15.07	13.66	13.87	5.36
$\lceil 0.05n \rceil$	15000	1.50	18.90	5.47	50.41	46.82	47.41	18.23
	20000	3.03	49.10	11.24	118.35	111.40	127.37	55.21
	25000	5.99	159.38	17.49	239.70	174.21	217.07	96.38

certain level, which indicates that LNA is not sensitive to the choice of the initial points for CS problems.

6.2 Sparse High-Order Portfolio Selection

This subsection is devoted to comparing LNA with successive convex approximation algorithm (SCA)[37] in sparse high-order portfolio selection problem (37) on real data sets.

6.2.1 Testing examples

Example 3 (Portfolio data sets) The data sets used in our experiments are selected from the Standard and Poor's 500 (USA) (S&P 500 for short)[6]. Firstly, we randomly select 100 stocks from S&P 500 Index components and randomly choose 500 continuous trading days from 2012-12-01 to 2018-12-01. The selected data is then normalized to raise precision of the model, and the sample moments[7] are computed. For fair comparison, we adopt the choices of model

[6] *http://cran.r-project.org/web/packages/portfolioBacktest/vignettes*

[7] *http://www.mathworks.com/matlabcentral/fileexchange/47839-co_moments-m.*

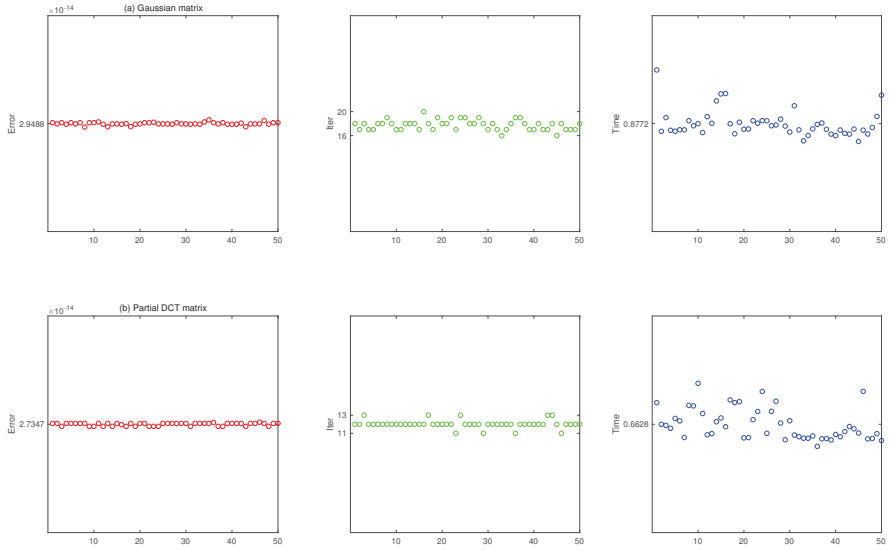

Fig. 3 Effects of initial points for Examples 1 and 2.

parameters in (37) from [7] with $\lambda_1 = 1,\ \lambda_2 = \xi/2,\ \lambda_3 = \xi(\xi+1)/6,\ \lambda_4 = \xi(\xi+1)(\xi+2)/24$, where $\xi = 5,\ 10$ is the risk aversion parameter. Direct calculations certify that these parameters satisfy the condition in Assumption 5. Additionally, the sparsity level s for $\xi = 5,\ 10$ will be varying among $\{5, 10, 15, 20, 25\}$ to generate a total of 10 testing instances.

6.2.2 Numerical comparisons

For portfolio data sets in Example 3, we find that different initial points lead to different output solutions. This is reasonable since the objective function is nonconvex and LNA is a locally convergent method. Note that there are various ways to find an initial point close to a strong β-Lagrangian stationary point. For instance, one may apply some first-order gradient descent methods or convex relaxation methods to obtain such an initial point. For simplicity, we initialize LNA with the origin $(x^0, y^0) = (0, 0)$ which is appropriate to our testing examples, and $\beta = 1$. For comparison purpose, SCA is called with the initial point $x^0 = 0$, along with other parameters as $\alpha = 0.2$, $\rho = 4 \times 10^{-3}, 3 \times 10^{-3}$ for $\xi = 5, 10$ respectively. The sparsity of a solution x generated by SCA will be recorded by $\hat{s} := \min\{t : \sum_{i=1}^{t} |x|_{(i)} \geq 0.99\|x\|_1\}$. Table 3 records s, $\hat{s}$, the objective function value (f-value) and CPU time when $\xi = 5,\ 10$ respectively.

Table 3 $\hat{s}$, f-value and CPU time (in seconds) for Example 3.

s	$\xi = 5$					$\xi = 10$				
	$\hat{s}$	f-value		CPU time		$\hat{s}$	f-value		CPU time	
		LNA	SCA	LNA	SCA		LNA	SCA	LNA	SCA
5	55	-1.11	-1.50	15.55	50.55	88	-0.70	-1.44	95.73	144.92
10	55	-3.08	-1.50	39.53	45.48	88	-1.55	-1.44	46.66	216.47
15	55	-2.81	-1.50	26.80	45.43	88	-1.78	-1.44	77.15	186.79
20	55	-4.27	-1.50	23.74	47.06	88	-2.80	-1.44	75.58	194.80
25	55	-4.35	-1.50	34.30	46.62	88	-3.01	-1.44	23.96	117.98

As one can see from Table 3, LNA outperforms SCA in computational time for all testing instances, and attains smaller f-value than that by SCA when $s > 5$. In particular, LNA provides s-sparse solutions while SCA fails. Furthermore, as s grows, f-value decreases by LNA and this indicates that LNA improves the solution quality as the search space becomes bigger. This is in contrast to almost no-improvement of f-value by SCA.

7 Conclusion

In this paper, we have designed a second-order greedy algorithm named the Lagrange-Newton Algorithm (LNA) for the sparse nonlinear programming (SNP) problem with sparsity and nonlinear equality constraints. The novelty is the introduction of the strong β-Lagrangian stationarity and Lagrangian equations. The resulting LNA has shown to be effective. It enjoys a locally quadratic convergence rate and low iterative complexity from the theoretical perspective, and good computational superiority from the numerical perspective.

There are some important issues left to be investigated. As LNA is a second-order local method heavily relying on the initial points in general, the first issue is whether a line search scheme would be incorporated into LNA, leading to global convergence. Another question is whether we can extend LNA to more general sparse optimization models with equality and inequality constraints. We leave these to our future research.

Acknowledgements We would like to thank AE and the two referees for their valuable comments, which have helped to shorten the paper and greatly improved its presentation. We also thank Dr. Shenglong Zhou of Imperial College, London for his great support on the numerical experiments.

References

1. Beck, A., Eldar, Y.C.: Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. SIAM Journal on Optimization **23**(3), 1480–1509 (2013)
2. Beck, A., Hallak, N.: On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. Mathematics of Operations Research **41**(1), 196–223 (2015)

3. Beck, A., Vaisbourd, Y.: The sparse principal component analysis problem: Optimality conditions and algorithms. Journal of Optimization Theory and Applications **170**(1), 119–143 (2016)
4. Blumensath, T., Davies, M.E.: Gradient pursuits. IEEE Transactions on Signal Processing **56**(6), 2370–2382 (2008)
5. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis **27**(3), 265–274 (2009)
6. Blumensath, T., Davies, M.E.: Normalized iterative hard thresholding: Guaranteed stability and performance. IEEE Journal of Selected Topics in Signal Processing **4**(2), 298–309 (2010)
7. Boudt, K., Lu, W., Peeters, B.: Higher order comoments of multifactor models and asset allocation. Finance Research Letters **13**, 225–233 (2015)
8. Červinka, M., Kanzow, C., Schwartz, A.: Constraint qualifications and optimality conditions for optimization problems with cardinality constraints. Mathematical Programming **160**(1), 353–377 (2016)
9. Chen, J., Gu, Q.: Fast Newton hard thresholding pursuit for sparsity constrained nonconvex optimization. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 757–766 (2017)
10. Chen, X., Ge, D., Wang, Z., Ye, Y.: Complexity of unconstrained $l_2 - l_p$ minimization. Mathematical Programming **143**(1-2), 371–383 (2014)
11. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal reconstruction. IEEE Transactions on Information Theory **55**(5), 2230–2249 (2009)
12. Donoho, D.L.: Compressed sensing. IEEE Transactions on Information Theory **52**(4), 1289–1306 (2006)
13. Elad, M.: Sparse and Redundant Representations. Springer, New York (2010)
14. Elad, M., Figueiredo, M.A., Ma, Y.: On the role of sparse and redundant representations in image processing. Proceedings of the IEEE **98**(6), 972–982 (2010)
15. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its Oracle properties. Journal of the American Statistical Association **96**(456), 1348–1360 (2001)
16. Foucart, S.: Hard thresholding pursuit: an algorithm for compressive sensing. SIAM Journal on Numerical Analysis **49**(6), 2543–2563 (2011)
17. Gao, J., Li, D.: Optimal cardinality constrained portfolio selection. Operations Research **61**(3), 745–761 (2013)
18. Gotoh, J.y., Takeda, A., Tono, K.: DC formulations and algorithms for sparse optimization problems. Mathematical Programming **169**(1), 141–176 (2018)
19. Han, S.P.: Superlinearly convergent variable metric algorithms for general nonlinear programming problems. Mathematical Programming **11**, 263–282 (1976)
20. Koh, K., Kim, S.J., Boyd, S.: An interior-point method for large-scale ℓ_1-regularized logistic regression. Journal of Machine Learning Research **8**, 1519–1555 (2007)
21. Kyrillidis, A., Becker, S., Cevher, V., Koch, C.: Sparse projections onto the simplex. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013, vol. 28, pp. 235–243 (2013)
22. Lu, Z.: Optimization over sparse symmetric sets via a nonmonotone projected gradient method. arXiv preprint arXiv:1509.08581 (2015)
23. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM Journal on Optimization **23**(4), 2448–2478 (2013)
24. Luo, Z., Sun, D., Toh, K.C., Xiu, N.: Solving the OSCAR and SLOPE models using a semismooth Newton-based augmented lagrangian method. Journal of Machine Learning Research **20**(106), 1–25 (2019)
25. Luo, Z.Q., Pang, J.S., Ralph, D.: Piecewise Sequential Quadratic Programming for Mathematical Programs with Nonlinear Complementarity Constraints, pp. 209–229. Springer (1998)
26. Misra, J.: Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Research **12**(7), 1112–1120 (2002)
27. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM Journal on Computing **24**(2), 227–234 (1995)
28. Needell, D., Tropp, J.A.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis **26**(3), 301–321 (2009)

29. Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B., et al.: A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statistical Science **27**(4), 538–557 (2012)
30. Pan, L., Luo, Z., Xiu, N.: Restricted Robinson constraint qualification and optimality for cardinality-constrained cone programming. Journal of Optimization Theory and Applications **175**(1), 104–118 (2017)
31. Pan, L., Xiu, N., Fan, J.: Optimality conditions for sparse nonlinear programming. Science China Mathematics **60**(5), 759–776 (2017)
32. Pan, L., Xiu, N., Zhou, S.: On solutions of sparsity constrained optimization. Journal of the Operations Research Society of China **3**(4), 421–439 (2015)
33. Pan, L., Zhou, S., Xiu, N., Qi, H.D.: Convergent iterative hard thresholding for sparsity and nonnegativity constrained optimization. Pacific Journal of Optimization **13**(2), 325–353 (2017)
34. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Proceedings of 27th Asilomar conference on signals, systems and computers, IEEE, pp. 40–44 (1993)
35. Powell, M.J.D.: Restart procedures for the conjugate gradient method. Mathematical Programming **12**, 241–254 (1977)
36. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on Information Theory **53**(12), 4655–4666 (2007)
37. Wang, J., Deng, Z., Zheng, T., So, A.M.C.: Sparse high-order portfolios via proximal DCA and SCA. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5425–5429 (2021)
38. Wilson, R.B.: A simplicial algorithm for concave programming. Ph.D. thesis, Graduate School of Business Administration, Harvard University (1963)
39. Xu, F., Lu, Z., Xu, Z.: An efficient optimization approach for a cardinality-constrained index tracking problem. Optimization Methods and Software **31**, 258–271 (2016)
40. Yin, P., Lou, Y., He, Q., Xin, J.: Minimization of ℓ_{1-2} for compressed sensing. SIAM Journal on Scientific Computing **37**(1), 536–563 (2015)
41. Yuan, X., Li, P., Zhang, T.: Gradient hard thresholding pursuit. Journal of Machine Learning Research **18**(166), 1–43 (2018)
42. Yuan, X., Liu, Q.: Newton greedy pursuit: A quadratic approximation method for sparsity-constrained optimization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 4122–4129 (2014)
43. Yuan, X., Liu, Q.: Newton-type greedy selection methods for ℓ_0-constrained minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2437–2450 (2017)
44. Zhou, S., Xiu, N., Qi, H.: Global and quadratic convergence of Newton hard-thresholding pursuit. Journal of Machine Learning Research **22**, 1–45 (2021)
45. Zhou, S., Xiu, N., Wang, Y., Kong, L., Qi, H.D.: A null-space-based weighted ℓ_1 minimization approach to compressed sensing. Information and Inference: A Journal of the IMA **5**(1), 76–102 (2016)
46. Zhou, T., Tao, D., Wu, X.: Manifold elastic net: a unified framework for sparse dimension reduction. Data Mining and Knowledge Discovery **22**(3), 340–371 (2011)
47. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational & Graphical Statistics **15**(2), 265–286 (2006)