Regularized Newton Methods for Convex Minimization Problems with Singular Solutions

Dong-Hui Li [1]
Department of Applied Mathematics
Hunan University
Changsha, China 410082
dhli@mail.hunu.edu.cn

Masao Fukushima [2]
Department of Applied Mathematics and Physics
Graduate School of Informatics
Kyoto University
Kyoto 606-8501, Japan
fuku@amp.i.kyoto-u.ac.jp

Liqun Qi [3]
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
maqilq@polyu.edu.hk

Nobuo Yamashita
Department of Applied Mathematics and Physics
Graduate School of Informatics
Kyoto University
Kyoto 606-8501, Japan
nobuo@amp.i.kyoto-u.ac.jp

February 28, 2002

Abstract

This paper studies convergence properties of regularized Newton methods for minimizing a convex function whose Hessian matrix may be singular everywhere. We show that if the objective function is LC^2, then the methods possess local quadratic convergence under a local error bound condition without the requirement of isolated nonsingular solutions. By using a backtracking line search, we globalize an inexact regularized Newton method. We show that the unit stepsize is accepted eventually.

[1]The author is partially supported by the National Natural Science Foundation of China via Grant 10171030 and a Hong Kong Polytechnic University Postdoctoral Fellowship.

[2]The author is partially supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan.

[3]The author is partially supported by the Research Grant Council of Hong Kong.

1 Introduction

Let $f : R^n \to R$ be LC^2 on an open subset S of R^n. That is, it is twice continuously differentiable, and the Hessian $\nabla^2 f$ of f is Lipschitz continuous on S, namely, there is a constant $L > 0$ such that

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\| \tag{1.1}$$

for all $x, y \in S$. Consider the minimization problem

$$\min f(x), \qquad x \in R^n. \tag{1.2}$$

Throughout we suppose that f is convex and denote by X the set of solutions of (1.2). Then X is convex. We suppose that X is nonempty and contained in S. Denote by $g(x) = \nabla f(x)$ and $G(x) = \nabla^2 f(x)$ the gradient and the Hessian matrix of f at x, respectively. It is well-known that f is convex if and only if $G(x)$ is positive semidefinite for all $x \in R^n$. Moreover, if f is convex, then x is a solution of (1.2) if and only if it is a solution of the system of nonlinear equations

$$g(x) = 0. \tag{1.3}$$

The classical Newton method is quite a way of solving (1.2) and (1.3). An attractive feature of Newton's method lies in its local quadratic convergence when $G(\bar{x})$ is nonsingular at a solution $\bar{x}$. We call a solution $\bar{x}$ of problem (1.2) or (1.3) nonsingular if $G(\bar{x})$ is a nonsingular matrix. Otherwise we call $\bar{x}$ singular. It is clear that a nonsingular solution is locally isolated. We are particularly interested in the case where problem (1.2) may have singular solutions. When Newton's method is applied to such a problem, the quadratic rate of convergence may no longer be guaranteed.

Recently, there has been some progress in convergence analysis of Newton-type methods for problem (1.3) with singular solutions. Yamashita and Fukushima [5] studied the Levenberg-Marquardt method for solving (1.3), in which the subproblem is the system of linear equations

$$(G(x_k)^T G(x_k) + \mu_k I)d + G(x_k)^T g(x_k) = 0, \tag{1.4}$$

where $\mu_k > 0$ is a parameter and I is the identity matrix, and proved that if $\|g(x)\|$ provides a local error bound for problem (1.3) and μ is chosen as $\mu_k = \|g(x_k)\|^2$, then the Levenberg-Marquardt method retains a quadratic convergence property. The concept of a local error bound is defined as follows.

Definition 1.1 *A function $F : R^n \to R$ is said to provide a local error bound for problem (1.2) near $\bar{x} \in X$ if there exist a neighborhood $\Omega(\bar{x})$ of $\bar{x}$ and a constant $m > 0$ such that for all $x \in \Omega(\bar{x})$*

$$F(x) \geq m \, \mathrm{dist}(x, X), \tag{1.5}$$

where $\mathrm{dist}(x, X)$ *denotes the distance from point x to the solution set X.*

In the remaining part of the paper, we will omit the phrase "for problem (1.2)" when we talk about a local error bound. It is clear that if X is a singleton, i.e., $X = \{\bar{x}\}$, and the Hessian matrix $G(\bar{x})$ is nonsingular, then $\|g(x)\|$ provides a local error bound near $\bar{x}$. However, the converse is not true in general. An example is given in [1], which shows that the local error bound condition is weaker than nonsingularity of $G(\bar{x})$. Note also that the condition in Definition 1.1 depends on a particular choice of $\bar{x}$. More specifically, let the function $f : R^2 \to R$ be defined by

$$f(x) = \tfrac{1}{2}x_1^2 + \phi(x_2)$$

with

$$\phi(x_2) = \begin{cases} (x_2 - 1)^4 & \text{if } x_2 \in [1, +\infty) \\ 0 & \text{if } x_2 \in (-1, 1) \\ (x_2 + 1)^4 & \text{if } x_2 \in (-\infty, -1]. \end{cases}$$

Clearly f is convex, and is LC^2 on any bounded set. The solution set of (1.2) is given by $X = \{x \in R^2 \,|\, x_1 = 0, -1 \leq x_2 \leq 1\}$, and every $\bar{x} \in X$ is singular. However, for any $\bar{x} \in X$ such that $-1 < \bar{x}_2 < 1$, $\|g(x)\|$ provides a local error bound near $\bar{x}$. In fact, for such $\bar{x}$, there is a neighborhood $\Omega(\bar{x})$ such that, for any $x = (x_1, x_2)^T \in \Omega(\bar{x})$, the point in X closest to x is given by $\hat{x} = (0, x_2)^T$. Thus we have $\text{dist}(x, X) = |x_1|$. Since $\|g(x)\| = \sqrt{x_1^2 + \phi'(x_2)^2} \geq |x_1|$, inequality (1.5) is satisfied with $m = 1$ for all $x \in \Omega(\bar{x})$. On the other hand, for $\bar{x} = (0,1)^T$ and $\bar{x} = (0,-1)^T$, $\|g(x)\|$ fails to provide a local error bound near $\bar{x}$. To see this, let $\bar{x} = (0,1)^T$ and consider a point $x^\epsilon = (0, 1+\epsilon)^T$ with $\epsilon > 0$. Then we have $\text{dist}(x^\epsilon, X) = \epsilon$ and $\|g(x^\epsilon)\| = 4\epsilon^3$. Since $\|g(x^\epsilon)\|/\text{dist}(x^\epsilon, X) = 4\epsilon^2 \to 0$ as $\epsilon \to 0$, there is no $m > 0$ such that (1.5) holds, no matter how small $\Omega(\bar{x})$ is chosen. Thus $\|g(x)\|$ cannot provide a local error bound near $\bar{x} = (0,1)^T$. In a similar manner, we can show that $\|g(x)\|$ fails to provide a local error bound near $\bar{x} = (0,-1)^T$.

Fan and Yuan [3] noted some possible defects of the choice $\mu_k = \|g(x_k)\|^2$ in the Levenberg-Marquardt method. If x_k is very close to X, then μ_k can be very small. As a result, μ_k may lose its role in (1.4). On the other hand, if x_k is far from X, then μ_k would be vary large and the solution of (1.4) may be very small. Taking this into account, Fan and Yuan [3] use $\mu_k = \|g(x_k)\|$ to improve the Yamashita-Fukushima method. With this choice of μ_k, it was proved [3] that the Levenberg-Marquardt method still possesses a local quadratic convergence property. Another improvement is due to Dan, Yamashita and Fukushima [1], where an inexact Levenberg-Marquardt method was studied. The subproblem of the Dan-Yamashita-Fukushima inexact method [1] is to find a vector d satisfying

$$(G(x_k)^T G(x_k) + \mu_k I)d + G(x_k)^T g(x_k) = r_k, \quad \|r_k\| = \mu_k\, o(\text{dist}(x_k, X)), \tag{1.6}$$

with $\mu_k = O(\|g(x_k)\|^\delta)$ for some $\delta \in (0, 2]$. The superlinear convergence of the inexact method was established. If in addition, $\mu_k = O(\|g(x_k)\|^2)$ and r_k satisfies

$$\|r_k\| = \mu_k O(\text{dist}(x_k, X)^2) = O(\text{dist}(x_k, X)^4), \tag{1.7}$$

then the inexact method converges quadratically. Combined with a line search strategy, the above methods can be globally convergent [1, 3, 5].

In this paper, we study the behavior of the regularized Newton methods for problem (1.2) with singular solutions. Of course, the above mentioned methods [1, 3, 5] can be used to solve (1.2) directly. However, there are some shortcomings when these methods are applied to (1.2). First, the system of linear equations (1.4) evaporates the sparsity the matrix $G(x_k)$ may have. Second, compared with Newton's method whose subproblem is the system of linear equations

$$G(x_k)d + g(x_k) = 0, \tag{1.8}$$

the coefficient matrix in (1.4) squares the condition number of $G(x_k)$. Third, the inexact rule (1.7) is very restrictive. Dembo, Eisenstat and Steihaug [2] proved that when applied to problem (1.3) with a nonsingular solution, the inexact Newton method is quadratically convergent if and only if r_k satisfies $\|r_k\| = O(\|x_k - x^*\|^2) = O(\text{dist}(x_k, X)^2)$, where x^* is the locally unique solution of (1.3). The purpose of this paper is to develop regularized Newton methods that overcome the above three shortcomings. Moreover, we prove that the unit stepsize is accepted eventually.

We summarize some attractive properties of the proposed methods as follows.

- The subproblem retains the possible sparsity of $G(x_k)$.
- The parameter μ_k is chosen to satisfy $\mu_k = O(\|g(x_k)\|)$.
- Global and quadratic convergence is achieved.
- Comparing the inexact version of the proposed method with the method of [1], the subproblem of the former method allows more inexactness, since it adopts an extension of the rule in [2].
- The unit stepsize is accepted eventually.

The paper is organized as follows. In the next section, we prove some useful properties of a local error bound. In Section 3, we propose exact and inexact versions of regularized Newton methods and establish quadratic convergence of the proposed methods. By using a backtracking line search, we globalize an inexact regularized Newton method in Section 4. We show the acceptance of the unit stepsize in Section 4.

2 Preliminaries

In this section, we prove some useful properties of a local error bound. In the rest of the paper, we assume for simplicity that f is LC^2 on the entire space R^n, although it is easy to extend the results to the more general case where f is LC^2 on a level set of f.

Let $\bar{x} \in X$. Since $G(\bar{x})$ is symmetric and positive semidefinite, there is an orthogonal matrix $\bar{Q} = (\bar{Q}_1, \bar{Q}_2)$ such that

$$G(\bar{x}) = (\bar{Q}_1, \bar{Q}_2) \begin{pmatrix} \bar{\Lambda} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{Q}_1^T \\ \bar{Q}_2^T \end{pmatrix}, \tag{2.1}$$

where $\bar{\Lambda}$ is a diagonal matrix with positive diagonals.

Let x be sufficiently close to $\bar{x}$ and $Q(x) = (Q_1, Q_2)$ be an orthogonal matrix conformable to $\bar{Q} = (\bar{Q}_1, \bar{Q}_2)$ such that

$$G(x) = (Q_1,\, Q_2) \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix}, \tag{2.2}$$

where Λ_1 and Λ_2 are diagonal matrices. By the theory of matrix perturbation [4, Chapter 7], we have $\Lambda_1 \to \bar{\Lambda}$ and $\Lambda_2 \to 0$ as $x \to \bar{x}$. So, there is a neighborhood $U(\bar{x})$ of $\bar{x}$ such that $\Lambda_1 \geq \frac{1}{2}\bar{\Lambda} > 0$ for all $x \in U(\bar{x})$. Let $\hat{x}$ be the projection of x onto X, i.e., $\hat{x} \in X$ and $\|x - \hat{x}\| = \mathrm{dist}(x, X)$. For simplicity, we let $\hat{Q}$ and Q denote $Q(\hat{x})$ and $Q(x)$, respectively, without confusing.

Lemma 2.1 *Let $\bar{x} \in X$ and (2.1) hold. Let Q and $\hat{Q}$ be orthogonal matrices satisfying (2.2) with $x = x$ and $x = \hat{x}$, respectively, where $\hat{x}$ is the projection of x onto X. Then there is a neighborhood $\Omega(\bar{x})$ of $\bar{x}$ such that the following statements are true:*

(i) It holds that

$$\|Q_2^T g(x)\| = o(\|x - \hat{x}\|) = o(\mathrm{dist}(x, X))$$

and

$$\|\hat{Q}_2^T g(x)\| = o(\|x - \hat{x}\|) = o(\mathrm{dist}(x, X)).$$

(ii) If $\|g(x)\|$ provides a local error bound near $\bar{x}$, then there is a constant $m_1 > 0$ such that for all $x \in \Omega(\bar{x})$

$$\|Q_1^T g(x)\| \geq m_1 \|x - \hat{x}\| = m_1\, \mathrm{dist}(x, X)$$

and

$$\|\hat{Q}_1^T g(x)\| \geq m_1 \|x - \hat{x}\| = m_1\, \mathrm{dist}(x, X).$$

(iii) If $\|g(x)\|$ provides a local error bound near $\bar{x}$, then there is a constant $m_2 > 0$ such that for all $x \in \Omega(\bar{x})$

$$\|Q_1^T (x - \hat{x})\| \geq m_2 \|x - \hat{x}\| = m_2\, \mathrm{dist}(x, X)$$

and

$$\|\hat{Q}_1^T (x - \hat{x})\| \geq m_2 \|x - \hat{x}\| = m_2\, \mathrm{dist}(x, X).$$

Proof We only prove the first equality and inequalities in (i) – (iii). The second equality and inequalities can be proved in a similar way.

Since f is LC^2, it is easy to deduce that

$$\begin{aligned}
\|Q_2^T g(x)\| &= \|Q_2^T (g(x) - g(\hat{x}))\| \\
&= \left\| Q_2^T \int_0^1 G(\hat{x} + \tau(x - \hat{x})) d\tau (x - \hat{x}) \right\| \\
&= \|Q_2^T G(x)(x - \hat{x})\| + o(\|x - \hat{x}\|) \\
&= \|\Lambda_2 Q_2^T (x - \hat{x})\| + o(\|x - \hat{x}\|) = o(\|x - \hat{x}\|).
\end{aligned}$$

This yields the first equality in statement (i). We have by (1.5) and (i) that

$$\begin{aligned} \|Q_1^T g(x)\|^2 &= \left\| \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} g(x) \right\|^2 - \|Q_2^T g(x)\|^2 \\ &= \|g(x)\|^2 + o(\|x-\hat{x}\|^2) \\ &\geq m^2\|x-\hat{x}\|^2 + o(\|x-\hat{x}\|^2). \end{aligned}$$

This implies the first inequality in statement (ii). In addition, it follows from (1.5) that

$$\begin{aligned} m\|x-\hat{x}\| &\leq \|g(x)\| = \|Q^T(g(x)-g(\hat{x}))\| \\ &= \|Q^T G(x)(x-\hat{x})\| + o(\|x-\hat{x}\|) \\ &= \|\Lambda_1 Q_1^T(x-\hat{x})\| + o(\|x-\hat{x}\|). \end{aligned}$$

This implies the first inequality in statement (iii). The proof is complete. Ξ

3 Local Quadratic Convergence Properties

In this section, we propose some exact and inexact versions of regularized Newton methods and study their local convergence properties. For simplicity, we use g_k, G_k etc. as abbreviations of $g(x_k)$, $G(x_k)$ etc. The subproblem of the exact version of regularized Newton methods is the system of linear equations

$$(G_k + \mu_k I)d + g_k = 0, \tag{3.1}$$

where $\mu_k > 0$. It is obvious that system (3.1) always has a unique solution because G_k is symmetric and positive semi-definite. Accordingly, the subproblem of an inexact version of the regularized Newton method is to find a vector d such that

$$(G_k + \mu_k I)d + g_k = r_k, \qquad \|r_k\| \leq \eta_k, \tag{3.2}$$

where $\eta_k > 0$ is a suitable parameter converging to zero as k goes to infinity. In our methods, we always take

$$\mu_k = C\|g(x_k)\|, \tag{3.3}$$

where $C > 0$ is a constant. If $\|g(x)\|$ provides a local error bound near $\bar{x} \in X$, (3.3) implies that there are positive constants $\bar{m} \leq \bar{M}$ such that

$$\bar{m}\,\mathrm{dist}(x_k, X) \leq \mu_k \leq \bar{M}\,\mathrm{dist}(x_k, X), \tag{3.4}$$

whenever x_k lies in a neighborhood of $\bar{x}$.

In the case where $G(\bar{x})$ is nonsingular, a possible choice is to let η_k satisfy $\eta_k \leq \nu_k\|g(x_k)\|$ with $\nu_k \to 0$. If η_k satisfies $\eta_k \leq C\|g(x_k)\|^2$, the iterative sequence generated by the inexact regularized Newton method converges quadratically [2]. For problems with singular solutions, Dan, Yamashita and Fukushima [1] choose η_k such that $\eta_k = o(\mathrm{dist}(x_k, X))$ to ensure superlinear convergence, and $\eta_k = O(\mathrm{dist}(x_k, X)^4)$ to obtain quadratic convergence of the method. These choices of η_k are obviously more restrictive than those in [2]. We are going to develop inexact regularized Newton methods whose inexact rules are the same as those in [2]. Specifically, we consider the following two inexact regularized Newton methods. Remember that the parameter μ_k is chosen as (3.3).

Inexact Regularized Newton Method I (IRM I):

$$(G_k + \mu_k I)d + g_k = r_k, \qquad \|r_k\| = O(\mu_k^2) = O(\|g_k\|^2). \tag{3.5}$$

Inexact Regularized Newton Method II (IRM II):

$$(G_k + \mu_k I)d + g_k = r_k, \qquad \|r_k\| \le \alpha\mu_k\|d_k\|, \tag{3.6}$$

where $\alpha \in (0, 1)$.

It is easy to see that these methods induce the exact regularized Newton method as a special case corresponding to $r_k = 0$. In the later part of the paper, we focus our attention on IRM I and IRM II.

Lemma 3.1 *Let $\bar{x} \in X$ and (2.1) hold. Suppose that $\|g(x)\|$ provides a local error bound near $\bar{x}$. Let d_k be a solution of (3.1) or (3.5) or (3.6). Then there is a neighborhood $\Omega(\bar{x})$ of $\bar{x}$ such that the following statements are true:*

(i) There is a constant $m_3 > 0$ such that if $x_k \in \Omega(\bar{x})$, then

$$\|Q_{1,k}^T d_k\| \ge m_3 \operatorname{dist}(x_k, X),$$

where $Q_k = Q(x_k) = (Q_{1,k}, Q_{2,k})$ is an orthogonal matrix such that (2.2) holds with $x = x_k$.

(ii) If $x_k \in \Omega(\bar{x})$, then

$$\|d_k\| = O(\operatorname{dist}(x_k, X)). \tag{3.7}$$

(iii) There are positive constants $m_4 \le M_4$ such that if $x_k \in \Omega(\bar{x})$, then

$$m_4 \operatorname{dist}(x_k, X) \le \|d_k\| \le M_4 \operatorname{dist}(x_k, X).$$

Proof For each k, let $\hat{x}_k$ denote the projection of x_k onto X. Since Q_k is orthogonal, we have $\|d_k\|^2 = \|Q_{1,k}^T d_k\|^2 + \|Q_{2,k}^T d_k\|^2$. Therefore, statement (iii) follows from (i) and (ii). So, we only need to show (i) and (ii).

Multiplying (3.5) or (3.6) by $Q_{1,k}^T$, we get

$$(\Lambda_{1,k} + \mu_k I)Q_{1,k}^T d_k + Q_{1,k}^T g_k = Q_{1,k}^T r_k, \tag{3.8}$$

where we use $Q_{1,k}^T G_k = \Lambda_{1,k} Q_{1,k}^T$. This implies

$$\begin{aligned} \|Q_{1,k}^T d_k\| &= \|(\Lambda_{1,k} + \mu_k I)^{-1}(Q_{1,k}^T g_k - Q_{1,k}^T r_k)\| \\ &\ge \|\Lambda_{1,k} + \mu_k I\|^{-1}(\|Q_{1,k}^T g_k\| - \|Q_{1,k}^T r_k\|) \\ &\ge \|\Lambda_{1,k} + \mu_k I\|^{-1}(m_1\|x_k - \hat{x}_k\| - \|r_k\|), \end{aligned} \tag{3.9}$$

where the last inequality follows from Lemma 2.1 (ii).

Multiplying (3.5) or (3.6) by $Q_{2,k}^T$, we get

$$(\Lambda_{2,k} + \mu_k I)Q_{2,k}^T d_k + Q_{2,k}^T g_k = Q_{2,k}^T r_k,$$

where we use $Q_{2,k}^T G_k = \Lambda_{2,k} Q_{2,k}^T$. This implies

$$\begin{aligned}
\|Q_{2,k}^T d_k\| &\leq \|(\Lambda_{2,k} + \mu_k I)^{-1} Q_{2,k}^T r_k\| + \|(\Lambda_{2,k} + \mu_k I)^{-1} Q_{2,k}^T g_k\| \\
&\leq \|(\Lambda_{2,k} + \mu_k I)^{-1} Q_{2,k}^T r_k\| + \|(\Lambda_{2,k} + \mu_k I)^{-1} Q_{2,k}^T (g_k - \hat{g}_k - G_k(x_k - \hat{x}_k))\| \\
&\quad + \|(\Lambda_{2,k} + \mu_k I)^{-1} Q_{2,k}^T G_k (x_k - \hat{x}_k)\| \\
&\leq \|(\Lambda_{2,k} + \mu_k I)^{-1}\| \cdot \|r_k\| + \|(\Lambda_{2,k} + \mu_k I)^{-1}\| O(\|x_k - \hat{x}_k\|^2) \\
&\quad + \|(\Lambda_{2,k} + \mu_k I)^{-1} \Lambda_{2,k} Q_{2,k}^T (x_k - \hat{x}_k)\|, \qquad (3.10)
\end{aligned}$$

where the second inequality follows from $\hat{g}_k = g(\hat{x}_k) = 0$, and the third inequality follows from $\|Q_{2,k}^T\| \leq 1$ and $Q_{2,k}^T G_k = \Lambda_{2,k} Q_{2,k}^T$. Note that (3.4) implies

$$\|(\Lambda_{2,k} + \mu_k I)^{-1}\| \leq \mu_k^{-1} \leq \bar{m}^{-1} \|x_k - \hat{x}_k\|^{-1}. \qquad (3.11)$$

Moreover, since $\Lambda_{2,k}$ is a diagonal matrix with nonnegative diagonals, we have

$$\|(\Lambda_{2,k} + \mu_k I)^{-1} \Lambda_{2,k}\| \leq 1,$$

and hence

$$\begin{aligned}
\|(\Lambda_{2,k} + \mu_k I)^{-1} \Lambda_{2,k} Q_{2,k}^T (x_k - \hat{x}_k)\| &\leq \|(\Lambda_{2,k} + \mu_k I)^{-1} \Lambda_{2,k}\| \cdot \|Q_{2,k}^T (x_k - \hat{x}_k)\| \\
&\leq \|Q_{2,k}^T (x_k - \hat{x}_k)\| = O(\|x_k - \hat{x}_k\|).
\end{aligned}$$

This along with (3.10) and (3.11) yields

$$\|Q_{2,k}^T d_k\| \leq O(\|x_k - \hat{x}_k\|) + \mu_k^{-1} \|r_k\|. \qquad (3.12)$$

For IRM I, we have $\|r_k\| = O(\|x_k - \hat{x}_k\|^2)$. Since $\Lambda_{1,k} \geq \frac{1}{2}\bar{\Lambda} > 0$ for any x_k sufficiently close to $\bar{x}$, $\{\Lambda_{1,k} + \mu_k I\}$ is uniformly positive definite. Moreover, since $\mu_k \leq \bar{M}\|x_k - \hat{x}_k\|$ by (3.4), (3.9) implies that statement (i) holds with some suitable constant $m_3 > 0$. On the other hand, it is easy to deduce from (3.8) that

$$\|Q_{1,k}^T d_k\| = O(\|x_k - \hat{x}_k\|) = O(\operatorname{dist}(x_k, X)). \qquad (3.13)$$

We also have from (3.4) and (3.12) that $\|Q_{2,k}^T d_k\| = O(\|x_k - \hat{x}_k\|) = O(\operatorname{dist}(x_k, X))$. This together with (3.13) implies (3.7), which shows statement (ii).

Consider IRM II. We get from (3.8)

$$\begin{aligned}
(Q_{1,k}^T d_k)^T \Lambda_{1,k} Q_{1,k}^T d_k &\leq (Q_{1,k}^T d_k)^T (\Lambda_{1,k} + \mu_k I) Q_{1,k}^T d_k \\
&= -(Q_{1,k}^T d_k)^T (Q_{1,k}^T g_k) + (Q_{1,k}^T d_k)^T (Q_{1,k}^T r_k) \\
&\leq \|Q_{1,k}^T d_k\| (\|Q_{1,k}^T g_k\| + \|Q_{1,k}^T r_k\|) \\
&\leq \|Q_{1,k}^T d_k\| (O(\|x_k - \hat{x}_k\|) + \alpha \mu_k \|d_k\|) \\
&\leq \|Q_{1,k}^T d_k\| (O(\|x_k - \hat{x}_k\|) + \alpha \mu_k \|Q_{1,k}^T d_k\| + \alpha \mu_k \|Q_{2,k}^T d_k\|),
\end{aligned}$$

where the third inequality follows from $\|Q_{1,k}^T g_k\| = \|Q_{1,k}^T(g_k - \hat{g}_k)\| = O(\|x_k - \hat{x}_k\|)$ and $\|r_k\| \leq \alpha\mu_k\|d_k\|$. Since $\{\Lambda_{1,k}\}$ is uniformly positive definite and $\mu_k = O(\mathrm{dist}(x_k, X))$, it is easy to deduce from the last inequality that

$$\|Q_{1,k}^T d_k\| = O(\|x_k - \hat{x}_k\|) + o(\|Q_{2,k}^T d_k\|). \tag{3.14}$$

On the other hand, we get from (3.12) and $\|r_k\| \leq \alpha\mu_k\|d_k\|$ that

$$\begin{aligned} \|Q_{2,k}^T d_k\| &\leq O(\|x_k - \hat{x}_k\|) + \alpha\|d_k\| \\ &\leq O(\|x_k - \hat{x}_k\|) + \alpha(\|Q_{1,k}^T d_k\| + \|Q_{2,k}^T d_k\|), \end{aligned}$$

which implies

$$\|Q_{2,k}^T d_k\| \leq \frac{\alpha}{1-\alpha}\|Q_{1,k}^T d_k\| + O(\|x_k - \hat{x}_k\|). \tag{3.15}$$

Consequently, we obtain (3.7) from (3.14) and (3.15). That is, statement (ii) holds. Since $\|r_k\| \leq \alpha\mu_k\|d_k\| = o(\|x_k - \hat{x}_k\|)$, statement (i) then follows from (3.9). The proof is complete. Ξ

The following theorem shows quadratic convergence of IRM I and IRM II.

Theorem 3.2 *Suppose that $\|g(x)\|$ provides a local error bound near $\bar{x} \in X$. Let $\{x_k\}$ be generated by the regularized Newton method (3.1) or IRM I or IRM II, in which x_{k+1} is determined by $x_{k+1} := x_k + d_k$. If the initial point x_0 is chosen sufficiently close to $\bar{x}$, then the entire sequence $\{x_k\}$ converges to a solution of problem (1.2). Moreover, the rate of convergence is quadratic.*

Proof Taking into account (3.7), it is easy to see that for either IRM I or IRM II, we always have $r_k = O(\mathrm{dist}(x_k, X)^2)$. Therefore, we have from (1.5) with $F(x) = \|g(x)\|$

$$\begin{aligned} m_1\,\mathrm{dist}(x_k + d_k, X) &\leq \|g(x_k + d_k)\| \\ &\leq \|g(x_k + d_k) - g_k - G_k d_k\| + \mu_k\|d_k\| + \|r_k\| \\ &\leq O(\|d_k\|^2) + O(\mathrm{dist}(x_k, X))\|d_k\| + O(\mathrm{dist}(x_k, X)^2) \\ &= O(\mathrm{dist}(x_k, X)^2), \end{aligned}$$

where the second inequality follows from (3.5) or (3.6). It then follows from Lemma 3.1 that $\{\|d_k\|\}$ converges to 0 quadratically, which in turn implies that it converges to 0 linearly, i.e., there exist a constant $\kappa \in (0,1)$ and a positive integer $\bar{k}$ such that for all $i \geq \bar{k}$

$$\|d_{i+1}\| \leq \kappa\|d_i\|.$$

Let $l > k \geq \bar{k}$. Then we have

$$\begin{aligned} \|x_l - x_k\| &= \|\sum_{i=k}^{l-1} d_i\| \\ &\leq \sum_{i=k}^{l-1} \|d_i\| \\ &\leq \|d_k\| \sum_{i=1}^{l-k} \kappa^{i-1} \\ &\leq \|d_k\|/(1-\kappa). \end{aligned}$$

Since $\{\|d_k\|\}$ tends to zero, the above inequalities show that $\{x_k\}$ is a Cauchy sequence, and hence converges.

Next, we investigate the rate of convergence. Let x^* be the limit point of $\{x_k\}$. Noting that

$$x^* = x_k + \sum_{i=k}^{\infty} d_i,$$

we have

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} = \frac{\|\sum_{i=k+1}^{\infty} d_i\|}{\|\sum_{i=k}^{\infty} d_i\|^2}. \tag{3.16}$$

Since $\{\|d_k\|\}$ converges to 0 quadratically, there exist $c_1 \in (0,1)$ and a positive integer $\bar{k}$ such that, for $k \geq \bar{k}$

$$c_1\|d_k\| \geq \Big\| \sum_{i=k+1}^{\infty} d_i \Big\|.$$

It then follows from the triangle inequality that, for $k \geq \bar{k}$

$$\Big\|\sum_{i=k}^{\infty} d_i\Big\| \geq \|d_k\| - \Big\| \sum_{i=k+1}^{\infty} d_i\Big\| \geq (1-c_1)\|d_k\| \tag{3.17}$$

and

$$\Big\| \sum_{i=k+1}^{\infty} d_i\Big\| \leq \|d_{k+1}\| + \Big\| \sum_{i=k+2}^{\infty} d_i\Big\| \leq (1+c_1)\|d_{k+1}\|. \tag{3.18}$$

Substituting (3.17) and (3.18) into (3.16), we have

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq \frac{(1+c_2)\|d_{k+1}\|}{(1-c_1)^2\|d_k\|^2}$$

for $k \geq \bar{k}$. Since $\{\|d_k\|\}$ converges to 0 quadratically, we get

$$\limsup_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} < +\infty,$$

that is, $\{x_k\}$ converges to x^* quadratically. Ξ

We note that, in the above theorem, the limit point x^* of the generated sequence $\{x_k\}$ need not coincide with the solution $\bar{x} \in X$ that appears in the assumption.

4 A Globally Convergent Inexact Regularized Newton Method

In this section, we develop a globally convergent inexact regularized Newton method. First, we point out that there should be no difficulty in globalizing IRM I in a way similar to Dan, Yamashita and Fukushima [1]. The method in [1] is a hybrid method that combines an inexact Newton method with a gradient method. In this section, we globalize IRM II that is a "pure" inexact regularized Newton method.

First, the following lemma shows that any vector d_k satisfying (3.6) provides a descent direction of f at x_k.

Lemma 4.1 *Any vector d_k satisfying (3.6) is a descent direction of f at x_k and satisfies*

$$g_k^T d_k \le -(1-\alpha)\mu_k \|d_k\|^2. \tag{4.1}$$

Proof By direct computation, we have

$$g_k^T d_k = -d_k^T (G_k + \mu_k I) d_k + d_k^T r_k \le -\mu_k \|d_k\|^2 + \alpha\mu_k \|d_k\|^2.$$

This implies (4.1). Ξ

Since d_k is a descent direction of f at x_k, we can use a backtracking line search to globalize IRM II. We state the steps of the globalized inexact regularized Newton method as follows.

Globalized Inexact Regularized Newton Method (GIRM):

Initialization: Choose constants $\alpha, \sigma, \rho \in (0,1)$ and $C > 0$. Select an initial point $x_0 \in R^n$. Let $k := 0$.

Step 1: Choose μ_k as in (3.3). Find a vector d_k satisfying (3.6). That is, d_k is a solution of the following inexact Newton subproblem:

$$\|(G_k + \mu_k I)d + g_k\| \le \alpha\mu_k \|d_k\|. \tag{4.2}$$

Step 2: Find the smallest nonnegative integer i such that

$$f(x_k + \rho^i d_k) \le f(x_k) + \sigma\rho^i g_k^T d_k. \tag{4.3}$$

Let $\lambda_k := \rho^i$.

Step 3: Let $x_{k+1} := x_k + \lambda_k d_k$ and $k := k+1$. Go to Step 1.

The next lemma shows the boundedness of vectors d_k obtained from (4.2).

Lemma 4.2 *The sequence $\{d_k\}$ generated by GIRNM is bounded.*

Proof It follows from (3.6) that

$$\|d_k\| \le \|(G_k + \mu_k I)^{-1} g_k\| + \|(G_k + \mu_k I)^{-1} r_k\| \le \mu_k^{-1}(\|g_k\| + \alpha\mu_k \|d_k\|),$$

which implies

$$(1-\alpha)\|d_k\| \le \mu_k^{-1}\|g_k\|.$$

Since $\alpha \in (0,1)$ and μ_k satisfies (3.3), the last inequality shows that $\{d_k\}$ is bounded. Ξ

Global convergence of GIRM is established in the following theorem.

Theorem 4.3 *Let $\{x_k\}$ be generated by GIRM. Suppose that f is bounded from below. Then every accumulation point of $\{x_k\}$ is a solution of problem (1.2).*

Proof It is easy to deduce from (4.3) that

$$\lim_{k\to\infty} \lambda_k g_k^T d_k = 0. \tag{4.4}$$

Let $\{x_k\}_{k\in K}$ converge to some point $\tilde{x}$. By Lemma 4.2, $\{d_k\}$ is bounded. Without loss of generality, we suppose that $\{d_k\}_{k\in K}$ converges to some vector $\tilde{d}$. Let $\tilde{\lambda} = \limsup_{k\in K, k\to\infty} \lambda_k$. Then $\tilde{\lambda} \ge 0$.

If $\tilde{\lambda} > 0$, then we get from (4.4) that

$$g(\tilde{x})^T \tilde{d} = 0. \tag{4.5}$$

Otherwise, we have $\lim_{k\in K, k\to\infty} \lambda_k = 0$. By the line search rule, we have for all $k \in K$ sufficiently large,

$$f(x_k + \lambda_k' d_k) > f(x_k) + \sigma \lambda_k' g_k^T d_k,$$

where $\lambda_k' = \lambda_k/\rho$. Dividing both sides by λ_k' and then taking limits as $k \to \infty$ with $k \in K$, we get

$$g(\tilde{x})^T \tilde{d} \ge \sigma g(\tilde{x})^T \tilde{d}.$$

Since $\sigma \in (0,1)$ and $g(\tilde{x})^T \tilde{d} \le 0$ by (4.1), the last inequality also implies (4.5).

Let $\tilde{\mu} = \lim_{k\in K, k\to\infty} \mu_k$. Then we have $\tilde{\mu} = C\|g(\tilde{x})\|$ by (3.3). Taking limits in (4.1) as $k \to \infty$ with $k \in K$, we deduce $\tilde{\mu}\|\tilde{d}\|^2 = 0$. So, we have either $\tilde{\mu} = 0$ or $\tilde{d} = 0$. If $\tilde{\mu} = 0$, we obviously have $g(\tilde{x}) = 0$, which shows that $\tilde{x}$ is a stationary point of (1.2). Consequently, the convexity of f implies that it is a solution of (1.2). If $\tilde{d} = 0$, it is also not difficult to get $g(\tilde{x}) = 0$ by taking limits in (3.6). The proof is then complete. Ξ

The next theorem shows that the unit stepsize is accepted for all k sufficiently large, which along with Theorem 3.2 establishes quadratic convergence of GIRM.

Theorem 4.4 *Let $\{x_k\}$ be generated by GIRM with $\sigma \in (0,1/2)$ and $\{x_k\}_{k\in K}$ be an arbitrary convergent subsequence whose limit is $\tilde{x} \in X$. If $\|g(x)\|$ provides a local error bound near $\tilde{x}$, then, for all $k \in K$ sufficiently large, $i = 0$ satisfies (4.3). Consequently, we have $\lambda_k = 1$ for all $k \in K$ sufficiently large.*

Proof It follows from Lemma 3.1 (ii) that when $k \in K$ is sufficiently large,

$$\|d_k\| = O(\mathrm{dist}(x_k, X)). \tag{4.6}$$

By Taylor's expansion, we have for all $k \in K$

$$\begin{aligned} f(x_k + d_k) - f(x_k) - \tfrac{1}{2} g_k^T d_k &= \tfrac{1}{2} d_k^T (G_k d_k + g_k) + o(\|d_k\|^2) \\ &= \tfrac{1}{2} d_k^T ((G_k + \mu_k I) d_k + g_k) - \tfrac{1}{2} \mu_k \|d_k\|^2 + o(\|d_k\|^2) \\ &\leq -\tfrac{1}{2}(1-\alpha)\mu_k \|d_k\|^2 + o(\|d_k\|^2), \end{aligned} \tag{4.7}$$

where the inequality follows from Cauchy-Schwarz inequality and (4.2). Since $\mu_k = C\|g_k\| \to 0$, it follows from (4.6) and (4.7) that

$$f(x_k + d_k) - f(x_k) - \tfrac{1}{2} g_k^T d_k = o(\mathrm{dist}(x_k, X)^2). \tag{4.8}$$

On the other hand, we have

$$\begin{aligned} -g_k^T d_k &= d_k^T (G_k + \mu_k I) d_k - d_k^T r_k \\ &= d_k^T G_k d_k + o(\|d_k\|^2) \\ &= (Q_{1,k}^T d_k)^T \Lambda_{1,k} (Q_{1,k}^T d_k) + (Q_{2,k}^T d_k)^T \Lambda_{2,k} (Q_{2,k}^T d_k) + o(\mathrm{dist}(x_k, X)^2), \end{aligned} \tag{4.9}$$

where the second inequality holds because $\mu_k \to 0$ and $\|r_k\| \leq \alpha \mu_k \|d_k\| = o(\|d_k\|)$. Since $\Lambda_{1,k} \geq \frac{1}{2}\tilde{\Lambda} > 0$ and $\Lambda_{2,k} \to 0$ as $k \to \infty$ with $k \in K$, (4.9) together with Lemma 3.1 (i) and (4.6) implies that there is a constant $c > 0$ such that

$$-g_k^T d_k \geq c\|Q_{1,k}^T d_k\|^2 \geq c m_2 \,\mathrm{dist}(x_k, X)^2. \tag{4.10}$$

Since $\sigma \in (0, 1/2)$, it follows from (4.8) and (4.10) that when $k \in K$ is sufficiently large,

$$f(x_k + d_k) \leq f(x_k) + \sigma g_k^T d_k.$$

This means that $i = 0$ satisfies (4.3). In other words, the unit stepsize is accepted. Ξ

5 Concluding Remark

We have developed regularized Newton and inexact Newton methods for convex minimization problems. We have shown that the methods are quadratically convergent without the requirement of isolated nonsingular solutions. We have also globalized an inexact regularized Newton method. The globalized method possesses global and quadratic convergence. It is possible to extend the methods to nonconvex minimization problems. For example, we can choose μ_k to satisfy

$$\mu_k = \max\{C\|g(x_k)\|, -2\lambda_1(G_k)\},$$

where $C > 0$ is a constant and $\lambda_1(G_k)$ is the minimum eigenvalue of G_k. It is not difficult to establish local quadratic convergence of the corresponding methods in a way similar to that in Section 3. However, the choice of μ_k depends on the computation of the minimum eigenvalue of $\lambda_1(G_k)$. Developing practicable Newton-like methods for nonconvex minimization problems with singular solutions will be a future research topic.

References

[1] H. Dan, N. Yamashita and M. Fukushima, Convergence properties of the inexact Levenberg-Marquardt method under local error bound conditions, Technical Report 2001-001, Department of Applied Mathematics and Physics, Kyoto University (January 2001).

[2] R.S. Dembo, S.C. Eisenstat and T. Steihaug, Inexact Newton method, *SIAM Journal on Numerical Analysis*, 19 (1982), pp. 400–408.

[3] J.Y. Fan and Y.X. Yuan, On the convergence of a new Levenberg-Marquardt method, Technical Report, AMSS, Chinese Academy of Sciences, 2001.

[4] G.W. Stewart and J.G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, NY, 1990.

[5] N. Yamashita and M. Fukushima, On the rate of convergence of the Levenberg-Marquardt method, *Computing*, 15 [Suppl] (2001), pp. 239–249.