Proximal Quasi-Newton Methods for Nondifferentiable Convex Optimization

Xiaojun Chen*
School of Mathematics
University of New South Wales
Sydney, NSW 2052, Australia
e-mail: X.Chen@unsw.edu.au

Masao Fukushima†
Graduate School of Information Science
Nara Institute of Science and Technology
Ikoma, Nara 630-01, Japan
e-mail: fuku@is.aist-nara.ac.jp

September, 1995

Abstract

This paper proposes an implementable proximal quasi-Newton method for minimizing a nondifferentiable convex function f in $\Re^n$. The method is based on Rockafellar's proximal point algorithms and a cutting-plane technique. At each step, we use an approximate proximal point $p^a(x_k)$ of x_k to define a $v_k \in \partial_{\epsilon_k} f(p^a(x_k))$ with $\epsilon_k \leq \alpha\|v_k\|$, where α is a constant. The method monitors the reduction in the value of $\|v_k\|$ to identify when a line search on f should be used. The quasi-Newton step is used to reduce the value of $\|v_k\|$. Without the differentiability of f, the method converges globally and the rate of convergence is Q-linear. Superlinear convergence is also discussed to extend the characterization result of Dennis and Moré.

Key words. nondifferentiable convex optimization, proximal point, quasi-Newton method, cutting-plane method.
AMS subject classifications. 65K05, 90C30
Abbreviated title. Proximal quasi-Newton method

*This author's work was supported by the Australian Research Council.
†This author's work was supported in part by the Scientific Research Grant-in-Aid from the Ministry of Education, Science and Culture, Japan.

1 Introduction

In this paper we consider the minimization problem

$$\min \{f(x) \mid x \in \Re^n\}, \tag{1.1}$$

where $f : \Re^n \to \Re$ is a possibly nondifferentiable convex function.

This problem can be transformed into a differentiable convex minimization problem

$$\min \{F(x) \mid x \in \Re^n\}, \tag{1.2}$$

where

$$F(x) := \min \{f(y) + \tfrac{1}{2}\mu\|y - x\|^2 \mid y \in \Re^n\}. \tag{1.3}$$

Here μ is a positive number and $\|\cdot\|$ denotes the Euclidean norm. The function F is a differentiable convex function defined on the whole space $\Re^n$ and is called the Moreau-Yosida regularization of f. Minimizing f and F are equivalent problems, in the sense that the solution sets of problems (1.1) and (1.2) coincide with each other [10]. We denote the unique minimizer in (1.3) by $p(x)$ and we call it the proximal point of x. The derivative of F is given by

$$G(x) := \nabla F(x) = \mu(x - p(x)) \in \partial f(p(x)), \tag{1.4}$$

where ∂f is the subdifferential mapping of f in the sense of convex analysis [21]. The function G is globally Lipschitz continuous with modulus μ. Recently some important properties have been studied in [13, 15, 18].

A quasi-Newton method for solving (1.2) is defined by

$$x_{k+1} = x_k - \tau_k B_k^{-1} G(x_k), \tag{1.5}$$

where τ_k is the stepsize and B_k is generated according to a quasi-Newton formula, using

$$s_k = x_{k+1} - x_k, \quad t_k = G(x_{k+1}) - G(x_k), \tag{1.6}$$

for example, the BFGS formula :

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{t_k t_k^T}{t_k^T s_k}. \tag{1.7}$$

Global convergence of the BFGS method with line search just requires Lipschitz continuity of G [16]. Superlinear convergence requires the strong differentiability of G at x^* [2, 3, 11].

The generalized Newton method with line search for (1.2) is defined by

$$x_{k+1} = x_k - \tau_k V_k^{-1} G(x_k), \tag{1.8}$$

where $V_k \in \partial G(x_k)$. The superlinear convergence of (1.8) requires the regularity and the semismoothness of G at a solution [17, 20].

Now come implementation issues: Since it is usually impossible to obtain the exact proximal point $p(x_k)$ for a given x_k, we need to approximate it by solving the minimization problem in (1.3) only approximately. Rockafellar [22] gave some approximation criteria for proximal point algorithms. Recently Bonnans, Gilbert, Lemaréchal and Sagastizábal [2] explored preliminary conceptual optimization methods combining the Moreau-Yosida regularization and quasi-Newton methods (1.5). Zhu [23] proved that the sequence of function values generated by the method in [2] converges linearly. Fukushima and Qi [9] first studied Q-superlinear convergence of optimization methods combining the Moreau-Yosida regularization and the generalized Newton method (1.8). To ensure global convergence, a step size $\tau_k = \gamma^{m_k}$ is determined by the following line search at each step of the algorithm in [9]: Find the smallest nonnegative integer $m = m_k$ satisfying

$$F^a(x_k + \gamma^m d_k, \epsilon_{k+1}) \leq F^a(x_k, \epsilon_k) + \sigma\gamma^m G^a(x_k, \epsilon_k)^T d_k + \epsilon_k, \tag{1.9}$$

where $\gamma \in (0,1), \epsilon_k > \epsilon_{k+1} \geq 0, d_k \in \Re^n, \sigma \in (0,1)$ and F^a and G^a are approximations of F and G, respectively. However, for each positive integer $m \leq m_k$, the inner optimization problem in (1.3) should be solved approximately to a required accuracy. Hence the line search (1.9) on F is not necessarily easy to implement. Furthermore their algorithm uses an approximation $\tilde{V}_k$ of a generalized Jacobian $V_k \in \partial G(x_k)$ but the choice of $\tilde{V}_k$ is not specified. Mifflin [14] presented an implementable quasi-second order proximal bundle method, in which a line search on f is used. However, convergence rate of the proximal bundle method is not be given in [14]. An important question remains to be studied: Can one have a globally and superlinearly convergent algorithm for (1.1) without using a line search on F or its approximation?

The purpose of this paper is to give an affirmative answer to this question by presenting a globally and superlinearly convergent algorithm for (1.1) with line search on f. In this algorithm, we first find an approximate proximal point $p^a(x_k)$ of x_k by utilizing a cutting-plane technique such that $v_k := \mu(x_k - p^a(x_k)) \in \partial_{\epsilon_k} f(p^a(x_k))$ for some $\epsilon_k \in (0, \alpha\|v_k\|]$, where α is a constant, and $\partial_\epsilon f(x)$ denotes for any $\epsilon \geq 0$ the *ϵ-subdifferential* of f at x, i.e.,

$$\partial_\epsilon f(x) = \{g \in \Re^n \mid f(y) \geq f(x) + g^T(y - x) - \epsilon,\ y \in \Re^n\}.$$

Then we use $G^a(x_k) := v_k$ as an approximation of $G(x_k)$ in a quasi-Newton method (1.5) and we construct B_{k+1} by using $t_k = G^a(x_{k+1}) - G^a(x_k)$ in (1.6) and (1.7). The stepsize τ_k is determined either from the value of $\|G^a(x_k)\|$ or by performing a line search on the function f. More specifically, if the value of $\|G^a(x_k)\|$ has been reduced sufficiently, we let $\tau_k = 1$; otherwise, a line search on f is used to determine $\tau_k \in (0, 1]$.

We establish global, Q-linear and Q-superlinear convergence of the proposed algorithm. Global convergence requires the boundedness of the level sets of f. Q-linear convergence requires the strong convexity of f and semismoothness of G at x^*. Q-superlinear convergence requires strong differentiability of G at x^*, which extends the characterization result of Dennis and Moré [5]. Furthermore if B_k is

a "good" approximation of $V_k \in \partial G(x_k)$ in the sense of [9], then Q-superlinear convergence only requires semismoothness of G at x^*.

2 Algorithm and global convergence

In this section we present the algorithm and show some important properties and global convergence of this algorithm. The proposed algorithm is doubly iterative. Each major iteration consists of inner iterations to solve the following minimization problem approximately by utilizing a cutting-plane technique:

$$\min_{y \in \Re^n} f(y) + \tfrac{1}{2}\mu \|y - x_k\|^2, \tag{2.1}$$

where x_k is a current iteration point given at the beginning of the major iteration k. More specifically, the inner iteration generates a sequence $\{y_j\}$ as follows: For $j = 1, 2, \ldots$, let y_j be the (unique) solution of the problem

$$\min_{y \in \Re^n} f_{k,j}(y) + \tfrac{1}{2}\mu \|y - x_k\|^2, \tag{2.2}$$

where $f_{k,j}$ is the polyhedral convex function defined by

$$f_{k,j}(y) := \max_{i=0,1,\ldots,j-1} [f(y_i) + g_i^T (y - y_i)] \tag{2.3}$$

with the pairs of n-vectors y_i and $g_i \in \partial f(y_i)$, $i = 0, 1, \ldots, j-1$, which constitute a bundle generated sequentially starting from $y_0 = x_k$ and $g_0 \in \partial f(x_k)$. Note that (2.2) is equivalent to the quadratic programming problem:

$$\begin{array}{ll} \min_{(y,w) \in \Re^{n+1}} & w + \tfrac{1}{2}\mu \|y - x_k\|^2 \\ \text{subject to} & f(y_i) + g_i^T (y - y_i) \le w, \quad i = 0, 1, \ldots, j-1. \end{array} \tag{2.4}$$

The inner iteration is terminated if one of the following two conditions is satisfied:

$$f(x_k) - f_{k,j}(y_j) < \rho, \tag{2.5}$$

$$f(y_j) \le f(x_k) - \sigma_k (f(x_k) - f_{k,j}(y_j)), \tag{2.6}$$

where $\rho > 0$ and $\sigma_k \in (0, 1)$ are parameters. Notice that the common parameter ρ is used throughout the whole (major) iterations of the algorithm, while σ_k is dependent on the particular iteration k and may be controlled in an appropriate manner.

Since the definition (2.3) of $f_{k,j}$ and the subgradient inequality imply

$$f_{k,j}(y) \le f(y) \quad \text{for all } y \in \Re^n, \tag{2.7}$$

we have the inequalities

$$\begin{aligned} f_{k,j}(y_j) &\le f_{k,j}(y_j) + \tfrac{1}{2}\mu \|y_j - x_k\|^2 \\ &= \min_{y \in \Re^n} f_{k,j}(y) + \tfrac{1}{2}\mu \|y - x_k\|^2 \\ &\le \min_{y \in \Re^n} f(y) + \tfrac{1}{2}\mu \|y - x_k\|^2 \\ &\le f(x_k) \end{aligned} \tag{2.8}$$

for all j. Therefore if (2.5) is satisfied, it follows from the inequalities (2.8) that

$$f_{k,j}(y_j) + \tfrac{1}{2}\mu\|y_j - x_k\|^2 \le \min_{y\in\Re^n} f(y) + \tfrac{1}{2}\mu\|y - x_k\|^2 + \rho,$$

which implies that, if ρ is sufficiently small, then y_j is a good approximation to the solution of problem (2.1), i.e., an approximate proximal point of x_k. However, (2.5) implies more than that. As shown in [8, Proposition 6] that, if (2.5) holds, then we have

$$f(x_k) \le f(x) + \sqrt{2\mu\rho}\|x - x_k\| + \rho \quad \text{for all } x \in \Re^n. \tag{2.9}$$

This indicates that x_k can be considered a good approximate solution of the original problem (1.1), provided that ρ is sufficiently small. Hence we may terminate the major iteration of the algorithm.

Now consider condition (2.6). It was shown in [8, Propositon 3] that, as j increases, $f(y_j)$ and $f_{k,j}(y_j)$ get closer to each other and y_j approaches the proximal point $p(x_k)$. Thus (2.6) will be satisfied when y_j becomes sufficiently close to $p(x_k)$. So, once (2.6) is satisfied, we put $p^a(x_k) := y_j$ to use it as an approximate proximal point of x_k. Note that the closer σ_k is to one, the better $p^a(x_k)$ approximates the proximal point $p(x_k)$.

Remark. Here we use the "pure" cutting-plane algorithm without any strategy for dropping useless cutting-planes. This is only for simplicity of presentation, because the argument on how to drop and aggregate the bundle is not an essential part (from a theoretical viewpoint) of our algorithm to be given shortly. Any improved cutting-plane type procedure [1, 4] can be used as well to generate $\{y_j\}$ converging to the proximal point $p(x_k)$.

Now we state the algorithm.

Algorithm 2.1 *Choose an initial point $x_0 \in \Re^n$, parameters $\rho > 0$, $0 < \sigma < 1$, $0 < c < 1$, $0 < \gamma < 1$, a sufficiently large constant $M \ge f(x_0)$, and a sequence $\{\sigma_k\}$ such that $\sigma < \sigma_k < 1$. Let $k := 0$.*

1. Solve subproblem (2.1) by the procedure (2.3)–(2.4) to obtain a point y_j satisfying either (2.5) or (2.6). If (2.5) is satisfied, then terminate. If (2.6) is satisfied, then put $p^a(x_k) := y_j$ and proceed to step 2.

2. Let

$$v_k := \mu(x_k - p^a(x_k)).$$

3. If $k = 0$, let $B_0 := I$. If $k \ge 1$, construct a symmetric positive definite matrix $B_k \in \Re^{n\times n}$ by using a quasi-Newton formula with $t_{k-1} := v_k - v_{k-1}$ and $s_{k-1} := x_k - x_{k-1}$.

4. Let $d_k := -(B_k^{-1} - \mu^{-1}I)v_k$. If $k = 0$, let $\eta_1 := \|v_0\|$ and go to step 5. For $k \ge 1$, if $\|v_k\| \le c\eta_k$ and $f(p^a(x_k) + d_k) \le M$, let $\tau_k := 1$ and $\eta_{k+1} := \|v_k\|$ and go to step 6; otherwise, let $\eta_{k+1} := \eta_k$ and go to step 5.

5. Let m_k be the smallest nonnegative integer m such that

$$f(p^a(x_k) + \gamma^m d_k) \le f(x_k) - \frac{\gamma^m\sigma}{\mu}\|v_k\|^2.$$

Set $\tau_k := \gamma^{m_k}$.

6. Set $x_{k+1} := p^a(x_k) + \tau_k d_k$, $k := k+1$ *and return to step 1.*

Note that, since $v_k = \mu(x_k - p^a(x_k))$ and $d_k = -(B_k^{-1} - \mu^{-1}I)v_k$, we have

$$\begin{aligned} x_{k+1} &= x_k - \mu^{-1}v_k - \tau_k(B_k^{-1} - \mu^{-1}I)v_k \\ &= x_k - (\tau_k B_k^{-1} + (1-\tau_k)\mu^{-1}I)v_k. \end{aligned} \tag{2.10}$$

In order to ensure that each step of the above algorithm is well-defined, we need to establish a few propositions. In what follows, we denote by f_k the polyhedral function $f_{k,j}$ for which condition (2.6) is satisfied. Thus $p^a(x_k)$ satisfies the inequality

$$f(p^a(x_k)) \leq f(x_k) - \sigma_k(f(x_k) - f_k(p^a(x_k))). \tag{2.11}$$

The next proposition shows that step 1 is always executed finitely.

Proposition 2.1 *At each iteration k of the algorithm, step 1 is executed in a finite number of steps, because procedure (2.2)–(2.3) produces y_j satisfying either (2.5) or (2.6) for some j.*

Proof: Suppose (2.5) is never satisfied for any j, i.e.,

$$f(x_k) - f_{k,j}(y_j) > \rho > 0.$$

But, as shown in [8, Proposition 3], we have

$$\lim_{j\to\infty} f(y_j) - f_{k,j}(y_j) = 0.$$

It is then easy to see that (2.6) must be satisfied for all j sufficiently large. This completes the proof. ∎

The following proposition gives the errors arising from the approximation function f_k and the approximate proximal point $p^a(x_k)$.

Proposition 2.2 *For each $k \geq 0$, let*

$$\begin{aligned} F^a(x_k) &= f(p^a(x_k)) + \tfrac{1}{2}\mu\|p^a(x_k) - x_k\|^2 \\ G^a(x_k) &= v_k, \end{aligned}$$

and

$$\epsilon_k = (1-\sigma_k)(f(x_k) - f_k(p^a(x_k))).$$

Then we have

$$F(x_k) \leq F^a(x_k) \leq F(x_k) + \epsilon_k, \tag{2.12}$$

$$\|p^a(x_k) - p(x_k)\| \leq \sqrt{\frac{2\epsilon_k}{\mu}}, \tag{2.13}$$

$$\|G^a(x_k) - G(x_k)\| \leq \sqrt{2\mu\epsilon_k}, \tag{2.14}$$

$$v_k \in \partial_{\epsilon_k} f(p^a(x_k)). \tag{2.15}$$

Furthermore, if the sequences $\{x_k\}$ *and* $\{p^a(x_k)\}$ *remain in a bounded set* D*, then*

$$\epsilon_k \leq \alpha \|v_k\|, \tag{2.16}$$

where $\alpha = (1-\sigma)L/\sigma\mu$ *and* L *is the Lipschitz constant of* f *relative to* D*.*

Proof: Obviously, $F(x_k) \leq F^a(x_k)$. By Proposition 3 in [8], we have

$$f_k(p^a(x_k)) + \tfrac{1}{2}\mu\|p^a(x_k) - x_k\|^2 \leq F(x_k).$$

Hence by (2.11),

$$\begin{aligned} F^a(x_k) &\leq f(x_k) - \sigma_k(f(x_k) - f_k(p^a(x_k))) + \tfrac{1}{2}\mu\|p^a(x_k) - x_k\|^2 \\ &= (1-\sigma_k)(f(x_k) - f_k(p^a(x_k))) + f_k(p^a(x_k)) + \tfrac{1}{2}\mu\|p^a(x_k) - x_k\|^2 \\ &\leq F(x_k) + \epsilon_k. \end{aligned}$$

Thus we obtain (2.12).

The inequalities (2.13) and (2.14) follow from Lemma 1 in [9].

Now we prove (2.15). Let $p^a(x_k) = y_{j_k}$ and $\lambda_i, i = 0, \ldots, j_k - 1$, be the corresponding multipliers of (2.4). Then by the KKT conditions for (2.4), we have

$$\sum_{i=0}^{j_k-1} \lambda_i g_i = -\mu(p^a(x_k) - x_k)$$

and

$$f_k(p^a(x_k)) = \sum_{i=0}^{j_k-1} \lambda_i \left(f(y_i) + g_i^T(p^a(x_k) - y_i)\right).$$

Since $g_i \in \partial f(y_i), \lambda_i \geq 0$ for $i = 0, \ldots, j_k - 1$, and $\sum_{i=0}^{j_k-1} \lambda_i = 1$, we have for all $z \in \Re^n$,

$$\begin{aligned} f(z) &\geq \sum_{i=0}^{j_k-1} \lambda_i (f(y_i) + g_i^T(z - y_i)) \\ &= f_k(p^a(x_k)) - \mu(p^a(x_k) - x_k)^T(z - p^a(x_k)). \end{aligned} \tag{2.17}$$

Setting $z = x_k$ in (2.17), we have

$$\tilde{\epsilon}_k := f(x_k) - f_k(p^a(x_k)) - \mu\|p^a(x_k) - x_k\|^2 \geq 0.$$

Then (2.17) can be written as

$$f(z) \geq f(x_k) - \mu(p^a(x_k) - x_k)^T(z - x_k) - \tilde{\epsilon}_k. \tag{2.18}$$

By the definition of ϵ-subdifferential of f at x_k, (2.18) implies

$$-\mu(p^a(x_k) - x_k) \in \partial_{\tilde{\epsilon}_k} f(x_k).$$

Moreover, it follows from (2.11) and (2.18) that, for all $z \in \Re^n$,

$$\begin{aligned} f(z) &\geq f(p^a(x_k)) - \mu(p^a(x_k) - x_k)^T(z - x_k) - \tilde{\epsilon}_k + \sigma_k(f(x_k) - f_k(p^a(x_k))) \\ &= f(p^a(x_k)) - \mu(p^a(x_k) - x_k)^T(z - p^a(x_k)) - (1 - \sigma_k)(f(x_k) - f_k(p^a(x_k))) \\ &= f(p^a(x_k)) - \mu(p^a(x_k) - x_k)^T(z - p^a(x_k)) - \epsilon_k. \end{aligned}$$

Hence (2.15) holds.

Finally, recall that v_k is given by $v_k = \mu(x_k - p^a(x_k))$. Then, taking into account $\sigma < \sigma_k$, (2.11) and the Lipschitz continuity of f, we have

$$\begin{aligned} \epsilon_k &\leq \frac{1 - \sigma_k}{\sigma_k}(f(x_k) - f(p^a(x_k))) \\ &\leq \frac{(1 - \sigma_k)L}{\sigma_k}\|x_k - p^a(x_k)\| \\ &\leq \frac{(1 - \sigma)L}{\sigma\mu}\|v_k\|. \end{aligned}$$

Hence we obtain (2.16). ∎

The line search criterion in step 5 is different from conventional line search criteria. In general, d_k need not be a descent direction of f at x_k. Nevertheless, the following proposition shows that at each iteration k of Algorithm 2.1, the integer m_k is well-defined and hence the stepsize τ_k can be determined finitely in step 5.

Proposition 2.3 *For every k, there exists $\bar{\tau}_k > 0$ such that*

$$f(p^a(x_k) + \tau d_k) \leq f(x_k) - \frac{\tau\sigma}{\mu}\|v_k\|^2 \tag{2.19}$$

for all $\tau \in (0, \bar{\tau}_k]$.

Proof: Since B_k is symmetric positive definite, B_k^{-1} exists and $d_k = -(B_k^{-1} - \mu^{-1}I)v_k$ is well-defined. Since

$$f(x_k) - f_k(p^a(x_k)) \geq \mu\|x_k - p^a(x_k)\|^2 = \frac{1}{\mu}\|v_k\|^2,$$

(2.11) implies

$$f(p^a(x_k)) \leq f(x_k) - \frac{\sigma_k}{\mu}\|v_k\|^2 < f(x_k) - \frac{\sigma}{\mu}\|v_k\|^2. \tag{2.20}$$

If $v_k = 0$, then $d_k = 0$ and for any $\tau > 0$, (2.19) holds. If $v_k \neq 0$, (2.20) implies that there exists $\bar{\tau}_k \in (0, 1]$ such that for any $\tau \in (0, \bar{\tau}_k]$

$$f(p^a(x_k) + \tau d_k) < f(x_k) - \frac{\sigma}{\mu}\|v_k\|^2 \leq f(x_k) - \frac{\tau\sigma}{\mu}\|v_k\|^2.$$

The proof is complete. ∎

By Proposition 2.2, if $p^a(x_k) - x_k = 0$ and $\epsilon_k = 0$, then

$$0 = v_k \in \partial f(p^a(x_k)) = \partial f(x_k).$$

This implies that x_k minimizes f. From $\epsilon_k = 0$, $f(x_k) - f_k(p^a(x_k)) = 0$, i.e., (2.5) holds and Algorithm 2.1 terminates. On the other hand, if x_k minimizes f, then there is no y_j such that (2.6) holds but (2.5) fails to hold. Furthermore, by Proposition 3 in [8], there is some y_j such that (2.5) holds and thus Algorithm 2.1 terminates at x_k.

The BFGS method is the most successful quasi-Newton method for solving convex optimization problems [6, 7] and has the property that B_{k+1} is symmetric positive definite if B_k is symmetric positive definite and $t_k^T s_k > 0$. Therefore we recommend to use the BFGS formula (1.7) in step 3 of Algorithm 2.1. Furthermore, if $\tau_k = 1$, then it follows from (2.10) and $v_k = G^a(x_k)$ that

$$x_{k+1} = x_k - B_k^{-1} G^a(x_k)$$

and

$$B_{k+1} = \begin{cases} B_k - \dfrac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \dfrac{t_k t_k^T}{t_k^T s_k}, & \text{if } s_k^T t_k > 0 \\ B_k, & \text{otherwise,} \end{cases} \tag{2.21}$$

where $t_k = G^a(x_{k+1}) - G^a(x_k)$ and $s_k = x_{k+1} - x_k$.

By Proposition 2.2, if $\tau_k = 1$ and $\epsilon_k = 0$ for all k, then Algorithm 2.1 with (2.21) reduces to the classical BFGS method

$$x_{k+1} = x_k - B_k^{-1} G(x_k)$$

for solving $G(x) = 0$.

Let us denote

$$K = \{0, 1, 2, \ldots\}$$

and

$$K_0 = \{0\} \cup \{k \in K \mid \text{step 5 is not applied at iteration } k\}.$$

Theorem 2.1 *Assume that the convex function f has a nonempty bounded set of minima and that the sequence $\{\|B_k^{-1}\|\}$ is bounded. Then Algorithm 2.1 terminates in a finite number of iterations.*

Proof: Note that the assumption that f has a bounded set of minima implies that both f and F have bounded level sets. By construction, for each k, $x_k, p^a(x_k)$ remain in the bounded set $D := \{x \mid f(x) \leq M\}$. Let $\bar{f}$ denote the minimal value of f.

Suppose that (2.5) is never satisfied, i.e.

$$f(x_k) - f_k(p^a(x_k)) > \rho \quad \text{for all } k. \tag{2.22}$$

By Proposition 2.1, Algorithm 2.1 generates two sequences $\{x_k\}$ and $\{p^a(x_k)\}$. We shall prove that any accumulation point of $\{x_k\}$ and $\{p^a(x_k)\}$ minimizes f.

First we consider the case where K_0 is infinite. Let K_0 consist of $k_0 = 0 < k_1 < k_2, \ldots$ By construction, we have

$$\|v_{k_\ell}\| \leq c\,\eta_{k_\ell} = c\,\|v_{k_{\ell-1}}\| \quad \text{for all } \ell = 1, 2, \ldots,$$

which together with $v_k = G^a(x_k)$ implies

$$\|v_{k_\ell}\| \leq c^\ell\,\|G^a(x_0)\| \quad \text{for all } \ell = 0, 1, 2, \ldots.$$

So we get

$$\lim_{\ell\to\infty} \|G^a(x_{k_\ell})\| \leq \lim_{\ell\to\infty} c^\ell \|G^a(x_0)\| = 0. \tag{2.23}$$

Let L be the Lipschitz constant of f relative to D. By (2.16) in Proposition 2.2, $\epsilon_{k_\ell} \leq \alpha\|G^a(x_{k_\ell})\|$, where $\alpha = (1-\sigma)L/\sigma\mu$. By (2.14) in Proposition 2.2, we then have

$$\begin{aligned} \|G(x_{k_\ell})\| &\leq \|G^a(x_{k_\ell})\| + \sqrt{2\mu\epsilon_{k_\ell}} \\ &\leq \|G^a(x_{k_\ell})\| + \sqrt{2\mu\alpha\|G^a(x_{k_\ell})\|}. \end{aligned} \tag{2.24}$$

Hence (2.23) implies

$$\lim_{\ell\to\infty} \|G(x_{k_\ell})\| = 0.$$

Since $x_{k_\ell}, p^a(x_{k_\ell}) \in D$ for all $\ell \geq 0$, there exists an infinite subset $K' \subset K_0$ such that $\lim_{k\to\infty, k\in K'} x_k = \bar{x}$ and $\lim_{k\to\infty, k\in K'} p^a(x_k) = \bar{p}$. Therefore from

$$\mu\|\bar{x} - \bar{p}\| = \lim_{\substack{k\to\infty \\ k\in K'}} \|G^a(x_k)\| = 0$$

and (2.24), we have $\bar{x} = \bar{p}$ and $G(\bar{x}) = 0$. Hence the accumulation point $\bar{x}$ is optimal. Now $\{\|G^a(x_{k_\ell})\|\}_{\ell=0}^\infty$ is decreasing and has a limit 0. Passing to the limit in (2.23) with $k \in K'$, we see that any accumulation point of the subsequence $\{x_k\}_{k\in K_0}$ is optimal.

Now let us show that any accumulation point of the entire sequence $\{x_k\}$ is an optimal solution. If there exists a $\bar{k}$ such that $k \in K_0$ for all $k \geq \bar{k}$, then from

$$\|G^a(x_{k+1})\| \leq c\|G^a(x_k)\|, \quad k \geq \bar{k},$$

it follows that any accumulation point of $\{x_k\}$ is optimal. On the other hand, if such $\bar{k}$ does not exist, then by construction, for any $k \notin K_0$ there exists the largest number $k_\ell \in K_0$ such that $k_\ell < k$ and

$$\bar{f} \leq f(x_{k+1}) \leq f(x_{k_\ell+1}) = f(x_{k_\ell} - B_{k_\ell}^{-1} v_{k_\ell}),$$

where the last equality follows from (2.10). Since $\{\|B_k^{-1}\|\}$ is bounded and since $\|v_{k_\ell}\| \to 0$ and $f(x_{k_\ell}) \to \bar{f}$, any accumulation point of $\{x_k\}$ is optimal.

Now we consider the case where K_0 is finite. Let $\hat{k} = \max_{k\in K_0} k$. Since the level sets of f are bounded, the sequences $\{x_k\}$ and $\{p^a(x_k)\}$ are bounded by construction. Then

$$\frac{\sigma}{\mu} \sum_{k=\hat{k}+1}^{\infty} \tau_k \|v_k\|^2 \leq f(x_{\hat{k}+1}) - \bar{f} < \infty, \tag{2.25}$$

which implies

$$\lim_{k\to\infty} \tau_k \|v_k\|^2 = 0. \tag{2.26}$$

Since $\sigma < \sigma_k$ and $\mu\|x_k - p^a(x_k)\|^2 \le f(x_k) - f_k(p^a(x_k))$, we have

$$\bar{f} \le f(p^a(x_k)) \le f(x_k) - \frac{\sigma}{\mu}\|v_k\|^2. \tag{2.27}$$

Moreover the boundedness of $\{x_k\}$ and $\{p^a(x_k)\}$ implies that there exists a subsequence $K_1 \subset K$ such that $\lim_{k\to\infty,k\in K_1} p^a(x_k) = \bar{p}$, $\lim_{k\to\infty,k\in K_1} x_k = \bar{x}$ and $\lim_{k\to\infty,k\in K_1} v_k = \bar{v}$. Since $\{\|B_k^{-1}\|\}$ is bounded, we may suppose by taking a further subsequence if necessary that $\lim_{k\to\infty,k\in K_1} d_k = \bar{d}$.

If $\liminf_{k\to\infty,k\in K_1} \tau_k > 0$, then (2.26) immediately implies $\bar{v} = 0$. On the other hand, if $\liminf_{k\to\infty,k\in K_1} \tau_k = 0$, then the definition of m_k in the line search rule gives

$$f(p^a(x_k) + \gamma^{m_k-1} d_k) > f(x_k) - \frac{\gamma^{m_k-1}\sigma}{\mu}\|v_k\|^2. \tag{2.28}$$

Since $\gamma^{m_k-1} = \tau_k/\gamma \to 0$, (2.28) implies

$$f(\bar{p}) \ge f(\bar{x}). \tag{2.29}$$

From (2.27), we have

$$f(\bar{p}) \le f(\bar{x}) - \frac{\sigma}{\mu}\|\bar{v}\|^2. \tag{2.30}$$

Combining (2.29) and (2.30), we have $\bar{v} = 0$ and $\bar{x} = \bar{p}$. Thus we get

$$\begin{aligned} \bar{\epsilon} &= \lim_{k\to\infty,k\in K_1} (1-\sigma_k)(f(x_k) - \hat{f}_k(p^a(x_k))) \\ &\le \frac{1-\sigma}{\sigma} \lim_{k\to\infty,k\in K_1} (f(x_k) - f(p^a(x_k))) = 0, \end{aligned}$$

implying that $\bar{\epsilon} = 0$. Since $v_k \in \partial_{\epsilon_k} f(p^a(x_k))$, the closedness of the ϵ-subdifferential mapping implies that $0 \in \partial f(\bar{p})$. Hence $\bar{x} = \bar{p}$ minimizes f and $f(\bar{x}) = \bar{f}$. Since

$$\bar{f} \le f(x_{k+1}) \le f(x_k) - \frac{\tau_k\sigma}{\mu}\|v_k\|^2, \quad \text{for} \quad k > \hat{k}, \tag{2.31}$$

the sequence $\{f(x_k)\}_{k>\hat{k}}$ is non-increasing and has a limit f^*. Passing to the limit in (2.31) with $k \in K_1$, we obtain $f^* = \bar{f}$. Hence any accumulation point of $\{x_k\}$ is also optimal.

Consequently, if (2.22) holds for all k, then any accumulation point of $\{x_k\}$ and $\{p^a(x_k)\}$ minimizes f and

$$\begin{aligned} f(p^a(x_k)) - f(x_k) &\le -\sigma_k(f(x_k) - f_k(p^a(x_k))) \\ &\le -\sigma(f(x_k) - f_k(p^a(x_k))), \quad \text{for all } k. \end{aligned}$$

However, this contradicts (2.22), because $f(p^a(x_k)) - f(x_k) \to 0$. Hence Algorithm 2.1 terminates finitely.

∎

3 Linear and superlinear convergence

In this section, we consider the convergence rate of Algorithm 2.1. For this purpose, we assume that x_k is not an optimal solution for any k and we ignore condition (2.5) in the algorithm, so that the algorithm generates infinite sequences $\{x_k\}$ and $\{p^a(x_k)\}$. Then by the proof of Theorem 2.1, we can deduce that any accumulation point of the sequences $\{x_k\}$ and $\{p^a(x_k)\}$ minimizes f.

We say that f is strongly convex with modulus $\nu > 0$, if

$$\lambda f(x) + (1-\lambda)f(y) - f(\lambda x + (1-\lambda)y) \geq \tfrac{1}{2}\nu\lambda(1-\lambda)\|x-y\|^2$$

for any $\lambda \in (0,1)$ and $x, y \in \Re^n$.

Lemaréchal and Sagastizábal [13] proved that strong convexity is transmitted between f and F. More specifically, if f is strongly convex with modulus $\nu > 0$, then F is also strongly convex with modulus $\tilde{\nu} \geq \mu\nu/(\mu+\nu)$. In the remainder of this section, we shall assume that f is strongly convex in $\Re^n$. It should be noted, however, that the arguments will remain valid, even if the assumption is weakened to the local strong convexity of f around the solution x^*. (For the local strong convexity, see [18].)

Lemma 3.1 *Suppose that f is strongly convex. Then there exists a unique solution x^* of $G(x) = 0$.*

Proof: Since f is strongly convex, f always attains its minimum and the set of minima is a singleton $\{x^*\}$. Since the sets of minima of f and F coincide with each other and F is convex and differentiable, we obtain the desired result. ∎

Lemma 3.2 *Suppose that the assumptions of Theorem 2.1 hold. Let x^* be an optimal solution of (1.1). For each $k \geq 0$, if $\tau_k = 1$, then*

$$\begin{aligned} \|x_{k+1} - x^*\| \;\leq\; & \|(I - B_k^{-1}V_k)(x_k - x^*)\| + \|B_k^{-1}\|\|G^a(x_k) - G(x_k)\| \\ & + \|B_k^{-1}\|\|G(x_k) - G(x^*) - V_k(x_k - x^*)\|, \end{aligned} \tag{3.1}$$

where $V_k \in \partial G(x_k)$.

Proof: If $\tau_k = 1$, then it follows from (2.10), $v_k = G^a(x_k)$ and $G(x^*) = 0$ that

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

which implies (3.1). ∎

Now we are ready to establish Q-linear convergence of the algorithm.

Theorem 3.1 *Suppose that f is strongly convex with modulus $\nu > 0$. Assume that G is semismooth at the minimizer x^* of (1.1). Assume that there exist $\tilde{k} \geq 0$, $\delta \geq 0, \beta \geq 0, \kappa > 0$, such that for $k \geq \tilde{k}$, it holds that*

$$\epsilon_k \leq \frac{1}{2\mu}\delta^2 \min\{\|G(x_k)\|^2, \|G^a(x_k)\|^2\}, \tag{3.2}$$

$\|B_k^{-1}\| \leq \kappa$ and $\|I - B_k^{-1}V_k\| \leq \beta$ for some $V_k \in \partial G(x_k)$. If

$$(1+\delta)^2(\beta + \kappa\mu\delta)(\nu + \mu) < \nu c, \tag{3.3}$$

where c is the parameter used in Algorithm 2.1, then there exists an integer $\hat{k} \geq \tilde{k}$, such that for all $k \geq \hat{k}$, we have $\tau_k \equiv 1$ and

$$\|x_{k+1} - x^*\| \leq c\|x_k - x^*\|, \tag{3.4}$$

i.e., the sequence $\{x_k\}$ converges to x^* Q-linearly.

Proof: Since the strong monotonicity of f implies that the level sets of f are compact, the sequence $\{x_k\}$ has at least one accumulation point. Moreover by Lemma 3.1 and the proof of Theorem 2.1, f has the unique minimum x^* and $\{x_k\}$ converges to x^*.

Since G is semismooth at x^*, by Theorem 2.3 in [20], we have

$$\|G(x^* + h) - G(x^*) - Vh\| = o(\|h\|), \tag{3.5}$$

where $V \in \partial G(x^* + h)$. Therefore there exists an open ball $S(x^*, r)$ with the center x^* and radius r such that for any $x \in S(x^*, r)$,

$$\kappa\|G(x) - G(x^*) - V(x - x^*)\| \leq \left(\frac{c\nu}{(1+\delta)^2(\nu+\mu)} - \beta - \kappa\mu\delta\right)\|x - x^*\|, \tag{3.6}$$

where $V \in \partial G(x)$.

Since the sequences $\{p^a(x_k)\}$ and $\{x_k\}$ converge to x^*, there exists a $\bar{k} \geq \tilde{k}$ such that $x_k, p^a(x_k) \in S(x^*, r)$ and $f(p^a(x_k) + d_k) \leq M$ for all $k \geq \bar{k}$, where M is the parameter used in Algorithm 2.1. By Theorem 2.2 in [13], the strong convexity of f with modulus ν implies that

$$(G(x) - G(y))^T(x - y) \geq \tilde{\nu}\|x - y\|^2, \text{ for } x, y \in \Re^n,$$

where

$$\tilde{\nu} \geq \frac{\mu\nu}{\mu + \nu}.$$

It then follows that

$$\tilde{\nu}\|x_k - x^*\|^2 \leq G(x_k)^T(x_k - x^*) \leq \|G(x_k)\|\|x_k - x^*\|,$$

and hence

$$\|x_k - x^*\| \leq \tilde{\nu}^{-1}\|G(x_k)\| \leq \frac{\mu + \nu}{\mu\nu}\|G(x_k)\|. \tag{3.7}$$

By Proposition 2.2 and (3.7), the condition (3.2) implies

$$\begin{aligned}\|x_k - x^*\| &\leq \frac{\mu+\nu}{\mu\nu}(\|G^a(x_k)\| + \sqrt{2\mu\epsilon_k}) \\ &\leq \frac{\mu+\nu}{\mu\nu}(\|G^a(x_k)\| + \delta\|G^a(x_k)\|) \\ &\leq \frac{\mu+\nu}{\nu\mu}(1+\delta)\|G^a(x_k)\|. \end{aligned} \tag{3.8}$$

Since $\|G^a(x_k)\| \to 0$, there exists $\hat{k} \geq \bar{k}$ such that $\|v_{\hat{k}}\| \leq c\eta_k$, i.e., $\hat{k} \in K_0$. Then $\tau_{\hat{k}} = 1$ and by (2.10),

$$x_{\hat{k}+1} = x_{\hat{k}} - B_{\hat{k}}^{-1} G^a(x_{\hat{k}}).$$

By Lemma 3.2, (3.6) and (3.8), we have

$$\begin{aligned}\|G^a(x_{\hat{k}+1})\| &\leq \|G^a(x_{\hat{k}+1}) - G(x_{\hat{k}+1})\| + \|G(x_{\hat{k}+1}) - G(x^*)\| \\ &\leq \sqrt{2\mu\epsilon_{\hat{k}+1}} + \mu\|x_{\hat{k}+1} - x^*\| \\ &\leq \delta\|G(x_{\hat{k}+1})\| + \mu\|x_{\hat{k}+1} - x^*\| \\ &\leq (1+\delta)\mu\|x_{\hat{k}+1} - x^*\| \\ &\leq (1+\delta)\mu(\beta\|x_{\hat{k}} - x^*\| + \kappa\|G(x_{\hat{k}}) - G(x^*) - V_{\hat{k}}(x_{\hat{k}} - x^*)\| \\ &\quad +\kappa\|G^a(x_{\hat{k}}) - G(x_{\hat{k}})\|) \\ &\leq (1+\delta)\mu(\beta\|x_{\hat{k}} - x^*\| + \kappa\|G(x_{\hat{k}}) - G(x^*) - V_{\hat{k}}(x_{\hat{k}} - x^*)\| \\ &\quad +\kappa\delta\|G(x_{\hat{k}})\|) \\ &\leq \frac{c\nu\mu}{(1+\delta)(\nu+\mu)}\|x_{\hat{k}} - x^*\| \\ &\leq c\|G^a(x_{\hat{k}})\|, \end{aligned}$$

where the third inequality follows from (3.2), the fourth inequality follows from the Lipschitz continuity of G and the last inequality follows from (3.8).

Hence $\hat{k}+1 \in K_0$ and $\tau_{\hat{k}+1} = 1$. This implies that for all $k \geq \hat{k}$, we have $k \in K_0$ and $\tau_k \equiv 1$. Furthermore for all $k \geq \hat{k}$,

$$\|x_{k+1} - x^*\| \leq \frac{c\nu}{(1+\delta)^2(\nu+\mu)}\|x_k - x^*\| \leq c\|x_k - x^*\|,$$

where the first inequality follows from (3.1) and (3.6).

∎

Now we proceed to establish superlinear convergence results for the algorithm. Fukushima and Qi [9] used the following two conditions in their superlinear convergence theorems.

(i) $\epsilon_k = o(\|G(x_k)\|^2)$,

(ii) $\lim_{k\to\infty} \mathrm{dist}(B_k, \partial G(x_k)) = 0$.

From Theorem 3.1, we can show that Algorithm 2.1 also converges to x^* superlinearly if the above two conditions hold.

Theorem 3.2 *Suppose that the conditions of Theorem 2.1 hold and f is strongly convex with modulus $\nu > 0$. Assume that G is semismooth at the minimizer x^* of f. If $\epsilon_k = o(\|G(x_k)\|^2)$ and $\lim_{k\to\infty} dist(B_k, \partial G(x_k)) = 0$, then there exists an integer $\hat{k} \geq 0$, such that $\tau_k \equiv 1$ for all $k \geq \hat{k}$. Furthermore the sequence $\{x_k\}$ converges to x^* Q-superlinearly.*

Proof: Note that

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

where $V_k \in \partial G(x_k)$. By assumption, we have $\|B_k^{-1}\| \leq \kappa$ and $\|I - B_k^{-1}V_k\| \to 0$, as $k \to \infty$. By (2.14) and $\epsilon_k = o(\|G(x_k)\|^2)$, we have

$$\begin{aligned}\|G^a(x_k) - G(x_k)\| \quad &\leq \quad \sqrt{2\mu\epsilon_k} \\ &= \quad o(\|G(x_k)\|) \\ &= \quad o(\|G(x_k) - G(x^*)\|) \\ &= \quad o(\|x_k - x^*\|),\end{aligned}$$

where the last equality follows from the Lipschitz continuity of G. By (3.5), we have

$$\|G(x_k) - G(x^*) - V_k(x_k - x^*)\| = o(\|x_k - x^*\|).$$

Hence we have $\|x_{k+1} - x^*\| = o(\|x_k - x^*\|)$.

∎

Now we consider the generalization of the well-known result of Dennis and Moré [5], which characterizes the Q-superlinear convergence of quasi-Newton methods for the system of differentiable equations.

We say that G is strongly differentiable at x^* if

$$\lim_{\substack{y\to x \\ z\to x}} \frac{\|G(z) - G(y) - G'(x)(z - y)\|}{\|z - y\|} = 0.$$

Theorem 3.3 *Suppose that the conditions of Theorem 2.1 hold and f is strongly convex with modulus $\nu > 0$. Assume that the sequence $\{\|B_k\|\}$ is bounded and that G is strongly differentiable at the minimizer x^* of f. If $\epsilon_k = o(\|G^a(x_k)\|^2)$, then the sequence $\{x_k\}$ converges Q-superlinearly to x^* if and only if*

$$\lim_{k\to\infty} \frac{\|(B_k - G'(x^*))s_k\|}{\|s_k\|} = 0. \tag{3.9}$$

Proof: Noting that

$$\left| \frac{\|G(x_k)\|}{\|G^a(x_k)\|} - 1 \right| \leq \frac{\sqrt{2\mu\epsilon_k}}{\|G^a(x_k)\|},$$

we see that the condition $\epsilon_k = o(\|G^a(x_k)\|^2)$ implies that $\epsilon_k = o(\|G(x_k)\|^2)$. Since $\|x_{k+1}-x_k\| \geq \|G^a(x_k)\|/\|B_k\|$, the boundedness of $\|B_k\|$ implies that $\epsilon_k = o(\|x_{k+1}-x_k\|^2)$. Thus

$$\|G^a(x_k) - G(x_k)\| \leq \sqrt{2\mu\epsilon_k} = o(\|x_{k+1} - x_k\|). \tag{3.10}$$

Assume first that (3.9) holds. By construction,

$$\begin{aligned}(B_k - G'(x^*))s_k &= B_k(x_{k+1} - x_k) - G'(x^*)(x_{k+1} - x_k) \\ &= G(x_{k+1}) - G(x_k) - G'(x^*)(x_{k+1} - x_k) \\ &\quad -G(x_{k+1}) + G(x_k) - G^a(x_k). \end{aligned} \tag{3.11}$$

The strong differentiability of G at x^* implies that

$$\|G(x_{k+1}) - G(x_k) - G'(x^*)(x_{k+1} - x_k)\| = o(\|x_{k+1} - x_k\|). \tag{3.12}$$

Hence (3.9), (3.10) and (3.11) imply

$$\lim_{k\to\infty} \frac{\|G(x_{k+1})\|}{\|x_{k+1} - x_k\|} = 0. \tag{3.13}$$

By Theorem 2.2 in [13], the strong convexity of f with modulus ν implies that F is strongly convex with modulus $\tilde{\nu} \geq \mu\nu/(\mu + \nu)$, which in turn implies that G is strongly monotone with modulus $\tilde{\nu}$. Thus

$$\|G(x_k) - G(x^*)\| \geq \tilde{\nu}\|x_k - x^*\|, \quad \text{for all large } k.$$

Hence for all large k,

$$\begin{aligned}\frac{\|G(x_{k+1})\|}{\|x_{k+1} - x_k\|} &= \frac{\|G(x_{k+1}) - G(x^*)\|}{\|x_{k+1} - x_k\|} \\ &\geq \frac{\tilde{\nu}\|x_{k+1} - x^*\|}{\|x_{k+1} - x^*\| + \|x_k - x^*\|} \\ &= \tilde{\nu}\frac{\|x_{k+1} - x^*\|/\|x_k - x^*\|}{1 + \|x_{k+1} - x^*\|/\|x_k - x^*\|}.\end{aligned}$$

It then follows from (3.13) that

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Conversely, assume that $\{x_k\}$ converges Q-superlinearly to x^*. By Lemma 2.1 in [5], $\lim_{k\to\infty} \|x_k - x^*\|/\|x_{k+1} - x_k\| = 1$. Since $\|G(x_{k+1}) - G(x^*)\| \leq \mu\|x_{k+1} - x^*\|$,

$$\frac{\|G(x_{k+1})\|}{\|x_{k+1} - x_k\|} = \frac{\|G(x_{k+1}) - G(x^*)\|}{\|x_k - x^*\|} \cdot \frac{\|x_k - x^*\|}{\|x_{k+1} - x_k\|} \to 0, \quad \text{as} \quad k \to \infty.$$

From (3.10), (3.11) and (3.12), condition (3.9) holds.

∎

Corollary 3.1 *Suppose that the conditions of Theorem 2.1 hold and that f is strongly convex with modulus $\nu > 0$. Assume that the sequence $\{\|B_k\|\}$ is bounded and that G is strongly differentiable at the minimizer x^* of f. If $\epsilon_k = o(\|G^a(x_k)\|^2)$ and $\|I - B_k^{-1}G'(x^*)\| < \nu c/(\nu+\mu)$ then the sequence $\{x_k\}$ converges Q-superlinearly to x^* if and only if condition (3.9) holds.*

Proof: Condition $\epsilon_k = o(\|G^a(x_k)\|^2)$ implies that there exist a $\tilde{k}$ and a small $\delta > 0$ such that (3.2) and (3.3) hold for $k \geq \tilde{k}$. By Theorem 3.1, $\tau_k \equiv 1$ for $k \geq \tilde{k}$. Then the conclusion follows from Theorem 3.3. ∎

Theorem 2 in [5] requires the continuous differentiability of G at x^* and $G(x_k) = G^a(x_k)$ for all k. Theorem 3.3 relaxes the two conditions.

4 Concluding remarks

In this paper, we have proposed an algorithm for solving nondifferentiable convex optimization problems. The proposed algorithm is related to the recent algorithms developed in [2, 12, 14, 9, 3]. In [2], a proximal quasi-Newton method is proposed and its convergence properties are discussed. However, it is a conceptual algorithm and its practical implementation is not given precisely. In [12, 14], implementable proximal point algorithms have been presented, but their rate of convergence is not discussed. In [9], an implementable algorithm for minimizing the Moreau-Yosida regularization F is proposed and its superlinear convergence is established. However, this algorithm makes use of the generalized Jacobian of F, instead of matrices B_k generated by a quasi-Newton formula. Moreover, the line search is performed on the function F, rather than f, which is usually quite expensive. In [3], a modification of the algorithm of [9] is presented and its local convergence properties are studied. The algorithm proposed in this paper uses a quasi-Newton update of B_k and line search is done on the function f. The resulting algorithm is globally and superlinearly convergent under suitable conditions, and it is implementable. These features altogether distinguish the proposed algorithm from those in [2, 12, 14, 9, 3].

Acknowledgment

The authors are grateful to R. Mifflin, L. Qi and R. Womersley for their helpful comments.

References

[1] A. Auslender, "Numerical methods for nondifferentiable convex optimization", *Mathematical Programming Study* 30 (1987) 102-126.

[2] J. Bonnans, J. Gilbert, C. Lemaréchal and C. Sagastizábal, "A family of variable metric proximal methods", *Mathematical Programming* 68 (1995) 15-47.

[3] X. Chen, "Convergence of the BFGS method for LC^1 convex constrained optimization", Applied Mathematics Report 94/26, Department of Applied Mathematics, The University of New South Wales (Sydney, 1994).

[4] R. Correa and C. Lemaréchal, "Convergence of some algorithms for convex minimization", *Mathematical Programming* 62 (1993) 261-275.

[5] J. E. Dennis, Jr. and J. J. Moré, "A characterization of superlinear convergence and its applications to quasi-Newton methods", *Mathematics of Computation* 28 (1974) 549-560.

[6] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Prentice-Hall, Inc., 1983).

[7] R. Fletcher, *Practical Methods of Optimization*, Second Edition, (John Wiley, Chichester and New York, 1987).

[8] M. Fukushima, "A descent algorithm for nonsmooth convex optimization", *Mathematical Programming* 30 (1984) 163-175.

[9] M. Fukushima and L. Qi, "A globally and superlinearly convergent algorithm for nonsmooth convex minimization", to appear in *SIAM J. Optimization*.

[10] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* (Springer-Verlag, Berlin-Heidelberg, 1993).

[11] C. M. Ip and J. Kyparisis, "Local convergence of quasi-Newton methods for B-differentiable equations", *Mathematical Programming* 56 (1992) 71-89.

[12] C. Lemaréchal and C. Sagastizábal, "An approach to variable metric bundle methods", Proceedings of the 16th IFIP Conference on System Modelling and Optimization (Compiegne, 1993).

[13] C. Lemaréchal and C. Sagastizábal, "Practical aspects of the Moreau-Yosida regularization: Theoretical preliminaries", Research Report No. 2250 (Revised Version), INRIA (Le Chesnay, 1995).

[14] R. Mifflin, "A quasi-second-order proximal bundle algorithm", Preprint (Revised version), Department of Pure and Applied Mathematics, Washington State University (Pullman, WA, 1995).

[15] R. A. Poliquin and R. T. Rockafellar, "Generalized Hessian properties of regularized nonsmooth functions", to appear in *SIAM Journal on Optimization*.

[16] M. J. D. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches", in *Nonlinear Programming*, SIAM-AMS Proceedings, Vol. IX, R. W. Cottle and C. E. Lemke, eds., American Mathematical Society (Providence, RI, 1976).

[17] L. Qi, "Convergence analysis of some algorithms for solving nonsmooth equations", *Mathematics of Operations Research* 18 (1993) 227-244.

[18] L. Qi, "Second-Order analysis of the Moreau-Yosida approximation of a convex function", Applied Mathematics Report, Department of Applied Mathematics, The University of New South Wales (Sydney, 1994).

[19] L. Qi and X. Chen, "A preconditioning proximal Newton method for nondifferentiable convex optimization", Applied Mathematics Report 95/20, Department of Applied Mathematics, The University of New South Wales (Sydney, 1995).

[20] L. Qi and J. Sun, "A nonsmooth version of Newton's method", *Mathematical Programming* 58 (1993) 353-368.

[21] R. T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970).

[22] R. T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming", *Mathematics of Operations Research* 1 (1976) 97-116.

[23] C. Zhu, "Asymptotic convergence analysis of some inexact proximal point algorithms for minimization", preprint, Mathematics and Computer Science Division, Argonne National Laboratory (Argonne, 1994).