Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 18, No. 1, pp. 367–393, 2022

A BUNDLE-TYPE QUASI-NEWTON METHOD FOR NONCONVEX NONSMOOTH OPTIMIZATION*

Chunming Tang, Huangyue Chen†, Jinbao Jian and Shuai Liu

Abstract: In this paper, we propose a bundle-type quasi-Newton method for minimizing a nonconvex nonsmooth function. The method is based on the redistributed bundle method with an on-the-fly convexification technique. At each iteration, the convexification parameter and the prox-parameter are suitably modified to guarantee that the proximal point of a piecewise affine model of a local convexification function approximates well-enough the proximal point of the objective function f at x^k. A quasi-Newton procedure is added at the end of each serious step. Specifically, we construct a suitable search direction d^k via the BFGS update and monitor the reduction in the norm of the approximate subgradient to recognize whether an Armijo-type line search on f should be executed. Global convergence of the algorithm is established in the sense that there exists an accumulation point of the serious iterations such that it is a stationary point of f. Superlinear convergence is proved under suitable assumptions. Preliminary numerical results are reported to illustrate that the method is efficient and has advantages over the redistributed bundle method.

Key words: *nonconvex nonsmooth optimization, proximal bundle method, quasi-Newton method, global convergence, superlinear convergence*

Mathematics Subject Classification: *90C30, 90C53, 65K05, 49M15*

1 Introduction

We consider the following unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} \ f(x), \tag{1.1}$$

where $f(x):\mathbb{R}^n\to\mathbb{R}$ is a nonconvex nonsmooth function. This type of problems arises in many fields of applications, such as signal denoising [39], compressed sensing [8], and machine learning [7, 16], etc.

In the past several decades, a great deal of efforts have been devoted to developing methods for problem (1.1), for example proximal alternating linearized minimization methods [1, 33], bundle-type methods [9, 10, 13], gradient sampling methods [3, 44], stochastic subgradient methods [6, 37]. Bundle methods were first introduced in [24, 46] and regarded as stabilized variants of cutting-plane methods [5, 19] from the viewpoint of primal approach [2, 29].

*This work was supported by the National Natural Science Foundation of China (Nos. 11761013; 11771383), Guangxi Natural Science Foundation (Nos. 2018GXNSFFA281007; 2020GXNSFDA238017; 2021GXNSFBA075012), and the special foundation for Guangxi Ba Gui Scholars.

†Corresponding author.

Let ℓ be the current iteration index, the classical bundle methods keep memory of previous candidate points y^i ($i \in J_\ell$), their function values $f(y^i)$ and subgradients $g^i \in \partial f(y^i)$ in a bundle of information

$$B_\ell := \bigcup_{i \in J_\ell} \{(y^i, f(y^i), g^i)\}, \tag{1.2}$$

where $J_\ell \subseteq \{1, \ldots, \ell\}$ denotes an index set corresponding to some previous candidate points, $\partial f(y)$ denotes the subdifferential of f at y, and each element $g \in \partial f(y)$ is called a subgradient. The linearizations of $f(x)$ at y^i are defined by

$$\bar{f}^i(x) := f(y^i) + \langle g^i, x - y^i \rangle, \quad i \in J_\ell,$$

and then the *cutting-planes (piecewise-affine) model* for f at the ℓth iteration is given by

$$\check{f}^\ell_{\mathrm{cp}}(x) := \max_{i \in J_\ell} \{\bar{f}^i(x)\}.$$

If f is convex, the linearization errors for f at any point x

$$\alpha_i(x) := f(x) - \bar{f}^i(x), \quad i \in J_\ell \tag{1.3}$$

are nonnegative, so $\check{f}^\ell_{\mathrm{cp}}(x)$ provides a lower approximation to f, i.e.,

$$\check{f}^\ell_{\mathrm{cp}}(x) \leq f(x),$$

which is crucial to the convergence analysis of most of the bundle methods. In the convex setting, a number of variants of bundle methods have been proposed in recent years, such as proximal bundle methods [21, 31, 43], level bundle methods [23, 25, 42], bundle-filter methods [17], alternating linearization bundle methods [22], and doubly stabilized bundle methods [32, 45], etc.

When f is nonconvex, the situation becomes much more complicated than the convex case, since the linearization errors α_i may be negative and thus the cutting-planes model is no longer a lower approximation to f. Nevertheless, some strategies have been proposed for overcoming this drawback. For instance, Kiwiel [20] proposed to replace the linearization errors with a more general nonnegative measure function; Schramm and Zowe [38] incorporated a trust region strategy into bundle method to restrict the cutting-planes model in a reliable region; Fuduli et al. [9] divided the linearization errors into two groups (nonnegative and negative), aiming to establish both a lower and an upper polyhedral approximations to the objective function. Hare and Sagastizábal [12, 13] proposed the redistributed proximal bundle method for minimizing nonconvex nonsmooth functions. In their method, the prox-parameter is split into two nonnegative terms, one is called the convexification parameter, and the other is the model prox-parameter; for the class of lower-C^2 functions, a cutting-plane model is used for approximating not the objective function f but a local convexification of f constructed via the convexification parameter. The local convexification is updated on the fly through verifying the nonnegativity of its linearization errors. Subsequently, the redistributed proximal bundle method is extended to inexact information setting [14], generalized variational inequality problems in Hilbert spaces [40], and DC programming [41].

Generally, the types of bundle methods for nonconvex optimization have only linear convergence rate. To the best of our knowledge, superlinear convergence rate have not been established for bundle methods in the nonconvex setting. Although in [27] a bundle-Newton method was shown to have global and superlinear convergence for nonsmooth functions, the superlinear convergence rate was shown under the assumption that the objective function is strongly convex. It is our interest in this paper to propose a bundle-type method with superlinear convergence speed. For nonsmooth convex problems, there are no shortages of methods that are shown to have superlinear rate of convergence [4, 11, 30]. Essentially, these methods utilize the second-order information via the Moreau-Yosida regularization. Specifically speaking, the minimization of a nonsmooth convex function $f(x)$ can be transformed equivalently to the minimization of a smooth convex function

$$e_R f(x) := \min_{y\in\mathbb{R}^n} \Big\{ f(y) + \frac{1}{2} R\|y-x\|^2 \Big\},$$

which is the Moreau-Yosida regularization of f with a positive parameter R. The unique minimizer of $f(y) + \frac{1}{2}R\|y-x\|^2$ is the proximal point $p_R f(x)$. The derivative of $e_R f(x)$ is given by $G(x) := \nabla e_R f(x) = R(x - p_R f(x))$. The approximate second-order information of $e_R f(x)$ can be obtained through the BFGS formula as in [30].

In this paper, our purpose is to combine the redistributed proximal bundle method [13] with the proximal quasi-Newton method [4], and develop a variant of bundle method for nonconvex optimization with superlinear convergence under suitable assumptions. Based on the redistributed proximal bundle method, our method has the same convexification process along the iterations. We add a quasi-Newton procedure at the end of a serious step. This procedure produces a direction d^k and possibly a step-size yielded from an Armijo-type line search on f. Then the prox-center is updated by $x^{k+1} = p^a(x^k) + \tau_k d^k$ where $p^a(x^k)$ is the proximal point of the cutting-planes model. Therefore, the sequence $\{x^k\}$ is no longer contained in the candidate point sequence $\{y^\ell\}$. We show that the proposed method is globally convergent in the sense that there exists an accumulation point of the serious iterations such that it is a stationary point of f. Superlinear convergence rate is established under some additional assumptions. Some preliminary numerical results demonstrate the good performance of our method compared with the redistributed bundle method.

The remainder of this paper is organized as follows. In Section 2, we review some definitions and results from variational analysis. An implementable algorithm and its global convergence are presented in Sections 3 and 4, respectively. The rate of convergence of the algorithm is studied in Section 5. Section 6 contains some encouraging preliminary numerical experiments and comparisons. Conclusions are presented in Section 7. The Euclidean inner product in $\mathbb{R}^n$ is denoted by $\langle x, y\rangle = x^T y$, and the associated norm by $\|\cdot\|$.

2 Preliminaries

In this section, we give some definitions in variational analysis [36] and show the properties of some objects for later use.

Firstly, we recall the definition of *regular subdifferential* in [36, Definition 8.3]. Consider a function $f : \mathbb{R}^n \to \overline{\mathbb{R}} := [-\infty, +\infty]$ and a point x with $f(x)$ finite, the regular subdifferential of f at x is defined by

$$\hat{\partial} f(x) := \left\{ g \in \mathbb{R}^n : \liminf_{y\to x,\ y\neq x} \frac{f(y) - f(x) - \langle g, y-x\rangle}{\|y-x\|} \geq 0 \right\}.$$

The subdifferential is defined by

$$\partial f(x) := \limsup_{y \xrightarrow[f]{} x} \hat{\partial} f(y),$$

where $y \xrightarrow[f]{} x$ implies $y \to x$ and $f(y) \to f(x)$. Furthermore, the relation $\hat{\partial} f(x) \subseteq \partial f(x)$ holds by [36, Theorem 8.6].

The function f is said to be *prox-bounded* [36, Definition 1.23 and Exercise 1.24] if there exists $R \geq 0$ such that the function $f(\cdot) + \frac{1}{2}R\|\cdot\|^2$ is bounded below. The threshold of prox-bounbedness is denoted by r_{pb} (≥ 0) which ensures $f(\cdot) + \frac{1}{2}R\|\cdot\|^2$ is bounded below for all $R \geq r_{pb}$.

Definition 2.1 ([36, Definition 1.22]). For a proper lower semicontinuous function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and parameter value $R > 0$, the *Moreau envelope function* $e_R f(x)$ and *proximal mapping* $p_R f(x)$ are defined by

$$e_R f(x) := \min_{y \in \mathbb{R}^n} \Big\{ f(y) + \frac{R}{2}\|y - x\|^2 \Big\},$$
$$p_R f(x) := \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big\{ f(y) + \frac{R}{2}\|y - x\|^2 \Big\}.$$

Definition 2.2 ([12, Remark 1] [36, Definition 10.29]). The function f is *lower-C^2* on an open set V if for each $\bar{x} \in V$ there is a neighbourhood V' of $\bar{x}$ upon which a representation $f(x) = \max_{t \in T} f_t(x)$ holds, where T is a compact set and the functions f_t are of class C^2 on V such that f_t, ∇f_t, and $\nabla^2 f_t$ depend continuously not just on $x \in V$ but jointly on $(t, x) \in T \times V$.

For convenience, we introduce an equivalent definition of lower-C^2 functions.

Definition 2.3 ([36, Theorem 10.33]). The function f is lower-C^2 on an open set V if f is finite valued on V, and for any point x in V there exists a threshold $\bar{\eta} > 0$ such that $f + \frac{\eta}{2}\|\cdot\|$ is convex on an open neighborhood V' of x for all $\eta \geq \bar{\eta}$.

We first state a basic assumption on the objective function f.

Assumption 2.4. The objective function f given in problem (1.1) is lower-C^2 on an open bounded set V. In addition, given a point $\bar{x}^0 \in \mathbb{R}^n$ and a parameter $M_0 \geq 0$, the level set $\Gamma := \{x \in \mathbb{R}^n : \ f(x) \leq f(\bar{x}^0) + M_0\}$ is a subset of V.

The following lemma collects some useful results with respect to Assumption 2.4.

Lemma 2.5. *If the objective function f satisfies Assumption 2.4, then the following statements are true:*

(1) *The level set Γ is nonempty and compact.*

(2) *The function f is bounded below and prox-bounded with threshold $r_{pb} = 0$.*

(3) *There exists a threshold $\bar{\eta} > 0$ such that the function $f(\cdot) + \frac{1}{2}\eta\|\cdot - x\|^2$ is convex on Γ, for any $\eta \geq \bar{\eta}$ and any given $x \in \Gamma$.*

(4) *The function f is Lipschitz continuous on Γ.*

(5) *In addition, if $R > \bar{\eta}$, then we have the following statements:*

(5.a) *The proximal mapping $p_R f(x)$ is single-valued (denoted by $p(x)$) and Lipschitz continuous on Γ.*

(5.b) *For all* $x \in \Gamma$, *the gradient of the Moreau envelope function* $e_R f$ *at* x *is given by* $G(x) = R(x - p_R f(x))$, *and the proximal point is uniquely determined by the relation*

$$p(x) = p_R f(x) \Longleftrightarrow R(x - p(x)) \in \partial f(p(x)). \tag{2.1}$$

(5.c) $x^* \in \Gamma$ *is a stationary point of* f *if* $x^* = p_R f(x^*)$.

Proof. The proof of (1)-(4) can be found in [13, Proposition 1].

To see item (5), combining Definition 2.1 and Assumption 2.4, for any $x \in \Gamma$, we get

$$\begin{aligned} p_R f(x) &:= \operatorname{argmin}\left\{f(y) + \frac{1}{2}R\|y - x\|^2, y \in \mathbb{R}^n\right\} \\ &\subseteq \left\{y \in \mathbb{R}^n : f(y) + \frac{1}{2}R\|y - x\|^2 \le f(x)\right\} \\ &\subseteq \{y \in \mathbb{R}^n : f(y) \le f(x)\} \\ &\subseteq \{y \in \mathbb{R}^n : f(y) \le f(\bar{x}^0) + M_0\} = \Gamma, \end{aligned}$$

which implies $p_R f(x) = \operatorname*{argmin}_{y\in\Gamma}\{f(y) + \frac{1}{2}R\|y - x\|^2\}$.

If $R > \bar{\eta}$, we further obtain that the proximal mapping $p_R f(x)$ and the Moreau envelope function $e_R f(x)$ can be rewritten as

$$\begin{aligned} p_R f(x) &= \operatorname*{argmin}_{y\in\Gamma}\left\{f(y) + \frac{1}{2}R\|y - x\|^2\right\} \\ &= \operatorname*{argmin}_{y\in\Gamma}\left\{f_x^+(y) + \frac{R - \bar{\eta}}{2}\|y - x\|^2\right\} = p_{R-\bar{\eta}} f_x^+(x), \end{aligned}$$

$$e_R f(x) = \min_{y\in\Gamma}\left\{f(y) + \frac{1}{2}R\|y - x\|^2\right\} = \min_{y\in\Gamma}\left\{f_x^+(y) + \frac{R - \bar{\eta}}{2}\|y - x\|^2\right\} = e_{R-\bar{\eta}} f_x^+(x),$$

where $f_x^+(y) := f(y) + \frac{1}{2}\bar{\eta}\|y - x\|^2$ is a convex function on level set Γ, for any $x \in \Gamma$.

By [36, Theorem 2.26], we know that $p_{R-\bar{\eta}} f_x^+(x)$ is single-valued and continuous. Moreover, $e_{R-\bar{\eta}} f_x^+(x)$ is continuously differentiable with the gradient

$$\nabla e_{R-\bar{\eta}} f_x^+(x) = R(x - p_{R-\bar{\eta}} f_x^+(x)). \tag{2.2}$$

Thus, the item (5.a) holds and $p(x) = p_R f(x)$. Furthermore, (2.2) can be rewritten as

$$\nabla e_R f(x) = G(x) = R(x - p_R f(x)),$$

and (2.1) holds by the first-order optimality conditions of $\min_{y\in\Gamma}\{f(y) + \frac{1}{2}R\|y - x\|^2\}$.

Taking $x = x^*$ and $p(x) = x^*$ in (2.1), we have $0 \in \partial f(x^*)$ which implies (5.c). □

Remark 2.6. (a) Obviously, C^2 functions (or maximum of a finite collection of C^2 functions) and any finite-valued convex function are lower-C^2 by Definitions 2.2 and 2.3, respectively. The class of lower-C^2 functions is a useful subset of locally Lipschitz continuous functions (see [36, Theorem 10.31]).

(b) Assumption 2.4 depends on a given $\bar{x}^0$ and a parameter M_0, where $\bar{x}^0$ plays the role of the first prox-center and starting point, while M_0 is an unacceptable increase parameter. Under this assumption, the objective function f needs not necessarily to be convex or smooth. Some examples of such functions can be found in [13].

(c) Suppose the threshold value $\bar{\eta}$ is known, then given any $R > \bar{\eta}$, we can utilize a convex bundle method for iteratively calculating a fixed point of proximal mapping $p_R f$ by using the relation

$$p_R f(x) = p_{R-\bar{\eta}}\left(f + \frac{1}{2}\bar{\eta}\|\cdot - x\|^2\right)(x),$$

but this is an ideal situation. In section 3, the threshold value $\bar{\eta}$ is estimated along the iterative process by using the bundle of information.

3 Algorithm Development

In this section, we propose our method which incorporates a quasi-Newton procedure into a proximal bundle method. The proposed method generates three sequences, namely, $\{y^\ell\}$, the sequence of candidate points generated from solving quadratic programming (QP) subproblems; $\{p^a(x^k)\}$, the sequence of candidate points that satisfy some descent condition; and $\{x^k\}$, the sequence of prox-centers obtained from executing quasi-Newton procedures.

3.1 Bundle for nonconvex function

In the convex case, from (1.3) the linearization errors for f at the current prox-center x^k are given by

$$\alpha_i^k := f(x^k) - f(y^i) - \langle g^i, x^k - y^i\rangle, \quad i \in J_\ell, \tag{3.1}$$

which are always nonnegative. Additionally the convexity of f gives $g^i \in \partial_{\alpha_i^k} f(x^k)$. In order to reduce storage, the bundle of information (1.2) can be rewritten as $B_\ell = \bigcup_{i\in J_\ell}\{(\alpha_i^k, g^i)\}$.

Whereas, for a nonconvex function f, the associated linearization errors may be negative which impairs convergence of the bundle method. Our method manages nonconvexity by the strategy proposed in [13] which utilizes the "locally convexification" function

$$h_\ell^k(\cdot) := f(\cdot) + \frac{1}{2}\eta_\ell\|\cdot - x^k\|^2,$$

where η_ℓ is the convexification parameter. The linearization errors for $h_\ell^k(\cdot)$ at x^k are

$$h_\ell^k(x^k) - [h_\ell^k(y^i) + \langle g^i + \eta_\ell(y^i - x^k), x^k - y^i\rangle] = \alpha_i^k + \frac{1}{2}\eta_\ell\|y^i - x^k\|^2, \quad i \in J_\ell. \tag{3.2}$$

Our method keeps memory of the previous iterations in a bundle of information with respect to $h_\ell^k(\cdot)$

$$\left\{(\alpha_i^k, g^i, \triangle_i^k, e_i^k),\ i \in J_\ell\right\} \tag{3.3}$$

where $\alpha_i^k := f(x^k) - f(y^i) - \langle g^i, x^k - y^i\rangle$, $g^i \in \partial f(y^i)$, $g^k \in \partial f(x^k)$, $\triangle_i^k := y^i - x^k$ and $e_i^k := \frac{1}{2}\|y^i - x^k\|^2$. A cutting-plane of $h_\ell^k(\cdot)$ at the current prox-center x^k is associated with a special quadruplet $(0, g^k, 0, 0)$ which can be obtained through replacing the y^i in (3.3) by x^k. Let the index i_k be such that $\left(\alpha_{i_k}^k, g_{i_k}^k, \Delta_{i_k}^k, e_{i_k}^k\right) := \left(0, g^k, 0, 0\right)$.

We construct the piecewise linear model $\check{h}_\ell^k(\cdot)$ of the function $h_\ell^k(\cdot)$ instead of f, based on the information in the bundle as follows

$$\begin{aligned}\check{h}_\ell^k(y) &= \max_{i\in J_\ell}\{h_\ell^k(y^i) + \langle g^i + \eta_\ell(y^i - x^k), y - y^i\rangle\} \\ &= f(x^k) + \max_{i\in J_\ell}\{-(\alpha_i^k + \eta_\ell e_i^k) + \langle g^i + \eta_\ell\triangle_i^k, y - x^k\rangle\}.\end{aligned} \tag{3.4}$$

The next candidate point $y^{\ell+1}$ is defined by

$$y^{\ell+1} := p_{\rho_\ell}\check{h}_\ell^k(x^k) = \operatorname*{argmin}_{y\in\mathbb{R}^n}\Big\{\check{h}_\ell^k(y) + \frac{1}{2}\rho_\ell\|y-x^k\|^2\Big\}, \tag{3.5}$$

where ρ_ℓ is the prox-parameter of the model. Since $\check{h}_\ell^k$ is a piecewise linear function, $y^{\ell+1}$ is uniquely obtained and (3.5) amounts to solving a QP subproblem. The optimality condition of (3.5) gives

$$y^{\ell+1} = x^k - \frac{1}{\rho_\ell}\sum_{i\in J_\ell}\lambda_i^\ell(g^i+\eta_\ell\triangle_i^k) \;\text{ and }\; \sum_{i\in J_\ell}\lambda_i^\ell(g^i+\eta_\ell\triangle_i^k) \in \partial\check{h}_\ell^k(y^{\ell+1}), \tag{3.6}$$

where λ^ℓ is the solution to the dual problem

$$\min_{\lambda\in\mathbb{R}_+^{|J_\ell|}} \;\frac{1}{2\rho_\ell}\Bigg\|\sum_{i\in J_\ell}\lambda_i(g^i+\eta_\ell\triangle_i^k)\Bigg\|^2 + \sum_{i\in J_\ell}\lambda_i(\alpha_i^k+\eta_\ell e_i^k), \;\text{ s.t. }\; \sum_{i\in J_\ell}\lambda_i = 1. \tag{3.7}$$

For convenience, we introduce the following notations:

$$I_\ell^{act} = \{i : \lambda_i^\ell > 0\}, \;\; g_{\eta_\ell}^{-\ell} = \sum_{i\in J_\ell}\lambda_i^\ell(g^i+\eta_\ell\triangle_i^k) = \sum_{i\in I_\ell^{act}}\lambda_i^\ell(g^i+\eta_\ell\triangle_i^k) \;\text{ and} \tag{3.8}$$

$$(\alpha_{-\ell}^k, g^{-\ell}, \triangle_{-\ell}^k, e_{-\ell}^k) = \sum_{i\in J_\ell}\lambda_i^\ell(\alpha_i^k, g^i, \triangle_i^k, e_i^k) = \sum_{i\in I_\ell^{act}}\lambda_i^\ell(\alpha_i^k, g^i, \triangle_i^k, e_i^k), \tag{3.9}$$

where $-\ell$ denotes the index of the aggregate bundle element.

By complementarity, we obtain

$$\check{h}_\ell^k(y^{\ell+1}) = f(x^k) - (\alpha_i^k+\eta_\ell e_i^k) + \langle g^i+\eta_\ell\triangle_i^k,\; y^{\ell+1}-x^k\rangle, \;\; \forall\, i\in I_\ell^{act}. \tag{3.10}$$

Furthermore, summing up instances of (3.10), and in view of (3.9), we obtain

$$\check{h}_\ell^k(y^{\ell+1}) = f(x^k) - (\alpha_{-\ell}^k+\eta_\ell e_{-\ell}^k) + \langle g^{-\ell}+\eta_\ell\triangle_{-\ell}^k,\; y^{\ell+1}-x^k\rangle. \tag{3.11}$$

From (3.6), (3.8) and (3.9), we obtain

$$\rho_\ell(x^k-y^{\ell+1}) = g_{\eta_\ell}^{-\ell} = g^{-\ell} + \eta_\ell\,\triangle_{-\ell}^k\,. \tag{3.12}$$

$\boxed{3.2}$ Convexification

Recalling that if the objective function f is convex, then its cutting-planes model is a lower approximation of f and therefore each of its linearization error α_i^k is always nonnegative with $g^i \in \partial_{\alpha_i^k} f(x^k)$. In case of a nonconvex f, a cutting-plane model of the "locally convexification" function h_ℓ^k is constructed and the nonnegativity of linearization errors with respect to h_ℓ^k serves as a necessary condition for the convexification technique. Each of the linearization errors has the form $\alpha_i^k+\eta_\ell e_i^k$ given in (3.2), and for any $i\in J_\ell$, it can be shown that

$$g^i+\eta_\ell\triangle_i^k \in \partial_{\alpha_i^k+\eta_\ell e_i^k}\check{h}_\ell^k(x^k) \;\text{ whenever }\; \alpha_i^k+\eta_\ell e_i^k \ge 0.$$

From Lemma 2.5(3) we see that if the value of η_ℓ associated with h_ℓ^k is greater than a certain threshold $\bar{\eta}$, then h_ℓ^k becomes convex on the level set Γ. Our adopt the convexification strategy proposed in [13], which iteratively updates the convexification parameter η_ℓ so that the linearization errors $\alpha_i^k+\eta_\ell e_i^k \ge 0$, for all $i\in J_\ell$, with the intention that η_ℓ can asymptotically approach the threshold $\bar{\eta}$. A minor difference is that our lower bound for η_ℓ is set to be always nonnegative.

Lemma 3.1. *Set the lower bound*

$$\bar{\eta}_\ell := \max\left\{\max\{-\alpha_i^k/e_i^k : i \in J_\ell,\, e_i^k > 0\}, 0\right\}. \tag{3.13}$$

Whenever $\eta_\ell \geq \bar{\eta}_\ell$, the linearization errors $\alpha_i^k + \eta_\ell e_i^k \geq 0$, for all $i \in J_\ell$.

Proof. If $e_i^k = 0$ then $y^i = x^k$ and by (3.1) $\alpha_i^k = 0$. Thus $\alpha_i^k + \eta_\ell e_i^k = 0$. Now consider the indices $i \in J_\ell$ such that $e_i^k > 0$. From (3.13) we see that if $\bar{\eta}_\ell = 0$ then $-\alpha_i^k/e_i^k \leq 0$ for all $i \in J_\ell$ such that $e_i^k > 0$ and therefore $\alpha_i^k \geq 0$. It follows that for all such i, $\alpha_i^k + \eta_\ell e_i^k \geq \alpha_i^k + \bar{\eta}_\ell e_i^k = \alpha_i^k \geq 0$. Now consider the case of a positive $\bar{\eta}_\ell$. If i is such that $e_i^k > 0$ and $-\alpha_i^k/e_i^k < 0$ then $\alpha_i^k > 0$ yielding $\alpha_i^k + \eta_\ell e_i^k \geq \alpha_i^k > 0$. For i such that $e_i^k > 0$ and $-\alpha_i^k/e_i^k \geq 0$ we have $\alpha_i^k + \eta_\ell e_i^k \geq \alpha_i^k + \bar{\eta}_\ell e_i^k \geq \alpha_i^k + e_i^k\left(-\alpha_i^k/e_i^k\right) = 0$. □

Because the convexification process tries to build up the model $\check{h}_\ell^k$ as a lower approximation of h_ℓ^k within the level set Γ, it is necessary to keep the bundle points y^i inside Γ. For the candidate point $y^{\ell+1}$ generated from (3.5), it may happen that $f(y^{\ell+1}) > f(x^k) + M_0$, i.e., $y^{\ell+1}$ may fall out of Γ. In this case, the bundle should be discarded and the algorithm should be restarted with an updated value of the model prox-parameter ρ_ℓ. The new value should be greater so that the next candidate point is closer to the prox-center x^k and therefore closer to Γ. It can be shown that such restarts happen a finite number of times (cf. Lemma 4.1 below).

3.3 The quasi-Newton procedure

In this subsection, we introduce a new quasi-Newton procedure based on [4] which can effectively improve the convergence rate of proximal bundle method. In the following, we describe our quasi-Newton procedure in detail.

```
**Procedure** quasi-Newton(QN)
1 Input $c, \gamma \in (0,1)$, $R_\ell$, $\eta_\ell$, $\rho_\ell$, $M_0$, $m_2 \in (0, m_1)$, $x^k$ and $p^a(x^k)$.
2 Calculate $G^a(x^k) := R_\ell(x^k - p^a(x^k))$, $s_{k-1} := x^k - x^{k-1}$, $t_{k-1} := G^a(x^k) - G^a(x^{k-1})$, and
  generate $B_k$ by the following formula

  $$B_k := \begin{cases} (1+R_\ell)I, & \text{if } k = 0, \\ B_{k-1}, & \text{if } k \geq 1 \text{ and } t_{k-1}^T s_{k-1} \leq 0, \\ B_{k-1} - \frac{B_{k-1}s_{k-1}s_{k-1}^T B_{k-1}}{s_{k-1}^T B_{k-1} s_{k-1}} + \frac{t_{k-1}t_{k-1}^T}{t_{k-1}^T s_{k-1}}, & \text{otherwise,} \end{cases} \tag{3.14}$$

  where $I$ is identity matrix. Calculate $d^k := -(B_k^{-1} - R_\ell^{-1} I)G^a(x^k)$.
3 If $k = 0$, let $\beta_1 := \|G^a(x^0)\|$. For $k \geq 1$, if

  $$\|G^a(x^k)\| \leq c\beta_k \text{ and } f(p^a(x^k) + d^k) \leq f(\bar{x}^0) + M_0, \tag{3.15}$$

  let $\beta_{k+1} := \|G^a(x^k)\|$, $\tau_k := 1$, $x^{k+1} := p^a(x^k) + \tau_k d^k$, and **terminate** the procedure;
  otherwise, let $\beta_{k+1} := \beta_k$.
4 Compute the step size $\tau_k := \max\{\tau \mid \tau = \gamma^j,\ j = 0, 1, 2, \dots\}$ such that

  $$f(p^a(x^k) + \tau_k d^k) \leq f(x^k) - \tau_k \frac{m_2(\eta_\ell + 2\rho_\ell)}{2R_\ell^2}\|G^a(x^k)\|^2 \tag{3.16}$$

  is satisfied. Set $x^{k+1} := p^a(x^k) + \tau_k d^k$, and **terminate** the procedure.
```

Remark 3.2. (a) $p^a(x^k)$ represents the approximate proximal point of x^k, which can be obtained by choosing a "good" candidate point (see Step 4 in Algorithm 1 below).

(b) The search direction d^k is reasonable. Particularly, by [26, Theorem 3.1], if f is a smooth convex function with invertible Hessian matrix $\nabla^2 f(p(x^k))$, then $B_k \approx \nabla^2 e_{R_\ell} f(x^k) = R_\ell I - R_\ell[\nabla^2 f(p(x^k))R_\ell^{-1} + I]^{-1}$. Together with $(I - ST)^{-1} = I + S(I - TS)^{-1}T$ (S, T are two matrices), we obtain $B_k^{-1} \approx \nabla^2 e_{R_\ell} f(x^k)^{-1} = \nabla^2 f(p(x^k))^{-1} + R_\ell^{-1} I$. Thus,

$$d^k \approx -(\nabla^2 e_{R_\ell} f(x^k)^{-1} - R_\ell^{-1} I)\nabla e_{R_\ell} f(x^k) = -\nabla^2 f(p(x^k))^{-1}\nabla f(p(x^k)).$$

Hence, d^k can be viewed as an approximate Newton direction of f at $p(x^k)$.

(c) By (3.14), we can ensure B_k is symmetric positive definite, for any $k \geq 0$. More specifically, B_0 is symmetric positive definite, and B_k inherits the positive definiteness of B_{k-1} for the second case in (3.14). The third case can refer to [47, Theorem 5.1.3].

(d) We obtain an equivalent update formula for x^{k+1}. In fact

$$\begin{aligned} x^{k+1} &= p^a(x^k) - \tau_k(B_k^{-1} - R_\ell^{-1} I)G^a(x^k) \\ &= x^k - R_\ell^{-1} G^a(x^k) - \tau_k(B_k^{-1} - R_\ell^{-1} I)G^a(x^k) \\ &= x^k - [\tau_k B_k^{-1} + (1 - \tau_k)R_\ell^{-1} I]G^a(x^k). \end{aligned} \tag{3.17}$$

(f) The second inequality of (3.15) can guarantee that x^{k+1} falls into Γ.

$\boxed{3.4}$ The algorithm

In this subsection, we present our bundle-type quasi-Newton method in Algorithm 1. As usual, we assume that there is an *oracle* that can compute the function value $f(x)$ and one arbitrary subgradient $g(x) \in \partial f(x)$ for any point $x \in \mathbb{R}^n$.

Remark 3.3. (1) In Step 1, λ^ℓ and $y^{\ell+1}$ can also be computed by solving (3.7) and (3.6).

(2) In view of (3.11) and (3.12), the predicted decrease $\delta_{\ell+1}$ defined in Step 1 has an equivalent expression

$$\begin{aligned} \delta_{\ell+1} &= f(x^k) + \frac{1}{2}\eta_\ell\|y^{\ell+1} - x^k\|^2 - \check{h}_\ell^k(y^{\ell+1}) \\ &= f(x^k) + \frac{1}{2}\eta_\ell\|y^{\ell+1} - x^k\|^2 - [f(x^k) - \alpha_{-\ell}^k - \eta_\ell e_{-\ell}^k + \langle g^{-\ell} + \eta_\ell \triangle_{-\ell}^k, y^{\ell+1} - x^k\rangle] \\ &= \alpha_{-\ell}^k + \eta_\ell e_{-\ell}^k + \rho_\ell\|y^{\ell+1} - x^k\|^2 + \frac{1}{2}\eta_\ell\|y^{\ell+1} - x^k\|^2 \\ &= \frac{R_\ell + \rho_\ell}{2}\|y^{\ell+1} - x^k\|^2 + \alpha_{-\ell}^k + \eta_\ell e_{-\ell}^k. \end{aligned} \tag{3.21}$$

(3) In Step 3, the bundle reset ensures that the sequences $\{x^k\}$ and $\{y^\ell\}$ eventually lie in the level set Γ (see Lemma 4.1 for details).

(4) In Step 7, the update rule for the convexification parameter can ensure $\eta_\ell \geq \bar{\eta}_\ell$, and thus $\alpha_{-\ell}^k + \eta_\ell e_{-\ell}^k \geq 0$, which in turn implies $\delta_{\ell+1} \geq 0$ by (3.21).

(5) The prox-center is updated by executing Procedure quasi-Newton and thus $\{x^k\}$ is not contained in the sequence $\{y^\ell\}$ of candidate points.

$\boxed{4}$ Convergence Analysis

In this section, we shall establish the global convergence of Algorithm 1. Naturally, we assume that the tolerance parameter $\epsilon = 0$. We are looking for only stationary points for

Algorithm 1: Bundle-type quasi-Newton method

Step 0. (Initialization) Choose a point $\bar{x}^0 \in \mathbb{R}^n$ and an unacceptable increase parameter $M_0 > 0$. Set the initial point $x^0 = \bar{x}^0$, and choose a stopping tolerance $\epsilon \geq 0$, a convexification growth parameter $\Theta > 1$, parameters $0 < m_2 < m_1 < 1$, $\rho_0 > 0$, and $c, \gamma \in (0,1)$. Initialize the iteration counter $\ell = 0$, the serious step counter $k = 0$ with $i_0 = 0$, and the index set of candidate point $J_0 = \{0\}$. Call the oracle to obtain $f(x^0)$ and $g^0 \in \partial f(x^0)$. Choose the starting parameter $\eta_0 = 0$ and $R_0 = \rho_0$.

Step 1. (Solve QP) Define the piecewise linear model $\check{h}_\ell^k$ based on J_ℓ and (3.4). Compute $y^{\ell+1}$ and $\check{h}_\ell^k(y^{\ell+1})$ by (3.5) and calculate the predicted decrease

$$\delta_{\ell+1} := f(x^k) + \frac{1}{2}\eta_\ell \|y^{\ell+1} - x^k\|^2 - \check{h}_\ell^k(y^{\ell+1}).$$

Step 2. (Stopping criterion) Call the oracle to obtain $f(y^{\ell+1})$ and $g^{\ell+1} \in \partial f(y^{\ell+1})$. If $\delta_{\ell+1} \leq \epsilon$, then stop the algorithm. The approximate minimal value is $\min\{f(x^k), f(y^{\ell+1})\}$ and the approximate minimal point is the point that yields less function value.

Step 3. (Bundle reset) If $f(y^{\ell+1}) > f(x^k) + M_0$, then the objective increase is unacceptable. Restart the algorithm by setting

$$\eta_0 := \eta_\ell,\ \rho_0 = \Theta\rho_\ell,\ R_0 := \eta_0 + \rho_0,\ x^0 := x^k,\ \ell := 0,\ k := 0,\ i_0 := 0,\ J_0 := \{0\}\,,$$

and going to Step 1.

Step 4. (Descent test) If the descent condition

$$f(y^{\ell+1}) \leq f(x^k) - m_1\delta_{\ell+1} \tag{3.18}$$

holds, then declare a serious step, set $p^a(x^k) := y^{\ell+1}$, and go to Step 5. Otherwise, declare a null step and go to Step 6.

Step 5. (Update prox-center) Compute x^{k+1} by executing the quasi-Newton procedure in Section 3.3.

Step 6. (Update bundle information) Select an index set $J_{\ell+1}$ satisfying

$$J_{\ell+1} \supseteq \{\ell+1\} \ \text{ and } \left\{ \begin{array}{ll} \text{either} & J_{\ell+1} \supseteq I_\ell^{act} \\ \text{or} & J_{\ell+1} \supseteq \{-\ell\}. \end{array} \right. \tag{3.19}$$

If a null step is declared then compute $\alpha_{\ell+1}^k$, $\Delta_{\ell+1}^k$, and $e_{\ell+1}^k$ by (3.3). Make sure i_k is in $J_{\ell+1}$ and go to Step 7. If a serious step is declared then call the oracle to obtain $f(x^{k+1})$ and $g^{k+1} \in \partial f(x^{k+1})$. Set $i_{k+1} = \ell + 1$ and $(\alpha_{\ell+1}^{k+1}, \triangle_{\ell+1}^{k+1}, e_{\ell+1}^{k+1}) := (0,0,0)$. Update the bundle according to the following rule, for all $\ell + 1 \neq i \in J_{\ell+1}$,

$$\left\{ \begin{array}{l} \alpha_i^{k+1} = \alpha_i^k + f(x^{k+1}) - f(x^k) + \langle g^i, x^k - x^{k+1} \rangle, \\ \triangle_i^{k+1} = \triangle_i^k + x^k - x^{k+1}, \\ e_i^{k+1} = e_i^k + \frac{1}{2}\|x^{k+1} - x^k\|^2 - \langle \triangle_i^k, x^{k+1} - x^k \rangle. \end{array} \right.$$

Set $k = k + 1$ and go to Step 7.

Step 7. (Update parameter η) Update the convexification parameter η,

$$\left\{ \begin{array}{ll} \eta_{\ell+1} = \eta_\ell, & \text{if } \ \bar{\eta}_{\ell+1} \leq \eta_\ell, \\ \eta_{\ell+1} = \Theta\bar{\eta}_{\ell+1}, R_{\ell+1} = \rho_\ell + \eta_{\ell+1}, & \text{otherwise.} \end{array} \right. \tag{3.20}$$

where $\bar{\eta}_{\ell+1}$ is given by (3.13). Set $\rho_{\ell+1} = \rho_\ell$, $\ell = \ell + 1$, and return to Step 1.

the nonconvex objective function f. In subsection 4.1, we will show that Algorithm 1 is well-defined and that the model function employed by Algorithm 1 satisfies some crucial conditions used in [12]. The global convergence of Algorithm 1 is established in subsection 4.2.

4.1 Well-definedness of Algorithm 1 and model properties

In order to ensure that each step of Algorithm 1 is well-defined, we need to give some lemmas. The first lemma shows that Algorithm 1 only has finite number of bundle resets, and the sequences $\{y^\ell\}$ and $\{x^k\}$ eventually lie in the level set Γ.

Lemma 4.1. *Consider the sequences of iterations $\{y^\ell\}$ and $\{x^k\}$ generated by Algorithm 1. If the objective function f satisfies Assumption 2.4, then Algorithm 1 has only a finite number of restarts in Step 3. Hence, eventually the sequences $\{y^\ell\}$ and $\{x^k\}$ lie in Γ, and the model prox-parameter sequence $\{\rho_\ell\}$ becomes constant.*

Proof. By Assumption 2.4, f is locally Lipschitz continuous on the open bounded set $V \supset \Gamma$, and therefore the compactness of Γ yields a constant L such that $L \geq \|g\|$ for any $g \in \partial f(x)$ with $x \in \Gamma$.

Algorithm 1 is initialized by $x^0 = \bar{x}^0 \in \Gamma$. For each $k > 0$ the prox-center x^k is obtained from the quasi-Newton procedure in Section 3.3. In view of (3.15) and (3.16), we have $f(x^{k+1}) \leq f(\bar{x}^0) + M_0$, and therefore $x^k \in \Gamma$ for all k.

Our sequence $\{y^\ell\}$ corresponds to the sequence $\{x^n\}$ in [13]. We apply [13, Lemma 1] to get that there can only be a finite number of restarts and after the last restart, we will always have $f(y^{\ell+1}) \leq f(x^k) + M_0 \leq f(\bar{x}^0) + M_0$ and therefore $y^{\ell+1} \in \Gamma$. As the value of ρ_ℓ will not be updated it will remain constant. □

If (3.18) holds, Algorithm 1 proceeds to Step 5. The following lemma shows that the step size τ_k is well-defined at each iteration of Algorithm 1.

Lemma 4.2. *For any k, if (3.18) holds and Step 5 is executed, then*

$$f(p^a(x^k)) \leq f(x^k) - \frac{m_2(\eta_\ell + 2\rho_\ell)}{2R_\ell^2}\|G^a(x^k)\|^2, \tag{4.1}$$

where m_2 is a real number in $(0, m_1)$. Moreover, there exists $\bar{\tau} \in (0, 1]$ such that

$$f(p^a(x^k) + \tau d_k) \leq f(x^k) - \tau\frac{m_2(\eta_\ell + 2\rho_\ell)}{2R_\ell^2}\|G^a(x^k)\|^2, \quad \textit{for all } \tau \in (0, \bar{\tau}]. \tag{4.2}$$

Proof. If the descent condition (3.18) holds and Step 5 is executed, then $p^a(x^k)$ satisfies the inequality

$$f(p^a(x^k)) \leq f(x^k) - m_1[f(x^k) + \frac{1}{2}\eta_\ell\|p^a(x^k) - x^k\|^2 - \check{h}_\ell^k(p^a(x^k))].$$

By (3.11) and (3.12), we obtain

$$\begin{aligned} f(x^k) - \check{h}_\ell^k(p^a(x^k)) &= f(x^k) - [f(x^k) - (\alpha_{-\ell}^k + \eta_\ell d_{-\ell}^k) + \langle g^{-\ell} + \eta_\ell \triangle_\ell^k, p^a(x^k) - x^k\rangle] \\ &\geq \rho_\ell\|p^a(x^k) - x^k\|^2. \end{aligned}$$

Combining the two inequalities mentioned above, in view of the fact that $0 < m_2 < m_1 < 1$, we have

$$\begin{aligned} f(p^a(x^k)) &\le f(x^k) - m_1\Big[f(x^k) + \frac{1}{2}\eta_\ell\|p^a(x^k) - x^k\|^2 - \check{h}_\ell^k(p^a(x^k))\Big] \\ &\le f(x^k) - m_1[\frac{1}{2}\eta_\ell\|p^a(x^k) - x^k\|^2 + \rho_\ell\|p^a(x^k) - x^k\|^2] \\ &= f(x^k) - m_1\frac{\eta_\ell + 2\rho_\ell}{2}\|p^a(x^k) - x^k\|^2 \\ &= f(x^k) - m_1\frac{\eta_\ell + 2\rho_\ell}{2R_\ell^2}\|G^a(x^k)\|^2 \\ &< f(x^k) - \frac{m_2(\eta_\ell + 2\rho_\ell)}{2R_\ell^2}\|G^a(x^k)\|^2. \end{aligned} \tag{4.3}$$

If $G^a(x^k) = 0$, then $d^k = 0$. Hence, for any $\tau > 0$, (4.2) holds.

If $G^a(x^k) \neq 0$, (4.3) and the continuity of f indicate that there exists $\bar{\tau} \in (0, 1]$ such that for any $\tau \in (0, \bar{\tau}]$

$$\begin{aligned} f(p^a(x^k) + \tau d_k) &\le f(x^k) - \frac{m_2(\eta_\ell + 2\rho_\ell)}{2R_\ell^2}\|G^a(x^k)\|^2 \\ &\le f(x^k) - \tau\frac{m_2(\eta_\ell + 2\rho_\ell)}{2R_\ell^2}\|G^a(x^k)\|^2. \end{aligned}$$

□

In the following lemma, we state some properties about the model function $\check{h}_\ell^k$ and Algorithm 1. These properties are crucial for the convergence of our method and the proofs are similar to those of [13, Lemma 2] and we omit its proof here.

Lemma 4.3. *Consider the family of the model functions $\check{h}_\ell^k$ given by (3.4), and the sequences of iterates $\{y^\ell\}$ and $\{x^k\}$ generated by Algorithm 1. Then the following statements are true:*

(1) *$\check{h}_\ell^k$ is a convex function.*

(2) *$\check{h}_\ell^k(x^k) \le f(x^k)$.*

(3) *For any prox-center x^k, if $y^{\ell+1}$ is a null step and $\eta_\ell = \eta_{\ell+1}$, then*
$$\check{h}_{\ell+1}^k(w) \ge \check{h}_\ell^k(y^{\ell+1}) + \rho_\ell\langle x^k - y^{\ell+1}, w - y^{\ell+1}\rangle \quad \forall w \in \mathbb{R}^n.$$

(4) *For any $w \in \mathbb{R}^n$, we have*
$$\check{h}_\ell^k(w) \ge f(y^\ell) + \eta_\ell e_\ell^k + \langle g^\ell + \eta_\ell \triangle_\ell^k, w - y^\ell\rangle, \quad \textit{where} \;\; g^\ell \in \partial f(y^\ell).$$

In Algorithm 1, the selection of index set $J_{\ell+1}$ given by (3.19) and the update rule (3.20) of η_ℓ in Step 7 ensure Lemma 4.3 (1), (2) and (4) hold. By contrast, Lemma 4.3 (3) holds only after the convexification parameter η_ℓ eventually stabilizes, which we state in the following lemma.

Lemma 4.4. *There exist an index $\ell_c > 0$ and a constant $\hat{\eta} \ge 0$ such that*

$$\eta_\ell \equiv \hat{\eta}, \quad \textit{for all } \ell \ge \ell_c.$$

In addition, if $\hat{\eta} \ge \bar{\eta}$, where $\bar{\eta}$ is given in Lemma 2.5(3), then

$$\check{h}_\ell^k(w) \le f(w) + \frac{\hat{\eta}}{2}\|w - x^k\|^2, \quad \textit{for all } \ell \ge \ell_c \textit{ and } w \in \Gamma.$$

Proof. See [13, Lemma 3]. □

Recalling that Lemma 4.1 and Lemma 4.4, we know that the convexification parameter sequence $\{\eta_\ell\}$ and the prox-parameter sequence $\{\rho_\ell\}$ stabilize eventually. There exists an index ℓ'' (not less than ℓ_c) such that

$$\eta_\ell \equiv \hat{\eta},\ \rho_\ell \equiv \hat{\rho} \text{ and } R_\ell \equiv \hat{R} := \hat{\eta} + \hat{\rho},\ \text{for all } \ell \geq \ell''.$$

$\boxed{4.2}$ Global convergence of Algorithm 1

In what follows, we will study the global convergence of our algorithm. We first set the tolerance parameter $\epsilon = 0$ and show the effectiveness of the stopping criterion given in Step 2. The following lemma follows from [13, page 2455].

Lemma 4.5. *If $\delta_{\ell+1} = 0$ and $\eta_\ell \geq \bar{\eta}$ then x^k is a stationary point of f.*

Let $\hat{\eta}$ and $\hat{\rho}$ be the stabilized values for the convexification parameter sequence $\{\eta_\ell\}$ and the model prox-parameter sequence $\{\rho_\ell\}$, respectively, as in Lemma 4.1 and Lemma 4.4. Denote

$$\hat{R} = \hat{\eta} + \hat{\rho}\,.$$

Just like previous bundle methods, the convergence analysis of our method includes two different asymptotic cases, depending on whether the Algorithm 1 produces a finite or an infinite number of serious steps. In the following theorem, we will prove that the sequence generated by Algorithm 1 converges to the stationary point for function f when the stabilized value $\hat{\eta}$ of the convexification parameter sequence is large enough. In the proof of the next theorem, we will invoke the result in [12, Theorem 2], of which a careful examination reveals that the result also holds for any $R > 0$ such that $R\left(x^0 - p\right) \in \partial f(p) \Rightarrow p = p_R f(x^0)$. This together with Lemma 2.5(5.b) implies that the conclusion in [12, Theorem 2] holds if $R > \bar{\eta}$.

For convenience, we introduce some notations. Let the minimal value of f be $\bar{f}$. Let $N := \{0, 1, 2, 3, \dots\}$ and

$$K_0 := \{0\} \cup \{k \in N : \text{ Line 4 of Procedure quasi-Newton \ is not excuted at iteration } k\}.$$

In view of Lemma 4.4, we denote

$$\check{h}^k_{\hat{\eta}}(\cdot) := \check{h}^k_\ell(\cdot),\ \text{for } \ell \text{ sufficiently large such that } \eta_\ell \equiv \hat{\eta}\,.$$

Given an infinite set $K \subseteq N$, we denote by $x^k \xrightarrow{K} x^*$ that $\{x^k\}$ converges to x^* with $k \in K$.

Theorem 4.6. *Let the objective function f satisfy Assumption 2.4. Suppose $\hat{\eta} \geq \bar{\eta}$ and $\{\|B_k^{-1}\|\}$ is bounded, then one of the following two conclusions holds.*

(1) *There is a last serious point $x^{\bar{k}}$, followed by infinitely many null steps. Then the sequence $\{y^\ell\}$ converges to $x^{\bar{k}}$, and $x^{\bar{k}}$ is a stationary point for f.*

(2) *There is an infinite number of serious steps. Then there exist an infinite subset $K' \subseteq N$ and a point x^* such that x^* is a stationary point of f with $x^k \xrightarrow{K'} x^*$ and $\|G^a(x^k)\| \xrightarrow{K'} 0$.*

Proof. (1) Suppose after the last serious point $x^{\bar{k}}$, an infinite sequence $\{y^\ell\}$ is generated by Algorithm 1. We apply [12, Theorem 2] to obtain that, as $\ell \to \infty$, the entire sequence $\{y^\ell\}$ converges to the proximal point $p = p_{\hat{R}} f(x^{\bar{k}})$ and

$$\lim_{\ell\to\infty} \check{h}^{\bar{k}}_{\hat{\eta}}(y^{\ell+1}) = f(p) + \frac{1}{2}\hat{\eta}\|p - x^{\bar{k}}\|^2.$$

Thus, as $\ell \to \infty$

$$\begin{aligned}\delta_{\ell+1} &= f(x^{\bar{k}}) + \frac{1}{2}\hat{\eta}\|y^{\ell+1} - x^{\bar{k}}\|^2 - \check{h}^{\bar{k}}_{\hat{\eta}}(y^{\ell+1}) \\ &\to f(x^{\bar{k}}) + \frac{1}{2}\hat{\eta}\|p - x^{\bar{k}}\|^2 - f(p) - \frac{1}{2}\hat{\eta}\|p - x^*\|^2 \\ &= f(x^{\bar{k}}) - f(p).\end{aligned}$$

Since the descent condition (3.18) dose not hold, we know that

$$f(y^{\ell+1}) > f(x^{\bar{k}}) - m_1\delta_{\ell+1},$$

which implies that $f(x^{\bar{k}}) \le f(p)$ by taking the limit as $\ell \to \infty$.

Furthermore, by the definition of proximal mapping given in Definition 2.1, $p = p_{\hat{R}} f(x^{\bar{k}})$ implies

$$f(p) + \frac{1}{2}\hat{R}\|p - x^{\bar{k}}\|^2 \le f(x^{\bar{k}}).$$

That is, $p = x^{\bar{k}} = p_{\hat{R}} f(x^{\bar{k}})$, so $x^{\bar{k}}$ is a stationary point for f by Lemma 2.5 (5.c).

(2) We denote by $\ell_k + 1$ the index of an iteration when a serious step is declared and the current prox-center is x^k, i.e., $y^{\ell_k+1} = p^a(x^k)$.

By Lemma 4.1, the infinite sequences $\{p^a(x^k)\}$ and $\{x^k\}$ eventually lie in Γ. Since Γ is a compact set, both $\{p^a(x^k)\}$ and $\{x^k\}$ have at least one accumulation point. Let x^* be an arbitrary accumulation point of $\{x^k\}$. Without loss of generality, suppose the infinite subset $K' \subseteq N$ and two vectors p^* and x^* are such that

$$p^a(x^k) \xrightarrow{K'} p^* \in \Gamma,\ x^k \xrightarrow{K'} x^* \in \Gamma\,. \tag{4.4}$$

As $\{\eta_\ell\}$ and $\{\rho_\ell\}$ are stabilized at $\hat{\eta}$ and $\hat{\rho}$, respectively, due to (3.20), we have $\{R_\ell\}$ stabilized at $\hat{R}$. From Line 2 of Procedure quasi-Newton , we see $G^a(x^k) = R_\ell(x^k - p^a(x^k))$ and thus

$$G^a(x^k) \xrightarrow{K'} \hat{R}\,(x^* - p^*)\ . \tag{4.5}$$

First, suppose the set K_0 is finite. Next, we show that $x^* = p^*$. Because K_0 is finite, eventually Line 4 in Procedure quasi-Newton is always executed. There exists an index $\hat{k} \in N$ such that

$$f(x^{k+1}) = f(p^a(x^k) + \tau_k d^k) \le f(x^k) - \tau_k \frac{m_2(\hat{\eta} + 2\hat{\rho})}{2\hat{R}^2}\|G^a(x^k)\|^2\,,\ \text{for all } k > \hat{k}\,. \tag{4.6}$$

We know that f is bounded below by Lemma 2.5. Then summing up the instances of (4.6) and taking the limit give

$$\frac{m_2(\hat{\eta} + 2\hat{\rho})}{2\hat{R}^2} \sum_{k=\hat{k}+1}^{\infty} \tau_k \|G^a(x^k)\|^2 \le f(x^{\hat{k}+1}) - \bar{f} < \infty,$$

which implies

$$\lim_{k\to\infty} \tau_k \|G^a(x^k)\|^2 = 0. \tag{4.7}$$

We claim that $x^* = p^*$. Suppose this is not the case. We proceed to find a contradiction. Because $x^* \neq p^*$, we have from (4.5) that $\lim_{K'\ni k\to\infty} \|G^a(x^k)\|^2 > 0$. In view of (4.7) we get

$$\lim_{K'\ni k\to\infty} \tau_k = 0\,. \tag{4.8}$$

By the definition of step size τ_k, we know that

$$f(p^a(x^k) + \frac{\tau_k}{\gamma} d^k) > f(x^k) - \frac{m_2(\hat{\eta}+2\hat{\rho})}{2\hat{R}^2}\frac{\tau_k}{\gamma}\|G^a(x^k)\|^2\,, \text{ for all } k > \hat{k}\,. \tag{4.9}$$

Since $\{\|B_k^{-1}\|\}$ is bounded, it follows from the definition of d^k in Line 2 of Procedure quasi-Newton and (4.5) that $\left\{d^k\right\}_{k\in K'}$ is bounded. Taking the limit of (4.9), in view of the continuity of f, (4.4), (4.8), the boundedness of d^k, and (4.7), we obtain $f(p^*) \geq f(x^*)$. On the other hand, taking the limit of (4.1) and combining (4.4) and (4.5) we obtain

$$f(p^*) \leq f(x^*) - \frac{m_1(\hat{\eta}+2\hat{\rho})}{2}\|x^* - p^*\|^2.$$

It must be the case that $x^* = p^*$. However, this contradicts the presumption that $x^* \neq p^*$. Consequently the presumption is false and we have $x^* = p^*$.

Second, we consider the case when K_0 is infinite. In this case, Line 3 of Procedure quasi-Newton is executed an infinite number of times. Let K_0 consist of $k_0 = 0 < k_1 < k_2 < \cdots$. By the construction of β_k, we have

$$\|G^a(x^{k_i})\| \leq c\|G^a(x^{k_{i-1}})\| \leq c^i \|G^a(x^0)\|, \; i = 1, 2, \ldots.$$

So we get

$$\lim_{i\to\infty} \|G^a(x^{k_i})\| \leq \lim_{i\to\infty} c^i \|G^a(x^0)\| = 0,$$

which implies that

$$\lim_{K_0\ni k\to\infty} \|G^a(x^k)\| = 0. \tag{4.10}$$

Since the infinite sequences $\{p^a(x^k)\}$ and $\{x^k\}$ lie eventually in Γ which is a compact set, there exists an infinite set $K'' \subseteq K_0 \subseteq N$ and two vectors $\hat{p}^*$ and $\hat{x}^*$ such that

$$p^a(x^k) \xrightarrow{K''} \hat{p}^* \in \Gamma \text{ and } x^k \xrightarrow{K''} \hat{x}^* \in \Gamma.$$

Therefore, the definition of $G^a(x^k)$ and (4.10) reveal $\hat{x}^* = \hat{p}^*$.

In summary, whether K_0 is finite or infinite, we can find a subsequence of $\left\{x^k\right\}$ such that the associated accumulation point equals that of $\left\{p^a(x^k)\right\}$. In the rest of the proof, for simplicity, we will unify the notations in the two cases as (4.4) with $p^* = x^*$.

Recall that in each serious step, $p^a(x^k)$ satisfies the descent condition,

$$f(p^a(x^k)) \leq f(x^k) - m_1\delta_{\ell_k+1}.$$

Taking the limit as $k \to \infty$, we get that

$$\delta_{\ell_k+1} \xrightarrow{K'} 0,$$

which further implies that $\alpha^k_{-\ell_k} + \hat{\eta} e^k_{-\ell_k}$ must converge to 0 due to an equivalent expression of δ_{ℓ_k+1} in (3.21).

Combining (3.11) and (3.12) and letting $\ell = -\ell_k$, we obtain

$$\check{h}^k_{\hat{\eta}}(p^a(x^k)) = f(x^k) - (\alpha^k_{-\ell_k} + \hat{\eta} e^k_{-\ell_k}) - \hat{\rho}\|p^a(x^k) - x^k\|^2.$$

Taking the limit as $k \to \infty$ gives

$$\lim_{K' \ni k \to \infty} \check{h}^k_{\hat{\eta}}(p^a(x^k)) = f(x^*).$$

By Definition 2.1, $p^a(x^k) = p_{\hat{\rho}} \check{h}^k_{\hat{\eta}}(x^k)$ implies that for any $w \in \Gamma$,

$$\check{h}^k_{\hat{\eta}}(p^a(x^k)) + \frac{1}{2}\hat{\rho}\|p^a(x^k) - x^k\|^2 \leq \check{h}^k_{\hat{\eta}}(w) + \frac{1}{2}\hat{\rho}\|w - x^k\|^2 \leq f(w) + \frac{1}{2}\hat{R}\|w - x^k\|^2,$$

where the second inequality follows from Lemma 4.4 and the fact that $\hat{\eta} \geq \bar{\eta}$. Taking the limit as $K' \ni k \to \infty$, we have that

$$f(x^*) \leq f(w) + \frac{1}{2}\hat{R}\|w - x^*\|^2, \text{ for any } w \in \Gamma.$$

We also know that $x^* \in \Gamma$ and thus

$$f(x^*) \leq f(x^0) + M_0 < f(w) \leq f(w) + \frac{1}{2}\hat{R}\|w - x^*\|^2, \text{ for any } w \notin \Gamma.$$

Hence,

$$f(x^*) \leq f(w) + \frac{1}{2}\hat{R}\|w - x^*\|^2, \text{ for any } w \in \mathbb{R}^n.$$

That is, $x^* = p_{\hat{R}} f(x^*)$, and hence x^* is a stationary point for f by Lemma 2.5 (5.c). □

5 Superlinear Convergence Rate

In this section, we establish the superlinear convergence rate of Algorithm 1. We first present some definitions and results about strong monotonicity and strong convexity from [36].

Definition 5.1 ([36, Definition 12.53])**.** A mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ is strongly monotone if there exists $\sigma > 0$ such that $T - \sigma I$ is monotone, or equivalently:

$$\langle v_x - v_y, x - y\rangle \geq \sigma\|x - y\|^2,$$

where $v_x \in T(x)$ and $v_y \in T(y)$.

Definition 5.2 ([36, Definition 12.58])**.** A proper function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex if there is a constant $\sigma > 0$ such that

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) - \frac{1}{2}\sigma\lambda(1-\lambda)\|x - y\|^2, \text{ for any } x,\ y \text{ when } \lambda \in (0,1).$$

Lemma 5.3 ([36, Exercise 12.59])**.** *For a function $f : \mathbb{R}^n \to \mathbb{R}$ and a constant $\sigma > 0$, the following properties are equivalent:*

(1) *∂f is strongly monotone with constant σ;*

(2) *f is strongly convex with constant σ;*

(3) $f - \frac{1}{2}\sigma|\cdot|^2$ *is convex.*

After η_ℓ and ρ_ℓ stabilize eventually, we define $p(\cdot) := p_{\hat{R}} f(\cdot)$,

$$G(x) := \hat{R}(x - p(x)) \text{ and } \zeta_k := (1 - m_1)\left(f(x^k) + \frac{\hat{\eta}}{2}\|p^a(x^k) - x^k\|^2 - \check{h}^k_{\hat{\eta}}(p^a(x^k))\right). \quad (5.1)$$

The following lemma gives the errors arising from the model function $\check{h}^k_\ell(\cdot)$ and the approximate proximal point $p^a(x^k)$.

Lemma 5.4. *For each $k \geq 0$, let*

$$e^a_{\hat{R}} f(x^k) := f(p^a(x^k)) + \frac{\hat{R}}{2}\|p^a(x^k) - x^k\|^2.$$

If all the parameters stabilize and $\hat{\eta} \geq \bar{\eta}$, then the following statements are true for k sufficiently large

$$e_{\hat{R}} f(x^k) \leq e^a_{\hat{R}} f(x^k) \leq e_{\hat{R}} f(x^k) + \zeta_k, \|p^a(x^k) - p(x^k)\| \leq \sqrt{2\zeta_k/\hat{\rho}},$$

$$\|G^a(x^k) - G(x^k)\| \leq \hat{R}\sqrt{2\zeta_k/\hat{\rho}} \text{ and } \zeta_k \leq \frac{L(1 - m_1)}{m_1 \hat{R}}\|G^a(x^k)\|. \quad (5.2)$$

Proof. It is easy to see from its definition that $e_{\hat{R}} f(x^k) \leq e^a_{\hat{R}} f(x^k)$. In each serious step, we know (3.18) holds and

$$\begin{aligned}
e^a_{\hat{R}} f(x^k) &= f(p^a(x^k)) + \hat{R}\|p^a(x^k) - x^k\|^2/2 \\
&\leq f(x^k) - m_1\left(f(x^k) + \hat{\eta}\|p^a(x^k) - x^k\|^2/2 - \check{h}^k_{\hat{\eta}}(p^a(x^k))\right) \\
&\quad + \hat{R}\|p^a(x^k) - x^k\|^2/2 \\
&= (1 - m_1)\left(f(x^k) + \frac{\hat{\eta}}{2}\|p^a(x^k) - x^k\|^2 - \check{h}^k_{\hat{\eta}}(p^a(x^k))\right) \\
&\quad + \check{h}^k_{\hat{\eta}}(p^a(x^k)) + \frac{\hat{\rho}}{2}\|p^a(x^k) - x^k\|^2 \\
&= \zeta_k + \check{h}^k_{\hat{\eta}}(p^a(x^k)) + \frac{\hat{\rho}}{2}\|p^a(x^k) - x^k\|^2.
\end{aligned}$$

By Lemma 4.1, eventually $\{x^k\}$ lies in Γ. As a proximal point $p(x^k)$ satisfies $f(p(x^k)) \leq f(x^k)$ and thus $p(x^k) \in \Gamma$. By the fact $p^a(x^k) = p_{\hat{\rho}} \check{h}^k_{\hat{\eta}}(x^k)$ and Lemma 4.4, we know that

$$\begin{aligned}
\check{h}^k_{\hat{\eta}}(p^a(x^k)) + \frac{\hat{\rho}}{2}\|p^a(x^k) - x^k\|^2 &\leq \check{h}^k_{\hat{\eta}}(p(x^k)) + \frac{\hat{\rho}}{2}\|p(x^k) - x^k\|^2 \\
&\leq f(p(x^k)) + \frac{\hat{\eta}}{2}\|p(x^k) - x^k\|^2 + \frac{\hat{\rho}}{2}\|p(x^k) - x^k\|^2 \\
&= e_{\hat{R}} f(x^k).
\end{aligned}$$

Consequently, $e^a_{\hat{R}} f(x^k) \leq e_{\hat{R}} f(x^k) + \zeta_k$.

By Lemma 2.5 (3), if $\hat{\eta} \geq \bar{\eta}$, then $f(y) + \hat{\eta}\|y - x^k\|^2/2$ is convex on the level set Γ. Therefore, Lemma 5.3 implies $\theta(y) := f(y) + \hat{\eta}\|y - x^k\|^2/2 + \hat{\rho}\|y - x^k\|^2/2$ is strongly convex on the level set Γ with constant $\hat{\rho}$. By [35, Proposition 6], we obtain

$$\theta(y) \geq \theta(z) + \langle \xi, y - z \rangle + \hat{\rho}\|y - z\|^2/2, \text{ for any } y,\ z \in \Gamma, \quad (5.3)$$

where $\xi \in \partial\theta(z)$. Let us put $y = p^a(x^k)$ and $z = p(x^k)$ in (5.3). We have

$$\theta(y) = \theta(p^a(x^k)) = e^a_{\hat{R}} f(x^k) \ \text{ and } \ \theta(z) = \theta(p(x^k)) = e_{\hat{R}} f(x^k).$$

Moreover, since $p(x^k)$ minimizes θ, let $\xi = 0 \in \partial\theta(p(x^k))$. So it follows from (5.3) that

$$e^a_{\hat{R}} f(x^k) \geq e_{\hat{R}} f(x^k) + \hat{\rho}\|p^a(x^k) - p(x^k)\|^2/2.$$

Therefore,

$$\|p^a(x^k) - p(x^k)\| \leq \sqrt{2\zeta_k/\hat{\rho}}.$$

Further,

$$\begin{aligned}\|G^a(x^k) - G(x^k)\| &= \|\hat{R}(x^k - p^a(x^k)) - \hat{R}(x^k - p(x^k))\| \\ &= \hat{R}\|p(x^k) - p^a(x^k)\| \\ &\leq \hat{R}\sqrt{2\zeta_k/\hat{\rho}}.\end{aligned}$$

Finally, recalling the descent condition (3.18) and the definition of ζ_k gives

$$f(p^a(x^k)) \leq f(x^k) - \frac{m_1}{1-m_1}\zeta_k,$$

which together with Lemma 2.5 (4) implies that

$$\zeta_k \leq \frac{1-m_1}{m_1}(f(x^k) - f(p^a(x^k))) \leq \frac{1-m_1}{m_1} L\|x^k - p^a(x^k)\| = \frac{(1-m_1)L}{m_1\hat{R}}\|G^a(x^k)\|.$$

□

In the following theorem, we will establish the superlinear convergence of Algorithm 1. In [4, Theorem 3], a proximal quasi-Newton method for minimizing convex functions was shown to have superlinear convergence. Here, without assuming convexity of the objective function, we are able to show the superlinear convergence speed with the same assumptions as in [4, Theorem 3] except that we require $\{x^k\}$ converges to x^*. Let $\hat{\eta}$ and $\hat{\rho}$ be the stabilized values for $\{\eta_\ell\}$ and $\{\rho_\ell\}$ used in Algorithm 1 so that $\hat{R} = \hat{\eta} + \hat{\rho}$.

Theorem 5.5. *Let f satisfy Assumption 2.4 and x^* be a stationary point of f. Suppose that there exists $\sigma_1 > 0$ such that $\hat{\eta} = \bar{\eta} + \sigma_1$, and the function $G(x)$ defined in (5.1) is semismooth at x^*. Suppose there exists a lower bound σ_4 such that $\|B_k^{-1}\| \leq \sigma_4$ for all $k \in N$. Suppose $x^k \to x^*$,*

$$\zeta_k = o(\|G^a(x^k)\|^2) \ \text{and} \ \lim_{k\to\infty} \operatorname{dist}(B_k, \partial G(x^k)) = 0.$$

Then there exists an integer $\bar{k} > 0$ such that for any $k \geq \bar{k}$, $\tau_k \equiv 1$ and the sequence $\{x^k\}$ converges to x^* superlinearly.

Proof. We first show that there exists an integer $\bar{k} > 0$ such that $\tau_k \equiv 1$ for any $k \geq \bar{k}$. Because $\hat{\eta} = \bar{\eta} + \sigma_1$, it follows from Lemma 2.5(3) and Lemma 5.3 that for any $x \in \Gamma$, the function defined as $h^x_{\hat{\eta}}(\cdot) := f(\cdot) + \frac{1}{2}\hat{\eta}\|\cdot - x\|^2$ is strongly convex on Γ with modulus σ_1. For each $x \in \Gamma$, the definition of the Moreau envelope function gives

$$e_{\hat{\rho}} h^x_{\hat{\eta}}(x) = \min_{y\in\mathbb{R}^n}\left\{f(y) + \frac{1}{2}\hat{\eta}\|y-x\|^2 + \frac{1}{2}\hat{\rho}\|y-x\|^2\right\} = e_{\hat{R}} f(x).$$

By Theorem 2.2 in [26], the strong convexity of $h_{\hat{\eta}}^x(x)$ with modulus σ_1 implies the strong convexity of $e_{\hat{\rho}} h_{\hat{\eta}}^x(x)$ with modulus $\sigma_0 \geq \frac{\sigma_1\hat{\rho}}{\sigma_1+\hat{\rho}}$. By Definition 5.1 and Lemma 5.3, we obtain

$$\langle G(x) - G(y), x - y\rangle \geq \frac{\sigma_1\hat{\rho}}{\sigma_1 + \hat{\rho}}\|x - y\|^2, \text{ for all } x, y \in \Gamma,$$

where $G(x) = \nabla e_{\hat{R}} f(x) = \nabla e_{\hat{\rho}} h_{\hat{\eta}}^x(x)$ and $G(y) = \nabla e_{\hat{R}} f(y) = e_{\hat{\rho}} h_{\hat{\eta}}^x(x)$. Taking $x = x^k$ and $y = x^*$, applying Lemma 2.5(5.c) we get $G(x^*) = 0$ and

$$\|x^k - x^*\| \leq \frac{\sigma_1 + \hat{\rho}}{\sigma_1\hat{\rho}}\|G(x^k)\|. \tag{5.4}$$

By $\lim\limits_{k\to\infty} \operatorname{dist}(B_k, \partial G(x^k)) = 0$ and $\|B_k^{-1}\| \leq \sigma_4$, for any $\sigma_3 > 0$ there exists $k_1 > 0$ such that for all $k \geq k_1$,

$$\|I - B_k^{-1}V_k\| \leq \|B_k^{-1}\|\|B_k - V_k\| \leq \sigma_3, \text{ for some } V_k \in \partial G(x^k). \tag{5.5}$$

The condition $\zeta_k = o(\|G^a(x^k)\|^2)$ and (5.2) imply

$$\left|\frac{\|G(x_k)\|}{\|G^a(x^k)\|} - 1\right| \leq \frac{\sqrt{2\mu_k\zeta_k}}{G^a(x_k)}$$

and therefore $\zeta_k = o(\|G(x^k)\|^2)$. It then follows that there exist $k_2 > 0$ and $\sigma_2 \geq 0$ such that

$$\max\left\{\frac{\zeta_k}{\|G^a(x^k)\|^2}, \frac{\zeta_k}{\|G(x^k)\|^2}\right\} \leq \frac{\hat{\rho}}{2\hat{R}}\sigma_2^2, \text{ for all } k > k_2,$$

i.e.,

$$\zeta_k \leq \frac{\hat{\rho}}{2\hat{R}^2}\sigma_2^2 \min\{\|G(x^k)\|^2,\ \|G^a(x^k)\|^2\}. \tag{5.6}$$

For any k sufficiently large, apply Lemma 5.4, (5.6), and (5.4) to see

$$\begin{aligned} \|x^k - x^*\| &\leq \frac{\sigma_1 + \hat{\rho}}{\sigma_1\hat{\rho}}\left(\|G^a(x^k)\| + \|G^a(x^k) - G(x^k)\|\right) \\ &\leq \frac{\sigma_1 + \hat{\rho}}{\sigma_1\hat{\rho}}(\|G^a(x^k)\| + \hat{R}\sqrt{2\zeta_k/\hat{\rho}}) \\ &\leq \frac{\sigma_1 + \hat{\rho}}{\sigma_1\hat{\rho}}(\|G^a(x^k)\| + \sigma_2\|G^a(x^k)\|) \\ &= \frac{\sigma_1 + \hat{\rho}}{\sigma_1\hat{\rho}}(1 + \sigma_2)\|G^a(x^k)\|. \end{aligned} \tag{5.7}$$

For k sufficiently large, the definition of x^{k+1} and the stabilization of R_ℓ imply

$$\begin{aligned} x^{k+1} - x^k &= p^a(x^k) - x^k + \tau_k d^k \\ &= p^a(x^k) - x^k - \tau_k\hat{R}\left(B_k{}^{-1} - \hat{R}^{-1}I\right)\left(x^k - p^a(x^k)\right) \\ &= \left[I + \tau_k\hat{R}\left(B_k{}^{-1} - \hat{R}^{-1}I\right)\right]\left(p^a(x^k) - x^k\right)\ . \end{aligned} \tag{5.8}$$

Because $x^k \to x^*$ we have $x^{k+1} - x^k \to 0$. In view of (5.8) and the boundedness of $\{\tau_k\}$ and $\left\{B_k{}^{-1}\right\}$ we get $p^a(x^k) - x^k \to 0$. Because $x^k \to x^*$ we have $p^a(x^k) \to x^*$. The definition of $G^a(x^k)$ and the stabilization of R_ℓ then give $\|G^a(x^k)\| \to 0$, and hence the operation in Line 3 of Procedure quasi-Newton implies the existence of an index $\bar{k}$ sufficiently large such that $\bar{k} \in K_0$, giving $\tau_{\bar{k}} = 1$.

By [15, Theorem XV.4.1.4] the convexity of $h^x_{\bar{\eta}}(\cdot)$ on Γ yields the Lipschitz continuity of $\nabla e_{\hat{\rho}} h^x_{\bar{\eta}}(x) = G(x)$ on Γ with modulus $\hat{\rho}$. From Lemma 5.4, (5.6), and the fact that $G(x^*) = 0$ we can get

$$\begin{aligned}\|G^a(x^{\bar{k}+1})\| &\le \|G^a(x^{\bar{k}+1}) - G(x^{\bar{k}+1})\| + \|G(x^{\bar{k}+1})\| \\ &\le \hat{R}\sqrt{2\zeta_{\bar{k}+1}/\hat{\rho}} + \|G(x^{\bar{k}+1})\| \\ &\le (1+\sigma_2)\|G(x^{\bar{k}+1})\| \\ &\le (1+\sigma_2)\|G(x^{\bar{k}+1}) - G(x^*)\| \\ &\le (1+\sigma_2)\hat{\rho}\|x^{\bar{k}+1} - x^*\|,\end{aligned} \tag{5.9}$$

where the last inequality follows from the Lipschitz continuity of $G(x)$ on Γ. The reasoning in (5.7) and (5.9) yields

$$\|G^a(x^{\bar{k}}) - G(x^{\bar{k}})\| \le \sigma_2\hat{\rho}\|x^{\bar{k}} - x^*\|\,. \tag{5.10}$$

From the definition of $G(x)$ and Lemma 2.5(5.b) we see that $G(x)$ is single-valued on Γ. Recalling the definition of a semismooth function [34, Page 3], the assumption that $G(x)$ is semismooth at x^* entails the condition that $x^* \in int\,\Gamma$. Therefore, there exists an open ball $B(x^*, r) \subseteq \Gamma$. As $G(x)$ is semismooth at x^*, applying Theorem 2.3 in [34] gives

$$\|G(x^* + h) - G(x^*) - Vh\| = o(\|h\|) \text{ with } V \in \partial G(x^* + h).$$

The indices k_1 and k_2 can be chosen sufficiently big such that the parameters σ_2 and σ_3 are sufficiently small and

$$\frac{c\sigma_1}{(\sigma_1+\hat{\rho})(1+\sigma_2)^2} - \sigma_3 - \sigma_4\sigma_2\hat{\rho} > 0\,. \tag{5.11}$$

Consequently, when x is sufficiently close to x^*,

$$\sigma_4\|G(x) - G(x^*) - V(x-x^*)\| \le \left(\frac{c\sigma_1}{(\sigma_1+\hat{\rho})(1+\sigma_2)^2} - \sigma_3 - \sigma_4\sigma_2\hat{\rho}\right)\|x-x^*\| \text{ where } V \in \partial G(x). \tag{5.12}$$

In view of the fact $\tau_{\bar{k}} = 1$, (3.17), (5.10), (5.12), (5.11), and (5.7), we have

$$\begin{aligned}\|x^{\bar{k}+1} - x^*\| &= \|x^{\bar{k}} - B_{\bar{k}}^{-1}G^a(x^{\bar{k}}) - x^*\| \\ &= \|(x^{\bar{k}} - x^*) - B_{\bar{k}}^{-1}V_{\bar{k}}(x^{\bar{k}} - x^*) - B_{\bar{k}}^{-1}(G^a(x^{\bar{k}}) - G(x^*) - V_{\bar{k}}(x^{\bar{k}} - x^*))\| \\ &\le \|I - B_{\bar{k}}^{-1}V_{\bar{k}}\|\|x^{\bar{k}} - x^*\| + \|B_{\bar{k}}^{-1}\|\|G^a(x^{\bar{k}}) - G(x^*) - V_{\bar{k}}(x^{\bar{k}} - x^*)\| \\ &\le \sigma_3\|x^{\bar{k}} - x^*\| + \sigma_4(\|G^a(x^{\bar{k}}) - G(x^{\bar{k}})\| + \|G(x^{\bar{k}}) - G(x^*) - V_{\bar{k}}(x^{\bar{k}} - x^*)\|) \\ &\le (\sigma_3 + \sigma_4\sigma_2 L)\|x^{\bar{k}} - x^*\| + \sigma_4\|G(x^{\bar{k}}) - G(x^*) - V_{\bar{k}}(x^{\bar{k}} - x^*)\| \\ &\le \frac{c\sigma_1}{(\sigma_1+\hat{\rho})(1+\sigma_2)^2}\|x^{\bar{k}} - x^*\| \\ &\le \frac{c}{(1+\sigma_2)\hat{\rho}}\|G^a(x^{\bar{k}})\|\,.\end{aligned} \tag{5.13}$$

Combining (5.9) and (5.13) reveals

$$\|G^a(x^{\bar{k}+1})\| \leq c\|G^a(x^{\bar{k}})\|. \tag{5.14}$$

The proven conditions $p^a(x^k) \to x^*$ and $G^a(x^k) \to 0$ together with the definition of d^k imply $d^k \to 0$. Consequently, the continuity of f gives $f\left(p^a(x^k) + d^k\right) \to f(x^*)$. The fact that $x^* \in int\ \Gamma$ gives $f(x^*) < f(\bar{x}^0) + M_0$ and therefore we have $f\left(p^a(x^k) + d^k\right) \leq f\left(\bar{x}^0\right) + M_0$ for k sufficiently large. Together with (5.14) and (3.15), we see that the sufficiently large $\bar{k}$ satisfies $\bar{k} + 1 \in K_0$. This implies that for all $k \geq \bar{k}$, we have $k \in K_0$ and $\tau_k \equiv 1$.

In what follows, we show that the sequence $\{x^k\}$ converges to x^* superlinearly. By Lemma 5.4 and $\zeta_k = o(\|G^a(x^k)\|^2)$, we have

$$\|G^a(x^k) - G(x^k)\| \leq \hat{R}\sqrt{2\zeta_k/\hat{\rho}} = o(\|G(x^k)\|) = o(\|x^k - x^*\|), \tag{5.15}$$

where the last equality follows from the Lipschitz continuity of $G(x)$. We know that

$$\|G(x^k) - G(x^*) - V_k(x^k - x^*)\| = o(\|x^k - x^*\|) \text{ with } V_k \in \partial G(x^k) \tag{5.16}$$

by the semismoothness of $G(x)$ at stationary point x^*.

For all $k \geq \bar{k}$, in view of the fact $\tau_k \equiv 1$ and (3.17), we obtain

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \|x^k - B_k^{-1}G^a(x^k) - x^*\| \\
&= \|(x^k - x^*) - B_k^{-1}V_k(x^k - x^*) - B_k^{-1}(G^a(x^k) - G(x^*) - V_k(x^k - x^*))\| \\
&\leq \|I - B_k^{-1}V_k\|\|x^k - x^*\| + \|B_k^{-1}\|\|G^a(x^k) - G(x^k)\| \\
&+ \|B_k^{-1}\|\|G(x^k) - G(x^*) - V_k(x^k - x^*)\|,
\end{aligned}$$

where $V_k \in \partial G(x^k)$ is such that $\|B_k - V_k\| = \text{dist}(B_k, \partial G(x^k))$. Combining (5.5), (5.15), and (5.16) we get

$$\|x^{k+1} - x^*\| = o(\|x^k - x^*\|).$$

□

6 Numerical Experiments

In this section, we evaluate the numerical performances of Algorithm 1. The algorithm was implemented in MATLAB (R2017a) with the QP (3.5) solved by the famous software MOSEK. We tested a set of 20 problems listed in Table 1. Problems 1 to 8 are from [28], and the remainder are from [18, Section 2] with three sets of dimensions. We used the same initial points of the problems as specified in those references.

We manage the size of the bundle via the following strategy. If the number of cutting-planes stored in the bundle is greater than the number $N := \min\{10n, 50\}$, then we delete from J_ℓ all the indices i such that $\lambda_i^\ell = 0$.

In our tests, the parameters were selected as $m_1 = 0.15$, $m_2 = 0.05$, $c = 0.99$, $\gamma = 0.4$, $\varepsilon =$ 1E-5; the unacceptable increase parameter $M_0 = 2000$ for `Wong1` and $M_0 = 10$ for other all test problems; the convexification growth parameter $\Theta = 3$ for `Wong1` and $\Theta = 2$ for other all test problems; the model prox-parameter $\rho_0 = 3$ for `EVD52`, $\rho_0 = 0.1$ for `Active Faces` and `Brown 2` and for other test problems ρ_0 was set according to the following formula,

$$\rho_0 = \begin{cases} 100, & \text{if } |f(x^0)| <= \text{2E-13}, \\ \frac{\|g^0\|}{0.2|f(x^0)|}, & \text{otherwise.} \end{cases}$$

Table 1: Tested nonconvex problems

No.	Problem	Dimension	Minimal Value
1	Crescent	2	0
2	Colville 1	5	-32.348679
3	HS78	5	-2.9197004
4	El-Attar	6	0.5598131
5	Gill	10	9.7857721
6	Steiner 2	12	16.703838
7	EVD52	3	3.5997193
8	Wong1	7	680.63006
9	Active Faces	2	0
10	Brown 2	2	0
11	Chained Crescent I	2	0
12	Chained Crescent II	2	0
13	Active Faces	10	0
14	Brown 2	10	0
15	Chained Crescent I	10	0
16	Chained Crescent II	10	0
17	Active Faces	100	0
18	Brown 2	100	0
19	Chained Crescent I	100	0
20	Chained Crescent II	100	0

To evaluate the efficiency of Algorithm 1, we also implemented a version of the algorithm without the quasi-Newton procedure, denoted by NoQN. The only difference is that this version skips the Step 6, i.e., if a serious step is declared then set $x^{k+1} = y^{\ell+1}$ and go directly to Step 7. All the other settings including bundle compression and parameters were the same as in the test for Algorithm 1. Essentially, the algorithm NoQN is just the redistributed bundle method developed in [13].

Table 2: Numerical results

No.	Alg. 1			NoQN [13]		
	`nf`	f^*	CPU	`nf`	f^*	CPU
1	53	8.03E-07	0.903	29	0.914173	2.641
2	102	-32.3487	0.410	135	-32.3486	1.024
3	193	-2.91968	0.676	27	-2.67144	0.256
4	172	0.559815	0.603	73	0.858586	0.571
5	402	9.786009	1.156	669	10.10416	5.107
6	288	16.70384	0.652	80	16.70385	0.676
7	67	3.599724	0.401	163	3.599921	1.668
8	215	680.6301	0.742	187	680.6305	1.583
9	12	1.24E-08	0.168	9	1.32E-08	0.082
10	17	9.22E-07	0.070	13	2.88E-08	0.076
11	53	8.03E-07	0.128	29	0.914173	0.158
12	53	8.03E-07	0.121	29	0.914173	0.176
13	27	0.003193	0.115	6	6.53E-07	0.033
14	66	3.88E-06	0.125	22	4.21E-06	0.122
15	220	7.89E-07	0.308	10	0.635473	0.054
16	225	7.80E-06	0.532	10	0.646517	0.055
17	71	0.003951	0.655	51	1.05E-05	0.385
18	71	2.48E-01	0.400	36	5.30E-06	0.247
19	242	3.96E-08	0.668	40	6.751946	0.258
20	534	7.67E-06	3.884	24	0.321092	0.147

The numerical results are listed in Table 2 where the `nf`, f^*, and CPU in the column header respectively refer to the number of function evaluations, the best value returned by the algorithm, and the CPU time in seconds consumed by the algorithm. A comparison of the results of Algorithm 1 and those of the algorithm NoQN suggests that Algorithm 1,

Table 3: Results for Algorithm NoQN run to the same nf with Alg. 1

No.	NoQN [13]		Alg. 1	
	f^*	CPU	f^*	CPU
1	9.1414E-01	0.523	8.0272E-07	0.903
2	-3.2337E+01	0.635	-3.2349E+01	0.410
3	-2.9000E+00	1.137	-2.9197E+00	0.676
4	5.5981E-01	1.100	5.5982E-01	0.603
5	1.0125E+01	2.709	9.7860E+00	1.156
6	1.6703838E+01	2.138	1.6703841E+01	0.652
7	3.6323E+00	0.448	3.5997E+00	0.401
8	6.806302E+02	1.519	6.806301E+02	0.742
9	1.3234E-08	0.064	1.2448E-08	0.168
10	1.1824E-09	0.086	9.2185E-07	0.070
11	9.1415E-01	0.287	8.0272E-07	0.128
12	9.1415E-01	0.287	8.0272E-07	0.121
13	6.5312E-07	0.146	3.1933E-03	0.115
14	3.4441E-10	0.362	3.8771E-06	0.125
15	1.1582E-09	1.237	7.8896E-07	0.308
16	9.9683E-03	1.296	7.8008E-06	0.532
17	1.0453E-05	0.539	3.9513E-03	0.655
18	1.1461E-08	0.510	2.4771E-01	0.400
19	2.4158E+00	1.933	3.9615E-08	0.668
20	1.2238E-03	6.933	7.6742E-06	3.884

with quasi-Newton procedure, consumes more function evaluations. However, it yields much more accuracy in the optimal values of the functions. Some of the optimal values returned by algorithm NoQN is very far away from the true optimal values.

To investigate the impact of the quasi-Newton procedure on the number of function evaluations, we changed the stopping criterion of NoQN to the following: the algorithm is terminated only if the consumed function evaluations reaches to the number `nf` recorded in Table 2. That is to say, we ran the algorithm NoQN without stop until it consumed the same number of function evaluations as did by Algorithm 1. Table 3 lists the results and Figure 1 is a comparison of the performance profiles of the two versions. From Figure 1 we can see that Algorithm 1 yielded better optimal values and less CPU time. This suggests that with a quasi-Newton procedure, our algorithm can have greater performance than the redistributed bundle method.

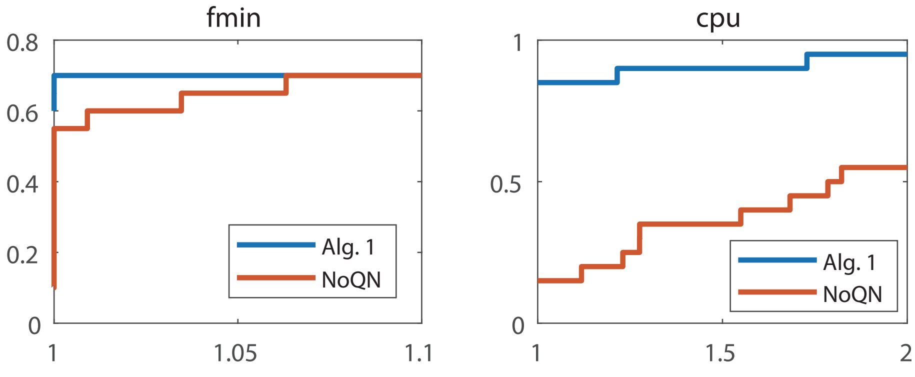

Figure 1: Performance profiles

$\boxed{7}$ Conclusions

In this paper we have proposed an accelerated bundle-type method for solving unconstrained optimization problem, and the objective function can be locally Lipschitz continuous but neither smooth nor convex. This method combines the redistributed bundle method and the idea of a proximal quasi-Newton method. At each iteration, the convexification parameter and the prox-parameter are suitably modified to guarantee that the proximal point of a piecewise affine model of a local convexification function approximates well-enough the proximal point of f at x^k. We have incorporated a quasi-Newton procedure at the end of a serious step. Our main results are the global convergence and superliner convergence of the proposed algorithm, which are stated in Theorem 4.6 and 5.5, respectively. Numerical results show that the proposed algorithm is promising.

Acknowledgements

The authors sincerely thank the referees as well as the associate editor for their constructive comments which have significantly improved the quality of the paper.

References

[1] J. Bolte, S. Sabach and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Math. Program.* 146 (2014) 459–494.

[2] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal and C. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*, Springer-Verlag, Berlin, Heidelberg, 2006.

[3] J.V. Burke, A.S. Lewis and M.L. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization, *SIAM J. Optim.* 15 (2005) 751–779.

[4] X.J. Chen and M. Fukushima, Proximal quasi-Newton methods for nondifferentiable convex optimization, *Math. Program.* 85 (1999) 313–334.

[5] E.W. Cheney and A.A. Goldstein, Newton's method for convex programming and Tchebycheff approximations, *Numer. Math.* 1 (1959) 253–268.

[6] D. Davis and B. Grimmer, Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems, *SIAM J. Optim.* 29 (2019) 1908–1930.

[7] T.M.T. Do and T. Artières, Regularized bundle methods for convex and non-convex risks, *J. Mach. Learn. Res.* 13 (2012) 3539–3583.

[8] D.L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* 52 (2006) 1289–1306.

[9] A. Fuduli, M. Gaudioso and G. Giovanni, Minimizing nonconvex nonsmooth functions via cutting planes and proximity control, *SIAM J. Optim.* 14 (2004) 743–756.

[10] A. Fuduli, M. Gaudioso and E.A. Nurminski, A splitting bundle approach for nonsmooth nonconvex minimization, *Optimization* 64 (2015) 1131–1151.

[11] M. Fukushima and L.Q. Qi, A globally and superlinearly convergent algorithm for nonsmooth convex minimization, *SIAM J. Optim.* 6 (1996) 1106–1120.

[12] W. Hare and C. Sagastizábal, Computing proximal points of nonconvex functions, *Math. Program.* 116 (2009) 221–258.

[13] W. Hare and C. Sagastizábal, A redistributed proximal bundle method for nonconvex optimization, *SIAM J. Optim.* 20 (2010) 2442–2473.

[14] W. Hare, C. Sagastizábal and M. Solodov, A proximal bundle method for nonsmooth nonconvex functions with inexact information, *Comput. Optim. Appl.* 63 (2016) 1–28.

[15] J.B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, Springer-Verlag, Berlin, Heidelberg, 1993.

[16] P. Jain and P. Kar, Non-convex optimization for machine learning, *Found. Trends Mach. Learn.* 10 (2017) 142–336.

[17] E. Karas, A. Ribeiro, C. Sagastizábal and M, Solodov, A bundle-filter method for nonsmooth convex constrained optimization, *Math. Program.* 116 (2007) 297–320.

[18] N. Karmitsa, Test problems for large-scale nonsmooth minimization, Department of Mathematical Information Technology, University of Jyväskylä, Finland, No. B. 4/2007 (2007) 297–320.

[19] J.E. Kelley, The cutting-plane method for solving convex programs, *J. Soc. Indust. Appl. Math.* 8 (1960) 703–712.

[20] K.C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture notes in mathematics 1133, Springer-Verlag, Berlin, Heidelberg, 1985.

[21] K.C. Kiwiel, A proximal bundle method with approximate subgradient linearizations, *SIAM J. Optim.* 16 (2006) 1007–1023.

[22] K.C. Kiwiel, An alternating linearization bundle method for convex optimization and nonlinear multicommodity flow problems, *Math. Program.* 130 (2011) 59–84.

[23] G.H. Lan, Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization, *Math. Program.* 149 (2015) 1–45.

[24] C. Lemaréchal, An extension of Davidon methods to nondifferentiable problems, in: *Nondifferentiable Optimization*, Mathematical Programming Study 3, Springer, Berlin, Heidelberg, 1975, 95–109.

[25] C. Lemaréchal, Y. Nesterov and A. Nemirovskii, New variants of bundle methods, *Math. Program.* 69 (1995) 111–147.

[26] C. Lemaréchal and C. Sagastizábal, Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries, *SIAM J. Optim.* 7 (1997) 367–385.

[27] L. Lukšan and J. Vlček, A bundle-Newton method for nonsmooth unconstrained minimization, *Math. Program.* 83 (1998) 373–391.

[28] L. Lukšan and J. Vlček, Test problems for nonsmooth unconstrained and linearly constrained optimization, Technical Report No.798, Institute of Computer Science, Academy of Science of the Czech Republic, Prague, 2000.

[29] M.M. Mäkelä, Survey of bundle methods for nonsmooth optimization, *Optim. Methods Softw.* 17 (2002) 1–29.

[30] R. Mifflin, D.F. Sun and L.Q. Qi, Quasi-Newton bundle-type methods for nondifferentiable convex optimization, *SIAM J. Optim.* 8 (1998) 583–603.

[31] W. de Oliveira, C. Sagastizábal and C. Lemaréchal, Convex proximal bundle methods in depth: a unified analysis for inexact oracles, *Math. Program.* 148 (2014) 241–277.

[32] W. de Oliveira and M. V. Solodov, A doubly stabilized bundle method for nonsmooth convex optimization, *Math. Program.* 156 (2016) 125–159.

[33] T. Pock and S. Sabach, Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems, *SIAM J. Imaging Sci.* 9 (2016) 1756–1787.

[34] L.Q. Qi and J. Sun, A nonsmooth version of Newton's method, *Math. Program.* 58 (1993) 353–367.

[35] R.T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM J. Control Optim.* 14 (1976) 877–898.

[36] R.T. Rockafellar and R.J-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, Heidelberg, 1998.

[37] A. Ruszczyński, A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization, *SIAM J. Control Optim.* 59 (2021) 2301–2320.

[38] H. Schramm and J. Zowe, A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results, *SIAM J. Optim.* 2 (1992), 121–152.

[39] I.W. Selesnick, A. Parekh and I. Bayram, Convex 1-d total variation denoising with non-convex regularization, *IEEE Signal Process. Lett.* 22 (2015) 141–144.

[40] J. Shen, Y.L. Gao, F.F. Guo and R. Zhao, A redistributed bundle algorithm for generalized variational inequality problems in Hilbert spaces, *Asia-Pac. J. Oper. Res.* 35 (2018): 1850019.

[41] J. Shen, J.T. Li, F.F. Guo and N. Xu, A redistributed bundle algorithm based on local convexification models for nonlinear nonsmooth DC programming, *J. Numer. Math.* 29 (2021) 159–170.

[42] C.M. Tang, B. He and Z.Z. Wang, Modified accelerated bundle-level methods and their application in two-stage stochastic programming. *Mathematics* 8 (2020): 265.

[43] C.M. Tang, J.B. Jian and G.Y. Li, A proximal-projection partial bundle method for convex constrained minimax problems, *J. Ind. Manag. Optim.* 15 (2019) 757–774.

[44] C.M. Tang, S. Liu, J.B. Jian and J.L. Li, A feasible SQP-GS algorithm for nonconvex, nonsmooth constrained optimization, *Numer. Algorithms* 65 (2014) 1–22.

[45] C.M. Tang, S. Liu, J. B. Jian and X.M. Ou, A multi-step doubly stabilized bundle method for nonsmooth convex optimization, *Appl. Math. Comput.* 376 (2020): 125154.

[46] P. Wolfe, A method of conjugate subgradients for minimizing nondifferentiable functions, in: *Nondifferentiable Optimization*, Mathematical Programming Study 3, Springer, Berlin, Heidelberg, 1975, pp. 145–173.

[47] Y.X. Yuan and W.Y. Sun, *Optimization Theory and Methods (in Chinese)*, Science Press, Beijing, 1997.

Manuscript received 5 June 2021
revised 16 September 2021
accepted for publication 13 October 2021

Chunming Tang
College of Mathematics and Information Science
Guangxi University, Nanning 530004, P.R. China
E-mail address: cmtang@gxu.edu.cn

Huangyue Chen
College of Mathematics and Information Science
Guangxi University, Nanning 530004, P.R. China
E-mail address: hychen_math@163.com

Jinbao Jian
College of Mathematics and Physics
Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis
Center for Applied Mathematics and Artificial Intelligence
Guangxi University of Nationalities
Nanning 530006, P. R. China E-mail address: jianjb@gxu.edu.cn

Shuai Liu
School of Software, South China Normal University
Nanhai Campus, Foshan 528225, P.R. China
E-mail address: shuai0liu@gmail.com