A Globally and Superlinearly Convergent Algorithm for Nonsmooth Convex Minimization

Masao Fukushima [1]
Graduate School of Information Science
Nara Institute of Science and Technology
Ikoma, Nara 630-01, Japan
E-mail: fuku@is.aist-nara.ac.jp

and

Liqun Qi [2]
School of Mathematics
The University of New South Wales
Sydney, New South Wales 2052, Australia
E-mail: L.Qi@unsw.edu.au

December 15, 1994

Abstract

It is well-known that a possibly nondifferentiable convex minimization problem can be transformed into a differentiable convex minimization problem by way of the Moreau-Yosida regularization. This paper presents a globally convergent algorithm that is designed to solve the latter problem. Under additional semismoothness and regularity assumptions, the proposed algorithm is shown to have a Q-superlinear rate of convergence.

[1] This author's work was supported in part by the Scientific Research Grant-in-Aid from the Ministry of Education, Science and Culture, Japan.

[2] This author's work was supported in part by the Australian Research Council.

1 Introduction

Let $f : \Re^n \to \Re$ be a possibly nondifferentiable convex function and consider the problem

$$\text{minimize} \quad f(x). \tag{1}$$

Associated with this problem is the problem

$$\text{minimize} \quad F(x), \tag{2}$$

where $F : \Re^n \to \Re$ is the so-called Moreau-Yosida regularization of f, which is defined by

$$F(x) = \min_{z \in \Re^n} \left\{ f(z) + \frac{1}{2\lambda} \|z - x\|^2 \right\}, \tag{3}$$

where λ is a positive parameter and $\|\cdot\|$ denotes the Euclidean norm. Throughout the paper we use a fixed parameter λ and we shall not make explicit the dependence of various functions and vectors on λ.

It is well-known that problems (1) and (2) are equivalent in the sense that the solution sets of the two problems coincide with each other. In fact, the classical proximal point algorithm [7, 15] is regarded as a gradient method for solving problem (2). A remarkable feature of problem (2) is that the objective function F is a differentiable convex function, even though the function f is nondifferentiable. Moreover F has a Lipschitz continuous gradient. Rockafellar showed that F is twice differentiable at a point if and only if f is twice differentiable in the general sense at a corresponding point relative to a subgradient at that corresponding point [16]. In general, F is not twice differentiable at some kinks and boundary points [16, 6]. However, it is shown recently that under some reasonable conditions, the gradient function of F is semismooth [12]. This paves a way to construct Q-superlinearly convergent algorithms for solving nonsmooth convex minimization problems.

The differentiability of F motivates us to develop a rapidly convergent algorithm for solving (2), particularly when f is nondifferentiable. In practice, however, it is difficult or even impossible to evaluate the functional value $F(x)$ and the gradient $\nabla F(x)$ exactly at any given point x. The purpose of this paper is to present an implementable algorithm that makes use of approximate function and gradient values together with an approximate generalized Hessian of F. The proposed algorithm is therefore considered a globalized approximate Newton method designed to solve problem (2). We prove global convergence and, under additional semismoothness and regularity assumptions, Q-superlinear convergence of the proposed algorithm.

2 Preliminaries

In this section, we review some basic results in convex analysis and nonsmooth analysis.

Since the minimand on the right-hand side of (3) is strongly convex, the minimum is attained uniquely for each $x \in \Re^n$. We denote the unique minimizer by $p(x)$. The results stated in the following propositions are fundamental and useful in the subsequent discussions.

Proposition 1 *The function F is finite-valued, convex and everywhere differentiable with gradient*

$$\nabla F(x) = \frac{1}{\lambda}(x - p(x)).$$

Moreover the gradient mapping $\nabla F : \Re^n \to \Re^n$ is co-coercive with modulus $1/\lambda$,i.e.,

$$\|\nabla F(x) - \nabla F(x')\|^2 \leq \frac{1}{\lambda}\langle \nabla F(x) - \nabla F(x'), x - x' \rangle \qquad \textit{for all } x, x' \in \Re^n,$$

and hence Lipschitz continuous with modulus $1/\lambda$, i.e.,

$$\|\nabla F(x) - \nabla F(x')\| \leq \frac{1}{\lambda}\|x - x'\| \qquad \textit{for all } x, x' \in \Re^n.$$

Proof. See [5, Theorem XV.4.1.4]. □

Proposition 2 *The following statements are equivalent:*
(i) x minimizes f;
(ii) $x = p(x)$;
(iii) $\nabla F(x) = 0$;
(iv) x minimizes F;
(v) $f(x) = f(p(x))$;
(vi) $f(x) = F(x)$.

Proof. See [5, Theorem XV.4.1.7]. □

In what follows, we denote the gradient mapping ∇F by g. By Proposition 1, the mapping $g : \Re^n \to \Re^n$ is globally Lipschitz continuous. Hence, by Rademacher's theorem, g is differentiable almost everywhere and the set

$$\partial_B g(x) = \{A \in \Re^{n \times n} \mid A = \lim_{x^k \to x} \nabla g(x^k), \quad x^k \in \Omega_g\}, \tag{4}$$

where $\Omega_g = \{x \in \Re^n \mid g \text{ is differentiable at } x\}$, is nonempty and bounded for each x. Notice that for a Lipschitz continuous function, there is no difference between Fréchet differentiability and Gâteaux differentiability. We say that g is BD-regular at x if all matrices $A \in \partial_B g(x)$ are nonsingular. Since g is a gradient mapping of the convex function F, its Jacobian $\nabla g(x)$ is symmetric and positive semidefinite for any $x \in \Omega_g$. The next proposition is a consequence of these observations.

Proposition 3 *For each $x \in \Re^n$, every $V \in \partial_B g(x)$ is a symmetric positive semidefinite matrix. Moreover, if g is BD-regular at x, then there exists a constant $c > 0$ and a neighborhood $\mathcal{N}$ of x such that for all $y \in \mathcal{N}$*

$$\langle d, Vd \rangle \geq c\|d\|^2 \qquad \textit{for all} \quad d \in \Re^n \quad \textit{and} \quad V \in \partial_B g(x).$$

□

3 Algorithm

Since F is a differentiable convex function with Lipschitz continuous gradient, it is tempting to apply a suitable gradient-based algorithm to solve problem (2). However, by (3) and Proposition 1, $F(x)$ and $\nabla F(x)$ are obtained through the optimal solution of another minimization problem

$$\text{minimize}_z \quad f(z) + \frac{1}{2\lambda}\|z - x\|^2, \tag{5}$$

which is difficult or even impossible to solve exactly. In practice, we may only expect to compute an approximation to the optimal solution $p(x)$ of problem (5). Thus we cannot apply an existing gradient-based method directly to problem (2), but need to develop an implementable algorithm that takes into account the inevitable approximate evaluation of the function value and gradient of F.

In the following, we suppose that, for each $x \in \Re^n$, we can evaluate $F(x)$ and $g(x)$ $(= \nabla F(x))$ only approximately but with any desired accuracy. To be more specific, we suppose that, for each $x \in \Re^n$ and any $\varepsilon > 0$, we can find a vector $p^a(x, \varepsilon) \in \Re^n$ such that

$$f(p^a(x,\varepsilon)) + \frac{1}{2\lambda}\|p^a(x,\varepsilon) - x\|^2 \leq F(x) + \varepsilon. \tag{6}$$

Implementable algorithms that are designed to find such $p^a(x, \varepsilon)$ for a nondifferentiable convex function f may be found in [4, 1, 2]. With vector $p^a(x, \varepsilon)$, define approximations to $F(x)$ and $\nabla F(x)$ by

$$F^a(x,\varepsilon) = f(p^a(x,\varepsilon)) + \frac{1}{2\lambda}\|p^a(x,\varepsilon) - x\|^2 \tag{7}$$

and

$$g^a(x,\varepsilon) = \frac{1}{\lambda}(x - p^a(x,\varepsilon)), \tag{8}$$

respectively.

Lemma 1 *Let $p^a(x, \varepsilon)$ be a vector satisfying (6), and $F^a(x, \varepsilon)$ and $g^a(x, \varepsilon)$ be given by (7) and (8), respectively. Then we have*

$$F(x) \leq F^a(x,\varepsilon) \leq F(x) + \varepsilon, \tag{9}$$

$$\|p^a(x,\varepsilon) - p(x)\| \le \sqrt{2\lambda\varepsilon} \tag{10}$$

and

$$\|g^a(x,\varepsilon) - g(x)\| \le \sqrt{\frac{2\varepsilon}{\lambda}}. \tag{11}$$

Proof. First, (9) is obvious from (6) and (7). Now define the function $\theta : \Re^n \to \Re$ by

$$\theta(z) = f(z) + \frac{1}{2\lambda}\|z - x\|^2.$$

It is easy to see that the function θ is strongly convex with modulus $1/\lambda$. Then, by [15, Proposition 6], we have the inequality

$$\theta(y) \ge \theta(z) + \langle y - z, w\rangle + \frac{1}{2\lambda}\|y - z\|^2 \qquad \text{for all } y, z \in \Re^n, \tag{12}$$

whenever $w \in \partial\theta(z)$. Let us put $y = p^a(x,\varepsilon)$ and $z = p(x)$ in (12). We have $\theta(y) = \theta(p^a(x,\varepsilon)) = F^a(x,\varepsilon)$ and $\theta(z) = \theta(p(x)) = F(x)$. Moreover, since $p(x)$ minimizes θ, $w = 0$ belongs to $\partial\theta(p(x))$. So it follows from (12) that

$$F^a(x,\varepsilon) \ge F(x) + \frac{1}{2\lambda}\|p^a(x,\varepsilon) - p(x)\|^2,$$

which together with the right-hand inequality in (9) yields (10). Finally, by Proposition 1 and (8), we have

$$\|g^a(x,\varepsilon) - g(x)\| = \frac{1}{\lambda}\|p^a(x,\varepsilon) - p(x)\|,$$

which together with (10) gives (11). □

The above lemma says that we can compute approximations $F^a(x,\varepsilon)$ and $g^a(x,\varepsilon)$, which may be made arbitrarily close to $F(x)$ and $g(x)$, respectively, by choosing parameter ε small enough. In order to design a rapidly convergent algorithm, we have to make use of a "second-order" approximation to F at each x. Recently much effort has been made to extend Newton method to semismooth equations [8, 13, 10] and SC^1 minimization problems [11, 9, 3]. The basic assumption made in those methods is that, in the context of the problem under consideration, a matrix $V \in \partial_B g(x)$ is available at every current point x. Unfortunately, like the function value $F(x)$ and gradient $g(x)$, an exact generalized Hessian $V \in \partial_B g(x)$ is in general not computable at the current point x. Again, we have to compute it approximately. Note that, when f is actually twice continuously differentiable at $p(x)$, then $\partial_B g(x)$ consists of the single element, which is precisely

$$\nabla^2 F(x) = \frac{1}{\lambda} I - \frac{1}{\lambda}\left[I + \lambda\nabla^2 f(p(x))\right]^{-1}. \tag{13}$$

Another important case is that the conjugate function f^* is twice continuously differentiable at $g(x)$. In this case, $\nabla^2 F(x)$ also exists and is given by

$$\nabla^2 F(x) = \left[\lambda I + \nabla^2 f^*(g(x))\right]^{-1}. \tag{14}$$

(See [16, 6, 12] for more details and further results on the second-order differentiability of F.) Since $p^a(x,\varepsilon)$ is available as an approximation to the uncomputable vector $p(x)$, it would be natural to use the matrix

$$V = \frac{1}{\lambda} I - \frac{1}{\lambda}\left[I + \lambda \nabla^2 f(p^a(x,\varepsilon))\right]^{-1}, \tag{15}$$

if f is twice differentiable at $p^a(x,\varepsilon)$, and possibly

$$V = \left[\lambda I + \nabla^2 f^*(g^a(x,\varepsilon))\right]^{-1}, \tag{16}$$

if f^* is twice differentiable at $g^a(x,\varepsilon)$, as a substitute for the Hessian $\nabla^2 F(x)$ given by (13) and (14), respectively. It is easy to see that matrices V given by (15) and (16) are symmetric and positive definite.

Now we state the algorithm.

Algorithm:

Step 0 (Initialization). Let $\lambda > 0$, $\sigma \in (0,1)$ and $\rho \in (0,1)$ be given constants. Choose $x^0 \in \Re^n$ and $\varepsilon_0 > 0$. Set $k := 0$.

Step 1 (Search direction). Compute $g^a(x^k, \varepsilon_k)$ and pick a positive semidefinite symmetric $n \times n$ matrix V_k and a scalar $\alpha_k > 0$. Compute the search direction

$$d^k := -(V_k + \alpha_k I)^{-1} g^a(x^k, \varepsilon_k).$$

Step 2 (Line search). Choose a scalar ε_{k+1} such that $0 < \varepsilon_{k+1} < \varepsilon_k$. Let m_k be the smallest nonnegative integer m such that

$$F^a(x^k + \rho^m d^k, \varepsilon_{k+1}) \le F^a(x^k, \varepsilon_k) + \sigma \rho^m \langle g^a(x^k, \varepsilon_k), d^k \rangle + \varepsilon_k.$$

Set $\tau_k := \rho^{m_k}$ and $x^{k+1} := x^k + \tau_k d^k$. Let $k := k+1$ and return to Step 1. □

In Step 1, there are some possibilities in choosing matrix V_k. For example, from the above arguments, we may set

$$V_k := \begin{cases} \frac{1}{\lambda} I - \frac{1}{\lambda}\left[I + \lambda \nabla^2 f(p^a(x,\varepsilon))\right]^{-1} & \text{if } \nabla^2 f(p^a(x,\varepsilon)) \text{ exists,} \\ V_{k-1} & \text{otherwise,} \end{cases}$$

with $V_{-1} = I$. Further we may use the formula (16) if possible. Alternatively, we may choose V_k by calculating $\nabla g(\hat{x}^k)$, where $\hat{x}^k$ is very close to x^k and g is differentiable at $\hat{x}^k$. This issue is discussed in [12]. By (4), $\nabla g(\hat{x}^k)$ can be made as close as possible to a member of $\partial_B g(x^k)$ by making $\hat{x}^k$ close to x^k. This justifies the condition (ii) of Theorem 3, which will be discussed in Section 5.

The next proposition ensures that, at each iteration k of the algorithm, m_k is well-defined and hence the step size $\tau_k > 0$ can be determined finitely in Step 2.

Proposition 4 *For every k, there exists $\bar{\tau}_k > 0$ such that*

$$F^a(x^k + \tau d^k, \varepsilon_{k+1}) \leq F^a(x^k, \varepsilon_k) + \sigma\tau\langle g^a(x^k, \varepsilon_k), d^k\rangle + \varepsilon_k \tag{17}$$

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

Proof. By Lemma 1, we have

$$F^a(x^k + \tau d^k, \varepsilon_{k+1}) \leq F(x^k + \tau d^k) + \varepsilon_{k+1}, \tag{18}$$

$$F(x^k) \leq F^a(x^k, \varepsilon_k), \tag{19}$$

and

$$\begin{aligned} \langle \nabla F(x^k), d^k\rangle - \langle g^a(x^k, \varepsilon_k), d^k\rangle &= \langle \nabla F(x^k) - g^a(x^k, \varepsilon_k), d^k\rangle \\ &\leq \|\nabla F(x^k) - g^a(x^k, \varepsilon_k)\|\,\|d^k\| \\ &\leq \sqrt{\frac{2\varepsilon_k}{\lambda}}\|d^k\|. \end{aligned} \tag{20}$$

Adding (18), (19) and (20) multiplied by τ, and rearranging terms, we obtain

$$\begin{aligned} F^a(x^k + \tau d^k, \varepsilon_{k+1}) \leq\ & F^a(x^k, \varepsilon_k) + \tau\langle g^a(x^k, \varepsilon_k), d^k\rangle + \varepsilon_{k+1} + \tau\sqrt{\frac{2\varepsilon_k}{\lambda}}\|d^k\| \\ & + F(x^k + \tau d^k) - F(x^k) - \tau\langle \nabla F(x^k), d^k\rangle. \end{aligned} \tag{21}$$

If $d^k = 0$, then (21) implies that (17) holds for any $\tau > 0$, because $\varepsilon_{k+1} < \varepsilon_k$. Consider the case $d^k \neq 0$. Then, since $V_k + \alpha_k I$ is positive definite, we have

$$\langle g^a(x^k, \varepsilon_k), d^k\rangle = -\langle (V_k + \alpha_k I)d^k, d^k\rangle < 0,$$

so that

$$\langle g^a(x^k, \varepsilon_k), d^k\rangle < \sigma\langle g^a(x^k, \varepsilon_k), d^k\rangle. \tag{22}$$

Moreover, since $\varepsilon_{k+1} < \varepsilon_k$, there exists $\bar{\tau}_k > 0$ such that

$$F(x^k + \tau d^k) - F(x^k) - \tau\langle \nabla F(x^k), d^k\rangle + \tau\sqrt{\frac{2\varepsilon_k}{\lambda}}\|d^k\| + \varepsilon_{k+1} < \varepsilon_k \tag{23}$$

for all $\tau \in (0, \bar{\tau}_k)$. That (17) holds for all $\tau \in (0, \bar{\tau}_k)$ now follows from (21) – (23). □

4 Global Convergence

The following theorem establishes global convergence of the algorithm.

Theorem 1 *Assume that the objective function f of problem (1) is bounded from below. Let $\{\alpha_k\}$ be a bounded sequence of positive numbers and suppose that there exists a constant $\gamma > 0$ such that*

$$\langle (V_k + \alpha_k I)d, d\rangle \geq \gamma \|d\|^2 \qquad \text{for all } d \in \Re^n \text{ and all } k. \tag{24}$$

If the sequence $\{\varepsilon_k\}$ of strictly decreasing positive numbers satisfies the condition $\sum_{k=0}^{\infty} \varepsilon_k < +\infty$, then any accumulation point of $\{x^k\}$ is an optimal solution of problem (1).

Proof. By the line search rule, it holds that

$$F^a(x^{k+1}, \varepsilon_{k+1}) \leq F^a(x^k, \varepsilon_k) + \sigma \tau_k \langle g^a(x^k, \varepsilon_k), d^k\rangle + \varepsilon_k \tag{25}$$

for all k. Since $\langle g^a(x^k, \varepsilon_k), d^k\rangle \leq 0$, it follows from (25) that

$$F^a(x^{k+1}, \varepsilon_{k+1}) \leq F^a(x^k, \varepsilon_k) + \varepsilon_k, \tag{26}$$

and hence

$$F^a(x^k, \varepsilon_k) \leq F^a(x^0, \varepsilon_0) + \sum_{i=0}^{k-1} \varepsilon_i,$$

which together with the assumption $\sum_{k=0}^{\infty} \varepsilon_k < +\infty$ implies that the sequence $\{F^a(x^k, \varepsilon_k)\}$ is bounded from above. On the other hand, f is bounded from below by assumption, and hence F is also bounded from below. Since $F^a(x^k, \varepsilon_k) \geq F(x^k)$ for all k, the sequence $\{F^a(x^k, \varepsilon_k)\}$ is bounded from below. Therefore the sequence $\{F^a(x^k, \varepsilon_k)\}$ has at least one accumulation point. In fact, it can be shown that the whole sequence $\{F^a(x^k, \varepsilon_k)\}$ is convergent. To see this, let $\{F^a(x^k, \varepsilon_k)\}$ have two accumulation points F_1^* and F_2^* such that $F_1^* < F_2^*$ and let $\{F^a(x^k, \varepsilon_k)\}_{k\in\kappa_1}$ and $\{F^a(x^k, \varepsilon_k)\}_{k\in\kappa_2}$ be subsequences converging to F_1^* and F_2^*, respectively. Put $\hat{\varepsilon} = (F_2^* - F_1^*)/3$. Then, there exist integers $\bar{k}_1, \bar{k}_2 > 0$ such that $F^a(x^k, \varepsilon_k) < F_1^* + \hat{\varepsilon}$ for all $k \in \kappa_1$ with $k > \bar{k}_1$ and $F^a(x^k, \varepsilon_k) > F_2^* - \hat{\varepsilon}$ for all $k \in \kappa_2$ with $k > \bar{k}_2$. Moreover, since $\sum_{k=0}^{\infty} \varepsilon_k < +\infty$, there exists an integer $\bar{k}_3 > 0$ such that $\sum_{k=\bar{k}_3}^{\infty} \varepsilon_k < \hat{\varepsilon}$. Pick an integer $\bar{k} \in \kappa_1$ such that $\bar{k} \geq \max\{\bar{k}_1, \bar{k}_2, \bar{k}_3\}$. Then, by (26), we have $F^a(x^k, \varepsilon_k) < F^a(x^{\bar{k}}, \varepsilon_{\bar{k}}) + \sum_{i=\bar{k}}^{k-1} \varepsilon_i < F_1^* + 2\hat{\varepsilon} = F_2^* - \hat{\varepsilon}$ for all $k > \bar{k}$. This is a contradiction, and hence $\{F^a(x^k, \varepsilon_k)\}$ has at most one accumulation point.

Then, since ε_k tends to zero and since $\langle g^a(x^k, \varepsilon_k), d^k\rangle \leq 0$, it follows from (25) that

$$\lim_{k\to\infty} \tau_k \langle g^a(x^k, \varepsilon_k), d^k\rangle = 0. \tag{27}$$

Now let $\bar{x}$ be an arbitrary accumulation point of $\{x^k\}$ and let $\{x^k\}_{k\in\kappa}$ be a subsequence converging to $\bar{x}$. Then by Lemma 1, we have

$$\lim_{k\to\infty, k\in\kappa} g^a(x^k, \varepsilon_k) = g(\bar{x}). \tag{28}$$

Since $\{(V_k+\alpha_k I)^{-1}\}$ is bounded, we may suppose by taking a further subsequence if necessary that

$$\lim_{k\to\infty,k\in\kappa} d^k = \bar{d} \tag{29}$$

for some vector $\bar{d} \in \Re^n$. Moreover, under the given assumption, we have

$$\lim_{k\to\infty,k\in\kappa} \langle g^a(x^k,\varepsilon_k), d^k\rangle = \langle g(\bar{x}), \bar{d}\rangle \le -\gamma\|\bar{d}\|^2. \tag{30}$$

If $\liminf_{k\to\infty,k\in\kappa} \tau_k > 0$, then (27) and (30) immediately yield

$$\bar{d} = 0. \tag{31}$$

On the other hand, if $\liminf_{k\to\infty,k\in\kappa} \tau_k = 0$, then the definition of m_k in the line search rule gives

$$F^a(x^k + \rho^{m_k-1}d^k, \varepsilon_{k+1}) > F^a(x^k,\varepsilon_k) + \sigma\rho^{m_k-1}\langle g^a(x^k,\varepsilon_k), d^k\rangle + \varepsilon_k. \tag{32}$$

By Lemma 1, we have

$$\begin{aligned} F^a(x^k,\varepsilon_k) &\ge F(x^k), \\ F^a(x^k + \rho^{m_k-1}d^k, \varepsilon_{k+1}) &\le F(x^k + \rho^{m_k-1}d^k) + \varepsilon_{k+1}. \end{aligned}$$

Combining these inequalities with (32) and taking $\varepsilon_k > \varepsilon_{k+1}$ into account yield

$$F(x^k + \rho^{m_k-1}d^k) > F(x^k) + \sigma\rho^{m_k-1}\langle g^a(x^k,\varepsilon_k), d^k\rangle,$$

that is,

$$\frac{F(x^k + \rho^{m_k-1}d^k) - F(x^k)}{\rho^{m_k-1}} > \sigma\langle g^a(x^k,\varepsilon_k), d^k\rangle. \tag{33}$$

Since $\rho^{m_k-1} = \tau_k/\rho \to 0$, we have

$$\lim_{k\to\infty,k\in\kappa} \frac{F(x^k + \rho^{m_k-1}d^k) - F(x^k)}{\rho^{m_k-1}} = \langle g(\bar{x}), \bar{d}\rangle.$$

So, by taking the limit in (33) on the subsequence $k \in \kappa$, we obtain

$$\langle g(\bar{x}), \bar{d}\rangle \ge \sigma\langle g(\bar{x}), \bar{d}\rangle.$$

In view of (30), the last inequality also gives

$$\bar{d} = 0.$$

Since $g^a(x^k,\varepsilon_k) = -(V_k+\alpha_k I)d^k$ and since $\{V_k+\alpha_k I\}$ is bounded, it follows from $\bar{d} = 0$ that

$$\lim_{k\to\infty,k\in\kappa} g^a(x^k,\varepsilon_k) = g(\bar{x}) = 0,$$

namely,

$$\nabla F(\bar{x}) = 0.$$

Therefore, by Proposition 2, $\bar{x}$ is a solution of problem (1). □

5 Sequential and Superlinear Convergence

Our next goal is to establish sequential and Q-superlinear convergence of the algorithm.

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

Proof. By the BD-regularity of g at $\bar{x}$ and the convexity of F, $\bar{x}$ is the unique optimal solution of (2); for a proof, see Theorem 2.2 of [14]. By Proposition 2, $\bar{x}$ is also the unique optimal solution of (1).

Let $\delta > 0$. By Theorem 1, there are only a finite number of k such that

$$\delta \leq \|x^k - \bar{x}\| \leq 2\delta.$$

Otherwise, there will be an accumulation point $\hat{x}$ of $\{x^k\}$ such that

$$\delta \leq \|\hat{x} - \bar{x}\| \leq 2\delta.$$

Then $\hat{x}$ must be an optimal solution of (1). This contradicts the fact that $\bar{x}$ is the unique optimal solution of (1). Hence, there is $k_0 > 0$ such that if $k \geq k_0$, either

$$\|x^k - \bar{x}\| \leq \delta$$

or

$$\|x^k - \bar{x}\| \geq 2\delta.$$

Let $\kappa_1 = \{k \,|\, \|x^k - \bar{x}\| \leq \delta, k \geq k_0\}$, $\kappa_2 = \{k \,|\, \|x^k - \bar{x}\| \geq 2\delta, k \geq k_0\}$, $\kappa_3 = \{k \,|\, k \in \kappa_1, k+1 \in \kappa_2\}$. Then, if $k \in \kappa_3$, we have

$$\|d^k\| \geq \|x^k - x^{k+1}\| \geq \delta. \tag{34}$$

Assume κ_2 is infinite. Then κ_3 is also infinite. By the same argument as above, $\{x^k \,|\, k \in \kappa_1\}$ cannot have an accumulation point other than $\bar{x}$. Hence, $\{x^k \,|\, k \in \kappa_1\}$ converges to $\bar{x}$. By the proof of Theorem 1, at least for a subsequence κ_4 of κ_3,

$$\lim_{k \to \infty, k \in \kappa_4} d^k = 0,$$

which contradicts (34). Hence, κ_2 is finite. This proves that the entire sequence $\{x^k\}$ converges to $\bar{x}$. □

By Proposition 3, we may let α_k tend to zero as k tends to infinity if we choose V_k as an approximate member of $\partial_B g(x^k)$. This will be useful in establishing Q-superlinear convergence of the algorithm.

Theorem 3 *Suppose that the conditions of Theorem 2 hold and that g is semismooth at the unique solution $\bar{x}$ to (1). Suppose furthermore that*
(i) $\varepsilon_k = o(\|g(x^k)\|^2)$;
(ii) $\lim_{k\to\infty} dist(V_k, \partial_B g(x^k)) = 0$;
(iii) $\lim_{k\to\infty} \alpha_k = 0$;
(iv) $\tau_k \equiv 1$ *for all large* k.
Then $\{x^k\}$ *converges to* $\bar{x}$ *Q-superlinearly.*

Proof. First note that, by Theorem 2, the sequence $\{x^k\}$ converges to $\bar{x}$. Then, by the condition (i), the inequality (11), Propositions 1 and 2, we have

$$\begin{aligned} \|g^a(x^k, \varepsilon_k) - g(x^k)\| &= O(\sqrt{\varepsilon_k}) \\ &= o(\|g(x^k)\|) \\ &= o(\|g(x^k) - g(\bar{x})\|) \\ &= o(\|x^k - \bar{x}\|). \end{aligned} \tag{35}$$

By the condition (ii), there is a $\bar{V}_k \in \partial_B g(x^k)$ such that

$$\|V_k - \bar{V}_k\| = o(1). \tag{36}$$

Since g is semismooth at $\bar{x}$, according to [10, 13], we have

$$\|g(x^k) - g(\bar{x}) - \bar{V}_k(x^k - \bar{x})\| = o(\|x^k - \bar{x}\|). \tag{37}$$

Notice that $\|(V_k + \alpha_k I)^{-1}\| = O(1)$. By the algorithm, (35), (36), (37), the conditions (iii) and (iv), for all large k, we have

$$\begin{aligned} \|x^{k+1} - \bar{x}\| &= \|x^k - \bar{x} - (V_k + \alpha_k I)^{-1} g^a(x^k, \varepsilon_k)\| \\ &\le \|(V_k + \alpha_k I)^{-1}\| \cdot \{\|g^a(x^k, \varepsilon_k) - g(x^k)\| + \|g(x^k) - g(\bar{x}) - \bar{V}_k(x^k - \bar{x})\| \\ &\quad + [\|V_k - \bar{V}_k\| + \alpha_k] \cdot \|x^k - \bar{x}\|\} \\ &= o(\|x^k - \bar{x}\|). \end{aligned}$$

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

In the above theorem, the condition (i) might seem somewhat problematic, since, in the algorithm, ε_k is chosen before x^k is determined. The next result shows that we may still ensure fast convergence even if the condition (i) is replaced by a more realistic condition.

Corollary 1 *Suppose that the conditions of Theorem 3 hold, except that the condition (i) is replaced by (i')* $\varepsilon_k = o(\|g(x^{k-1})\|^2)$. *Then* $\{x^k\}$ *converges to* $\bar{x}$ *at least 2-step Q-superlinearly.*

Proof. Under condition (i'), we have the estimate

$$\|g^a(x^k, \varepsilon_k) - g(x^k)\| = o(\|x^{k-1} - \bar{x}\|),$$

instead of (35). Then, an argument similar to the proof of Theorem 3 yields

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

That is, the rate of convergence is at least 2-step Q-superlinear. □

We now prove that the condition (iv) in Theorem 3 will be satisfied automatically, i.e., the Maratos effect will not occur.

Theorem 4 *Suppose that the conditions of Theorem 3, except the condition (iv), hold and that $\sigma < \frac{1}{2}$. Then the condition (iv) holds and $\{x^k\}$ converges to the unique solution $\bar{x}$ to (1) Q-superlinearly.*

Proof. Theorem 3 shows that

$$\|x^k + d^k - \bar{x}\| = o(\|x^k - \bar{x}\|), \tag{38}$$

which implies

$$\|d^k\| = \|x^k - \bar{x}\| + o(\|x^k - \bar{x}\|). \tag{39}$$

By Theorem 1 of [9], we have

$$\begin{aligned} F(x^k + d^k) &= F(\bar{x}) + \langle g(\bar{x}), x^k + d^k - \bar{x}\rangle \\ &\quad + \frac{1}{2}\langle \hat{V}_k(x^k + d^k - \bar{x}), x^k + d^k - \bar{x}\rangle + o(\|x^k + d^k - \bar{x}\|^2) \\ &= F(\bar{x}) + o(\|x^k - \bar{x}\|^2), \end{aligned} \tag{40}$$

where $\hat{V}_k \in \partial_B g(x^k + d^k)$ and the last equality follows from (38), the local boundedness of $\partial_B g$ at $\bar{x}$ and $g(\bar{x}) = 0$. Similarly,

$$\begin{aligned} F(x^k) &= F(\bar{x}) + \langle g(\bar{x}), x^k - \bar{x}\rangle + \frac{1}{2}\langle \bar{V}_k(x^k - \bar{x}), x^k - \bar{x}\rangle + o(\|x^k - \bar{x}\|^2) \\ &= F(\bar{x}) + \frac{1}{2}\langle \bar{V}_k(x^k - \bar{x}), x^k - \bar{x}\rangle + o(\|x^k - \bar{x}\|^2), \end{aligned} \tag{41}$$

where $\bar{V}_k \in \partial_B g(x^k)$. By (38), (40) and (41), we have

$$\begin{aligned} & F(x^k + d^k) - F(x^k) - \frac{1}{2}\langle g(x^k), d^k\rangle \\ = & -\frac{1}{2}\langle \bar{V}_k(x^k - \bar{x}), x^k - \bar{x}\rangle - \frac{1}{2}\langle g(x^k), d^k\rangle + o(\|x^k - \bar{x}\|^2) \\ = & -\frac{1}{2}\langle g(x^k) - g(\bar{x}) - \bar{V}_k(x^k - \bar{x}), d^k\rangle \\ & -\frac{1}{2}\langle \bar{V}_k(x^k - \bar{x}), x^k + d^k - \bar{x}\rangle + o(\|x^k - \bar{x}\|^2). \end{aligned} \tag{42}$$

By (37) and (39),

$$\langle g(x^k) - g(\bar{x}) - \bar{V}_k(x^k - \bar{x}), d^k \rangle = o(\|x^k - \bar{x}\|^2), \tag{43}$$

while by (38) and the local boundedness of $\partial_B g$ at $\bar{x}$,

$$\langle \bar{V}_k(x^k - \bar{x}), x^k + d^k - \bar{x} \rangle = o(\|x^k - \bar{x}\|^2). \tag{44}$$

Combining (42), (43) and (44), we have

$$F(x^k + d^k) - F(x^k) - \frac{1}{2}\langle g(x^k), d^k \rangle = o(\|x^k - \bar{x}\|^2). \tag{45}$$

By (20), the condition (i) of Theorem 3 and (39),

$$\begin{aligned} \frac{1}{2}\langle g(x^k) - g^a(x^k, \varepsilon_k), d^k \rangle &\leq \sqrt{\frac{\varepsilon_k}{2\lambda}}\|d^k\| \\ &= o(\|g(x^k)\|)\|x^k - \bar{x}\| \\ &= o(\|g(x^k) - g(\bar{x})\|)\|x^k - \bar{x}\| \\ &= o(\|x^k - \bar{x}\|^2), \end{aligned} \tag{46}$$

where the last equality follows from Proposition 1.

Adding (46), (45), (19) and (18) with $\tau = 1$, we obtain

$$\begin{aligned} F^a(x^k + d^k, \varepsilon_{k+1}) &\leq F^a(x^k, \varepsilon_k) + \sigma\langle g^a(x^k, \varepsilon_k), d^k \rangle + \varepsilon_{k+1} \\ &\quad + \left(\frac{1}{2} - \sigma\right)\langle g^a(x^k, \varepsilon_k), d^k \rangle + o(\|x^k - \bar{x}\|^2) \\ &\leq F^a(x^k, \varepsilon_k) + \sigma\langle g^a(x^k, \varepsilon_k), d^k \rangle + \varepsilon_k \\ &\quad - \left(\frac{1}{2} - \sigma\right)\langle (V_k + \alpha_k I)d^k, d^k \rangle + o(\|x^k - \bar{x}\|^2). \end{aligned}$$

However, by (24) and (39),

$$\begin{aligned} -\left(\frac{1}{2} - \sigma\right)\langle (V_k + \alpha_k I)d^k, d^k \rangle + o(\|x^k - \bar{x}\|^2) &\leq -\gamma\left(\frac{1}{2} - \sigma\right)\|x^k - \bar{x}\|^2 + o(\|x^k - \bar{x}\|^2) \\ &< 0 \end{aligned}$$

for all large k. Hence, for all large k,

$$F^a(x^k + d^k, \varepsilon_{k+1}) \leq F^a(x^k, \varepsilon_k) + \sigma\langle g^a(x^k, \varepsilon_k), d^k \rangle + \varepsilon_k.$$

Therefore, $m_k \equiv 0$, i.e, $\tau_k \equiv 1$ in Step 2 of the algorithm for all large k. Q-superlinear convergence of $\{x^k\}$ to $\bar{x}$ now follows from Theorem 3. This proves the theorem. □

6 Conclusion

We have proposed an algorithm for solving a convex minimization problem via the Moreau-Yosida regularization. We have shown that the algorithm is globally convergent and, under some additional conditions, the convergence rate is superlinear. We note that the algorithm inherently involves the inner iteration for finding a vector $p^a(x, \varepsilon)$ satisfying (6) for given x and ε. There are a number of iterative methods [4, 1, 2] that can be used as a procedure to obtain $p^a(x, \varepsilon)$. In general, those methods tend to spend more (inner) iterations as the outer iteration proceeds. Nevertheless, we emphasize that the algorithm proposed in this paper is the first nonsmooth optimization method that is provably superlinearly convergent.

References

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

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

[3] F. Facchinei, Minimization of SC^1 functions and the Maratos effect, *Operations Research Letters* forthcoming.

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

[5] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II*, Springer-Verlag, Berlin-Heidelberg, 1993.

[6] C. Lemaréchal and C. Sagastizábal, Practical aspects of the Moreau-Yosida regularization I: Theoretical properties, Research Report No. 2250, INRIA, Le Chesnay, France, 1994.

[7] B. Martinet, Régularisation d'inéquations variationelles par approximations successives, *Revue Française d'Automatique, Informatique et Recherche Opérationnelle* 4 (1970) 154-159.

[8] J.-S. Pang and L. Qi, Nonsmooth equations: Motivation and algorithms, *SIAM Journal on Optimization* 3 (1993) 443–465.

[9] J.-S. Pang and L. Qi, A globally convergent Newton method for convex SC^1 minimization problems, *Journal of Optimization Theory and Applications* forthcoming.

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

[11] L. Qi, Superlinearly convergent approximate Newton methods for LC^1 optimization problems, *Mathematical Programming* 64 (1994) 277–294.

[12] L. Qi, Second-order analysis of the Moreau-Yosida approximation of a convex function, Applied Mathematics Report AMR94/20, School of Mathematics, The University of New South Wales, Sydney, Australia, 1994.

[13] L. Qi and J. Sun, A nonsmooth version of Newton's method, *Mathematical Programming* 58 (1993) 353–367.

[14] L. Qi and R.S. Womersley, An SQP algorithm for extended linear problems in stochastic programming, Applied Mathematics Preprint AM92/23, School of Mathematics, The University of New South Wales, Sydney, Australia, 1992.

[15] R.T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM Journal on Control and Optimization* 14 (1976) 877-898.

[16] R.T. Rockafellar, Maximal monotone relations and the second derivatives of nonsmooth functions, *Ann. Inst. H. Poincaré Analyse non linéaire* 2 (1985) 167-184.