J.V. Burke* · Maijian Qian

On the superlinear convergence of the variable metric proximal point algorithm using Broyden and BFGS matrix secant updating

Abstract. In previous work, the authors provided a foundation for the theory of variable metric proximal point algorithms in Hilbert space. In that work conditions are developed for global, linear, and super–linear convergence. This paper focuses attention on two matrix secant updating strategies for the finite dimensional case. These are the Broyden and BFGS updates. The BFGS update is considered for application in the symmetric case, e.g., convex programming applications, while the Broyden update can be applied to general monotone operators. Subject to the linear convergence of the iterates and a quadratic growth condition on the inverse of the operator at the solution, super–linear convergence of the iterates is established for both updates. These results are applied to show that the Chen–Fukushima variable metric proximal point algorithm is super–linearly convergent when implemented with the BFGS update.

Key words. maximal monotone operator – proximal point methods – variable metric – global convergence – super-linear convergence

1. Introduction

In [2], we introduced the variable metric proximal point algorithm (VMPPA) for general monotone operators and established a basic convergence theory. The algorithm builds on the classical proximal point algorithm and can be viewed as a Newton–like method for solving inclusions of the form

$$0 \in T(z)$$

where T is a maximal monotone operator on a Hilbert space. In this paper, we focus on the finite dimensional case and consider two quasi–Newton updating strategies for generating the Newton–like iterates: the BFGS and Broyden updates. The BFGS update is appropriate for application to convex programming and the Broyden update is suitable for other applications such as mini–max problems. We develop a local convergence theory for these updates. In particular, we establish conditions for the super–linear convergence of the iterates. These results are then used to establish the super–linear convergence of the Chen–Fukushima VMPPA for convex programming when implemented with the BFGS update.

J.V. Burke: Department of Mathematics, Box # 354350, University of Washington, Seattle, Washington 98195–4350

M. Qian: Department of Mathematics, California State University, Fullerton, CA 92834

Recently, the drive to develop a VMPPA in the context of finite–valued, finite dimensional convex programming has been joined by several authors [1,2,5,7,11–14,17,18]. In convex programming, the goal is to derive a variable metric method for minimizing the Moreau–Yosida regularization of a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$:

$$f_\lambda(z) := \min_{y \in \mathbb{R}^n} \left\{ \lambda f(y) + \frac{1}{2} \|y - z\|^2 \right\} \tag{1}$$

(in the finite–valued case, f cannot take the value $+\infty$). Here the operator T is the convex subdifferential of the function f, denoted ∂f. The inclusion $0 \in \partial f(z)$ corresponds to the first–order optimality conditions for the convex programming problem

$$\mathcal{P} : \quad \min_{x \in \mathbb{R}^n} f(x) .$$

It is well known that the convex program $\mathcal{P}$ has the same solution set as the convex program

$$\mathcal{P}_\lambda : \quad \min_{x \in \mathbb{R}^n} f_\lambda(x) ,$$

for every $\lambda > 0$. In addition, the function f_λ is continuously differentiable with Lipschitz continuous derivative even if the function f is neither differentiable or finite–valued. The goal of all research into VMPPAs is to develop a super–linearly convergent method for solving $\mathcal{P}_\lambda$ for a fixed value of the proximation parameter λ. The method should not require precise values for either f_λ or its derivative, and should not require excessively strong smoothness hypotheses on the function f.

We study matrix secant updating strategies for the VMPPA in the operator setting. Other than [2], all previous work on the VMPPA is focused on the convex programming case where the objective function is assumed to be finite valued. In the convex programming case, matrix secant updating strategies are studied in [1], [5], [11], [12], and [14]. In [1], Bonnans, Gilbert, Lemaréchal, and Sagastizábal consider an algorithmic pattern modeled on the approach suggested by Qian in [18]. In this regard, the quasi–Newton updates are applied to the function f instead of f_λ. This approach allows one to circumvent the technical difficulties associated with varying the value of λ in f_λ. The authors provide an adaptation of the Dennis–Moré characterization theorem for the super–linear convergence of Newton–like methods in nonlinear programming [8]. The authors also establish the super–linear convergence of the PSB, DFP, and BFGS updates. The results in [1] require that the function f is continuously differentiable at a unique solution with Lipschitz continuous derivative, the derivative is strongly directionally differentiable, and the directional derivative operator is *positive definite*. In addition, the results for the BFGS update require that the directional derivative satisfy a radially–Lipschitz condition [1, Inequality 3.11].

In [5], Chen and Fukushima use a bundle strategy for approximating f_λ and its gradient. They use a line search based on the function f instead of f_λ. This is an important practical innovation since the evaluation of approximations to f_λ can be costly. They establish an analog of Dennis and Moré's characterization theorem for super–linear convergence under the assumption that f is strongly convex, f_λ is strongly twice differentiable at a unique solution, the derivative approximations and their inverses

are bounded, and the error in the approximation to f_λ converges to zero much faster than sum of the squares of the errors in the approximation to ∇f_λ. None of the results in [5] depend on quasi-Newton updating. However, the authors encourage the use of the BFGS update.

In [11], Lemaréchal and Sagastizábal consider a scalar quasi–Newton update. Following the classical result of Rockafellar [20], the authors establish super–linear convergence by showing that the proximation parameter λ diverges to $+\infty$. The analysis requires that f is locally continuously differentiable with Lipschitz continuous derivative and that f satisfies a quadratic growth condition near the solution set. In [12], Lemaréchal and Sagastizábal establish another analog of the Dennis–Moré characterization theorem under a technical assumption on the structure of the derivative approximation for ∂f. The authors then study the SR1 update and a scalar quasi–Newton update which they label a *poor man's* update. A convergence rate is not established for the SR1 update. Super–linear convergence is established for the poor man's update along the lines of [11] by showing that the inverse of the proximation parameter converges to zero. The super–linear convergence results we obtain in Sect. 3 do not require that the proximation parameter λ diverges to $+\infty$. However, we do require that this parameter is sufficiently large (see hypothesis (H4) in Sect. 3.1).

In [14], Mifflin, Sun, and Qi obtain the first super–linear convergence result for a variable metric proximal point algorithm using the BFGS matrix secant update in the setting of finite dimensional finite–valued convex programming. Their algorithm uses a line search based on approximations to the function f_λ and requires that f_λ is strongly convex with ∇f_λ Fréchet differentiable at the unique global solution to the convex program. In addition, the main super–linear convergence result for the Mifflin–Sun–Qi algorithm [14, Theorem 5.3] assumes that the iterates satisfy an approximation property involving the Hessian $\nabla^2 f_\lambda$ [14, Theorem 5.3]. Using very different techniques and hypotheses no stronger than those used in [14], we are able to avoid such a hypothesis. Indeed, we are able to show that the iterates in our algorithm automatically satisfy an even stronger approximation property (Lemma 3).

Our goal is to establish the super–linear convergence of the VMPPA as presented in [2] using the Broyden and BFGS updates. This algorithm applies to general monotone operators on $I\!\!R^n$ and is not confined to applications in convex programming. In the context of convex programming, our convergence hypotheses differ from those used in [1,5,11,12] since they are imposed on the operator $(\partial f)^{-1}$ rather than ∂f. This difference is significant since it allows us to handle the non–finite valued case, i.e., constrained convex programs. We apply these general results to the Chen–Fukushima VMPPA in the final section. Further details concerning the application of the VMPPA in the context of convex programming can be found in [3] along with some preliminary numerical results.

The paper is organized as follows. In Sect. 2, we recall the basic features of the VMPPA. In Sect. 3, we study the Broyden and BFGS updating strategies for the VMPPA and provide conditions under which super–linear convergence is achieved. In Sect. 4, we consider the Chen–Fukushima VMPPA for finite–valued convex programming. We begin by establishing the linear convergence of the algorithm using hypotheses consistent with the analysis provided in Sect. 3. We then apply the results of Sect. 3 to the Chen–Fukushima algorithm.

A word about our notation is in order. Let $\mathbb{R}^n$ denote real n–dimensional Euclidean space and let $\mathbb{R}^{n \times n}$ denote the set of all real $n \times n$ matrices. We use the standard Euclidean inner product and norm on $\mathbb{R}^n$, i.e., for $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = x^T y$ and $\|x\| = \sqrt{x^T x}$. For a matrix $H \in \mathbb{R}^{n \times n}$, the norm is $\|H\| = \sup_{\|x\|=1} \|Hx\|$. We denote the closed unit ball in $\mathbb{R}^n$ by $\mathbb{B}$. Then the ball with center a and radius r is denoted by $a + r\mathbb{B}$. Given a set $Z \subset \mathbb{R}^n$ and an element $z \in \mathbb{R}^n$, the distance of z to Z is $\text{dist}\,(z, Z) = \inf\{\|z - z'\| : z' \in Z\}$.

Given a *multi-function* (also referred to as a *mapping* or an *operator* depending on the context) $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ (here the double arrows $\rightrightarrows$ are used to signify the fact that T is a multi–function), the *graph* of T, graph (T), is the subset of the product space $\mathbb{R}^n \times \mathbb{R}^n$ defined by graph $T = \{(z, w) \in \mathbb{R}^n \times \mathbb{R}^n | w \in T(z)\}$. The *domain* of T is the set dom $T := \{z \in \mathbb{R}^n | T(z) \neq \emptyset\}$. The identity mapping will be denoted by I. The *inverse* of an operator T is defined by $T^{-1}(w) := \{z \in \mathbb{R}^n | (z, w) \in \text{graph}\,T\}$.

Given a lower semi-continuous convex function $f : \mathbb{R}^n \to \mathbb{R} \bigcup \{+\infty\}$, the *conjugate* of f is defined by $f^*(z^*) = \sup_{z \in \mathbb{R}^n} \{\langle z^*, z \rangle - f(z)\}$, and the subdifferential of f is the multi-function defined by $\partial f(z) = \{y \in \mathbb{R}^n : f(z') \geq f(z) + \langle y, z' - z \rangle$ for all $z' \in \mathbb{R}^n\}$.

2. The variable metric proximal point algorithm

The multi-function $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is said to be *monotone* if for every (z, w) and (z', w') in graph (T) we have $\langle z - z', w - w' \rangle \geq 0$. The monotone operator T is said to be *maximal* if its graph is not properly contained in the graph of any other monotone operator. The proximal point algorithm for solving the inclusion $0 \in T(z)$ generates a sequence $\{z^k\}$ satisfying the approximation rule

$$z^{k+1} \approx (I + \lambda_k T)^{-1}(z^k)$$

for a given sequence of positive scalars $\{\lambda_k\}$.

In the case of convex programming, the function f_λ defined in (1) is continuously differentiable [15] with $\nabla f_\lambda(z) = -w(\lambda, z)$ where $w(\lambda, z) = P(\lambda, z) - z$ and $P(\lambda, z)$ is the unique solution to the minimization problem in (1). The proximal point iteration has the form

$$z^{k+1} = z^k + w^k, \quad \text{where } w^k \approx -\nabla f_{\lambda_k}(z^k).$$

That is, it is a method of steepest descent with unit step size applied to the function f_{λ_k} with λ_k varying between iterations. The algorithm for a general maximal monotone operator T can be formally derived from this iteration by replacing ∂f by T and $-\nabla f_{\lambda_k}$ by the operator

$$D_k = \left[(I + \lambda_k T)^{-1} - I \right]. \tag{2}$$

The operator D_k corresponds to the negative of the gradient operator and yields a direction analogous to the direction of steepest descent. The proximal point algorithm takes

the form $z^{k+1} = z^k + w^k$, with $w^k \approx D_k(z^k)$. A Newton–like variation on this iteration yields the VMPPA.

The Variable Metric Proximal Point Algorithm:

Let $z^0 \in \mathbb{R}^n$, $H_0 \in \mathbb{R}^{n \times n}$, and $\lambda_0 \geq 1$ be given. Having $z^k \in \mathbb{R}^n$, $H_k \in \mathbb{R}^{n \times n}$, and $\lambda_k \geq 1$, set

$$z^{k+1} := z^k + H_k w^k \quad \text{where} \quad w^k \approx D_k(z^k)$$

and choose $H_{k+1} \in \mathbb{R}^{n \times n}$ and $\lambda_{k+1} \geq 1$.

Remark. The condition $\lambda_k \geq 1$ is used in [2] to establish global convergence. A more stringent condition on the magnitude of the λ_k's is required for the local analysis in Sect. 3 (e.g. see Theorems 1 and 2).

Our local analysis uses the following approximation criterion for the w^k's:

$$(\mathcal{L}) \quad \left\| w^k - D_k(z^k) \right\| \leq \delta_k \left\| w^k \right\| \quad \text{with} \quad \sum_{k=0}^{\infty} \delta_k < +\infty .$$

This criterion is stronger than a similar condition used in [2]. The stronger condition is required to establish super–linear convergence.

In general, criterion $(\mathcal{L})$ is not implementable. However, it is shown in [2, Proposition 3.1] that the condition $(\mathcal{L})$ is implied by the approximation criterion

$$(\mathcal{L}') \quad \text{dist}\,(0, S_k(w^k)) \leq \frac{\delta_k}{\lambda_k} \left\| w^k \right\| \quad \text{with} \quad \sum_{k=0}^{\infty} \delta_k < +\infty ,$$

where $S_k(w) = T(z^k + w) + \frac{1}{\lambda_k} w$. This condition is implementable. In particular, one need not obtain an element in $S_k(w^k)$ of least norm in order to guarantee that $(\mathcal{L}')$ is satisfied.

Before leaving this section we recall from [20] a few properties of the operators D_k and $P_k := D_k + I$ that are essential in the analysis to follow.

Proposition 1 [20, Proposition 1].

a) *The operator D_k can be expressed as $D_k = -(I + T^{-1}\frac{1}{\lambda_k})^{-1}$ and for any $z \in \mathbb{R}^n$,*
   $-\frac{1}{\lambda_k} D_k(z) \in T(P_k(z))$.
b) *For any $z, z' \in \mathbb{R}^n$, $\langle P_k(z) - P_k(z'), D_k(z) - D_k(z') \rangle \leq 0$.*
c) *For any $z, z' \in \mathbb{R}^n$, $\| P_k(z) - P_k(z') \|^2 + \| D_k(z) - D_k(z') \|^2 \leq \| z - z' \|^2$.*

Remark. An important consequence of Part c) above is that the operators P_k and D_k are non–expansive. We make free use of this fact in subsequent sections.

3. Matrix secant updating

We now consider the local behavior of the VMPPA when the matrices H_k are updated using the Broyden and BFGS formula. Our goal is to establish the super–linear convergence of the iterates beginning with the assumption that the iterates converge at a global linear rate. Conditions that guarantee global linear convergence can be found in [2, Theorem 7.1]. To establish super–linear convergence, we need to assume that the operator T^{-1} satisfies the *quadratic growth condition*.

Definition 1.   *We say that an operator* $\Psi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ *is Lipschitz continuous at a point* $\bar{w}$ *(with modulus* $\alpha \geq 0$ *) if the set* $\Psi(\bar{w})$ *is nonempty and there is a* $\tau > 0$ *such that*

$$\Psi(w) \subset \Psi(\bar{w}) + \alpha \|w - \bar{w}\| \, \mathbb{B} \quad \text{whenever} \ \|w - \bar{w}\| \leq \tau \ .$$

We say that Ψ *is differentiable at a point* $\bar{w}$ *if* $\Psi(\bar{w})$ *consists of a single element* $\bar{z}$ *and there is a continuous linear transformation* $J : \mathbb{R}^n \rightarrow \mathbb{R}^n$ *such that*

$$\emptyset \neq \Psi(w) - \bar{z} - J(w - \bar{w}) \subset o(\|w - \bar{w}\|)\mathbb{B} \ ,$$

and write $J = \nabla\Psi(\bar{w})$, *where the function* $o \colon \mathbb{R} \mapsto \mathbb{R}_+$ *depends on the point* $\bar{w}$ *and satisfies* $\lim\limits_{t \searrow 0} t^{-1}o(t) = 0$. *Finally, we say that the operator* Ψ *satisfies the quadratic growth condition at* $\bar{w}$ *if* Ψ *is differentiable at* $\bar{w}$ *and*

$$\Psi(w) - \Psi(\bar{w}) - \nabla\Psi(\bar{w})(w - \bar{w}) \subset O(\|w - \bar{w}\|^2)\mathbb{B} \ ,$$

where $O \colon \mathbb{R} \mapsto \mathbb{R}_+$ *depends on the point* $\bar{w}$ *and satisfies* $\limsup\limits_{t \searrow 0} t^{-1}|O(t)| < +\infty$.

Remarks. 1) Rockafellar [20, Theorem 2] was the first to use Lipschitz continuity to establish rates of convergence for the proximal point algorithm.
2) When the set $\Psi(\bar{w})$ is restricted to be a singleton $\{\bar{z}\}$, the differentiability of Ψ at $\bar{w}$ implies the Lipschitz continuity of Ψ at $\bar{w}$. Moreover, one can take $\alpha(\tau) \rightarrow \|J\|$ as $\tau \rightarrow 0$. This observation is verified in [20, Proposition 4].
3) This notion of differentiability corresponds to the usual notion of differentiability in the case when Ψ is single–valued.
4) It follows from the definition of monotonicity that if T is a maximal monotone operator, then the operator $\nabla T(x)$ is positive semi–definite, if it exists.
5) In [2, Sect. 4], we give an example of a convex function f for which ∂f^{-1} is Lipschitz continuous but not differentiable. In [2, Sect. 4], we show that it is possible to choose f so that ∂f^{-1} is differentiable at the origin, but does not satisfy the quadratic growth condition there.

The quadratic growth condition is a strong smoothness property. However, in the case of convex programming, this condition is weaker than the standard hypothesis used for establishing the rapid local convergence of optimization algorithms. More pointedly, the quadratic growth condition is weaker than the conditions typically employed for the local analysis of variable metric proximal point algorithms [1,5,11,12]. For convex programs, the operator T is the subdifferential of the essential objective function $\bar{f}$. In this case,

the assumption that the standard second–order sufficiency condition is satisfied at the solution to the convex program implies that the operator $\partial \bar{f}^{-1}$ satisfies the quadratic growth condition at the origin (see [20, Proposition 2] and [2, Theorem 4.3]). Thus, it is not surprising that we require a condition of this type in our local convergence analysis.

3.1. Statement of updates and convergence results

We now recall the BFGS and Broyden update formulas. The BFGS update is suitable for convex programming applications since it preserves both symmetry and positive definiteness. Broyden's update is suitable for mini-max problems since in many of these applications the Jacobian $\nabla(T^{-1})$ is non–symmetric when it exists.

Symmetric updating with BFGS:

Let $H_0 \in \mathbb{R}^{n \times n}$ be any positive definite symmetric matrix and for $k \geq 0$, set $y^k = w^k - w^{k+1}$ and $s^k = z^{k+1} - z^k$. If $y^{k^T} s^k \leq 0$, set $H_{k+1} = H_k$; otherwise, set

$$H_{k+1} = H_k + \frac{\left(s^k - H_k y^k\right)s^{k^T} + s^k\left(s^k - H_k y^k\right)^T}{y^{k^T} s^k} - \frac{\left(s^k - H_k y^k\right)^T y^k s^k s^{k^T}}{\left(y^{k^T} s^k\right)^2} . \tag{3}$$

Non-symmetric updating with Broyden's formula:

Let $H_0 = I$ and for $k \geq 0$, set $y^k = w^k - w^{k+1}$ and $s^k = z^{k+1} - z^k$. If $s^{k^T} H_k y^k = 0$, set $H_{k+1} = H_k$; otherwise, set

$$H_{k+1} = H_k + \frac{\left(s^k - H_k y^k\right)s^{k^T} H_k}{s^{k^T} H_k y^k} . \tag{4}$$

Remarks. 1. The updating formula in (3) is the formula for updating the inverse in BFGS updating. The condition $y^{k^T} s^k > 0$ in the BFGS update not only insures the existence of the inverse, but also insures that the updates are positive definite. The corresponding formula for direct approximation of $\nabla D_k(z^k)$ (when it exists) is given by

$$B_{k+1} = B_k - \frac{B_k s^k s^{k^T} B_k}{s^{k^T} B_k s^k} + \frac{y^k y^{k^T}}{y^{k^T} s^k}, \tag{5}$$

where $B_k = H_k^{-1}$ for all $k \geq 0$.
2. The updating formula in (4) is the formula for updating the inverse of Broyden's update. The condition $s^{k^T} H_k y^k \neq 0$ is satisfied if and only if the inverse Broyden updates are well–defined and nonsingular, in which case

$$A_{k+1} = A_k + \frac{\left(y^k - A_k s^k\right)s^{k^T}}{s^{k^T} s^k}, \tag{6}$$

where $A_k = H_k^{-1}$ for all $k \geq 0$.

We require that the following hypotheses hold in our convergence analysis:

(H1) The operator T^{-1} satisfies the quadratic growth condition at the origin with $J := \nabla(T^{-1})(0)$ and $T^{-1}(0) = \{\bar{z}\}$.

(H2) The approximation criteria $(\mathcal{L})$ is satisfied at every iteration.

(H3) The sequence $\{z^k\}$ converges linearly to $\bar{z}$.

(H4) There is an iteration index $\bar{k}$ such that $\lambda_k \equiv \lambda > 2\,\|J\|$ for all $k \geq \bar{k}$.

Theorem 1 (Symmetric updating).

Let $\{z^k\}$ be any sequence generated by the variable metric proximal point algorithm using the symmetric updating strategy and suppose that the hypotheses (H1)–(H4) are all satisfied. If J is symmetric, then

(i)   for all k sufficiently large $y^{k^T} s^k > 0$,
(ii) the sequences $\{\|H_k\|\}$ and $\{\|H_k^{-1}\|\}$ are bounded,
(iii) $H_k = B_k^{-1}$ for all k sufficiently large, and
(iv) the sequence converges to $\bar{z}$ at a super–linear rate.

Theorem 2 (Non–symmetric updating).

Let $\{z^k\}$ be any sequence generated by the variable metric proximal point algorithm using the non–symmetric updating strategy and suppose that the hypotheses (H1)–(H4) are all satisfied. If it is further assumed that there exists $\bar{k} > 0$ such that $\lambda_k = \lambda > 3\,\|J\|$ for all $k \geq \bar{k}$, then

(i)   there is a $\hat{k} \geq \bar{k}$ such that for all $k \geq \hat{k}$ we have $s^{k^T} H_k y^k \neq 0$ and H_k is updated using Broyden's formula,
(ii) the sequences $\{\|H_k\|\}$ and $\{\|H_k^{-1}\|\}$ are bounded, and
(iii) the sequence converges to $\bar{z}$ at a super–linear rate.

Our proofs of these results are based on an extension of the Dennis–Moré characterization theorem for superlinear convergence [8]. This result is stated below for the readers convenience.

Theorem 3 [2, Theorem 7.2, Super–Linear Convergence].

Let $\{z^k\}$ be any sequence generated by the variable metric proximal point algorithm satisfying criterion $(\mathcal{L})$ for all k. Suppose that the operator T^{-1} is differentiable at the origin with $T^{-1}(0) = \{\bar{z}\}$ and $\nabla T^{-1}(0) = J$. If $\lim_k \|D_k(z^k)\| = 0$, then $\{z^k\}$ converges to the solution $\bar{z}$ super–linearly if and only if

$$\frac{\left[I - \left(I + \frac{1}{\lambda_k}J\right)H_k^{-1}\right](z^{k+1} - z^k)}{\|z^{k+1} - z^k\|} \to 0 \quad \text{as } k \to \infty .$$

3.2. Convergence proofs

Several technical lemmas are required to prepare the way for the proofs of Theorems 1 and 2. We begin with three lemmas that depend only on structure of the algorithm and not on the specific choice of the updates $\{H_k\}$. The fact that these lemmas do not depend on the choice of the updates $\{H_k\}$ also plays an important role in the proof of Theorem 7 in Sect. 4.

Lemma 1. *Under hypotheses (H2) and (H3), there are positive numbers L_0, L_1, L_2, L_3 and L_4 such that for all k sufficiently large*

(a) $\left\| w^k \right\| \leq L_0 \left\| s^k \right\|$ *and* $\left\| w^k \right\| \leq L_0 \left\| D_k(z^k) \right\|$,

(b) $L_1 \left\| z^k - \bar{z} \right\| \leq \left\| s^k \right\| \leq L_2 \left\| z^k - \bar{z} \right\|$,

(c) $\left\| D_k(z^k) \right\| \leq L_3 \left\| s^k \right\|$,

(d) $\left\| s^{k+1} \right\| \leq L_4 \left\| s^k \right\|$, *and*

(e) $\left\| w^k - D_k(z^k) \right\| \leq \delta_k \left\| w^k \right\| \leq \delta_k L_0 \left\| s^k \right\|$.

If it is further assumed that hypotheses (H1) and (H4) hold, then for all k sufficiently large

(f) $\left\| D_k(z^k) - D_{k+1}(z^{k+1}) \right\| \geq \frac{1}{2} \left\| s^k \right\|$.

Remark. The constants L_0, L_1, L_2, L_3 and L_4 in the above lemma are closely related to each other. Due to the assumption of linear convergence, there exists $\theta \in (0, 1)$ such that $\left\| z^{k+1} - \bar{z} \right\| \leq \theta \left\| z^k - \bar{z} \right\|$ for all k sufficiently large. We show that one can take

$$L_0 = (1 - \theta)^{-2}, \ L_1 = 1 - \theta, \ L_2 = 1 + \theta, \ L_3 = (1 - \theta)^{-1}, \text{ and } L_4 = \theta\frac{1 + \theta}{1 - \theta} \ .$$

Proof. Both inequalities of (b) follow from the linear convergence of z^k to $\bar{z}$. Since there exists $0 < \theta < 1$ such that $\left\| z^{k+1} - \bar{z} \right\| \leq \theta \left\| z^k - \bar{z} \right\|$, we have

$$\left\| s^k \right\| \leq (1 + \theta)\left\| z^k - \bar{z} \right\| \quad \text{and} \quad \left\| z^k - \bar{z} \right\| \leq \frac{1}{1 - \theta}\left\| s^k \right\| \tag{7}$$

for all k sufficiently large. Hence $L_1 = 1 - \theta$ and $L_2 = 1 + \theta$.

By (H2) and Proposition 1 (c) (applied with $z = z^k$ and $z' = \bar{z}$ so that $D_k(\bar{z}) = 0$ for all k), we have

$$\left\| w^k \right\| \leq \frac{1}{1 - \delta_k}\left\| D_k(z^k) \right\| \leq \frac{1}{1 - \delta_k}\left\| z^k - \bar{z} \right\|. \tag{8}$$

Since $\delta_k \to 0$ as $k \to \infty$, we can assume that $\delta_k < \theta$ for all k sufficiently large. To see (a), just combine (7) and (8) to obtain $L_0 = (1 - \theta)^{-2}$.

The relation (c) follows from Proposition 1 (c) and the first inequality in (b) with $L_3 = 1/L_1 = (1 - \theta)^{-1}$, while (d) follows from the linear convergence of $\{z^k\}$ and both inequalities in (b) with $L_4 = \theta L_2/L_1 = \theta(1 + \theta)(1 - \theta)^{-1}$. Thus, (a)–(d) have been established.

Part (e) follows immediately from (H2) and Part (a).

We now show Part (f). For this, we only need to consider $k \geq \bar{k}$, where $\bar{k}$ is defined by (H4). Then $D_k \equiv D_\lambda$ for all such k. Clearly, $\|D_\lambda(z^k)\| \to 0$ by (H3) and (c). Hence the differentiability of T^{-1} at the origin (which is implied by (H1)) implies that for all k sufficiently large

$$T^{-1}\left(-\frac{1}{\lambda}D_\lambda(z^k)\right) - \bar{z} - J\left(-\frac{1}{\lambda}D_\lambda(z^k)\right) \subset o\big(\|D_\lambda(z^k)\|\big)\mathbb{B}\ ,\quad \text{or}$$

$$\left(I + T^{-1}\frac{1}{\lambda}\right)\big(-D_\lambda(z^k)\big) - \bar{z} + \left(I + \frac{1}{\lambda}J\right)D_\lambda(z^k) \subset o\big(\|D_\lambda(z^k)\|\big)\mathbb{B}\ .$$

But $D_\lambda(z^k) = -(I + T^{-1}\frac{1}{\lambda})^{-1}(z^k)$, hence $z^k \in (I + T^{-1}\frac{1}{\lambda})(-D_\lambda(z^k))$. This yields

$$z^k - \bar{z} + \left(I + \frac{1}{\lambda}J\right)D_\lambda(z^k) \in o\big(\|D_\lambda(z^k)\|\big)\mathbb{B}\ . \tag{9}$$

From (9) and (c), we conclude that

$$z^k - \bar{z} + \left(I + \frac{1}{\lambda}J\right)D_\lambda(z^k) \in o(\|s^k\|)\mathbb{B}\ .$$

Hence, for all k large,

$$s^k \in -\left(I + \frac{1}{\lambda}J\right)\big(D_\lambda(z^{k+1}) - D_\lambda(z^k)\big) + \big[o(\|s^k\|) + o(\|s^{k+1}\|)\big]\mathbb{B}\ .$$

Therefore, by (d) and (H4), we have

$$\|s^k\| < \frac{3}{2}\|D_{k+1}(z^{k+1}) - D_k(z^k)\| + o\big(\|s^k\|\big)\ ,$$

for all k large. This establishes (f).

$\square$

Lemma 2. *If hypotheses (H1)–(H4) hold, then $y^{k^T}s^k > 0$ for all k sufficiently large.*

Proof. By (H4), $\lambda_k \equiv \lambda$ and $D_k \equiv D_\lambda$ for all k large. Letting $z = z^{k+1}$ and $z' = z^k$ in Proposition 1 (b) and recalling that $P_\lambda = I + D_\lambda$ yields

$$\big[(I + D_\lambda)(z^{k+1}) - (I + D_\lambda)(z^k)\big]^T\big[D_\lambda(z^k) - D_\lambda(z^{k+1})\big] \geq 0\ ,\quad \text{or}$$

$$\big[z^{k+1} - z^k + D_\lambda(z^{k+1}) - D_\lambda(z^k)\big]^T\big[D_\lambda(z^k) - D_\lambda(z^{k+1})\big] \geq 0\ .$$

Hence, by (f) of Lemma 1,

$$s^{k^T}\big[D_\lambda(z^k) - D_\lambda(z^{k+1})\big] \geq \|D_\lambda(z^k) - D_\lambda(z^{k+1})\|^2 \geq \frac{1}{4}\|s^k\|^2$$

for all k sufficiently large. Therefore

$$s^{k^T}y^k \geq \frac{1}{4}\|s^k\|^2 + s^{k^T}\big(w^k - D_\lambda(z^k)\big) - s^{k^T}\big(w^{k+1} - D_\lambda(z^{k+1})\big)$$

$$\geq \frac{1}{4}\|s^k\|^2 - \|s^k\|\big[\|w^k - D_\lambda(z^k)\| + \|w^{k+1} - D_\lambda(z^{k+1})\|\big]\ . \tag{10}$$

Hence, by (10) and Lemma 1 (d) and (e),

$$s^{k^T} y^k \geq \frac{1}{4} \|s^k\|^2 - \|s^k\| \left[\delta_k L_0 \|s^k\| + \delta_{k+1} L_0 \|s^{k+1}\| \right] \geq \left(\frac{1}{4} - \delta_k L_0 - \delta_{k+1} L_0 L_4 \right) \|s^k\|^2 .$$

Now, for all k sufficiently large, $L_0(\delta_k + \delta_{k+1} L_4) < 1/4$, hence, for such k, $y^{k^T} s^k > 0$.

$\square$

Lemma 3. *If (H1)–(H4) hold, then there exist positive numbers L_5, L_6, and L_7 such that*

$$\frac{\left\| y^k - \left(I + \frac{1}{\lambda} J \right)^{-1} s^k \right\|}{\|s^k\|} \leq L_5 \|s^k\| + L_6 \delta_k + L_7 \delta_{k+1} \tag{11}$$

for all k sufficiently large. In particular,

$$\sum_{k=0}^{\infty} \frac{\left\| y^k - \left(I + \frac{1}{\lambda} J \right)^{-1} s^k \right\|}{\|s^k\|} < \infty . \tag{12}$$

Proof. By (H4), $\lambda_k \equiv \lambda$ and $D_k \equiv D_\lambda$ for all k large. By (H1) and (H3),

$$T^{-1}\left(-\frac{1}{\lambda} D_\lambda(z^k) \right) - \bar{z} - J\left(-\frac{1}{\lambda} D_\lambda(z^k) \right) \in O\left(\|D_\lambda(z^k)\|^2 \right) \mathbb{B}$$

for all k sufficiently large. Thus, as in (9), we have

$$z^k - \bar{z} + \left(I + \frac{1}{\lambda} J \right) D_\lambda(z^k) \in O\left(\|D_\lambda(z^k)\|^2 \right) \mathbb{B} \subset O\left(\|s^k\|^2 \right) \mathbb{B}$$

where the second inclusion follows from Lemma 1 (c). Therefore, by Lemma 1 (d),

$$s^k + \left(I + \frac{1}{\lambda} J \right) \left(D_\lambda(z^{k+1}) - D_\lambda(z^k) \right) \in O\left(\|s^k\|^2 \right) \mathbb{B} , \quad \text{or}$$

$$\left(I + \frac{1}{\lambda} J \right)^{-1} s^k - D_\lambda(z^k) + D_\lambda(z^{k+1}) \in O\left(\|s^k\|^2 \right) \mathbb{B} .$$

Hence, for some $L_5 > 0$,

$$\left(I + \frac{1}{\lambda} J \right)^{-1} s^k - y^k \in \left(D_\lambda(z^k) - w^k \right) - \left(D_\lambda(z^{k+1}) - w^{k+1} \right) + L_5 \|s^k\|^2 \mathbb{B} .$$

By Lemma 1 (d) and (e),

$$\left\| y^k - \left(I + \frac{1}{\lambda} J \right)^{-1} s^k \right\| \leq \left(L_0 \delta_k + L_0 L_4 \delta_{k+1} + L_5 \|s^k\| \right) \|s^k\| .$$

Therefore (11) holds for all k large.

Since $z^k \to \bar{z}$ linearly, Lemma 1 (b) implies that $\sum_{k=0}^{\infty} \|s^k\| < \infty$. Thus, (11) and the hypothesis that $\sum_{k=1}^{\infty} \delta_k < \infty$ imply (12).

$\square$

The proof of Theorem 1 now follows from a result due to Byrd and Nocedal [4].

Theorem 4 (Byrd and Nocedal, 1989). *Let $\{B_k\}$ be generated by the BFGS formula*

$$B_{k+1} = B_k - \frac{B_k s^k s^{k^T} B_k}{s^{k^T} B_k s^k} + \frac{y^{k^T} y^k}{y^{k^T} s^k} \ ,$$

where B_1 is symmetric and positive definite, and where $y^{k^T} s^k > 0$ for all k. Furthermore assume that $\{s^k\}$ and $\{y^k\}$ are such that

$$\frac{\left\| y^k - Gs^k \right\|}{\left\| s^k \right\|} \leq \epsilon_k \ , \tag{13}$$

for some symmetric and positive definite matrix G, and for some sequence $\{\epsilon_k\}$ with the property $\sum_{k=1}^{\infty} \epsilon_k < \infty$. Then $\lim_{k \to \infty} \frac{\|(B_k - G)s^k\|}{\|s^k\|} = 0$, and the sequences $\{\|B_k\|\}$, $\{\|B_k^{-1}\|\}$ are bounded.

Remark. It is important to note that the sequences $\{s^k\}$ and $\{y^k\}$ do not necessarily depend on the matrices B_k. This fact is used in the proof of our super–linear convergence result Theorem 7.

Proof of Theorem 1. By Lemma 2 and (H4), there is a $k_0 > 0$ such that $\lambda_k = \lambda > 2 \|J\|$, $y^{k^T} s^k > 0$, and $H_k = B_k^{-1}$ for all $k \geq k_0$. Consider the sequences $\{\tilde{z}^k\}$, $\{\tilde{y}^k\}$, $\{\tilde{s}^k\}$, and $\{\tilde{B}_k\}$ given by

$$\tilde{z}^k = z^{k_0 + k}, \quad \tilde{y}^k = y^{k_0 + k}, \quad \tilde{s}^k = s^{k_0 + k}, \quad \text{and} \quad \tilde{B}_k = B_{k_0 + k},$$

for $k = 0, 1, 2, \ldots$. Since $\lambda > \|J\|$ and J is symmetric, $(I + \frac{1}{\lambda}J)^{-1}$ is symmetric and positive definite. Lemma 3 implies that (13), with $\tilde{y}^k$ and $\tilde{s}^k$ replacing y^k and s^k, respectively, is satisfied with $G = (I + \frac{1}{\lambda}J)^{-1}$. Consequently, by Theorem 4 both $\{\|\tilde{B}_k\|\}$ and $\{\|\tilde{B}_k^{-1}\|\}$ are bounded and $\frac{\left\| (\tilde{B}_k - (I + \frac{1}{\lambda}J)^{-1})\tilde{s}^k \right\|}{\|\tilde{s}^k\|} \to 0$, or equivalently, both $\{\|B_k\|\}$ and $\{\|B_k^{-1}\|\}$ are bounded and

$$\frac{\left\| \left(I - (I + \frac{1}{\lambda}J)B_k\right)s^k \right\|}{\|s^k\|} \to 0 \ . \tag{14}$$

Hypothesis (H3) implies that $z^k \to \bar{z}$ at a linear rate. Therefore, by Lemma 1 Parts (b) and (c), we have that $D_k(z^k) \to 0$ at a linear rate as well. By combining this fact with (14), Theorem 3 can be applied to find that $z^k \to \bar{z}$ at a super–linear rate.
$\square$

The proof of Theorem 2 uses two more technical lemmas.

Lemma 4 [9, Lemma 8.2.5]. *Let $s \in \mathbb{R}^n$ be nonzero, $E \in \mathbb{R}^{n \times n}$, and let $\|\cdot\|_F$ denote the Frobenious norm, then*

$$\left\| E\left(I - \frac{ss^T}{s^T s}\right) \right\|_F = \left(\|E\|_F^2 - \left(\frac{\|Es\|}{\|s\|}\right)^2 \right)^{1/2} \leq \|E\|_F - \frac{1}{2\|E\|_F}\left(\frac{\|Es\|}{\|s\|}\right)^2 \ .$$

The proof of the next lemma follows the line of proof given in [9, Lemma 8.2.1] and so is omitted.

Lemma 5. *Let A_0, A_1, ... , A_k be generated by the Broyden update formula (6). Then for any matrix G we have*

$$A_{k+1} - G = (A_k - G)\left(I - \frac{s^k s^{k^T}}{s^{k^T} s^k}\right) + \frac{(y^k - Gs^k)s^{k^T}}{s^{k^T} s^k} \quad and \quad (15)$$

$$\|A_{k+1} - G\| \leq \|A_0 - G\| + \sum_{j=0}^{k} \frac{\|y^j - Gs^j\|}{\|s^j\|} . \quad (16)$$

Proof of Theorem 2. Set $G = (I + \frac{1}{\lambda}J)^{-1}$. Then, by the Banach Lemma,

$$\|I - G\| \leq \sum_{i=1}^{\infty} \left(\frac{1}{\lambda}\|J\|\right)^i < \frac{1}{2} , \quad (17)$$

hence $1/2 - \|I - G\| > 0$. By (12), there is a $k_0 \geq \bar{k}$ such that

$$\sum_{j=k_0}^{\infty} \frac{\|y^j - Gs^j\|}{\|s^j\|} \leq \left(\frac{1}{2} - \|I - G\|\right) . \quad (18)$$

To show (i), we need only show that there exists a $\hat{k} \geq k_0$ such that the matrices A_k defined in (6) are nonsingular for all $k \geq \hat{k}$. If we cannot take $\hat{k} = k_0$, then there is a $\hat{k} > k_0$ such that $H_{\hat{k}} = I$ (i.e. $s^{\hat{k}-1^T} H_{\hat{k}-1} y^{\hat{k}-1} = 0$). We claim that for this choice of $\hat{k}$ the matrices A_k are non-singular for all $k \geq \hat{k}$. To see this, note that for all $k \geq \hat{k}$

$$\|A_k - I\| \leq \|I - G\| + \|A_k - G\|$$

$$\leq 2\|I - G\| + \sum_{j=\hat{k}}^{k-1} \frac{\|y^j - Gs^j\|}{\|s^j\|}$$

$$< \frac{1}{2} + \|I - G\| < 1$$

by Lemma 5 (replacing A_0, A_1, ... , A_k by $A_{\hat{k}}$, $A_{\hat{k}+1}$, ... , A_{k-1} and recalling that $A_{\hat{k}} = I$), (18), and (17). Therefore, A_k is non–singular for all $k \geq \hat{k}$ with

$$\|A_k\| \leq 2 \quad and \quad \|A_k^{-1}\| \leq \frac{1}{\frac{1}{2} - \|I - G\|},$$

which also verifies (ii).

We now show (iii). Set $E_k := A_k - G$ and $\sigma_k := \frac{\|y^k - Gs^k\|\,\|s^k\|}{\|s^k\|^2}$, and recall that for any vectors $u, v \in \mathbb{R}^n$, $\|uv^T\|_F = \|u\|\,\|v\|$. Let $k \geq \hat{k}$. From (15), we have

$$\|E_{k+1}\|_F \leq \left\|E_k\left(I - \frac{s^k s^{k^T}}{s^{k^T} s^k}\right)\right\|_F + \frac{\|y^k - Gs^k\|\,\|s^k\|}{\|s^k\|^2} = \left\|E_k\left(I - \frac{s^k s^{k^T}}{s^{k^T} s^k}\right)\right\|_F + \sigma_k .$$

By Lemma 4,

$$\|E_{k+1}\|_F \leq \|E_k\|_F - \frac{1}{2\|E_k\|_F}\left(\frac{\|E_k s^k\|}{\|s^k\|}\right)^2 + \sigma_k , \text{ or}$$

$$\frac{\|E_k s^k\|^2}{\|s^k\|^2} \leq 2\|E_k\|_F\left(\|E_k\|_F - \|E_{k+1}\|_F + \sigma_k\right) \leq \bar{L}\left(\|E_k\|_F - \|E_{k+1}\|_F + \sigma_k\right) .$$

Note that $\bar{L}$ does not depend on k due to (16) and (12). Hence

$$\sum_{k=\hat{k}}^{N} \frac{\|E_k s^k\|^2}{\|s^k\|^2} \leq \bar{L}\left(\|E_0\|_F - \|E_{N+1}\|_F + \sum_{k=\hat{k}}^{N}\sigma_k\right) .$$

From (12), this means

$$\sum_{k=0}^{\infty} \frac{\|E_k s^k\|^2}{\|s^k\|^2} < \infty ,$$

and so

$$\frac{\left\|\left(A_k - \left(I + \frac{1}{\lambda}J\right)^{-1}\right)s^k\right\|}{\|s^k\|} \to 0 . \tag{19}$$

Since $H_k = A_k^{-1}$ for $k \geq \hat{k}$, the result follows from Theorem 3.

$\square$

4. Application to the Chen–Fukushima algorithm

We now give an example of how the results of the previous section can be applied to obtain a local super–linear convergence result for a VMPPA that uses matrix secant updating. The example is based on the Chen–Fukushima VMPPA [5]. There are three basic steps in the application process: first, one must verify that the local approximation criteria ($\mathcal{L}$) is satisfied (Proposition 2); second, one must establish the global linear convergence of the iterates (Theorem 6); and third, the updating strategy must be designed so that eventually the algorithm generates the same iterates as the algorithm described in Sect. 3.

The Chen–Fukushima [5] VMPPA is designed to solve the convex programming problem $\mathcal{P}$, where $f: \mathbb{R}^n \mapsto \mathbb{R}$ is a lower semi–continuous convex function. In this context, the VMPPA can be viewed as a variable metric algorithm applied to f_λ, the Moreau–Yosida regularization of f. For this reason, the BFGS updating strategy is used to approximate the second–order behavior of f_λ.

The Chen–Fukushima algorithm uses bundle techniques [6,10] to approximate both f_λ and its gradient. Global convergence is obtained with the aid of a line–search procedure based on the function f rather than f_λ. This is an exceptional feature of the method since approximating f_λ can be costly. Chen and Fukushima characterize the super–linear convergence of their VMPPA in a result that parallels the landmark result

of Dennis and Moré [8]. In addition, they state that the BFGS updating strategy can be used to approximate the second–order behavior of f_λ. However, the convergence theory in [5] applies to general variable metric updating with no explicit dependence on quasi-Newton techniques. In particular, Chen and Fukushima do not show that BFGS updating yields super–linear convergence. We now fill this gap.

As in Sect. 3, we assume that the proximation parameter λ_k remains fixed from some point on. For simplicity, we assume $\lambda_k \equiv \lambda$ throughout the iteration process.

4.1. Statement of the algorithm

The algorithm in [5] is doubly iterative in the sense that each major iteration k consists of a finite sequence of inner iterations to approximate $f_\lambda(z^k)$ and $D(z^k) = -\nabla f_\lambda(z^k)$. Denoting z^k as the current iterate of the major iteration, the inner iterations approximately solve the problems

$$\min_{u \in \mathbb{R}^n} \lambda f(u) + \frac{1}{2} \left\| u - z^k \right\|^2 \tag{20}$$

by generating a sequence $\{u^j\}$ as follows: set $u^0 = z^k$ and choose $g^0 \in \partial f(z^k)$; for $j = 1, 2, \ldots$, let u^j be the (unique) solution of the problem

$$\min_{u \in \mathbb{R}^n} \lambda f_{k,j}(u) + \frac{1}{2} \left\| u - z^k \right\|^2, \tag{21}$$

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

$$f_{k,j}(u) := \max_{i=0,1,\ldots,j-1} [f(u^i) + \langle g^i, u - u^i \rangle] \tag{22}$$

with $g^i \in \partial f(u^i)$ for $i = 0, 1, \ldots, j - 1$. The pairs (u^i, g^i), $i = 0, 1, \ldots, j - 1$, constitute a *bundle* generated sequentially starting from $u^0 = z^k$ and $g^0 \in \partial f(z^k)$. Note that (21) is equivalent to the quadratic programming problem

$$\begin{aligned} &\min_{(u,v) \in R^{n+1}} \lambda v + \frac{1}{2} \left\| u - z^k \right\|^2 \\ &\text{subject to} \quad f(u^i) + \langle g^i, y - y^i \rangle \leq v, \ i = 0, 1, \ldots, j - 1. \end{aligned} \tag{23}$$

The inner iteration is terminated if

$$f(u^j) \leq f(z^k) - \sigma_k \big(f(z^k) - f_{k,j}(u^j) \big), \tag{24}$$

where $\sigma_k \in (0, 1)$ is pre–specified.

We now state the Chen–Fukushima algorithm.

Algorithm [5, Algorithm 2.1]. *Choose an initial point $z^0 \in \mathbb{R}^n$, parameters $\sigma, \rho, \gamma \in (0, 1)$, a sufficiently large constant $M \geq f(z^0)$, and a sequence $\{\sigma_k\}$ such that $\sigma < \sigma_k < 1$. Let $k := 0$.*

1. *Approximate the solution to subproblem (20) by the procedure (22)–(23) to obtain a point $u^{j(k)}$ satisfying (24). Let $p^a(z^k) := u^{j(k)}$.*
2. *Let $w^k := p^a(z^k) - z^k$.*

3. *Choose a symmetric positive definite matrix $H_k \in I\!\!R^{n \times n}$.*
4. *Let $d^k := (H_k - I)w^k$. If $k = 0$, let $\eta_1 := \|w^0\|$ and go to Step 5. For $k \geq 1$, if $\|w^k\| \leq \rho\eta_k$ and $f(p^a(z^k) + d^k) \leq M$, let $\tau_k := 1$ and $\eta_{k+1} := \|w^k\|$ and go to Step 6; otherwise, let $\eta_{k+1} := \eta_k$ and go to Step 5.*
5. *Set $\tau_k := \gamma^{m_k}$, where m_k is the smallest nonnegative integer m such that*

$$f(p^a(z^k) + \gamma^m d^k) \leq f(z^k) - \frac{\gamma^m \sigma}{\lambda} \|w^k\|^2 .$$

6. *Set $z^{k+1} := p^a(z^k) + \tau_k d^k$, $k := k + 1$ and return to Step 1.*

Remarks. (1) In Step 3 of their algorithm, Chen and Fukushima specify that the matrix H_k should be constructed using the quasi–Newton formula $H_k y^{k-1} = s^{k-1}$ with $y^{k-1} := w^{k-1} - w^k$ and $s^{k-1} := z^k - z^{k-1}$. However, they go on to explain in the second paragraph following the statement of the algorithm that quasi–Newton updating is not required for any of their convergence analysis. All that is required in Step 3 is that the matrices H_k are symmetric and positive definite. The only connection between the Chen-Fukushima algorithm and quasi-Newton methods is that the BFGS update can be implemented in a way that preserves the symmetry and positive definiteness of the updates.

(2) We have $z^{k+1} = z^k + \tilde{H}_k w^k$ where $\tilde{H}_k = (\tau_k H_k + (1 - \tau_k)I)$. If we regard w^k as an approximation to $D_k(z^k)$, then this algorithm is an instance of the VMPPA given in Sect. 2, and, when $\tau_k = 1$, this update is identical to an update generated by the variable metric proximal point method proposed in Sect. 3. One of our objectives is to show that eventually the step length τ_k always takes the value 1 since then the local analysis developed in Sect. 3 applies.

(3) For the reader's convenience, we relate our notation to that used in [5]: the objects λH_k, λ, w^k, and p^a given above correspond to the objects B_k^{-1}, $1/\mu$, $-v_k/\mu$, and p^a as defined in [5], respectively.

It is shown in [5] that the algorithm is well–defined and that the procedures in Steps 1 and 5 of the algorithm are finitely terminating as long as z^k is not a solution to $\mathcal{P}$. Chen and Fukushima give the following global convergence result.

Theorem 5 [5, Theorem 2.1]. *Assume that the convex function f has a nonempty bounded set of minima and that the sequence $\{\|H_k w^k\|/\|w^k\|\}$ is bounded. Let the sequence $\{z^k\}$ be generated by the algorithm. Then all accumulation points of the sequences $\{z^k\}$ and $\{p^a(z^k)\}$ minimize f.*

Remarks. (1) The exact statement of Theorem 2.1 in [5] is that the algorithm terminates finitely. Finite termination occurs since the iteration process is stopped if it is ever the case that $f(z^k) - f_{k,j}(p^a(z^k)) \leq \kappa$ for a given value of $\kappa > 0$. It is explained in the first paragraph of [5, Sect. 3] that one may set $\kappa = 0$ in which case all accumulation points of the sequences $\{z^k\}$ and $\{p^a(z^k)\}$ minimize f.

(2) In Theorem 5, we have replaced the hypothesis that the sequence $\{\|H_k\|\}$ is bounded with the weaker condition that the sequence $\{\|H_k w^k\|/\|w^k\|\}$ is bounded. The proof given in [5] remains valid with this replacement.

Our local analysis assumes that the operator $(\partial f)^{-1}$ is differentiable at the origin. Since $(\partial f)^{-1} = \partial f^*$, where f^* is the convex conjugate of f, this implies that the origin is in the interior of the domain of f^*. This in turn implies that f is *inf–compact* [19, Theorem 27.1], that is, the sets $\{z \in \mathbb{R}^n \ : \ f(z) \leq \alpha\}$ are bounded for all $\alpha \in \mathbb{R}$. In this case, the iterates remain bounded, and so there is at least one accumulation point. Since this accumulation point must be the unique global minimizer, it follows that the entire sequence converges to the unique minimizer of f.

Corollary 1. *If, in addition to the hypotheses of Theorem 5, it is assumed that $(\partial f)^{-1}$ is differentiable at the origin, then the sequence generated by the Chen–Fukushima algorithm converges to the unique global minimizer of f with both $\{D(z^k)\}$ and $\{w^k\}$ converging to zero.*

Remark. The statement concerning the sequences $\{D(z^k)\}$ and $\{w^k\}$ is inserted to facilitate the discussion to follow. The convergence of $D(z^k)$ to the origin follows from the continuous differentiability of f_λ, while the convergence of $\{w^k\}$ to the origin follows from [6, Proposition 3].

4.2. Linear convergence

In [5], linear convergence is established using techniques from the theory of non–smooth equations [16]. Chen and Fukushima assume that the function f is strongly convex, the operator D is *semi–smooth* [16], and the matrices H_k satisfy a refined approximation criteria. However, they note that the strong convexity hypothesis is stronger than required. All that is needed is the Lipschitz continuity of the derivative of f^* near the origin, whereas the strong convexity of f implies that the conjugate f^* is differentiable with globally Lipschitz continuous derivative (with Lipschitz constant equal to the modulus of strong convexity).

We present a different approach to linear convergence based on the techniques developed in [2, Theorem 7.1]. This gives us ready access to the results of Sect. 3. The first step in our analysis is to refine the termination criteria used in the inner iteration of Step 1. In addition to the termination criteria (24), we require that the inner iteration terminate only if the condition

$$f(u^j) \leq f_{k,j}(u^j) + \frac{\delta_k^2}{2\lambda}\left\|u^j - z^k\right\|^2, \tag{25}$$

is also satisfied. We now show that this additional requirement on termination in the inner iteration guarantees that the local convergence criteria $(\mathcal{L})$ is satisfied at each iteration.

Proposition 2. *(i) If z^k does not solve $\mathcal{P}$, then the procedure (22)–(23) produces u^j satisfying both (24) and (25) in a finite number of steps.*
(ii) If the termination criterion (25) is satisfied, then we have

$$\left\|w^k - D(z^k)\right\| \leq \delta_k \left\|w^k\right\|, \tag{26}$$

which implies that

$$(1 - \delta_k)\|w^k\| \le \|D(z^k)\| \le (1 + \delta_k)\|w^k\|. \tag{27}$$

Proof. (i) The fact that condition (24) can be satisfied finitely is established in [5, Proposition 2.1]. That condition (25) can also be satisfied finitely follows immediately from [6, Proposition 3].

(ii) First note that by (25) and [6, Proposition 3], we have

$$\lambda f(p^a(z^k)) + \frac{1}{2}\|w^k\|^2 \le \lambda f_{k,j(k)}(p^a(z^k)) + \frac{1 + \delta_k^2}{2}\|w^k\|^2$$

$$\le f_\lambda(z^k) + \frac{\delta_k^2}{2}\|w^k\|^2. \tag{28}$$

Next, let

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

Then θ is strongly convex with modulus 1. The subdifferential inequality yields the inequality

$$\theta(u) \ge \theta(v) + \langle g, u - v \rangle + \frac{1}{2}\|u - v\|^2 \quad \forall u, v \in \mathbb{R}^n$$

whenever $g \in \partial\theta(v)$. Let $u = p^a(z^k)$ and $v = p(z^k) := D(z^k) + z^k$, we have $\theta(u) = \lambda f(p^a(z^k)) + \frac{1}{2}\|w^k\|^2$ and $\theta(v) = f_\lambda(z^k)$. Also notice that $0 \in \partial\theta(p(z^k))$. Hence we have

$$\lambda f(p^a(z^k)) + \frac{1}{2}\|w^k\|^2 \ge f_\lambda(z^k) + \frac{1}{2}\|p^a(z^k) - p(z^k)\|^2$$

$$= f_\lambda(z^k) + \frac{1}{2}\|w^k - D(z^k)\|^2. \tag{29}$$

By combining (28) and (29) we obtain (26) which easily yields (27).

$\square$

We now establish conditions for the linear convergence of the Chen–Fukushima algorithm.

Theorem 6. *Suppose that $(\partial f)^{-1}$ is differentiable at the origin with $(\partial f)^{-1}(0) = \{\bar{z}\}$ and $\nabla(\partial f)^{-1}(0) = J$. Let $\{z^k\}$ be any sequence generated by the algorithm such that in Step 1, in addition to (24), the termination criterion (25) must also be satisfied on each outer iteration. In addition, assume that $z^k \ne \bar{z}$ for all $k = 0, 1, 2, \dots$. Define*

$$\beta_k := \frac{\|(H_k - (I + \frac{1}{\lambda}J))w^k\|}{\|w^k\|}. \tag{30}$$

Assume that there is a $\bar{k} > 0$ such that for all $k \ge \bar{k}$

$$\delta_k \le \delta < \frac{1}{2}, \tag{31}$$

and there is an $\bar{\epsilon} > 0$ such that

$$\bar{\epsilon} + \beta_k + \left(1 + \frac{\|J\|}{\lambda}\right)\delta \le \rho(1 - \delta) , \tag{32}$$

where $\delta = \delta_{\bar{k}}$ and ρ is the parameter used in Step 4 of the algorithm. Then there exists an index $\hat{k}$ such that for all $k \ge \hat{k}$, we have $\tau_k \equiv 1$ and

$$\left\|z^{k+1} - \bar{z}\right\| \le \rho\left\|z^k - \bar{z}\right\| ,$$

that is, the convergence rate is linear.

The proof of Theorem 6 requires the following technical lemma.

Lemma 6. *Under the hypotheses of Theorem 6 we have*

(a) $(\partial f)^{-1}(\frac{-1}{\lambda}D(z^k)) - \bar{z} - J(\frac{-1}{\lambda}D(z^k)) \subset o(\|D(z^k)\|)\,I\!\!B$ *for all k sufficiently large, and*

(b) *if $\tau_k = 1$, then $\left\|(I + \frac{1}{\lambda}J)H_k^{-1}(z^{k+1} - (I + H_kD)(z^k))\right\| \le \delta_k(1 + \frac{\|J\|}{\lambda})\|w^k\|$ and $(I + \frac{1}{\lambda}J)H_k^{-1}(z^{k+1} - (I + H_kD)(z^k)) \in O(\delta_k\|D(z^k)\|)\,I\!\!B$.*

Proof. First note that conditions (30) and (32) imply that $\{\|H_kw^k\|/\|w^k\|\}$ is bounded. Hence the conditions in Theorem 5 and Corollary 1 are all satisfied. In particular, $z^k \to \bar{z}$ with $D(z^k) \to 0$ and $w^k \to 0$. Let $\bar{\delta} > 0$ be such that

$$(\partial f)^{-1}(v) - \bar{z} - Jv \subset o(\|v\|)\,I\!\!B \tag{33}$$

whenever $\|v\| < \bar{\delta}$. Let $\bar{k}_1$ be such that whenever $k > \bar{k}_1$, $\|D(z^k)\| \le \bar{\delta}$. When $k > \bar{k}_1$, the inclusion (33) implies that

$$(\partial f)^{-1}\left(\frac{-1}{\lambda}D(z^k)\right) - \bar{z} + J\left(\frac{1}{\lambda}D(z^k)\right) \subset o(\|D(z^k)\|)\,I\!\!B ,$$

which proves (a).

We now show (b). If $\tau_k = 1$, then $z^{k+1} = z^k + H_kw^k$. Hence

$$H_k^{-1}(z^{k+1} - (I + H_kD)(z^k)) = H_k^{-1}(H_kw^k - H_kD(z^k)) = w^k - D(z^k) .$$

By Proposition 2,

$$\left\|w^k - D(z^k)\right\| \le \delta_k\left\|w^k\right\| .$$

Therefore,

$$\left\|\left(I + \frac{1}{\lambda}J\right)H_k^{-1}(z^{k+1} - (I + H_kD)(z^k))\right\| \le \left(1 + \frac{\|J\|}{\lambda}\right)\left\|w^k - D(z^k)\right\|$$

$$\le \delta_k\left(1 + \frac{\|J\|}{\lambda}\right)\|w^k\| \le \frac{\delta_k}{1 - \delta_k}\left(1 + \frac{\|J\|}{\lambda}\right)\|D(z^k)\| ,$$

or equivalently,

$$\left(I + \frac{1}{\lambda}J\right)H_k^{-1}(z^{k+1} - (I + H_kD)(z^k)) \in O(\delta_k\|D(z^k)\|)\,I\!\!B .$$

$\square$

Proof of Theorem 6. Conditions (30) and (32) imply that $\{\|H_k w^k\|/\|w^k\|\}$ is bounded. Hence the conditions in Theorem 5 and Corollary 1 are all satisfied. In particular, $z^k \to \bar{z}$ with $D(z^k) \to 0$ and $w^k \to 0$.

Let $\bar{k}_1 \geq \bar{k}$ be such that Part (a) of Lemma 6 holds for all $k \geq \bar{k}_1$. Then, by Lemma 6(a), (27), and (31), there must exist a $\tilde{k} > \bar{k}_1$ such that for $k > \tilde{k}$,

$$
\left\| (\partial f)^{-1}\left(\frac{-1}{\lambda}D(z^k)\right) - \bar{z} + J\left(\frac{1}{\lambda}D(z^k)\right) \right\| \leq \delta\bar{\epsilon}\|D(z^k)\| \leq \delta(1+\delta)\bar{\epsilon}\|w^k\| \leq \bar{\epsilon}\|w^k\| .
\tag{34}
$$

Since $z^k \to \bar{z}$ and $\|w^k\| \to 0$, Step 4 of the algorithm implies that there must be a $\hat{k} > \tilde{k}$ such that $\|w^{\hat{k}}\| \leq \rho\eta_{\hat{k}}$, i.e., $\tau_{\hat{k}} = 1$ and $z^{\hat{k}+1} = z^{\hat{k}} + H_{\hat{k}}w^{\hat{k}}$. Let $\tilde{z}^{\hat{k}+1} := (I + H_{\hat{k}}D)(z^{\hat{k}})$, or equivalently, $\tilde{z}^{\hat{k}+1} = z^{\hat{k}} - H_{\hat{k}}(I + (\partial f)^{-1}\frac{1}{\lambda})^{-1}(z^{\hat{k}})$ (by Proposition 1 (a)). Therefore,

$$
\begin{aligned}
z^{\hat{k}} &\in \left(I + (\partial f)^{-1}\frac{1}{\lambda}\right)\left[H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1})\right] \\
&= H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1}) + (\partial f)^{-1}\left[\frac{1}{\lambda}H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1})\right] .
\end{aligned}
$$

By re–arranging this inclusion, we obtain the inclusion

$$
\begin{aligned}
z^{\hat{k}+1} - \bar{z} &= z^{\hat{k}} - \bar{z} + (z^{\hat{k}+1} - z^{\hat{k}}) \\
&\in (\partial f)^{-1}\left(\frac{1}{\lambda}H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1})\right) - \bar{z} + (z^{\hat{k}+1} - z^{\hat{k}}) + H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1}) \\
&= \left[(\partial f)^{-1}\left(\frac{1}{\lambda}H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1})\right) - \bar{z} - J\left(\frac{1}{\lambda}H_{\hat{k}}^{-1}(z^{\hat{k}} - \tilde{z}^{\hat{k}+1})\right)\right] \\
&\quad + \left[I - \left(I + \frac{1}{\lambda}J\right)H_{\hat{k}}^{-1}\right](z^{\hat{k}+1} - z^{\hat{k}}) \\
&\quad + \left(I + \frac{1}{\lambda}J\right)H_{\hat{k}}^{-1}(z^{\hat{k}+1} - \tilde{z}^{\hat{k}+1}) \\
&= \left[(\partial f)^{-1}\left(\frac{-1}{\lambda}D_{\hat{k}}(z^{\hat{k}})\right) - \bar{z} - J\left(\frac{-1}{\lambda}D_{\hat{k}}(z^{\hat{k}})\right)\right] \\
&\quad + \left[I - \left(I + \frac{1}{\lambda}J\right)H_{\hat{k}}^{-1}\right](z^{\hat{k}+1} - z^{\hat{k}}) \\
&\quad + \left(I + \frac{1}{\lambda}J\right)H_{\hat{k}}^{-1}(z^{\hat{k}+1} - \tilde{z}^{\hat{k}+1}) .
\end{aligned}
$$

Now consider the final three terms in the sum on the right hand side of this inclusion. By (34), the first of these terms is bounded by $\bar{\epsilon}\|w^{\hat{k}}\|$. The definition of β_k bounds the second by $\beta_{\hat{k}}\|w^{\hat{k}}\|$ and Lemma 6(b) bounds the third by $(1 + \frac{\|J\|}{\lambda})\delta\|w^{\hat{k}}\|$. Therefore,

$$
\|z^{\hat{k}+1} - \bar{z}\| \leq \left(\bar{\epsilon} + \beta_{\hat{k}} + \left(1 + \frac{\|J\|}{\lambda}\right)\delta\right)\|w^{\hat{k}}\| .
\tag{35}
$$

But then

$$\left\| w^{\hat{k}+1} \right\| \leq \tfrac{1}{1-\delta} \left\| D(z^{\hat{k}+1}) \right\| \qquad \text{(by (27))}$$

$$\leq \tfrac{1}{1-\delta} \left\| z^{\hat{k}+1} - \bar{z} \right\| \qquad \text{(Proposition 1 (c))}$$

$$\leq \tfrac{\bar{\epsilon} + \beta_{\hat{k}} + (1 + \frac{\|J\|}{\lambda})\delta}{1-\delta} \left\| w^{\hat{k}} \right\| \qquad \text{(by (35))}$$

$$\leq \rho \left\| w^{\hat{k}} \right\| \qquad \text{(by (32))}.$$

Hence $\tau_{\hat{k}+1} = 1$. Proceeding as above, we find that this implies that $\tau_k = 1$ for all $k \geq \hat{k}$. Therefore,

$$\left\| z^{k+1} - \bar{z} \right\| \leq \rho(1-\delta) \left\| w^k \right\| \qquad \text{(by (32) and (35))}$$

$$\leq \rho \left\| D(z^k) \right\| \qquad \text{(by (27))}$$

$$\leq \rho \left\| z^k - \bar{z} \right\|, \qquad \text{(Proposition 1 (c))}$$

for all $k \geq \hat{k}$.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

4.3. Super–linear convergence with BFGS updating

So far, our discussion of the Chen–Fukushima algorithm has been independent of quasi–Newton updating and the BFGS update. Other than the hypothesis on the boundedness of the sequence $\{\|H_k w^k\|/\|w^k\|\}$ in Theorem 5 and the hypothesis on the parameters β_k in Theorem 6, our discussion thus far only assumes that the matrices H_k are symmetric and positive definite. We now bring the BFGS update into the discussion by specifying precisely how it is to be used in the selection of the matrices H_k in Step 3 of the Chen–Fukushima algorithm.

BFGS updating in Step 3 of the Chen–Fukushima algorithm:

Choose $0 < \hat{\epsilon} < 0.1\rho$ where ρ is defined in the algorithm. Choose $H_0 = \hat{H}_0 = I$. For $k \geq 0$, set $y^k = w^k - w^{k+1}$, $s^k = z^{k+1} - z^k$, and

$$\hat{H}_{k+1} = \begin{cases} \hat{H}_k + \frac{(s^k - \hat{H}_k y^k)s^{k^T} + s^k(s^k - \hat{H}_k y^k)^T}{\langle y^k, s^k \rangle} - \frac{\langle s^k - \hat{H}_k y^k, y^k \rangle s^k s^{k^T}}{\langle y^k, s^k \rangle^2} & , \text{ if } y^{k^T} s^k > 0, \\ \hat{H}_k & , \text{ otherwise.} \end{cases}$$

Then set

$$H_{k+1} = \begin{cases} \hat{H}_{k+1} & , \text{ if } \left\| (I - \hat{H}_{k+1})w^{k+1} \right\| \leq \frac{\rho - \hat{\epsilon}}{2} \left\| w^{k+1} \right\|, \\ I & , \text{ otherwise.} \end{cases}$$

In our final result, we use the results of Sect. 3 to show that this updating scheme produces iterates that converge super–linearly to a solution of the inclusion $0 \in \partial f(z)$ under suitable local hypotheses. The proof has two stages. The first stage establishes that hypotheses (H1)–(H4) are satisfied with the majority of the effort devoted to showing (H3), i.e., the iterates converge linearly. For this we apply Theorem 6. The first stage

of the proof does not make use of the fact that the $\hat{H}_k$'s are being updated using the BFGS formula. This stage is a global convergence result that only requires that each iterate not deviate too much from the corresponding iterate produced by an iteration of the classical proximal point algorithm. Deviation from a classical proximal point iterate is regulated by the condition

$$\left\| (I - H_k)w^k \right\| \leq \frac{\rho - \hat{\epsilon}}{2} \left\| w^k \right\|, \tag{36}$$

which is required to hold on every iteration. If $H_k \neq \hat{H}_k$, then $\hat{H}_k$ failed to satisfy (36), in which case $H_k = I$ and z^{k+1} is the result of a classical proximal point iteration since $d^k = 0$. Condition (36), with H_k replaced by $\hat{H}_k$, acts as a switch that toggles the iteration process between a globally convergent first–order method (the classical proximal point algorithm) and a locally convergent second–order method (the VMPPA of Sect. 2).

An equally important consequence of Theorem 6 is that $\tau_k \equiv 1$ for all k greater than some $\tilde{k}$. That is, eventually the unit step is always accepted and no line search is performed. The significance of this is that from iteration $\tilde{k}$ onward the algorithm is an instance of the VMPPA introduced in Sect. 2. Consequently, the technical results Lemma 1, Lemma 2, and Lemma 3 can be applied to reveal further insight into the behavior of the iteration process.

In the second stage of the proof, we show that $H_k = \hat{H}_k$ from some iteration $\hat{k}$ onward. This, combined with the fact that eventually the line search parameter τ_k always has the value 1, implies that there is an iteration k_0 such that from k_0 onward the iterates generated by the Chen–Fukushima algorithm are identical to the iterates generated by the algorithm of Sect. 3 initiated at z^{k_0}. Once this is accomplished, then the superlinear convergence of the iterates follows from Theorem 1.

The second stage of the proof hinges on the theorem of Byrd and Nocedal (Theorem 4). Here it is important to note that the sequences $\{y^k\}$ and $\{s^k\}$ in Theorem 4 need not depend on the sequence $\{B_k\}$ (or, equivalently, $\{\hat{H}_k\}$). Consequently, we can derive properties of the sequence $\{\hat{H}_k\}$ even though $\hat{H}_k$ may not have been used in the computation of s^k. We then use these properties to show that eventually it must be the case that $H_k = \hat{H}_k$. This technique of proof is the reason why we continue to update the sequence $\hat{H}_k$ even though it may not be used in the computation of s^k.

Theorem 7. *Let $\rho \in (0, 1)$ be as in the statement of the Chen–Fukushima algorithm and choose $\hat{\epsilon} \in (0, 0.1\rho)$. Suppose that the following conditions are satisfied:*

(i)   The operator $(\partial f)^{-1}$ satisfies the quadratic growth condition at the origin with

$$(\partial f)^{-1}(0) = \{\bar{z}\} \quad and \quad \nabla(\partial f)^{-1}(0) = J.$$

(ii) $\lambda > \frac{2}{\rho - \hat{\epsilon}} \|J\|$.

(iii) In Step 1 of the Chen–Fukushima algorithm, the point $u^{j(k)}$ must satisfy both of the conditions (24) and (25) where the non–increasing sequence $\{\delta_k\}$ is chosen to satisfy $\sum_{k=0}^{\infty} \delta_k < \infty$.

Let $\{z^k\}$ be any sequence generated by the Chen–Fukushima algorithm using the BFGS updating strategy formulated above for Step 3 of the algorithm. Assume that $z^k \neq \bar{z}$ for all $k = 0, 1, 2, \ldots$. Then hypotheses (H1)–(H4) of Sect. 3.1 are satisfied with $T = \partial f$ and there is an iteration k_0 such that

(a) $y^{k^T} s^k > 0$ *for all* $k \geq k_0$,
(b) $H_k = \hat{H}_k$ *for all* $k \geq k_0$,
(c) the sequences $\{\|H_k\|\}$ *and* $\{\|H_k^{-1}\|\}$ *are bounded, and*
(d) the sequence $\{z^k\}$ *converges to* $\bar{z}$ *at a super–linear rate.*

Proof. Hypotheses (H1) follows from hypotheses (i), hypothesis (H4) follows from hypothesis (ii) and the choice of $\hat{\epsilon}$ and ρ, and hypothesis (H2) follows from (iii) and Proposition 2 (ii). We now proceed to show that (H3) follows from Theorem 6. To apply Theorem 6, we must show that (32) is satisfied for all k sufficiently large where the sequence $\{\beta_k\}$ is defined in (30). First note that the BFGS updating strategy for the Chen–Fukushima algorithm guarantees that

$$\left\|(H_k - I)w^k\right\| \leq \frac{\rho - \hat{\epsilon}}{2}\|w^k\| \tag{37}$$

for all k. In particular, this implies that the sequence $\{\|H_k w^k\|/\|w^k\|\}$ is bounded by $(1+\frac{\rho-\epsilon}{2})$. Hypothesis (H1) implies that $(\partial f)^{-1}(0) = \{\bar{z}\}$. Therefore, Corollary 1 implies that

$$z^k \to \bar{z}, \quad D(z^k) \to 0, \quad \text{and} \quad w^k \to 0 .$$

By (37) and (ii), we have

$$\left\|\left(H_k - \left(I + \frac{1}{\lambda}J\right)\right)w^k\right\| \leq \left\|(H_k - I)w^k\right\| + \frac{\|J\|}{\lambda}\|w^k\| \leq (\rho - \hat{\epsilon})\|w^k\| .$$

Hence $\beta_k \leq \rho - \hat{\epsilon} < \rho - \frac{\hat{\epsilon}}{2}$. But then, since $\delta_k \to 0$, condition (32) is eventually satisfied for all k sufficiently large with $\bar{\epsilon} = \frac{\hat{\epsilon}}{2}$. We can now apply Theorem 6 to say that $z^k \to \bar{z}$ at a linear rate and there exists $\hat{k}$ such that $\tau_k \equiv 1$ for all $k \geq \hat{k}$. Therefore, hypotheses (H1)–(H4) are all satisfied and $\tau_k = 1$ for all $k \geq \hat{k}$.

We now show (a). With no loss of generality, we may assume that the algorithm was initiated at $z^{\hat{k}}$. If necessary, one can re-label the sequences so that

$$z^k = z^{\hat{k}+k}, \quad w^k = w^{\hat{k}+k}, \quad s^k = s^{\hat{k}+k}, \quad y^k = y^{\hat{k}+k},$$
$$H^k = H^{\hat{k}+k}, \quad \text{and} \quad \hat{H}^k = \hat{H}^{\hat{k}+k}$$

for $k = 0, 1, 2, \ldots$. With this re-labeling, we have $\tau_k \equiv 1$ for all k. Since no line search is being performed, we have that for this particular sequence the Chen-Fukushima algorithm is an instance of the VMPPA of Sect. 2. In addition, since the hypotheses (H1)–(H4) hold, we have that the technical results Lemma 1, Lemma 2, and Lemma 3 also apply as these lemmas only require the symmetry and positive definiteness of the matrices H_k. In particular, we have from Lemma 2 that there is a $\tilde{k}$ such that $y^{k^T} s^k > 0$ for all $k \geq \tilde{k}$ which verifies (a), and, from Lemma 3, (12) holds.

We now show (b) and (c). Again, with no loss of generality, we may assume that the algorithm was initiated at $z^{\tilde{k}}$. As above, this only involves a re-labeling of the sequences so that

$$z^k = z^{\tilde{k}+k}, \quad w^k = w^{\tilde{k}+k}, \quad s^k = s^{\tilde{k}+k}, \quad y^k = y^{\tilde{k}+k},$$
$$H^k = H^{\tilde{k}+k}, \quad \text{and} \quad \hat{H}^k = \hat{H}^{\tilde{k}+k}$$

for $k = 0, 1, 2, \ldots$. Since (a) and (12) hold, we can apply Theorem 4 with $G = (I + \frac{1}{\lambda}J)^{-1}$ and $B_k = \hat{H}_k^{-1}$ to say that the sequences $\{\|\hat{H}_k\|\}$ and $\{\|\hat{H}_k^{-1}\|\}$ are bounded and

$$\frac{\left\|\left(\hat{H}_k^{-1} - \left(I + \frac{1}{\lambda}J\right)^{-1}\right)s^k\right\|}{\|s^k\|} \to 0,$$

or equivalently,

$$\frac{\left\|\left(I - \left(I + \frac{1}{\lambda}J\right)\hat{H}_k^{-1}\right)s^k\right\|}{\|s^k\|} \to 0.$$

Therefore,

$$\frac{\left\|\hat{H}_k s^k - s^k + \frac{1}{\lambda}\hat{H}_k J \hat{H}_k^{-1} s^k\right\|}{\|s^k\|} \leq \frac{\|\hat{H}_k\|\left\|\left(I - \left(I + \frac{1}{\lambda}J\right)\hat{H}_k^{-1}\right)s^k\right\|}{\|s^k\|} \to 0.$$

But then there is a sequence $\zeta_k \to 0$ such that

$$\left\|\hat{H}_k s^k - s^k + \frac{1}{\lambda}\hat{H}_k J \hat{H}_k^{-1} s^k\right\| \leq \zeta_k \|s^k\|,$$

which in turn implies that

$$\left\|(I - \hat{H}_k)s^k\right\| \leq \left(\frac{1}{\lambda}\|\hat{H}_k J \hat{H}_k^{-1}\| + \zeta_k\right)\|s^k\|$$
$$= \left(\frac{1}{\lambda}\|J\| + \zeta_k\right)\|s^k\| \leq \frac{\rho - \hat{\epsilon}}{2}\|s^k\| \tag{38}$$

for all k sufficiently large since $\lambda > \frac{2}{\rho - \hat{\epsilon}}\|J\|$ by (ii). If $H_k \neq \hat{H}_k$, then it must be the case that

$$\left\|(I - \hat{H}_k)s^k\right\| > \frac{\rho - \hat{\epsilon}}{2}\|s^k\|,$$

since the algorithm sets $s^k = w^k$ (recall $\tau_k \equiv 1$). By (38) this cannot occur for k sufficiently large, therefore eventually $H_k = \hat{H}_k$ establishing (b). Note that this also verifies (c) since it has already been shown that the sequences $\{\|\hat{H}_k\|\}$ and $\{\|\hat{H}_k^{-1}\|\}$ are bounded.

So far we have shown the existence of an iteration $k_0 \geq \hat{k} + \tilde{k}$ such that for all $k \geq k_0$ we have that $y^{k^T} s^k > 0$, the line search parameter τ_k always takes the value 1, and the inverse Hessian approximations H_k are always updated by the formula given in (3). Thus, if the variable metric proximal point algorithm given in Sect. 2 using the symmetric updating scheme of Sect. 3.1 were initiated at z^{k_0} with initial Hessian

estimate H_{k_0}, then this algorithm would produce exactly the same iterates as given by the VMPPA sequence $\{z^k : k \geq k_0\}$ using formula (3). Consequently, since (H1)–(H4) are satisfied, (d) follows from Theorem 1.

$\square$

References

1. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C. (1995): A family of variable metric proximal point methods. Math. Program. **68**, 15–47
2. Burke, J.V., Qian, M. (1998): A variable metric proximal point algorithm for monotone operators. SIAM J. Control Optim. **37**, 353–375
3. Burke, J.V., Qian, M. (1998): On the local super-linear convergence of a matrix secant implementation of the variable metric proximal point algorithm. In: Qi, L., Fukushima, M., eds., Reformulation - Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, pp. 317–334. Kluwer Academic Publishers
4. Byrd, R.H., Nocedal, J. (1989): A tool for the analysis of quasi–Newton methods with application to unconstrained minimization. SIAM J. Numer. Anal. **26**, 727–739
5. Chen, X., Fukushima, M. (1999): Proximal quasi–Newton methods for nondifferentiable convex optimization. Math. Program. **85**, 313–334
6. Fukushima, M. (1984): A descent algorithm for nonsmooth convex minimization. Math. Program. **30**, 163–175
7. Fukushima, M., Qi, L. (1996): A globally and superlinearly convergent algorithm for nonsmooth convex minimization. SIAM J. Optim. **30**, 1106–1120
8. Dennis Jr, J.E., Moré, J.J. (1974): A characterization of superlinear convergence and its application to quasi–Newton methods. Math. Comp. **28**, 549–560
9. Dennis Jr, J.E., Schnabel, R.B. (1983): Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice Hall, New Jersey
10. Lemaréchal, C. (1978): Bundle methods in nonsmooth optimization. In: Lemaréchal, C., Mifflin, R., eds., Nonsmooth Optimization. Pergamon Press, Oxford
11. Lemaréchal, C., Sagastizábal, C. (1994): An approach to variable metric bundle methods. In: Henry, J., Yvon, J.P., eds., Systems Modeling and Optimization, Vol. 197 of LNCIS, pp. 144–162. Springer, Berlin
12. Lemaréchal, C., Sagastizábal, C. (1997): Variable metric bundle methods: from conceptual to implementable forms. Math. Program. **76**, 393–410
13. Mifflin, R. (1996): A quasi–second–order proximal bundle algorithm. Math. Program. **73**, 51–72
14. Mifflin, R., Sun, D., Qi, L. (1998): Quasi–Newton bundle–type methods for nondifferentiable convex optimization. SIAM J. Optim. **8**, 583–603
15. Moreau, J.J. (1965): Proximité et dualité dans un espace Hilbertien. Bull. Soc. Math. France **93**, 273–299
16. Pang, J.-S., Qi, L. (1993): Nonsmooth equations: Motivation and algorithms. SIAM J. Optim. **3**, 443–465
17. Qi, L., Chen, X. (1995): A preconditioning proximal Newton method for nondifferentiable convex optimization. Math. Program. **76**, 411–430
18. Qian, M. (1992): The Variable Metric Proximal Point Algorithm: Theory and Application. Ph.D. Thesis, University of Washington, Seattle, WA
19. Rockafellar, R.T. (1970): Convex Analysis. Princeton University Press, Princeton
20. Rockafellar, R.T. (1976): Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898