Mathematical Programming
https://doi.org/10.1007/s10107-025-02266-5

FULL LENGTH PAPER

Series A

Composite optimization with indicator functions: stationary duality and a semismooth newton method

Penghe Zhang[1] · **Naihua Xiu**[2] · **Houduo Qi**[3]

Received: 28 October 2024 / Accepted: 23 July 2025

Abstract

Indicator functions of taking values of zero or one are essential to numerous applications in machine learning and statistics. The corresponding primal optimization model has been researched in several recent works. However, its dual problem is a more challenging topic that has not been well addressed. One possible reason is that the Fenchel conjugate of any indicator function is finite only at the origin. This work aims to explore the dual optimization for the sum of a strongly convex function and a composite term with indicator functions on positive intervals. For the first time, a dual problem is constructed by extending the classic conjugate subgradient property to the indicator function. This extension further helps us establish the equivalence between the primal and dual solutions. The dual problem turns out to be a sparse optimization with a ℓ_0 regularizer and a nonnegative constraint. The proximal operator of the sparse regularizer is used to identify a dual subspace to implement gradient and/or semismooth Newton iteration with low computational complexity. This gives rise to a dual Newton-type method with both global convergence and local superlinear (or quadratic) convergence rate under mild conditions. Finally, when applied to AUC maximization and sparse multi-label classification, our dual Newton method demonstrates satisfactory performance on computational speed and accuracy.

This work was supported by the National Key R&D Program of China (2023YFA1011100), the National Natural Science Foundation of China (12131004), and Hong Kong Research Grants Council Projects PolyU/15303124, PolyU/15309223.

✉ Houduo Qi
houduo.qi@polyu.edu.hk

Penghe Zhang
penghe.zhang@polyu.edu.hk

Naihua Xiu
nhxiu@bjtu.edu.cn

[1] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, China

[2] School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China

[3] Department of Data Science and Artificial Intelligence and Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

Published online: 06 August 2025

Keywords Composite optimization · Indicator function · Dual stationarity · Semismooth Newton method · Global convergence · Local superlinear convergence rate

Mathematics Subject Classification 90C26 · 49M37 · 65K10

1 Introduction

In this work, we aim to solve the following nonconvex composite optimization

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \mathcal{I}(A\mathbf{x}+\mathbf{b}), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a σ_f-strongly convex function (possibly nonsmooth), and $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are given data. For $\mathbf{u} = (u_1, \cdots, u_m)^\top$, we define $\mathcal{I}(\mathbf{u}) = \lambda \sum_{i=1}^m \mathbf{1}_{(0,\infty)}(u_i)$, where λ is a positive constant and $\mathbf{1}_{(0,\infty)}$ is the indicator function[1] on the positive interval (also called zero-one loss), defined as

$$\mathbf{1}_{(0,\infty)}(t) = \begin{cases} 1, & \text{if } t > 0, \\ 0, & \text{if } t \le 0. \end{cases}$$

Problem (1) arises from extensive applications, including multi-label classification [66], area under the receiver operating characteristic curve (AUC) [23, 28], maximum ranking correlation [26, 33], to mention a few. The strongly convex function f often serves as a regularizer in these tasks, including squared ℓ_2 norm for margin maximization (e.g. [18, 59]) or (group) elastic net for feature selection (e.g. [48, 63, 71]). Problem (1), in particular, covers the composite ℓ_0-norm: $\|B\mathbf{x}+\mathbf{b}\|_0 = \mathcal{I}(B\mathbf{x}+\mathbf{b}) + \mathcal{I}(-B\mathbf{x}-\mathbf{b})$. The matrix A becomes $[B; -B]$ (Matlab notation) and this implies that the rows of A are dependent. Therefore, it is essential we do not assume that A has a full-row rank property. And we do not assume that f is smooth.

When f is smooth, several progresses have been made in solving the primal nonconvex composition problem involving the indicator function or a lower semi-continuous (l.s.c) function. These methods include, but are not limited to the Newton method [70], augmented Lagrangian-based method [10, 11, 38, 61], lifted reformulation [19, 27], and complementarity reformulation [22, 32]. However, the composition of a nonconvex, discontinuous indicator function and a large-scale matrix brings significant difficulty in both convergence and computation. For example, certain regularity conditions (such as a full row-rank matrix A) are often required for the convergence analysis of the aforementioned algorithms. Moreover, it still remains a difficult task to globalize a primal Newton method while retaining local quadratic convergence. Finally, we should stress that the function f in (1) is not necessarily smooth. This together

[1] This paper adopts the terminology "indicator function" from the statistical learning theory [60, Page 19], differing from the usage in optimization where the function takes 0 on a given set and ∞ elsewhere. It is also called the characteristic function [64] in other fields.

with the combinatorial nature of $\mathcal{I}$ poses significant barriers to designing a globally convergent algorithm for (1).

This paper offers an alternative methodology, strongly motivated by the classical duality theory of convex optimization. We will propose a dual reformulation of (1) with a conducive structure for numerical computation. In the following, we first explain the motivation that leads to a framework for constructing the dual problem. We then briefly summarize the main contributions of the paper.

1.1 Motivation from convex optimization

Let us recall the Fenchel duality in convex optimization (see [56, Chap. 31] and [4, 14]):

$$\inf_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) + \varphi(A\mathbf{x}+\mathbf{b}) \right\} = \sup_{\mathbf{z}\in\mathbb{R}^m} \left\{ -f^*(-A^\top \mathbf{z}) + \langle \mathbf{b}, \mathbf{z}\rangle - \varphi^*(\mathbf{z}) \right\}. \tag{2}$$

where both f and φ are closed, proper, and convex, and f^* and φ^* are their respective conjugate functions. For simplicity of discussion, we omit the relative interior conditions for the duality to hold. The dual problem is often advantageous to solve due to a few features. Firstly, the matrix A has been separated from the possibly nonsmooth function φ while composited with f^*. Secondly, f^* has Lipschitzian continuous gradient when f is strongly convex. Thirdly, in many applications, φ^* has a closed-form proximal operator that would allow proximal gradient method to be developed, see [4]. An impressive progress has been made on applying the dual approach to LASSO-type problems and highly efficient semismooth Newton algorithms are developed, see [39, 68] and the references therein. Recently, [24] proposed the semismoothness* of a set-valued mapping, which is an improvement of classic semismoothness. For minimization of a prox-regular function, [46] designed a generalized Newton method based on the second-order subdifferential [42] and proved its local superlinear convergence under semismoothness* and tilt stability assumptions. This method is further developed for solving composite optimization [34–36] and difference programming [1] under some assumptions. A more comprehensive introduction to this method can be found in the monograph [45]. These works motivated us to investigate whether a dual approach is viable for (1) and whether a generalized Newton method can be constructed.

However, if we blindly apply the dual formulation (2) to (1) with $\varphi = \mathcal{I}$, we would not gain much. This is because the conjugate function $\mathcal{I}^*$ only takes finite value at $\{0\}$:

$$\mathcal{I}^*(\mathbf{z}) := \sup_{\mathbf{u}\in\mathbb{R}^m} \left\{ \langle \mathbf{z}, \mathbf{u}\rangle - \mathcal{I}(\mathbf{u}) \right\} = \begin{cases} 0, & \text{if } \mathbf{z} = 0, \\ \infty, & \text{otherwise.} \end{cases}$$

This means that the dual problem would have only one feasible point $\mathbf{z} = 0$ and everything we care about in duality breaks down.

Now let us take a step back to recall one essential property for the conjugate function pair (φ, φ^*) when φ is proper, closed, and convex:

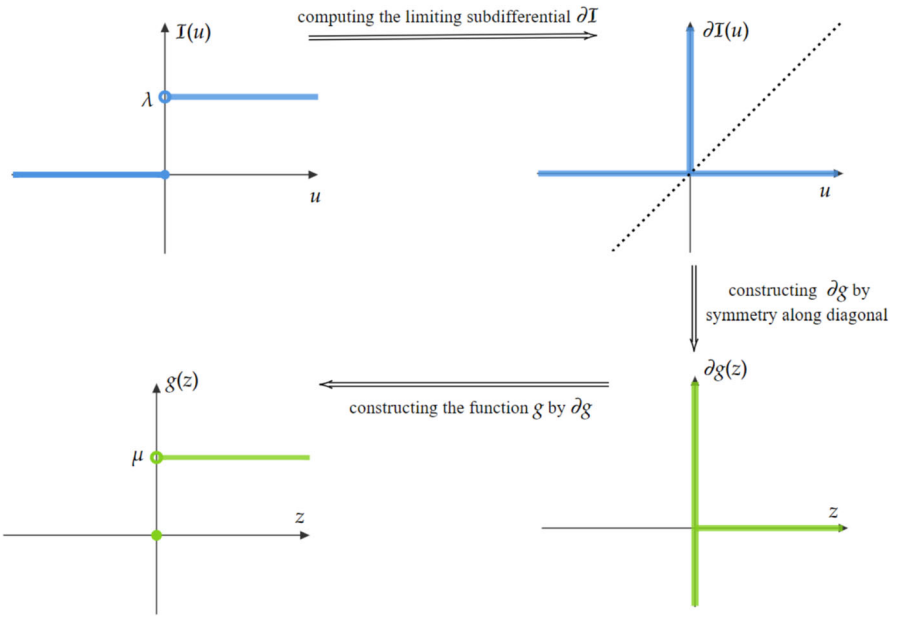

Fig. 1 Constructing a function g such that $z \in \partial\mathcal{I}(u)$ if and only if $u \in \partial g(z)$ for any given $z, u \in \mathbb{R}$

$$\mathbf{z} \in \partial\varphi(\mathbf{u}) \iff \mathbf{u} \in \partial\varphi^*(\mathbf{z}),$$

where $\partial\varphi(\cdot)$ is the subdifferential of $\varphi(\cdot)$ in convex analysis [56]. We call this property the stationary duality property. We would like to find a function $g(\cdot)$ that plays the role of $\mathcal{I}^*(\cdot)$ and satisfies the stationary duality property:

$$\mathbf{z} \in \partial\mathcal{I}(\mathbf{u}) \iff \mathbf{u} \in \partial g(\mathbf{z}), \tag{3}$$

where $\partial\mathcal{I}(\cdot)$ is the limiting subdifferential in variational analysis [44, 57]. In the meantime, we require the function $g(\mathbf{z})$ to inherit some useful information from the primal Problem (1). For the one-dimensional case, the construction of such function $g(z)$ is demonstrated in Fig. 1.

1.2 Stationary duality and main contributions

Extension of the one-dimensional case to the multi-dimensional case leads to the following function:

$$g(\mathbf{z}) := \mu\|\mathbf{z}\|_0 + \boldsymbol{\delta}_+(\mathbf{z}), \tag{4}$$

where $\mu > 0$ is a parameter and $\boldsymbol{\delta}_+(\cdot)$ is defined as follows:

$$\boldsymbol{\delta}_+(\mathbf{z}) = \begin{cases} 0, & \text{if } \mathbf{z} \geq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Using the characterizations in (6) and (7), it is easy to verify that the stationary dual relationship (3) holds for $g(\cdot)$. This gives rise to the stationary dual problem:

$$\max_{\mathbf{z}\in\mathbb{R}^m} \; -h(\mathbf{z}) - g(\mathbf{z}) \qquad \text{with} \quad h(\mathbf{z}) := f^*(-A^\top \mathbf{z}) - \langle \mathbf{z}, \mathbf{b}\rangle. \tag{5}$$

Although we cannot expect the perfect duality in convex optimization, we are able to establish the following important results.

(i) There is one-to-one correspondence between local minimizers of the primal problem (1) and that of the dual problem (5). See Thm. 1 for the detailed formula for computing the corresponding solution from the other. This is achieved through the study of proximal-type ("P-" for short) stationary point of (5) and its correspondence to the KKT (Karush-Kuhn-Tucker) point of a local convex reformulation of (1) at a local minimizer. This result is the basis for further numerical development.

(ii) Since f is strongly convex, its conjugate f^* has Lipschitzian continuous gradient and well-defined generalized Jacobian. More importantly, the nonsmooth function g is a sparse regularizer with a closed-form proximal operator, which helps us identify a dual subspace to perform low-complexity iteration. In this reduced subspace, we can first implement a gradient step and then accelerate it by a semismooth Newton step. In this process, easy-to-check conditions are proposed to ensure the dual objective function in (5) ascent. These steps form the basic framework of our subspace gradient semismooth Newton (SGSN) method, see Alg. 1.

(iii) In addition to the low per-iteration complexity, our SGSN has global convergence with local superliner (or quadratic) rate. If the generated sequence is bounded and the dual objective function in (5) satisfies Polyak-Łojasiewicz-Kurdyka (PŁK) property, then the sequence converges to a P-stationary point of (5) (Thm. 2). Thereby, a local minimizer of the primal Problem (1) is found. This result does not require the regularity condition that the matrix A has full-row rank. If we further assume that the gradient function ∇f^* is (strongly) semismooth and a second-order sufficient condition holds, then the sequence converges to a P-stationary point of (5) with local superlinear (or quadratic) rate. To the best of our knowledge, SGSN is the first algorithm that enjoys both global convergence and local superlinear rate for Problem (1) (Thm. 3). Finally, we demonstrate the use of SGSN on two important applications: AUC maximization and the sparse multi-label classification. The experiments show that SGSN has satisfactory performance on computational speed and solution accuracy when benchmarking against several competitive algorithms.

1.3 Organization

The paper is organized as follows. Some frequently used symbols and notations are introduced in Section 2. The one-to-one correspondence between solutions of primal and dual problems is established in Section 3. To find a dual solution, we describe the framework of SGSN and prove its global convergence with a local superlinear (or

quadratic) rate in Section 4. The numerical experiments on AUC maximization and multi-label classification are conducted in Section 5. We conclude the paper in Section 6 with some discussion on possible future research.

2 Preliminaries

In this part, we first describe the notation used in the paper. We then recall the definition of the limiting subdifferential and apply it to the function $g(\cdot)$ in (4). We also characterize its proximal operator. Finally, we introduce some function classes.

2.1 Notation

The boldfaced lowercase letter $\mathbf{x} \in \mathbb{R}^n$ denotes a column vector of size n and $\mathbf{x}^\top$ is its transpose. Let x_i or $[\mathbf{x}]_i$ denote the ith element of $\mathbf{x}$. We let $\mathbb{R}^n_+$ denote the non-negative orthant in $\mathbb{R}^n$. The norm $\|\cdot\|$ denotes the ℓ_2 norm of $\mathbf{x}$ or the Frobenius norm for a matrix A, and thus we always have $\|A\mathbf{x}\| \le \|A\|\|\mathbf{x}\|$. A neighborhood of $\mathbf{x}^* \in \mathbb{R}^n$ is often denoted as $\mathcal{N}(\mathbf{x}^*)$, while $\mathcal{N}(\mathbf{x}^*, \delta) := \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}^*\| < \delta\}$ when the radius $\delta > 0$ is given. We denote I as the identity matrix of appropriate dimension. Let $\mathbf{e}$ represent a vector of all 1's with dimension implied by the context. We denote ceiling and floor functions as $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ respectively.

Let $[m]$ denote the set of indices $\{1, \ldots, m\}$. For a subset $\Gamma \subset [m]$, $\overline{\Gamma}$ consists of those indices of $[m]$ not in Γ and $|\Gamma|$ denotes the number of elements in Γ (cardinality of Γ). For vector $\mathbf{z} \in \mathbb{R}^m$ (resp. matrix $A \in \mathbb{R}^{m\times n}$), $\mathbf{z}_\Gamma$ (resp. $A_{\Gamma:}$) denotes the subvector of $\mathbf{z}$ indexed by Γ (resp. the submatrix of A with rows indexed by Γ). Particularly, for a differentiable function h, we denote $\nabla_\Gamma h(\mathbf{z}) := (\nabla h(\mathbf{z}))_\Gamma$. For a symmetric matrix $H \in \mathbb{R}^{m\times m}$, H_Γ is the submatrix with both rows and columns indexed by Γ. Meanwhile, the largest and smallest eigenvalue of H are denoted by $\boldsymbol{\lambda}_{\max}$ and $\boldsymbol{\lambda}_{\min}$ respectively. Given vector $\mathbf{z} \in \mathbb{R}^m$, we denote its support set by $\mathcal{S}(\mathbf{z}) := \{i \in [m] : z_i \neq 0\}$, where ":=" means "define". For a set $\Omega \subset \mathbb{R}^m$, $\mathrm{co}\{\Omega\}$ is the convex hull of Ω.

2.2 Limiting subdifferential

The major reference for this part is [43, 44, 57]. Let us consider a proper, lower semicontinuous (l.s.c.) and bounded-below function $\varphi : \mathbb{R}^m \to (-\infty, +\infty]$. The regular subdifferential of φ at $\overline{\mathbf{z}} \in \mathrm{dom}\,\varphi := \{\mathbf{x} \in \mathbb{R}^m : \varphi(\mathbf{x}) < \infty\}$ is defined by

$$\widehat{\partial}\varphi(\overline{\mathbf{z}}) := \left\{ \mathbf{v} \in \mathbb{R}^m : \liminf_{\substack{\mathbf{z}\to\overline{\mathbf{z}} \\ \mathbf{z}\neq\overline{\mathbf{z}}}} \frac{\varphi(\mathbf{z}) - \varphi(\overline{\mathbf{z}}) - \langle \mathbf{v}, \mathbf{z} - \overline{\mathbf{z}}\rangle}{\|\mathbf{z} - \overline{\mathbf{z}}\|} \ge 0 \right\}.$$

For $\overline{\mathbf{z}} \notin \mathrm{dom}\,\varphi$, we set $\partial\varphi(\overline{\mathbf{z}}) = \varnothing$. The limiting or Mordukhovich subdifferential of φ at $\overline{\mathbf{z}} \in \mathbb{R}^m$ is defined by

$$\partial\varphi(\overline{\mathbf{z}}) := \{\mathbf{v} \in \mathbb{R}^m : \exists \mathbf{z}^k \to \overline{\mathbf{z}},\ \varphi(\mathbf{z}^k) \to \varphi(\overline{\mathbf{z}}) \text{ and } \widehat{\partial}\varphi(\mathbf{z}^k) \ni \mathbf{v}^k \to \mathbf{v}\}.$$

It was initially proposed in [41] in an equivalent formulation. Particularly, the limiting subdifferential of $\mathcal{I}(\cdot)$ takes the following form [70]:

$$\partial\mathcal{I}(\mathbf{u}) = \left\{\mathbf{v} \in \mathbb{R}^m \;\middle|\; v_i \begin{cases} \geq 0, \text{ if } u_i = 0 \\ = 0, \text{ if } u_i \neq 0, \end{cases}, \quad i \in [m] \right\}. \tag{6}$$

For given $\tau > 0$, the proximal operator of φ is defined by

$$\mathrm{Prox}_{\tau\varphi(\cdot)}(\mathbf{u}) := \arg\min_{\mathbf{z}\in\mathbb{R}^n} \left\{\varphi(\mathbf{z}) + \frac{1}{2\tau}\|\mathbf{z} - \mathbf{u}\|^2\right\}.$$

The following result reveals the relationship of $\partial\mathcal{I}$, ∂g and $\mathrm{Prox}_{\tau g(\cdot)}(\cdot)$, and they will be utilized to establish the equivalence of solutions between the primal and dual problems. Although it is straightforward, a proof is included for easy reference.

Lemma 1 *Given points $\mathbf{z} \in \mathbb{R}^m_+$ and $\mathbf{u} \in \mathbb{R}^m$, the following results hold.*

(i) The limiting subdifferential of g at $\mathbf{z}$ can be computed by

$$\partial g(\mathbf{z}) = \left\{\mathbf{v} \in \mathbb{R}^m : v_i \in \begin{cases} \mathbb{R}, & \text{if } z_i = 0 \\ \{0\}, & \text{if } z_i > 0 \end{cases}\right\}. \tag{7}$$

(ii) The proximal operator of g with $\tau > 0$ can be represented as

$$\Big[Prox_{\tau g(\cdot)}(\mathbf{u})\Big]_i = \begin{cases} 0, & u_i < \sqrt{2\tau\mu}, \\ \{0, u_i\}, & u_i = \sqrt{2\tau\mu}, \\ u_i, & u_i > \sqrt{2\tau\mu}, \end{cases} \quad i = 1, \ldots, m. \tag{8}$$

(iii) $\mathbf{z} \in \partial\mathcal{I}(\mathbf{u})$ if and only if $\mathbf{u} \in \partial g(\mathbf{z})$.

(iv) $\mathbf{z} \in Prox_{\tau g(\cdot)}(\mathbf{z} + \tau\mathbf{u})$ holds with some $\tau > 0$ if and only if $\mathbf{u}$ and $\mathbf{z}$ take the following value

$$(u_i, z_i) \in \Omega_1 \cup \Omega_2, \quad \forall i \in [m], \tag{9}$$

where $\Omega_1 := \{0\} \times [\sqrt{2\mu\tau}, \infty)$ and $\Omega_2 := (-\infty, \sqrt{2\mu/\tau}] \times \{0\}$.

(v) If $\mathbf{z} \in Prox_{\tau g(\cdot)}(\mathbf{z} + \tau\mathbf{u})$ with some $\tau > 0$, then $\mathbf{u} \in \partial g(\mathbf{z})$. Conversely, if $\mathbf{u} \in \partial g(\mathbf{z})$, then $\mathbf{z} \in Prox_{\tau g(\cdot)}(\mathbf{z} + \tau\mathbf{u})$ holds with $\tau \in (0, \min\{\boldsymbol{\tau}_1(\mathbf{z}), \boldsymbol{\tau}_2(\mathbf{u})\})$, where

$$\boldsymbol{\tau}_1(\mathbf{z}) = \begin{cases} +\infty, & \mathbf{z} = 0, \\ \min\left\{\dfrac{z_i^2}{2\mu} : z_i > 0\right\}, & \text{otherwise}, \end{cases}$$

$$\boldsymbol{\tau}_2(\mathbf{u}) := \begin{cases} +\infty, & \mathbf{u} \leq 0, \\ \min\left\{ \dfrac{2\mu}{u_i^2} : u_i > 0 \right\}, & \textit{otherwise}. \end{cases}$$

(vi) If $\mathbf{z} \in Prox_{\tau g(\cdot)}(\mathbf{z}+\tau\mathbf{u})$ *with some* $\tau > 0$*, then for any* $\tau' \in (0, \tau)$*,* $Prox_{\tau' g(\cdot)}(\mathbf{z}+\tau'\mathbf{u})$ *is single-valued and it holds that* $\mathbf{z} = Prox_{\tau' g(\cdot)}(\mathbf{z}+\tau'\mathbf{u})$.

Proof (i) Let us consider the regular and limiting subdifferential of a l.s.c. function $g_1 : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, defined by

$$g_1(t) := \begin{cases} \mu, & \text{if } t > 0, \\ 0, & \text{if } t = 0, \\ \infty, & \text{otherwise.} \end{cases}$$

Obviously, $g(\mathbf{z}) = \sum_{i=1}^m g_1(z_i)$ holds and it follows from [57, Prop. 10.5] that

$$\widehat{\partial} g(\mathbf{z}) = \widehat{\partial} g_1(z_1) \times \cdots \times \widehat{\partial} g_1(z_m) \quad \text{and} \quad \partial g(\mathbf{z}) = \partial g_1(z_1) \times \cdots \times \partial g_1(z_m). \tag{10}$$

Given $t \geq 0$, we claim that

$$\widehat{\partial} g_1(t) = \left\{ v \in \mathbb{R} : v \in \begin{cases} \mathbb{R}, & \text{if } t = 0, \\ \{0\}, & \text{if } t > 0 \end{cases} \right\}. \tag{11}$$

Indeed, when $t > 0$, g_1 is differentiable and then $\widehat{\partial} g_1(t) = g_1'(t) = 0$ from [57, Exercise 8.8]. If $t = 0$, then for any sequence $t_k \downarrow 0$ and $t_k \neq 0$, we have $g_1(t_k) - g_1(0) \geq \mu$. Thus, given any $v \in \mathbb{R}$, we have $(g_1(t_k) - g_1(0) - \langle v, t_k \rangle)/|t_k| \to +\infty$. Based on these facts, (11) can be derived by the definition of regular subdifferential. We can further prove $\partial g_1(t) = \widehat{\partial} g_1(t)$ by the definition of limiting subdifferential. Combining this result with (10), we obtain (7).

(ii) By the definition of the proximal operator and the separable property of $\|\cdot\|^2$ and $\|\cdot\|_0$, we only need to solve the following optimization problem for each $i \in [m]$

$$\left[\text{Prox}_{\tau g(\cdot)}(\mathbf{u})\right]_i = \arg\min_{z_i \geq 0} \frac{1}{2}(z_i - u_i)^2 + \tau\mu\|z_i\|_0. \tag{12}$$

When $z_i = 0$, the objective function value is $\texttt{obj}_1 = u_i^2/2$. Now we consider a constrained quadratic optimization

$$\min_{z_i \geq 0} \frac{1}{2}(z_i - u_i)^2 + \tau\mu.$$

Its optimal solution is $z_i = \max\{0, u_i\}$ and the objective function value is $\texttt{obj}_2 = (\min\{u_i, 0\})^2/2 + \tau\mu$. Finally, we just need to compare $\texttt{obj}_1$ and $\texttt{obj}_2$ to determine the solution to (12).

Case I: if $\texttt{obj}_1 < \texttt{obj}_2$, this means $u_i < \sqrt{2\tau\mu}$ and $[\text{Prox}_{\tau g(\cdot)}(\mathbf{u})]_i = 0$.

Case II: $\texttt{obj}_1 > \texttt{obj}_2$ is equivalent to $u_i > \sqrt{2\tau\mu}$ and $[\mathrm{Prox}_{\tau g(\cdot)}(\mathbf{u})]_i = \max\{0, u_i\} = u_i$.

Case III: if $\texttt{obj}_1 = \texttt{obj}_2$, this leads to $u_i = \sqrt{2\tau\mu}$ and $[\mathrm{Prox}_{\tau g(\cdot)}(\mathbf{u})]_i = \{0, u_i\}$. We arrive at the conclusion of (ii).

(iii) This result can be directly obtained from (6) and (7).

(iv) Suppose $\mathbf{z} \in \mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z} + \tau\mathbf{u})$ holds, we can use (8) to derive the range of $(\mathbf{u}, \mathbf{z})$. If $z_i + \tau u_i < \sqrt{2\tau\mu}$, then $z_i = 0$ and $u_i < \sqrt{2\mu/\tau}$. If $z_i + \tau u_i > \sqrt{2\tau\mu}$, then $z_i = z_i + \tau u_i$, leading to $u_i = 0$ and $z_i > \sqrt{2\mu\tau}$. For $z_i + \tau u_i = \sqrt{2\tau\mu}$, we have $z_i \in \{0, z_i + \tau u_i\}$, yielding $(u_i, z_i) = (0, \sqrt{2\tau\mu})$ or $(u_i, z_i) = (\sqrt{2\mu/\tau}, 0)$.

Now let us prove the converse counterpart. If $(u_i, z_i) \in \Omega_1$, then $z_i + \tau u_i \geq \sqrt{2\tau\mu}$ and thus $z_i = z_i + \tau u_i \in [\mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z}+\tau\mathbf{u})]_i$. If $(u_i, z_i) \in \Omega_2$, then $z_i + \tau u_i \leq \sqrt{2\tau\mu}$ and thus $z_i = 0 \in [\mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z} + \tau\mathbf{u})]_i$. Overall, we complete the proof of (iii).

(v) The first part of this assertion directly follows from (i) and (iv). To prove the converse part, it suffices to prove that $\mathbf{u} \in \partial g(\mathbf{z})$ implies (9) with $\tau \in (0, \min\{\boldsymbol{\tau}_1(\mathbf{z}), \boldsymbol{\tau}_2(\mathbf{u})\})$. Indeed, for any i_0 such that $z_{i_0} > 0$, we have $\tau < \boldsymbol{\tau}_1(\mathbf{z}) < z_{i_0}^2/(2\mu)$, implying $z_{i_0} > \sqrt{2\tau\mu}$. Meanwhile, for any i_0 such that $u_{i_0} > 0$, we have $\tau < \boldsymbol{\tau}_2(\mathbf{u}) < (2\mu)/u_{i_0}^2$, leading to $u_{i_0} < \sqrt{2\mu/\tau}$. These results together with $\mathbf{u} \in g(\mathbf{z})$ yields $(u_i, z_i) \in \Omega_1 \cup \Omega_2$ for $i \in [m]$, and we obtain (9).

(vi) It suffices to prove $\mathbf{z} = \mathrm{Prox}_{\tau' g(\cdot)}(\mathbf{z}+\tau'\mathbf{u})$. By (iv), we have $(u_i, z_i) \in \Omega_1 \cup \Omega_2$ for $i \in [m]$. Then let us consider the following two cases.

Case I: If $(u_i, z_i) \in \Omega_1$, then $u_i = 0$ and $z_i + \tau' u_i \geq \sqrt{2\mu\tau} > \sqrt{2\mu\tau'}$. Using (8), we can derive $[\mathrm{Prox}_{\tau' g(\cdot)}(\mathbf{z} + \tau'\mathbf{u})]_i = z_i + \tau' u_i = z_i$.

Case II: If $(u_i, z_i) \in \Omega_2$, then $z_i = 0$ and $u_i \leq \sqrt{2\mu/\tau}$. Hence, $z_i + \tau' u_i \leq \tau'\sqrt{2\mu/\tau} < \sqrt{2\mu\tau}$. Using (8), we can derive $[\mathrm{Prox}_{\tau' g(\cdot)}(\mathbf{z} + \tau'\mathbf{u})]_i = 0 = z_i$.

Therefore, we obtain the desired conclusion. □

2.3 Function classes

This paper involves a few classes of functions. We briefly introduce them below.

(A) Strongly convex functions, L-smooth functions and conjugate functions. We say a function $\psi : \mathbb{R}^n \to (-\infty, +\infty]$ is σ_ψ-strongly convex for $\sigma_\psi > 0$, if dom ψ is convex and the following inequality holds for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}$ and $s \in [0, 1]$,

$$\psi(s\mathbf{x} + (1-s)\mathbf{y}) \leq s\psi(\mathbf{x}) + (1-s)\psi(\mathbf{y}) - \frac{\sigma_\psi}{2}s(1-s)\|\mathbf{x}-\mathbf{y}\|^2.$$

We say ψ is ℓ_ψ-smooth for $\ell_\psi > 0$ if it is continuously differentiable and

$$\|\nabla\psi(\mathbf{x}) - \nabla\psi(\mathbf{y})\| \leq \ell_\psi\|\mathbf{x}-\mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Lemma 2 (Descent Lemma in [3]) *Let $\psi : \mathbb{R}^n \to (-\infty, \infty]$ be a ℓ_ψ-smooth function. Then we have*

$$\psi(\mathbf{x}) \leq \psi(\mathbf{y}) + \langle\nabla\psi(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{\ell_\psi}{2}\|\mathbf{x}-\mathbf{y}\|^2, \quad \forall\, \mathbf{x},\ \mathbf{y} \in \mathbb{R}^n.$$

For a function $\varphi : \mathbb{R}^m \to (-\infty, +\infty]$, its conjugate function $\varphi^* : \mathbb{R}^m \to (-\infty, +\infty]$ is defined by

$$\varphi^*(\mathbf{v}) := \sup_{\mathbf{z}\in\mathbb{R}^m} \langle \mathbf{v}, \mathbf{z}\rangle - \varphi(\mathbf{z}).$$

Some properties of conjugate function from [3] are summarized as follows.

Lemma 3 *Let φ be a proper, l.s.c. and convex function, then we have*

(i) *φ^* is proper, l.s.c. and convex.*
(ii) *$\mathbf{u} \in \partial\varphi(\mathbf{z})$ if and only if $\mathbf{z} \in \partial\varphi^*(\mathbf{u})$.*
(iii) *If φ is σ_φ-strongly convex, then φ^* is $\frac{1}{\sigma_\varphi}$-smooth.*
(iv) *If φ is ℓ_φ-smooth, then φ^* is $\frac{1}{\ell_\varphi}$-strongly convex.*

(B) PŁK functions. They are also widely known as Kurdyka-Łojasiewicz (KŁ) functions defined by KŁ inequality and popularized in [2, 7]. During the review process of this paper, we were brought to a recent paper [5], which attributes the origin of KŁ inequality to Polyak [52] and suggested the term Polyak-Łojasiewicz-Kurdyka (PŁK) inequality to reflect the order of the time the inequality and its variants were first proposed. We refer to the introduction part of [5] for a short historical account of PŁK inequality.

Let $s \in (0, +\infty]$, we denote Ψ_s as the class of all concave and continuous functions: $\psi : [0, s) \to \mathbb{R}_+$ which satisfies the following conditions

(i) $\psi(0) = 0$;
(ii) ψ is continuously differentiable on $(0, s)$ and continuous at 0;
(iii) $\psi'(t) > 0$ for all $t \in (0, s)$.

Next let us define Polyak-Łojasiewicz-Kurdyka (PŁK) property. A proper and l.s.c. function $\varphi : \mathbb{R}^m \to (-\infty, +\infty]$ is said to have the PŁK property at $\overline{\mathbf{z}} \in \operatorname{dom}\partial\varphi := \{\mathbf{z} \in \mathbb{R}^m : \partial\varphi(\mathbf{z}) \neq \varnothing\}$ if there exists $s \in (0, +\infty]$, a neighborhood $\mathcal{N}(\overline{\mathbf{z}})$ of $\overline{\mathbf{z}}$ and a function $\psi \in \Psi_s$, such that the following inequality holds

$$\psi'(\varphi(\mathbf{z}) - \varphi(\overline{\mathbf{z}}))\operatorname{dist}(0, \partial\varphi(\mathbf{z})) \geq 1, \quad \forall \mathbf{z} \in \mathcal{N}(\overline{\mathbf{z}}) \cap [\varphi(\overline{\mathbf{z}}) < \varphi(\mathbf{z}) < \varphi(\overline{\mathbf{z}}) + s] \quad (13)$$

where for $\overline{c}_1, \overline{c}_2 > 0$, $[\overline{c}_1 < \varphi(\mathbf{z}) < \overline{c}_2] := \{\mathbf{z} \in \mathbb{R}^m : \overline{c}_1 < \varphi(\mathbf{z}) < \overline{c}_2\}$. If φ satisfies the PŁK property at every point of dom $\partial\varphi$, then we call φ a PŁK function.

The PŁK property reveals that the values of φ can be reparameterized to sharpness [53, 54] and this is crucial for global convergence analysis of algorithms in nonconvex settings. One attractive aspect of PŁK functions is that they are ubiquitous in applications. The common classes of PŁK functions include real subanalytic, semi-algebraic, convex functions with growth conditions and so on. For more illustrating examples about PŁK properties, we recommend [2] and [8].

(C) Semismooth functions. Let $\Phi : \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitzian function. Suppose that D_Φ is the set containing all the differentiable points of Φ and $J\Phi(\mathbf{z})$ is the Jacobian at a differentiable point $\mathbf{z}$. Then the generalized Jacobian of Φ is defined

by

$$\partial_C \Phi(\bar{\mathbf{z}}) := \mathrm{co}\Big\{ V \mid \text{there exists a sequence } \mathbf{z}^k \in D_\Phi \text{ such that } V = \lim_{\mathbf{z}^k \to \bar{\mathbf{z}}} J\Phi(\mathbf{z}^k) \Big\}.$$

Furthermore, Φ is said to be semismooth [20, 21, 55] if for any $\mathbf{z} \in \mathbb{R}^m$, Φ is directionally differentiable and it holds

$$\Phi(\mathbf{z}+\mathbf{d}) - \Phi(\mathbf{z}) - H\mathbf{d} = o(\|\mathbf{d}\|) \quad \forall\, H \in \partial_C \Phi(\mathbf{z}+\mathbf{d}),\ \mathbf{d} \in \mathbb{R}^m. \tag{14}$$

For a continuously differentiable function $\psi : \mathbb{R}^m \mapsto (-\infty, +\infty]$, we say it is SC^1 function if the gradient function $\nabla\psi(\cdot)$ is semismooth.

3 Optimality of primal and dual problems

In this section, we study the optimality conditions of the primal and dual problems. The ultimate result is the equivalence of local minimizers between the two problems in the sense that one can be obtained from the other. This one-to-one correspondence can be seen as the duality theorem for the problem (1). For convenience, we consider the following equivalent reformulation of (1)

$$\min_{\mathbf{x},\mathbf{u}} f(\mathbf{x}) + \mathcal{I}(\mathbf{u}), \quad s.t. \quad (\mathbf{x},\mathbf{u}) \in \mathcal{F} := \{(\mathbf{x},\mathbf{u}) \mid A\mathbf{x}+\mathbf{b} = \mathbf{u}\}. \tag{15}$$

We interchangeably refer to them as the primal problem depending on whether the variable $\mathbf{u}$ is needed or not.

3.1 Equivalence of local minimizers and KKT points of the primal problem

We note that the strong convexity of f ensures that the objective of (1) is l.s.c and coercive, then there must exist an optimal solution for the primal problem (1) and (15). The KKT point of (15) with a continuously differentiable f has been introduced in [70]. By utilizing limiting subdifferential, we can extend this definition to the case when f is nonsmooth.

Definition 1 A point $\mathbf{w}^* := (\mathbf{x}^*, \mathbf{u}^*) \in \mathbb{R}^{n+m}$ is a KKT point of (15) if there exists a multiplier $\mathbf{z}^*$ such that the following system holds:

$$\begin{cases} \partial f(\mathbf{x}^*) + A^\top \mathbf{z}^* \ni 0, \\ -\mathbf{z}^* + \partial \mathcal{I}(\mathbf{u}^*) \ni 0, \\ A\mathbf{x}^* + \mathbf{b} - \mathbf{u}^* = 0. \end{cases} \tag{16}$$

We call $(\mathbf{w}^*, \mathbf{z}^*)$ a KKT pair of (15).

The convexity of f and the piecewise linear structure of the composite term $\mathcal{I}(A\mathbf{x}+\mathbf{b})$ allow us to characterize a local minimizer in terms of KKT point.

Proposition 1 *For problem* (15), $\mathbf{w}^* := (\mathbf{x}^*, \mathbf{u}^*) \in \mathbb{R}^{n+m}$ *is a local minimizer if and only if it is a KKT point.*

Proof First, for a given reference point $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*)$, define the following linearly constrained convex programming:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \text{s.t.} \quad (\mathbf{x}, \mathbf{u}) \in \mathcal{F}^* := \left\{(\mathbf{x}, \mathbf{u}) \mid A\mathbf{x} + \mathbf{b} - \mathbf{u} = 0, \ \ \mathbf{u}_{\mathcal{J}_-^*} \le 0\right\}, \tag{17}$$

where $\mathcal{J}_-^* := \{i \in [m] : u_i^* \le 0\}$. A minimizer $(\mathbf{x}, \mathbf{u})$ of (17) must satisfy the following system together with the multiplier $(\mathbf{z}, \mathbf{z}_{\mathcal{J}_-^*})$

$$\begin{cases} \partial f(\mathbf{x}) + A^\top \mathbf{z} \ni 0, \\ \mathbf{z}_{\overline{\mathcal{J}_-^*}} = 0, \ \mathbf{z}_{\mathcal{J}_-^*} \ge 0, \ \mathbf{u}_{\mathcal{J}_-^*} \le 0, \ \langle \mathbf{z}_{\mathcal{J}_-^*}, \ \mathbf{u}_{\mathcal{J}_-^*} \rangle = 0, \\ A\mathbf{x} + \mathbf{b} - \mathbf{u} = 0. \end{cases} \tag{18}$$

Comparing (16) and (18), and utilizing representation (6), we see that a KKT point $\mathbf{w}^*$ of (15) must be a global minimizer of the convex problem (17). With this equivalence in hand, we prove the claim in the proposition.

Necessity: Let $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*)$ be a local minimizer of (15), then there exists $\delta > 0$ such that

$$f(\mathbf{x}) + \boldsymbol{\mathcal{I}}(\mathbf{u}) \ge f(\mathbf{x}^*) + \boldsymbol{\mathcal{I}}(\mathbf{u}^*), \quad \forall \mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta) \cap \mathcal{F}.$$

By the definition of $\mathcal{J}_-^*$, we have $\boldsymbol{\mathcal{I}}(\mathbf{u}) \le \boldsymbol{\mathcal{I}}(\mathbf{u}^*)$ for any $\mathbf{u} \in \mathcal{F}^*$, and then we can obtain

$$f(\mathbf{x}) \ge f(\mathbf{x}^*), \quad \forall \mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta) \cap \mathcal{F}^*.$$

This means $\mathbf{w}^*$ is also a local minimizer of (17). Since Problem (17) is convex, $\mathbf{w}^*$ is also a global minimizer of (17), which in turn is equivalent to a KKT point of (15).

Sufficiency: Let $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*)$ be a KKT point of (15). We have proved at the beginning that $\mathbf{w}^*$ is also a global minimizer of (17). Thus we have

$$f(\mathbf{x}) \ge f(\mathbf{x}^*), \ \ \forall\, \mathbf{w} = (\mathbf{x}, \mathbf{u}) \in \mathcal{F}^*. \tag{19}$$

Now let us consider a sufficiently small radius δ^* such that for any $\mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta^*)$, we have

$$\{i \in [m] : u_i > 0\} \supseteq \overline{\mathcal{J}_-^*} \quad \text{and} \quad f(\mathbf{x}) - f(\mathbf{x}^*) > -\lambda/2, \tag{20}$$

where the second inequality follows from the lower semicontinuity of f. Then for any $\mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta^*) \cap \mathcal{F}^*$, (20) leads to $\boldsymbol{\mathcal{I}}(\mathbf{u}) \ge \boldsymbol{\mathcal{I}}(\mathbf{u}^*)$. This together with (19) implies

$$f(\mathbf{x}) + \boldsymbol{\mathcal{I}}(\mathbf{u}) \ge f(\mathbf{x}^*) + \boldsymbol{\mathcal{I}}(\mathbf{u}^*), \quad \forall \mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta^*) \cap \mathcal{F}^*.$$

If $\mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta^*) \cap (\mathcal{F} \backslash \mathcal{F}^*)$, then there exists $i_0 \in \mathcal{J}_-^*$ such that $u_{i_0} > 0$. Combining this with (20) leads to $\mathcal{I}(\mathbf{u}) \geq \mathcal{I}(\mathbf{u}^*) + \lambda$. Taking the inequality in (20) into account, we have

$$f(\mathbf{x}) + \mathcal{I}(\mathbf{u}) \geq f(\mathbf{x}^*) + \mathcal{I}(\mathbf{u}^*) + \lambda/2, \quad \forall \mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \delta^*) \cap \mathcal{F}^*.$$

Hence, we proved that $\mathbf{w}^*$ is also a local minimizer of (15). □

3.2 Equivalence of local minimizers and KKT points of the dual problem

We rewrite the dual problem (5) as a minimization problem:

$$\min_{\mathbf{z} \in \mathbb{R}^m} F(\mathbf{z}) := h(\mathbf{z}) + g(\mathbf{z}), \tag{21}$$

Since f is σ_f-strongly convex, it follows from Lem. 3 that h is convex and ℓ_h-smooth with $\ell_h = \|A\|^2/\sigma_f$. Utilizing the definition of ∂g and $\mathrm{Prox}_{\tau g(\cdot)}(\cdot)$, we introduce the KKT and P-stationary points of (21). These two types of stationary points have also been given in [30, 47] for minimizing the sum of two functions.

Definition 2 Let us consider a reference point $\mathbf{z}^* \in \mathbb{R}^m$.

(i) $\mathbf{z}^*$ is said to be a KKT point of (21) if

$$0 \in \nabla h(\mathbf{z}^*) + \partial g(\mathbf{z}^*). \tag{22}$$

(ii) $\mathbf{z}^*$ is called a P-stationary point of (21) if there is $\tau > 0$ such that

$$\mathbf{z}^* \in \mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z}^* - \tau \nabla h(\mathbf{z}^*)). \tag{23}$$

In comparison to the definition of KKT point, a P-stationary point involves a positive parameter $\tau > 0$. For a given $\mathbf{z}^*$, let $\mathbf{u}^* = -\nabla h(\mathbf{z}^*)$. The conditions (22) and (23) can be rewritten respectively as

$$\mathbf{u}^* \in \partial g(\mathbf{z}^*) \quad \text{and} \quad \mathbf{z}^* \in \mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z}^* + \tau \mathbf{u}^*).$$

It then follows from Lem. 1(v) that a KKT point of (21) is also its P-stationary point as stated in the result below.

Lemma 4 *If $\mathbf{z}^*$ is a P-stationary point of (21), then it is also a KKT point of (21). Conversely, if $\mathbf{z}^*$ is a KKT point of (21), then it is a P-stationary point with $\tau \in (0, \min\{\boldsymbol{\tau}_1(\mathbf{z}^*), \boldsymbol{\tau}_2(-\nabla h(\mathbf{z}^*))\})$, where $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ have been defined in Lem. 1.*

Moreover, every P-stationary point (KKT point) of (21) is also its local minimizer, as proved below.

Proposition 2 *For problem (21), $\mathbf{z}^*$ is a local minimizer if and only if it is a P-stationary point.*

Proof If $\mathbf{z}^*$ is a local minimizer of (21), then using [57, Thm. 10.1, Exercise 10.10], we can derive (22), which means $\mathbf{z}^*$ is a P-stationary point of (21) as we have claimed in Lem. 4.

Now let us prove the sufficiency. Suppose $\mathbf{z}^*$ is a P-stationary point of (21). We must have $\mathbf{z}^* \in \mathbb{R}^m_+$ due to the function $g(\mathbf{z})$. We choose $\widetilde{\delta}^*$ sufficiently small such that

$$\mathcal{S}(\mathbf{z}) \supseteq \mathcal{S}(\mathbf{z}^*) := \mathcal{S}^* \quad \text{and} \quad |h(\mathbf{z}) - h(\mathbf{z}^*)| < \mu/2, \quad \forall\, \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \widetilde{\delta}^*), \tag{24}$$

where we used the inclusion property of the support-set function $\mathcal{S}(\cdot)$ near $\mathbf{z}^*$ and the inequality used the continuity of h. Let $\Theta^* := \{\mathbf{z} \in \mathbb{R}^m : \mathcal{S}(\mathbf{z}) = \mathcal{S}(\mathbf{z}^*)\}$. We consider two cases.

Case I: $\mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \widetilde{\delta}^*) \cap \mathbb{R}^m_+ \cap \Theta^*$. For this case, $g(\mathbf{z}) = g(\mathbf{z}^*)$. Denote $\mathbf{u}^* = -\nabla h(\mathbf{z}^*)$, by the convexity of h, we obtain the following inequality:

$$h(\mathbf{z}) - h(\mathbf{z}^*) \geq -\langle \mathbf{u}^*, \mathbf{z} - \mathbf{z}^* \rangle = -\langle \mathbf{u}^*_{\mathcal{S}^*}, (\mathbf{z} - \mathbf{z}^*)_{\mathcal{S}^*} \rangle = 0,$$

where the last two equations follows from $\mathbf{z} \in \Theta^*$ and Lem. 1 (iv) respectively.

Case II: $\mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \widetilde{\delta}^*) \cap \mathbb{R}^m_+ \backslash \Theta^*$. Then (24) implies that $\mathcal{S}(\mathbf{z})$ contains at least one more element than $\mathcal{S}(\mathbf{z}^*)$. This in turn implies $g(\mathbf{z}) \geq g(\mathbf{z}^*) + \mu$. It follows from (24) that $h(\mathbf{z}) \geq h(\mathbf{z}^*) - \mu/2$. The two inequalities just obtained means $h(\mathbf{z}) + g(\mathbf{z}) \geq h(\mathbf{z}^*) + g(\mathbf{z}^*) + \mu/2$.

In summary, we have proved the following:

$$\begin{cases} g(\mathbf{z}) = g(\mathbf{z}^*),\ h(\mathbf{z}) \geq h(\mathbf{z}^*), & \text{for } \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \widetilde{\delta}^*) \cap \mathbb{R}^m_+ \cap \Theta^*, \\ h(\mathbf{z}) + g(\mathbf{z}) \geq h(\mathbf{z}^*) + g(\mathbf{z}^*) + \mu/2, & \text{for } \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \widetilde{\delta}^*) \cap \mathbb{R}^m_+ \backslash \Theta^*, \\ F(\mathbf{z}) = +\infty, & \text{for } \mathbf{z} \notin \mathbb{R}^m_+. \end{cases} \tag{25}$$

Consequently, we have $F(\mathbf{z}) \geq F(\mathbf{z}^*)$ for $\mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \widetilde{\delta}^*)$ and $\mathbf{z}^*$ is a local minimizer of (21). □

We finish this subsection with a result on the relationship between global minimizers and P-stationary points of (21).

Proposition 3 *For problem* (21)*, the following assertions hold:*

(i) If $\mathbf{z}^*$ is a global minimizer, then it is a P*-stationary point with $\tau \in (0, 1/\ell_h)$.*
(ii) If h is σ_h-strongly convex and $\mathbf{z}^*$ is a P*-stationary point with some $\tau > 1/\sigma_h$, then it is the unique global minimizer.*

Proof (i) directly follows from [35, Prop. 4.4]. Now we prove (ii). By the definitions of P-stationary point and proximal point, we have

$$\frac{1}{2\tau}\|\mathbf{z} - (\mathbf{z}^* - \tau\nabla h(\mathbf{z}^*))\|^2 + g(\mathbf{z}) \geq \frac{1}{2\tau}\|\tau\nabla h(\mathbf{z}^*)\|^2 + g(\mathbf{z}^*).$$

After simplification, we obtain

$$g(\mathbf{z}) - g(\mathbf{z}^*) \geq -\frac{1}{2\tau}\|\mathbf{z} - \mathbf{z}^*\|^2 - \langle \nabla h(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle.$$

The σ_h-strong convexity of h implies

$$h(\mathbf{z}) - h(\mathbf{z}^*) \geq \langle \nabla h(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle + \frac{\sigma_h}{2} \|\mathbf{z} - \mathbf{z}^*\|^2.$$

Adding the two inequalities above leads to

$$F(\mathbf{z}) - F(\mathbf{z}^*) \geq \left(\frac{\sigma_h}{2} - \frac{1}{2\tau} \right) \|\mathbf{z} - \mathbf{z}^*\|^2 > 0, \text{ whenever } \mathbf{z} \neq \mathbf{z}^*.$$

which means that $\mathbf{z}^*$ is the unique global minimizer of (21). □

3.3 Equivalence of local minimizers of primal and dual problems

The equivalence results in the two preceding subsections lead to the main result in this section.

Theorem 1 (Equivalence Characterization of Local Minimizer) *If* $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*)$ *is a local minimizer of the primal problem* (15) *with associated multiplier* $\mathbf{z}^*$*, then* $\mathbf{z}^*$ *is a local minimizer of the dual problem* (21)*. Conversely, if* $\mathbf{z}^*$ *is a local minimizer of the dual problem* (21)*, then* $\mathbf{w}^* := (\nabla f^*(-A^\top \mathbf{z}^*), -\nabla h(\mathbf{z}^*))$ *is a local minimizer of the primal problem* (15)*.*

Proof In Prop. 1 for the primal problem (15), we established the equivalence of its local minimizer and KKT point. Combining Lem. 4 and Prop. 2 for the dual problem (21), we established the equivalence of its local minimizer and its KKT point. We only need to prove the equivalence of the KKT point for both primal and the dual problems.

Suppose $\mathbf{z}^*$ is a KKT point of the dual problem (21). Let $\mathbf{x}^* := \nabla f^*(-A^\top \mathbf{z}^*)$ and $\mathbf{u}^* := -\nabla h(\mathbf{z}^*)$. Then by using Lem. 3(ii) and $h(\mathbf{z}) = f^*(-A^\top \mathbf{z}) - \langle \mathbf{b}, \mathbf{z} \rangle$, we know that the KKT condition (22) for the dual problem is equivalent to

$$\begin{cases} \partial f(\mathbf{x}^*) + A^\top \mathbf{z}^* \ni 0, \\ -\mathbf{u}^* + \partial g(\mathbf{z}^*) \ni 0, \\ A\mathbf{x}^* + \mathbf{b} - \mathbf{u}^* = 0. \end{cases} \tag{26}$$

Comparing (26) with (16), we just need to prove their second lines are equivalent. This directly follows from Lem. 1(iii). Consequently, $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{u}^*)$ is a KKT point of the primal problem. This establishes the equivalence result. □

At the end of this section, we introduce a second-order sufficient condition (SOSC) for (21). Just like the case of classic smooth nonlinear programming (see e.g. [50]), we will show that our SOSC is useful to derive a quadratic growth condition for a P-stationary point and local superlinear convergence rate for a second-order algorithm.

Given a P-stationary point $\mathbf{z}^*$, denote the support set $\mathcal{S}^* := \mathcal{S}(\mathbf{z}^*)$. One can verify that it is also a KKT point with multiplier $(\nabla_{\mathcal{S}^*} h(\mathbf{z}^*), \nabla_{\overline{\mathcal{S}}^*} h(\mathbf{z}^*))$ of the following convex programming

$$\min_{\mathbf{z} \in \mathbb{R}^m} h(\mathbf{z}), \text{ s.t. } \mathbf{z}_{\mathcal{S}^*} \geq 0, \ \mathbf{z}_{\overline{\mathcal{S}}^*} = 0. \tag{27}$$

The critical cone of the above problem at $\mathbf{z}^*$ is $\mathcal{C} := \{\mathbf{d} \in \mathbb{R}^m : \mathbf{d}_{\overline{\mathcal{S}}^*} = 0\}$ (it happens to be a subspace). Since h is ℓ_h-smooth, for any $\mathbf{z} \in \mathbb{R}^m$, $\partial^2 h(\mathbf{z}) := \partial_C(\nabla h(\mathbf{z})) \neq \varnothing$. Then we can give the following definition based on the second-order sufficient condition for the convex problem (27) (see e.g. [21]).

Definition 3 Given a P-stationary point $\mathbf{z}^* \in \mathbb{R}^m$ of (21), we say the second-order sufficient condition (SOSC) holds at $\mathbf{z}^*$ if

$$\mathbf{d}^\top H^*_{\mathcal{S}^*}\mathbf{d} > 0, \quad \forall H^* \in \partial^2 h(\mathbf{z}^*) \text{ and } \mathbf{d} \in \mathbb{R}^{|\mathcal{S}^*|}\backslash\{0\}. \tag{28}$$

Similar to the results for SC^1 functions (see [21, Prop. 7.4.12]), we can prove the quadratic-growth condition for (21) at a P-stationary point $\mathbf{z}^*$ satisfying SOSC.

Proposition 4 (Quadratic Growth Condition) *If ∇h is semismooth and $\mathbf{z}^* \in \mathbb{R}^m$ is a P-stationary point of (21) satisfying SOSC, then there exist $\overline{\delta}^*, c^* > 0$ such that the following quadratic growth at $\mathbf{z}^*$ holds:*

$$F(\mathbf{z}) \geq F(\mathbf{z}^*) + c^*\|\mathbf{z} - \mathbf{z}^*\|^2, \qquad \forall\, \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \overline{\delta}^*) \cap \mathbb{R}^m_+. \tag{29}$$

Proof As we have mentioned before Def. 3, $\mathbf{z}^*$ is also a KKT point of (27). Considering that (28) holds, it follows from [21, Prop. 7.4.12] that there exists $\overline{\delta}^*_1$ and $c^* > 0$ such that

$$h(\mathbf{z}) \geq h(\mathbf{z}^*) + c^*\|\mathbf{z} - \mathbf{z}^*\|^2, \ \forall \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \overline{\delta}^*_1) \cap \{\mathbf{z} \in \mathbb{R}^m : \mathbf{z}_{\mathcal{S}^*} \geq 0, \mathbf{z}_{\overline{\mathcal{S}}^*} = 0\}. \tag{30}$$

Then we can take $\overline{\delta}^* = \min\{\widetilde{\delta}^*, \overline{\delta}^*_1, \sqrt{\mu/(2c^*)}\}$, where $\widetilde{\delta}^*$ has been defined in the proof of Prop. 2. Then (24) holds. Furthermore, for any $\mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \overline{\delta}^*)$, we have $\{\mathbf{z} \in \mathbb{R}^m : \mathbf{z}_{\mathcal{S}^*} \geq 0,\ \mathbf{z}_{\overline{\mathcal{S}}^*} = 0\} = \mathbb{R}^m_+ \cap \Theta^*$, where $\Theta^* = \{\mathbf{z} \in \mathbb{R}^m : \mathcal{S}(\mathbf{z}) = \mathcal{S}(\mathbf{z}^*)\}$. Combination of (30) and (25) leads to

$$\begin{aligned} &F(\mathbf{z}) - F(\mathbf{z}^*) \geq c^*\|\mathbf{z} - \mathbf{z}^*\|^2, \quad \forall \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \overline{\delta}^*) \cap \mathbb{R}^m_+ \cap \Theta^*, \\ &F(\mathbf{z}) - F(\mathbf{z}^*) \geq \mu/2 \geq c^*\|\mathbf{z} - \mathbf{z}^*\|^2, \quad \forall \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \overline{\delta}^*) \cap \mathbb{R}^m_+ \backslash \Theta^*. \end{aligned}$$

We arrived at the claimed result. □

4 A subspace gradient semismooth newton method

Since we have established the equivalence of solutions between the primal problem (15) and the dual problem(21), we now develop an efficient method for the dual problem. It is essentially a subspace method, which first identifies a subspace and then within this subspace applies a gradient step followed by a semismooth Newton step if certain conditions are met. We call the resulting method a subspace gradient semismooth Newton (SGSN) method. The key challenge is to establish its global convergence as well as its local superlinear convergence rate under mild assumptions. We now describe the framework for SGSN followed by its convergence analysis.

4.1 Steps of SGSN

Suppose $\mathbf{z}^k$ is the current iterate. First, we perform the classical proximal gradient step

$$\mathbf{v}^k \in \mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z}^k - \tau \nabla h(\mathbf{z}^k)). \tag{31}$$

The aim of this step is to ensure global convergence and to identify the subspace $\mathbb{R}^{|T_k|}$ indexed by

$$T_k := \{i \in [m] : [\mathbf{z}^k - \tau \nabla h(\mathbf{z}^k)]_i > \sqrt{2\tau\mu}\}.$$

From (8), once such $\mathbf{v}^k$ is given by

$$\mathbf{v}^k_{T_k} = [\mathbf{z}^k - \tau \nabla h(\mathbf{z}^k)]_{T_k} \quad \text{and} \quad \mathbf{v}^k_{\overline{T}_k} = 0. \tag{32}$$

In other words, $\mathbf{v}^k$ takes a gradient descent on $\mathbb{R}^{|T_k|}$ while setting the complementary part to zero.

The second step is to compute a Newton direction along the subspace identified in the first step. Specifically, we minimize a second-order Taylor approximation of $h(\mathbf{v}^k + \mathbf{d})$ with $\mathbf{d}_{\overline{T}_k} = 0$:

$$\mathbf{d}^k := \arg\min_{\mathbf{d}\in\mathbb{R}^m} \ \langle \nabla h(\mathbf{v}^k), \mathbf{d}\rangle + \frac{1}{2}\mathbf{d}^\top (H^k + \gamma_k I)\mathbf{d}, \quad \text{s.t. } \mathbf{d}_{\overline{T}_k} = 0.$$

where $\gamma_k > 0$ is a regularization parameter, $H^k \in \widehat{\partial}^2 h(\mathbf{v}^k)$, and the set $\widehat{\partial}^2 h(\mathbf{z})$ is defined by

$$\widehat{\partial}^2 h(\mathbf{z}) := \{AQA^\top \mid Q \in \partial^2 f^*(-A^\top \mathbf{z})\}. \tag{33}$$

Since f^* is $(1/\sigma_f)$-smooth, $\widehat{\partial}^2 h(\mathbf{z})$ is always nonempty. The reason to use $\widehat{\partial}^2 h$ instead of $\partial^2 h$ is because h includes the composition term $f^* \circ (-A)$ and it is usually difficult to characterize the exact form of $\partial^2 h$. The relationship between the two sets is given in [16, Thm. 2.6.6]:

$$\partial^2 h(\mathbf{z})\overline{\mathbf{d}} \subseteq \widehat{\partial}^2 h(\mathbf{z})\overline{\mathbf{d}}, \qquad \forall\, \overline{\mathbf{d}} \in \mathbb{R}^m. \tag{34}$$

Since H^k is positive semidefinite, $(H^k + \gamma_k I)$ is positive definite and hence $\mathbf{d}^k$ is well-defined.

The final step is to measure whether the Newton direction is "sufficiently" good. Let

$$\widetilde{\mathbf{z}}^{k+1} := \mathbf{v}^k + \alpha_k \mathbf{d}^k,$$

where $\alpha_k > 0$ can be explicitly computed. We accept $\widetilde{\mathbf{z}}^{k+1}$ as our next iterate if it satisfies the following two conditions:

$$F(\mathbf{v}^k) - F(\widetilde{\mathbf{z}}^{k+1}) \geq c_1 \|\widetilde{\mathbf{z}}^{k+1} - \mathbf{v}^k\|^2 \tag{C1}$$

$$\|\nabla_{T_k} h(\widetilde{\mathbf{z}}^{k+1})\| \leq c_2 \|\widetilde{\mathbf{z}}^{k+1} - \mathbf{v}^k\|, \tag{C2}$$

where $c_1, c_2 > 0$. Otherwise, we take $\mathbf{v}^k$ as our next iterate. Those steps are summarized in Alg. 1.

Algorithm 1 SGSN: Subspace Gradient Semismooth Newton Method

Initialization: Take $\tau \in (0, 1/\ell_h)$, $c_1, c_2, \gamma > 0$, and then start with $\mathbf{z}^0 \in \mathbb{R}^m$.
for $k = 0, 1, \cdots$ **do**
1. Proximal gradient: Compute T_k, and then update the gradient iterate $\mathbf{v}^k$ by (32).
2. Semismooth Newton: Denote $\gamma_k := \gamma \|\nabla_{T_k} h(\mathbf{v}^k)\|$, and choose $H^k \in \widehat{\partial}^2 h(\mathbf{v}^k)$. Compute Newton direction $\mathbf{d}^k_{T_k}$ by solving

$$(H^k_{T_k} + \gamma_k I)\mathbf{d}^k_{T_k} = -\nabla_{T_k} h(\mathbf{v}^k). \tag{35}$$

Update Newton iterate by

$$\widetilde{\mathbf{z}}^{k+1}_{T_k} = \mathbf{v}^k_{T_k} + \alpha_k \mathbf{d}^k_{T_k} \text{ and } \widetilde{\mathbf{z}}^{k+1}_{\overline{T}_k} = 0. \tag{36}$$

where $\alpha_k = \min\{1, \beta_k\}$ with

$$\beta_k := \begin{cases} 1, & \text{if } \mathbf{d}^k_{T_k} \geq 0 \\ \min\{-v^k_i / d^k_i : i \in T_k \text{ and } d^k_i < 0\}, & \text{otherwise.} \end{cases} \tag{37}$$

3. Update step: Update $\mathbf{z}^k$ either by the semismooth Newton step or the proximal gradient step as follows:

$$\mathbf{z}^{k+1} = \begin{cases} \widetilde{\mathbf{z}}^{k+1}, & \text{if } (\mathbf{C1}) \text{ and } (\mathbf{C2}) \text{ hold} \\ \mathbf{v}^k, & \text{otherwise.} \end{cases} \tag{38}$$

end for

Remark 1 (i) From (32), one can observe that T_k exactly includes all nonzero entries of $\mathbf{v}^k$. Together with (36), we have

$$T_k = \mathcal{S}(\mathbf{v}_k) \supseteq \mathcal{S}(\widetilde{\mathbf{z}}^{k+1}). \tag{39}$$

(ii) Denote $\mathbf{x}^k := \nabla f^*(-A^\top \mathbf{z}^k)$, then $-\nabla h(\mathbf{z}^k) = A\mathbf{x}^k + \mathbf{b}$ and $\mathbf{v}^{k+1}_{T_k} = [\mathbf{z}^k + \tau(A\mathbf{x}^k + \mathbf{b})]_{T_k}$.

Following this setting, the computational cost of the proximal gradient step is $\mathcal{O}(|T_k|n)$. Regarding the Newton step, let us consider a simple case when f is twice continuously differentiable and so is f^*. Then the Hessian is $H^k = AQ^kA^\top$ with $Q^k := \nabla^2 f^*(-A^\top \mathbf{z}^k)$ and Newton equation (35) will be

$$(\gamma_k I + A_{T_k:} Q^k A^\top_{T_k:})\mathbf{d}^k_{T_k} = -\nabla_{T_k} h(\mathbf{v}^k).$$

In many applications, including the AUC maximization and multi-label classification, f is often a separable regularizer. Thereby, Q^k is diagonal. Then the above equation can be efficiently solved by linear conjugate gradient method and the computational cost is $\mathcal{O}(n|T_k|^2)$. In the case $n < |T_k|$ and the inverse of Q^k is cheap to compute,

we can make use of the Sherman-Morrison-Woodbury formula so that the overall computational cost is in the order of $\mathcal{O}(n^2|T_k|)$.

(iii) At the k-th iteration, if it happens that $\mathbf{v}^k = \mathbf{z}^k$, then $\nabla_{T_k} h(\mathbf{v}^k) = 0$ and $\mathbf{v}^k_{\overline{T}_k} = 0$, which further implies $0 \in \nabla h(\mathbf{v}^k) + \partial g(\mathbf{v}^k)$ and $\widetilde{\mathbf{z}}^{k+1} = \mathbf{v}^k$. This means that $\mathbf{z}^k$ is a KKT point (see (22) in Def. 2). Consequently, after the k-th iteration, the iterate would not change and is a P-stationary point of (21) by Lem. 4.

4.2 Global convergence

In this subsection, our main goal is to prove the sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ generated by SGSN converges to a P-stationary point of (21). We first state some basic properties of the generated sequence.

Proposition 5 *Let $\{(\mathbf{z}^k, \mathbf{v}^k)\}_{k\in\mathbb{N}}$ be the sequence generated by SGSN. The following holds.*

(i) The sequence of objective function value satisfies sufficient descent

$$F(\mathbf{z}^k) - F(\mathbf{v}^k) \geq \eta_0 \|\mathbf{v}^k - \mathbf{z}^k\|^2, \quad F(\mathbf{v}^k) - F(\mathbf{z}^{k+1}) \geq c_1 \|\mathbf{v}^k - \mathbf{z}^{k+1}\|^2, \tag{40}$$

where $\eta_0 = 1/(2\tau) - \ell_h/2 > 0$.

(ii) There exist $\mathbf{p}^k \in \partial F(\mathbf{v}^k)$ such that $\|\mathbf{p}^k\| \leq (1/\tau + \ell_h)\|\mathbf{v}^k - \mathbf{z}^k\|$. If Newton iterate $\widetilde{\mathbf{z}}^{k+1}$ is accepted, then there exists $\widetilde{\mathbf{q}}^{k+1} \in \partial F(\widetilde{\mathbf{z}}^{k+1})$ such that $\|\widetilde{\mathbf{q}}^{k+1}\| \leq \eta_1 \|\mathbf{z}^{k+1} - \mathbf{v}^k\|$, where $\eta_1 := \max\{c_2, 1/\tau + \ell_h\}$.

(iii) Suppose the function sequence $\{F(\mathbf{z}^k)\}_{k\in\mathbb{N}}$ is bounded from below, then there exists $C \in \mathbb{R}$ such that

$$\lim_{k\to\infty} F(\mathbf{z}^k) = \lim_{k\to\infty} F(\mathbf{v}^k) = C. \tag{41}$$

Furthermore, it holds

$$\lim_{k\to\infty} \|\mathbf{v}^k - \mathbf{z}^k\| = 0 \quad \text{and} \quad \lim_{k\to\infty} \|\mathbf{v}^k - \mathbf{z}^{k+1}\| = 0. \tag{42}$$

(iv) Let $\mathbf{z}^*$ be an accumulation point of $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$, then $\mathbf{z}^*$ is a P-stationary point of (21) and $F(\mathbf{z}^*) = C$. The corresponding conclusions for sequence $\{\mathbf{v}^k\}_{k\in\mathbb{N}}$ also hold.

Proof (i) By (31) and the definition of the proximal operator, we have

$$\|\mathbf{v}^k - (\mathbf{z}^k - \tau\nabla h(\mathbf{z}^k))\|^2/2 + \tau g(\mathbf{v}^k) \leq \|\mathbf{z}^k - (\mathbf{z}^k - \tau\nabla h(\mathbf{z}^k))\|^2/2 + \tau g(\mathbf{z}^k),$$

which implies

$$\|\mathbf{v}^k - \mathbf{z}^k\|^2/(2\tau) + \langle \nabla h(\mathbf{z}^k), \mathbf{v}^k - \mathbf{z}^k\rangle + g(\mathbf{v}^k) \leq g(\mathbf{z}^k).$$

Since h is ℓ_h-smooth, using Lem. 2, we obtain

$$h(\mathbf{v}^k) \leq h(\mathbf{z}^k) + \langle \nabla h(\mathbf{z}^k), \mathbf{v}^k - \mathbf{z}^k \rangle + (\ell_h/2)\|\mathbf{v}^k - \mathbf{z}^k\|^2.$$

Adding the above two inequalities leads to $F(\mathbf{z}^k) - F(\mathbf{v}^k) \geq \eta_0 \|\mathbf{v}^k - \mathbf{z}^k\|^2$. If $\mathbf{z}^{k+1} = \widetilde{\mathbf{z}}^{k+1}$, then (40) follows from (**C1**), and (40) automatically holds otherwise.

(ii) From (31) and the definition of the proximal operator, we have

$$\mathbf{v}^k \in \arg\min_{\mathbf{z}} \|\mathbf{z} - (\mathbf{z}^k - \tau \nabla h(\mathbf{z}^k))\|^2/2 + \tau g(\mathbf{z}).$$

By [57, Exercise 8.8], we have

$$0 \in \mathbf{v}^k - (\mathbf{z}^k - \tau \nabla h(\mathbf{z}^k)) + \tau \partial g(\mathbf{v}^k).$$

Denoting $\mathbf{p}^k := -(\mathbf{v}^k - \mathbf{z}^k)/\tau + \nabla h(\mathbf{v}^k) - \nabla h(\mathbf{z}^k)$, then $\mathbf{p}^k \in \nabla h(\mathbf{v}^k) + \partial g(\mathbf{v}^k) = \partial F(\mathbf{v}^k)$. We can also estimate $\|\mathbf{p}^k\| \leq (1/\tau + \ell_h)\|\mathbf{v}^k - \mathbf{z}^k\|$ by using ℓ_h-smoothness of h.

If $\mathbf{z}^{k+1} = \widetilde{\mathbf{z}}^{k+1}$, let us take $\widetilde{\mathbf{q}}^{k+1} = (\nabla_{T_k} h(\widetilde{\mathbf{z}}^{k+1}); 0_{\overline{T}_k})$. Noticing that $\widetilde{\mathbf{z}}^{k+1}_{\overline{T}_k} = 0$ from (36), we have $(0_{T_k}; -\nabla_{\overline{T}_k} h(\widetilde{\mathbf{z}}^{k+1})) \in \partial g(\widetilde{\mathbf{z}}^{k+1})$. This means that $\widetilde{\mathbf{q}}^{k+1} \in \nabla h(\widetilde{\mathbf{z}}^{k+1}) + \partial g(\widetilde{\mathbf{z}}^{k+1}) = \partial F(\widetilde{\mathbf{z}}^{k+1})$. In this case, the upper bounded of $\|\widetilde{\mathbf{q}}^{k+1}\|$ follows from (**C2**). Overall, we finish the proof of (ii).

(iii) Since $\{F(\mathbf{z}^k)\}_{k\in\mathbb{N}}$ is assumed to be bounded below, (40) implies $F(\mathbf{z}^{k+1}) \leq F(\mathbf{v}^k) \leq F(\mathbf{z}^k)$, and thus $\{F(\mathbf{v}^k)\}_{k\in\mathbb{N}}$ is also bounded below. Then it follows from the monotone convergence theorem that (41) holds. Using (40) and (41), we can derive (42).

(iv) Since $\mathbf{z}^*$ is an accumulation point, there exists a subsequence $\{\mathbf{z}^{k_j}\}_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty} \mathbf{z}^{k_j} = \mathbf{z}^*$. Then we have $\lim_{j\to\infty} \mathbf{z}^{k_j} - \tau \nabla h(\mathbf{z}^{k_j}) = \mathbf{z}^* - \tau \nabla h(\mathbf{z}^*)$, as well as $\lim_{j\to\infty} \mathbf{v}^{k_j} = \mathbf{z}^*$ from (42). It follows from the Proximal Behavior theorem [57, Thm. 1.25] that $\mathbf{z}^*$ is a P-stationary point of (21).

Now let us prove $F(\mathbf{z}^*) = C$. From (31) and the definition of proximal operator, we have

$$\|\mathbf{v}^{k_j} - (\mathbf{z}^{k_j} - \tau \nabla h(\mathbf{z}^{k_j}))\|^2/2 + \tau g(\mathbf{v}^{k_j}) \leq \|\mathbf{z}^* - (\mathbf{z}^{k_j} - \tau \nabla h(\mathbf{z}^{k_j}))\|^2/2 + \tau g(\mathbf{z}^*).$$

Meanwhile, the second inequality of (40) implies

$$h(\mathbf{z}^{k_j+1}) + g(\mathbf{z}^{k_j+1}) + c_1 \|\mathbf{v}^{k_j} - \mathbf{z}^{k_j+1}\|^2 \leq h(\mathbf{v}^{k_j}) + g(\mathbf{v}^{k_j}).$$

Considering that $\lim_{j\to\infty} \mathbf{z}^{k_j} = \lim_{j\to\infty} \mathbf{z}^{k_j+1} = \lim_{j\to\infty} \mathbf{v}^{k_j} = \mathbf{z}^*$ holds, taking limit superior on above two inequalities implies $\limsup_{j\to\infty} g(\mathbf{z}^{k_j}) \leq \limsup_{j\to\infty} g(\mathbf{v}^{k_j}) \leq g(\mathbf{z}^*)$. This together with lower semi-continuity of g means $\lim_{j\to\infty} g(\mathbf{z}^{k_j}) = g(\mathbf{z}^*)$. Considering the continuity of h, we have $F(\mathbf{z}^*) = \lim_{j\to\infty} F(\mathbf{z}^{k_j}) = C$. By a similar procedure, we can prove the counterpart conclusion for sequence $\{\mathbf{v}^k\}_{k\in\mathbb{N}}$. □

Remark 2 The assumption of the boundedness of the function sequence $\{F(\mathbf{z}^k)\}_{k\in\mathbb{N}}$ and the existence of an accumulation point of the sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ can be ensured by the following assumption:

Assumption 1 The sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ generated by SGSN is bounded.

This sequence boundedness assumption is commonly used for convergence analysis of algorithms in nonconvex settings (see e.g. [8, 9, 15, 25]). Under Assumption 1, the generated sequence $\{\mathbf{v}^k\}_{k\in\mathbb{N}}$ is also bounded due to (42).

Prop. 5 presents some basic properties about the sequence $\{(\mathbf{z}^k, \mathbf{v}^k)\}_{k\in\mathbb{N}}$ in separate cases. When Condition (**C1**) or (**C2**) is not satisfied at the k-th iteration, we would have $\mathbf{z}^{k+1} = \mathbf{v}^k$ (repeating points). We now remove those repeating points from $\{(\mathbf{z}^k, \mathbf{v}^k)\}_{k\in\mathbb{N}}$ and relabel the sequence by $\mathbf{y}^j$:

$$\{\mathbf{y}^j\}_{j\in\mathbb{N}} := \{\mathbf{z}^0, \mathbf{v}^0, \cdots, \mathbf{z}^k, \mathbf{v}^k, \cdots\} \backslash \{\mathbf{z}^{k+1} : \mathbf{z}^{k+1} = \mathbf{v}^k, k \in \mathbb{N}\}.$$

We can see that $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ is a subsequence of $\{\mathbf{y}^j\}_{j\in\mathbb{N}}$, because if $\mathbf{z}^{k+1} = \mathbf{v}^k$, then $\mathbf{v}^k \in \{\mathbf{y}^j\}_{j\in\mathbb{N}}$, otherwise $\mathbf{z}^{k+1} \in \{\mathbf{y}^j\}_{j\in\mathbb{N}}$. We will show that $\{\mathbf{y}^j\}_{j\in\mathbb{N}}$ is convergent and so is $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$. The benefit of removing the repeated points and relabeling allow us to restate the properties established in Prop. 5 in a unified fashion:

- Denote $\eta_2 := \min\{\eta_0, c_1\}$, then we have

$$F(\mathbf{y}^j) - F(\mathbf{y}^{j+1}) \geq \eta_2 \|\mathbf{y}^{j+1} - \mathbf{y}^j\|^2. \tag{43}$$

- There exists $\mathbf{q}^j \in \partial F(\mathbf{y}^j)$ such that

$$\|\mathbf{q}^j\| \leq \eta_1 \|\mathbf{y}^{j+1} - \mathbf{y}^j\|. \tag{44}$$

- The Ostrowski's condition holds

$$\lim_{j\to\infty} \|\mathbf{y}^{j+1} - \mathbf{y}^j\| = 0. \tag{45}$$

These crucial properties enable us to utilize the PŁK property for convergence analysis of $\{\mathbf{y}^j\}_{j\in\mathbb{N}}$.

Assumption 2 $F(\cdot)$ is a PŁK function.

Particularly, the semi-algebraic function, which remains stable under basic operations, is an important class of the PŁK function (see e.g. [8]). We can give the following sufficient condition for Assumption 2.

Lemma 5 *If f is semi-algebraic, then F is semi-algebraic, and thus it is a PŁK function.*

Proof Since f is semi-algebraic, it follows from [8, Example 2] that $\widetilde{f}(\mathbf{v}, \mathbf{z}) := \langle \mathbf{v}, \mathbf{z}\rangle - f(\mathbf{z})$ is semi-algebraic, and thus $f^*(\mathbf{v}) := \sup_{\mathbf{v}\in\mathbb{R}^m} \widetilde{f}(\mathbf{v}, \mathbf{z})$ is also semi-algebraic. As h is the composition of f^* and a linear mapping, it is semi-algebraic. Finally, taking semi-algebraicity of $\boldsymbol{\delta}_+$ and $\|\cdot\|_0$ (see [8, Example 4]) into account, $F(\cdot) = h(\cdot) + \boldsymbol{\delta}_+(\cdot) + \mu\|\cdot\|_0$ is semi-algebraic, which implies its PŁK property. □

Lemma 6 *Suppose Assumption 1 holds and let $\mathcal{Y}$ be the set consisting of all the accumulation points of $\{\mathbf{y}^j\}_{j\in\mathbb{N}}$. Then we have*

(i) $\lim_{j\to\infty} dist(\mathbf{y}^j, \mathcal{Y}) = 0$.
(ii) *$\mathcal{Y}$ is a nonempty, compact and connected set.*
(iii) *Each element $\mathbf{y}^* \in \mathcal{Y}$ is a P-stationary point of (21) with $F(\mathbf{y}^*) = \lim_{j\to\infty} F(\mathbf{y}^j) = C$, where the constant C has been defined in Prop. 5 (ii).*
(iv) *If Assumption 2 holds, then there exist $\epsilon > 0$, $s > 0$ and $\psi \in \Psi_s$ such that for any $\mathbf{z} \in \mathbb{B}^* := \{\mathbf{z} \in \mathbb{R}^m : dist(\mathbf{z}, \mathcal{Y}) < \epsilon\} \cap [C < F(\mathbf{z}) < C + s]$, the following inequality holds*

$$\psi'(F(\mathbf{z}) - C)\, dist(0, \partial F(\mathbf{z})) \geq 1. \tag{46}$$

Proof Since Ostrowski's condition (45) holds, (i) and (ii) follow from [8, Lem. 5]. (iii) follows from Prop. 5. Under Assumption 2, F is a PŁK function with a constant value on $\mathcal{Y}$. Combining this with the compactness of $\mathcal{Y}$, we can derive (iv) by [8, Lem. 6]. □

Lem. 6(iv) is also known as uniform PŁK property. This is a more favorable property than those introduced in Section 2, because the data ϵ, s and ψ is the same for all the points in $\mathbb{B}^*$, making the formula (46) more applicable for convergence analysis.

Theorem 2 *Suppose that Assumptions 1 and 2 hold. Then the sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ generated by SGSN converges to a P-stationary point of (21).*

Proof Since $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ is a subsequence of $\{\mathbf{y}^j\}_{j\in\mathbb{N}}$, we can arrive at the desired conclusion by proving that $\{\mathbf{y}^j\}_{j\in\mathbb{N}}$ converges to a P-stationary point of (21). First, let us consider a trivial case that there exists $\overline{j} \in \mathbb{N}$ satisfying $F(\mathbf{y}^{\overline{j}}) = C$. Taking $\lim_{j\to\infty} F(\mathbf{y}^j) = C$ and the sufficient descent (43) into account, we have $\mathbf{y}^j = \mathbf{y}^{\overline{j}}$ for all $j \geq \overline{j}$. As we have discussed in Remark 1(iii), $\mathbf{y}^{\overline{j}}$ actually is a P-stationary point of (21).

Now let us consider the case when $F(\mathbf{y}^j) > C$ for any $j \in \mathbb{N}$. Since $\lim_{j\to\infty} \mathrm{dist}(\mathbf{y}^j, \mathcal{Y}) = 0$ and $\lim_{j\to\infty} F(\mathbf{y}^j) = C$, it follows from Lem. 6(iv) that when j is sufficiently large, we have

$$\psi'(F(\mathbf{y}^j) - C)\, \mathrm{dist}(0, \partial F(\mathbf{y}^j)) \geq 1.$$

This together with (44) leads to

$$\psi'(F(\mathbf{y}^j) - C) \geq 1/(\eta_1 \|\mathbf{y}^j - \mathbf{y}^{j-1}\|). \tag{47}$$

Given $l_1, l_2 \in \mathbb{N}$, denote $\Delta_{l_1,l_2} := \psi(F(\mathbf{y}^{l_1}) - C) - \psi(F(\mathbf{y}^{l_2}) - C)$, then by the concavity of ψ, (47) and (43), we can derive

$$\Delta_{j,j+1} \geq \psi'(F(\mathbf{y}^j) - C)(F(\mathbf{y}^j) - F(\mathbf{y}^{j+1})) \geq \frac{\eta_2}{\eta_1} \frac{\|\mathbf{y}^{j+1} - \mathbf{y}^j\|^2}{\|\mathbf{y}^j - \mathbf{y}^{j-1}\|},$$

which leads to $\|\mathbf{y}^{j+1}-\mathbf{y}^{j}\| \le \sqrt{(\eta_1/\eta_2)\Delta_{j,j+1}\|\mathbf{y}^{j}-\mathbf{y}^{j-1}\|}$. Using the inequality of arithmetic and geometric means, we have

$$2\|\mathbf{y}^{j+1}-\mathbf{y}^{j}\| \le \|\mathbf{y}^{j}-\mathbf{y}^{j-1}\| + (\eta_1/\eta_2)\Delta_{j,j+1}.$$

Now let us sum up the above inequality from sufficiently large $\underline{l}$ to $\bar{l}$ to get

$$\begin{aligned} 2\sum_{j=\underline{l}}^{\bar{l}}\|\mathbf{y}^{j+1}-\mathbf{y}^{j}\| &\le \sum_{j=\underline{l}}^{\bar{l}}\|\mathbf{y}^{j}-\mathbf{y}^{j-1}\| + \sum_{j=\underline{l}}^{\bar{l}}\frac{\eta_1}{\eta_2}\Delta_{j,j+1} \\ &= \sum_{j=\underline{l}}^{\bar{l}}\|\mathbf{y}^{j+1}-\mathbf{y}^{j}\| + \|\mathbf{y}^{\underline{l}}-\mathbf{y}^{\underline{l}-1}\| + \frac{\eta_1}{\eta_2}\Delta_{\underline{l},\bar{l}+1}. \end{aligned}$$

This further leads to

$$\sum_{j=\underline{l}}^{\bar{l}}\|\mathbf{y}^{j+1}-\mathbf{y}^{j}\| \le \|\mathbf{y}^{\underline{l}}-\mathbf{y}^{\underline{l}-1}\| + (\eta_1/\eta_2)(\psi(F(\mathbf{y}^{\underline{l}})-C) - \psi(F(\mathbf{y}^{\bar{l}+1})-C)).$$

Taking $\bar{l}\to\infty$, we can derive

$$\sum_{j=1}^{\infty}\|\mathbf{y}^{j+1}-\mathbf{y}^{j}\| < \infty. \tag{48}$$

This indicates that $\{\mathbf{y}^{j}\}_{j\in\mathbb{N}}$ is a Cauchy sequence. It must converge to a P-stationary point of (21), and so does $\{\mathbf{z}^{k}\}_{k\in\mathbb{N}}$. We have finished the proof. □

4.3 Local superlinear convergence rate

In this Subsection, we will derive the local superlinear convergence rate of SGSN under the following assumption.

Assumption 3 $\nabla f^*(\cdot)$ is a semismooth function, and there exists an accumulation point $\mathbf{z}^*$ of the sequence $\{\mathbf{z}^{k}\}_{k\in\mathbb{N}}$ satisfying

$$\mathbf{d}^{\top}H^*_{\mathcal{S}^*}\mathbf{d} > 0, \quad \forall H^* \in \widehat{\partial}^2 h(\mathbf{z}^*) \text{ and } \mathbf{d}\in\mathbb{R}^{|\mathcal{S}^*|}\backslash\{0\}. \tag{49}$$

By the relationship (34), we can derive that (49) is stronger than (28). Assumption 3 is mainly used to ensure the nonsingularity for each element of the generalized Jacobian in the neighborhood of the reference point $\mathbf{z}^*$. We have the following result.

Lemma 7 *If Assumption 3 holds, there exist radius $\epsilon^* > 0$ and $\eta_3, \eta_4 > 0$ such that*

$$\eta_3 \le \inf_{\substack{H\in\widehat{\partial}^2 h(\mathbf{z}),\\ \mathbf{z}\in\mathcal{N}(\mathbf{z}^*,\epsilon^*)}} \boldsymbol{\lambda}_{\min}(H_{\mathcal{S}^*}) \le \sup_{\substack{H\in\widehat{\partial}^2 h(\mathbf{z}),\\ \mathbf{z}\in\mathcal{N}(\mathbf{z}^*,\epsilon^*)}} \boldsymbol{\lambda}_{\max}(H_{\mathcal{S}^*}) \le \eta_4. \tag{50}$$

Proof Under Assumption 3, for any $H^* \in \widehat{\partial}^2 h(\mathbf{z}^*)$, $H^*_{\mathcal{S}^*}$ is positive definite. According to [21, Prop. 7.1.4], $\partial^2 f^*(-A^\top \mathbf{z}^*)$ is a nonempty and compact set. Then the same property is true for $\widehat{\partial} h(\mathbf{z}^*)$ by the definition (33). Therefore, there exist $\eta_3, \eta_4 > 0$ such that

$$\inf_{H^*\in\widehat{\partial}^2 h(\mathbf{z}^*)} \boldsymbol{\lambda}_{\min}(H^*_{\mathcal{S}^*}) \ge 2\eta_3, \qquad \sup_{H^*\in\widehat{\partial}^2 h(\mathbf{z}^*)} \boldsymbol{\lambda}_{\max}(H^*_{\mathcal{S}^*}) \le \frac{\eta_4}{2}. \tag{51}$$

By the upper semi-continuity of $\partial^2 f^*$ (see [21, Prop. 7.1.4]), there exists $\epsilon^* > 0$ such that

$$\partial^2 f^*(-A^\top \mathbf{z}) \subseteq \partial^2 f^*(-A^\top \mathbf{z}^*) + \mathcal{N}(0, \delta_1^*), \ \forall \mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \epsilon^*),$$

where $\delta_1^* := \min\{\eta_3, \eta_4/2\}/\|A_{\mathcal{S}^*:}\|^2$. Then for any $\mathbf{z} \in \mathcal{N}(\mathbf{z}^*, \epsilon^*)$ and $H = AQA^\top \in \widehat{\partial} h(\mathbf{z})$, we can find $H^* = AQ^*A^\top \in \widehat{\partial} h(\mathbf{z}^*)$ with $\|Q - Q^*\| \le \delta_1^*$. Using the Weyl's perturbation theorem (see [6, Corollary III.2.6]), we have

$$\begin{aligned}&\max\{|\boldsymbol{\lambda}_{\min}(H_{\mathcal{S}^*}) - \boldsymbol{\lambda}_{\min}(H^*_{\mathcal{S}^*})|, |\boldsymbol{\lambda}_{\max}(H_{\mathcal{S}^*}) - \boldsymbol{\lambda}_{\max}(H^*_{\mathcal{S}^*})|\} \le \|H_{\mathcal{S}^*} - H^*_{\mathcal{S}^*}\| \\ &\le \|A_{\mathcal{S}^*:}(Q - Q^*)(A_{\mathcal{S}^*:})^\top\| \le \min\{\eta_3, \eta_4/2\}.\end{aligned}$$

Combining this with (51), we can derive (50). □

Remark 3 (i) (34) implies that for any $\overline{\mathbf{d}} \in \mathbb{R}^m$, $\overline{\mathbf{d}}^\top(\partial^2 h(\mathbf{z}))\overline{\mathbf{d}} \subseteq \overline{\mathbf{d}}^\top(\widehat{\partial}^2 h(\mathbf{z}))\overline{\mathbf{d}}$. Then by the Rayleigh-Ritz theorem (see [40, Chapter 5]), we can also obtain

$$0 < \eta_3 \le \inf_{\substack{H\in\partial^2 h(\mathbf{z}),\\ \mathbf{z}\in\mathcal{N}(\mathbf{z}^*,\epsilon^*)}} \boldsymbol{\lambda}_{\min}(H_{\mathcal{S}^*}) \le \sup_{\substack{H\in\partial^2 h(\mathbf{z}),\\ \mathbf{z}\in\mathcal{N}(\mathbf{z}^*,\epsilon^*)}} \boldsymbol{\lambda}_{\max}(H_{\mathcal{S}^*}) \le \eta_4. \tag{52}$$

(ii) Semismoothness of ∇f^* implies semismoothness of ∇h. This can be verified by [21, Thm. 7.4.3]. Therefore, Assumption 3 ensures the SOSC in Prop. 4.

(iii) As we have shown in Prop. 5, if Assumption 1 holds, $\mathbf{z}^*$ will be a P-stationary point and F is constant on all the accumulation points of $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$. If we further take Assumption 3 into consideration, SOSC holds and Prop. 4 ensure that each accumulation point of $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ is isolated. Since $\lim_{k\to\infty} \|\mathbf{z}^{k+1} - \mathbf{z}^k\| = 0$ (derived by (42)), it follows from [31] that $\lim_{k\to\infty} \mathbf{z}^k = \mathbf{z}^*$. Then from (42), we also have $\lim_{k\to\infty} \mathbf{v}^k = \mathbf{z}^*$.

To derive the local superlinear convergence of SGSN, we need to overcome several obstacles. First, SGSN performs Newton iterate on changing subspace $\mathbb{R}^{|T_k|}$. The technique for the convergence rate of classic Newton method is unavailable unless

$\mathcal{S}^*$ is exactly identified by T_k. Moreover, the full-Newton step is only implemented when (**C1**), (**C2**) and $\alpha_k = 1$ are met. In the subsequent lemma, we show that these conditions hold after finite iterations.

Lemma 8 *Suppose that Assumptions 1 and 3 hold. Then there exists a sufficiently large $k^* \in \mathbb{N}$ such that for $k \geq k^*$*

(i) The support set of $\mathbf{z}^*$ is identified by T_k, i.e.

$$T_k = \mathcal{S}(\mathbf{v}_k) = \mathcal{S}^*. \tag{53}$$

(ii) The step length α_k always equals to 1.

(iii) If $c_1 \leq \eta_3/4$ and $c_2 \geq \ell_h + \eta_4 + 1$, then the Newton iterate $\widetilde{\mathbf{z}}^{k+1}$ is always accepted.

Proof (i) As we just discussed, $\lim_{k\to\infty} \mathbf{v}^k = \mathbf{z}^*$ holds. Then when k is sufficiently large, we have $\mathcal{S}(\mathbf{v}^k) \supseteq \mathcal{S}^*$. Moreover, according to Prop. 5, $\lim_{k\to\infty} F(\mathbf{v}^k) = F(\mathbf{z}^*)$ holds, which further leads to $\lim_{k\to\infty} \|\mathbf{v}^k\|_0 = \|\mathbf{z}^*\|_0$. Noticing that the values of $\|\cdot\|_0$ are from the set $\{0, 1, \ldots, m\}$, therefore $\|\mathbf{v}^k\|_0 = \|\mathbf{z}^*\|_0$ must be true when k is large enough. Combining this with $\mathcal{S}(\mathbf{v}^k) \supseteq \mathcal{S}^*$, we can derive assertion (i).

(ii) Now let us consider k is large enough such that $T_k = \mathcal{S}^*$ holds. From Lem. 7, $H^k_{T_k}$ is nonsingular and we can estimate

$$\|\mathbf{d}^k_{T_k}\| \leq \|(H^k_{T_k} + \gamma_k I)^{-1} \nabla_{T_k} h(\mathbf{v}^k)\| \leq \|\nabla_{T_k} h(\mathbf{v}^k)\|/\eta_3.$$

Since $\nabla_{\mathcal{S}^*} h(\mathbf{z}^*) = 0$ holds from (22), we have $\lim_{k\to\infty} \|\mathbf{d}^k_{T_k}\| = 0$. Meanwhile, since for any $i \in T_k = \mathcal{S}^*$, we have $\lim_{k\to\infty} v^k_i = z^*_i > 0$. Then according to (37), when k is large enough $\beta_k \geq 1$, and thus $\alpha_k = 1$.

(iii) We need to show that (**C1**) and (**C2**) hold when k is large enough. Let us begin with some preliminary estimations. Using $\nabla_{T_k} h(\mathbf{z}^*) = 0$ and Lipschitz continuity of h, we can estimate

$$\gamma_k = \|\nabla_{T_k} h(\mathbf{v}^k)\| = \|\nabla_{T_k} h(\mathbf{v}^k) - \nabla_{T_k} h(\mathbf{z}^*)\| \leq \ell_h \|\mathbf{v}^k - \mathbf{z}^*\|. \tag{54}$$

Let us denote $\Phi := \nabla f^*$ and $H^k = AQ^kA^\top$ with $Q^k \in \partial^2 f^*(-A^\top \mathbf{v}^k) = \partial_C \Phi(-A^\top \mathbf{v}^k)$. Since $\alpha_k = 1$ when k is sufficiently large, from (35) and (36), we have

$$\begin{aligned}
&\|\widetilde{\mathbf{z}}^{k+1} - \mathbf{z}^*\| \\
=&\|(\widetilde{\mathbf{z}}^{k+1} - \mathbf{z}^*)_{T_k}\| = \|(\mathbf{v}^k - \mathbf{z}^*)_{T_k} - (H^k_{T_k} + \gamma_k I)^{-1} \nabla_{T_k} h(\mathbf{v}^k)\| \\
\leq&\frac{1}{\eta_3} \|(H^k_{T_k} + \gamma_k I)(\mathbf{v}^k - \mathbf{z}^*)_{T_k} - (\nabla h(\mathbf{v}^k) - \nabla h(\mathbf{z}^*))_{T_k}\| \\
=&\frac{1}{\eta_3} \|H^k_{T_k}(\mathbf{v}^k - \mathbf{z}^*)_{T_k} - (\nabla h(\mathbf{v}^k) - \nabla h(\mathbf{z}^*))_{T_k}\| + o(\|\mathbf{v}^k - \mathbf{z}^*\|) \\
=&\frac{1}{\eta_3} \|A_{T_k:} Q^k A^\top_{T_k:} (\mathbf{v}^k - \mathbf{z}^*)_{T_k} + A_{T_k:}(\Phi(-A^\top \mathbf{v}^k) - \Phi(-A^\top \mathbf{z}^*))\| + o(\|\mathbf{v}^k - \mathbf{z}^*\|)
\end{aligned}$$

$$\begin{aligned}&\leq(\|A\|/\eta_3)\|\Phi(-A^\top \mathbf{v}^k)-\Phi(-A^\top \mathbf{z}^*)-Q^k(-A^\top(\mathbf{v}^k-\mathbf{z}^*))\|+o(\|\mathbf{v}^k-\mathbf{z}^*\|)\\&=o(\|\mathbf{v}^k-\mathbf{z}^*\|),\end{aligned}\tag{55}$$

where the second and last equations follow from (54) and the semismoothness property (14) respectively. The above relationship also implies $\lim_{k\to\infty}\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|=0$. From (35) and (54), we can also estimate

$$\begin{aligned}\|\nabla_{T_k}h(\mathbf{v}^k)\|=&\|(H_{T_k}^{k+1/2}+\gamma_k I)(\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k)_{T_k}\|\\\leq&\eta_4\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|+o(\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|).\end{aligned}$$

Now we are ready to prove (**C2**). When k is sufficiently large, it follows from $o(\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|)\leq\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|$ and Lipschitz continuity of h that

$$\begin{aligned}\|\nabla_{T_k}h(\widetilde{\mathbf{z}}^{k+1})\|\leq&\|\nabla_{T_k}h(\mathbf{v}^k)\|+\|\nabla_{T_k}h(\mathbf{v}^k)-\nabla_{T_k}h(\widetilde{\mathbf{z}}^{k+1})\|\\\leq&(\ell_h+\eta_4+1)\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|.\end{aligned}$$

Finally, let us prove (**C1**). According to [21, Prop. 7.4.10] and (55), we have

$$\begin{aligned}h(\widetilde{\mathbf{z}}^{k+1})=&h(\mathbf{z}^*)+\langle\nabla h(\mathbf{z}^*),\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*\rangle+\frac{1}{2}(\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*)^\top\widetilde{H}^{k+1}(\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*)\\&+o(\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*\|^2)\\=&h(\mathbf{z}^*)+\langle\nabla h(\mathbf{z}^*),\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*\rangle+o(\|\mathbf{v}^k-\mathbf{z}^*\|^2)\end{aligned}\tag{56}$$

$$\begin{aligned}h(\mathbf{v}^k)=&h(\mathbf{z}^*)+\langle\nabla h(\mathbf{z}^*),\mathbf{v}^k-\mathbf{z}^*\rangle+\frac{1}{2}(\mathbf{v}^k-\mathbf{z}^*)^\top P^k(\mathbf{v}^k-\mathbf{z}^*)\\&+o(\|\mathbf{v}^k-\mathbf{z}^*\|^2),\end{aligned}\tag{57}$$

for any $\widetilde{H}^{k+1}\in\partial^2h(\widetilde{\mathbf{z}}^{k+1})$ and $P^k\in\partial^2h(\mathbf{v}^k)$. Moreover, using $T^k=\mathcal{S}^*$ and $\nabla_{S^*}h(\mathbf{z}^*)=0$, we have $\langle\nabla h(\mathbf{z}^*),\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*\rangle=\langle\nabla_{T_k}h(\mathbf{z}^*),(\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*)_{T_k}\rangle=0$ and $\langle\nabla h(\mathbf{z}^*),\mathbf{v}^k-\mathbf{z}^*\rangle=\langle\nabla_{T_k}h(\mathbf{z}^*),(\mathbf{v}^k-\mathbf{z}^*)_{T_k}\rangle=0$. Then subtracting (56) and (57) leads to

$$h(\mathbf{v}^k)-h(\widetilde{\mathbf{z}}^{k+1})\geq\frac{\eta_3}{2}\|\mathbf{v}^k-\mathbf{z}^*\|^2+o(\|\mathbf{v}^k-\mathbf{z}^*\|^2).$$

From $\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|\leq\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{z}^*\|+\|\mathbf{v}^k-\mathbf{z}^*\|$ and (55), we have $\|\widetilde{\mathbf{z}}^{k+1}-\mathbf{v}^k\|^2\leq\|\mathbf{v}^k-\mathbf{z}^*\|^2+o(\|\mathbf{v}^k-\mathbf{z}^*\|^2)$, and thus we can verify

$$\begin{aligned}h(\mathbf{v}^k)-h(\widetilde{\mathbf{z}}^{k+1})\geq&\frac{\eta_3}{4}\|\mathbf{v}^k-\mathbf{z}^*\|^2+\frac{\eta_3}{4}\|\mathbf{v}^k-\widetilde{\mathbf{z}}^{k+1}\|^2+o(\|\mathbf{v}^k-\mathbf{z}^*\|^2)\\\geq&\frac{\eta_3}{4}\|\mathbf{v}^k-\widetilde{\mathbf{z}}^{k+1}\|^2.\end{aligned}$$

Since $\widetilde{\mathbf{z}}^{k+1}_{T_k} = 0$ and $T_k = \mathcal{S}(\mathbf{v}^k)$, we have $g(\widetilde{\mathbf{z}}^{k+1}) \le g(\mathbf{v}^k)$. Finally, we can prove (**C1**) by

$$F(\mathbf{v}^k) - F(\widetilde{\mathbf{z}}^{k+1}) \ge h(\mathbf{v}^k) - h(\widetilde{\mathbf{z}}^{k+1}) \ge \frac{\eta_3}{4}\|\mathbf{v}^k - \widetilde{\mathbf{z}}^{k+1}\|^2.$$

Overall, we have completed the proof. □

Lem. 8(iii) indicates that c_1 and $1/c_2$ should be small enough to ensure the acceptance of Newton iterate $\widetilde{\mathbf{z}}^{k+1}$. In fact, we can give a more specific lower bound for c_2. By the $(1/\sigma_f)$-smoothness of f^*, the upper bound in (51) will be $\sup_{H^*\in\widehat{\partial}^2 h(\mathbf{z}^*)} \boldsymbol{\lambda}_{\max}(H^*_{\mathcal{S}^*}) \le \|A\|^2/\sigma_f = \ell_h := \eta_4/2$. Then we can set $c_2 \ge 3\ell_h + 1$. Moreover, an upper bound for c_1 can be computed in a special case that f is ℓ_f-smooth and $A_{\mathcal{S}^*:}$ has full row rank. In this case, f^* is $(1/\ell_f)$-strongly convex, and thus the lower bound in (51) will be $\inf_{H^*\in\widehat{\partial}^2 h(\mathbf{z}^*)} \boldsymbol{\lambda}_{\min}(H^*_{\mathcal{S}^*}) \ge \sqrt{\boldsymbol{\lambda}_{\min}(A_{\mathcal{S}^*:}A^\top_{\mathcal{S}^*:})}/\ell_f := 2\eta_3$. Then we can set $c_1 \le \sqrt{\boldsymbol{\lambda}_{\min}(A_{\mathcal{S}^*:}A^\top_{\mathcal{S}^*:})}/(8\ell_f)$.

We are ready to state the local superlinear convergence of SGSN.

Theorem 3 *Suppose that Assumptions 1 and 3 hold. If we set $c_1 \le \eta_3/4$ and $c_2 \ge \ell_h + \eta_4 + 1$, then the sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ converges to $\mathbf{z}^*$ with local superliner rate, i.e. when k is sufficiently large, we have*

$$\|\mathbf{z}^{k+1} - \mathbf{z}^*\| \le o(\|\mathbf{z}^k - \mathbf{z}^*\|).$$

Proof Since we have proved (55) and $\mathbf{z}^{k+1} = \widetilde{\mathbf{z}}^{k+1}$ when k is sufficiently large, we have

$$\|\mathbf{z}^{k+1} - \mathbf{z}^*\| \le o(\|\mathbf{v}^k - \mathbf{z}^*\|). \tag{58}$$

It suffices to find the relationship between $\|\mathbf{z}^k - \mathbf{z}^*\|$ and $\|\mathbf{v}^k - \mathbf{z}^*\|$. Using (32), $\nabla_{T_k} h(\mathbf{z}^*) = 0$, and Lipschitz continuity of h, we have

$$\begin{aligned}\|\mathbf{v}^k - \mathbf{z}^*\| =& \|(\mathbf{v}^k - \mathbf{z}^*)_{T_k}\| = \|(\mathbf{z}^k - \mathbf{z}^*)_{T_k} - \tau(\nabla h(\mathbf{z}^k) - \nabla h(\mathbf{z}^*))_{T_k}\| \\ \le& (\ell_h\tau + 1)\|\mathbf{z}^k - \mathbf{z}^*\|.\end{aligned}$$

Combining this with (58), we can derive our desired conclusion. □

Remark 4 If the semismoothness of ∇f^* in Assumption 3 is strengthened to strong semismoothness (see e.g. [21]), then similar to the classical result in [55], the convergence rate in Thm. 3 can be improved to be quadratic.

As emphasized in Introduction, various types of semismooth Newton methods have been recently developed for nonsmooth equations arising from optimization, see e.g., [30, 35, 36, 39, 67, 69]. We make detailed comments below on three such methods: Classical semismooth Newton method [55], Coderivative-based Newton method [35], and Subspace Newton method [69]. The comparisons will highlight the unique features

of our method and will be also useful when conducting numerical comparison in the next section.

Remark 5 (**Motivation from classical semismooth Newton method**) In general, classical semismooth Newton method serves a guidance for constructing new methods for nonsmooth equations. Let us consider the proximal gradient equation:

$$0 = G(\mathbf{z}) := \mathbf{z} - \mathrm{Prox}_{\tau g(\cdot)}(\mathbf{z} - \tau \nabla h(\mathbf{z})). \tag{59}$$

This equation is well-defined at any P-stationary point when τ is sufficiently small, see Lem. 1(vi). We assume that $G(\cdot)$ is locally Lipschitzian and semismooth. A semismooth/generalized Newton method gives the following update at $\mathbf{z}^k$:

$$\begin{cases} (A^k + \gamma_k I)\mathbf{d}^k = -G(\mathbf{z}^k), \quad A^k \in \partial_C G(\mathbf{z}^k) \\ \mathbf{z}^{k+1} = \mathbf{z}^k + \mathbf{d}^k, \end{cases} \tag{60}$$

where $\partial_C G(\cdot)$ is the generalized Jacobian in the sense of Clarke [16] and $\gamma_k \geq 0$ is a regularization parameter. Various generalized Newton methods mainly differ in three aspects: (i) using different generalized Jacobian instead of $\partial_C G(\cdot)$, (ii) proposing approximation property of $G(\cdot)$ in terms of the proposed Jacobian, and (iii) developing globalization strategy. For (i), we used the generalized Jacobian $\widehat{\partial}^2 h(\mathbf{z})$ defined in (33). For (ii), we assumed the semismoothness of $\nabla f^*(\cdot)$. We make more comments on Point (iii).

It is often a nontrivial task to globalize the Newton method in (60). A common strategy is to use Newton direction $\mathbf{d}^k$ if certain conditions hold and to use a gradient descent direction otherwise. Our method roughly follows this strategy. Firstly, we work with the set-valued mapping $\mathrm{Prox}_{\tau g}(\cdot)$ and select an element $\mathbf{v}^k$ from the set of proximal gradients defined in (31). Secondly, according to [35, Prop. 2.3], $\mathrm{Prox}_{\tau g}(\cdot)$ is Lipschitz continuous and single-valued around $\mathbf{z}^* - \tau \nabla h(\mathbf{z}^*)$ when τ is sufficiently small and $\mathbf{z}^*$ is a P-stationary point. This result is useful because locally the Newton equation (35) in SGSN can be derived from (60) using the index set T_k and the associated properties established in Lem. 1. In our SGSN, τ can take any value in the range $(0, 1/\ell_h)$. Thirdly, this is probably the most difficult part and it is to decide when to use the gradient direction $\mathbf{v}^k$ or Newton direction $\mathbf{d}^k$. Our selection is decided by the two conditions (**C1**) and (**C2**), which are carefully designed for the Newton step to be eventually accepted and for the convergence to hold for a large class of PŁK functions $F(\mathbf{z}) = h(\mathbf{z}) + g(\mathbf{z})$.

In summary, our SGSN algorithm is motivated by the classical Semismooth Newton method. But its convergence does not follow from existing convergence theory due to some innovative strategies proposed, especially on (**C1**) and (**C2**).

Remark 6 (**Comparison with a coderivative-based Newton method**) When $h \in \mathrm{C}^2$ (twice continuously differentiable) and g is prox-regular, the coderivative-based Newton (CBN) method [35, Alg. 4] can be applied for Problem (21). It first implements a proximal gradient step (32) to compute $\mathbf{v}^k$, and followed by a Newton step to compute direction $\mathbf{d}^k$:

$$-\mathbf{s}^k - H^k\mathbf{d}^k \in \partial^2 g(\mathbf{v}^k, \mathbf{r}^k)(\mathbf{d}^k), \tag{61}$$

where $H^k = \nabla^2 h(\mathbf{v}^k)$, $\mathbf{s}^k := \nabla h(\mathbf{v}^k) - \nabla h(\mathbf{z}^k) + (\mathbf{z}^k - \mathbf{v}^k)/\tau$, $\mathbf{r}^k := \mathbf{s}^k - \nabla h(\mathbf{v}^k)$ and the second-order subdifferential (see e.g. [44, Def. 3.17]) of g at $\mathbf{v}$ relative to $\mathbf{r} \in \partial g(\mathbf{v})$ is defined by

$$\partial^2 g(\mathbf{v}, \mathbf{r})(\mathbf{d}) = \left\{ \mathbf{q} \in \mathbb{R}^m \;\middle|\; q_i \in \begin{cases} \{t \mid td_i = 0\}, & \text{if } v_i = r_i = 0, \\ \mathbb{R}, & \text{if } v_i = 0, r_i \neq 0, d_i = 0, \\ \{0\}, & \text{if } v_i > 0, \\ \emptyset, & \text{otherwise.} \end{cases} \right\} \tag{62}$$

using the notation $T_k = \mathcal{S}(\mathbf{v}^k)$ and $\mathbf{v}^k_{\overline{T}_k} = 0$ from (39), we can compute a solution of (61) by solving the following linear equation:

$$H^k_{T_k}\mathbf{d}^k_{T_k} = -\mathbf{s}^k_{T_k} \text{ and } \mathbf{d}^k_{\overline{T}_k} = 0.$$

Ignoring the regularization term $\gamma_k I$ in (35), CBN and SGSN appear similar. But they are very different in a few aspects.

(A) Globalization strategies are different. Although both methods share the same coefficient matrix $H^k_{T_k}$, the right-hand side vectors are different, $-\mathbf{s}^k_{T_k}$ vs $-\nabla_{T_k} h(\mathbf{v}^k)$. Moreover, CBN adopts a line search technique based on a sufficient descent on a forward-back envelope [35, Def. 5.1] for globalization. In contrast, SGSN uses (37) and Conditions (**C1**) or (**C2**).

(B) Assumptions on global convergence are different. CBN requires $h \in C^2$ and g being prox-regular for its global convergence under the additional assumption that an accumulation point is isolated. In contrast, SGSN assume PŁK property of the objective function F, allowing the accumulation points to be non-isolated.

(C) Assumptions on superlinear convergence are different. Under the premise of global convergence to a P-stationary point $\mathbf{z}^*$, the local superlinear convergence of CBN further requires

(i) g is continuously prox-regular (see e.g. [44, Def. 3.27]) at $\mathbf{z}^*$ for $-\nabla h(\mathbf{z}^*)$.
(ii) $\nabla^2 h$ is strictly differentiable (see e.g. [57, Def. 9.17]) at $\mathbf{z}^*$.
(iii) $\mathbf{z}^*$ is a tilt-stable local minimizer (see e.g. [51, Def. 1.1]) of (21).

Similar to the proof of [35, Thm. 7.3], (i) holds if and only if $[\mathbf{z}^* + \nabla h(\mathbf{z}^*)]_i \neq 0$ for all $i \in [m]$, and then through the use of representation (62), [43, Prop. 1.121] and [51, Thm. 1.3(b)], one can derive that (iii) is equivalent to positive definiteness of $H^*_{\mathcal{S}^*}$, where $H^* := \nabla^2 h(\mathbf{z}^*)$. Meanwhile, the local superlinear convergence of SGSN is guaranteed by Assumption 3. Particularly, in the case that h is C^2-smooth, this assumption becomes positive definiteness of $H^*_{\mathcal{S}^*}$. Overall, convergence of SGSN is established under weaker assumptions.

Remark 7 **(Comparison with a subspace Newton method)** [69] proposed a Newton method for ℓ_0-regularized optimization (NL0R). It also implements Newton step

on a subspace determined by certain index sets. Moreover, it enjoys global convergence with local quadratic rate under suitable conditions. However, both methods are significantly different.

(i) Problem types: NL0R solves the minimization of a C^2-smooth function with ℓ_0 regularization, while SGSN is designed for a more challenging problem (21), where h is SC^1 and g comprises a ℓ_0 regularizer and a nonnegative constraint. NL0R cannot be directly applied to (21). The nonsmoothness of ∇h and the constraint in g demand more sophisticated algorithmic design to ensure both feasibility and convergence of the iterates.

(ii) Globalization strategy: NL0R adopts Armijo line search on Newton direction. In comparison, SGSN combines proximal gradient and semismooth Newton steps with the aid of conditions (**C1**) and (**C2**). This helps us utilize PŁK property to globalize SGSN.

(iii) Convergence: The whole sequence convergent of NL0R requires existence of an isolated accumulation point, whereas SGSN adopts the PŁK property. For local quadratic convergence to a solution $\mathbf{z}^*$, NL0R requires that the smooth part in the objective function has positive definite and Lipschitz continuous Hessian around $\mathbf{z}^*$. In comparison, SGSN requires weaker assumptions that h is strongly semismooth and its generalized Hessian at $\mathbf{z}^*$ is positive definite on a subspace indexed by the support set $\mathcal{S}^*$.

5 Numerical experiments

In this section, we demonstrate the performance of SGSN on AUC maximization and sparse multi-label classification. All the experiments are implemented on Matlab 2022a by a laptop with 32GB memory and Intel CORE i7 2.6 GHz CPU.

5.1 AUC maximization

The area under the receiver operating characteristic curve (AUC) [28] is a widely used evaluation metric in imbalanced classification and anomaly detection. Let $X^+ = [\mathbf{x}_1^+, \cdots, \mathbf{x}_{q_+}^+]^\top \in \mathbb{R}^{q_+ \times n}$ and $X^- = [\mathbf{x}_1^-, \cdots, \mathbf{x}_{q_-}^-]^\top \in \mathbb{R}^{q_- \times n}$ be the positive and negative sample matrices respectively, then the AUC associated with parameter vector $\mathbf{x} \in \mathbb{R}^n$ can be represented as

$$\mathrm{AUC}(\mathbf{x}) := \frac{1}{q_+ q_-} \sum_{i=1}^{q_+} \sum_{j=1}^{q_-} \mathbf{1}_{(0,\infty)} \left(\langle \mathbf{x}_i^+, \mathbf{x} \rangle - \langle \mathbf{x}_j^-, \mathbf{x} \rangle \right).$$

After an opportune scaling of $\mathbf{x}$ (a trick commonly used in support vector machines), each term $(\langle \mathbf{x}_i^+, \mathbf{x} \rangle - \langle \mathbf{x}_j^-, \mathbf{x} \rangle)$ is replaced by $(\langle \mathbf{x}_i^+, \mathbf{x} \rangle - \langle \mathbf{x}_j^-, \mathbf{x} \rangle - 1)$, see e.g. [23]. This scaling operation also removes the trivial solution $\mathbf{x}^* = 0$.

Therefore, AUC can be maximized by solving composite optimization (1) with the following setting:

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2,\ A = (\mathbf{e}_{q_+} \otimes I_{q_-})X^- - (I_{q_+} \otimes \mathbf{e}_{q_-})X^+ \in \mathbb{R}^{m\times n},\ \mathbf{b} = \mathbf{e}_m, \quad (63)$$

where $\otimes$ is the Kroneker product and $m := q_+q_-$. Then the conjugate function $f^*(\mathbf{v}) = \|\mathbf{v}\|^2/2$, and thereby the associated dual problem (21) can be written as

$$\min_{\mathbf{z}\in\mathbb{R}^m} F(\mathbf{z}) = \frac{1}{2}\|A^\top \mathbf{z}\|^2 - \langle \mathbf{b}, \mathbf{z}\rangle + g(\mathbf{z}). \quad (64)$$

In this case, f is semi-algebraic, f^* is strongly convex and $\nabla^2 f^*$ is Lipschitz continuous. When SGSN is applied to solving (64), provided that the generated sequence is bounded, then it must converge to a P-stationary point $\mathbf{z}^*$ of (64) according to Thm. 2. Additionally, if $A_{\mathcal{S}^*:}$ is a full row-rank matrix, then the sequence converges with local quadratic rate by Remark 4.

In the experiment of AUC maximization, we select four other algorithms for comparison. NM01 [70] is a subspace Newton-type method designed for solving the primal Problem (1) with a twice continuously differentiable f. CBN [35, Alg. 4] was already commented in Remark 7. PRSVM [13] is a semismooth Newton method which optimizes AUC with a squared hinge surrogate loss function. OPAUC [23] is a one-pass gradient descent method for maximizing AUC with a least square surrogate loss function.

In the setting of (63) and (64), we compute the Lipschitzian constant of ∇h by $\ell_h = \|A\|^2 = q_+\|X^-\|^2 + q_-\|X^+\|^2 - 2\langle \mathbf{e}_{q_-}^\top X^-, \mathbf{e}_{q_+}^\top X^+\rangle$. Then for SGSN, we set $\mu = \tau = 1/(2\ell_h), \gamma = 10^{-1}, c_1 = 1/(3\ell_h)$ and $c_2 = 3\ell_h$. In our own implementation of CBN [35, Alg. 4] with parameter fine-tuning, we particularly set the proximal gradient step length $\tau = 1/(2\ell_h)$, the descent quantity coefficient $\sigma = 1/(20\ell_h)$, and the scaling factor of line search $\beta = 0.1$. The parameter setting of other algorithms will be given in the comparison part.

5.1.1 Convergence test

In this subsection, our main goal is to examine the convergence performance of SGSN. The following numerical example generates various types of simulated datasets for convergence tests.

Example 1 We first generate vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^n$ with independent and identically distributed (i.i.d.) elements being selected from standard normal distribution $N(0, 1)$. Each element of diagonal covariance matrices $\Lambda_1, \Lambda_2 \in \mathbb{R}^{n\times n}$ is also i.i.d., following a folded normal distribution. Then positive and negative samples are normally distributed with $N(\boldsymbol{\mu}_1, \Lambda_1)$ and $N(\boldsymbol{\mu}_2, \Lambda_2)$ respectively. Let q be the total number of samples and $p \in (0, 1)$ be the proportion of positive samples, then $q_+ = \lceil pq \rceil$ and $q_- = q - q_+$. Given noise ratio $r \in (0, 1)$, $\lfloor rq_+ \rfloor$ positive and negative samples are chosen to be marked with reverse labels respectively.

We use the following violation of dual optimality (VDO) to measure the closeness between a dual variable $\mathbf{z}$ and a P-stationary point of (21)

$$\text{VDO} := \text{dist}(\mathbf{z},\ \text{Prox}_{\tau g(\cdot)}(\mathbf{z} - \tau\nabla h(\mathbf{z})))/\tau.$$

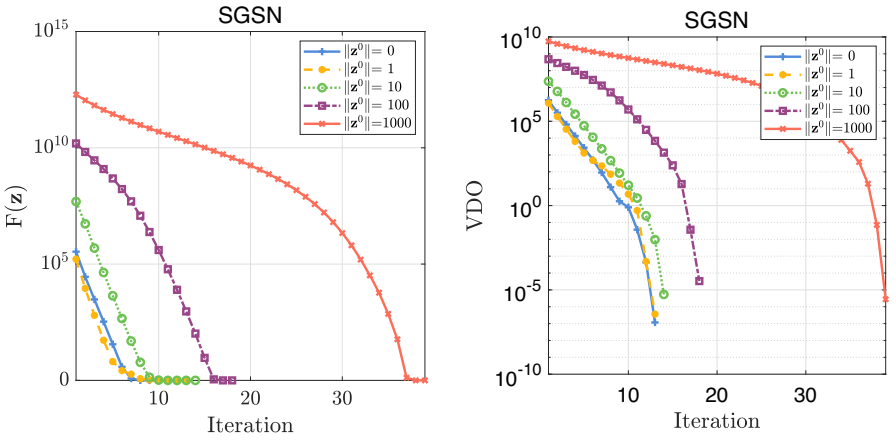

Fig. 2 Convergence test for SGSN with different initial points

We also record the value of objective function $F(\mathbf{z})$ and computational time (TIME) of the algorithm.

• **Test 0: On change of $\mathbf{z}^0$.** This test aims to observe the basic convergence property of SGSN with different initial points $\mathbf{z}^0$. It is conducted on a simulated dataset with $q = 1000, n = 100, p = 0.05$ and $r = 0$. To produce different initial points for SGSN, we first generate a vector $\bar{\mathbf{z}}^0 \in \mathbb{R}^n$ with entries being i.i.d. samples from the uniform distribution $U(0, 1)$. Then we compute $\mathbf{z}^0 = \rho\bar{\mathbf{z}}^0/\|\bar{\mathbf{z}}^0\|$ with $\rho \in \{0, 1, 10, 100, 1000\}$. SGSN starts with different $\mathbf{z}^0$ and the corresponding VDO and $F(\mathbf{z})$ are recorded in Figure 2. The left panel in Figure 2 demonstrates the sufficient descent of $\{F(\mathbf{z}^k)\}_{k\in\mathbb{N}}$, which has been proved in Prop. 5. The right panel shows the descent of VDO , with a faster rate during the last few iterations. This phenomenon illustrates the global and local quadratic convergence of $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ given in Thms. 2 and 3.

Next, we test SGSN with $\mathbf{z}^0 = 0$ on simulated datasets with various q (total number of samples), n (number of features), p (proportion of positive samples), and r (noise ratio). For comparison, we select NM01 and CBN to solve the primal problem (1) and the dual problem (21) respectively. Their convergence performance can be measured by VDO and violation of primal optimality (VPO) respectively, where VPO is defined as follows:

$$\texttt{VPO} := \max\left\{\|\nabla f(\mathbf{x}) + A^\top\mathbf{z}\|,\ \operatorname{dist}\left(A\mathbf{x}+\mathbf{b},\ \operatorname{Prox}_{\xi\boldsymbol{\mathcal{I}}(\cdot)}(A\mathbf{x}+\mathbf{b}+\xi\mathbf{z})\right)/\xi\right\},$$

where ξ is a proximal parameter adjusted as the defaulted setting of NM01.

• **Test 1: On change of q.** In this test, we generate datasets with $q \in \{5000, 10000, 20000\}$, $n = 100$, $p = 0.2$ and $r = 0$. In this setting, m (the number of rows in matrix A) is taken from $\{4 \times 10^6, 1.6 \times 10^7, 6.4 \times 10^7\}$. SGSN and CBN demonstrate better convergence performance when dealing with these large-scale matrices. From Figure 3, we can also see that SGSN converges faster at the initial stage compared with CBN. At the final few steps, when the iterate is close to a P-stationary point of (21), VDO of SGSN rapidly drops below 10^{-4}. In comparison,

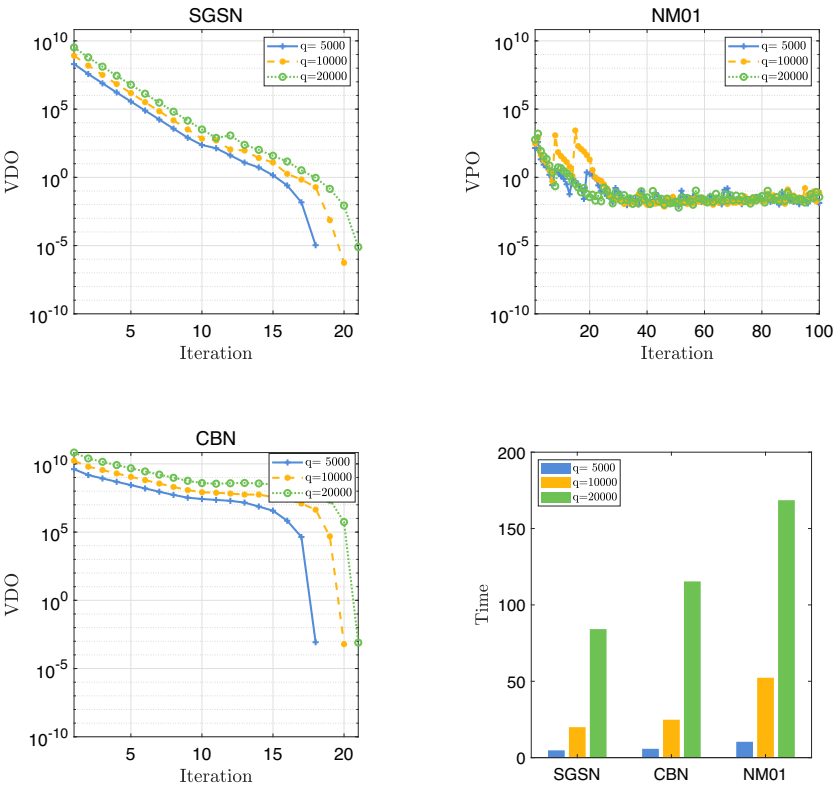

Fig. 3 Convergence comparison on the simulated data with change of q

the VPO of NM01 decreases relatively faster during the first several iterations and then stagnates around 10^{-2}. In this test, SGSN spends the shortest TIME to achieve the smallest VDO.

• **Test 2: On change of n.** We take $n \in \{5000, 10000, 20000\}$, $q = 1000$, $p = 0.2$ and $r = 0$. Figure 4 shows that all the algorithms converge very fast in the cases of $n \geq 10000$. When $n = 5000$, NM01 requires more iterations for VPO to get below 10^{-4}.

• **Test 3: On change of p.** We alter $p \in \{0.01, 0.05, 0.1\}$, and fix $q = 1000$, $n = 10000$ and $r = 0$. In all the cases, SGSN and CBN demonstrate significant decrease on VDO along with iteration. We also observe from Figure 5 that NM01 spends more TIME and requires more iterations when $p = 0.05$ and 0.1.

• **Test 4: On change of r.** Finally, we test cases of $r \in \{0.01, 0.05, 0.1\}$, $q = 1000$, $n = 10000$ and $p = 0.2$. As shown in Figure 6 again, compared with the other two algorithms, SGSN has the smallest VDO with the shortest TIME. The VPO of NM01 decreases faster over the first several iterations, and then fluctuates around 10^{-3}.

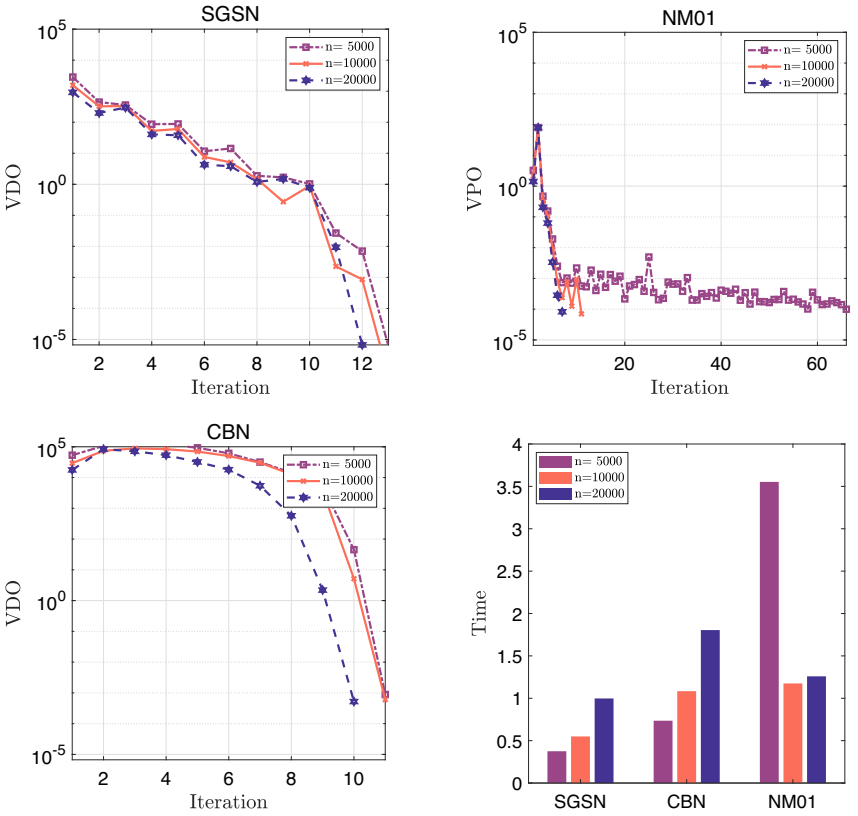

Fig. 4 Convergence comparison on the simulated data with change of n

5.1.2 Experiments on real dataset

In this part, we test SGSN and the three selected algorithms on the real datasets given in Table 1. For each dataset, five-fold cross-validation is conducted and the average value of AUC and TIME for each algorithm are recorded. As for the parameters of the comparison algorithms, NM01 adopts the default setting. The weighted parameter for the loss function in PRSVM and OPAUC is chosen from $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$ with the highest AUC. Moreover, we stop SGSN and CBN when $\mathrm{VDO}_k/\mathrm{VDO}_1 \leq 10^{-3}$ or $|\mathrm{VDO}_k - \mathrm{VDO}_{k-1}| < 10^{-3}$, where VDO_k denotes the violation of dual optimality at the k-th iterate. NM01 is terminated by a similar rule, replacing the above VDO with VPO.

Example 2 The details of real datasets[2345] for AUC maximization are summarized in Table 1. All the datasets are preprocessed by feature-wise scaling to the interval $[-1, 1]$.

[2] https://jundongl.github.io/scikit-feature/

[3] https://www.openml.org/

[4] https://www.refine.bio/

[5] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

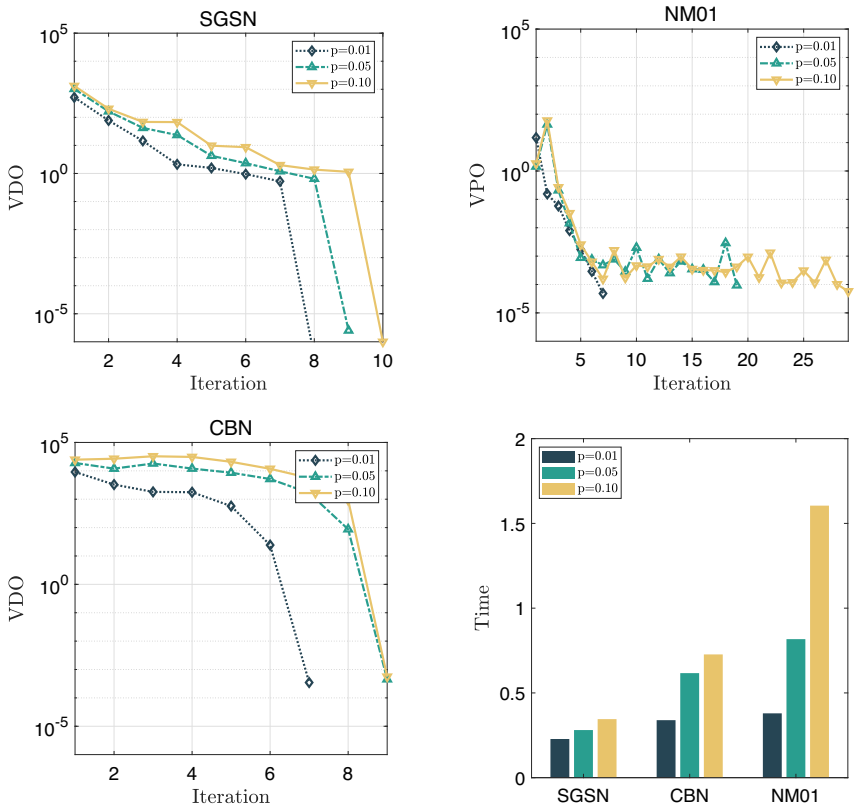

Fig. 5 Convergence comparison on the simulated data with change of p

The comparison results are summarized in Table 2. We can observe that SGSN has a similar AUC but shorter TIME than NM01 or CBN because it is more effective in reducing the violation of optimality. Although the other two algorithms are also competitive on AUC, their TIME are longer than that of SGSN on most of the datasets. For example, on the dataset ovk, our SGSN only requires about 10% of the TIME of PRSVM to achieve a higher AUC.

5.2 Sparse multi-label classification

For a multi-label classification (MLC) problem with ℓ labels, let $\mathbf{c}_i \in \mathbb{R}^d$, $i \in [q]$ be the feature vector with its last entry $[\mathbf{c}_i]_d = 1$ and class label $\mathbf{y}_i \in \{-1, 1\}^\ell$ for $i \in [q]$. Then we denote matrices $\mathbf{C} := [\mathbf{c}_1, \cdots, \mathbf{c}_q]^\top \in \mathbb{R}^{q\times d}$ and $\mathbf{Y} := [\mathbf{y}_1, \cdots, \mathbf{y}_q]^\top \in \mathbb{R}^{q\times \ell}$. Given arbitrary feature vector $\mathbf{c}$, a linear binary relevance classifier $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}^\ell$ returns a label vector by the following rule [12, 62]:

$$\mathcal{H}(\mathbf{c}) := \left(\operatorname{sgn}(\mathbf{c}^\top \mathbf{x}_1), \cdots, \operatorname{sgn}(\mathbf{c}^\top \mathbf{x}_\ell)\right),$$

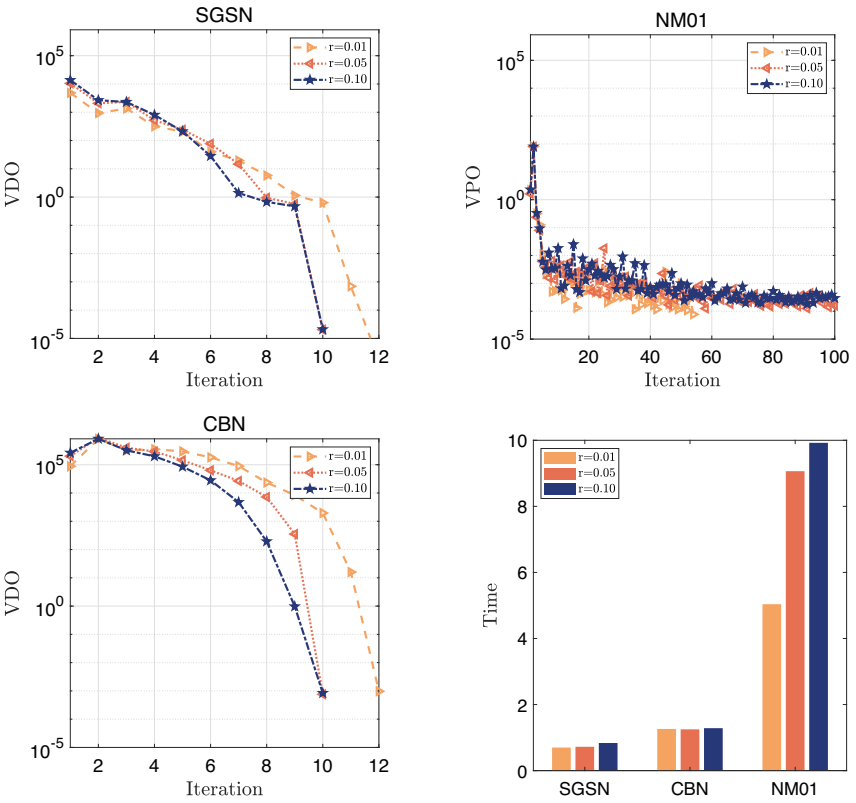

Fig. 6 Convergence comparison on the simulated data with change of r

where $\mathbf{x} := (\mathbf{x}_1; \cdots; \mathbf{x}_\ell) \in \mathbb{R}^{d\ell}$ is a weighted vector to be determined through optimizing certain loss function, and $\mathrm{sgn} : \mathbb{R} \to \mathbb{R}$ is the sign function, returning 1 when the argument variable is positive and -1 otherwise. The Hamming loss (HL, see [66]) is an important evaluation metric in MLC. Let $\mathbf{y}^{(k)} = (y_1^{(k)}, \cdots, y_q^{(k)}) \in \mathbb{R}^q$ be the k-th column of the matrix $\mathbf{Y}$, then Hamming loss for classifier $\mathcal{H}$ is defined as follows:

$$\mathtt{HL} := \frac{1}{q\ell} \sum_{i=1}^{q} \sum_{k=1}^{\ell} \mathbf{1}_{(0,\infty)}(-y_i^{(k)} \mathbf{c}_i^\top \mathbf{x}_k).$$

When optimizing HL on the training data, a practical technique is to replace the above $-y_i^{(k)} \mathbf{c}_i^\top \mathbf{x}_k$ with $-y_i^{(k)} \mathbf{c}_i^\top \mathbf{x}_k + 1$, see e.g. [62]. This helps to avoid trivial solution and improve generalization ability of the classifier. Meanwhile, when the dimension of the parameter $\mathbf{x} \in \mathbb{R}^{d\ell}$ is large, we can use sparse regularization to alleviate the risk of overfitting and reduce the dimension. Overall, we consider minimizing the Hamming loss with an elastic net regularizer. Denoting $m := q\ell$ and $n := d\ell$, this model is a special case of (1) with the following setting:

Table 1 Real datasets for AUC maximization

Abbreviation	Dataset	Domain	q_-	q_+	m	n
all	Allaml	Text	47	25	1175	7129
aou	AP_Ovary_Uterus	Biology	124	198	24552	10936
bas	Basehock	Text	999	994	993006	4862
col	Colon_Cancer	Biology	40	22	880	2000
duk	Duke	Biology	21	23	483	7129
gli	Gli85	Biology	26	59	1534	22283
gs2	Gse2280	Biology	5	22	110	11868
gs7	Gse7670	Biology	27	27	729	11868
leu	Leukemia	Biology	47	25	1175	7070
mad	Madelon	Artificial	1000	1000	1000000	500
ove	Ova_Endometrium	Biology	61	1484	90524	10936
ovk	Ova_Kidney	Biology	260	1285	334100	10936
ovo	Ova_Ovary	Biology	198	1347	266706	10936
pcm	Pcmac	Text	961	982	943702	3289
pro	Prostate_Ge	Biology	50	52	2600	5966
rel	Relathe	Text	648	779	504792	4322
smk	Smk_Can187	Biology	90	97	8730	19993
spl	Splice	Biology	483	517	249711	60

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 + \lambda_1\|\mathbf{x}\|_1,\ A := \begin{bmatrix} -(\mathbf{y}^{(1)}\mathbf{e}_d^\top) \odot \mathbf{C} & & \\ & \ddots & \\ & & -(\mathbf{y}^{(\ell)}\mathbf{e}_d^\top) \odot \mathbf{C} \end{bmatrix},\ \mathbf{b} = \mathbf{e}_m,$$

where $\|\cdot\|_1$ is the ℓ_1 norm and $\odot$ is the Hadamard product.

We can compute the conjugate function

$$f^*(\mathbf{v}) = \sum_{i=1}^{n} \varphi(v_i) \quad \text{with } \varphi(t) := \begin{cases} 0, & \text{if } |t| \le \lambda_1, \\ (|t| - \lambda_1)^2/2, & \text{if } |t| > \lambda_1, \end{cases}$$

and thus the corresponding gradient and generalized Jacobian are as follows:

$$\nabla f^*(\mathbf{v}) = \left\{ \mathbf{r} \in \mathbb{R}^n : r_i = \begin{cases} v_i - \lambda_1, & \text{if } v_i > \lambda_1 \\ 0, & \text{if } |v_i| \le \lambda_1 \\ v_i + \lambda_1, & \text{if } v_i < -\lambda_1 \end{cases} \right\}$$

$$\partial^2 f^*(\mathbf{v}) = \left\{ \mathrm{Diag}(\mathbf{r}) : \mathbf{r} \in \mathbb{R}^n, r_i \in \begin{cases} 1, & \text{if } |v_i| > \lambda_1 \\ [0,1], & \text{if } |v_i| = \lambda_1 \\ 0, & \text{if } |v_i| < \lambda_1 \end{cases} \right\}.$$

Table 2 Numerical results on real datasets for AUC maximization

	AUC (%) ↑					TIME (sec) ↓				
	SGSN	NM01	CBN	PRSVM	OPAUC	SGSN	NM01	CBN	PRSVM	OPAUC
all	99.00	99.00	99.00	98.00	98.50	2.677e-2	3.285e-2	3.980e-2	1.438e-1	1.961e+1
aou	94.48	94.75	94.46	94.18	94.16	2.773e-1	3.585e-1	4.819e-1	2.821e+0	1.950e+2
bas	98.82	97.82	98.46	98.73	98.76	1.868e+0	1.017e+1	3.077e+0	4.701e+0	2.335e+2
col	87.92	85.63	87.92	86.04	85.83	2.061e-2	1.085e-1	3.083e-2	7.766e-2	3.407e+0
duk	89.54	90.79	89.54	89.54	84.54	3.082e-2	4.377e-2	4.797e-2	1.517e-1	3.031e+1
gli	95.76	95.76	95.76	94.18	93.33	8.698e-2	1.100e-1	1.422e-1	4.148e-1	2.701e+2
gs2	61.67	61.67	61.67	61.67	56.67	1.556e-2	1.513e-2	2.031e-2	6.896e-2	1.999e+1
gs7	98.40	98.40	98.40	98.40	97.60	4.164e-2	3.406e-2	7.484e-2	1.333e-1	3.964e+1
leu	99.00	99.00	99.00	98.50	65.00	2.344e-2	2.690e-2	4.478e-2	1.665e-1	1.951e+1
mad	55.68	53.66	55.41	54.33	58.26	5.120e+0	1.300e+1	1.134e+1	3.154e+0	9.349e+0
ove	95.81	94.52	95.21	95.42	93.69	7.186e-1	2.379e+0	1.492e+0	2.012e+1	1.209e+3
ovk	99.16	98.95	99.11	99.09	98.26	1.251e+0	2.151e+0	2.314e+0	1.844e+1	1.288e+3
ovo	93.30	91.74	93.32	93.16	91.52	1.517e+0	2.367e+0	2.375e+0	4.341e+1	1.263e+3
pcm	91.11	88.73	90.56	90.40	90.26	1.669e+0	2.782e+1	5.701e+0	8.364e+0	1.468e+2
pro	96.00	95.60	96.00	95.80	90.33	3.279e-2	3.787e-1	5.935e-2	1.196e-1	2.401e+1
rel	93.28	90.88	92.14	92.57	92.44	1.336e+0	4.266e+0	6.050e+0	3.166e+0	1.456e+2
smk	78.76	77.93	77.81	78.04	76.49	1.217e+0	1.791e+0	4.312e-1	9.226e-1	4.819e+2
spl	88.28	88.39	88.23	88.23	87.19	1.909e-1	5.908e-1	5.429e-1	1.597e-1	1.845e-1

↑" means that a larger metric value corresponds to better algorithm performance, while "↓" has the opposite meaning

Let $\mathbf{a}^{(i)}$ be the i-th column of A, our SGSN will solve the following dual problem:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \; F(\mathbf{z}) = \sum_{i=1}^{n} \varphi(-\langle \mathbf{a}^{(i)}, \mathbf{z}\rangle) - \langle \mathbf{b}, \mathbf{z}\rangle + g(\mathbf{z}). \tag{65}$$

Next, we discuss the convergence property when solving (65). One can verify that the elastic net regularizer is semi-algebraic by [8, Def. 5]. Assuming the sequence $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ is bounded, the whole sequence converges to a P-stationary point $\mathbf{z}^*$ of (65) according to Prop. 5. ∇f^* is semismooth. Denoting $\mathcal{S}^* := \mathcal{S}(\mathbf{z}^*)$ and $\Gamma^* := \{i \in [n] : |\langle \mathbf{a}^{(i)}, \mathbf{z}^*\rangle| > \lambda_1\}$, if we further assume that $A_{\mathcal{S}^*\Gamma^*}$ has full row rank, then for any $H \in \widehat{\partial}^2 h(\mathbf{z}^*)$, we have $\mathbf{d}^\top H_{\mathcal{S}^*}\mathbf{d} = \|A^\top_{\mathcal{S}^*\Gamma^*}\mathbf{d}\|^2 > 0$ for any $\mathbf{d} \in \mathbb{R}^{|\mathcal{S}^*|}\backslash\{0\}$. This means Assumption 3 holds and $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$ converges to $\mathbf{z}^*$ with local superlinear rate by Thm. 3.

In the following numerical experiments, we set $\mathbf{z}^0 = 0, \mu = 10^{-5}, \gamma = 10^{-2}, c_1 = 10^{-4}$ and $c_2 = 10^8$ for SGSN. The proximal parameter is adaptively chosen as $\tau_k = 0.1^j$, where j is the smallest natural number such that $F(\mathbf{v}^k) - F(\mathbf{z}^k) \geq 10^{-4}\|\mathbf{v}^k - \mathbf{z}^k\|^2$. The regularization parameter λ_1 is selected from $\{2^{-6}, 2^{-5}, \cdots, 2^6\}$. Other setting for SGSN is the same as that in the last subsection. We select three comparison algorithms for the MLC problem with ℓ_1 norm or group sparse regularizer: LLFS from [29], MDFS from [65] and RFS from [49]. We adjust the sparse regularization parameters of the three algorithms to achieve better performance. They are selected from $\{10^{-6}, 10^{-5}, \cdots, 10^6\}$ for RFS and MDFS, and $\{2^{-6}, 2^{-5}, \cdots, 2^6\}$ for LLFS. Other parameters of the three algorithms are set the same as their default values. We use the Hamming loss (HL), computational time (TIME) and number of nonzero elements (NNE) of the solution to evaluate the performance of the algorithms.

5.2.1 Experiments on simulated dataset

To observe how the change of dimension influence the performance, we test all four algorithms on simulated datasets with various q, d and ℓ generated as follows.

Example 3 We first generate a matrix $\overline{\mathbf{C}} \in \mathbb{R}^{q\times(d-1)}$ with i.i.d elements being chosen from standard normal distribution $N(0, 1)$, and then the feature matrix is computed by $\mathbf{C} = [\overline{\mathbf{C}}, \mathbf{e}_q]$. Next, we generate a matrix $W \in \mathbb{R}^{d\times\ell}$ with each element chosen from the uniform distribution $U(-1, 1)$, and then the label matrix is computed by $\mathbf{Y} = \mathrm{Sgn}(\mathbf{C}W)$, where $\mathrm{Sgn}(\cdot)$ is the sign function for a matrix. Finally, 90% of samples are randomly selected as training data and the rest is regarded as testing data.

• **On change of** q. We fix $d = 6000, \ell = 3$, and select $q \in \{1000, 2000, \cdots, 6000\}$. We can observe from Table 3 that SGSN takes the shortest TIME in all the cases. In particular, SGSN is at least 6 times faster than the second fastest algorithm LLFS. SGSN also finds the solution with the smallest NNE when $q > 1000$. TIME and NNE increases when q rises, whereas HL shows the opposite trend.

• **On change of** d. We fix $q = 2500, \ell = 3$, and select $d \in \{2500, 4000, \cdots, 10000\}$. In this test, SGSN achieves the best performance in all the cases. When feature dimension d grows, SGSN has a relatively small increase on TIME. This indicates that SGSN may be more advantageous when applied to large-scale datasets.

Table 3 Numerical results on Example 3 with different q, d and ℓ

	HL ↓				TIME (sec) ↓				NNE ↓			
	SGSN	RFS	LLFS	MDFS	SGSN	RFS	LLFS	MDFS	SGSN	RFS	LLFS	MDFS
q	$d = 6000, \ell = 3$											
1000	0.424	0.424	0.445	0.453	1.186	4.908	22.10	23.47	2874	7293	2750	2970
2000	0.397	0.419	0.394	0.434	2.295	19.82	32.94	42.41	4625	12870	5312	4716
3000	0.368	0.423	0.373	0.410	4.371	49.11	42.54	75.31	6097	15174	7620	6192
4000	0.339	0.428	0.340	0.402	7.818	93.03	52.79	126.8	7470	16235	9679	7560
5000	0.310	0.431	0.321	0.390	11.21	154.0	64.08	200.0	8524	16797	12726	8586
6000	0.300	0.428	0.319	0.395	14.15	239.6	86.30	295.3	9303	17161	17046	9378
d	$q = 2500, \ell = 3$											
2500	0.304	0.438	0.331	0.376	2.809	19.37	8.118	20.14	4205	6362	5593	4237
4000	0.346	0.420	0.354	0.406	2.880	25.36	16.23	33.98	4942	9772	4988	5004
5500	0.375	0.422	0.370	0.419	3.235	30.94	29.96	49.35	5290	13152	5476	5362
7000	0.401	0.426	0.404	0.437	2.857	36.96	50.58	69.50	5416	16499	5755	5544
8500	0.406	0.431	0.401	0.437	3.311	41.69	77.83	95.52	5830	19755	6035	5967
10000	0.407	0.433	0.418	0.451	3.761	47.29	106.8	128.6	6147	23134	6322	6330
ℓ	$q = 1000, d = 2000$											
5	0.370	0.462	0.372	0.403	1.176	2.186	4.604	2.756	4207	5364	4732	4270
7	0.372	0.468	0.378	0.415	1.698	2.252	5.854	2.453	5872	8014	6455	5950
9	0.373	0.450	0.374	0.398	2.117	2.26	7.887	2.428	7528	10855	8030	7614
11	0.371	0.441	0.375	0.403	2.479	2.294	9.458	2.069	9166	13749	9832	9284
13	0.369	0.463	0.379	0.407	3.143	2.389	10.82	2.191	10872	16795	11601	10972
15	0.364	0.467	0.374	0.403	3.503	2.378	12.78	2.092	12560	19877	13194	12720

"↓" means that a smaller metric value corresponds to better algorithm performance

Table 4 Real multi-label classification datasets with higher dimension

Abbreviation	Dataset	Domain	q	d	ℓ	$q \times d \times \ell$
bbc	3s-bbc1000	Text	352	1000	6	2.11e+06
int	3s-inter3000	Text	169	3000	6	3.04e+06
eug	EukaryoteGO	Biology	7766	12689	22	2.17e+09
gng	GnegativeGO	Biology	1392	1717	8	1.91e+07
gnp	GnegativePseAAC	Biology	1392	440	8	4.90e+06
gpg	GpositiveGO	Biology	519	912	4	1.89e+06
gpp	GpositivePseAAC	Biology	519	440	4	9.13e+05
hup	HumanPseAAC	Biology	3106	440	14	1.91e+07
ima	Image	Image	2000	294	5	2.94e+06
med	Medical	Text	978	1449	45	6.38e+07
plg	PlantGO	Biology	978	3091	12	3.63e+07
plp	PlantPseAAC	Biology	978	440	12	5.16e+06
sch	Stackex_chess	Text	1675	585	227	2.22e+08
sco	Stackex_coffee	Text	225	1763	123	4.88e+07
vig	VirusGO	Biology	207	749	6	9.30e+05
vip	VirusPseAAC	Biology	207	440	6	5.46e+05
yar	Yahoo_Arts	Text	7484	23146	26	4.50e+09
yed	Yahoo_Education	Text	12030	27534	33	1.09e+10
yrr	Yahoo_Recreation	Text	12830	30324	22	8.56e+09
yrf	Yahoo_Reference	Text	8027	39679	33	1.05e+10

• **On change of** ℓ. We fix $q = 1000$, $d = 2000$, and select $\ell \in \{5, 7, \cdots, 15\}$. From Table 3, we can observe that SGSN has the best HL and NNE. Its TIME is also the shortest when $\ell \leq 9$, and slightly larger than that of RFS and MDFS when $\ell \geq 11$.

5.2.2 Experiments on real data

In this part,

we test all the algorithms on real MLC problems listed in Table 4.

Example 4 The real datasets in Table 4 are available on the multi-label classification dataset repository[6]. For each dataset, the partition in 67% train and 33% test is performed by following the Iterative Stratification method proposed by [58].

We report the performance of all the algorithms in terms of HL, TIME and NNE. Particularly, on the large-scale datasets yar, yed, yrr, yrf, we restrict the TIME of LLFS within 3000 seconds, because this algorithm takes longer time to terminate.

• On HL: We can see from Table 5 that SGSN has the best HL on almost all the datasets. LLFS also shows competitive HL, especially on datasets gpg and sco. RFS and MDFS have relatively larger HL.

[6] https://www.uco.es/kdis/mllresources/

Table 5 Numerical results of all algorithms on real multi-label classification datasets

	HL ↓				TIME (sec) ↓				NNE ↓			
	SGSN	RSF	LLSF	MFDS	SGSN	RSF	LLSF	MFDS	SGSN	RSF	LLSF	MFDS
bbc	1.949e-1	1.949e-1	1.994e-1	2.426e-1	3.856e-2	1.364e-1	1.541e-1	8.103e-2	13	4224	58	5394
int	1.871e-1	1.871e-1	1.871e-1	1.988e-1	6.921e-2	1.169e-1	1.674e+0	4.506e-1	10	804	109	8814
gng	1.410e-2	4.935e-2	1.437e-2	6.291e-2	1.704e-1	2.078e+0	1.042e+0	3.584e+0	181	11160	79	72
gnp	7.674e-2	8.026e-2	8.080e-2	1.020e-1	9.605e-1	1.431e+0	8.913e-2	1.207e+0	120	3520	976	3512
gpg	3.343e-2	8.866e-2	3.052e-2	5.087e-2	5.279e-2	2.027e-1	1.170e-1	4.915e-1	77	2664	49	24
gpp	1.599e-1	2.137e-1	1.817e-1	2.965e-1	4.249e-1	1.650e-1	1.841e-2	2.311e-1	20	1760	490	56
hup	8.221e-2	8.242e-2	8.235e-2	9.286e-2	1.585e+0	9.824e+0	1.634e-1	8.181e+0	123	378	224	6146
ima	2.192e-1	2.473e-1	2.473e-1	2.834e-1	8.593e-1	1.026e+1	3.932e-2	3.695e+0	129	105	156	1175
med	1.024e-2	1.268e-2	1.031e-2	5.726e-2	7.167e-1	1.030e+0	1.001e+0	2.196e+0	2132	53505	1809	4545
plg	3.643e-2	8.658e-2	4.990e-2	8.832e-2	4.637e-1	1.503e+0	3.650e+0	5.607e+0	1515	4408	2584	3336
plp	8.458e-2	8.658e-2	8.658e-2	1.742e-1	2.042e+0	7.156e-1	4.372e-2	5.203e-1	25	192	52	5268
sch	9.958e-3	1.049e-2	9.640e-3	1.315e-2	2.187e+0	3.015e+0	1.774e+0	5.188e+0	115	21565	1910	227
sco	1.570e-2	1.581e-2	1.558e-2	1.581e-2	6.203e-1	2.562e-1	1.562e+0	1.426e+0	25	13644	123	1845
vig	2.412e-2	5.482e-2	3.289e-2	8.772e-2	7.366e-2	3.952e-2	6.714e-2	1.815e-1	288	2232	229	180
vip	1.842e-1	1.864e-1	1.908e-1	1.930e-1	1.866e-1	3.388e-2	2.749e-2	3.306e-1	25	114	50	1314
eug	2.015e-2	2.883e-2	2.022e-2	8.562e-2	1.496e+1	3.214e+2	1.131e+2	7.526e+2	2784	239316	1558	13948
yar	5.850e-2	6.469e-2	6.159e-2	6.205e-2	2.724e+1	4.590e+02	2.242e+3	1.133e+3	1046	4420	36491	2418
yed	3.980e-2	4.470e-2	4.070e-2	4.722e-2	4.426e+1	1.631e+03	2.761e+3	1.531e+3	680	13728	34962	891
yrr	5.173e-2	6.568e-2	6.357e-2	7.984e-2	4.367e+1	1.965e+03	2.703e+3	1.568e+3	1825	9108	20036	2684
yrf	2.628e-2	3.385e-2	2.671e-2	3.281e-2	3.479e+1	8.487e+02	2.924e+3	1.744e+3	1300	11451	85839	2640

"↓" means that a smaller metric value corresponds to better the algorithm performance

• On TIME: SGSN shows significant advantage on TIME when solving large-scale dataset, such as eug, yar, yed, yrr, yrf. Particularly, on these datasets, TIME of SGSN is less than 1/10 of other algorithms. However, on the small dimensional datasets with $d < 600$, LLFS has shorter TIME.

• On NNE: SGSN has the smallest NNE on most datasets in Table 5, especially on the high-dimensional datasets. The smaller NNE is, the faster SGSN runs because the subspace identified becomes smaller. This fact also accounts for the shortest TIME of SGSN used on the datasets yar, yed, yrr, yrf. For some low-dimensional datasets, such as gng and gpg, LLFS and MDFS find solutions with smaller NNE.

6 Conclusions

This paper proposes a new stationary duality approach for a class of composite optimization with indicator functions. The dual problem is constructed as ℓ_0 regularized and nonnegativity constrained sparse optimization. The fundamental one-to-one correspondence between solutions of primal and dual problems is established. To find a dual solution, a subspace gradient semismooth Newton (SGSN) method is developed. It has low per-iteration complexity by exploiting the proximal operator of the sparse regularizer in the dual objective function. Moreover, SGSN enjoys global convergence with a local superlinear (or quadratic) rate under suitable conditions. Extensive experiments on the AUC maximization and multi-label classification problems demonstrate its competitive performance against top solvers for the respective problems.

Although the developed algorithm SGSN explicitly requires the gradient information of the conjugate function $f^*(\cdot)$, the stationary duality result does not rely on the availability of this piece of information. In theory, the stationary duality applies to general convex function f as long as the dual problem admits an optimal solution. Numerically, this calls for further investigation when the gradient function cannot be cheaply computed. For instance, it would be interesting to see if derivative-free algorithms [17, 37] is a viable choice for the dual problem in this situation. If successful, it would significantly enlarge the applications of the stationary duality approach. We also note that the dual problem provides valuable information for the Lagrangian multiplier of the primal problem. It would be interesting to investigate whether a primal-dual method is possible. Those questions are not easy to answer and we leave them to our future research.

Acknowledgements The authors are very grateful to the co-editor and two anonymous referees for their insightful comments and suggestions, which has greatly enhanced the quality of the paper.

Funding Open access funding provided by The Hong Kong Polytechnic University

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aragón-Artacho, F.J., Mordukhovich, B.S., Pérez-Aros, P.: Coderivative-based semi-Newton method in nonsmooth difference programming. Math. Program. 1–48 (2024)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
3. Beck, A.: First-Order Methods in Optimization. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia (2017)
4. Beck, A., Teboulle, M.: A fast dual proximal gradient algorithm for convex minimization and applications. Oper. Res. Lett. **42**(1), 1–6 (2014)
5. Bento, G., Mordukhovich, B., Mota, T., Nesterov, Y.: Convergence of descent optimization algorithms under Polyak-Łojasiewicz-Kurdyka conditions (2025). https://optimization-online.org/wp-content/uploads/2025/02/Optimization-ModelFev2025.pdf
6. Bhatia, R.: Matrix analysis. Springer Science & Business Media, Berlin (2013)
7. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
8. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1), 459–494 (2014)
9. Bolte, J., Sabach, S., Teboulle, M.: Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence. Math. Oper. Res. **43**(4), 1210–1232 (2018)
10. Boţ, R.I., Csetnek, E.R., Nguyen, D.K.: A proximal minimization algorithm for structured nonconvex and nonsmooth problems. SIAM J. Optim. **29**(2), 1300–1328 (2019)
11. Boţ, R.I., Nguyen, D.K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. Math. Oper. Res. **45**(2), 682–712 (2020)
12. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recogn. **37**(9), 1757–1771 (2004)
13. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with SVMs. Inf. Retr. **13**, 201–215 (2010)
14. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**(1), 81–101 (1994)
15. Chen, X., Guo, L., Lu, Z., Ye, J.J.: An augmented Lagrangian method for non-Lipschitz nonconvex programming. SIAM J. Numer. Anal. **55**(1), 168–193 (2017)
16. Clarke, F.H.: Optimization and nonsmooth analysis. SIAM (1990)
17. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to derivative-free optimization. SIAM (2009)
18. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
19. Cui, Y., Liu, J., Pang, J.S.: The minimization of piecewise functions: pseudo stationarity. arXiv preprint arXiv:2305.14798 (2023)
20. Cui, Y., Pang, J.S.: Modern nonconvex nondifferentiable optimization. SIAM (2021)
21. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems. Springer, Berlin (2003)
22. Feng, M., Mitchell, J.E., Pang, J.S., Shen, X., Wächter, A.: Complementarity formulations of ℓ_0-norm optimization problems. Pac. J. Optim. **14**, 273–305 (2018)
23. Gao, W., Jin, R., Zhu, S., Zhou, Z.H.: One-pass AUC optimization. In: International conference on machine learning, pp. 906–914. PMLR (2013)
24. Gfrerer, H., Outrata, J.V.: On a semismooth* Newton method for solving generalized equations. SIAM J. Optim. **31**(1), 489–517 (2021)
25. Hallak, N., Teboulle, M.: An adaptive Lagrangian-based scheme for nonconvex composite optimization. Math. Oper. Res. **48**(4), 2337–2352 (2023)

26. Han, A.K.: Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. J. Econom. **35**(2–3), 303–316 (1987)
27. Han, S., Cui, Y., Pang, J.S.: Analysis of a class of minimization problems lacking lower semicontinuity. Math. Oper. Res. (2024). https://doi.org/10.1287/moor.2023.0295
28. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)
29. Huang, J., Li, G., Huang, Q., Wu, X.: Learning label-specific features and class-dependent labels for multi-label classification. IEEE Trans. Knowl. Data Eng. **28**(12), 3309–3323 (2016)
30. Kanzow, C., Lechner, T.: Globalized inexact proximal Newton-type methods for nonconvex composite functions. Comput. Optim. Appl. **78**, 377–410 (2021)
31. Kanzow, C., Qi, H.D.: A QP-free constrained Newton-type method for variational inequality problems. Math. Program. **85**(1), 81–106 (1999)
32. Kanzow, C., Schwartz, A., Weiß, F.: The sparse (st) optimization problem: reformulations, optimality, stationarity, and numerical results. Comput. Optim. Appl. **90**(1), 77–112 (2025)
33. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1–2), 81–93 (1938)
34. Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: A generalized Newton method for subgradient systems. Math. Oper. Res. **48**(4), 1811–1845 (2023)
35. Khanh, P.D., Mordukhovich, B.S., Phat, V.T.: Coderivative-based Newton methods in structured nonconvex and nonsmooth optimization. arXiv preprint arXiv:2403.04262 (2024)
36. Khanh, P.D., Mordukhovich, B.S., Phat, V.T., Tran, D.B.: Globally convergent coderivative-based generalized Newton methods in nonsmooth optimization. Math. Program. **205**(1), 373–429 (2024)
37. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numer **28**, 287–404 (2019)
38. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**(4), 2434–2460 (2015)
39. Li, X., Sun, D., Toh, K.C.: A highly efficient semismooth Newton augmented Lagrangian method for solving lasso problems. SIAM J. Optim. **28**(1), 433–458 (2018)
40. Lütkepohl, H.: Handbook of matrices. John Wiley & Sons, Hoboken (1997)
41. Mordukhovich, B.S.: Maximum principle in the problem of time optimal response with nonsmooth constraints. J. Appl. Math. Mech. **40**(6), 960–969 (1976)
42. Mordukhovich, B.S.: Sensitivity analysis in nonsmooth optimization. Theor. Asp. Ind. Des. **58**, 32–46 (1992)
43. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation. I. Basic Theory, II. Applications. Springer, Berlin (2006)
44. Mordukhovich, B.S.: Variational Analysis and Applications. Springer (2018)
45. Mordukhovich, B.S.: Second-Order Variational Analysis in Optimization, Variational Stability, and Control: Theory, Algorithms. Applications, Springer Nature (2024)
46. Mordukhovich, B.S., Sarabi, M.E.: Generalized Newton algorithms for tilt-stable minimizers in nonsmooth optimization. SIAM J. Optim. **31**(2), 1184–1214 (2021)
47. Mordukhovich, B.S., Yuan, X., Zeng, S., Zhang, J.: A globally convergent proximal Newton-type method in nonsmooth convex optimization. Math. Program. **198**(1), 899–936 (2023)
48. Münch, M.M., Peeters, C.F., Van Der Vaart, A.W., Van De Wiel, M.A.: Adaptive group-regularized logistic elastic net regression. Biostatistics **22**(4), 723–737 (2021)
49. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. Advances in neural information processing systems **23**, (2010)
50. Nocedal, J., Wright, S.: Numerical optimization. Springer series in operations research and financial engineering. Springer, New York (2006)
51. Poliquin, R., Rockafellar, R.T.: Tilt stability of a local minimum. SIAM J. Optim. **8**(2), 287–299 (1998)
52. Polyak, B.T.: Gradient methods for the minimisation of functionals. USSR Comput. Math. Math. Phys. **3**(4), 864–878 (1963)
53. Polyak, B.T.: Sharp minima. Institute of Control Sciences Lecture Notes, Moscow,: Presented at the IIASA Workshop on Generalized Lagrangians and Their Applications. Laxenburg, Austria (1979)
54. Polyak, B.T.: Introduction to optimization. Optimization Software, New York (1987)
55. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**(1–3), 353–367 (1993)
56. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
57. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Fundamental Principles of Mathematical Sciences. Springer, Berlin (1998)

58. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multi-label data. In: Machine Learning and Knowledge Discovery in Databases: European Conference, pp. 145–158. Springer (2011)
59. Shivaswamy, P.K., Jebara, T.: Maximum relative margin and data-dependent regularization. J. Mach. Learn. Res. **11**(2), 747–788 (2010)
60. Vapnik, V.: The nature of statistical learning theory. Springer Science & Business Media, Berlin (1999)
61. Wang, F., Cao, W., Xu, Z.: Convergence of multi-block Bregman ADMM for nonconvex composite problems. Sci. China Inform. Sci. **61**(12), 1–12 (2018)
62. Wu, G., Tian, Y., Zhang, C.: A unified framework implementing linear binary relevance for multi-label learning. Neurocomputing **289**, 86–100 (2018)
63. Xie, W., Yang, H.: Group sparse recovery via group square-root elastic net and the iterative multivariate thresholding-based algorithm. AStA Adv. Stat. Anal. **107**(3), 469–507 (2023)
64. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
65. Zhang, J., Luo, Z., Li, C., Zhou, C., Li, S.: Manifold regularized discriminative feature selection for multi-label learning. Pattern Recogn. **95**, 136–150 (2019)
66. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. **26**(8), 1819–1837 (2013)
67. Zhang, P., Xiu, N.: Global convergence of inexact augmented Lagrangian method for zero-one composite optimization. arXiv preprint arXiv:2112.00440 (2021)
68. Zhang, Y., Zhang, N., Sun, D., Toh, K.C.: An efficient Hessian based algorithm for solving large-scale sparse group lasso problems. Math. Program. **179**, 223–263 (2020)
69. Zhou, S., Pan, L., Xiu, N.: Newton method for ℓ_0-regularized optimization. Numer. Algorithms pp. 1–30 (2021)
70. Zhou, S., Pan, L., Xiu, N., Qi, H.D.: Quadratic convergence of smoothing Newton's method for 0/1 loss optimization. SIAM J. Optim. **31**(4), 3184–3211 (2021)
71. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Methodol. **67**(2), 301–320 (2005)