SIAM J. CONTROL OPTIM.
Vol. 37, No. 2, pp. 353–375
© 1998 Society for Industrial and Applied Mathematics

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

A VARIABLE METRIC PROXIMAL POINT ALGORITHM FOR MONOTONE OPERATORS*

J. V. BURKE† AND MAIJIAN QIAN‡

Abstract. The proximal point algorithm (PPA) is a method for solving inclusions of the form $0 \in T(z)$, where T is a monotone operator on a Hilbert space. The algorithm is one of the most powerful and versatile solution techniques for solving variational inequalities, convex programs, and convex-concave mini-max problems. It possesses a robust convergence theory for very general problem classes and is the basis for a wide variety of decomposition methods called *splitting methods*. Yet the classical PPA typically exhibits slow convergence in many applications. For this reason, acceleration methods for the PPA algorithm are of great practical importance. In this paper we propose a variable metric implementation of the proximal point algorithm. In essence, the method is a Newton-like scheme applied to the Moreau–Yosida resolvent of the operator T. In this article, we establish the global and linear convergence of the proposed method. In addition, we characterize the superlinear convergence of the method. In a companion work, we establish the superlinear convergence of the method when implemented with Broyden updating (the nonsymmetric case) and BFGS updating (the symmetric case).

Key words. maximal monotone operator, proximal point methods, variable metric, global convergence, convergence rates

AMS subject classifications. Primary, 90C25; Secondary, 49J45, 47H05, 49M45

PII. S0363012992235547

1. Introduction.

The proximal point algorithm (PPA) is one of the most powerful and versatile solution techniques for problems of convex programming and mini-max convex-concave programming. It possesses a robust convergence theory for very general problem classes in finite- and infinite-dimensions (e.g., see [11, 16, 21, 22, 23, 28, 32, 41, 40]) and is the basis for a wide variety of decomposition methods called *splitting methods* (e.g., see [4, 9, 12, 43, 44]). Yet, the classical PPA typically exhibits slow convergence in many applications. For this reason, acceleration methods for the PPA are of great practical importance. In this paper we propose a variable metric implementation of the proximal point algorithm. Our approach extends and refines results that originally appeared in [38] and is in the spirit of several recent articles [3, 7, 10, 18, 20, 24, 25, 36]. However, there is a fundamental difference between the method presented here and those studied in [3, 7, 10, 18, 20, 24, 25, 36]. This difference has a profound impact on the methodology applied in this article. All previous work on this topic (except [38]) applies exclusively to monotone operators that arise as the subdifferential of a finite-valued, finite dimensional convex function. The results of this article apply to general monotone operators on a Hilbert space. The resulting difference in methodology roughly corresponds to the difference between methods for function minimization and methods for solving systems of equations.

There are both advantages and disadvantages to the more general approach. The advantages are that the method applies to a much broader class of problems. This is

*Received by the editors August 12, 1992; accepted for publication (in revised form) September 30, 1997; published electronically November 23, 1998.
http://www.siam.org/journals/sicon/37-2/23554.html

†Department of Mathematics, Box 354350, University of Washington, Seattle, WA 98195–4350 (burke@math.washington.edu). The research of this author was supported by National Science Foundation grant DMS-9303772.

‡Department of Mathematics, California State University, Fullerton, CA 92834 (mqian@fullerton.edu).

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

so not only because the theory is developed in the Hilbert space setting, but, more important, because many monotone operators cannot be represented as the subdifferential of a finite-valued, finite dimensional convex function. General monotone operators do not possess many of the rich structural properties associated with the subdifferential of a convex function (e.g., subdifferentials of convex functions are the only maximal cyclically monotone operators [33]). In addition, in the case where the operator is the subdifferential of a convex function, we do not require the usual assumption that the underlying function be finite-valued.

The disadvantages of our general approach arise from the fact that the method cannot make use of the additional structure present when the operator is the subdifferential of a convex function. This complicates both the structure of the method and its analysis. Of particular note in this regard is the complexity of our global convergence result. If the operator is the subdifferential of a convex function, then solving the inclusion $0 \in T(x)$ is equivalent to minimizing the underlying convex function. The global convergence of a method is then typically driven by a line-search routine (e.g., see [3, 7, 10, 18, 20, 24, 25, 36]). In the general setting we do not have direct recourse to this strategy. This complicates both the structure of the algorithm and its convergence theory. Nonetheless, the proof technique developed in this paper can be refined in the convex programming setting, thereby significantly simplifying both the global and the local convergence results [5, 6].

Notwithstanding these differences in methodology, our approach is still nicely motivated by recalling the behavior of the PPA in the context of convex programming:

$$\min_{z \in \mathcal{H}} f(z) \ , \tag{1}$$

where $\mathcal{H}$ is a Hilbert space and $f: \mathcal{H} \mapsto \mathbb{R} \cup \{+\infty\}$ is a lower semicontinuous convex function that is not identically $+\infty$. Define the Moreau–Yosida regularization of f to be the function $f_\lambda: \mathcal{H} \mapsto \mathbb{R}$ given by

$$f_\lambda(\bar{z}) := \min_{z \in \mathcal{H}} \left\{ \lambda f(z) + \frac{1}{2}\|z - \bar{z}\|^2 \right\}.$$

The set of solutions to (1) corresponds precisely to the set of points at which f_λ attains its minimum value. The function f_λ is continuously Fréchet differentiable [28, Proposition 7.d]. The PPA applied to (1) is *approximately* the steepest descent algorithm applied to f_λ [11]. This analogy immediately suggests that a variable metric approach could be applied to the function f_λ to accelerate the method. This idea was first studied in [38] and is the basis of the acceleration techniques described in [3, 7, 10, 18, 20, 24, 25, 36].

In [3], Bonnans et al. develop methods along an algorithmic pattern originally suggested by Qian in [38]. This pattern circumvents many of the difficulties associated with a variable metric approach applied directly to the function f_λ. The key is to employ a matrix secant update based on the function f instead of f_λ. The local convergence results in [3, Section 3] require some smoothness assumptions. In particular, linear convergence is established when the function f is differentiable with Lipschitz continuous derivative, and superlinear convergence is established when f is twice strictly Fréchet differentiable at a unique solution $\bar{z}$, where the second derivative is positive definite (we speak only of quotient or q-rate of convergence).

In [18, 20, 24, 25], the authors apply the bundle concept for nonsmooth convex minimization [17] to approximate the Moreau–Yosida regularization f_λ and its derivative. Variable metric updates, in particular, quasi-Newton updates, are then applied

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

using these approximate values. The superlinear convergence results in the papers [18, 20, 24] require either strong smoothness assumptions on the function f (such as the Lipschitz continuity of ∇f) or that the regularization parameter λ diverges to $+\infty$. In [20], Lemaréchal and Sagastizábal propose a clever *reversal* quasi-Newton formula which uses the value of the gradient of f_λ at a variety of points other than those strictly obtained by the iterates. This promising idea deserves further theoretical and numerical study.

In [10] and [36], the authors develop an approach based on Newton's method for semismooth functions as developed in [30, 31, 37, 34]. Properly speaking, these methods are neither an adaptation of the PPA algorithm nor a variable metric method. Nonetheless, the flavor of both of these methodologies is present. In order to obtain superlinear convergence, smoothness hypotheses are again required; however, these hypotheses are of a somewhat more technical nature. Specifically, it is required that

(a) the function f be *semismooth* at a unique solution to (1) [37],

(b) every element of the set-valued mapping

$$\partial_B^2 f(z) := \{ \lim_{y_k \to z} \nabla^2 f_\lambda(y_k) \,:\, y_k \to x, \nabla f(y_k) \text{ exists for all } k = 1, 2, \ldots \}$$

be nonsingular at the unique solution $\bar{z}$, and

(c) the sequence of Hessian approximates $\{V_k\}$ used to generate the iterates $\{z_k\}$ satisfy

$$\lim_{k \to \infty} \operatorname{dist}(V_k, \partial_B^2 f(z_k)) = 0 \,. \tag{2}$$

One can show that the semismoothness hypotheses are satisfied in many cases of interest when f is finite-valued. Moreover, by Rademacher's theorem on the differentiability of Lipschitz continuous functions, it follows that the set-valued mapping $\partial_B^2 f(z)$ is always well-defined and compact-valued in the finite dimensional, finite-valued case, with the nonsingularity property being closely tied to the usual hypothesis of strong convexity. Although the limiting hypotheses on the V_k's is a bit strong, it is not entirely unreasonable in the absence of differentiability. In [36], Qi and Chen propose a very nice preconditioning technique wherein an exact value for the gradient of a *shifted* Moreau–Yosida regularization can be computed from inexact values for the gradient of f_λ. This technique is similar in spirit to the *reversal* quasi-Newton formula found in [20]. Both of these techniques should prove useful in numerical implementations.

The algorithm presented in this paper is most closely related to the methods proposed by Chen and Fukushima [7] and Mifflin, Sun, and Qi [25]. However, there are several fundamental distinctions, the foremost of which is that the methods in [7, 25] are restricted to finite dimensional finite-valued convex programming problems. Within this framework, these authors use bundle strategies to approximate f_λ and its gradient and establish the global convergence of their methods with the aid of a line search routine. Chen and Fukushima establish global and linear convergence results along with a generalization of the Dennis–Moré characterization theorem for superlinear convergence [14]. One of the most important features of the Chen–Fukushima algorithm is that the line search is based on the function f rather than approximations to the function f_λ. This is very important in practice since obtaining sufficiently accurate approximations to the function f_λ is usually quite time consuming. Their linear and superlinear convergence results blend bundle techniques with the theory of nonsmooth equations. Consequently, the convergence hypotheses are reminiscent of those employed in [10] and [36]; in particular, they require semismoothness, CD-regularity,

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

and the strong approximation property (2). In [6], the methods of this paper are applied to the Chen–Fukushima algorithm to obtain the superlinear convergence of the method when BFGS matrix secant updating is employed.

In [25], Mifflin, Sun, and Qi obtain the first superlinear 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 proposed algorithm uses a line search based on approximations to the function f_λ and requires that the function f_λ is strongly convex with ∇f_λ Fréchet differentiable at the unique global solution to the convex program. In addition it is assumed that the iterates satisfy a certain approximation property involving the gradient ∇f_λ. In section 4 of this paper, we discuss how these hypotheses are related to those that are also required in our convergence analysis.

In this paper, we provide a general theory for a variable metric proximal point algorithm (VMPPA) applied to maximal monotone operators from a Hilbert space to itself. In the important special case of convex programming, where T is taken to be the subdifferential of the function f, we do not assume that f is finite-valued or differentiable on the whole space. However, to obtain superlinear convergence, we do require certain smoothness hypotheses at a unique global solution $\bar{z}$. These smoothness hypotheses differ from those assumed in [3, 18, 20, 24] since they are imposed on the operator T^{-1} rather than T. In this regard, they are reminiscent of the hypotheses employed in [25]. The choice of smoothness hypotheses has deep significance in the context of convex programming. Differentiability hypotheses on $T = \partial f$ imply the second-order differentiability of f, whereas differentiability hypotheses on $T^{-1} = (\partial f)^{-1}$ are related to the standard strong second-order sufficiency conditions of convex programming [40, Proposition 2] and thus reduce to the standard hypotheses used in local analysis of convergence. In particular, the differentiability of $(\partial f)^{-1}$ does not imply that ∂f is single-valued or differentiable, nor does it imply that f is finite-valued.

Our smoothness hypotheses also differ from those that appear in [7, 10, 36]. These methods rely on the theory of nonsmooth equations and require hypotheses such as semismoothness and nonsingularity of the elements of $\partial_B^2 f$. In addition, the proof theory for these methods specifically requires that the underlying convex function be finite-valued in a neighborhood of the unique solution to (1) (again, these methods assume that the function is finite-valued on all of $\mathbb{R}^n$). This limits direct application to constrained problems since in the constrained case solutions typically lie on the boundary of the constraint region (i.e., on the boundary of the domain of the essential objective function).

Throughout the paper we illustrate many of the ideas and results by applying them to the case of convex programming. Our purpose here is not only to show how the results can be applied, but also to ground them in the familiar surroundings of this concrete application. Further details on the application of these results to the case of convex programming can be found in [5].

The paper is structured as follows. We begin with a review of the classic proximal point algorithm in section 2. The VMPPA is introduced in section 3. This section contains the approximation criteria that must be satisfied at each iteration. Two criteria are presented. The first is required to obtain global convergence and the second is required to accelerate the local convergence of the method. This division into global and local criteria is one of the recurring themes of the paper. On the global level the method behaves like a steepest descent method, while at the local

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

level it becomes more Newton-like. This feature is common to most general purpose methods in nonlinear programming, such as the nonmonotone descent methods, the dogleg method, and trust-region methods. In section 4 we discuss the smoothness hypotheses required for the local analysis. We also extend some of the differentiability results appearing in [19, 35] to maximal monotone operators. In section 5, we study the operators $\mathcal{N}_k$ associated with the Newton-like iteration proposed in section 3. The focus of this section is to provide conditions under which the operators $\mathcal{N}_k$ are nonexpansive at a solution to the inclusion $0 \in T(z)$. A global convergence result paralleling Rockafellar's 1976 result [41] is given in section 6. In section 7 we study local convergence rates. Linear convergence is established under a Lipschitz continuity assumption on T^{-1}, and a characterization of superlinear convergence for the VMPPA is also given. This characterization is modeled on the landmark characterization of superlinear convergence of variable metric methods in nonlinear programming due to Dennis and Moré [14]. In [6], we use this characterization result to establish the superlinear convergence of the method when the derivatives are approximated using the BFGS and Broyden updating strategies.

A word about our notation is in order. We denote the closed unit ball in the Hilbert space $\mathcal{H}$ by $\mathbb{B}$. Then the ball with center a and radius r is denoted by $a + r\mathbb{B}$. Given a set $Z \subset \mathcal{H}$ and an element $z \in \mathcal{H}$, the distance of z to Z is $\operatorname{dist}(z, Z) = \inf\{\|z - z'\| : z' \in Z\}$.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces. Given a *multifunction* (also referred to as a *mapping* or an *operator* depending on the context) $T : \mathcal{H}_1 \rightrightarrows \mathcal{H}_2$, the *graph* of T, $\operatorname{gph} T$, is the subset of the product space $\mathcal{H}_1 \times \mathcal{H}_2$ defined by $\operatorname{gph} T = \{(z, w) \in \mathcal{H}_1 \times \mathcal{H}_2 | w \in T(z)\}$. The *domain* of T is the set $\operatorname{dom} T := \{z \in \mathcal{H}_1 | T(z) \neq \emptyset\}$. The identity mapping will be denoted by I. The *inverse* of an operator T is defined by $T^{-1}(w) := \{z \in \mathcal{H}_1 | (z, w) \in \operatorname{gph} T\}$.

Given a lower semicontinuous convex function $f : \mathcal{H} \to \mathbb{R} \bigcup \{+\infty\}$, the *conjugate* of f is defined by $f^*(z^*) = \sup_{z \in \mathcal{H}} \{\langle z^*, z\rangle - f(z)\}$.

2. Monotone operators and the classic algorithm. Given a real Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$, we say that the multifunction $T : \mathcal{H} \rightrightarrows \mathcal{H}$ is *monotone* if for every z and z' in $\operatorname{dom} T$, and $w \in T(z)$ and $w' \in T(z')$, we have $\langle z - z', w - w' \rangle \geq \kappa \|z - z'\|^2$ for some $\kappa \geq 0$. If $\kappa > 0$, then T is said to be *strongly monotone with modulus* κ. The monotone operator T is said to be *maximal* if its graph is not properly contained in the graph of any other monotone operator. An important example of a monotone operator is the *subgradient* of a convex function (see Minty [27] and Moreau [28]).

We are concerned with solving inclusions of the form

$$0 \in T(z), \tag{3}$$

where T is a maximal monotone operator. In the case of the convex programming problem (1), the operator T is the subdifferential of the convex function f, and the inclusion (3) characterizes the points z at which f attains its minimum value. A wide variety of other problems can be cast in this framework, e.g., variational inequalities, complementary problems, and mini-max problems. Existence results for inclusion (3) can be found in [41].

In 1962, Minty [27] showed that, when the operator T is maximal monotone, the *Moreau–Yosida resolvent* of T,

$$P_\lambda = (I + \lambda T)^{-1} \text{ with } \lambda > 0,$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

is single-valued and nonexpansive on $\mathcal{H}$. This result suggests that a solution to the inclusion $0 \in T(z)$ can be iteratively approximated by the recursion $z^{k+1} = P_\lambda(z^k)$. One can modify this scheme by varying the scalar λ and by choosing the iterates z^{k+1} to be an approximate solution to the equation $(I + \lambda_k T)(z) = z^k$. The PPA applies precisely these ideas. The algorithm, starting from any point z^0, generates a sequence $\{z^k\}$ in $\mathcal{H}$ by the approximation rule

$$z^{k+1} \approx (I + c_k T)^{-1}(z^k) \; . \tag{4}$$

The principal difficulty in applying the PPA lies in executing the operators $P_k = (I + c_k T)^{-1}$. In the case of convex programming, the iteration (4) reduces to the iteration

$$z^{k+1} \approx \arg\min_{z \in \mathcal{H}} \left\{ c_k f(z) + \frac{1}{2}\|z - z^k\|^2 \right\} \; .$$

Notice that executing the algorithm exactly (i.e., with "=" instead of "≈" in the above algorithm) can be as difficult as solving the original problem directly. Hence it is critical that the convergence results are obtained under the assumption of approximation.

In [22] and [23], Martinet proved the convergence of the *exact* PPA for certain cases of the operator T with fixed $c_k \equiv c$. The first theorem on the convergence of the general PPA was proved by Rockafellar [41] in 1976. His theorem not only insures the global convergence under an approximating rule, but also describes the global behavior when the inclusion $0 \in T(z)$ has no solution.

The convergence rate of the PPA depends on properties of the operator T, the choice of the sequence $\{c_k\}$, and the accuracy of the approximation in (4). The first rate of convergence results were also obtained by Rockafellar [41] in 1976, under the assumption that the solution set is a singleton $\{\bar{z}\}$. He proved that if the sequence $\{c_k\}$ is bounded away from 0, and $T^{-1}(w)$ is bounded by a linear function of $\|w\|$ when w is near 0, then the rate of convergence is at least linear. Luque [21] extended Rockafellar's theorem to the case where $T^{-1}(0)$ is not required to be a singleton, and showed that such an estimate of the convergence rate is tight.

3. The algorithm and approximation criteria. The algorithm proposed in this section is a Newton-like iteration for solving the resolvent equation $z = P_\lambda(z)$. In the context of the convex programming problem, the iteration takes the form

$$z^{k+1} = z^k - H_k \nabla f_\lambda(z^k) \; ,$$

where the operator H_k is used to approximate second-order properties of the function f_λ. If f_λ is twice differentiable with $[\nabla^2 f_\lambda(z^k)]^{-1}$ bounded, then for Newton's method one sets $H_k = [\nabla^2 f_\lambda(z^k)]^{-1}$. However, in general, f_λ is only known to be differentiable with Lipschitz continuous gradient [28]. Thus, in the finite dimensional case, the Hessian $\nabla^2 f_\lambda(x)$ is guaranteed to exist only on a dense subset by Rademacher's theorem. Further results on the second-order properties of f_λ can be found in [19, 35, 42].

It is well known that the negative gradient $-\nabla f_\lambda(z^k)$ is the unique element w^k solving the problem

$$\min_{w \in \mathcal{H}} \left\{ \lambda f(z^k + w) + \frac{1}{2}\|w\|^2 \right\}$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

or, equivalently, satisfying the inclusion

$$0 \in \lambda \partial f(z^k + w^k) + w^k \ . \tag{5}$$

The PPA for a general maximal monotone operator T can be formally derived from (5) by replacing λ, z^k, and ∂f by c_k, z^k, and T, respectively, to obtain

$$0 \in c_k T(z^k + w^k) + w^k \ ,$$

or equivalently,

$$w^k = [(I + c_k T)^{-1} - I](z^k) \ ,$$

where equality follows from the fact that w^k is unique. This motivates us to define the operator

$$D_k := (I + c_k T)^{-1} - I \ . \tag{6}$$

This operator provides the analogue of the direction of steepest descent in the operator setting.

The algorithm we propose for solving the inclusion $0 \in T(z)$ can be succinctly stated as follows.

The Variable Metric Proximal Point Algorithm.

Let $z^0 \in \mathcal{H}$ and $c_0 \geq 1$ be given. Having z^k, set

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

and choose $c_{k+1} \geq 1$.

As mentioned in the previous section, it is critical that the convergence results are obtained under the assumption that $D_k(z^k)$ can only be approximated. We use the following approximation criteria:

$$(\mathcal{G}) \qquad \|w^k - D_k(z^k)\| \leq \min\left\{1, \frac{1}{\|H_k\|}\right\} \epsilon_k \ \ \text{with} \ \ \sum_{k=0}^{\infty} \epsilon_k < \infty$$

and

$$(\mathcal{L}) \qquad \|w^k - D_k(z^k)\| \leq \delta_k \|w^k\| \ \ \text{with} \ \ \lim_{k \to \infty} \delta_k = 0 \ .$$

The approximation criterion $(\mathcal{G})$ is used to establish global convergence properties, while criterion $(\mathcal{L})$ is used to obtain local rates of convergence.

Although these criteria are used in the proof of convergence, they are impractical from the perspective of implementation. In their stead, we provide criteria that are implementable. To obtain these criteria we recall the following result from Rockafellar [41].

Proposition 1 (see [41, Proposition 3]). *Let $S_k(w) := T(z^k + w) + \frac{1}{c_k} w$. Then $0 \in S_k(w^k) \Leftrightarrow w^k = D_k(z^k)$. Moreover, for all $w \in \mathcal{H}$ we have the bound*

$$\|w - D_k(z^k)\| \leq c_k \text{dist}\,(0, S_k(w)) \ . \tag{7}$$

Proposition 1 yields the following alternative approximation criteria for the w^k's. Since this result is an immediate consequence of Proposition 1, its proof is omitted.

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

PROPOSITION 2. *Consider the following acceptance criteria for the w^k's:*

$$(\mathcal{G}') \qquad \operatorname{dist}(0, S_k(w^k)) \le \min\left\{1, \frac{1}{\|H_k\|}\right\} \frac{\epsilon_k}{c_k} \quad \textit{with} \quad \sum_{k=0}^{\infty} \epsilon_k < \infty$$

and

$$(\mathcal{L}') \qquad \operatorname{dist}(0, S_k(w^k)) \le \frac{\delta_k}{c_k}\|w^k\| \quad \textit{with} \quad \lim_{k\to\infty} \delta_k = 0 \ .$$

We have $(\mathcal{G}')$ implies $(\mathcal{G})$ and $(\mathcal{L}')$ implies $(\mathcal{L})$.

Remark. Note that to satisfy either $(\mathcal{G}')$ or $(\mathcal{L}')$ it is not necessary to find an element of $S_k(w^k)$ of least norm.

Before leaving this section we recall from [41] a few properties of the operators D_k and $P_k := D_k + I$ that are essential in the analysis to follow.

PROPOSITION 3 (see [41, Proposition 1]).

a) *The operator D_k can be expressed as*

$$D_k = -\left(I + T^{-1}\frac{1}{c_k}\right)^{-1}, \tag{8}$$

and for any $z \in \mathcal{H}$, $-\frac{1}{c_k} D_k(z) \in T(P_k(z))$.

b) *For any $z, z' \in \mathcal{H}$, $\langle P_k(z) - P_k(z'), D_k(z) - D_k(z')\rangle \le 0$.*

c) *For any $z, z' \in \mathcal{H}$, $\|P_k(z) - P_k(z')\|^2 + \|D_k(z) - D_k(z')\|^2 \le \|z - z'\|^2$.*

Remark. An important consequence of part c) above is that the operators P_k and D_k are Lipschitz continuous with Lipschitz constant 1; that is, they are nonexpansive. Henceforth, we make free use of this fact.

4. On the differentiability of T^{-1} and D_k. Just as Newton's method for minimization locates roots of the gradient, one can view the VMPPA as a Newton-like method for locating roots of the operator D_k. This perspective motivates our approach to the local convergence analysis. For this analysis, we require that the operator T^{-1} possesses certain smoothness properties. These properties in turn imply the smoothness of the operators D_k. Smoothness hypotheses are used in the convergence analysis in much the same way as they are used in the convergence analysis for Newton's method. For example, recall that to ensure the quadratic convergence of Newton's method one requires the derivative at a solution to be both locally Lipschitz and nonsingular. Nonsingularity ensures that the iterates are well-defined and can be bounded, while the Lipschitzian hypothesis guarantees that the error in the linearization is quadratically bounded (see [29, sections 3.2.12 and 10.2.2]). We make use of similar properties in our analysis.

In order to discuss the smoothness of T^{-1} and D_k, we recall various notions of differentiability for multivalued functions from the literature. For a more thorough treatment of these ideas in the context of monotone operators, we refer the reader to [1, 19, 26, 35, 42].

DEFINITION 4. *We say that an operator $\Psi : \mathcal{H} \rightrightarrows \mathcal{H}$ is Lipschitz continuous at a point $\bar{w}$ (with modulus $\alpha \ge 0$) if the set $\Psi(\bar{w})$ is nonempty and there is a $\tau > 0$ such that*

$$\Psi(w) \subset \Psi(\bar{w}) + \alpha\|w - \bar{w}\|\mathbb{B} \qquad \textit{whenever } \|w - \bar{w}\| \le \tau \ .$$

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 $J : \mathcal{H} \to \mathcal{H}$ such that for some $\delta > 0$,

$$\emptyset \ne \Psi(w) - \bar{z} - J(w - \bar{w}) \subset o(\|w - \bar{w}\|)\mathbb{B} \qquad \textit{whenever } \|w - \bar{w}\| \le \delta \ .$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We then write $J = \nabla\Psi(\bar{w})$.

Remarks. 1) These definitions of Lipschitz continuity and differentiability for multifunction are taken from [41, pp. 885 and 887] (also see [2, p. 41]). Note that these notions of Lipschitz continuity and differentiability correspond to the usual notions when Ψ is single-valued.

2) Rockafellar [41, Theorem 2] was the first to use Lipschitz continuity to establish rates of convergence for the PPA.

3) 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 $\alpha(\tau) \to \|J\|$ as $\tau \to 0$. This observation is verified in [41, Proposition 4].

4) It follows from the definition of monotonicity that if T is a maximal monotone operator, then the operator $\nabla T(z)$ is positive semidefinite whenever it exists.

We now give a result that relates the differentiability of a multivalued function to the differentiability of its inverse. The proof is omitted since it parallels the proof of a similar result for single-valued functions.

LEMMA 5. *Assume that* $\Psi : \mathcal{H} \Longrightarrow \mathcal{H}$ *is differentiable at* $\bar{z}$ *with* $\Psi(\bar{z}) = \{\bar{w}\}$ *and* $\nabla\Psi(\bar{z}) = J$ *with* J^{-1} *bounded. Also assume that* Ψ^{-1} *is Lipschitz continuous at* $\bar{w}$ *with* $\Psi^{-1}(\bar{w}) = \{\bar{z}\}$. *Then* Ψ^{-1} *is differentiable at* $\bar{w}$ *with* $\nabla\Psi^{-1}(\bar{w}) = J^{-1}$.

In the two examples that follow, we examine the concepts introduced in Definition 4 when the operator in question is the subdifferential of a convex function. The first example illustrates that ∂f^{-1} can be Lipschitz continuous but not differentiable at the origin, while in the second example ∂f^{-1} is differentiable at the origin, but ∂f is not differentiable on $(\partial f)^{-1}(0)$.

EXAMPLE 6. *Let*

$$f(z) := \begin{cases} 0 & \text{if } z < 0, \\ z & \text{if } z \geq 0, \end{cases} \qquad \text{and} \qquad T(z) := \partial f(z) = \begin{cases} 0 & \text{if } z < 0, \\ [0,1] & \text{if } z = 0, \\ 1 & \text{if } z > 0\ . \end{cases}$$

$$\text{Then} \quad T^{-1}(y) = \begin{cases} \emptyset & \text{if } y < 0 \text{ or } y > 1, \\ (-\infty, 0] & \text{if } y = 0, \\ \{0\} & \text{if } y \in (0,1), \\ [0,\infty) & \text{if } y = 1\ . \end{cases}$$

T^{-1} *is Lipschitz continuous at* 0 *but is not differentiable at* 0.

EXAMPLE 7. *Let*

$$f(z) := \begin{cases} -z & \text{if } z < 0, \\ z^{5/3} & \text{if } z \geq 0, \end{cases} \qquad \text{and} \qquad T(z) := \partial f(z) = \begin{cases} -1 & \text{if } z < 0, \\ [-1,0] & \text{if } z = 0, \\ \frac{5}{3}z^{2/3} & \text{if } z > 0. \end{cases}$$

$$\text{Then} \quad T^{-1}(y) = \begin{cases} \emptyset & \text{if } y < -1, \\ (-\infty, 0] & \text{if } y = -1, \\ \{0\} & \text{if } y \in (-1,0), \\ \frac{3}{5}y^{3/2} & \text{if } y \geq 0\ . \end{cases}$$

T^{-1} *is differentiable at* 0 *with* $J = 0$, *but* T *is not differentiable on* $T^{-1}(0)$.

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The superlinear convergence result of section 7 requires the assumption that the operator T^{-1} be differentiable at the origin. Although this is a severe restriction on the applicability of these results, it turns out that in the case of convex programming it is a consequence of the standard second-order sufficiency conditions for constrained mathematical programs. This and related results were established by Rockafellar in [40, Proposition 2]. In this context, it is important to note that the second-order sufficiency condition is the standard hypothesis used in the mathematical programming literature to ensure the rapid local convergence of numerical methods. So, at least in the context of constrained convex programming, such a differentiability hypothesis is not as severe an assumption as one might at first suspect. To the contrary, it is a bit weaker than the standard hypothesis employed for such results. For the sake of completeness, we recall a portion of Rockafellar's result below.

THEOREM 8. *Consider the convex programming problem* (1)*, where* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *is given by*

$$f(z) = \begin{cases} f_0(z) & \text{if } f_i(z) \le 0 \text{ for } i = 1, 2, \ldots, m, \\ +\infty & \text{otherwise,} \end{cases}$$

with $f_i : \mathbb{R}^n \to \mathbb{R}$ *convex for* $i = 0, 1, \ldots, m$*. Suppose that the following conditions are satisfied:*

(i) *The functions* f_i *for* $i = 0, 1, \ldots, m$ *are* $k \ge 2$ *times continuously differentiable in a neighborhood of a point* $\bar{z} \in \mathbb{R}^n$*.*

(ii) *There is a Kuhn–Tucker vector* $\bar{y} \in \mathbb{R}^m$ *for* $\bar{z}$ *such that* $\bar{y}_i > 0$ *for* $i \in I(\bar{z}) = \{i : f_i(\bar{z}) = 0,\ i = 1, 2, \ldots, m\}$*.*

(iii) *The gradients* $\{\nabla f_i(\bar{z}) : i \in I(\bar{z})\}$ *are linearly independent.*

(iv) *The matrix* $H = \nabla^2 f_0(\bar{z}) + \sum_{i=1}^m \bar{y}_i \nabla^2 f_i(\bar{z})$ *satisfies* $u^T H u > 0$ *for every nonzero* $u \in \mathbb{R}^n$ *such that* $\nabla f_0(\bar{z})^T u = 0$*, and* $\nabla f_i(\bar{z})^T u = 0$ *for* $i \in I(\bar{z})$*.*

Then the operator ∂f^{-1} *is* $(k-1)$ *times continuously differentiable in a neighborhood of the origin.*

Remark. Theorem 8 follows by applying the implicit function theorem to the Kuhn–Tucker conditions for the parameterized problems $\min\{f(z) - \langle w, z\rangle\}$ in a neighborhood of $w = 0$. The relationship to ∂f^{-1} comes from the fact that $\partial f^{-1}(w) = \operatorname{argmin}\{f(z) - \langle w, z\rangle\}$. Rockafellar establishes the result only for $k = 2$. The extension to $k > 2$ follows trivially from the implicit function theorem.

We now examine the differentiability properties of the mapping D_k. Two results in this direction are given. The first uses (8) to relate the differentiability of the operators T^{-1} and D_k, while the second uses the definition of D_k given in (6) to relate the differentiability of the operators T and D_k.

PROPOSITION 9. *Let* $T : \mathcal{H} \rightrightarrows \mathcal{H}$ *be maximal monotone and* $\lambda > 0$*. Define*

$$D(z) = -\left(I + T^{-1}\frac{1}{\lambda}\right)^{-1}(z)\ . \tag{9}$$

Let $\bar{z} \in \mathcal{H}$ *and set* $\bar{w} = D(\bar{z})$ *and* $\bar{y} = -\frac{1}{\lambda}\bar{w}$*. The operator* T^{-1} *is differentiable at* $\bar{y}$ *with* $[I + \frac{1}{\lambda}\nabla(T^{-1})(\bar{y})]^{-1}$ *bounded if and only if the operator* D *is differentiable at* $\bar{z}$ *with* $(\nabla D(\bar{z}))^{-1}$ *bounded. In either case, we have*

$$\nabla D(\bar{z}) = -\left[I + \frac{1}{\lambda}\nabla(T^{-1})(\bar{y})\right]^{-1}\ . \tag{10}$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. First assume that T^{-1} is differentiable at $\bar{y}$ with $\nabla(T^{-1})(\bar{y})$ bounded. The differentiability of T^{-1} at $\bar{y}$ clearly implies that of D^{-1} at $\bar{w}$ with

$$\nabla[D^{-1}](\bar{w}) = -\left(I + \frac{1}{\lambda}\nabla[T^{-1}](\bar{y})\right) .$$

Since D is Lipschitzian with $D(\bar{z}) = \bar{w}$, Lemma 5 implies that D is differentiable at $\bar{z}$ with derivative given by (10). Since $\nabla[D^{-1}](\bar{w}) = (\nabla D(\bar{z}))^{-1}$, we conclude that the latter is bounded.

Conversely, assume that D is differentiable at $\bar{z}$ with $(\nabla D(\bar{z}))^{-1}$ bounded. We show that D^{-1} is single-valued and Lipschitzian at $\bar{w}$. The result will then follow from Lemma 5.

Let $\delta > 0$ be as in Definition 4 for $\nabla D(\bar{z})$. Since D is single-valued and $\nabla D(\bar{z})$ is surjective (it is invertible), we may apply a standard open mapping result from functional analysis (e.g., [8, Theorem 15.5]) to obtain the existence of a $\rho > 0$ and a $0 < \hat{\delta} < \delta$ such that

$$\bar{w} + \rho\mathbb{B} \subset D(\bar{z} + \hat{\delta}\mathbb{B}) . \tag{11}$$

Hence for each $w \in \bar{w} + \rho\mathbb{B}$ and $z \in D^{-1}(w) \cap (\bar{z} + \hat{\delta}\mathbb{B}) \neq \emptyset$ we have

$$w = \bar{w} + \nabla D(\bar{z})(z - \bar{z}) + o(\|z - \bar{z}\|) . \tag{12}$$

Since $(\nabla D(\bar{z}))^{-1}$ is bounded, there is a $\kappa > 0$ such that

$$\|w - \bar{w}\| + o(\|z - \bar{z}\|) = \|\nabla D(\bar{z})(z - \bar{z})\| \geq \kappa\|z - \bar{z}\| .$$

Hence, by reducing ρ and $\hat{\delta}$ if necessary, we may assume that

$$\|w - \bar{w}\| \geq \frac{\kappa}{2}\|z - \bar{z}\| \geq \frac{\kappa}{2}\|w - \bar{w}\|$$

for $w \in \bar{w} + \rho\mathbb{B}$, where the second inequality follows since D is nonexpansive. Therefore, we can assume that $o(\|z - \bar{z}\|) = o(\|w - \bar{w}\|)$ for all $w \in \bar{w} + \rho\mathbb{B}$ and $z \in D^{-1}(w) \cap (\bar{z} + \hat{\delta}\mathbb{B})$. By substituting this into (12) and rearranging, we obtain

$$\begin{aligned} &z = \bar{z} + (\nabla D(\bar{z}))^{-1}(w - \bar{w}) + o(\|w - \bar{w}\|) \\ &\text{for all } w \in \bar{w} + \rho\mathbb{B} \text{ and } z \in D^{-1}(w) \cap (\bar{z} + \hat{\delta}\mathbb{B}). \end{aligned} \tag{13}$$

We now show that (13) implies the existence of an $\epsilon > 0$ such that $D^{-1}(\bar{w} + \epsilon\mathbb{B}) \subset \bar{z} + \hat{\delta}\mathbb{B}$. Indeed, if this were not the case, then there would exist sequences $\{w_i\}$ and $\{z_i\}$ such that $z_i \in D^{-1}(w_i)$, $\|z_i - \bar{z}\| > \hat{\delta}$, and $w_i \to \bar{w}$. Since D^{-1} is itself maximal monotone, its images are convex; hence, by (11), there exists a sequence $\{\hat{z}_i\}$ with $\hat{z}_i \in D^{-1}(w_i)$ and $\|\hat{z}_i - \bar{z}\| = \hat{\delta}$ for all $i = 1, 2, \ldots$. But then (13) implies that

$$\hat{z}_i = \bar{z} + (\nabla D(\bar{z}))^{-1}(w_i - \bar{w}) + o(\|w_i - \bar{w}\|)$$

for all $i = 1, 2, \ldots$. This contradicts the fact that $w_i \to \bar{w}$ and $\|\hat{z}_i - \bar{z}\| = \hat{\delta}$ for all $i = 1, 2, \ldots$, and so such an $\epsilon > 0$ must exist. This fact combined with (13) implies that D^{-1} is Lipschitzian at $\bar{w}$ with $D^{-1}(\bar{w}) = \{\bar{z}\}$. Lemma 5 now applies to yield the result. ∎

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

PROPOSITION 10. *Let D be defined as in* (9). *Let $\bar{z} \in \mathcal{H}$ and set $\bar{y} = (I + D)(\bar{z})$. The operator T is differentiable at $\bar{y}$ with $[I + \lambda \nabla T(\bar{y})]^{-1}$ bounded if and only if the operator D is differentiable at $\bar{z}$ with $[I + \nabla D(\bar{z})]^{-1}$ bounded. In either case we have the formula*

$$\nabla D(\bar{z}) = [I + \lambda \nabla T(\bar{y})]^{-1} - I \ .$$

Proof. Replace D by $P := I + D = (I + \lambda T)^{-1}$ and observe that D is differentiable at $\bar{z}$ with $[I + \nabla D(\bar{z})]^{-1}$ bounded if and only if P is differentiable at $\bar{z}$ with $[\nabla P(\bar{z})]^{-1}$ bounded. The proof now follows the same argument as in the proof of Proposition 9 with D replaced by P, T^{-1} replaced by T, and $\bar{w}$ replaced by $\bar{y}$. ∎

Propositions 9 and 10 say quite different things about the differentiability of D_k. To illustrate this difference, observe that in Example 7 the operator T is not differentiable at 0, while T^{-1} and D are differentiable at 0. On the other hand, if we take $T = \partial f$ with $f(x) = |x|^3$, then T^{-1} is not differentiable at 0, while T and D are differentiable at 0. It is also important to note that even if neither T nor T^{-1} is differentiable, D may be differentiable. But, in this case, we know from Propositions 9 and 10 that if D is differentiable and neither T nor T^{-1} is differentiable, then both $\nabla D(\bar{z})$ and $\nabla P(\bar{z})$ have to be singular or have unbounded inverses. For a further discussion of these issues in the context of finite dimensional convex programming, see [35].

When T is assumed to be the subdifferential of a convex function f, Propositions 9 and 10 can be refined by making use of the relation $\partial f^{-1} = \partial f^*$, where f^* is the convex conjugate of f [39, Corollary 12A]. This allows us to extend [35, Theorem 1] and [35, Theorem 2] to the Hilbert space setting (also see [19, Theorem 3.1]). However, some caution in terminology is required since f^* is not necessarily twice differentiable in the classical sense at points where ∂f^* is differentiable in the sense of Definition 4. Indeed, ∂f^* may be multivalued arbitrarily close to a point of differentiability. The best way to interpret this result is through Alexandrov's theorem [1], which states that at almost every point $\bar{z}$ in the interior of the domain of a convex function $f : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ there is a quadratic function $q_{\bar{z}}$ such that $f(x) = q_{\bar{z}}(x) + o(\|x - \bar{z}\|^2)$. In [19] and [35], the matrix $\nabla^2 q_{\bar{z}}$ is called a *generalized Hessian* and is denoted $Hf(x)$. Note that the existence of a generalized Hessian at the point $\bar{z}$ guarantees that f is strictly differentiable at $\bar{z}$. Moreover, if $\partial f(x)$ is single-valued in a neighborhood of a point $\bar{z}$ at which $Hf(\bar{z})$ exists, then $\nabla^2 f(\bar{z})$ exists and equals $Hf(\bar{z})$. We extend this terminology to the Hilbert space setting with the following definition.

DEFINITION 11. *Let $\phi : \mathcal{H} \mapsto \mathbb{R} \cup \{\infty\}$ be a function on the Hilbert space $\mathcal{H}$. We say that ϕ is twice differentiable in the generalized sense at a point $\bar{z} \in \mathcal{H}$ if there is a continuous quadratic functional $q_{\bar{z}}$ such that $\phi(x) = q_{\bar{z}}(x) + o(\|x - \bar{z}\|^2)$. The operator $\nabla^2 q_{\bar{z}}$ is called a generalized Hessian of ϕ at $\bar{z}$ and is denoted by $H\phi(\bar{z})$.*

With this terminology in hand, we apply Propositions 9 and 10 to the case of convex programming. The proofs of these results are not required since they are a direct translation of Propositions 9 and 10 into the terminology of convex programming.

COROLLARY 12. *Let $f : \mathcal{H} \to \mathbb{R} \bigcup \{+\infty\}$ be lower semicontinuous and convex. Let $\bar{z} \in \mathcal{H}$ and set $\bar{w} = \nabla f_\lambda(\bar{z})$ and $\bar{y} = \frac{1}{\lambda}\bar{w}$. Then f_λ is twice (Fréchet) differentiable at $\bar{z}$ with $[\nabla^2 f_\lambda(\bar{z})]^{-1}$ bounded if and only if f^* has a generalized Hessian at $\bar{y}$ with $[I + \frac{1}{\lambda} Hf^*(\bar{y})]^{-1}$ bounded. In either case we have*

$$\nabla^2 f_\lambda(\bar{z}) = \left[I + \frac{1}{\lambda} Hf^*(\bar{y}) \right]^{-1} \ .$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

COROLLARY 13. *Let $f : \mathcal{H} \to \mathbb{R} \bigcup \{+\infty\}$ be lower semicontinuous and convex. Let $\bar{z} \in \mathcal{H}$ and set $\bar{y} = \bar{z} - \nabla f_\lambda(\bar{z})$. Then f_λ is twice (Fréchet) differentiable at $\bar{z}$ with $[I + \nabla^2 f_\lambda(\bar{z})]^{-1}$ bounded if and only if f is twice differentiable in the generalized sense at $\bar{y}$ with $[I + \lambda H f(\bar{y})]^{-1}$ bounded. In either case we have*

$$\nabla^2 f_\lambda(\bar{z}) = I - [I + \lambda H f(\bar{y})]^{-1} .$$

Remark. As observed earlier, the generalized Hessian is necessarily positive semidefinite. This observation can be used to further refine the statement of Corollaries 12 and 13.

5. Newton operators. In this section we study the operators associated with the variable metric proximal point iteration:

$$\mathcal{N}_k := I + H_k D_k = P_k + (H_k - I) D_k . \tag{14}$$

This notation emphasizes the fact that these operators produce Newton-like iterates. Just as in the case of the classical Newton's method for equation solving [29, section 12.6], one of the keys to the convergence analysis is to show that these operators are contractive with respect to the solution set $T^{-1}(0)$. Clearly the operators $\mathcal{N}_k$ are single-valued. Moreover, fixed points of the operators $\mathcal{N}_k$ are solutions to the inclusion $0 \in T(z)$ since

$$0 \in T(z) \;\Leftrightarrow\; P_k(z) = z \;\Leftrightarrow\; D_k(z) = 0 \;\Leftrightarrow\; \mathcal{N}_k(z) = z.$$

Thus, conditions that ensure that the operators $\mathcal{N}_k$ are nonexpansive with respect to $T^{-1}(0)$ are important for the global analysis of the variable metric proximal point iteration. To obtain this property, we impose the following conditions on the linear transformations $\{H_k\}$:

(H1) Each H_k is a continuous linear transformation with continuous inverse.

(H2) There is a nonempty closed bounded subset Γ of $T^{-1}(0)$ such that

$$\|(H_k - I) D_k(z^k)\| \le \gamma_k \|D_k(z^k)\| \quad \text{for all } k,$$

where

$$\gamma_k := \frac{\|D_k(z^k)\|}{2\sigma_k + 3\|D_k(z^k)\|} \quad \text{with } \sigma_k = \sup\{\|z^k - z\| \,:\, z \in \Gamma\}.$$

Remark. The set Γ in (H2) is used to guarantee the boundedness of the sequence $\{z^k\}$. By taking $\Gamma = \{\bar{z}\}$, one can show that every weak cluster point of the sequence $\{z^k\}$ is an element of $T^{-1}(0)$. It was observed by Iusem [13] that if $T^{-1}(0)$ is bounded and one takes $\Gamma = T^{-1}(0)$, then the sequence $\{z^k\}$ has a weak limit $z^\infty \in T^{-1}(0)$ (see Theorem 17 and [41, Theorem 1]).

Hypothesis (H1) is standard and is automatically satisfied in the finite dimensional case. On the other hand, hypothesis (H2) is quite technical and requires careful examination. This hypothesis is problematic since it specifies that the matrices H_k satisfy a condition that depends on the unknown values σ_k and $\|D_k(z^k)\|$. We will show that in certain cases it is possible to satisfy (H2) without direct knowledge of these unknown values. This is done in two steps. First it is shown in Lemma 14 that if T^{-1} is Lipschitz continuous or differentiable at the origin, then γ_k is bounded below by a positive constant (which can be taken to be 1/6 as $\|D_k(z^k)\|$ approaches zero).

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Then, in Lemma 15, it is shown that (H2) is satisfied if a related condition in terms of H_k and w^k is satisfied. Taken together, these results imply that at least locally (H2) can be satisfied by checking a condition based on known quantities.

Further insight into hypothesis (H2) can be gained by considering the case in which T^{-1} is differentiable at the origin. In this case H_k is intended to approximate $-(\nabla D_k(0))^{-1} = (I + c_k^{-1}J)$, where $J = \nabla(T^{-1})(0)$ (by Proposition 9). Hence, if $H_k \approx -(\nabla D_k(0))^{-1}$, then $(H_k - I) \approx c_k^{-1}J$. Therefore, one can guarantee that (H2) is satisfied by choosing c_k sufficiently large and $H_k \approx I$. This fact is used in [6] to establish the superlinear convergence of the method when the H_k's are obtained via matrix secant updating techniques.

The purpose of hypothesis (H2) is to globalize what is essentially a local algorithm (Newton's method). In the context of convex programming, one commonly obtains global convergence properties with the aid of a line search routine applied to the objective function f or its regularization f_λ. However, in the operator setting there is no natural underlying objective function to which a line search can be applied. This is a key difference between the approach taken in this paper and those in [3, 7, 10, 18, 20, 24, 36]. In the convex programming setting, the global convergence of the VMPPA is driven by a line search routine applied to the objective function f (or its regularization f_λ). In the operator setting, hypothesis (H2) replaces the line search and the associated hypotheses needed to make the line search strategy effective (such as the finite-valuedness of the objective function f and the boundedness of the sequence $\{H_k\}$). On the other hand, when it is known that the operator T is the subdifferential of a finite-valued finite dimensional convex function, then the algorithm of this paper can be modified to include the line search routine of Chen and Fukushima [7], thereby avoiding the need for hypothesis (H2) [6].

We now show three cases where the γ_k's are bounded away from zero.

LEMMA 14. *Suppose $T^{-1}(0)$ is nonempty.*

(i) *If the operator T is strongly monotone with modulus κ, then $T^{-1}(0) = \{\bar{z}\}$,*

$$\|z^k - \bar{z}\| \le \left(1 + \frac{1}{\kappa c_k}\right) \|D_k(z^k)\| ,$$

and $\gamma_k \ge \frac{1}{5+\frac{2}{\kappa c_k}} \ge \frac{1}{5+2/\kappa}$ for all k.

(ii) *If the operator T^{-1} is Lipschitz continuous at the origin with modulus α, then*

$$\operatorname{dist}(z^k, T^{-1}(0)) \le \left(1 + \frac{\alpha}{c_k}\right) \|D_k(z^k)\| , \tag{15}$$

for all k such that $\|D_k(z^k)\| \le \tau$, where τ is given in Definition 4. Moreover, if $T^{-1}(0) = \{\bar{z}\}$, then $\gamma_k \ge \frac{1}{5+2\alpha/c_k} \ge \frac{1}{5+2\alpha}$ for all k such that $\|D_k(z^k)\| \le \tau$.

(iii) *If T^{-1} is differentiable at the origin with derivative J, then $T^{-1}(0) = \{\bar{z}\}$, there is a $\delta > 0$ such that for all k with $\|D_k(z^k)\| \le \tau$ we have*

$$\|z^k - \bar{z}\| \le \left(1 + \frac{\|J\|}{c_k} + \sigma(\|D_k(z^k)\|)\right) \|D_k(z^k)\| ,$$

and $\gamma_k \ge \frac{1}{5+2\frac{\|J\|}{c_k}+\sigma(\|D_k(z^k)\|)}$ for all k, where $\sigma(\tau) \to 0$ as $\tau \to 0$.

Proof.

(i) If T is strongly monotone with modulus κ, then $\|z - z'\| \le \frac{1}{\kappa}\|w - w'\|$ for any z, z', w, w' such that $w \in T(z)$ and $w' \in T(z')$. That is, T^{-1} is single-valued

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

and Lipschitz continuous. Let $z = P_k(z^k)$ and $z' = \bar{z}$, where $\{\bar{z}\} = T^{-1}(0)$. By Proposition 3 a) we have $-\frac{1}{c_k} D_k(z^k) \in T(P_k(z^k))$. Hence

$$\|z^k - \bar{z}\| \le \|z^k - P_k(z^k)\| + \|P_k(z^k) - \bar{z}\| \le \left(1 + \frac{1}{\kappa c_k}\right) \|D_k(z^k)\| \ ,$$

since $D_k = P_k - I$. By the definition of γ_k,

$$\begin{aligned} \gamma_k &= \frac{\|D_k(z^k)\|}{2\|z^k - \bar{z}\| + 3\|D_k(z^k)\|} \\ &\ge \frac{\|D_k(z^k)\|}{2(1 + \frac{1}{\kappa c_k})\|D_k(z^k)\| + 3\|D_k(z^k)\|} \ge \frac{\kappa c_k}{5\kappa c_k + 2} \ . \end{aligned}$$

This establishes the result since $c_k \ge 1$ for all k.

(ii) If $\|D_k(z^k)\| \le \tau$, Definition 4 implies that

$$T^{-1}\left(-\frac{1}{c_k} D_k(z^k)\right) \subset T^{-1}(0) + \alpha \left\| \frac{1}{c_k} D_k(z^k) \right\| \mathbb{B} = T^{-1}(0) + \frac{\alpha}{c_k} \|D_k(z^k)\| \mathbb{B} \ ,$$

or

$$\left(I + T^{-1} \frac{1}{c_k}\right)(-D_k(z^k)) + D_k(z^k) \subset T^{-1}(0) + \frac{\alpha}{c_k} \|D_k(z^k)\| \mathbb{B} \ .$$

Since $D_k(z^k) = -(I + T^{-1} \frac{1}{c_k})^{-1}(z^k)$, we have $z^k \in (I + T^{-1} \frac{1}{c_k})(-D_k(z^k))$, and so

$$z^k \in T^{-1}(0) - D_k(z^k) + \frac{\alpha}{c_k} \|D_k(z^k)\| \mathbb{B}.$$

Hence (15) holds. If $T^{-1}(0) = \{\bar{z}\}$, then the lower bound on γ_k follows as in part (i).

(iii) This result follows as in part (ii) using the second remark after Definition 4. ∎

When $w^k \approx D_k(z^k)$, one can establish the inequality in hypothesis (H2) from a related condition on the vectors w^k. A specific technique for accomplishing this is given in the following lemma.

LEMMA 15. *Let $\xi, \hat{\gamma}_k, \delta_k \in \mathbb{R}_+$ be such that*

$$0 \le \xi < 1, \ \delta_k \le \min\left\{1, \|H_k\|^{-1}\right\} \frac{3}{7}(1 - \xi)\hat{\gamma}_k, \ \text{and } \hat{\gamma}_k \le \frac{1}{3}, \tag{16}$$

and let H_k be a continuous linear transformation from $\mathcal{H}$ to itself. If $z^k, w^k \in \mathcal{H}$ satisfy

$$\|(I - H_k)w^k\| \le \xi\hat{\gamma}_k \|w^k\| \ \text{and} \ \|w^k - D_k(z^k)\| \le \delta_k \|w^k\|, \tag{17}$$

then $\|(I - H_k)D_k(z^k)\| \le \hat{\gamma}_k \|D_k(z^k)\|$. Therefore, if (H1) *and criterion* ($\mathcal{L}$) *are satisfied, and if ξ and the sequence $\{(\hat{\gamma}_k, \delta_k)\} \subset \mathbb{R}^2$ satisfy* (16)*, with $\hat{\gamma}_k \le \gamma_k$ for all k (where γ_k is defined in* (H2)*), then hypothesis* (H2) *is satisfied.*

Proof. From (16) and (17), we have

$$\|w^k\| \le \|D_k(z^k)\| + \|w^k - D_k(z^k)\| \le \|D_k(z^k)\| + \frac{3}{7}(1 - \xi)\hat{\gamma}_k \|w^k\|;$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

hence

$$\|w^k\| \leq \frac{1}{1-\frac{3}{7}(1-\xi)\hat{\gamma}_k}\|D_k(z^k)\| \ .$$

Again by (17),

$$\begin{aligned}\|(I-H_k)D_k(z^k)\| &\leq \|(I-H_k)w^k\| + \|H_k\|\|w^k - D_k(z^k)\| + \|w^k - D_k(z^k)\| \\ &\leq \xi\hat{\gamma}_k\|w^k\| + (\|H_k\|+1)\delta_k\|w^k\| \leq \left(\xi + \frac{6}{7}(1-\xi)\right)\hat{\gamma}_k\|w^k\| \\ &\leq \frac{\xi+\frac{6}{7}(1-\xi)}{1-\frac{3}{7}(1-\xi)\hat{\gamma}_k}\hat{\gamma}_k\|D_k(z^k)\| \leq \hat{\gamma}_k\|D_k(z^k)\|\end{aligned}$$

since the inequality $\hat{\gamma}_k \leq \frac{1}{3}$ implies that $\frac{\xi+\frac{6}{7}(1-\xi)}{1-\frac{3}{7}(1-\xi)\hat{\gamma}_k} = \frac{6+\xi}{7-3(1-\xi)\hat{\gamma}_k} \leq 1$. ∎

We conclude this section by showing that the operators $\mathcal{N}_k$ are nonexpansive with respect to the set $T^{-1}(0)$.

PROPOSITION 16. *Assume $T^{-1}(0)$ is nonempty. If the sequence of linear transformations $\{H_k\}$ satisfies hypotheses* (H1) *and* (H2)*, then for all k we have $\|H_kD_k(z^k)\| \leq \frac{3}{2}\|D_k(z^k)\|$ and*

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 \leq \|z^k - \bar{z}\|^2 \quad \textit{for all } \bar{z} \in \Gamma. \tag{18}$$

Proof. Let $\bar{z} \in \Gamma$. From the definitions of P_k and $\mathcal{N}_k$, we have

$$\|P_k(z^k) - \bar{z}\| = \|\mathcal{N}_k(z^k) - (H_k - I)D_k(z^k) - \bar{z}\| \geq |\|\mathcal{N}_k(z^k) - \bar{z}\| - \|(H_k - I)D_k(z^k)\||; \tag{19}$$

hence

$$\|P_k(z^k) - \bar{z}\|^2 \geq \|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \|(H_k - I)D_k(z^k)\|^2 - 2\|(H_k - I)D_k(z^k)\|\|\mathcal{N}_k(z^k) - \bar{z}\| \ . \tag{20}$$

From hypothesis (H2), we have

$$\|H_kD_k(z^k)\| \leq \|D_k(z^k)\| + \|(H_k - I)D_k(z^k)\| \leq (1+\gamma_k)\|D_k(z^k)\| \leq \frac{3}{2}\|D_k(z^k)\| \ .$$

Hence

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \|z^k - \bar{z}\| + \|H_kD_k(z^k)\| \leq \sigma_k + \frac{3}{2}\|D_k(z^k)\| \ .$$

Then, again by hypothesis (H2),

$$\|(H_k - I)D_k(z^k)\| \leq \gamma_k\|D_k(z^k)\| = \frac{\|D_k(z^k)\|^2}{2\sigma_k + 3\|D_k(z^k)\|} \leq \frac{\|D_k(z^k)\|^2}{2\|\mathcal{N}_k(z^k) - \bar{z}\|} \ . \tag{21}$$

Thus, from (20) and (21),

$$\|P_k(z^k) - \bar{z}\|^2 \geq \|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \|(H_k - I)D_k(z^k)\|^2 - \|D_k(z^k)\|^2 \ . \tag{22}$$

Letting $z = z^k$ and $z' = \bar{z}$ in Proposition 3 c) yields

$$\|P_k(z^k) - \bar{z}\|^2 + \|D_k(z^k)\|^2 \leq \|z^k - \bar{z}\|^2 \ . \tag{23}$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

From (22) and (23) we have

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \|(H_k - I)D_k(z^k)\|^2 \leq \|z^k - \bar{z}\|^2 \ . \tag{24}$$

We now consider $\alpha_k = \frac{\|(H_k - I)D_k(z^k)\|}{\|D_k(z^k)\|}$. If $\alpha_k \geq \frac{\gamma_k}{2}$, then (18) holds by (24). Suppose that $\alpha_k < \frac{\gamma_k}{2}$. From (19), we have

$$\|P_k(z^k) - \bar{z}\| \geq \|\mathcal{N}_k(z^k) - \bar{z}\| - \frac{\gamma_k}{2}\|D_k(z^k)\| \ .$$

Therefore, by (23),

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \sqrt{\|z^k - \bar{z}\|^2 - \|D_k(z^k)\|^2} + \frac{\gamma_k}{2}\|D_k(z^k)\| \ .$$

Using the inequality $\sqrt{a^2 - b^2} \leq a - \frac{b^2}{2a}$ for $a > b > 0$,

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \|z^k - \bar{z}_k\| - \frac{\|D_k(z^k)\|^2}{2\|z^k - \bar{z}\|} + \frac{\gamma_k}{2}\|D_k(z^k)\| \ .$$

But $\frac{\|D_k(z^k)\|}{2\|z^k - \bar{z}\|} \geq \gamma_k$; thus

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \|z^k - \bar{z}\| - \frac{\gamma_k}{2}\|D_k(z^k)\|$$

or

$$\|\mathcal{N}_k(z^k) - \bar{z}\| + \frac{\gamma_k}{2}\|D_k(z^k)\| \leq \|z^k - \bar{z}\| \ . \tag{25}$$

From (25) we again obtain (18). □

6. Global convergence. The statement and proof of the global convergence result given below parallels the development given by Rockafellar in [41, Theorem 1] for the classical PPA.

THEOREM 17. *Let $\{z^k\}$ be any sequence generated by the VMPPA under criterion $(\mathcal{G})$ (or $(\mathcal{G}')$). Suppose that the solution set $T^{-1}(0)$ is nonempty and the sequence of linear transformations $\{H_k\}$ satisfies the hypotheses* (H1) *and* (H2). *Then the sequence $\{z^k\}$ is bounded, each weak cluster point of this sequence is an element of $T^{-1}(0)$, and $\lim_k D_k(z^k) = 0$. If it is also assumed that $T^{-1}(0)$ is bounded and $\Gamma = T^{-1}(0)$ in* (H2), *then there is a $\bar{z} \in T^{-1}(0)$ such that $\{z^k\}$ converges weakly to $\bar{z}$.*

In order to establish this result we require the following technical lemma, whose proof is straightforward and so is omitted.

LEMMA 18. *Suppose the nonnegative sequences $\{\epsilon_k\}$ satisfy $\sum_{k=0}^{\infty} \epsilon_k < +\infty$. If $\{u_k\}$ is a nonnegative sequence satisfying $u_{k+1} \leq \epsilon_k + u_k$, then $\{u_k\}$ is a Cauchy sequence.*

Proof of Theorem 17. We begin by showing that the limit $\lim_k \|z^k - \bar{z}\| = \mu(\bar{z})$ exists for every $\bar{z} \in \Gamma$. To this end let $\bar{z} \in \Gamma$ and observe that the definition of $\mathcal{N}_k$ and Proposition 16 imply that

$$\begin{aligned}\|z^{k+1} - \bar{z}\| &= \|z^{k+1} - \mathcal{N}_k(z^k) + \mathcal{N}_k(z^k) - \bar{z}\| \leq \|z^{k+1} - \mathcal{N}_k(z^k)\| + \|\mathcal{N}_k(z^k) - \bar{z}\| \\ &\leq \|H_k\|\|w^k - D_k(z^k)\| + \|z^k - \bar{z}\| \leq \epsilon_k + \|z^k - \bar{z}\| \ .\end{aligned}$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Therefore, Lemma 18 implies that the sequence $\{\|z^k - \bar{z}\|\}$ is Cauchy, and so $\mu(\bar{z})$ exists for every $\bar{z} \in \Gamma$. An immediate consequence of the existence of these limits is the boundedness of the sequences $\{z^k\}$ and σ_k.

We now show that the sequence $\{D_k(z^k)\}$ converges strongly to the origin. Indeed, if this is not the case, then there is a subsequence $J \subset \{1, 2, \ldots\}$ such that $\inf_J \|D_k(z^k)\| = \beta_1 > 0$. This in turn implies that $\inf_J \gamma_k = \beta_2 > 0$ since otherwise $\lim_J \|D_k(z^k)\| = 0$ due to the boundedness of the sequence $\{\sigma_k\}$. Let $\bar{z} \in \Gamma$. By Proposition 16,

$$\begin{aligned}
&\frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 - \|z^k - \bar{z}\|^2 + \|z^{k+1} - \bar{z}\|^2 \le \|z^{k+1} - \bar{z}\|^2 - \|\mathcal{N}_k(z^k) - \bar{z}\|^2 \\
&= \langle z^{k+1} - \mathcal{N}_k(z^k), z^{k+1} - \bar{z} + \mathcal{N}_k(z^k) - \bar{z}\rangle \\
&\le \|z^{k+1} - \mathcal{N}_k(z^k)\|(\|z^{k+1} - \bar{z}\| + \|\mathcal{N}_k(z^k) - \bar{z}\|) \\
&\le \|H_k\|\|w^k - D_k(z^k)\|(\|z^{k+1} - \bar{z}\| + \|z^k - \bar{z}\|) \le \epsilon_k(\|z^{k+1}\| + 2\|\bar{z}\| + \|z^k\|) = \epsilon_k C_k \ ,
\end{aligned}$$

with $\{C_k\}$ bounded, where the final inequality follows from criterion $(\mathcal{G})$. Hence

$$\frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 \le \|z^k - \bar{z}\|^2 - \|z^{k+1} - \bar{z}\|^2 + \epsilon_k C_k,$$

whereby we obtain the contradiction

$$\begin{aligned}
0 < \frac{\beta_1^2\beta_2^2}{4} &\le \limsup_J \frac{\gamma_k^2}{4}\|D_k(z^k)\|^2 \\
&\le \lim_J(\|z^k - \bar{z}\|^2 - \|z^{k+1} - \bar{z}\|^2 + \epsilon_k C_k) = \mu(\bar{z}) - \mu(\bar{z}) + 0 = 0 \ .
\end{aligned}$$

Therefore, $\lim_k \|D_k(z^k)\| = 0$.

Next let $J \subset \{1, 2, \ldots\}$ be such that the subsequence $\{z^k\}_J$ converges weakly to z^∞, i.e., z^∞ is a weak cluster point of the sequence $\{z^k\}$. We show that z^∞ must be an element of $T^{-1}(0)$. From Proposition 3 a), we have that $-\frac{1}{c_k}D_k(z^k) \in T(P_k(z^k))$ for all k; hence $0 \le \langle z - P_k(z^k), w + \frac{1}{c_k}D_k(z^k)\rangle$, or equivalently, $\langle z - z^k - D_k(z^k), w + \frac{1}{c_k}D_k(z^k)\rangle \ge 0$ for all k and z, w with $w \in T(z)$. Taking the limit over J yields the inequality $\langle z - z^\infty, w\rangle \ge 0$ for all z, w with $w \in T(z)$. Since T is maximal monotone, we get $0 \in T(z^\infty)$.

Under the assumption that $\Gamma = T^{-1}(0)$, the argument showing that there is no more than one weak cluster point of $\{z^k\}$ is identical to the one given by Rockafellar in [41, Theorem 1]. □

Remark. To ensure the strong convergence of the sequence $\{z^k\}$, one again requires a growth condition on the inverse mapping T^{-1} in a neighborhood of the origin. Rockafellar has shown that Lipschitz continuity of T^{-1} at the origin suffices for this purpose [41, Theorem 2]. Other conditions can be found in the work of Luque [21, Proposition 1.2]. The results of Rockafellar and Luque are easily extended to the VMPPA.

7. Convergence rates.

7.1. Linear convergence. Just as in Rockafellar [41, Theorem 2], we require that the operator T^{-1} is Lipschitz continuous at the origin in order to establish that the convergence rate is at least linear.

THEOREM 19. *Let $\{z^k\}$ be any sequence generated by the VMPPA satisfying both criteria $(\mathcal{G})$ and $(\mathcal{L})$ for all k. Assume that T^{-1} is Lipschitz continuous at the*

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

origin with modulus α *and the solution set* $T^{-1}(0)$ *is a singleton* $\{\bar{z}\}$. *If the sequence* $\{H_k\}$ *satisfies the hypotheses* (H1) *and* (H2) *with* $\delta_k\|H_k\| \to 0$, *then the sequence* $\{z^k\}$ *strongly converges to the solution and there is an index* $\bar{k}$ *such that*

$$\|z^{k+1} - \bar{z}\| \leq \sigma_k \|z^k - \bar{z}\| \quad \textit{for all } k \geq \bar{k} \ ,$$

where σ_k *satisfies* $\limsup_{k\to\infty} \sigma_k < 1$. *That is, the convergence rate is linear.*

Proof. By Theorem 17, we have $\|D_k(z^k)\| \to 0$. Hence, Part (ii) of Lemma 14 implies that $\{z^k\}$ converges strongly to $\bar{z}$. We now establish the linear rate.

Let $\tau > 0$ be as in Definition 4, and let $\tilde{k}$ be such that $\|\frac{1}{c_k} D_k(z^k)\| \leq \tau$ for all $k \geq \tilde{k}$. By Proposition 3 a) and the Lipschitz continuity of T^{-1} at 0, we have

$$\|P_k(z^k) - \bar{z}\| \leq \frac{\alpha}{c_k} \|D_k(z^k)\| \ . \tag{26}$$

Hence relation (14) and hypothesis (H2) yield

$$\begin{aligned} \|\mathcal{N}_k(z^k) - \bar{z}\| &= \|P_k(z^k) + (H_k - I)D_k(z^k) - \bar{z}\| \\ &\leq \|P_k(z^k) - \bar{z}\| + \gamma_k \|D_k(z^k)\| \ . \end{aligned} \tag{27}$$

Let $a_k := \frac{\alpha}{c_k} + \gamma_k$. Using (26) and (27),

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \left(\frac{\alpha}{c_k} + \gamma_k \right) \|D_k(z^k)\| = a_k \|D_k(z^k)\| \ . \tag{28}$$

Let $\gamma := \frac{1}{2(5+2\alpha)}$. By Proposition 16 and Lemma 14 we have, for $k \geq \tilde{k}$, that

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 + \gamma^2 \|D_k(z^k)\|^2 \leq \|z^k - \bar{z}\|^2 \ . \tag{29}$$

By (28) and (29), when $k \geq \tilde{k}$,

$$\|\mathcal{N}_k(z^k) - \bar{z}\|^2 \leq {a_k}^2 \|D_k(z^k)\|^2 \leq \frac{{a_k}^2}{\gamma^2} \|z^k - \bar{z}\|^2 - \frac{{a_k}^2}{\gamma^2} \|\mathcal{N}_k(z^k) - \bar{z}\|^2 \ . \tag{30}$$

Let $\mu_k := \frac{a_k}{\sqrt{a_k^2 + \gamma^2}}$. From (30) we have

$$\|\mathcal{N}_k(z^k) - \bar{z}\| \leq \mu_k \|z^k - \bar{z}\| \ . \tag{31}$$

By (31), criterion $(\mathcal{L})$ (or $(\mathcal{L}')$), and Proposition 3 c),

$$\begin{aligned} &\|z^{k+1} - \bar{z}\| \leq \|z^{k+1} - \mathcal{N}_k(z^k)\| + \|\mathcal{N}_k(z^k) - \bar{z}\| \\ &\leq \delta_k \|H_k\| \|w^k\| + \mu_k \|z^k - \bar{z}\| \leq \frac{\delta_k \|H_k\|}{1 - \delta_k} \|D_k(z^k)\| + \mu_k \|z^k - \bar{z}\| \\ &\leq \left(\frac{\delta_k \|H_k\|}{1 - \delta_k} + \mu_k \right) \|z^k - \bar{z}\| = \sigma_k \|z^k - \bar{z}\| \ , \end{aligned}$$

where $\sigma_k := \frac{\delta_k \|H_k\|}{1 - \delta_k} + \mu_k$. Since there is a $\tilde{\delta} > 0$ such that $\mu_k < 1 - \tilde{\delta}$ for any k, and $\delta_k \|H_k\| \to 0$, we have $\sigma_k < 1$ for k sufficiently large. Moreover, we have $\limsup_{k\to\infty} \sigma_k = \limsup_{k\to\infty} \mu_k \leq 1 - \tilde{\delta}$. ∎

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

7.2. Superlinear convergence. We now give an analogue of Dennis and Moré's [14] characterization theorem for the superlinear convergence of variable metric methods in nonlinear programming that applies to the VMPPA. This result is used in [6] to establish the superlinear convergence of the VMPPA when the Broyden (nonsymmetric case) or the BFGS (symmetric case) updating formula is used to generate the matrices H_k.

THEOREM 20. *Let $\{z^k\}$ be any sequence generated by the VMPPA 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 $\bar{z}$ superlinearly if and only if*

$$\frac{[I - (I + \frac{1}{c_k}J)H_k^{-1}](z^{k+1} - z^k)}{\|z^{k+1} - z^k\|} \to 0 \quad \text{as } k \to \infty \ . \tag{32}$$

Remark. By Proposition 9 we have $\nabla D(\bar{z}) = -(I + \frac{1}{c}J)^{-1}$. Consequently, condition (32) can be recast in the more familiar form given in [15, Theorem 8.2.4]. Note that the assumption in (32) on the sequence $\{H_k\}$ is much weaker than assuming that this sequence converges. Specific choices of the linear transformations H_k satisfying (32) are discussed in [6].

The proof of Theorem 20 requires the following lemma.

LEMMA 21. *Under the conditions in Theorem* 20 *we have*

(a) $T^{-1}(\frac{-1}{c_k}D_k(z^k)) - \bar{z} - J(\frac{-1}{c_k}D_k(z^k)) \subset o(\|z^k - \bar{z}\|)\mathbb{B}$, *and*

(b) $(I + \frac{1}{c_k}J)H_k^{-1}(z^{k+1} - \mathcal{N}_k(z^k)) \in o(\|z^k - \bar{z}\|)\mathbb{B}$,

for all k sufficiently large.

Proof. For part (a), let $\delta > 0$ be such that

$$T^{-1}(w) - Jw - \bar{z} \subset o(\|w\|)\mathbb{B} \tag{33}$$

whenever $\|w\| < \delta$. Let $\bar{k}_1$ be such that whenever $k > \bar{k}_1$, $\|D_k(z^k)\| \le \delta$. Then, by (33) and Proposition 3 c), when $k > \bar{k}_1$,

$$T^{-1}\left(\frac{-1}{c_k}D_k(z^k)\right) - \bar{z} - J\left(\frac{-1}{c_k}D_k(z^k)\right) \subset o(\|D_k(z^k)\|)\mathbb{B} \subset o(\|z^k - \bar{z}\|)\mathbb{B} \ .$$

We now prove (b). Note that $\mathcal{N}_k(z^k) = (I + H_kD_k)(z^k)$; hence by criterion $(\mathcal{L})$

$$\begin{aligned}\left\|\left(I + \frac{1}{c_k}J\right)H_k^{-1}(z^{k+1} - \mathcal{N}_k(z^k))\right\| &= \left\|\left(I + \frac{1}{c_k}J(w^k - D_k(z^k)\right)\right\| \\ &\le (1 + \|J\|)\|w^k - D_k(z^k)\| \le \delta_k(1 + \|J\|)\|w^k\| \\ &\le \frac{\delta_k(1 + \|J\|)}{1 - \delta_k}\|D_k(z^k)\| \ . \end{aligned} \tag{34}$$

Therefore by (34) and Proposition 3 c),

$$\left(I + \frac{1}{c_k}J\right)H_k^{-1}(z^{k+1} - \mathcal{N}_k(z^k)) \in o(\|D_k(z^k)\|)\mathbb{B} \subset o(\|z^k - \bar{z}\|)\mathbb{B}. \qquad \square$$

Proof of Theorem 20. Let $\tilde{z}^{k+1} := \mathcal{N}_k(z^k) = (I + H_kD_k)(z^k)$. By Proposition 3 a) we have $\tilde{z}^{k+1} = z^k - H_k(I + T^{-1}\frac{1}{c_k})^{-1}(z^k)$. Hence

$$\begin{aligned} z^k &\in \left(I + T^{-1}\frac{1}{c_k}\right)[H_k^{-1}(z^k - \tilde{z}^{k+1})] \\ &= H_k^{-1}(z^k - \tilde{z}^{k+1}) + T^{-1}\left[\frac{1}{c_k}H_k^{-1}(z^k - \tilde{z}^{k+1})\right] \ , \end{aligned}$$

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

or equivalently,

$$\begin{aligned}
z^{k+1}-\bar z &= z^k-\bar z+(z^{k+1}-z^k)\\
&\in \left[T^{-1}\left(\frac{1}{c_k}H_k^{-1}(z^k-\tilde z^{k+1})\right)-\bar z+(z^{k+1}-z^k)+H_k^{-1}(z^k-\tilde z^{k+1})\right]\\
&= \left[T^{-1}\left(\frac{1}{c_k}H_k^{-1}(z^k-\tilde z^{k+1})\right)-\bar z-J\left(\frac{1}{c_k}H_k^{-1}(z^k-\tilde z^{k+1})\right)\right]\\
&\quad+\left[I-\left(I+\frac{1}{c_k}J\right)H_k^{-1}\right](z^{k+1}-z^k)\\
&\quad+\left(I+\frac{1}{c_k}J\right)H_k^{-1}(z^{k+1}-\tilde z^{k+1})\\
&= \left[T^{-1}\left(\frac{-1}{c_k}D_k(z^k)\right)-\bar z-J\left(\frac{-1}{c_k}D_k(z^k)\right)\right]\\
&\quad+\left[I-\left(I+\frac{1}{c_k}J\right)H_k^{-1}\right](z^{k+1}-z^k)\\
&\quad+\left(I+\frac{1}{c_k}J\right)H_k^{-1}(z^{k+1}-\tilde z^{k+1})\ . \qquad (35)
\end{aligned}$$

By Lemma 21 the first and third of the three terms appearing on the right-hand side of this inclusion can be bounded by an expression of the form $o(\|z^k-\bar z\|)\mathbb{B}$. If (32) holds, then $[I-(I+\frac{1}{c_k}J)H_k^{-1}](z^{k+1}-z^k)\in o(\|z^{k+1}-z^k\|)\mathbb{B}$. Therefore there are positive sequences $\{\alpha_{1k}\}$ and $\{\alpha_{2k}\}$, each converging to zero such that, for $k>\bar k_1$,

$$\begin{aligned}
\|z^{k+1}-\bar z\| &\le \alpha_{1k}\|z^{k+1}-z^k\|+\alpha_{2k}\|z^k-\bar z\|\\
&\le \alpha_{1k}(\|z^k-\bar z\|+\|z^{k+1}-\bar z\|)+\alpha_{2k}\|z^k-\bar z\|\\
&= \alpha_{1k}\|z^{k+1}-\bar z\|+(\alpha_{1k}+\alpha_{2k})\|z^k-\bar z\|.
\end{aligned}$$

Let $\bar k_2>\bar k_1$ be such that $\alpha_{1k}<\frac12$ for all $k>\bar k_2$. Then, denoting $\frac{\alpha_{1k}+\alpha_{2k}}{1-\alpha_{1k}}$ by τ_k,

$$\|z^{k+1}-\bar z\|\le\frac{\alpha_{1k}+\alpha_{2k}}{1-\alpha_{1k}}\|z^k-\bar z\|=\tau_k\|z^k-\bar z\|$$

whenever $k>\bar k_2$, and $\tau_k\to0$ as $k\to\infty$. Therefore $\{z^k\}$ converges to $\bar z$ superlinearly.

Conversely, suppose that

$$\lim_{k\to\infty}\frac{\|z^{k+1}-\bar z\|}{\|z^k-\bar z\|}=0\ . \qquad (36)$$

Divide (35) by $\|z^k-\bar z\|$ and let $k\to\infty$. From (36) and Lemma 21 we obtain

$$\frac{[I-(I+\frac{1}{c_k}J)H_k^{-1}](z^{k+1}-z^k)}{\|z^k-\bar z\|}\to0\quad\text{as } k\to\infty\ .$$

However, from (36) we have

$$\frac{\|z^k-\bar z\|}{\|z^{k+1}-z^k\|}\le\frac{\|z^k-\bar z\|}{\|z^k-\bar z\|-\|z^{k+1}-\bar z\|}=\frac{1}{1-\frac{\|z^{k+1}-\bar z\|}{\|z^k-\bar z\|}}\to1$$

as $k\to\infty$. Hence (32) holds. ∎

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

8. Concluding remarks. In this paper, we introduced a new PPA for solving the inclusion $0 \in T(x)$, where T is an arbitrary maximal monotone operator. The global convergence of the algorithm is demonstrated with an inexact solution at each step. This is important in practice, since solving for the exact solution at each step is impractical and may in fact be almost as difficult as solving the original problem. If it is assumed that T^{-1} is Lipschitz continuous at the origin, then the method is shown to be linearly convergent. If it is further assumed that T^{-1} is differentiable at the origin, then the classical characterization of superlinear convergence due to Dennis and Moré also holds for the VMPPA. In [6], this characterization of superlinear convergence is applied to establish the super-linear convergence of the method when certain matrix secant updating strategies are employed to generate the matrices H_k. In [5], we give some of the implementation details in the case of convex programming. We show how to apply the method to solve the associated primal, dual, and Lagrangian saddle point problems. In particular, it is shown how the bundle technique [17] can be applied to satisfy the approximation criteria $(\mathcal{L})$ and $(\mathcal{G})$ in both the primal and saddle point solution techniques. Preliminary numerical results comparing these three approaches are also presented.

Acknowledgments. The authors would like to thank the reviewers for their thorough work. Their comments and suggestions have greatly contributed to our exposition. In particular, we would like to thank Professor Alfredo Iusem for observing an error in an earlier version of Theorem 17 and for his suggested revision of this result when the set $T^{-1}(0)$ is assumed to be bounded.

REFERENCES

[1] A.D. Alexandrov, *The existence almost everywhere of the second differential of a convex function and some associated properties of convex surfaces*, Ucenye Zapiski Leningr. Gos. Univ. Ser. Mat., 37 (1939), pp. 3–35 (in Russian).

[2] J.P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[3] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal, *A family of variable metric proximal point methods*, Math. Programming, 68 (1995), pp. 15–47.

[4] L.M. Bregman, *The method of successive projection for finding a common point of convex sets*, Soviet Mathematics Doklady, 162 (1965), pp. 487–490.

[5] J.V. Burke and M. Qian, *On the local super-linear convergence of a matrix secant implementation of the variable metric proximal point algorithm for monotone operators*, in Reformulation—Nonsmooth, Piecewise Smooth, Semi-smooth, and Smoothing Methods, L. Qi and M. Fukushima, eds., Kluwer Academic Publishers, Norwell, MA, 1998, pp. 317–334.

[6] J.V. Burke and M. Qian, *On the super-linear convergence of the variable metric proximal point algorithm using Broyden and BFGS matrix secant updating*, Math. Programming, 1999, to appear.

[7] X. Chen and M. Fukushima, *Proximal Quasi-Newton Methods for Nondifferentiable Convex Optimization*, Math. Programming, 1999, to appear.

[8] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1980.

[9] J. Eckstein, *Splitting Methods for Monotone Operators with Application to Parallel Optimization*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[10] M. Fukushima and L. Qi, *A globally and superlinearly convergent algorithm for nonsmooth convex minimization*, SIAM J. Optim., 6 (1996), pp. 1106–1120.

[11] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim., 2 (1992), pp. 649–664.

[12] S. Han, *A decomposition method and its application to convex programming*, Math. Oper. Res., 14 (1989), pp. 237–248.

[13] A. Iusem, private communication, IMPA, Rio de Janeiro, Brazil, 1996.

[14] J.E. Dennis, Jr., and J.J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[15] J.E. Dennis, Jr., and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[16] G. Kassay, *The proximal points algorithm for reflexive Banach spaces*, Studia Univ. Babes-Bolyai Math., 30 (1930), pp. 9–17.
[17] C. Lemaréchal, *Bundle methods in nonsmooth optimization*, in Nonsmooth Optimization, C. Lemaréchal and R. Mifflin, eds., Pergamon Press, Oxford, 1978.
[18] C. Lemaréchal and C. Sagastizábal, *An approach to variable metric bundle methods*, in IFIP Proceedings, Systems Modeling and Optimization, J. Henry and J.P. Yuan, eds., Springer, Berlin, 1994, pp. 144–162.
[19] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385.
[20] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Programming, 76 (1997), pp. 393–410.
[21] F.J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22 (1984), pp. 277–293.
[22] B. Martinet, *Regularisation d'inequations variationelles par approximations successive*, Revue Française d'Informatique et de Recherche Opérationelle, 4 (1970), pp. 154–158.
[23] B. Martinet, *Determination approachée d'un point fixe d'une application pseudo-contractante. cas de l'application prox*, Comptes Rendus de l'Académie des Sciences, Paris, Série A, 274 (1972), pp. 163–165.
[24] R. Mifflin, *A quasi–second-order proximal bundle algorithm*, Math. Programming, 73 (1996), pp. 51–72.
[25] R. Mifflin, D. Sun, and L. Qi, *Quasi-Newton Bundle-Type Methods for Nondifferentiable Convex Optimization*, Technical report AMR 96/21, Dept. of Applied Math., University of New South Wales, Sydney, New South Wales, Australia, 1996.
[26] F. Mignot, *Control dan les inequations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
[27] G.J. Minty, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341–346.
[28] J.J. Moreau, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
[29] J.M. Ortega and W.G. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[30] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
[31] J.-S. Pang and L. Qi, *A globally convergent Newton method for SC^1 problems*, J. Optim. Theory Appl., 85 (1995), pp. 633–648.
[32] G.B. Passty, *Weak convergence theorems for nonexpansive mappings in Banach spaces*, J. Math. Anal. Appl., 67 (1979), pp. 274–276.
[33] R.R. Phelps, *Convex Functions, Monotone Operators, and Differentiability*, Lecture Notes in Math., Springer-Verlag, New York, 1989.
[34] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[35] L. Qi, *Second-order analysis of the Moreau-Yosida regularization of a convex function*, Technical report AMR 94/20, Dept. of Applied Math., University of New South Wales, Sydney, New South Wales, Australia, 1994.
[36] L. Qi and X. Chen, *A preconditioning proximal Newton method for nondifferentiable convex optimization*, Math. Programming, 76 (1995), pp. 411–430.
[37] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 66 (1994), pp. 25–43.
[38] M. Qian, *The Variable Metric Proximal Point Algorithm: Theory and Application*, Ph.D. thesis, University of Washington, Seattle, WA, 1992.
[39] R.T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.
[40] R.T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[41] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[42] R.T. Rockafellar, *Maximal monotone relations and the second derivatives of nonsmooth functions*, Ann. Inst. H. Poincaré Analyse Non Linéaire, 2 (1985), pp. 167–184.
[43] J.E. Spingarn, *Partial inverse of a monotone operator*, Appl. Math. Optim., 10 (1983), pp. 247–265.
[44] J.E. Spingarn, *Applications of the methods of partial inverses to convex programming: Decomposition*, Math. Programming, 32 (1985), pp. 199–223.

Downloaded 07/20/15 to 205.175.118.98. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php