Comput Optim Appl (2008) 39: 161–181
DOI 10.1007/s10589-007-9067-3

Self-adaptive inexact proximal point methods

William W. Hager · Hongchao Zhang

Received: 23 June 2006 / Revised: 23 June 2006 / Published online: 21 September 2007

Abstract We propose a class of self-adaptive proximal point methods suitable for degenerate optimization problems where multiple minimizers may exist, or where the Hessian may be singular at a local minimizer. If the proximal regularization parameter has the form $\mu(\mathbf{x}) = \beta\|\nabla f(\mathbf{x})\|^\eta$ where $\eta \in [0, 2)$ and $\beta > 0$ is a constant, we obtain convergence to the set of minimizers that is linear for $\eta = 0$ and β sufficiently small, superlinear for $\eta \in (0, 1)$, and at least quadratic for $\eta \in [1, 2)$. Two different acceptance criteria for an approximate solution to the proximal problem are analyzed. These criteria are expressed in terms of the gradient of the proximal function, the gradient of the original function, and the iteration difference. With either acceptance criterion, the convergence results are analogous to those of the exact iterates. Preliminary numerical results are presented using some ill-conditioned CUTE test problems.

Keywords Proximal point · Degenerate optimization · Multiple minima · Self-adaptive method

1 Introduction

We consider the unconstrained optimization problem

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

This material is based upon work supported by the National Science Foundation under Grant Nos. 0203270, 0619080, and 0620286.

W.W. Hager (✉)
Department of Mathematics, University of Florida, P.O. Box 118105, Gainesville, FL 32611-8105, USA
e-mail: hager@math.ufl.edu

H. Zhang
Institute for Mathematics and Its Applications (IMA), University of Minnesota, 400 Lind Hall, 207 Church Street S.E., Minneapolis, MN 55455-0436, USA
e-mail: hozhang@ima.umn.edu

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuous and the set of minimizers for (1.1), denoted $\mathbf{X}$, is nonempty. The convergence rate of algorithms for solving an unconstrained optimization problem often depends on the eigenvalues of the Hessian matrix at a local minimizer. As the ratio between largest and smallest eigenvalues grows, convergence rates can degrade. In this paper, we consider ill conditioned problems where the smallest eigenvalue of the Hessian can be zero, and where $\mathbf{X}$ may have more than one element.

The proximal point method is one strategy for dealing with degeneracy at a minimum. The iterates $\mathbf{x}_k$, $k \geq 1$, are generated by the rule:

$$\mathbf{x}_{k+1} \in \arg\min\{F_k(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}, \tag{1.2}$$

where

$$F_k(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}\mu_k \|\mathbf{x} - \mathbf{x}_k\|^2.$$

Here $\mathbf{x}_0 \in \mathbb{R}^n$ is an initial guess for a minimizer, $\|\cdot\|$ is the Euclidean norm, and the parameters μ_k, $k \geq 0$, are positive scalars. Due to the quadratic term, F_k is strictly convex at a local minimizer. Hence, the proximal point method improves the conditioning at the expense of replacing the single minimization (1.1) by a sequence of minimizations (1.2).

The proximal point method, first proposed by Martinet [13, 14], has been studied in many papers including [4, 10, 12, 16, 17]. In [17] Rockafellar shows that if f is strongly convex at a solution of (1.1), then the proximal point method converges linearly when μ_k is bounded away from zero, and superlinearly when μ_k tends to zero. Here we develop linear and superlinear convergence results for problems that are not necessarily strongly convex.

Since the solution to (1.2) approximates a solution to (1.1), we do not need to solve (1.2) exactly. We analyze two criteria for the accuracy with which we solve (1.2). The first criterion is that an iterate $\mathbf{x}_{k+1}$ is acceptable when

$$\text{(C1)} \quad F_k(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \quad \text{and} \quad \|\nabla F_k(\mathbf{x}_{k+1})\| \leq \mu_k \|\nabla f(\mathbf{x}_k)\|.$$

The second acceptance criterion is

$$\text{(C2)} \quad F_k(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \quad \text{and} \quad \|\nabla F_k(\mathbf{x}_{k+1})\| \leq \theta\mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|,$$

where $\theta < 1/\sqrt{2}$. In either case, we show, for μ_k sufficiently small, that

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq C\mu_k \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|, \tag{1.3}$$

where C is a constant depending on local properties of f and $\bar{\mathbf{x}}$ is any element of $\mathbf{X}$ for which

$$\|\bar{\mathbf{x}} - \mathbf{x}\| = \min_{\chi \in \mathbf{X}} \|\chi - \mathbf{x}\|.$$

Since f is continuous, the set of minimizers $\mathbf{X}$ of (1.1) is closed, and $\bar{\mathbf{x}}$ exists. By taking $\mu_k = \|\nabla f(\mathbf{x}_k)\|$ in (1.3), we obtain quadratic convergence of the approximate proximal iterates to the solution set $\mathbf{X}$, while the sequence of iterates approaches a limit at least linearly.

In [17] Rockafellar studies the acceptance condition

$$\|\nabla F_k(\mathbf{x}_{k+1})\| \leq \epsilon_k \mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|,$$

where $\sum_k \epsilon_k < \infty$ (see his condition B′). Our criterion (C2) corresponds to the case $\sum_k \epsilon_k = \infty$. (C2) is also studied in [8] as well as a criterion similar to (C1) except that the μ_k factor is removed. The authors in [8] give an introduction to proximal point algorithm, borrowing ideas from descent methods for unconstrained optimization. In [8] global convergence for convex functions is established, while here we obtain local convergence rates for general nonlinear functions.

Our analysis of the approximate proximal iterates makes use of a local error bound condition employed when the Hessian is singular at a minimizer of f—see [3, 18, 21, 22]. Referring to [3] and [22], we have the following terminology: ∇f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ if there exist positive constants α and ρ such that

$$\|\nabla f(\mathbf{x})\| \geq \alpha \|\mathbf{x} - \bar{\mathbf{x}}\| \quad \text{whenever} \quad \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \rho. \tag{1.4}$$

Using this condition, Yamashita and Fukushima [22], and Fan and Yuan [3], study the Levenberg-Marquardt method to solve a system of nonlinear equations. When their approach is applied to (1.1), the following linear system is solved in each subproblem:

$$(\mathbf{H}(\mathbf{x}_k)^2 + \mu_k \mathbf{I})\mathbf{d} + \mathbf{H}(\mathbf{x}_k)\mathbf{g}(\mathbf{x}_k) = \mathbf{0}, \tag{1.5}$$

where $\mu_k > 0$ is the regularization parameter, $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})^{\mathsf{T}}$ is the gradient (a column vector), $\mathbf{H}(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ is the Hessian, and $\mathbf{d}$ is the search direction at step k. In [3] and [22], the authors choose $\mu_k = \|\mathbf{g}(\mathbf{x}_k)\|$ and $\mu_k = \|\mathbf{g}(\mathbf{x}_k)\|^2$, respectively. They show that if $\nabla f(\mathbf{x})$ provided a local error bound, then the iterates associated with (1.5) are locally quadratically convergent.

Li, Fukushima, Qi and Yamashita point out in [11] that the linear system of equations (1.5) may lose sparsity when $\mathbf{H}(\mathbf{x}_k)$ is squared; moreover, squaring the matrix squares the condition number of $\mathbf{H}(\mathbf{x}_k)$. Hence, they consider a search direction $\mathbf{d}$ chosen to satisfy:

$$(\mathbf{H}(\mathbf{x}_k) + \mu_k \mathbf{I})\mathbf{d} + \mathbf{g}(\mathbf{x}_k) = \mathbf{0}, \tag{1.6}$$

where $\mu_k = c\|\mathbf{g}(\mathbf{x}_k)\|$ for some constant $c > 0$. When $f(\mathbf{x})$ is convex and $\nabla f(\mathbf{x})$ provides a local error bound, they establish a local quadratic convergence result for iterates generated by the approximate solution of (1.6) followed by a line search along the approximate search direction.

When the problem dimension is large, computing the Hessian and solving (1.6) can be expensive. In this paper, we consider an algorithm which employs an iterative method to approximately solve (1.2). The iteration is stopped whenever criterion (C1) or (C2) is satisfied. As a specific example, we give numerical results based on our newly developed conjugate gradient code CG_DESCENT [5–7]. One advantage of the conjugate gradient approach is that there is no need to evaluate the Hessian. In the case where the Hessian is easy to evaluate, either the update (1.6) or the conjugate gradient strategy might be faster, depending on the nonzero structure of the Hessian matrix and the distribution of its eigenvalues.

We also point out the related papers [2, 9, 15] in which the authors relax some of the assumptions underlying the convergence of the proximal point method. In these papers, convergence results are given for an operator whose inverse is hypomonotone; that is, monotone when a multiple of the identity is added. In [15] results are obtained for an exact algorithm in which each iterate is gotten by a step of the proximal point method; in [9], the authors analyze an algorithm in which an approximate proximal iteration $\mathbf{y}_k$ is computed, and $\mathbf{x}_{k+1}$ is obtained by subtracting from $\mathbf{x}_k$ a multiple of the gradient at $\mathbf{y}_k$. In [2] the authors generalize the results in [9, 15] by considering the problem of finding a common zero of countably many cohypomonotone operators. In our paper, we focus on the special case where the operator is the derivative of a function.

We now compare in more detail the difference between our results and results in [9, 15], in the special case where the operator is the derivative of a function. The exact algorithm analyzed in [15] should be compared to our corresponding analysis in Theorem 3.1. Neither smoothness nor convexity assumptions are used in either result. The principal assumption in our analysis is that the function provides a local error bound. In [15], it is assumed that a Yosida regularization of the derivative is maximal monotone. As shown in Proposition 2 of [9], this is equivalent to the hypomonotonicity of the inverse of the derivative of f.

In [15] linear convergence is established either when the stationary point is locally unique and the local error bound condition holds, or when the inverse of the derivative has a Lipschitz localization. This later condition implies that the inverse is locally single valued and Lipschitz. In our Theorem 3.1, the assumption that the function provides a local error bound allows us to consider a solution set and a derivative whose inverse is not locally unique.

In [9] the authors consider a method which is different, but related to the method analyzed in our paper; after computing an approximate proximal point $\mathbf{y}_k$, the authors apply the update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{\mu_k}\mathbf{g}(\mathbf{y_k}).$$

For accuracy criteria in the spirit of our (C1) and (C2), convergence is established when the inverse of the derivative of f is maximal hypomonotone. With the additional assumption that the inverse of the derivative is Lipschitz continuous (which implies that the inverse is single-valued), convergence rate results are established. For their criterion similar to (C1), the authors establish an error decay factor proportional to $\sqrt{\mu_k}$, while for the direct proximal point method analyzed in this paper ($\mathbf{x}_{k+1} = \mathbf{y}_k$), the decay factor is proportional to μ_k. For a criterion in the spirit of (C2), the authors in [9] establish linear convergence, independent of μ_k, while the error decay factor for our proximal step satisfying (C2) is again proportional to μ_k. Thus we obtain stronger convergence rate results than those obtained in [9]. On the other hand, we assume that f is locally convex and twice continuously differentiable, which implies the hypomonotonicity condition in [9]. Our convergence rate results apply to problems with multiple solutions, while [9] assumes that the inverse of the derivative is single-valued.

Our paper is organized as follows: In Sect. 2 we develop an equivalent formulation of the local error bound condition. In Sect. 3 the exact iteration (1.2) is analyzed,

while Sect. 4 analyzes approximate iterates satisfying (C1) or (C2). Section 5 gives a global convergence result, and Sect. 6 gives preliminary numerical results using ill-conditioned problems from the CUTE library and a problem from [11].

1.1 Notation

Throughout this paper, we use the following notation. $\|\cdot\|$ is the Euclidean norm of a vector. The gradient $\nabla f(\mathbf{x})$ is a row vector while $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})^\mathsf{T}$ is a column vector; here $^\mathsf{T}$ denotes transpose. The gradient at the iterate $\mathbf{x}_k$ is $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$. We let $\nabla^2 f(\mathbf{x})$ denote the Hessian of f at $\mathbf{x}$. The ball with center $\mathbf{x}$ and radius ρ is denoted $\mathcal{B}_\rho(\mathbf{x})$.

2 Preliminaries

Since the minimum of F_k is bounded from above by $F_k(\bar{\mathbf{x}}_k)$, we should restrict the minimization in (1.2) to those $\mathbf{x}$ satisfying:

$$\begin{aligned} F_k(\mathbf{x}) &= f(\mathbf{x}) + \frac{1}{2}\mu_k\|\mathbf{x}-\mathbf{x}_k\|^2 \le F_k(\bar{\mathbf{x}}_k) = f(\bar{\mathbf{x}}_k) + \frac{1}{2}\mu_k\|\bar{\mathbf{x}}_k-\mathbf{x}_k\|^2 \\ &\le f(\mathbf{x}) + \frac{1}{2}\mu_k\|\bar{\mathbf{x}}_k-\mathbf{x}_k\|^2. \end{aligned}$$

It follows that

$$\|\mathbf{x}-\mathbf{x}_k\| \le \|\bar{\mathbf{x}}_k-\mathbf{x}_k\|. \tag{2.1}$$

The $\mathbf{x}$ satisfying (2.1) forms a bounded set. By continuity of f, the set of minimizers of F_k is nonempty, and $\mathbf{x}_{k+1}$ in (1.2) exists. These observations are summarized as follows:

Proposition 2.1 *If f is continuous and its set of minimizers $\mathbf{X}$ is nonempty, then for each k, F_k has a minimizer $\mathbf{x}_{k+1}$ and we have*

$$\|\mathbf{x}_{k+1}-\mathbf{x}_k\| \le \|\bar{\mathbf{x}}_k-\mathbf{x}_k\|.$$

We let $\mathcal{B}_\rho(\mathbf{y})$ denote the ball with center $\mathbf{y}$ and radius ρ. With this notation, Proposition 2.1 can be stated

$$\mathbf{x}_{k+1} \in \mathcal{B}_\rho(\mathbf{x}_k), \qquad \text{where } \rho = \|\bar{\mathbf{x}}_k-\mathbf{x}_k\|.$$

In this paper, it is more convenient to employ a local error bound based on the function value rather than the function gradient used in (1.4). We say that f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ if there exist positive constants α and ρ such that

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) \ge \alpha\|\bar{\mathbf{x}}-\mathbf{x}\|^2 \quad \text{whenever} \quad \|\mathbf{x}-\hat{\mathbf{x}}\| \le \rho. \tag{2.2}$$

We now show that these two conditions are equivalent:

Lemma 2.2 *If f is twice continuously differentiable in a neighborhood of $\hat{\mathbf{x}} \in \mathbf{X}$, then f provides a local error bound at $\hat{\mathbf{x}}$ is equivalent to ∇f provides a local error bound at $\hat{\mathbf{x}}$.*

Proof Suppose f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying (2.2). Choose ρ smaller if necessary so that f is twice continuously differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$. Define $r = \rho/2$. Given $\mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}})$, the triangle inequality implies that

$$\|\bar{\mathbf{x}} - \hat{\mathbf{x}}\| \le \|\bar{\mathbf{x}} - \mathbf{x}\| + \|\mathbf{x} - \hat{\mathbf{x}}\| \le 2r = \rho. \tag{2.3}$$

Since both $\mathbf{x}$ and $\bar{\mathbf{x}} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$, we can expand $f(\mathbf{x})$ in a Taylor series around $\bar{\mathbf{x}}$ and apply (2.2) to obtain:

$$\begin{aligned} \alpha\|\mathbf{x} - \bar{\mathbf{x}}\|^2 &\le f(\mathbf{x}) - f(\bar{\mathbf{x}}) \\ &= \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^{\mathsf{T}}\bar{\mathbf{H}}(\mathbf{x} - \bar{\mathbf{x}}) + R_2(\mathbf{x}, \bar{\mathbf{x}}), \end{aligned} \tag{2.4}$$

where R_2 is the remainder term and $\bar{\mathbf{H}} = \nabla^2 f(\bar{\mathbf{x}})$ is the Hessian at $\bar{\mathbf{x}}$. Choose ρ small enough that

$$|R_2(\mathbf{x}, \bar{\mathbf{x}})| \le \frac{\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\|^2 \quad \text{whenever} \quad \mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}}).$$

In this case, (2.4) gives

$$\frac{4\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\|^2 \le (\mathbf{x} - \bar{\mathbf{x}})^{\mathsf{T}}\bar{\mathbf{H}}(\mathbf{x} - \bar{\mathbf{x}}) \le \|\mathbf{x} - \bar{\mathbf{x}}\|\|\bar{\mathbf{H}}(\mathbf{x} - \bar{\mathbf{x}})\|.$$

Hence, we have

$$\frac{4\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\| \le \|\bar{\mathbf{H}}(\mathbf{x} - \bar{\mathbf{x}})\|. \tag{2.5}$$

Now expand $\nabla f(\mathbf{x})$ in a Taylor series around $\bar{\mathbf{x}}$ to obtain

$$\nabla f(\mathbf{x}) = \nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}}) = \bar{\mathbf{H}}(\mathbf{x} - \bar{\mathbf{x}}) + \mathbf{R}_1(\mathbf{x}, \bar{\mathbf{x}}), \tag{2.6}$$

where $\mathbf{R}_1$ is the remainder term. Choose ρ smaller if necessary so that

$$\|R_1(\mathbf{x}, \bar{\mathbf{x}})\| \le \frac{\alpha}{3}\|\mathbf{x} - \bar{\mathbf{x}}\| \quad \text{whenever} \quad \mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}}). \tag{2.7}$$

Combining (2.5–2.7) yields

$$\|\nabla f(\mathbf{x})\| \ge \alpha\|\mathbf{x} - \bar{\mathbf{x}}\|.$$

Hence, ∇f provides a local error bound at $\hat{\mathbf{x}}$ with constants α and $\rho/2$.

Conversely, suppose ∇f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying (1.4). Choose ρ smaller if necessary so that f is twice continuously differentiable in $\mathcal{B}_\rho(\hat{\mathbf{x}})$ and (2.7) holds. Combining the Taylor expansion (2.6),

the fact that $\bar{\mathbf{H}}$ is positive semidefinite, the bound (2.7) on the remainder, and the local error bound condition (1.4), we obtain

$$\frac{2\alpha}{3}\|\mathbf{x}-\bar{\mathbf{x}}\| \le \|\bar{\mathbf{H}}(\mathbf{x}-\bar{\mathbf{x}})\| \le \|\bar{\mathbf{H}}^{1/2}\|\|\bar{\mathbf{H}}^{1/2}(\mathbf{x}-\bar{\mathbf{x}})\|.$$

Squaring both sides gives

$$\frac{4\alpha^2}{9}\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \le \|\bar{\mathbf{H}}\|\|\bar{\mathbf{H}}^{1/2}(\mathbf{x}-\bar{\mathbf{x}})\|^2 = \|\bar{\mathbf{H}}\|(\mathbf{x}-\bar{\mathbf{x}})^{\mathsf{T}}\bar{\mathbf{H}}(\mathbf{x}-\bar{\mathbf{x}}).$$

Consequently,

$$(\mathbf{x}-\bar{\mathbf{x}})^{\mathsf{T}}\bar{\mathbf{H}}(\mathbf{x}-\bar{\mathbf{x}}) \ge \left(\frac{4\alpha^2}{9\|\bar{\mathbf{H}}\|}\right)\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \ge \left(\frac{4\alpha^2}{9\lambda}\right)\|\mathbf{x}-\bar{\mathbf{x}}\|^2, \tag{2.8}$$

where λ is a bound for the Hessian of f over $\mathcal{B}_\rho(\hat{\mathbf{x}})$. Similar to (2.4), we expand f in a Taylor expansion around $\bar{\mathbf{x}}$ and utilize (2.8) to obtain

$$f(\mathbf{x}) - f(\bar{\mathbf{x}}) - R_2(\mathbf{x},\bar{\mathbf{x}}) = \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^{\mathsf{T}}\bar{\mathbf{H}}(\mathbf{x}-\bar{\mathbf{x}}) \ge \beta\|\mathbf{x}-\bar{\mathbf{x}}\|^2, \quad \text{where } \beta = \frac{2\alpha^2}{9\lambda}.$$

To complete the proof, choose ρ smaller if necessary so that

$$|R_2(\mathbf{x},\bar{\mathbf{x}})| \le \frac{\beta}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2$$

whenever $\mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. □

3 Convergence analysis for exact minimization

We first analyze the proximal point method when the iterates are exact solutions of (1.2).

Theorem 3.1 *Assume the following conditions are satisfied*:

(E1) f *provides a local error bound at* $\hat{\mathbf{x}} \in \mathbf{X}$ *with positive scalars* α *and* ρ *satisfying* (2.2).
(E2) $\mu_k \le \beta$ *for each* k *where* $\beta < 2\alpha/3$.
(E3) $\mathbf{x}_0$ *is close enough to* $\hat{\mathbf{x}}$ *that*

$$\|\mathbf{x}_0-\hat{\mathbf{x}}\|\left(1+\frac{1}{1-\gamma}\right) \le \rho, \quad \text{where } \gamma = \frac{2\beta}{2\alpha-\beta}. \tag{3.1}$$

Then the proximal iterates $\mathbf{x}_k$ *are all contained in* $\mathcal{B}_\rho(\hat{\mathbf{x}})$ *and they approach a minimizer* $\mathbf{x}^* \in \mathbf{X}$; *for each* k, *we have*

$$\begin{aligned} \|\mathbf{x}_k-\mathbf{x}^*\| &\le \frac{\gamma^k}{1-\gamma}\|\bar{\mathbf{x}}_0-\mathbf{x}_0\| \quad \text{and} \\ \|\mathbf{x}_{k+1}-\bar{\mathbf{x}}_{k+1}\| &\le \frac{2\mu_k}{2\alpha-\mu_k}\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|. \end{aligned} \tag{3.2}$$

Proof For $j = 0$, we have

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \le \rho \quad \text{and} \quad \|\bar{\mathbf{x}}_j - \mathbf{x}_j\| \le \gamma^j \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|. \tag{3.3}$$

Proceeding by induction, suppose that (3.3) holds for all $j \in [0, k]$ and for some $k \ge 0$. We show that (3.3) holds for all $j \in [0, k+1]$. By Proposition 2.1 and the induction hypothesis,

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| \le \gamma^k \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|.$$

By (E2), we have $\beta < 2\alpha/3$ and hence,

$$\gamma = \frac{2\beta}{2\alpha - \beta} < 1.$$

By the triangle inequality, Proposition 2.1, and the induction hypothesis, it follows that

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_0\| &\le \sum_{j=0}^{k} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \le \sum_{j=0}^{k} \|\bar{\mathbf{x}}_j - \mathbf{x}_j\| \\ &\le \sum_{j=0}^{k} \gamma^j \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| \le \frac{1}{1-\gamma} \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| \le \frac{1}{1-\gamma} \|\mathbf{x}_0 - \hat{\mathbf{x}}\|.\end{aligned}$$

Again, by the triangle inequality and (3.1),

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \le \|\mathbf{x}_{k+1} - \mathbf{x}_0\| + \|\mathbf{x}_0 - \hat{\mathbf{x}}\| \le \left(1 + \frac{1}{1-\gamma}\right) \|\mathbf{x}_0 - \hat{\mathbf{x}}\| \le \rho. \tag{3.4}$$

Observe that

$$\begin{aligned}&\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_k\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad = (\bar{\mathbf{x}}_{k+1} + \mathbf{x}_{k+1} - 2\mathbf{x}_k)^{\mathsf{T}} (\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}) \\ &\quad \le (\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\| + 2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|) \|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\|.\end{aligned} \tag{3.5}$$

Combining (3.5) with the relation $F_k(\mathbf{x}_{k+1}) \le F_k(\bar{\mathbf{x}}_{k+1})$ gives

$$\begin{aligned}&f(\mathbf{x}_{k+1}) - f(\bar{\mathbf{x}}_{k+1}) \\ &\quad \le \frac{1}{2} \mu_k (\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_k\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2) \\ &\quad \le \frac{1}{2} \mu_k (\|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\| + 2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|) \|\bar{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1}\|.\end{aligned} \tag{3.6}$$

By (3.4), $\mathbf{x}_{k+1} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$. Since f provides a local error bound at $\hat{\mathbf{x}}$, we conclude that

$$\alpha \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 \le f(\mathbf{x}_{k+1}) - f(\bar{\mathbf{x}}_{k+1}). \tag{3.7}$$

Combining this with (3.6) gives

$$\left(\alpha - \frac{1}{2}\mu_k\right) \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \mu_k \|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{3.8}$$

Due to the assumption $\mu_k < 2\alpha/3$, the coefficient $(\alpha - \frac{1}{2}\mu_k)$ in (3.8) is positive and

$$2\mu_k/(2\alpha - \mu_k) \le 2\beta/(2\alpha - \beta) = \gamma < 1.$$

Hence, (3.8), Proposition 2.1, and (3.3), with $j = k$, give

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \left(\frac{2\mu_k}{2\alpha - \mu_k}\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \left(\frac{2\mu_k}{2\alpha - \mu_k}\right) \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| \tag{3.9}$$

$$\le \gamma \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| \le \gamma^{k+1} \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|. \tag{3.10}$$

Relations (3.4) and (3.10) complete the proof of the induction step. Relation (3.9) gives estimate (3.3).

By (3.3) and Proposition 2.1, we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| \le \gamma^k \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|.$$

Hence, the proximal iterates $\mathbf{x}_k$ form a Cauchy sequence, which has a limit denoted $\mathbf{x}^*$. Again, it follows from (3.3) and Proposition 2.1 that

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le \sum_{j=k}^{\infty} \|\mathbf{x}_{j+1} - \mathbf{x}_j\| \le \sum_{j=k}^{\infty} \|\bar{\mathbf{x}}_j - \mathbf{x}_j\|$$

$$\le \sum_{j=k}^{\infty} \gamma^j \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| = \frac{\gamma^k}{1-\gamma} \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|. \tag{3.11}$$

By (3.3) and (3.11), we have

$$\|\bar{\mathbf{x}}_k - \mathbf{x}^*\| \le \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| + \|\mathbf{x}_k - \mathbf{x}^*\| \le \left(\gamma^k + \frac{\gamma^k}{1-\gamma}\right) \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|.$$

Hence, $\bar{\mathbf{x}}_k$ approaches $\mathbf{x}^*$ as k tends to ∞. Since $\mathbf{X}$ is closed and $\bar{\mathbf{x}}_k \in \mathbf{X}$ for each k, it follows that the limit $\mathbf{x}^* \in \mathbf{X}$. □

Note that neither smoothness nor convexity assumptions enter into the convergence results of Theorem 3.1. In a further extension of these results, let us consider the case where the regularization sequence μ_k of Theorem 3.1 is expressed as a function of the current iterate. That is, we assume that $\mu_k = \mu(\mathbf{x}_k)$ where $\mu(\cdot)$ is defined on $\mathbb{R}^n$.

Corollary 3.2 *We make the following assumptions*:

(Q1) *f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying (2.2).*

(Q2) ∇f *is Lipschitz continuously differentiable in* $\mathcal{B}_\rho(\hat{\mathbf{x}})$ *with Lipschitz constant* L.
(Q3) ρ *is small enough that for some scalar* β, *we have*

$$\|\nabla f(\mathbf{x})\| \le \beta < \frac{2\alpha}{3} \quad \text{for all } \mathbf{x} \in \mathcal{B}_r(\hat{\mathbf{x}}) \text{ where } r = \rho/2.$$

(Q4) $\mathbf{x}_0$ *is close enough to* $\hat{\mathbf{x}}$ *that*

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \frac{1}{1-\gamma}\right) \le r, \quad \text{where } \gamma = \frac{2\beta}{2\alpha - \beta}.$$

Then for the choice $\mu(\mathbf{x}) = \|\nabla f(\mathbf{x})\|$ *and* $\mu_k = \mu(\mathbf{x}_k)$, *the proximal iterates* (1.2) *are all contained in* $\mathcal{B}_r(\hat{\mathbf{x}})$ *and they approach a minimizer of* f. *Moreover, we have*

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \left(\frac{3L}{2\alpha}\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \tag{3.12}$$

for each k.

Proof The proof is identical to the inductive proof of Theorem 3.1 except that we append the condition $\mu_j \le \beta$ for each $j \in [0, k]$ to the induction hypothesis (3.3). That is, we assume that for all $j \in [0, k]$, we have

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \le r, \qquad \|\bar{\mathbf{x}}_j - \mathbf{x}_j\| \le \gamma^j \|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|, \quad \text{and} \quad \mu_j \le \beta. \tag{3.13}$$

Since $\mathbf{x}_0 \in \mathcal{B}_r(\hat{\mathbf{x}})$, it follows by (Q3) that $\mu_0 = \|\nabla f(\mathbf{x}_0)\| \le \beta$. Hence, (3.13) is satisfied for $j = 0$. In the proof of Theorem 3.1, we show that if $\mu_j \le \beta$ for $j \in [0, k]$, then the first two conditions in (3.13) hold for $j = k + 1$. Also, as in (3.4), $\mathbf{x}_{k+1} \in \mathcal{B}_r(\hat{\mathbf{x}})$. Consequently, $\mu_{k+1} = \|\nabla f(\mathbf{x}_{k+1})\| \le \beta$, which implies that the last condition in (3.13) holds for $j = k + 1$. This completes the induction step; hence, (3.13) holds for all $j \ge 0$.

Since $\mathbf{x}_k \in \mathcal{B}_r(\hat{\mathbf{x}})$, it follows that $\bar{\mathbf{x}}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ (see (2.3)). Since $\mathbf{x}_k$ and $\bar{\mathbf{x}}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$, we have

$$\mu_k = \|\nabla f(\mathbf{x}_k)\| = \|\nabla f(\mathbf{x}_k) - \nabla f(\bar{\mathbf{x}}_k)\| \le L\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|, \tag{3.14}$$

where L is the Lipschitz constant for ∇f in $\mathcal{B}_\rho(\hat{\mathbf{x}})$. By estimate (3.3) in Theorem 3.1,

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \left(\frac{2\mu_k}{2\alpha - \mu_k}\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|.$$

Using the bound (3.14) in the numerator and the bound $\mu_k \le \beta < 2\alpha/3$ in the denominator, we obtain (3.12). □

4 Convergence analysis for approximate minimization

We now analyze the situation where the proximal point iteration (1.2) is implemented inexactly; the approximation to a solution of (1.2) need only satisfy (C1) or (C2). The following property of a convex function is used in the analysis.

Proposition 4.1 *If $\mathbf{x}^*$ is a local minimizer of F_k and f is convex and continuously differentiable in a convex neighborhood $\mathcal{N}$ of $\mathbf{x}^*$, then*

$$F_k(\mathbf{x}) \leq F_k(\mathbf{x}^*) + \frac{\|\nabla F_k(\mathbf{x})\|^2}{\mu_k}$$

for all $\mathbf{x} \in \mathcal{N}$.

Proof The convexity of f in $\mathcal{N}$ implies that F_k is convex in $\mathcal{N}$ and

$$F_k(\mathbf{x}^*) \geq F_k(\mathbf{x}) + \nabla F_k(\mathbf{x})(\mathbf{x}^* - \mathbf{x}).$$

Since $\mathbf{x}^*$ is a local minimizer of F_k, we have

$$\begin{aligned} \nabla F_k(\mathbf{x})(\mathbf{x} - \mathbf{x}^*) &= (\nabla F_k(\mathbf{x}) - \nabla F_k(\mathbf{x}^*))(\mathbf{x} - \mathbf{x}^*) \\ &= (\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*))(\mathbf{x} - \mathbf{x}^*) + \mu_k \|\mathbf{x} - \mathbf{x}^*\|^2. \end{aligned} \tag{4.1}$$

Since f is convex in $\mathcal{N}$, the monotonicity condition

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*))(\mathbf{x} - \mathbf{x}^*) \geq 0$$

holds. Combining this with (4.1), we obtain

$$\nabla F_k(\mathbf{x})(\mathbf{x} - \mathbf{x}^*) \geq \mu_k \|\mathbf{x} - \mathbf{x}^*\|^2. \tag{4.2}$$

By the Cauchy-Schwarz inequality,

$$\|\mathbf{x} - \mathbf{x}^*\| \leq \frac{\|\nabla F_k(\mathbf{x})\|}{\mu_k}. \tag{4.3}$$

Combining (4.2) and (4.3), the proof is complete. □

Our convergence result for the inexact proximal point iterates is now established.

Theorem 4.2 *Assume that the following conditions are satisfied*:

(A1) *f provides a local error bound at $\hat{\mathbf{x}} \in \mathbf{X}$ with positive scalars α and ρ satisfying (2.2).*

(A2) *f is twice continuously differentiable and convex throughout $\mathcal{B}_\rho(\hat{\mathbf{x}})$; let L be a Lipschitz constant for ∇f in $\mathcal{B}_\rho(\hat{\mathbf{x}})$.*

(A3) *The parameter $\beta = \sup\{\mu(\mathbf{x}) : \mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}})\}$ satisfies*

$$\beta < \alpha/\Lambda, \tag{4.4}$$

where

$$\Lambda = L + \tau \quad \textit{and} \quad \tau^2 = 1 + 2L^2 \quad \text{if acceptance criterion (C1) is used,}$$

while

$$\Lambda = \tau(1 + \theta) \quad \textit{and} \quad \tau^2 = \frac{1}{1 - 2\theta^2} \quad \text{if acceptance criterion (C2) is used.}$$

(A4) *The parameter*

$$\epsilon = \tau\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \frac{1}{1-\gamma}\right), \quad \textit{where } \gamma = \frac{\beta\Lambda}{\alpha},$$

satisfies

$$\epsilon + \sup\left\{\sqrt{\frac{\lambda\|\mathbf{x} - \bar{\mathbf{x}}\|^2}{2\mu(\mathbf{x})}} : \mathbf{x} \in \mathcal{B}_\epsilon(\hat{\mathbf{x}}), \mathbf{x} \notin \mathbf{X}\right\} \le r, \quad \textit{where } r = \rho/2,$$

and λ *is any upper bound for the largest eigenvalue of* $\nabla^2 f(\mathbf{x})$ *over* $\mathbf{x} \in \mathcal{B}_\rho(\hat{\mathbf{x}})$.

If the approximate proximal iterates $\mathbf{x}_k$ *satisfy either* (C1) *or* (C2) *with* $\mu_k = \mu(\mathbf{x}_k)$, *then the iterates are all contained in* $\mathcal{B}_\epsilon(\hat{\mathbf{x}})$, *they approach a minimizer* $\mathbf{x}^* \in \mathbf{X}$, *and for each* k, *we have*

$$\|\mathbf{x}_k - \mathbf{x}^*\| \le \frac{\tau\gamma^k}{1-\gamma}\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| \quad \textit{and} \tag{4.5}$$

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \left(\frac{\Lambda\mu_k}{\alpha}\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|, \tag{4.6}$$

where Λ *is defined in* (A3).

Proof The following relations hold trivially for $j = 0$ since the index range for the summation is vacuous and $\tau \ge 1$:

$$\|\mathbf{x}_j - \hat{\mathbf{x}}\| \le \tau\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \sum_{l=0}^{j-1}\gamma^l\right) \quad \text{and} \quad \|\bar{\mathbf{x}}_j - \mathbf{x}_j\| \le \gamma^j\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|. \tag{4.7}$$

Proceeding by induction, suppose that (4.7) holds for all $j \in [0, k]$ and for some $k \ge 0$. We show that (4.7) holds for all $j \in [0, k+1]$.

The condition $F_k(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k)$ in (C1) or (C2) implies that

$$\mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \le f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) - f(\bar{\mathbf{x}}_k). \tag{4.8}$$

Since $\gamma < 1$ by (4.4), the first half of (4.7) with $j = k$ implies that $\mathbf{x}_k \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$, where $\epsilon \le r = \rho/2$. Thus we have $\bar{\mathbf{x}}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ (see (2.3)). Expanding f in (4.8) in a Taylor series around $\bar{\mathbf{x}}_k$ and using the fact that $\nabla f(\bar{\mathbf{x}}_k) = \mathbf{0}$ gives

$$\mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \le \frac{\lambda}{2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2, \tag{4.9}$$

where λ is the bound for the Hessian of f over $\mathcal{B}_\rho(\hat{\mathbf{x}})$. By the triangle inequality, we have

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \le \|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \|\mathbf{x}_k - \hat{\mathbf{x}}\|. \tag{4.10}$$

Combining (4.10) with the condition $\mathbf{x}_k \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$, (4.9), and (A4) yields:

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \leq \epsilon + \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \epsilon + \sqrt{\frac{\lambda}{2\mu_k}}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \leq r.$$

Hence, $\mathbf{x}_{k+1} \in \mathcal{B}_r(\hat{\mathbf{x}})$.

Let $\hat{\mathbf{x}}_{k+1}$ denote an exact proximal point iterate:

$$\hat{\mathbf{x}}_{k+1} \in \arg\min\{F_k(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}.$$

By Proposition 2.1 and (4.7), we have

$$\|\hat{\mathbf{x}}_{k+1} - \mathbf{x}_k\| \leq \|\bar{\mathbf{x}}_k - \mathbf{x}_k\| \leq \gamma^k\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|. \tag{4.11}$$

By the triangle inequality, (4.11), the fact that $\tau \geq 1$, and (4.7), we obtain

$$\begin{aligned}
\|\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}\| &\leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{x}_k\| + \|\mathbf{x}_k - \hat{\mathbf{x}}\| \\
&\leq \tau\gamma^k\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| + \|\mathbf{x}_k - \hat{\mathbf{x}}\| \\
&\leq \tau\gamma^k\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| + \tau\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|\left(1 + \sum_{l=0}^{k-1}\gamma^l\right) \\
&\leq \tau\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \sum_{l=0}^{k}\gamma^l\right).
\end{aligned}$$

Referring to the definition of ϵ, we have $\hat{\mathbf{x}}_{k+1} \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$, where $\epsilon \leq r = \rho/2$.

By assumption (A1), f provides a local error bound with constants α and ρ. Hence, by Lemma 2.2, ∇f provides a local error bound with constants α and $r = \rho/2$. Since $\nabla F(\mathbf{x}_{k+1}) = \nabla f(\mathbf{x}_{k+1}) + \mu_k(\mathbf{x}_{k+1} - \mathbf{x}_k)^\mathsf{T}$ and $\mathbf{x}_{k+1} \in \mathcal{B}_r(\hat{\mathbf{x}})$, the local error bound condition gives

$$\alpha\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \|\nabla f(\mathbf{x}_{k+1})\| \leq \|\nabla F_k(\mathbf{x}_{k+1})\| + \mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{4.12}$$

If (C1) is used, then $\|\nabla F_k(\mathbf{x}_{k+1})\| \leq \mu_k\|\nabla f(\mathbf{x}_k)\|$, and (4.12) implies that

$$\begin{aligned}
\alpha\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| &\leq \mu_k(\|\nabla f(\mathbf{x}_k)\| + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|) \\
&\leq \mu_k(L\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \|\mathbf{x}_{k+1} - \mathbf{x}_k\|).
\end{aligned} \tag{4.13}$$

If (C2) is used, then $\|\nabla F_k(\mathbf{x}_{k+1})\| \leq \theta\mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$, and by (4.12), we have

$$\alpha\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \mu_k(1+\theta)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{4.14}$$

We now derive a bound for the $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ term in either (4.13) or (4.14). Since both $\mathbf{x}_{k+1}$ and $\hat{\mathbf{x}}_{k+1}$ lie in $\mathcal{B}_\rho(\hat{\mathbf{x}})$, we can apply Proposition 4.1 to the difference $F_k(\mathbf{x}_{k+1}) - F_k(\hat{\mathbf{x}}_{k+1})$ to obtain

$$F_k(\mathbf{x}_{k+1}) = F_k(\hat{\mathbf{x}}_{k+1}) + (F_k(\mathbf{x}_{k+1}) - F_k(\hat{\mathbf{x}}_{k+1}))$$

$$\leq F_k(\hat{\mathbf{x}}_{k+1}) + \frac{1}{\mu_k}\|\nabla F_k(\mathbf{x}_{k+1})\|^2$$
$$\leq F_k(\bar{\mathbf{x}}_k) + \frac{1}{\mu_k}\|\nabla F_k(\mathbf{x}_{k+1})\|^2. \tag{4.15}$$

Above we observed that $\mathbf{x}_k \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$ where $\epsilon \leq r = \rho/2$. In (2.3) we show that $\bar{\mathbf{x}}_k \in \mathcal{B}_\rho(\hat{\mathbf{x}})$ when $\mathbf{x}_k \in \mathcal{B}_r(\hat{\mathbf{x}})$. Hence, by (A2) we have

$$\|\nabla f(\mathbf{x}_k)\| = \|\nabla f(\mathbf{x}_k) - \nabla f(\bar{\mathbf{x}}_k)\| \leq L\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|. \tag{4.16}$$

If (C1) is used, then by (4.16),

$$\|\nabla F_k(\mathbf{x}_{k+1})\| \leq \mu_k\|\nabla f(\mathbf{x}_k)\| \leq \mu_k L\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|. \tag{4.17}$$

Using the relation $f(\bar{\mathbf{x}}_k) \leq f(\mathbf{x}_{k+1})$ in (4.15) gives:

$$\frac{\mu_k}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq \frac{\mu_k}{2}\|\bar{\mathbf{x}}_k - \mathbf{x}_k\|^2 + \frac{1}{\mu_k}\|\nabla F_k(\mathbf{x}_{k+1})\|^2. \tag{4.18}$$

Combining this with (4.17), we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq (1 + 2L^2)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2. \tag{4.19}$$

Similarly, if criterion (C2) is used, then $\|\nabla F_k(\mathbf{x}_{k+1})\| \leq \theta\mu_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$, and by (4.18), we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq \frac{1}{1 - 2\theta^2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2. \tag{4.20}$$

Combining (4.19) and (4.20) and referring to the definition of τ in (A3), we conclude that

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \tau\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \tag{4.21}$$

if either acceptance criterion (C1) or (C2) is used.

Inserting the bound (4.21) in (4.13) or (4.14) yields (4.6). By (4.6) and the definition of β in (A3), we have

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \left(\frac{\Lambda\mu_k}{\alpha}\right)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \leq \gamma\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \leq \gamma^{k+1}\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|.$$

This establishes the second half of (4.7) for $j = k + 1$.

For the first half of (4.7), we use the triangle inequality, (4.21), and the induction hypothesis to obtain:

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}\| \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \|\mathbf{x}_k - \hat{\mathbf{x}}\|$$
$$\leq \tau\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \|\mathbf{x}_k - \hat{\mathbf{x}}\|$$
$$\leq \tau\gamma^k\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\| + \|\mathbf{x}_k - \hat{\mathbf{x}}\|$$

$$\leq \tau\gamma^k\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\| + \tau\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \sum_{l=0}^{k-1}\gamma^l\right)$$

$$\leq \tau\|\mathbf{x}_0 - \hat{\mathbf{x}}\|\left(1 + \sum_{l=0}^{k}\gamma^l\right).$$

This completes the proof of the induction step, and in particular, it shows that $\mathbf{x}_{k+1} \in \mathcal{B}_\epsilon(\hat{\mathbf{x}})$, where ϵ is defined in (A4).

By (4.7) and (4.21), we have

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \tau\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \leq \tau\gamma^k\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|. \tag{4.22}$$

Hence, the $\mathbf{x}_k$ form a Cauchy sequence, which has a limit denoted $\mathbf{x}^*$. By the triangle inequality and (4.22),

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \sum_{j=k}^{\infty}\|\mathbf{x}_{j+1} - \mathbf{x}_j\| \leq \tau\sum_{j=k}^{\infty}\|\bar{\mathbf{x}}_j - \mathbf{x}_j\|$$

$$\leq \tau\sum_{j=k}^{\infty}\gamma^j\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\| = \frac{\tau\gamma^k}{1-\gamma}\|\bar{\mathbf{x}}_0 - \mathbf{x}_0\|.$$

This establishes (4.5). □

Remark In our analysis of the exact proximal point algorithm, our proof of Theorem 3.1 could rely on the bound provided by Proposition 2.1 for the step size $\mathbf{x}_{k+1} - \mathbf{x}_k$. In Theorem 4.2, we obtain a similar bound using either condition (C1) or (C2) along with the relationship between the regularization parameter μ_k and the current iterate $\mathbf{x}_k$. Condition (A4) is satisfied if

$$\lim_{\mathbf{x}\to\hat{\mathbf{x}}}\frac{\|\mathbf{x} - \bar{\mathbf{x}}\|^2}{\mu(\mathbf{x})} = 0 \tag{4.23}$$

and $\mathbf{x}_0$ is sufficiently close to $\hat{\mathbf{x}}$. For example, if $\mu(\mathbf{x}) = \beta\|\nabla f(\mathbf{x})\|^\eta$ where $\eta \in [0, 2)$ and $\beta > 0$ is a constant, and if ∇f provides a local error bound at $\hat{\mathbf{x}}$, then

$$\frac{\|\mathbf{x} - \bar{\mathbf{x}}\|^2}{\mu(\mathbf{x})} = \frac{\|\mathbf{x} - \bar{\mathbf{x}}\|^2}{\beta\|\nabla f(\mathbf{x})\|^\eta} \leq \left(\frac{1}{\beta\alpha^\eta}\right)\|\mathbf{x} - \bar{\mathbf{x}}\|^{2-\eta}.$$

Hence, (4.23) holds for $\eta \in [0, 2)$. Also, (A3) holds if either $\eta \in (0, 2)$ and ρ is sufficiently small, or $\eta = 0$ and β is sufficiently small. For this choice of $\mu(\cdot)$, it follows from Theorem 4.2, that the convergence rate to the set of minimizers $\mathbf{X}$ is linear for $\eta = 0$, superlinear for $\eta \in (0, 1)$, and at least quadratic for $\eta \in [1, 2)$.

5 Global convergence

In Sect. 4, we analyzed the local convergence of approximate minimizers of (1.2). We now establish a global convergence result for an algorithm which has the follow-

ing structure: If $\mathbf{x}_k$ denotes the current iterate, one iteration of some descent method is used to generate a point $\mathbf{y}_k$ by performing a Wolfe line search along a search direction $\mathbf{d}_k$. The point $\mathbf{x}_{k+1}$ is either $\mathbf{y}_k$, or any point for which $F_k(\mathbf{x}_{k+1}) \le F_k(\mathbf{y}_k)$. If $\mathbf{x}_{k+1}$ satisfies (C1) or (C2), then the local convergence theory of Sect. 4 applies. To establish global convergence, we make the following assumptions:

(L1) There exist positive constants c_1 and c_2 satisfying both the sufficient descent condition

$$\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \le -c_1\|\mathbf{g}_k\|^2,$$

and the bound

$$\|\mathbf{d}_k\| \le c_2\|\mathbf{g}_k\|,$$

where $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^{\mathsf{T}} = \nabla F_k(\mathbf{x}_k)$.

(L2) The Wolfe conditions [19, 20] hold. That is, if α_k denotes the stepsize and $\mathbf{y}_k = \mathbf{x}_k + \alpha_k\mathbf{d}_k$, then

$$F_k(\mathbf{y}_k) \le F_k(\mathbf{x}_k) + \delta\alpha_k\nabla F_k(\mathbf{x}_k)\mathbf{d}_k = F_k(\mathbf{x}_k) + \delta\alpha_k\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k,$$

and

$$\nabla F_k(\mathbf{y}_k)\mathbf{d}_k \ge \sigma\nabla F_k(\mathbf{x}_k)\mathbf{d}_k = \sigma\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k,$$

where δ and σ are constants satisfying $0 < \delta < \sigma < 1$.

(L3) The next iterate $\mathbf{x}_{k+1}$ is any point satisfying $F_k(\mathbf{x}_{k+1}) \le F_k(\mathbf{y}_k)$.

Our global convergence result is as follows:

Theorem 5.1 *Let $\mathbf{x}_k$, $k \ge 1$, denote a sequence of inexact proximal point iterates associated with (1.2). We assume that the algorithm used to approximately solve (1.2) satisfies* (L1)–(L3). *If the iterates are all contained in a convex set $\mathcal{C}$ where ∇f is Lipschitz continuous, and if $\mu_k \le \beta\|\nabla f(\mathbf{x}_k)\|^\eta$ for some $\eta \in [0, 2)$ and some constant β, then we have*

$$\lim_{k\to\infty} \mathbf{g}(\mathbf{x}_k) = \mathbf{0}.$$

Proof Since $\mathbf{X}$ is nonempty, f is bounded from below. By (L3) and the definition of F_k,

$$f(\mathbf{x}_{k+1}) \le F_k(\mathbf{x}_{k+1}) \le F_k(\mathbf{y}_k) \le F_k(\mathbf{x}_k) = f(\mathbf{x}_k).$$

Hence, we have

$$\infty > \sum_{k=0}^{\infty} f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) = \sum_{k=0}^{\infty} F_k(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$$

$$\ge \sum_{k=0}^{\infty} F_k(\mathbf{x}_k) - F_k(\mathbf{x}_{k+1})$$

$$\geq \sum_{k=0}^{\infty} F_k(\mathbf{x}_k) - F_k(\mathbf{y}_k). \tag{5.1}$$

By (L1) and the first Wolfe condition in (L2), it follows that

$$F_k(\mathbf{y}_k) \leq F_k(\mathbf{x}_k) - \delta c_1 \alpha_k \|\mathbf{g}_k\|^2.$$

Combining this with (5.1) gives

$$\sum_{k=0}^{\infty} \alpha_k \|\mathbf{g}(\mathbf{x}_k)\|^2 < \infty. \tag{5.2}$$

By the second Wolfe condition in (L2) and (L1), we have

$$\begin{aligned}(\nabla F_k(\mathbf{y}_k) - \nabla F_k(\mathbf{x}_k))\mathbf{d}_k &\geq -(1-\sigma)\nabla F_k(\mathbf{x}_k)\mathbf{d}_k \\ &= -(1-\sigma)\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \\ &\geq (1-\sigma)c_1\|\mathbf{g}_k\|^2. \end{aligned} \tag{5.3}$$

If L is the Lipschitz constant for ∇f in $\mathcal{C}$, then we have

$$\begin{aligned}(\nabla F_k(\mathbf{y}_k) - \nabla F_k(\mathbf{x}_k))\mathbf{d}_k &= (\nabla f(\mathbf{y}_k) - \nabla f(\mathbf{x}_k) + \mu_k(\mathbf{y}_k - \mathbf{x}_k)^{\mathsf{T}})\mathbf{d}_k \\ &\leq (L+\mu_k)\|\mathbf{y}_k - \mathbf{x}_k\|\|\mathbf{d}_k\| = \alpha_k(L+\mu_k)\|\mathbf{d}_k\|^2 \\ &\leq c_2^2\alpha_k(L+\mu_k)\|\mathbf{g}_k\|^2.\end{aligned}$$

In the last inequality, we utilize (L1). Combining this with (5.3) yields:

$$\alpha_k \geq \frac{c_1(1-\sigma)}{c_2^2(L+\mu_k)}.$$

By (5.2) and the bound $\mu_k \leq \beta\|\nabla f(\mathbf{x}_k)\|^{\eta}$,

$$\begin{aligned}\infty > \sum_{k=0}^{\infty} \frac{\|\mathbf{g}(\mathbf{x}_k)\|^2}{L+\mu_k} &\geq \frac{1}{2}\sum_{k=0}^{\infty} \min\left(\frac{\|\mathbf{g}(\mathbf{x}_k)\|^2}{L}, \frac{\|\mathbf{g}(\mathbf{x}_k)\|^2}{\mu_k}\right) \\ &\geq \frac{1}{2}\sum_{k=0}^{\infty} \min\left(\frac{\|\mathbf{g}(\mathbf{x}_k)\|^2}{L}, \frac{\|\mathbf{g}(\mathbf{x}_k)\|^{2-\eta}}{\beta}\right).\end{aligned}$$

Since $\eta < 2$, we conclude that $\mathbf{g}(\mathbf{x}_k)$ approaches 0. □

Table 1 $\|\mathbf{g}_k\|$ and total number of conjugate gradient iterations (It) versus iteration number k

k	Problem 1				Problem 2			
	(C1)		(C2)		(C1)		(C2)	
	$\|\mathbf{g}_k\|$	It	$\|\mathbf{g}_k\|$	It	$\|\mathbf{g}_k\|$	It	$\|\mathbf{g}_k\|$	It
1	4.5e−01	6	5.4e−01	5	4.7e−01	2	1.3e−01	4
2	7.2e−02	12	1.0e−01	10	1.3e−02	6	3.2e−03	8
3	2.6e−03	22	7.1e−03	16	1.5e−04	13	2.5e−05	14
4	3.5e−06	32	2.6e−05	25	4.2e−07	32	4.5e−08	27
5	6.4e−12	48	4.3e−10	38	1.8e−10	151	1.2e−11	61

6 Numerical results

We now present numerical examples to illustrate the convergence theory. To illustrate the quadratic convergence when $\mu(\mathbf{x}) = \beta\|\nabla f(\mathbf{x})\|$, we consider the following two problems (the first is introduced in [11]):

$$\text{(P1)} \quad f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \frac{1}{12}\sum_{i=1}^{n-1}\alpha_i(x_i - x_{i+1})^4,$$

$$\text{(P2)} \quad f(\mathbf{x}) = \sum_{i=1}^{n} b_i(x_i - 1)^2 + \sum_{i=1}^{n}(x_i - 1)^4.$$

In (P1) we take $n = 10$, $\alpha_i = 1$, and the starting guess $x_i = i$, $1 \le i \le n$. The set of minimizers are given by $x_1 = x_2 = \cdots = x_n$. In (P2) we take $n = 10$, $b_i = e^{-4i}$, and the starting guess $x_i = 1 + 1/i$, $1 \le i \le n$. For this problem, the minimizer is unique, however, the condition number of the Hessian at the solution is around .5e16. Table 1 gives the gradient norms corresponding to $\mu(\mathbf{x}) = .05\|\nabla f(\mathbf{x})\|$, and acceptance criterions (C1) and (C2). In other words, we took $\mu(\mathbf{x}) = \beta\|\nabla f(\mathbf{x})\|^\eta$ with $\beta = .05$ and $\eta = 1$. For (C2), we chose $\theta = .66$. The subproblems (1.2) were solved using our conjugate gradient routine CG_DESCENT [5, 6], stopping when either (C1) or (C2) is satisfied. We also tabulate the total number of conjugate gradient iterations as a function of the proximal iteration number (column "It" in the Table 1). The number of conjugate gradient iterations associated with any particular proximal iteration is obtained by subtracting the cumulative number of conjugate gradient iteration for two consecutive proximal iterations.

In the next series of numerical experiments, we solve some ill-conditioned problems from the CUTE library [1]. In the numerical experiments we performed when evaluating the code CG_DESCENT in [6], there were several problems with relatively small dimension where the conjugate gradient method was relatively slow, and the Hessian at the final solution was very ill conditioned. These problems are listed in Table 2. For these problems, we find that it is more efficient to use the proximal strategy in a neighborhood $\mathcal{N}$ of an optimum; outside $\mathcal{N}$, we apply CG_DESCENT to the original problem (1.1). In our numerical experiments, our method for choosing $\mathcal{N}$ was the following: We applied the conjugate gradient method to the original

Table 2 Solution statistics for ill-condition CUTE test problems and CG_DESCENT

Problem	Dim	Cond	No proximal point			With proximal point		
			It	NF	NG	It	NF	NG
SPARSINE	2000	2.6e17	12,528	25,057	12,529	10,307	20,615	10,308
SPARSINE	1000	3.4e14	4,657	9,315	4,658	3,760	7,521	3,761
NONDQUAR	1000	6.6e10	4,004	8,015	4,152	3,013	6,032	3,068
NONDQUAR	500	1.0e10	3,027	6,074	3,185	2,526	5,062	2,676
EIGENALS	420	1.2e06	1,792	3,591	1,811	1,464	2,935	1,482
EIGENBLS	420	3.0e05	5,087	10,185	5,099	2,453	4,910	2,458
EIGENCLS	420	8.2e04	1,733	3,484	1,754	1,774	3,566	1,795
NCB20	510	3.7e16	1,631	3,048	2,372	1,251	2,262	1,684

problem (1.1) until the following condition was satisfied:

$$\|\mathbf{g}(\mathbf{x})\|_\infty \leq 10^{-2}(1 + |f(\mathbf{x})|). \tag{6.1}$$

When this condition holds, we continue to apply the conjugate gradient method until an estimate for the condition number exceeds 10^3; then we switch to the proximal point method (1.2), using CG_DESCENT to solve the subproblems. We estimate the condition number by first estimating the second derivative of the function along the normalized search direction. Our estimate for the condition number is the ratio between the maximum and minimum second derivative, during the iterations after (6.1) is satisfied.

Table 2 gives convergence results for ill-condition problems from CUTE [1]. The "exact condition numbers" are computed in the following way: We solve the problem and output the Hessian matrix at the solution. The extreme eigenvalues of the Hessian were computed using Matlab, and the eigenvalue ratio appears in the column labeled "Cond" of Table 2. For the proximal point iteration, we used acceptance criterion (C2). The iterations were continued until the stopping condition $\|\nabla f(\mathbf{x})\|_\infty \leq 10^{-6}$ was satisfied. The number of iteration (It), number of function evaluations (NF), and number of gradient evaluations (NG) are given in the table. Observe that the reduction in the number of function and gradient evaluations varies from almost nothing in EIGENCLS to about 50% for EIGENBLS.

In our numerical experiments, we found that $\eta = 1$ often yields the best results. For comparison with the $\eta = 1$ results in Table 2, we give in Table 3 the convergence results obtained by taking $\eta = 0.5$ and $\eta = 2.0$.

As these experiments indicate, proximal regularization can lead to faster solution times for certain ill-conditioned problems. On the other hand, for a well-conditioned problem, one may spend as much time solving a single proximal regularization as is spent in solving the original problem. The development of an effective general strategy for deciding when to turn on the proximal regularization is a topic for future research. Possibly, our estimate for the Hessian condition number along the search directions could be developed into a general rule for activating the proximal regularization.

Table 3 Solution statistics for ill-condition CUTE test problems with different choices for η

Problem	Dim	$\eta = 0.5$			$\eta = 2.0$		
		It	NF	NG	It	NF	NG
SPARSINE	2000	13,716	27,433	13,717	11,851	23,703	11,852
SPARSINE	1000	5,153	10,307	5,154	3,776	7,553	3,777
NONDQUAR	1000	3,392	6,789	3,510	5,198	10,403	5,469
NONDQUAR	500	4,751	9,514	5,027	4,063	8,138	4,359
EIGENALS	420	1,411	2,829	1,424	1,302	2,611	1,316
EIGENBLS	420	7,025	14,060	7,036	2,583	5,170	2,588
EIGENCLS	420	1,778	3,574	1,799	1,762	3,542	1,783
NCB20	510	3,290	5,918	5,004	823	1,638	986

7 Final discussion

In this paper, we analyzed the convergence of the proximal point method for problems that may not be strongly convex at an optimum. A new local error bound based on the function value is introduced in Sect. 2. We show that it is equivalent to the standard local error bound based on the gradient. Two criteria (C1) and (C2) for accepting an approximate solution to the proximal point problem (1.2) were introduced. With either criterion, the convergence results obtained in Theorem 4.2 for the approximate iterates are analogous to those obtained in Theorem 3.1 for the exact iteration. By taking $\mu(\mathbf{x}) = \beta\|\nabla f(\mathbf{x})\|^\eta$ where $\eta \in [0, 2)$ and $\beta > 0$ is a constant, we obtain convergence to the set of minimizers $\mathbf{X}$ that is linear for $\eta = 0$ and β sufficiently small, superlinear for $\eta \in (0, 1)$, and at least quadratic for $\eta \in [1, 2)$.

References

1. Bongartz, I., Conn, A.R., Gould, N.I.M., Toint, P.L.: CUTE: constrained and unconstrained testing environments. ACM Trans. Math. Softw. **21**, 123–160 (1995)
2. Combettes, P.L., Pennanen, T.: Proximal methods for cohypomonotone operators. SIAM J. Control **43**, 731–742 (2004)
3. Fan, J., Yuan, Y.: On the convergence of the Levenberg-Marquardt method without nonsingularity assumption. Computing **74**, 23–39 (2005)
4. Ha, C.D.: A generalization of the proximal point algorithm. SIAM J. Control **28**, 503–512 (1990)
5. Hager, W.W., Zhang, H.: CG_DESCENT user's guide. Tech. Rep., Department of Mathematics, University of Florida, Gainesville (2004)
6. Hager, W.W., Zhang, H.: A new conjugate gradient method with guaranteed descent and an efficient line search. SIAM J. Optim. **16**, 170–192 (2005)
7. Hager, W.W., Zhang, H.: Algorithm 851: CG_DESCENT, a conjugate gradient method with guaranteed descent. ACM Trans. Math. Softw. **32**, 113–137 (2006)
8. Humes, C., Silva, P.: Inexact proximal point algorithms and descent methods in optimization. Optim. Eng. **6**, 257–271 (2005)
9. Iusem, A.N., Pennanen, T., Svaiter, B.F.: Inexact variants of the proximal point algorithm without monotonicity. SIAM J. Optim. **13**, 1080–1097 (2003)
10. Kaplan, A., Tichatschke, R.: Proximal point methods and nonconvex optimization. J. Glob. Optim. **13**, 389–406 (1998)
11. Li, D., Fukushima, M., Qi, L., Yamashita, N.: Regularized Newton methods for convex minimization problems with singular solutions. Comput. Optim. Appl. **28**, 131–147 (2004)

12. Luque, F.J.: Asymptotic convergence analysis of the proximal point algorithm. SIAM J. Control **22**, 277–293 (1984)
13. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Francaise Inf. Rech. Oper. Ser. R-3 **4**, 154–158 (1970)
14. Martinet, B.: Determination approachée d'un point fixe d'une application pseudo-contractante. C.R. Séances Acad. Sci. **274**, 163–165 (1972)
15. Pennanen, T.: Local convergence of the proximal point algorithm and multiplier methods without monotonicity. Math. Oper. Res. **27**, 170–191 (2002)
16. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **2**, 97–116 (1976)
17. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control **14**, 877–898 (1976)
18. Tseng, P.: Error bounds and superlinear convergence analysis of some Newton-type methods in optimization. In: Pillo, G.D., Giannessi, F. (eds.) Nonlinear Optimization and Related Topics, pp. 445–462. Kluwer, Dordrecht (2000)
19. Wolfe, P.: Convergence conditions for ascent methods. SIAM Rev. **11**, 226–235 (1969)
20. Wolfe, P.: Convergence conditions for ascent methods II: some corrections. SIAM Rev. **13**, 185–188 (1971)
21. Yamashita, N., Fukushima, M.: The proximal point algorithm with genuine superlinear convergence for the monotone complementarity problem. SIAM J. Optim. **11**, 364–379 (2000)
22. Yamashita, N., Fukushima, M.: On the rate of convergence of the Levenberg-Marquardt method. In: Topics in Numerical Analysis. Comput. Suppl., vol. 15, pp. 239–249. Springer, New York (2001)