Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 697474, 7 pages
http://dx.doi.org/10.1155/2013/697474

Research Article

An Approximate Quasi-Newton Bundle-Type Method for Nonsmooth Optimization

Jie Shen,[1] Li-Ping Pang,[2] and Dan Li[2]

[1] *School of Mathematics, Liaoning Normal University, Dalian 116029, China*
[2] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*

Correspondence should be addressed to Jie Shen; tt010725@163.com

Received 22 January 2013; Revised 31 March 2013; Accepted 1 April 2013

Academic Editor: Gue Lee

An implementable algorithm for solving a nonsmooth convex optimization problem is proposed by combining Moreau-Yosida regularization and bundle and quasi-Newton ideas. In contrast with quasi-Newton bundle methods of Mifflin et al. (1998), we only assume that the values of the objective function and its subgradients are evaluated approximately, which makes the method easier to implement. Under some reasonable assumptions, the proposed method is shown to have a Q-superlinear rate of convergence.

1. Introduction

In this paper we are concerned with the unconstrained minimization of a real-valued, convex function $f : R^n \to R$, namely,

$$\begin{aligned} &\min \quad f(x) \\ &\text{s.t.} \quad x \in R^n, \end{aligned} \tag{1}$$

and in general f is nondifferentiable. A number of attempts have been made to obtain convergent algorithms for solving (1). Fukushima and Qi [1] propose an algorithm for solving (1) under semismoothness and regularity assumptions. The proposed algorithm is shown to have a Q-superlinear rate of convergence. An implementable BFGS method for general nonsmooth problems is presented by Rauf and Fukushima [2], and global convergence is obtained based on the assumption of strong convexity. A superlinearly convergent method for (1) is proposed by Qi and Chen [3], but it requires the semismoothness condition. He [4] obtains a globally convergent algorithm for convex constrained minimization problems under certain regularity and uniform continuity assumptions. Among methods for nonsmooth optimization problems, some have superlinear rate of convergence, for instance, see Mifflin and Sagastizábal [5] and Lemaréchal et al. [6]. They propose two conceptual algorithms with superlinear convergence for minimizing a class of convex functions, and the latter demands that the objective function f should be differentiable in a certain space U (the subspace along which $\partial f(p)$ has 0 breadth at a given point p), but sometimes it is difficult to decompose the space. Besides these methods mentioned above, there is a quasi-Newton bundle type method proposed by Mifflin et al. [7] it has superlinear rate of convergence, but the exact values of the objective function and its subgradients are required. In this paper, we present an implementable algorithm by using bundle and quasi-Newton ideas and Moreau-Yosida regularization, and the proposed algorithm can be shown to have a superlinear rate of convergence. An obvious advantage of the proposed algorithm lies in the fact that we only need the approximate values of the objective function and its subgradients.

It is well known that (1) can be solved by means of the Moreau-Yosida regularization $F : R^n \to R$ of f, which is defined by

$$F(x) = \min_{z \in R^n} \left\{ f(z) + (2\lambda)^{-1} \|z - x\|^2 \right\}, \tag{2}$$

where λ is a fixed positive parameter and $\|\cdot\|$ denotes the Euclidean norm or its induced matrix norm on $R^{n\times n}$. The problem of minimizing $F(x)$, that is,

$$\begin{aligned} &\min \quad F(x) \\ &\text{s.t.} \quad x \in R^n, \end{aligned} \tag{3}$$

is equivalent to (1) in the sense that $x \in R^n$ solves (1) if and only if it solves (3), see Hiriart-Urruty and Lemaréchal [8]. The problem (3) has a remarkable feature that the objective function F is a differentiable convex function, even though f is nondifferentiable. Moreover F has a Lipschitz continuous gradient

$$G(x) = \lambda^{-1}(x - p(x)) \in \partial f(p(x)), \tag{4}$$

where $p(x)$ is the unique minimizer of (2) and ∂f is the subdifferential mapping of f. Hence, by Rademacher's theorem, G is differentiable almost everywhere and the set

$$\partial_B G(x) = \left\{ D \in R^{n\times n} \mid D = \lim_{x^k \to x} \nabla G\left(x^k\right), \text{ where } G \text{ is differentiable at } x^k \right\} \tag{5}$$

is nonempty and bounded for each x. We say G is BD-regular at x if all matrices $D \in \partial_B G(x)$ are nonsingular. It is reasonable to pay more attention to the problem (3) since F has such good properties. However, because the Moreau-Yosida regularization itself is defined through a minimization problem involving f, the exact values of F and its gradient G at an arbitrary point x are difficult or even impossible to compute in general. Therefore, we attempt to explore the possibility of utilizing the approximations of these values.

Several attempts have been made to combine quasi-Newton idea with Moreau-Yosida regularization to solve (1). For related works on this subject, see Chen and Fukushima [9] and Mifflin [10]. In particular, Mifflin et al. [7] consider using bundle ideas to approximate linearly the values of f in order to approximate F in which the exact values of f and one of its subgradients g at some points are needed. In this paper we assume that for given $x \in R^n$ and $\varepsilon \geq 0$, we can find some $\widetilde{f} \in R$ and $g^a(x,\varepsilon) \in R^n$ such that

$$\begin{aligned} f(x) \geq \widetilde{f} \geq f(x) - \varepsilon, \\ f(z) \geq \widetilde{f} + \left\langle g^a(x,\varepsilon), z - x \right\rangle, \quad \forall z \in R^n, \end{aligned} \tag{6}$$

which means that $g^a(x,\varepsilon) \in \partial_\varepsilon f(x)$. This setting is realistic in many applications, see Kiwiel [11]. Let us see some examples. Assume that f is strongly convex with modulus $\mu > 0$, that is,

$$\begin{aligned} f(x) + g(x)^T (z - x) + \frac{\mu}{2}\|z - x\|^2 \leq f(z), \\ \forall z, x \in R^n, \quad g(x) \in \partial f(x), \end{aligned} \tag{7}$$

and that $f(x) = w(v(x))$ with $v : R^n \to R^m$ continuously differentiable and $w : R^m \to R$ convex. By the chain rule we have $\partial f(x) = \{\sum_{i=1}^m \xi_i \nabla v_i(x) \mid \xi = (\xi_1, \xi_2, \ldots, \xi_n)^T \in \partial w(v(x))\}$. Now assume that we have an approximation $\nabla_h v(x)$ of $\nabla v(x)$ such that $\|\nabla_h v(x) - \nabla v(x)\| \leq \kappa(h)$, $h > 0$. Such an approximation may be obtained by using finite differences. In this case, typically $\kappa(h) \to 0$ for $h \to 0$. Let $g_h(x) = \sum_{i=1}^m \xi_i \nabla_h v_i(x)$, $\xi \in \partial w(v(x))$. Then, we have

$$\begin{aligned} & f(x) + g_h(x)^T (z - x) \\ & \quad \leq f(x) + g(x)^T (z - x) \\ & \qquad + \|\xi\| \left\|\nabla_h v(x) - \nabla v(x)\right\| \|z - x\| \\ & \quad \leq f(z) - \frac{\mu}{2}\|z - x\|^2 + \kappa(h) \|\xi\| \|z - x\| \end{aligned} \tag{8}$$

for all $x, z \in R^n$ and $g(x) = \sum_{i=1}^m \xi_i \nabla v_i(x) \in \partial f(x)$. Some simple manipulations show that

$$\begin{aligned} & -\frac{\mu}{2}\|z - x\|^2 + \kappa(h) \|\xi\| \|z - x\| \\ & \quad \leq \frac{1}{2\mu}\|\xi\|^2 \kappa(h)^2 =: \varepsilon_h, \quad \forall x, z \in R^n. \end{aligned} \tag{9}$$

By the definition of ξ, the bound ε_h depends on x, we obtain

$$f(x) + g_h(x)^T (z - x) \leq f(z) + \varepsilon_h, \quad \forall z \in R^n. \tag{10}$$

From the local boundedness of $\partial w(v(x))$, we infer that $\varepsilon_h > 0$ is locally bounded. Thus, $g_h(x)$ is an ε_h-subgradient of f at x, see Hintermüller [12]. As for the approximate function values, if f is a max-type function of the form

$$f(x) = \sup\{\phi_u(x) \mid u \in U\}, \quad \forall x \in R^n, \tag{11}$$

where each $\phi_u : R^n \to R$ is convex and U is an infinite set, then it may be impossible to calculate $f(x)$. However, for any positive ε one can usually find in finite time an ε-solution to the maximization problem (11), that is, an element $u_\varepsilon \in U$ satisfying $\phi_{u_\varepsilon} \geq f(x) - \varepsilon$. Then one may set $f_\varepsilon(x) = \phi_{u_\varepsilon}(x)$. On the other hand, in some applications, calculating u_ε for a prescribed $\varepsilon \geq 0$ may require much less work than computing u_0. This is, for instance, the case when the maximization problem (11) involves solving a linear or discrete programming problem by the methods of Gabasov and Kirilova [13]. Some people have tried to solve (1) by assuming the values of the objective function, and its subgradients can only be computed approximately. For example, Solodov [14] considers the proximal form of a bundle algorithm for (1), assuming the values of the function and its subgradients are evaluated approximately, and it is shown how these approximations should be controlled in order to satisfy the desired optimality tolerance. Kiwiel [15] proposes an algorithm for (1), and the algorithm utilizes the approximation evaluations of the objective function and its subgradients; global convergence of the method is obtained. Kiwiel [11] introduces another method for (1); it requires only the approximate evaluations of f and its ε-subgradients, and this method converges globally. It is in evidence that bundle methods with superlinear convergence for solving (1) by using approximate values of the objective and its subgradients are seldom obtained. Compared with the methods mentioned above, the method proposed in this paper is not only implementable but also has a superlinear

rate of convergence under some additional assumptions, and it should be noted that we only use the approximate values of the objective function and its subgradients which makes the algorithm easier to implement.

Some notations are listed below for presenting the algorithm.

(i) $\partial f(x) = \{\xi \in R^n \mid f(z) \geq f(x) + \xi^T(z-x),\ \forall z \in R^n\}$, the subdifferential of f at x, and each such ξ is called a subgradient of f at x.

(ii) $\partial_\varepsilon f(x) = \{\eta \in R^n \mid f(z) \geq f(x) + \eta^T(z-x) - \varepsilon\}$, the ε-subdifferential of f at x, and each such η is called an ε-subgradient of f at x.

(iii) $p(x) = \arg\min_{z\in R^n}\{f(z)+(2\lambda)^{-1}\|z-x\|^2\}$, the unique minimizer of (2).

(iv) $G(x) = \lambda^{-1}(x - p(x))$, the gradient of F at x.

This paper is organized as follows: in Section 2, to approximate the unique minimizer $p(x)$ of (2), we introduce the bundle idea, which uses approximate values of the objective function and its subgradients. The approximate quasi-Newton bundle-type algorithm is presented in Section 3. In the last section, we prove the global convergence and, under additional assumptions, Q-superlinear convergence of the proposed algorithm.

2. The Approximation of $p(x)$

Let $x = x^k$ and $s = z - x^k$, where x^k is the current iterate point of AQNBT algorithm presented in Section 3, then (13) has the form

$$F\left(x^k\right) = \min_{s\in R^n}\left\{f\left(x^k+s\right) + (2\lambda)^{-1}\|s\|^2\right\}. \tag{12}$$

Now we consider approximating $f(x^k+s)$ by using the bundle idea. Suppose we have a bundle J^k generated sequentially starting from x^k and possibly a subset of the previous set used to generate x^k. The bundle includes the data $(z^i, \widetilde{f}^i, g^a(z^i,\varepsilon_i))$, $i \in J^k$, where $z^i \in R^n$, $\widetilde{f}^i \in R$, and $g^a(z^i,\varepsilon_i) \in R^n$ satisfy

$$\begin{aligned} f\left(z^i\right) &\geq \widetilde{f}^i \geq f\left(z^i\right) - \varepsilon_i, \\ f(z) &\geq \widetilde{f}^i + \left\langle g^a\left(z^i,\varepsilon_i\right), z - z^i\right\rangle, \quad \forall z \in R^n. \end{aligned} \tag{13}$$

Suppose that the elements in J^k can be arranged according to the order of their entering the bundle. Without loss of generality we may suppose $J^k = \{1,\ldots,j\}$. ε_i is updated by the rule $\varepsilon_{i+1} = \gamma\varepsilon_i$, $0 < \gamma < 1$, $i \in J^k$. The condition (13) means $g^a(z^i,\varepsilon_i) \in \partial_{\varepsilon_i} f(z^i)$, $i \in J^k$. By using the data in the bundle we construct a polyhedral function $f_a(x^k+s)$ defined by

$$f_a\left(x^k+s\right) = \max_{i\in J^k}\left\{\widetilde{f}^i + g^a\left(z^i,\varepsilon_i\right)^T\left(x^k+s-z^i\right)\right\}. \tag{14}$$

Obviously $f_a(x^k+s)$ is a lower approximation of $f(x^k+s)$, so $f_a(x^k+s) \leq f(x^k+s)$. We define a linearization error by

$$\alpha\left(x^k,z^i,\varepsilon_i\right) = \widetilde{f}^{x^k} - \widetilde{f}^i - g^a\left(z^i,\varepsilon_i\right)^T\left(x^k - z^i\right), \tag{15}$$

where $\widetilde{f}^{x^k} \in R$ satisfies

$$f\left(x^k\right) \geq \widetilde{f}^{x^k} \geq f\left(x^k\right) - \varepsilon_{x^k}, \quad \text{for given } \ \varepsilon_{x^k} \geq 0. \tag{16}$$

Then $f_a(x^k+s)$ can be written as

$$f_a\left(x^k+s\right) = \widetilde{f}^{x^k} + \max_{i\in J^k}\left\{g^a\left(z^i,\varepsilon_i\right)^T s - \alpha\left(x^k,z^i,\varepsilon_i\right)\right\}. \tag{17}$$

Let

$$\begin{aligned} F_a\left(x^k\right) &= \min_{s\in R^n}\left\{f_a\left(x^k+s\right) + (2\lambda)^{-1}\|s\|^2\right\} \\ &= \widetilde{f}^{x^k} + \min_{s\in R^n}\left\{\max_{i\in J^k}\left\{g^a\left(z^i,\varepsilon_i\right)^T s - \alpha\left(x^k,z^i,\varepsilon_i\right)\right\}\right. \\ &\qquad \left. + (2\lambda)^{-1}s^T s\right\}. \end{aligned} \tag{18}$$

The problem (18) can be dealt with by solving the following quadratic programming:

$$\begin{aligned} &\min \quad v + \lambda(2)^{-1}s^T s, \\ &\text{s.t.} \quad g^a\left(z^i,\varepsilon_i\right)^T s - \alpha\left(x^k,z^i,\varepsilon_i\right) \leq v \quad \forall i \in J^k. \end{aligned} \tag{19}$$

As iterations go along, the number of elements in bundle J^k increases. When the size of the bundle becomes too big, it may cause serious computational difficulties in the form of unbounded storage requirement. To overcome these difficulties, it is necessary to compress the bundle and clean the model. Wolfe [16] and Lemaréchal [17], for the first time, introduce the aggregation strategy, which requires storing only a limited number of subgradients, see Kiwiel and Mifflin [18–20]. Aggregation strategy is the synthesis mechanism that condenses the essential information of the bundle into one single couple $(\widetilde{g}_\varepsilon^k, \widehat{\alpha}_{\widetilde{k}})$ (defined below). The corresponding affine function, inserted in the model when there is compression, is called aggregate linearization (defined below). This function summarizes all the information generated up to iteration k. Suppose $J_{\max}$ is the upper bound of the number of elements in J^k, $k = 1, 2, \ldots$. If $|J^k|$ reaches the prescribed $J_{\max}$, two or more of those elements are deleted from the bundle J^k; that is, two or more linear pieces in the constraints of (19) are discarded (notice that different selections of discarded linear pieces may result in different speed of convergence), and introduce the aggregate linearization associated with the aggregate ε-subgradient and linearization error into bundle. Define the aggregate linearization as

$$f_t\left(x^k+s\right) = \widetilde{f}^{x^k} + \left\langle \widetilde{g}_\varepsilon^k, s\right\rangle - \widehat{\alpha}_{\widetilde{k}}, \tag{20}$$

where $\widetilde{g}_\varepsilon^k = \sum_{i\in J^k}\mu_i g^a(z^i,\varepsilon_i)$, $\widehat{\alpha}_{\widetilde{k}} = \sum_{i\in J^k}\mu_i\alpha(x^k,z^i,\varepsilon_i)$. Multiplier $\mu = (\mu_i)_{i\in J^k}$ is the optimal solution of dual problem for (19), see Solodov [14]. By doing so, the surrogate aggregate linearization maintains the information of the deleted linear

pieces and at the same time the problem (19) is manageable since the number of the elements in J^k is limited. Suppose $s(x^k)$ solves the problem (19), and let $p^a(x^k) = x^k + s(x^k)$ be an approximation of $p(x^k)$ and $\varepsilon_{p^a(x^k)} = \varepsilon_{j+1} = \gamma\varepsilon_j$. Let

$$F^a\left(x^k\right) = \widetilde{f}^{p^a(x^k)} + \varepsilon_{p^a(x^k)} + (2\lambda)^{-1} s\left(x^k\right)^T s\left(x^k\right), \quad (21)$$

where $\widetilde{f}^{p^a(x^k)} \in R$ is chosen to satisfy

$$f\left(p^a\left(x^k\right)\right) \geq \widetilde{f}^{p^a(x^k)} \geq f\left(p^a\left(x^k\right)\right) - \varepsilon_{p^a(x^k)}. \quad (22)$$

The results stated below are fundamental and useful in the subsequent discussions.

(P1) $F_a(x^k) \leq F(x^k) \leq F^a(x^k)$.

(P2) $F^a(x^k) = F(x^k)$ if and only if $p^a(x^k) = p(x^k)$ and $\widetilde{f}^{p^a(x^k)} = f(p(x^k))$.

Note that $p(x^k)$ is the unique minimizer of (2) and (P1) and (P2) can be obtained by the definitions of $F^a(x^k)$, $F_a(x^k)$, and $F(x^k)$.

(P3)

(i) If we define $F_{ea}(x^k) = \min_{s\in R^n}\{\max_{i\in J^k}\{f(z^i) + g(z^i)^T(x^k + s - z^i)\} + (2\lambda)^{-1}s^T s\}$, where $g(z^i) \in \partial f(z^i)$, then $F_a(x^k) \to F_{ea}(x^k)$ as the new point $z^{j+1} = x^k + s(x^k)$ is appended into the bundle J^k infinitely.

(ii) Let $\varepsilon = \max_{i\in J^k}\{\varepsilon_i\}$. if $g^a(z^i, \varepsilon_i) = g(z^i) \in \partial f(z^i)$, then $F_a(x^k) \geq F_{ea}(x^k) - \varepsilon$.

Because $\varepsilon_i \to 0$ by the update rule $\varepsilon_{i+1} = \gamma\varepsilon_i$, $0 < \gamma < 1$, we have $g^a(z^i, \varepsilon_i) \to g(z^i)$ and $\widetilde{f}^i \to f(z^i)$. Thus $f_a(x^k + s) \to \max_{i\in J^k}\{f(z^i) + g(z^i)^T(x^k + s - z^i)\}$, so $F_{ea}(x^k) \to F_a(x^k)$. It is easy to see that $f_a(x^k + s) = \max_{i\in J^k}\{\widetilde{f}^i + g^a(z^i, \varepsilon_i)^T(x^k + s - z^i)\} \geq \max_{i\in J^k}\{f(z^i) + g^a(z^i, \varepsilon_i)^T(x^k + s - z^i) - \varepsilon_i\} \geq \max_{i\in J^k}\{f(z^i) + g(z^i)^T(x^k + s - z^i)\} - \varepsilon$. Therefore, $F_a(x^k) = \min_{s\in R^n}\{f_a(x^k + s) + (2\lambda)^{-1}\|s\|^2\} \geq F_{ea}(x^k) - \varepsilon$.

Let

$$a\left(x^k\right) = F^a\left(x^k\right) - F_a\left(x^k\right). \quad (23)$$

We accept $p^a(x^k)$ as an approximation of $p(x^k)$ based on the following rule:

$$a\left(x^k\right) < m\left(x^k\right)\min\left\{\lambda^{-2}s\left(x^k\right)^T s\left(x^k\right), L\right\}, \quad (24)$$

where $m(x^k)$ and L are given positive numbers and $m(x^k)$ is fixed during one bundling process; that is, $m(x^k)$ depends on x^k, see Step 1 in AQNBT algorithm presented in Section 3. If (24) is not satisfied, we let $z^{j+1} = x^k + s(x^k)$ and $\varepsilon_{j+1} = \gamma\varepsilon_j$, $0 < \gamma < 1$, and take $\widetilde{f}^{j+1} = \widetilde{f}^{p^a(x^k)}$ and $g^a(z^{j+1}, \varepsilon_{j+1}) \in R^n$ satisfying

$$f\left(z^{j+1}\right) \geq \widetilde{f}^{j+1} \geq f\left(z^{j+1}\right) - \varepsilon_{j+1},$$
$$f(z) \geq \widetilde{f}^{j+1} + \left\langle g^a\left(z^{j+1}, \varepsilon_{j+1}\right), z - z^{j+1}\right\rangle, \quad \forall z \in R^n, \quad (25)$$

and then append a new piece $\widetilde{f}^{j+1} + g^a(z^{j+1}, \varepsilon_{j+1})^T(x^k + s - z^{j+1})$ to (14), replace j by $j+1$, and solve (19) for finding a new $s(x^k)$ and $a(x^k)$ to be tested in (24). If this bundle process does not terminate, we have the following conclusion.

(P4) Suppose that x^k is not the minimizer of f. If (24) is never satisfied, then $a(x^k) \to 0$ as the new point z^{j+1} is appended into the bundle J^k infinitely.

Suppose that $|J^k| = |\{1, 2, \ldots, j\}| = j < J_{\max}$. Define the functions ϕ and φ_{j+1}, $j = 1, 2, \ldots$ by

$$\phi(z) = f(z) + (2\lambda)^{-1}\left\|z - x^k\right\|^2,$$
$$\varphi_{j+1}(z) = \max_{i\in J^k=\{1,2,\ldots,j\}}\left\{\widetilde{f}^i + g^a\left(z^i, \varepsilon_i\right)^T\left(z - z^i\right)\right\} + (2\lambda)^{-1}\left\|z - x^k\right\|^2. \quad (26)$$

Let z^{j+1} be the unique minimizer of $\min_{z\in R^n}\varphi_{j+1}(z)$, and let z^{j+2} be the unique minimizer of $\min_{z\in R^n}\varphi_{j+2}(z)$, where $\varphi_{j+2}(z) = \max_{i\in J^{k+1}}\{\widetilde{f}^i + g^a(z^i, \varepsilon_i)^T(z - z^i)\} + (2\lambda)^{-1}\|z - x^k\|^2$. Note that if $|\{1, 2, \ldots, j+1\}| = j + 1 < J_{\max}$, then let $J^{k+1} = \{1, 2, \ldots, j+1\}$, so $\varphi_{j+1}(z^{j+1}) \leq \varphi_{j+2}(z^{j+2})$; if $|\{1, 2, \ldots, j+1\}| = j + 1 = J_{\max}$, delete at least two elements from $\{1, 2, \ldots, j+1\}$, say q_1, q_2, and $q_1 \neq j+1$, $q_2 \neq j+1$, the order of the other elements in $\{1, 2, \ldots, j+1\}$ are left intact. Introduce an additional index $\widetilde{k}$ associated with the aggregated ε-subgradient and linearization error into J^{k+1} and let $J^{k+1} = \{1, 2, \ldots, q_1 - 1, q_1 + 1, \ldots, q_2 - 1, q_2 + 1, \ldots, \widetilde{k}, j+1\}$, so $|J^{k+1}| = j < J_{\max}$. By adjusting λ appropriately, we can make sure that z^{j+1} and z^{j+2} are not far away from x^k. According to the proof of Proposition 3, see Fukushima [21], we find that $\phi(z^j)$ has limit, say ϕ^*, and $\varphi_{j+1}(z^{j+1})$ also converges to ϕ^* as $j \to \infty$. By the definitions of $F(x^k)$ and $F^a(x^k)$ we have $F_a(x^k) \to F(x^k)$ and $F^a(x^k) \to F(x^k)$ as $j \to \infty$, so $a(x^k) \to 0$ as $j \to \infty$.

In the next part we give the definition of $G^a(x^k)$, which is the approximation of $G(x^k)$,

$$G^a\left(x^k\right) = \lambda^{-1}\left(x^k - p^a\left(x^k\right)\right) = -\lambda^{-1}s\left(x^k\right), \quad (27)$$

and some properties of $G^a(x^k)$ are discussed. It is easy to see that the approximation of $G(x^k)$ is associated with $F(x^k)$:

(P5) $\|G(x^k) - G^a(x^k)\| = \|\lambda^{-1}(p(x^k) - p^a(x^k))\| \leq \sqrt{2a(x^k)/\lambda}$.

By the strong convexity of $\phi(z)$, we have $\phi(p^a(x^k)) \geq \phi(p(x^k)) + (2\lambda)^{-1}\|p(x^k) - p^a(x^k)\|^2$. From the definitions of $F^a(x^k)$ and $p(x^k)$, we obtain $F^a(x^k) = \widetilde{f}^{p^a(x^k)} + \varepsilon_{p^a(x^k)} + (2\lambda)^{-1}\|p^a(x^k) - x^k\|^2 \geq f(p^a(x^k)) + (2\lambda)^{-1}\|p^a(x^k) - x^k\|^2 = \phi(p^a(x^k)) \geq \phi(p(x^k)) + (2\lambda)^{-1}\|p(x^k) - p^a(x^k)\|^2 = F(x^k) + (2\lambda)^{-1}\|p(x^k) - p^a(x^k)\|^2$. By (P1), (P5) holds.

By (P4) and (P5), we have the following (P6). In fact, (P6) says that the bundle subalgorithm for finding $s(x^k)$ terminates in finite steps.

(P6) If x^k does not minimize f, then we can find one solution $s(x^k)$ of (18) such that (24) holds.

3. Approximate Quasi-Newton Bundle-Type Algorithm

For presenting the algorithm, we use the following notations: $a_k = a(x^k)$, $s^k = s(x^k)$, and $m_k = m(x^k)$. Given positive numbers δ, υ, γ, and L such that $0 < \delta < 1$, $0 < \upsilon < 1$, $0 < \gamma < 1$, and one symmetric $n \times n$ positive definite matrix N.

Approximate Quasi-Newton Bundle-Type Algorithm (AQNBT Alg):

Step 1 (initialization). Let x^1 be a starting point, and let B_1 be an $n \times n$ symmetric positive definite matrix. Let ε_1 and λ be positive numbers. Choose a sequence of positive numbers $\{m_k\}_{k=1}^{\infty}$ such that $\sum_{k=1}^{\infty} m_k < \infty$. Set $k = 1$. Find $s^1 \in R^n$ and a_1 such that

$$a_1 \le m_1 \min\left\{\lambda^{-2}\left(s^1\right)^T s^1, L\right\}. \tag{28}$$

Let $G^a(x^1) = -\lambda^{-1}s^1$, $z^1 = x^1$, $j = 1$, and j be the running index of bundle subalgorithm.

Step 2 (finding a search direction). If $\|G^a(x^k)\| = 0$, stop with x^k optimal. Otherwise compute

$$d^k = -B_k^{-1}G^a\left(x^k\right). \tag{29}$$

Step 3 (line search). Starting with $u = 1$, let i_k be the smallest nonnegative integer u such that

$$F_a\left(x^k + \upsilon^u d^k\right) \le F^a\left(x^k\right) + \delta\upsilon^u\left(d^k\right)^T G^a\left(x^k\right), \tag{30}$$

where $\varepsilon_{u+1} = \gamma\varepsilon_u$ corresponds to the approximations $F_a(x^k + \upsilon^u d^k)$ and $F^a(x^k + \upsilon^u d^k)$ of F at $x^k + \upsilon^u d^k$; $F_a(x^k + \upsilon^u d^k)$ satisfies

$$\begin{aligned} &F^a\left(x^k + \upsilon^u d^k\right) - F_a\left(x^k + \upsilon^u d^k\right) \\ &\quad \le m_{k+1}\min\left\{\lambda^{-2}s\left(x^k + \upsilon^u d^k\right)^T s\left(x^k + \upsilon^u d^k\right), L\right\}, \end{aligned} \tag{31}$$

and $s(x^k + \upsilon^u d^k)$ is the solution of (19), in which x^k is replaced by $x^k + \upsilon^u d^k$, and the expression of $F^a(x^k + \upsilon^u d^k)$ is similar to (21), but x is replaced by $x^k + \upsilon^u d^k$. Set $t^k = \upsilon^{i_k}$ and $x^{k+1} = x^k + t^k d^k$.

Step 4 (computing the approximate gradient). Compute $G^a(x^{k+1}) = -\lambda^{-1}s^{k+1}$.

Step 5 (updating B_k). Let $\Delta x^k = x^{k+1} - x^k$ and $\Delta g^k = G^a(x^{k+1}) - G^a(x^k)$. Set

$$B_{k+1} = \begin{cases} N, & \text{if } \left(\Delta x^k\right)^T \Delta g^k \le 0, \\ \left(\text{symmetric, positive definite and satisfies } B_{k+1}\Delta x^k = \Delta g^k\right) & \text{otherwise.} \end{cases} \tag{32}$$

Set $k = k + 1$, and go to Step 2.

End of AQNBT algorithm.

4. Convergence Analysis

In this section we prove the global convergence of the algorithm described in Section 3, and furthermore under the assumptions of semismoothness and regularity, we show that the proposed algorithm has a Q-superlinear convergence rate. Following the proof of Theorem 3, see Mifflin et al. [7], we can show that, at each iteration k, i_k is well defined, and hence the stepsize $t^k > 0$ can be determined finitely in Step 4. We assume the proposed algorithm does not terminate in finite steps, so the sequence $\{x^k\}_{k=1}^{\infty}$ is an infinite sequence. Since the sequence $\{m_k\}_{k=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} m_k < \infty$, there exists a constant W such that $\sum_{k=1}^{\infty} m_k \le W$. Let $D_a = \{x \in R^n \mid F(x) \le F(x^1) + 2LW\}$. By making a slight change of the proof of Lemma 1, see Mifflin et al. [7], we have the following lemma.

Lemma 1. *$F(x^{k+1}) \le F(x^k) + L(m_k + m_{k+1})$ for all $k \ge 1$ and $x^k \in D_a$.*

Theorem 2. *Suppose f is bounded below and there exists a constant β such that*

$$\left\langle B_k\, d, d\right\rangle \ge \beta\|d\|^2, \quad \forall d \in R^n,\ \forall k. \tag{33}$$

Then any accumulation point of $\{x^k\}$ is an optimal solution of problem (1).

Proof. According to the first part of the proof of Theorem 3, see Mifflin et al., [7], we have $\lim_{k\to\infty} F(x^k) = F^*$. Since $m_k \to 0$, from (P1) we obtain $a_k \to 0$ as $k \to \infty$, and $\lim_{k\to\infty} F_a(x^k) = \lim_{k\to\infty} F^a(x^k) = F^*$. Thus

$$\lim_{k\to\infty} t^k\left(d^k\right)^T G^a\left(x^k\right) = 0. \tag{34}$$

Let $\overline{x}$ be an arbitrary accumulation point of $\{x^k\}$, and let $\{x^k\}_{k\in K}$ be a subsequence converging to $\overline{x}$. By (P5) we have

$$\lim_{k\in K, k\to\infty} G^a\left(x^k\right) = G\left(\overline{x}\right). \tag{35}$$

Since $\{B_k^{-1}\}$ is bounded, we may suppose

$$\lim_{k\in K, k\to\infty} d^k = \overline{d} \tag{36}$$

for some $\overline{d} \in R^n$. Moreover we have

$$\lim_{k\in K, k\to\infty} \left\langle G^a\left(x^k\right), d^k\right\rangle = \left\langle G\left(\overline{x}\right), \overline{d}\right\rangle \le -\beta\left\|\overline{d}\right\|^2. \quad (37)$$

If $\liminf_{k\to\infty} t^k > 0$, then $\overline{d} = 0$. Otherwise, if $\liminf_{k\to\infty} t^k = 0$, by taking a subsequence if necessary we may assume $t^k \to 0$ for $k \in K$. The definition of i_k in the line search rule gives

$$F_a\left(x^k + v^{i_k-1}d^k\right) > F^a\left(x^k\right) + \delta v^{i_k-1}\left(d^k\right)^T G^a\left(x^k\right), \quad (38)$$

where $v^{i_k-1} = t^k/v$. So by (P1) we obtain

$$\frac{F\left(x^k + v^{i_k-1}d^k\right) - F\left(x^k\right)}{v^{i_k-1}} > \delta\left(d^k\right)^T G^a\left(x^k\right). \quad (39)$$

By taking the limit in (39) on the subsequence $k \in K$, we have

$$\overline{d}^T G\left(\overline{x}\right) \ge \delta\overline{d}^T G\left(\overline{x}\right). \quad (40)$$

In view of (37), the last inequality also gives $\overline{d} = 0$. Since $G^a(\overline{x}) = -B_k d^k$ and B_k is bounded, it follows from $\overline{d} = 0$ that

$$\lim_{k\to\infty, k\in K} G^a\left(x^k\right) = G\left(\overline{x}\right) = 0. \quad (41)$$

Therefore, $\overline{x}$ is an optimal solution of problem (1). □

In the next part, we focus our attention on establishing Q-superlinear convergence of the proposed algorithm.

Theorem 3. *Suppose that the conditions of Theorem 2 hold and $\overline{x}$ is an optimal solution of* (1)*. Assume that G is BD-regular at $\overline{x}$. Then $\overline{x}$ is the unique optimal solution of* (1) *and the entire sequence $\{x^k\}$ converges to $\overline{x}$.*

Proof. By the convexity and BD-regularity of G at $\overline{x}$, $\overline{x}$ is the unique optimal solution of (3); for the proof, see Qi and Womersley [22]. So $\overline{x}$ is also the unique optimal solution of (1). This implies that both f and F must have compact level sets. By Lemma 1 $\{x^k\}$ has at least one accumulation point, and from Theorem 2 we know this accumulation point must be $\overline{x}$ since $\overline{x}$ is the unique solution of (1). Next following the proof of Theorem 5.1, see Fukushima and Qi [1], we can prove that the entire sequence $\{x^k\}$ converges to $\overline{x}$. □

The condition that the Lipschitz continuous gradient G of F is semismooth at the unique optimal solution of (1) is required in the next theorem. This condition is identified if f is the maximum of several affine functions or f satisfies the constant rank constraint qualification.

Theorem 4. *Suppose that the conditions of Theorem 3 hold and G is semismooth at the unique optimal solution $\overline{x}$ of* (1)*. Suppose further that*

(i) $a_k = o(\|G(x^k)\|^2)$,

(ii) $\lim_{k\to\infty} \operatorname{dist}(B_k, \partial_B G(x^k)) = 0$,

(iii) $t^k \equiv 1$, *for all large k.*

Then $\{x^k\}$ converges to $\overline{x}$ Q-superlinearly.

Proof. Firstly we have $\{x^k\}$ converges to $\overline{x}$ by Theorem 3. Then by condition (i) and (P5), we have

$$\begin{aligned}&\left\|G^a\left(x^k\right) - G\left(x^k\right)\right\| \\ &\quad = O\left(\left\|\sqrt{a_k}\right\|\right) = o\left(\left\|G\left(x^k\right)\right\|\right) = o\left(\left\|x^k - \overline{x}\right\|\right).\end{aligned} \quad (42)$$

By condition (ii), there is a $\overline{B}_k \in \partial_B G(x^k)$ such that

$$\left\|B_k - \overline{B}_k\right\| = o(1). \quad (43)$$

Since G is semismooth at $\overline{x}$, we have, according to Qi and Sun [13],

$$\left\|G\left(x^k\right) - G\left(\overline{x}\right) - \overline{B}_k\left(x^k - \overline{x}\right)\right\| = o\left(\left\|x^k - \overline{x}\right\|\right). \quad (44)$$

Notice that $\| B_k^{-1} \| = O(1)$, (42)–(44) and condition (iii), for all large k, we have

$$\begin{aligned}&\left\|x^{k+1} - \overline{x}\right\| \\ &\quad = \left\|x^k - \overline{x} - B_k^{-1}G^a\left(x^k\right)\right\| \\ &\quad = \left\|B_k^{-1}\left[G^a\left(x^k\right) - G\left(x^k\right) + G\left(x^k\right) - G\left(\overline{x}\right)\right.\right. \\ &\qquad \left.\left. - \overline{B}_k\left(x^k - \overline{x}\right) + \left(\overline{B}_k - B_k\right)\left(x^k - \overline{x}\right)\right]\right\| \\ &\quad \ge \left\|\overline{B}_k^{-1}\right\|\left[\left\|G^a\left(x^k\right) - G\left(x^k\right)\right\|\right. \\ &\qquad + \left\|G\left(x^k\right) - G\left(\overline{x}\right) - \overline{B}_k\left(x^k - \overline{x}\right)\right\| \\ &\qquad \left. + \left\|\overline{B}_k - B_k\right\|\left\|x^k - \overline{x}\right\|\right].\end{aligned} \quad (45)$$

This establishes Q-superlinear convergence of $\{x^k\}$ to $\overline{x}$. □

Condition (i) can be replaced by a more realistic condition $a_k = o(\| G(x^{k-1})\|^2)$ without impairing the convergence result since a_k is chosen before x_k is generated. For condition (ii), Fukushima and Qi [1] suggest one of possible choices of B_k, we may expect B_k to provide a reasonable approximation to an element in $\partial_B G(x^k)$, but it may be far from what we should approximate. There are some approaches to overcome this phenomenon, see Mifflin [10] and Qi and Chen [3]. For condition (iii) we can make sure that if the conditions of Theorem 4, except (iii), hold and $0 < \delta < 1/2$, then condition (iii) holds automatically.

Acknowledgment

This research was partially supported by the National Natural Science Foundation of China (Grants no. 11171049 and no. 11171138).

References

[1] M. Fukushima and L. Qi, "A globally and superlinearly convergent algorithm for nonsmooth convex minimization," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1106–1120, 1996.

[2] A. I. Rauf and M. Fukushima, "Globally convergent BFGS method for nonsmooth convex optimization," *Journal of Optimization Theory and Applications*, vol. 104, no. 3, pp. 539–558, 2000.

[3] L. Qi and X. Chen, "A preconditioning proximal Newton method for nondifferentiable convex optimization," *Mathematical Programming*, vol. 76, no. 3, pp. 411–429, 1997.

[4] Y. R. He, "Minimizing and stationary sequences of convex constrained minimization problems," *Journal of Optimization Theory and Applications*, vol. 111, no. 1, pp. 137–153, 2001.

[5] R. Mifflin and C. Sagastizábal, "A VU-proximal point algorithm for minimization," in *Numerical Optimization*, Universitext, Springer, Berlin, Germany, 2002.

[6] C. Lemaréchal, F. Oustry, and C. Sagastizábal, "The U-Lagrangian of a convex function," *Transactions of the American Mathematical Society*, vol. 352, no. 2, pp. 711–729, 2000.

[7] R. Mifflin, D. Sun, and L. Qi, "Quasi-Newton bundle-type methods for nondifferentiable convex optimization," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 583–603, 1998.

[8] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer, Berlin, Germany, 1993.

[9] X. Chen and M. Fukushima, "Proximal quasi-newton methods for nondifferentiable convex optimization," Applied Mathematics Report 95/32, School of Mathematics, The University of New South Wales, Sydney, Australia, 1995.

[10] R. Mifflin, "A quasi-second-order proximal bundle algorithm," *Mathematical Programming*, vol. 73, no. 1, pp. 51–72, 1996.

[11] K. C. Kiwiel, "Approximations in proximal bundle methods and decomposition of convex programs," *Journal of Optimization Theory and Applications*, vol. 84, no. 3, pp. 529–548, 1995.

[12] M. Hintermüller, "A proximal bundle method based on approximate subgradients," *Computational Optimization and Applications*, vol. 20, no. 3, pp. 245–266, 2001.

[13] R. Gabasov and F. M. Kirilova, *Methods of Linear Programming, Part 3, SpecialProblems*, Izdatel'Stov BGU, Minsk, Belarus, 1980 (Russian).

[14] M. V. Solodov, "On approximations with finite precision in bundle methods for nonsmooth optimization," *Journal of Optimization Theory and Applications*, vol. 119, no. 1, pp. 151–165, 2003.

[15] K. C. Kiwiel, "An algorithm for nonsmooth convex minimization with errors," *Mathematics of Computation*, vol. 45, no. 171, pp. 173–180, 1985.

[16] P. Wolfe, "A method of conjugate subgradients for minimizing nondifferentiable functions," *Mathematical Programming Study*, vol. 3, pp. 145–173, 1975.

[17] C. Lemaréchal, "An extension of davidon methods to non differentiable problems," *Mathematical Programming Study*, vol. 3, pp. 95–109, 1975.

[18] K. C. Kiwiel, *A Variable Metric Method of Centres for Nonsmooth Minimization*, International Institute for Applied Systems Analysis, Laxemnburg, Austria, 1981.

[19] K. C. Kiwiel, *Efficient algorithms for nonsmooth optimization and their applications [Ph.D. thesis]*, Department of Electronics, Technical University of Warsaw, Warsaw, Poland, 1982.

[20] R. Mifflin, "A modification and extension of Lemarechal's algorithm for nonsmooth minimization," *Mathematical Programming Study*, vol. 17, pp. 77–90, 1982.

[21] M. Fukushima, "A descent algorithm for nonsmooth convex optimization," *Mathematical Programming*, vol. 30, no. 2, pp. 163–175, 1984.

[22] L. Q. Qi and R. S. Womersley, "An SQP algorithm for extended linear-quadratic problems in stochastic programming," *Annals of Operations Research*, vol. 56, pp. 251–285, 1995.