Journal of Nonlinear and Convex Analysis 22/11 (2021), 2347–2374

ADVANCES IN CONVERGENCE AND SCOPE OF THE PROXIMAL POINT ALGORITHM

R. Tyrrell Rockafellar[1]

Abstract

The proximal point algorithm, as an approach to finding a zero of a maximal monotone mapping, is well known for its role in numerical optimization, such as in methods of multipliers (ALM). Although originally designed for global reach, versions of the algorithm have been developed that can operate with only local maximal monotonocity, which is essential for applications to nonconvex optimization. Here such local articulation is investigated with inexact iterations involving trust regions. The main accomplishment is the development of much sharper criteria for when the convergence to a solution will be linear or superlinear, even if the solution might not be locally unique. The case of subgradient mappings receives extra attention.

The results are furthermore extended to a variable metric version of the localized algorithm that potentially offers additional improvements in the rate of linear convergence. For problems where the solution has a number of components, this version supports an articulation of the iterations that admits separate proximal parameters for each of those components.

Keywords: *numerical optimization, proximal point algorithm, local maximal monotonicity, variational convexity, nonunique solutions, linear convergence, directional convergence, variable metric implementation*

i

(preprint of 9 July 2021 with post-publication fixes in Section 4 indicated in footnotes)

[1]University of Washington, Department of Mathematics, Box 354350, Seattle, WA 98195-4350;
E-mail: *rtr@uw.edu*, URL: www.math.washington.edu/∼rtr/mypage.html

1 Introduction

An enormous range of problems in optimization and equilibrium modeling can be reduced to solving a generalized equation for a possibly set-valued mapping $T : \mathcal{H} \rightrightarrows \mathcal{H}$ in a real Hilbert space $\mathcal{H}$:

$$\text{find } \bar{z} \text{ such that } 0 \in T(\bar{z}), \text{ or equivalently, } \bar{z} \in Z \text{ for } Z = T^{-1}(0). \tag{1.1}$$

Especially prominent among such problems are those in which T is *maximal monotone*, in which case T and T^{-1} are closed-convex-valued. Monotonicity of T means that

$$\langle w_1 - w_0, z_1 - z_0 \rangle \geq 0 \text{ when } (z_i, w_i) \in \operatorname{gph} T, \tag{1.2}$$

where $\operatorname{gph} T$ is the graph set $\{ (z, w) \mid w \in T(z) \}$, and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}$. Maximality adds that there is no monotone mapping $T' : I\!R^n \rightrightarrows I\!R^n$ with $\operatorname{gph} T' \supset \operatorname{gph} T$ and $\operatorname{gph} T' \neq \operatorname{gph} T$. Then a solution to (1.1), if any exists, can be determined by the proximal point algorithm as laid out in 1976 [20] which, starting from any initial z^0, takes[2]

$$z^{k+1} \approx P_k(z^k) \text{ for } P_k = (I + c_k T)^{-1}, \text{ where } 1 \leq c_k \to c_\infty \in (0, \infty]. \tag{1.3}$$

Although T may be set-valued, maximal monotonicity makes the resolvant mappings P_k in (1.3) be single-valued and globally Lipschitz continuous with modulus 1, i.e., nonexpansive — in fact with the stronger property of being firmly nonexpansive:

$$||P_k(z) - P_k(z')||^2 + ||Q_k(z) - Q_k(z')||^2 \leq ||z - z'||^2 \text{ with } Q_k = I - P_k = (I + (c_k T)^{-1})^{-1}. \tag{1.4}$$

Solutions to (1.1) are the fixed points of the mappings P_k.

Although it may be hard to imagine how the iterations in (1.3) could be executed in a practical manner, if (1.1) itself can't just be solved directly, the long history of applications in optimization [21], [18], and elsewhere, testifies the opposite. The procedure has the remarkable property of generating a sequence of points z^k that weakly converges to some particular solution $\bar{z} \in Z$, even when the set Z isn't a singleton. Questions of rate of convergence have remained open, however.

In [20], only the singleton case of Z was supplied with a criterion for linear convergence. The non-singleton case is important, though, for applications of the proximal point algorithm to augmented Lagrangian methods. In that context the procedure is invoked in solving a dual problem in the multiplier space. Requiring the multiplier vector to be unique may be inconvenient — the associated solution to the primal problem may be all that really matters.

Luque [11] in 1984 provided a linear convergence criterion for the non-singleton case of Z, although with a slightly weakened result. He obtained linear convergence to 0 not of the distance of z^k from the limit point $\bar{z} \in Z$, just the distance of z^k from Z. His condition on Z was global in character, and that nowadays raises a further issue because of advances in localization. Pennanen [14] in 2003 showed (under some restrictions that no longer appear necessary) that the proximal point algorithm could manage with global maximal monotonicity replaced by maximal monotonicity relative a neighborhood of a pair $(z, 0)$ with $z \in Z$. The iterations only need to start from a point z^0 near enough to such a z and take care not to stray too far from it. In such an execution of the procedure, a global condition on the solution set Z isn't appropriate.

A virtue of the kind of localization introduced by Pennanen is that it enables the proximal point algorithm to be applied to problems in optimization and equilibrium modeling in which the local

[2] In [20], the requirement on the c_k values was just that $\limsup_k c_k < \infty$ and $\liminf_k c_k > 0$, but here it will help us to have a limit c_∞. The restriction to $c_k \geq 1$ harmlessly simplifies some considerations.

monotonicity is hidden from sight, yet can be "elicited." Decomposition schemes presented in [24] in 2019 especially underscore this and strongly motivate further efforts.

Here we develop alternative guarantees for linear and superlinear convergence which avoid the globality of Luque's condition, but depend on the sequence $\{z^k\}$ generated by the algorithm converging strongly to a solution point in $\bar{z} \in Z$, not just weakly. To support this, we assume throughout that[3]

$$\text{the Hilbert space } \mathcal{H} \text{ is finite-dimensional.}$$

In other improvements, we supply missing details on carrying out the localized iterations. We show how the focus of the convergence rate analysis can be returned to the distance of the points z^k from their eventual limit $\bar{z}$, instead of their distance from Z. We then specialize our conditions to problems of minimization.

The criterion for z^{k+1} being a good enough appoximation to $P_k(z^k)$ in the iterations (1.3) will use error parameters

$$\varepsilon_k \in (0,1) \;\text{ with }\; \sum\nolimits_{k=0}^{\infty} \varepsilon_k = \sigma < \infty. \tag{1.5}$$

Three levels of increasing tightness can be invoked:

$$||z^{k+1} - P_k(z^k)|| \le \begin{cases} \text{(a)} & \varepsilon_k \\ \text{(b)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||\} \\ \text{(c)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||^2\} \end{cases} \tag{1.6}$$

or alternatively

$$c_k \operatorname{dist}(0, S_k(z^{k+1})) \le \begin{cases} \text{(a)} & \varepsilon_k \\ \text{(b)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||\} \\ \text{(c)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||^2\} \end{cases} \tag{1.6$'$}$$

through the estimate in [20, Proposition 3] that

$$||z^{k+1} - P_k(z^k)|| \le c_k \operatorname{dist}(0, S_k(z^{k+1})) \;\text{ for }\; S_k(z) = T(z) + c_k^{-1}(z - z^k). \tag{1.7}$$

All are satisfied of course in exact execution of the algorithm, where $z^{k+1} = P_k(z^k)$.

Pennanen in [14] allowed for additional inexactness by incorporating the relaxations of Eckstein and Bertsekas [8], which replace $P_k(z^k)$, equaling $z^k - Q_k(z^k)$ in (1.3), by $z^k - \theta_k Q^k(z^k)$ for stepsizes $\theta_k \in (0,2)$. In his convergence analysis, aimed at the singleton case of Z as in [20] and limiting the relaxations to $\theta_k \in [1,2)$, he found that with $\theta_k \to \theta_\infty$ the best linear rate would only be achieved when $\theta_\infty = 1$. This suggests that little advantage can be gained in linear convergence from this feature, and so we keep here to the original iterations in (1.3). The applications of the proximal point algorithm in [24], which particularly motivate us, seem anyway unable to work with Eckstein-Bertsekas relaxations.

Note that if at some stage actually $z^k \in Z$, as signaled by having $P_k(z^k) = z^k$ or equivalently $Q_k(z^k) = 0$, the procedure can come to a halt and under (1.6bc) or (1.6$'$bc) automatically does. This is the case of *finite termination*, which is welcome but doesn't demand further attention. Therefore, *in our asymptotic convergence analysis, we will always take for granted that* $z^k \notin Z$ *for all* k.

Embracing localization of the kind pioneered by Pennanen [14], we suppose

$$Z \cap \mathcal{Z} \neq \emptyset \;\text{ and }\; 0 \in \mathcal{W} \;\text{ for open convex sets } \mathcal{Z} \text{ and } \mathcal{W} \tag{1.8}$$

[3] For the sake of applications such as in stochastic optimization, where inner products based on expectations are essential, we keep to the Hilbert space context, rather than reverting to $I\!R^n$ with its canonical Euclidean norm.

and henceforth *only require the monotonicity in (1.2) and its maximality to hold relative to the elements of* $\mathcal{Z}\times\mathcal{W}$. It will emerge that, when the algorithm is initiated at a point $z^0\in\mathcal{Z}$ near enough to Z, the iterations in (1.3) will be feasible and never draw on aspects of $\operatorname{gph} T$ from outside of $\mathcal{Z}\times\mathcal{W}$.

Such localization raises an issue that must be settled before we continue. Although maximal monotone mappings T and their inverses T^{-1} are closed-convex-valued, that's under the global definition of maximal monotonicity. What do we know about the closedness and convexity of $T^{-1}(0)$ merely under the assumption that T is maximal monotone relative to the open convex set $\mathcal{Z}\times\mathcal{W}$ in (1.8)? The answer is that $Z=T^{-1}(0)$ *is closed convex relative to* $\mathcal{Z}$, i.e.,

$$Z\cap\mathcal{Z}\ \text{ is convex and agrees with }\ \mathcal{Z}\cap\operatorname{cl}[Z\cap\mathcal{Z}]. \tag{1.9}$$

The reason is that any monotone mapping can be extended to one that is maximal monotone globally, cf. [26, 12.6]. In application to the monotone mapping T' with $\operatorname{gph} T'=[\mathcal{Z}\times\mathcal{W}]\cap\operatorname{gph} T$, which is already maximal monotone relative to $\mathcal{Z}\times\mathcal{W}$, this tells us that

$$\exists\text{ maximal monotone }\overline{T}\text{ such that }\ [\mathcal{Z}\times\mathcal{W}]\cap\operatorname{gph}\overline{T}=[\mathcal{Z}\times\mathcal{W}]\cap\operatorname{gph} T. \tag{1.10}$$

That fact was the key to the results on localizing the proximal point algorithm in Pennanen [14] and also in our recent papers [23] and [24].

In the foundational results of [20], the stopping criterion (1.6a) was enough to ensure convergence of the sequence of points z^k generated by the proximal point algorithm to some particular solution $\bar z\in Z$. For a linear rate of convergence the stopping criteria (1.6b) was combined with assuming a "one-sided property of Lipschitz continuity" of T^{-1} at 0 (nowadays instead called the property of *calmness* [26]), namely

$$\exists\,\delta>0,\ a\geq 0,\ \text{ such that }\ \ z\in T^{-1}(w),\ ||w||<\delta\ \implies\ ||z-\bar z||\leq a||w||. \tag{1.11}$$

Then the solution set Z consists just of $\bar z$. Luque [11] showed in 1984, though, that linear convergence can be obtained for the global version of the algorithm without Z being a singleton, and under the same stopping criterion (1.6b), provided that the condition in (1.11) is modified to[4]

$$\exists\,\delta>0,\ a\geq 0,\ \text{ such that }\ \ z\in T^{-1}(w),\ ||w||<\delta\ \implies\ \operatorname{dist}(z,Z)\leq a||w||. \tag{1.12}$$

His linear convergence to a solution $\bar z\in Z$ was not that of $||z^k-\bar z||$ to 0, but that of $\operatorname{dist}(z^k,Z)$ to 0. Moreover, in formulating his condition with localization in w but not in z, he triggered an unsuspected limitation. The convex set $\operatorname{cl}[\operatorname{dom} T^{-1}]$ must have a property at 0 akin to polyhedrality.[5] An appeal to the finite-dimensionality of $\mathcal{H}$ will allow us to avoid this limitation simply by modifying (1.12) to

$$\exists\,\delta>0,\ a\geq 0,\ \text{ such that }\ \ z\in T^{-1}(w),\ ||w||<\delta,\ ||z-\bar z||<\delta\ \implies\ \operatorname{dist}(z,Z)\leq a||w||. \tag{1.13}$$

In utilizing the stopping criterion (1.6c), however, we will be able to go much further than this, even recovering linear convergence of $||z^k-\bar z||$ to 0, while refining (1.13) sharply in our localized setting

[4]Luque's formulation of his result asked to have $c_k\nearrow c_\infty$, but his proof never made use of having $c_{k+1}\geq c_k$ except for getting the existence of c_∞ as a limit. The condition on c_k in (1.3) is therefore adequate.

[5]In (1.12), in contrast to (1.11), the sets $T^{-1}(w)$ may be unbounded. Since the recession cone for $T^{-1}(w)$ is the normal cone to $\operatorname{cl}[\operatorname{dom} T^{-1}]$ at w [26, 12.37], it may vary with w in its direction. That would be incompatible with the globally enforced bound in (1.11), unless the normal cones to $\operatorname{cl}[\operatorname{dom} T^{-1}]$ are locally around 0 all contained in the normal cone at 0. Such normal cone behavior can be counted on for a polyhedral convex set, but otherwise only occurs in unusual situations. The same trouble comes up with the closely similar condition of "upper Lipschitz continuity" of T^{-1} at 0 that was furnished by Robinson [17] in 1999 as supporting linear convergence when T is a subgradient mapping. Robinson did not refer to Luque's earlier work.

with $\bar{z} \in Z \cap \mathcal{Z}$. The refinement centers on the normal cone $N_Z(\bar{z})$ to Z at $\bar{z}$ (this being the same as the normal cone to the convex set $Z \cap \mathcal{Z}$ at $\bar{z}$ because $\bar{z} \in \mathcal{Z}$):

$$\begin{array}{l}\exists\, \varepsilon > 0,\ \delta > 0,\ a \geq 0, \text{ such that} \\ \left.\begin{array}{c} z \in T^{-1}(w),\ \ 0 < |w| < \delta,\ \ 0 < ||z - \bar{z}|| < \delta, \\ \operatorname{dist}\Big(\dfrac{w}{||w||}, N_Z(\bar{z})\Big) < \varepsilon,\ \operatorname{dist}\Big(\dfrac{z - \bar{z}}{||z - \bar{z}||}, N_Z(\bar{z})\Big) < \varepsilon \end{array}\right\} \implies \operatorname{dist}(z, Z) \leq a||w||.\end{array} \tag{1.14}$$

Note here that, when $||z - \bar{z}|| < \delta$ for δ small enough, the distance of z from Z is the same as its distance from $Z \cap \mathcal{Z}$. The restriction in (1.14) to $||w|| > 0$ and $||z - \bar{z}|| > 0$ is needed to make sense of $w/||w||$ and $(z - \bar{z})/||z - \bar{z}||$, but is inconsequential anyway. Having $z \in T^{-1}(w)$ when $w = 0$ results in $\operatorname{dist}(z, Z) = 0$ on the right.

The secret to this refinement is establishing that, under (1.6c), the convergence of the proximal point algorithm to a particular solution $\bar{z}$ has a *directional* property. The angle of approach of z^k to $\bar{z}$ must eventually come close to aligning with the normal cone in question, and the same also for the angles of the vectors $z^k - z^{k+1}$. Until now, this has escaped notice, but it has powerful implications for analyzing convergence in various applications of the algorithm. Of course, when $\bar{z}$ is the unique element of $Z \cap \mathcal{Z}$, the normal cone $N_Z(\bar{z})$ is the whole space $\mathcal{H}$. Then (1.14) reverts to the original condition of [20] in (1.11).

Although the directionality innovation in (1.14) comes with the price that the stopping criterion (1.6c) must be deployed instead of (1.6b), that only would be called for in the final stretch as z^k gets near to $\bar{z}$. On the other hand, (1.14) also offers something intriguingly different for theoretical developments in this subject. We will be able to recast it into a "pointwise" condition in terms of the graphical derivative of the mapping T relative to the pair $(\bar{z}, 0) \in \operatorname{gph} T$. In specialization to subgradient mappings, it will come out as a quadratic growth condition just over a normal cone.

In the end, we will also undertake improvements in another direction, where the mappings $c_k T$ in (1.3) are replaced by mappings T_k enriched by other algorithmic parameters than just c_k. One important incentive for this comes out of applications in which the space $\mathcal{H}$ is a product $\mathcal{H}_1 \times \cdots \times \mathcal{H}_s$ of Hilbert spaces $\mathcal{H}_j$, with $z = (z_1, \ldots, z_s)$ for $z_j \in \mathcal{H}_j$, and T has the structure

$$T(z) = (T_1(z_1, \ldots, z_s), \ldots, T_s(z_1, \ldots, z_s)). \tag{1.15}$$

In the proximal point algorithm as formulated in (1.3), $P_k(z^k)$ is obtained by solving the generalized equation

$$0 \in T(z) + c_k^{-1}[z - z^k] \tag{1.16}$$

for z, and in (1.15) with $z = (z_1, \ldots, z_s)$ that means solving

$$0 \in T_j(z_1, \ldots, z_s) + c_k^{-1}[z_j - z_j^k] \ \text{ for } \ j = 1, \ldots, s. \tag{1.17}$$

The different behaviors of the different components z_j may suggest allowing the flexibility of a separate c_{jk} for each j, with these proximal parameters c_{jk} converging to different values $c_{j\infty} \in (0, \infty]$, some faster and some slower, depending on numerical insights. A particular instance of this can be envisioned in applications of the proximal point algorithm to calculating saddle points in convex optimization [20, Section 5], which is the origin of the proximal method of multipliers [21]. Then z is divided into a primal component and a dual component, and it ought to be possible to compute without the primal and dual proximal terms moving in lockstep.

But with the door open to different c_{jk} parameters in (1.17), other ideas enter with their own attractions. The generalized equation (1.16) might be replaced by $0 \in C_k^{-1}[z - z^k] + T(z)$ for positive-definite linear transformations C_k. Then (1.17) would be just a special case — except for awkwardness over not covering situations in which some of the parameters c_{jk} might have limits in k that are infinite, as an echo of the original version allowing $c_k \to \infty$.

Out of this we adopt a compromise in which (1.16) is generalized to $0 \in T(z) + c_k^{-1} B_k[z - z^k]$ for linear transformations $B_k : \mathcal{H} \to \mathcal{H}$. The mappings $P_k = (I + c_k T)^{-1}$ are replaced then by $P'_k = (I + c_k B_k^{-1} T)^{-1}$, so that z^{k+1} is to be obtained as an approximate solution to $0 \in T(z) + c_k^{-1} B_k[z - z^k]$ instead of (1.16). An obvious possibility would be to have B_k serve in rescaling to enhance the rate of linear convergence. If T is differentiable at a solution $\bar{z}$, for example, B_k could be aimed at approximating the Jacobian $\nabla T(\bar{z})$ in a sort of quasi-Newton manner. We insist, though, on each B_k being *self-adjoint and positive-definite*, so as to give rise to an auxiliary inner product

$$\langle z, w \rangle_{B_k} := \langle B_k z, w \rangle = \langle z, B_k w \rangle \text{ with associated norm } ||z||_{B_k} = \langle z, B_k z \rangle^{1/2}. \tag{1.18}$$

Maximal monotonicity of T, local or global, with respect to the given inner product $\langle \cdot, \cdot \rangle$ of $\mathcal{H}$ carries over to the mapping $T'_k = B_k^{-1} T$ having that property with respect to $\langle \cdot, \cdot \rangle_{B_k}$. From this perspective, the modified iterations constitute a *variable metric* version of the proximal point algorithm to which existing theory can in some ways still be applied. Of course, because $\mathcal{H}$ is finite-dimensional here, all norms on it are equivalent in the sense of providing the same standards for whether a sequence converges or not. But they provide different standards for *rates* of convergence.

Variable metric versions of the proximal point algorithm are not new and have already been the subject of much research, starting in 1999 with Burke and Qian [3], [4], [5], but eventually in a more complex format. For the general case of the algorithm in determining a zero of a monotone mapping, that higher stage was first reached in the 2008 paper of Parente, Lotito and Solodov [13], although similar advances had been made earlier, at least in part, in the context of convex minimization problems in [2], [6], [9], [16]. There is also a huge, active literature on variable metric *proximal-type* methods, which however does not deal with extensions of the proximal point algorithm itself; see for instance [10], [12], and their references.

Our contribution to the topic differs by not insisting on global monotonicity, thereby making the method available for application to nonconvex problems where monotonicity can merely be elicited locally around a solution. We show how our localized results on rates and directions of convergence can be replicated in this context with adaptations of the inexactness rules. It remains, of course, that these results only uncover how the algorithm behaves when initiated close enough to the solution set. The construction of a globally convergent algorithm that first finds a way to get close enough to the solution set is still a major challenge that isn't addressed here.

2 Central results

The original version of the proximal point algorithm, for mappings T that are maximal monotone globally, will also benefit from some of the convergence results we are about to obtain, but our particular challenge is the localized version. For that, the starting point z^0 must be close enough to the solution set $Z = T^{-1}(0)$ in a region $\mathcal{Z} \times \mathcal{W}$ (1.8) where maximal monotonicity prevails locally, and the iterations must keep within that region. More about that needs to be pinned down next.

We suppose the initial point z^0 is close enough to Z that, with respect to the σ employed in (1.5),

$$\exists\, \rho > \operatorname{dist}(z^0, Z) + \sigma \text{ such that } (z, w) \in \mathcal{Z} \times \mathcal{W} \text{ if } ||z - z^0|| < 3\rho,\ ||w|| < 2\rho, \tag{2.1}$$

and we add to the tests for admissibility of the choice of z^{k+1} in (1.6) the "trust region" stipulation that

$$||z^{k+1} - z^k|| < \rho. \tag{2.2}$$

This will later be seen to ensure that the procedure only utilizes elements (z,w) of $(\mathcal{Z}\cap\mathcal{W})\cap\operatorname{gph}T$. Observe in (2.1) that

$$\operatorname{dist}(z^0,Z)=|z^0-\bar z^0| \text{ for } \bar z^0=\operatorname{proj}(z^0,Z), \text{ the point of } Z \text{ closest to } z^0. \tag{2.3}$$

That projection is well defined under (2.1) and the convexity in (1.9). Here we don't exclude from (2.1) and (2.2) the possibility that $\rho=\infty$ with $\mathcal{Z}\times\mathcal{W}=\mathcal{H}\times\mathcal{H}$, because that provides a way of keeping the global version of the algorithm covered.

Theorem 2.1 (local convergence of the proximal point algorithm)**.** *In the circumstances of (1.8) for the solution set $Z=T^{-1}(0)$, let T be maximal monotone relative to $\mathcal{Z}\times\mathcal{W}$. Suppose the proximal point algorithm (1.3) is initiated at a point z^0 satisfying (2.1) with the approximations controlled by (1.6a) or (1.6′a) under (1.5), augmented by (2.2), and let $\bar z^0$ be the point of Z in (2.3).*

Then the iterations will be feasible and, without need of any elements of gph T *outside of $\mathcal{Z}\times\mathcal{W}$, will generate a sequence of points z^k that converges to a point*

$$\bar z\in Z\cap\mathcal{Z} \quad \textit{such that} \quad ||\bar z-\bar z^0||<\rho \tag{2.4}$$

while ensuring also that $P_k(z^k)\to\bar z$ and $Q_k(z^k)\to 0$.

Proof. This is so close to [20, Theorem 1] that we can take the proof of that as our platform and concentrate just on what needs updating. The localization will ultimately pass through applying convergence results under global maximal monotonicity to an extension $\overline{T}$ of T as provided by (1.10), so we begin by explaining how things would work if T itself were maximal monotone globally. A simplification is that here we already know $Z\neq\emptyset$ and have in sight one of its elements, namely $\bar z^0$.

The proof of [20, Theorem 1] establishes under global maximal monotonicity that, from any starting point z^0, the iterations (1.3) utilizing (1.6a) generate a sequence $\{z^k\}$ satisfying

$$||z^{k+1}-P_k(z^k)||\le\varepsilon_k \text{ with } (P_k(z^k),c_k^{-1}Q_k(z^k))\in\operatorname{gph}T \tag{2.5}$$

that converges to some $\bar z\in Z$, hence also $P_k(z^k)\to\bar z$ and $Q_k(z^k)\to 0$, since $||z^{k+1}-P_k(z^k)||\to 0$ and $Q_k(z^k)=z^k-P_k(z^k)$. A crucial detail in the convergence argument in [20] is the estimate

$$||z^k-\bar z^0||\le||z^0-\bar z^0||+\sigma_k=\operatorname{dist}(\bar z^0,Z)+\sigma_k<\rho \text{ for } \sigma_k=\varepsilon_0+\varepsilon_1+\cdots+\varepsilon_{k-1}, \tag{2.6}$$

where the upper bound ρ supplied at the end comes from (2.1). In the limit as $z^k\to\bar z$, this yields the inequality in (2.4), which in particular confirms through (2.1) that $\bar z\in\mathcal{Z}$.

The estimate in (2.6) provides information on how much of gph T actually comes under consideration in the iterations. The nonexpansiveness of both P_k and Q_k in (1.4) lets us compare $P_k(z^k)$ and $Q_k(z^k)$ with $P_k(\bar z^0)=\bar z^0$ and $Q_k(\bar z^0)=0$ to see that

$$||P_k(z^k)-\bar z^0||\le||z^k-\bar z^0|| \text{ and } ||Q_k(z^k)||\le||z^k-\bar z^0||. \tag{2.7}$$

Together with (2.6), this tells us that the elements of gph T in (2.5) have

$$||P_k(z^k)-z^0||\le||P_k(z^k)-\bar z^0||+||\bar z^0-z^0||<2\rho, \text{ and } ||c_k^{-1}Q_k(z^k)||<c_k^{-1}\rho\le\rho,$$

since $c_k\ge 1$ in (1.3). Those elements thus always lie in $\mathcal{Z}\times\mathcal{W}$ by (2.1). Next, because $z^{k+1}-z^k=z^{k+1}-[P_k(z^k)+Q_k(z^k)]$, we have through (2.5), (2.6) and (2.7) that

$$\begin{aligned}||z^{k+1}-z^k||&\le||Q_k(z^k)||+||z^{k+1}-P_k(z^k)||\le||z^k-\bar z^0||+\varepsilon_k\\&\le\operatorname{dist}(z^0,Z)+\sigma_k+\varepsilon_k=\operatorname{dist}(z^0,Z)+\sigma_{k+1}<\rho.\end{aligned}$$

Thus, when T is maximal monotone globally, initiation of the algorithm in the circumstances of (2.1) will cause the extra condition in (2.2) on the choice of z^{k+1} to be satisfied automatically. Moreover the "trust region" $\{ z \,|\, ||z - z^k|| < \rho \}$ behind (2.2) will lie entirely in $\mathcal{Z}$, because of (2.6) and having $||z^k - z^0|| \le ||z^k - \bar{z}^0|| + ||\bar{z}^0 - z^0|| < \rho + \operatorname{dist}(z^0, Z) \le 2\rho$.

This confirms that the algorithm will not draw on elements of gph T outside of $\mathcal{Z} \times \mathcal{W}$, at least in working directly with the error level (1.6a). But even the alternative in (1.6′a) in terms of $S_k(z^{k+1})$ meets this test. According to its definition in (1.7), we have

$$S_k(z^{k+1}) = \{ w \,|\, (z^{k+1}, c_k^{-1}(z^k - z^{k+1}) + w) \in \operatorname{gph} T \},$$

so in replacing (1.6a) by (1.6′a) we are concerned with the elements $(z^{k+1}, c_k^{-1}(z^k - z^{k+1}) + w)$ belonging to gph T that have $||w|| < \varepsilon_k$. We already know here that $z^{k+1} \in \mathcal{Z}$, and also that $||z^{k+1} - z^k|| < \rho$. Then $||c_k^{-1}(z^k - z^{k+1}) + w|| < \rho + \varepsilon_k$ since $c_k \ge 1$, and therefore $c_k^{-1}(z^k - z^{k+1}) + w \in \mathcal{W}$ by (2.1).

We have been supposing up to now that T is maximal monotone globally, not just locally. But our observations about the procedure never utilizing more of gph T than lies in $\mathcal{Z} \times \mathcal{W}$, when initiated in tune with (2.1) according to the closeness of z^0 to $\bar{z}$ and the size of σ, indicate that the global property isn't needed for local convergence. The argument could be applied to any extension $\overline{T}$ of T beyond $\mathcal{Z} \times \mathcal{W}$ as in (1.10), and in the end it is only T in its maximal monotone *localization* that matters. □

In what follows, we use the terminology that a sequence of values $\alpha_k > 0$ converges to 0 *at a linear rate bounded by* β if

$$\limsup_{k \to \infty} \frac{\alpha_{k+1}}{\alpha_k} \le \beta.$$

Theorem 2.2 (linear convergence to solution set)**.** *Let the stopping criterion in Theorem 2.1 be strengthened to (1.6b), and suppose the property in (1.13) holds for T at the limit $\bar{z}$ of the sequence of points z^k. Then*

$$\operatorname{dist}(z^k, Z) \to 0 \quad \textit{at a linear rate bounded by} \quad \frac{a}{\sqrt{a^2 + c_\infty^2}}, \tag{2.8}$$

this being superlinear convergence when $c_\infty = \infty$. Moreover, (1.13) is sure to hold at $\bar{z}$ if actually

$$\exists\, \delta > 0,\ a \ge 0, \ \ \textit{such that} \ \ z \in T^{-1}(w),\ ||w|| < \delta,\ ||z - \bar{z}^0|| < \rho \implies \operatorname{dist}(z, Z) \le a||w|| \tag{2.9}$$

Proof. We can build on the proof given by Luque [11] that invoked (1.12) in place of the original (1.11) of [20]. All that's necessary is to demonstrate, in the light of Theorem 2.1 (and our restriction to finite-dimensionality), that his proof merely requires (1.13). The juncture where his (1.12) enters is in ensuring that the pair $(P_k(z^k), c_k^{-1} Q_k(z^k))$ in (2.5) has

$$\operatorname{dist}(P_k(z^k), Z) \le a||c_k^{-1} Q_k(z^k)||.$$

Thus, (1.12) needs only to be invoked when $z = P_k(z^k)$ and $w = c_k^{-1}(z^k - P_k(z^k))$. Because $P_k(z^k) \to \bar{z}$ and $Q_k(z^k) \to 0$ while $c_k^{-1} \le 1$, it's enough to have the localization of (1.12) in (1.13).

Although (1.13) is specific to the limit point $\bar{z}$, the broader condition in (2.9) guarantees through the bound in (2.4) that it will hold regardless of the particular $\bar{z}$ that is reached. □

The next theorem reveals a local characteristic of convergence in the proximal point algorithm that had previously not been detected, but allows linear convergence to Z to be sharpened to linear convergence to $\bar{z}$.

Theorem 2.3 (linear convergence to solution point). *When the stopping criterion in the execution of the proximal point algorithm in Theorem 2.1 is tightened to (1.6c), the convergence of z^k to $\bar{z}$ has the directional property that*

$$\operatorname{dist}\left(\frac{z^k-\bar{z}}{||z^k-\bar{z}||}, N_Z(\bar{z})\right) \to 0, \qquad \operatorname{dist}\left(\frac{z^k-z^{k+1}}{||z^k-z^{k+1}||}, N_Z(\bar{z})\right) \to 0. \tag{2.10}$$

Then the linear convergence (2.8) in Theorem 2.2 is obtained without assuming (1.13), but just (1.14), and moreover entails

$$||z^k-\bar{z}|| \to 0 \;\; \text{at a linear rate bounded by} \;\; \frac{a}{\sqrt{a^2+c_\infty^2}}, \tag{2.11}$$

this being superlinear convergence when $c_\infty = \infty$.

Proof. Denoting by $I\!B_\rho(z)$ the closed unit ball of radius ρ centered at a point z. Let $Z_\rho = Z \cap I\!B_\rho(\bar{z})$ for the ρ in (2.1), which ensures through (2.4) in Theorem 2.1 that $Z_\rho \subset \mathcal{Z}$. Then Z_ρ is a compact convex set having at $\bar{z}$ the same normal cone $N_Z(\bar{z})$. The function

$$h_\rho(w) = \sup\{\, \langle w, z-\bar{z}\rangle \mid z \in Z_\rho\} \tag{2.12}$$

is positively homogeneous, convex and finite (hence also continuous), and moreover $h_\rho(w) \geq 0$ for all w, with equality if and only if $w \in N_Z(\bar{z})$. The directional property in (2.10) can therefore be expressed in terms of h_ρ by

$$h_\rho\left(\frac{z^k-\bar{z}}{||z^k-\bar{z}||}\right) \to 0, \qquad h_\rho\left(\frac{z^k-z^{k+1}}{||z^k-z^{k+1}||}\right) \to 0, \tag{2.13}$$

and this is what our analysis will aim at confirming. The localization established in Theorem 2.1 allows us, for simplicity of exposition, to assume with the support of (1.10) that T is maximal monotone globally, since nothing outside of $\mathcal{Z} \times \mathcal{W}$ enters the current picture anyway.

Consider now first the case of exact execution of the proximal point algorithm, where $z^{k+1} = P_k(z^k)$ and $z^k - z^{k+1} = Q_k(z^k)$. Because P_k is firmly nonexpansive with the elements of Z_ρ among its fixed points, we have

$$||z^{k+1}-z|| < ||z^k - z|| \;\; \text{for all} \;\; z \in Z_\rho. \tag{2.14}$$

In expanding $||z^k - z||^2 = ||z^{k+1}-z) + (z^k - z^{k+1})||^2$ we get from (2.14) that

$$2\langle z^k - z^{k+1}, z^{k+1}-z\rangle + ||z^k - z^{k+1}||^2 = ||z^k - z||^2 - ||z^{k+1}-z||^2 > 0. \tag{2.15}$$

By then substituting $(z^{k+1}-\bar{z}) - (z-\bar{z})$ for $z^{k+1}-z$ on the left, we arrive at

$$\langle z^k - z^{k+1}, z-\bar{z}\rangle < \langle z^k - z^{k+1}, z^{k+1}-\bar{z}\rangle + \tfrac{1}{2}||z^k - z^{k+1}||^2 = \tfrac{1}{2}||z^k-\bar{z}||^2 - \tfrac{1}{2}||z^{k+1}-\bar{z}||^2$$

for all $z \in Z_\rho$, where the equation on the right comes from specializing (2.15) to $z = \bar{z}$. Therefore

$$h_\rho(z^k - z^{k+1}) \leq \langle z^k - z^{k+1}, z^{k+1}-\bar{z}\rangle + \tfrac{1}{2}||z^k - z^{k+1}||^2 = \tfrac{1}{2}||z^k-\bar{z}||^2 - \tfrac{1}{2}||z^{k+1}-\bar{z}||^2, \tag{2.16}$$

and in particular $h_\rho(z^k - z^{k+1}) \leq ||z^k - z^{k+1}|| \cdot ||z^k - \bar{z}|| + \frac{1}{2}||z^k - z^{k+1}||^2$. In dividing both sides of the latter by $||z^k - z^{k+1}||$ and utilizing the positive homogeneity of h_ρ and the fact that z^k and z^{k+1} tend to $\bar{z}$, we arrive at the second of the limits in (2.13).

Turning toward the first of the limits in (2.13), we observe through (2.16), in terms of having $z^j - z^i = \sum_{k=j}^{i-1}(z^k - z^{k+1})$ when $i-1 > j$, that

$$h_\rho(z^j - z^i) \le \sum_{k=j}^{i-1} h_\rho(z^k - z^{k+1}) < \tfrac{1}{2}\sum_{k=j}^{i-1}\Big(||z^k - \bar{z}||^2 - ||z^{k+1} - \bar{z}||^2\Big) = \tfrac{1}{2}||z^j - \bar{z}||^2 - \tfrac{1}{2}||z^i - \bar{z}||^2.$$

Passing to the limit as $i \to \infty$, we get $h_\rho(z^j - \bar{z}) \le \frac{1}{2}||z^j - \bar{z}||^2$. Dividing by $||z^j - \bar{z}||$ we see from $z^j \to \bar{z}$ that the first of the limits in (2.13) is likewise correct.

So far, we have been focused on exact execution of the proximal point algorithm, but the only property guaranteed by exactness that we've actually made use of is (2.14). Therefore, in order to prove that the limits in (2.13) are valid even with inexact execution under the stopping criterion (1.6c), we only need to show that (2.14) continues to hold at that reduced level of inexactness, at least for k sufficiently high.

We begin by noting that $||z^k - z^{k+1}|| = ||P_k(z^k) + Q_k(z^k) - z^{k+1})| \le ||Q_k(z^k)|| + ||z^{k+1} - P_k(z^k)|| \le ||Q_k(z^k)|| + \varepsilon_k ||z^k - z^{k+1}||^2$ under (1.6c), so that

$$||Q_k(z^k)|| \ge \delta_k(1 - \varepsilon_k\delta_k) \text{ in the case of } \delta_k = ||z^k - z^{k+1}||. \tag{2.17}$$

Another thing we know, because $P_k(z^k) \to \bar{z}$, is that

$$||P_k(z^k) - z|| < \rho \text{ for all } z \in Z_\rho \text{ when } k \text{ is high enough.} \tag{2.18}$$

For any $z \in Z_\rho$ we have $||z^k - z||^2 > ||P_k(z^k) - z||^2 + ||Q_k(z^k)||^2$, but on the other hand from $z^{k+1} - z = (P_k(z^k) - z) + (z^{k+1} - P_k(z^k))$ that

$$||z^{k+1} - z||^2 = ||P_k(z^k) - z||^2 + 2\langle P_k(z^k) - z, z^{k+1} - P_k(z^k)\rangle + ||z^{k+1} - P_k(z^k)||^2,$$

and consequently, with the criterion (1.6c) again coming into play along with (2.17) and (2.18),

$$\begin{aligned} ||z^k - z||^2 - ||z^{k+1} - z||^2 &> ||Q_k(z^k)||^2 - 2||P_k(z^k) - z||\cdot||z^{k+1} - P_k(z^k)|| - ||z^{k+1} - P_k(z^k)||^2 \\ &\ge \delta_k^2(1 - \varepsilon_k\delta_k)^2 - 2\rho\varepsilon_k\delta_k^2 - \varepsilon_k^2\delta_k^4 = \delta_k^2[(1 - \varepsilon_k\delta_k)^2 - 2\rho\varepsilon_k - \varepsilon_k^2\delta_k^2]. \end{aligned}$$

This reveals that our wish for (2.14) will be granted if, for high k,

$$0 < (1 - \varepsilon_k\delta_k)^2 - 2\rho\varepsilon_k - \varepsilon_k^2\delta_k^2 = 1 - 2\varepsilon_k\delta_k - 2\rho\varepsilon_k.$$

That does hold, inasmuch as $\varepsilon_k \to 0$ and $\delta_k \to 0$.

With (2.10) thereby secured under the stopping criterion (1.6c), it's obvious that the condition (1.14) assumed in Theorem 2.2 is no longer fully needed. The reduced property in (1.15) is enough.

All that remains is demonstrating that linear convergence in the sense of (2.8) is equivalent in these circumstances to linear convergnece in the sense of (2.11). That can be accomplished by proving that the ratio of $\text{dist}(z^k, Z)$ to $||z^k - \bar{z}||$ approaches 1 as $k \to \infty$. Because all is local, we can again, for convenience based on (1.10), imagine that T is maximal monotone globally and thus avoid grappling with a truncation of $\text{gph}\, T$.

Express z^k as $\bar{z} + \tau_k\zeta^k$ with $\tau_k = ||z^k - \bar{z}||$ and $||\zeta^k|| = 1$. Denoting the closed convex cone $N_Z(\bar{z})$ just by N for simplicity, let $\bar{\zeta}^k = \text{proj}(\zeta^k, N)$, so that

$$\text{dist}(\zeta^k, N) = ||\zeta^k - \bar{\zeta}^k|| \to 0. \tag{2.19}$$

Let $\bar{z}^k = \text{proj}(z^k, Z)$ and $\bar{\tau}_k = \text{dist}(z^k, Z) = ||z^k - \bar{z}^k||$. Then $\bar{\tau}_k/\tau_k$ is the ratio we want to approach 1. Because $\bar{\zeta}^k \in N$, the nearest point of Z to $\bar{z} + \tau_k \bar{\zeta}^k$ is $\bar{z}$. Hence

$$\begin{aligned} ||\bar{z}^k - \bar{z}|| &= ||\,\text{proj}(\bar{z} + \tau_k \zeta^k, Z) - \text{proj}(\bar{z} + \tau_k \bar{\zeta}^k, Z)|| \\ &\leq ||(\bar{z} + \tau_k \zeta^k) - (\bar{z} + \tau_k \bar{\zeta}^k)|| = \tau_k ||\zeta^k - \bar{\zeta}^k|| = \tau_k \,\text{dist}(\zeta_k, N), \end{aligned}$$

and then in consequence of (2.19),

$$\eta_k \to 0 \;\text{ for the vectors }\; \eta_k = (\bar{z}^k - \bar{z})/\tau_k. \tag{2.20}$$

Next write $z^k - \bar{z}^k$ as $(z^k - \bar{z}) - (\bar{z}^k - \bar{z}) = \tau_k \zeta^k - \tau_k \eta^k$ to identify $\bar{\tau}_k = ||z^k - \bar{z}||$ with $\tau_k ||\zeta^k - \eta^k||$. The ratio $\bar{\tau}_k/\tau_k$ comes out then as $||\zeta^k - \eta^k||$, where $||\zeta^k|| = 1$ by choice. In view of (2.20), it clear then that $\bar{\tau}_k/\tau_k \to 1$, as we had to show. □

It needs to be kept in mind that the directionality in Theorem 2.3 only has an influence when the localized solution set $Z \cap \mathcal{Z}$ isn't a singeton. In the singleton case, the normal cone $N_Z(\bar{z})$ is the entire Hilbert space $\mathcal{H}$. But there may be intermediate grades of the phenomenon. If $Z \cap \mathcal{Z}$ isn't full-dimensional and $\bar{z}$ belongs to its relative interior, $N_Z(\bar{z})$ will be the product of the subspace M orthogonal to the affine hull of $Z \cap \mathcal{Z}$ and a pointed convex cone in its complement $M^\perp$. Then the points z^k will be comprised of a component in M and a component in $M^\perp$. The directional limits in (2.10) will only act in that case to stablize the second components and have no effect on the first.

It could be wondered whether, regardless of the nature of the normal cone, the difference between the unit vectors $(z^k - \bar{z})/||z^k - \bar{z}||$ and $(z^k - z^{k+1})/||z^k - z^{k+1}||$ might anyway always tend to 0. No, this is dispelled by simple examples. Consider for instance, in the case of $\mathcal{H} = I\!R^2$ under the usual Euclidean norm, the (maximal) monotone linear mapping $T : (z_1, z_2) \to (-z_2, z_1)$, for which the solution set Z is the singleton $\{(0,0)\}$. Take $c_k \equiv 1$, so that $P_k \equiv P := (I + T)^{-1}$ and $Q_k \equiv Q := I - P$ with

$$P : (z_1, z_2) \to \tfrac{1}{2}(z_1 - z_2, z_1 + z_2), \qquad Q : (z_1, z_2) \to \tfrac{1}{2}(z_1 + z_2, -z_1 + z_2).$$

The algorithm in exact execution has $z^{k+1} = P(z^k)$ and $z^k - z^{k+1} = Q(z^k)$, and the sequence must converge to $\bar{z} = (0,0)$. The question therefore is whether the angle between $P(z^k)$ and $Q(z^k)$ will tend to 0, but in fact $P(z^k) \perp Q(z^k)$ always.

The next result will tie the property in (1.14) to a kind of generalized directional differentiation of T through a change of variables. Expressing (z, w) as $(\bar{z}, 0) + \tau(\zeta, \omega)$, we can translate having $z \in T^{-1}(w)$ into having $\tau\omega \in T(\bar{z} + \tau\zeta)$. A key issue then is what can happen to the set of such possible pairs (ζ, ω) as $\tau \searrow 0$. For that we appeal to the notion of the *graphical derivative* of the set-valued mapping T at a point $\bar{z}$ in its domain $\text{dom}\, T$ with respect to an element $\bar{w} \in T(\bar{z})$. This is the mapping $DT(\bar{z}\,|\,\bar{w})$ having as its graph the tangent cone to $\text{gph}\, T$ at $(\bar{z}, \bar{w})$ [26]. In limits,

$$\omega \in DT(\bar{z}\,|\,\bar{w})(\zeta) \iff \exists \tau_k \searrow 0 \;\text{ and }\; (\zeta^k, \omega^k) \to (\zeta, \omega) \;\text{ with }\; \bar{w} + \tau_k \omega^k \in T(\bar{z} + \tau_k \zeta^k), \tag{2.21}$$

and the graphical derivative mapping $DT(\bar{z}\,|\,\bar{w})$ thus has closed graph and is positively homogeneous,

$$\omega \in DT(\bar{z}\,|\,\bar{w})(\zeta),\; \lambda > 0 \implies \lambda\omega \in DT(\bar{z}\,|\,\bar{w})(\lambda\zeta). \tag{2.22}$$

The directions of the pairs (ζ, ω) in (2.21) are the directions from which a sequence of pairs (z^k, w^k) in $\text{gph}\, T$ can approach $(\bar{z}, \bar{w})$. In connection with (1.15), of course, our interest lies with $\bar{w} = 0$ and limit pairs $(\zeta, \omega) \in N_Z(\bar{z}) \times N_Z(\bar{z})$. Assistence comes from the fact that

$$\omega \in DT(\bar{z}\,|\,0)(\zeta) \implies \omega \in N_Z(\bar{z}), \quad \langle \omega, \zeta \rangle \geq 0. \tag{2.23}$$

This is a consequence of the monotonicity of T in $\mathcal{Z}\times\mathcal{W}$ as applied to having $\bar{w}+\tau_k\omega^k\in T(\bar{z}+\tau_k\zeta^k)$ in (2.21) with $\bar{w}=0$ and $0\in T(z)$ when $z\in Z$. That yields $0\leq\langle\tau_k\omega_k-0,\bar{z}+\tau_k\zeta-z\rangle$ when $z\in Z\cap\mathcal{Z}$, hence $\langle\omega^k,z-\bar{z}\rangle\leq\tau_k\langle\omega^k,\zeta^k\rangle$, and indicates beyond (2.23) that the distance of $\omega^k/||\omega_k||$ from $N_Z(\bar{z})$ must go to 0 (as seen from the early part of the proof of Theorem 2.3 in terms of the function h_ρ).

Theorem 2.4 (graphical derivative criterion)**.** *In terms of the graphical derivative $DT(\bar{z}\,|\,0)$ with $\bar{z}\in Z$, the condition in (1.14) holds if and only if*

$$\zeta\in N_Z(\bar{z}),\;\; 0\in DT(\bar{z}\,|\,0)(\zeta)\quad\Longrightarrow\quad\zeta=0. \tag{2.24}$$

That is equivalent in turn to having

$$\infty>\bar{a}:=\inf\Big\{\,a\geq 0\,\Big|\,||\zeta||\leq a||\omega||\;\;\textit{when}\;\;\zeta\in N_Z(\bar{z}),\;\omega\in DT(\bar{z}\,|\,0)(\zeta)\Big\}. \tag{2.25}$$

More specifically, (1.14) entails $a\geq\bar{a}$, while on the other hand (1.14) is sure to hold for any $a>\bar{a}$ when ε and δ are small enough. Thus, $\bar{a}$ can replace a in (2.11).

Proof. The equivalence between (2.24) and (2.25) can be taken care of first. It is based on $N_Z(\bar{z})$ and the graph of $DT(\bar{z}\,|\,0)$ being closed cones, which allows both conditions to be focused on $||\zeta||=1$. Then (2.24) can be restated as

$$0<\min\Big\{\,||\omega||\,\Big|\,\omega\in DT(\bar{z}\,|\,0)(\zeta)\;\text{ for some }\;\zeta\in N_Z(\bar{z})\;\text{ with }\;||\zeta||=1\Big\}$$

and identified that way obviously with (2.25).

Next we look at the relationship between (2.25) and (1.14). Changing the notation from $z\in T^{-1}(w)$ to $\tau\omega\in T(\bar{z}+\tau\zeta)$ with $\tau>0$ allows (1.14) to be posed as the existence of $\varepsilon>0$, $\delta>0$, $a\geq 0$, such that, when $\tau>0$,

$$\left.\begin{array}{c}\tau\omega\in T(\bar{z}+\tau\zeta),\;\;0<\tau||\omega||<\delta,\;\;0<\tau||\zeta||<\delta\\ \operatorname{dist}\Big(\dfrac{\zeta}{||\zeta||},N_Z(\bar{z}))\Big)<\varepsilon,\;\operatorname{dist}\Big(\dfrac{\omega}{||\omega||},N_Z(\bar{z}))\Big)<\varepsilon\end{array}\right\}\Longrightarrow\;\;||\zeta||\leq a||\omega||.$$

This can be translated into a limit condition, namely the existence of $a\geq 0$ such that

$$\left.\begin{array}{c}\tau_k\omega^k\in T(\bar{z}+\tau_k\zeta^k),\;\;\tau_k\searrow 0,\;\;(\zeta^k,\omega^k)\to(\zeta,\omega)\neq(0,0)\\ \operatorname{dist}\Big(\dfrac{\zeta^k}{||\zeta^k||},N_Z(\bar{z}))\Big)\to 0,\;\operatorname{dist}\Big(\dfrac{\omega^k}{||\omega^k||},N_Z(\bar{z}))\Big)\to 0\end{array}\right\}\Longrightarrow\;\;||\zeta^k||\leq a||\omega^k||\;\text{ for high }k.$$

In comparison, the existence of $a\geq 0$ such that

$$\left.\begin{array}{c}\tau_k\omega^k\in T(\bar{z}+\tau_k\zeta^k),\;\;\tau_k\searrow 0,\;\;(\zeta^k,\omega^k)\to(\zeta,\omega)\neq(0,0)\\ \operatorname{dist}\Big(\dfrac{\zeta^k}{||\zeta^k||},N_Z(\bar{z}))\Big)\to 0\end{array}\right\}\Longrightarrow\;\;||\zeta^k||\leq a||\omega^k||\;\text{ for high }k$$

corresponds to (2.21) with the relationship between a and $\bar{a}$ indicated in the theorem. But the ζ limit in the latter automatically entails the ω limit in the former. This was observed in discussion just ahead of the theorem. □

It might be hoped that the condition (2.24) in Theorem 2.4 would automatically persist to points of Z in an neighborhood of $\bar{z}$, but it may not. An illustration of this will come at the end of the next section, once function versions of the conditions here are available. Still, investigations may uncover classes of functions where there is a form of such persistence, which could produce a guarantee that the point $\bar{z}$, even if not known in advance, will enjoy (1.14).

3 Specialization to subgradient mappings

Our attention focuses now on having T be the subgradient mapping ∂f associated with a lower semicontinuous (lsc) function $f : \mathcal{H} \to (-\infty, \infty]$, $f \not\equiv \infty$, so that

$$Z = T^{-1}(0) = \{\, z \,|\, 0 \in \partial f(z)\}. \tag{3.1}$$

When f is convex, ∂f is maximal monotone globally and $Z = \operatorname{argmin} f$ (the set of points minimizing f over all of $\mathcal{H}$). The proximal point algorithm is well understood in that setting as a method for determining a global minimizer of f. It can be applied indirectly also in solving implicit dual problems concave problems of maximization. Here, though, our interest goes beyond such global optimization to concern also localizations in which the graph of $T = \partial f$ is truncated by intersection with a set $\mathcal{Z} \times \mathcal{W}$. Then subgradients that generalize those of convex analysis have to enter the picture.

As laid out in [26], the inequality $f(z') \geq \langle w, z' - z\rangle + o(||z' - z||)$ describes a *regular* subgradient of f at z, indicated by $w \in \widehat{\partial} f(z)$. General subgradients $w \in \partial f(z)$ are defined in terms of taking limits of regular subgradients:

$$\exists\, w^k \in \widehat{\partial} f(z^k) \;\text{ with }\; w^k \to w,\; z^k \to z,\; f(z^k) \to f(z).$$

For many functions, not just convex, $f(z^k) \to f(z)$ is automatic from the rest in this definition. That's *subdifferential continuity* of f at z for w. Many classes of functions anyway have $\partial f(z) = \widehat{\partial} f(z)$. For them taking limits yields nothing more. (See "subdifferential regularity" in [26].)

With this extension of the subgradient concept, the functions f for which the mapping ∂f is maximal monotone globally are still precisely the convex functions [15]. But what can be said about the maximal monotonicity of ∂f with respect to an open convex set $\mathcal{Z} \times \mathcal{W}$? When $\mathcal{W}$ is the whole space $\mathcal{H}$, that holds if f is convex on $\mathcal{Z}$, but otherwise something less obvious characterizes the situation. We showed in [23] that for such local maximal monotonicity of ∂f it is sufficient that f be *variationally convex* with respect to $\mathcal{Z} \times \mathcal{W}$, and necessary when the subgradients in question are regular, which anyway is what we'll be occupied with in our context of minimization. The property of variational convexity is simpler to explain under regularity, where it means that

$$\begin{array}{c}\text{there exists a proper lsc convex function } \widehat{f} \leq f \text{ on } \mathcal{Z} \text{ such that}\\ [\mathcal{Z} \times \mathcal{W}] \cap \operatorname{gph} \partial \widehat{f} \;=\; [\mathcal{Z} \times \mathcal{W}] \cap \operatorname{gph} \partial f\\ \text{and, for } (z, w) \text{ belonging to this common set, also } \widehat{f}(z) = f(z).\end{array} \tag{3.2}$$

In other words, in localization, the subgradients of f and the function values associated with them can't be distinguished from those coming from a convex function. In particular then, if $z \in \mathcal{Z}$ every $w \in \partial f(z) \cap \mathcal{W}$ will be a regular subgradient: $w \in \widehat{\partial} f(z)$.

Many implications of this property are explained, along with telling examples, in [23] and more recently [25]. The main fact of importance to us here concerns the points $\bar{z} \in \mathcal{Z}$ with $0 \in \partial f(\bar{z})$. Under the variational convexity condition (3.2), f attains its minimum over $\mathcal{Z}$ at $\bar{z}$.

Implementations of the proximal point algorithm in such localized subgradient circumstances as a method of iterative local minimization have already been explored in [23] and further in [24], but only for exact execution. Here we extend the theory to inexact execution,[6] benefiting from details in Theorem 2.1 involving the "trust region" condition introduced in (2.2). When $T = \partial f$, the iteration rule in (1.3) specializes with

$$P_k(z^k) = \{\, z \,|\, 0 \in \partial f^k(z)\}, \;\text{ where }\; f^k(z) = f(z) + \frac{1}{2c_k}||z - z^k||^2. \tag{3.3}$$

[6] The localization results of Pennanen [14] for inexact execution apply also to subgradients, but only to such mappings directly, i.e., without the iterations being identified with steps of local minimization.

For f convex, f^k is strongly convex and $P_k(z^k) = \operatorname{argmin} f^k$, a singleton. That remains important for applying the new results on a linear rate of convergence, but in our localization of ∂f we are interested in tying (3.3) to *local* minimization. Over what set, however? That was left somewhat vague in [23] and [24], with $\mathcal{Z}$ indicated despite it not necessarily being known.

Here, through (2.2), the answer will be a ball of radius ρ around z^k. The iteration (1.3) under (2.2), will thus take the form

$$\text{get } z^{k+1} \text{ from } z^k \text{ by approximately minimizing } f^k(z) \text{ subject to } ||z - z^k|| < \rho. \tag{3.4}$$

The standards for approximation come from adaptations of the stopping criteria (1.6′) to this context. Such adaptations go back to [20] and utilize (1.7), taking advantage of the fact that the set $S_k(z) = T(z) + c_k^{-1}(z - z^k)$ reduces to $\partial f^k(z)$ when $T = \partial f$. The resulting version of (1.6′) is

$$c_k \operatorname{dist}(0, \partial f^k(z^{k+1})) \le \begin{cases} \text{(a)} \quad \varepsilon_k \\ \text{(b)} \quad \varepsilon_k \min\{1, ||z^{k+1} - z^k||\} \\ \text{(c)} \quad \varepsilon_k \min\{1, ||z^{k+1} - z^k||^2\} \end{cases} \tag{3.5}$$

For differentiable f, this test thus concerns the nearness to 0 of the gradient of f^k at z^{k+1}.

Theorem 3.1 (local convergence of the proximal point algorithm in minimization)**.** *For $T = \partial f$ and accordingly Z as in (3.1), let f have the variational convexity property in (3.2) (or simply be convex itself, as the case where $\mathcal{Z} \times \mathcal{W} = \mathcal{H} \times \mathcal{H}$). Suppose the proximal point algorithm in the mode of (3.4), with approximations controlled by (3.5a), is initiated at a point z^0 satisfying (2.1), and let $\bar{z}^0$ be the point of Z in (2.3).*

Then in each iteration there is a unique point giving the exact minimum in (3.4), so an approximate minimizer z^{k+1} fulfilling (3.5a) will always exist (and be able to be determined without any involvement of elements of $\operatorname{gph} \partial f$ outside of $\mathcal{Z} \times \mathcal{W}$). The generated sequence of points z^k will converge to a particular point $\bar{z}$, as in (2.4), which will minimize f over $\mathcal{Z}$ and in particular have $f(z) \ge f(\bar{z})$ when $||z - \bar{z}|| < \rho$. Moreover, the sequence of values $f(z^k)$ will be finite and converge to $f(\bar{z})$.

Proof. Under (3.2), the algorithm acts on f as if it were acting on the convex function $\hat{f}$ instead, and as if the minimization steps were based on the modifications $\hat{f}^k$ of $\hat{f}$ corresponding to f^k in (3.3). Those functions $\hat{f}^k$ are strongly convex on $\mathcal{Z}$, so there is a uniquely determined exact minimizer in each iteration (3.4). The claims are justified that way from Theorem 3.1 through the connections explained above, except for the convergence of $f(z^k)$ to $f(\bar{z})$. For that, consider on the basis of (3.5a) vectors $w^k \in \partial f^k(z^{k+1})$ with $c_k|w^k| \le \varepsilon_k$, hence $w^k \to 0$. The subgradient inequality $f^k(\bar{z}) \ge f^k(z^{k+1}) + \langle w^k, \bar{z} - z^{k+1} \rangle$ gives through (3.3) the estimate that

$$f(z^{k+1}) - f(\bar{z}) \le \frac{1}{2c_k}||\bar{z} - z^k||^2 - \frac{1}{2c_k}||z^{k+1} - z^k||^2 - \langle w^k, \bar{z} - z^{k+1} \rangle,$$

were all the terms on the right tend to 0. Therefore $\limsup_k f(z^{k+1}) \le f(\bar{z})$. Since $f(\bar{z}) = \min f$, this means that $f(z^{k+1}) \to f(\bar{z})$. □

The question needing to be answered next is how the extra properties in (1.13) and (1.14) that have featured in our linear convergence results are manifested in terms of f when $T = \partial f$. In [20] with f convex on $\mathcal{H}$, it was demonstrated that the original property offered for linear convergence in (1.11) corresponded to a quadratic growth condition holding at the unique minimizing point $\bar{z}$. Here, in looking beyond uniqueness, a modified quadratic growth condition will enter for a point $\bar{z} \in Z$, namely

$$\exists a > 0,\ \lambda > 0, \text{ such that } f(z) \ge \mu + \frac{1}{a} \operatorname{dist}^2(z, Z) \text{ when } ||z - \bar{z}|| < \lambda, \tag{3.6}$$

where

$$\mu = \text{ minimum value of } f \text{ over } \mathcal{Z}, \text{ taken on at every } z \in Z \cap \mathcal{Z} \text{ under (3.2).} \tag{3.7}$$

That modified condition (3.6) happens in fact to be *equivalent* to the condition on $T = \partial f$ in (1.13), which in the terminology of [26, Sec. 9I] is the *calmness* of T^{-1} at $\bar{w}$ for $\bar{z}$ and can also be identified with the property of T being *metrically subregular* at $\bar{z}$ for $\bar{w}$ [7, Theorem 3H.3]. But Aragon and Geoffroy, in a comprehensive analysis of metric regularity-type properties of subdifferential mappings in 2008 [1], proved that this property of ∂f is equivalent to (3.6).

Theorem 3.2 (linear convergence in minimization)**.** *For f satisfying the variational convexity condition (3.2) and $\bar{z} \in Z \cap \mathcal{Z}$, the growth condition in (3.6) ensures the property (1.13) that supports the linear rate of convergence in Theorem 2.2. Moreover, in the circumstances of Theorem 3.1 concerning initiation at z^0, (3.6) is guaranteed at $\bar{z}$ under the broader growth condition that*

$$\exists a > 0 \;\textit{ such that }\; f(z) \geq \mu + \frac{1}{a}\operatorname{dist}^2(z, Z) \;\textit{ when }\; ||z - \bar{z}^0|| < \rho. \tag{3.8}$$

Proof. For convex f, the condition (1.13) for $T = \partial f$ is equivalent to (3.6), as explained ahead of the statement of this theorem in citing [1]. But (1.13) is local in the graph of ∂f, so when ∂f is just variationally convex at $\bar{z}$ for 0, it corresponds to the same property holding for the associated function $\widehat{f}$ as in (3.2). Because $\widehat{f} \leq f$ locally, the quadratic growth condition (3.6) on $\widehat{f}$ implies the same for f, with a possible shrinking of the neighborhood. The guarantee under (3.8) then follows accordingly from the estimates in Theorem 2.1. □

For the property in (1.14) that supports the linear convergence in Theorem 2.3, we can work with its characterization in Theorem 2.4. That comes down to a condition on the graphical derivative of the mapping ∂f at $\bar{z}$ with respect to 0 being a subgradient there. Graphical derivatives of subgradient mappings are known however to be closely related to generalized second derivatives of the functions themselves, and that's what we will be able to exploit. The *second-order epi-derivative* of f at $\bar{z}$ and $\bar{w} \in \partial f(\bar{z})$ is the function $d^2 f(\bar{z}\,|\,\bar{w})$ defined by

$$d^2 f(\bar{z}\,|\,\bar{w})(\zeta) = \liminf_{\substack{\zeta' \to \zeta \\ \tau \searrow 0}} \Delta^2_\tau f(\bar{z}\,|\,\bar{w})(\zeta'), \;\text{ where }\; \Delta^2_\tau f(\bar{z}\,|\,\bar{w})(\zeta') = \frac{f(\bar{z} + \tau\zeta') - f(\bar{z}) - \tau\langle \bar{w}, \zeta'\rangle}{\frac{1}{2}\tau^2}. \tag{3.9}$$

In many situations this definition can draw on a further property with respect to any ζ for which $d^2 f(\bar{z}\,|\,\bar{w})(\zeta) < \infty$: for every sequence $\tau_k \searrow 0$ there is a sequence $\zeta^k \to \zeta$ such that $\Delta^2_{\tau_k} f(\bar{z}\,|\,\bar{w})(\zeta^k) \to d^2 f(\bar{z}\,|\,\bar{w})(\zeta)$. Then $\Delta^2_\tau f(\bar{z}\,|\,\bar{w})$ *epi-converges* to $d^2 f(\bar{z}\,|\,\bar{w})$ (the epigraphs of these functions set-converge). All of this is explained in depth in [26, Chapter 13]. The powerful consequence then is a formula for the graphical derivative of ∂f at $\bar{z}$ for $\bar{w}$, namely that it is the subgradient mapping associated with the function $\frac{1}{2} d^2 f(\bar{z}\,|\,\bar{w})$.

A more subtle connection will developed here, however, in not assuming twice epi-differentiability of f, but just utilizing (3.9) by itself. In set-convergence terms, (3.9) says that the epigraph of $d^2 f(\bar{z}\,|\,\bar{w})$ is the outer limit of the epigraphs of the functions $\Delta^2_\tau f(\bar{z}\,|\,\bar{w})$ as $\tau \searrow 0$. The cluster description of outer limits in [26, 4.19] tells us that

$$\operatorname{epi}\left[\tfrac{1}{2} d^2 f(\bar{z}\,|\,\bar{w})\right] = \bigcup \{\operatorname{epi} q \,|\, \exists\, \tau_k \searrow 0 \;\text{ such that }\; \tfrac{1}{2}\Delta^2_{\tau_k} f(\bar{z}\,|\,\bar{w}) \;\text{ epiconverges to }\; q\}, \tag{3.10}$$

and therefore

$$\tfrac{1}{2} d^2 f(\bar{z}\,|\,\bar{w})(\zeta) = \min\{ q(\zeta) \,|\, \exists\, \tau_k \searrow 0 \;\text{ such that }\; \tfrac{1}{2}\Delta^2_{\tau_k} f(\bar{z}\,|\,\bar{w}) \;\text{ epiconverges to }\; q\}. \tag{3.11}$$

In a similar vein, the definition of the graphical derivative $DT(\bar{z}\,|\,\bar{w})$ in (2.18) says in set-convergence terms that $\operatorname{gph} DT(\bar{z}\,|\,\bar{w})$ is the outer limit, as $\tau \searrow 0$, of the graphs of the difference quotient mappings $\Delta_\tau T(\bar{z}\,|\,\bar{w})$ given by

$$\Delta_\tau T(\bar{z}\,|\,\bar{w})(\zeta) = \tau^{-1}[T(\bar{z}+\tau\zeta) - \bar{w}]. \tag{3.12}$$

The cluster description of outer limits in [26, 4.19] lets us pose this in terms of mappings Q that are graphical limits of extracted sequences,

$$\operatorname{gph} DT(\bar{z}\,|\,\bar{w}) = \bigcup \{\, \operatorname{gph} Q \,|\, \exists\, \tau_k \searrow 0 \text{ such that } \Delta_{\tau_k} T(\bar{z}\,|\,\bar{w}) \text{ converges graphically to } Q\}. \tag{3.13}$$

In application to $T = \partial f$ in our context of variational convexity and $\bar{w} = 0$, we will be able to tie (3.13) to (3.11) through $Q = \partial q$.

Theorem 3.3 (second-derivative characterization of the convergence criterion)**.** *For f satisfying the variational convexity condition (3.2) and $\bar{z} \in Z$, the graphical derivative condition (2.21) on $T = \partial f$, which in Theorem 2.4 captures the property (1.14) that supports the convergence in Theorem 2.3, is equivalent in terms of second-order epiderivatives (3.9) to*

$$d^2 f(\bar{z}\,|\,0)(\zeta) > 0 \text{ for all nonzero } \zeta \in N_Z(\bar{z}), \tag{3.14}$$

which holds in particular when

$$d^2 f(\bar{z}\,|\,0)(\zeta) = 0 \text{ only for } \zeta \text{ in the tangent cone } T_Z(\bar{z}). \tag{3.15}$$

The value $\bar{a}$ in Theorem 2.4 then satisfies

$$\bar{a} \le \hat{a} \text{ for } \hat{a} = 1/\min\Big\{\, \tfrac{1}{2} d^2 f(\bar{z}\,|\,0)(\zeta) \,\Big|\, \zeta \in N_Z(\bar{z}),\ ||\zeta|| = 1 \Big\} < \infty, \tag{3.16}$$

where also

$$\hat{a} = \inf\Big\{\, a > 0 \,\Big|\, \exists\, \lambda > 0 \text{ such that } f(z) \ge f(\bar{z}) + \frac{1}{a}||z - \bar{z}||^2 \text{ if } z \in N_Z(\bar{z}),\ ||z-\bar{z}|| < \lambda \Big\}. \tag{3.17}$$

Proof. The variational convexity puts us in a situation where f is locally (primally and dually) indistinguishable from a convex function. Adding to it the indicator of $\operatorname{cl} Z$, if need be, we can just as well suppose that f itself is convex on the whole space $\mathcal{H}$. The subgradient mapping $\partial[\frac{1}{2}\Delta_\tau^2 f(\bar{z}\,|\,0)]$ comes out as $\Delta_\tau \partial f(\bar{z}\,|\,0)$, and this leads to an important connection between (3.13) for $T = \partial f$ and the formulas in (3.10) and (3.11). By Attouch's Theorem [26, 12.35], the graphical convergence in the former identifies with the epiconvergence in the latter: we have $Q = \partial q$.

What does this reveal about the condition in (2.21)? Having $0 \in DT(\bar{z}\,|\,0)(\zeta)$ means having $0 \in Q(\zeta)$ for some Q in (3.13), but that can be identified now with $0 \in \partial q(\zeta)$ for some q as in (3.11). Recall now that f is minimized by $\bar{z}$, so that the functions $\frac{1}{2}\Delta_\tau^2 f(\bar{z}\,|\,0)$, likewise convex, have minimum value 0, attained at $\zeta = 0$. Hence $0 \in \partial q(\zeta)$ if and only if $q(\zeta) = 0$. Altogether then, having $0 \in DT(\bar{z}\,|\,0)$ is equivalent through (3.11) to having $\frac{1}{2} d^2 f(\bar{z}\,|\,0)(\zeta) = 0$. Thus, (2.21) can correctly be identified here with (3.14), which holds in particular under (3.15) because of the rule in convex analysis that the cones $T_Z(\bar{z})$ and $N_Z(\bar{z})$ are polar to each other.

Under (3.14) and lower semicontinuity, the minimum in (3.16) is positive, so $\hat{a}$ is well defined and not ∞. The restriction to $||\zeta|| = 1$ takes advantage of $d^2 f(\bar{z}\,|\,0)$ being positively homogeneous of degree 2; it yields

$$\tfrac{1}{2} d^2 f(\bar{z}\,|\,0)(\zeta) \ge \frac{1}{\hat{a}}||\zeta||^2 \text{ for all } \zeta \in N_Z(\bar{z}), \tag{3.18}$$

and going back to (3.11), the same bound for each q indicated there. Then such a q has for $\omega \in \partial q(\zeta)$ with $\zeta \in N_Z(\bar{z})$ that $0 = q(0) \geq q(\zeta) - \langle \omega, \zeta \rangle$, and consequently $\frac{1}{a}||\zeta||^2 \leq \langle \omega, \zeta \rangle \leq ||\omega||{\cdot}||\zeta||$. Thus, $||\zeta|| \leq a||\omega||$ for $\omega \in \partial q(\zeta)$ when $\zeta \in N_Z(\bar{z})$. But, as seen earlier, $\partial q = Q$ for Q in (3.13) in the case of $T = \partial f$ and $\bar{w} = 0$. Hence, $a > \hat{a}$ implies that $||\zeta|| \leq a||\omega||$ when $\omega \in DT(\bar{z}\,|\,0)(\zeta)$ with $\zeta \in N_Z(\bar{z})$, as we needed to show.

Finally, (3.18) yields the claimed equivalent description of $\hat{a}$ in (3.17) on the basis of the definition (3.9) of $d^2 f(\bar{z}\,|\,0)$. □

Although Theorem 3.3 only addresses circumstances at the limit $\bar{z}$ of the z^k sequence generated by the algorithm, which usually isn't known in advance, it does purvey some broader information along the lines of (3.8) in Theorem 3.2. If we can be sure that the condition in (3.14) holds for every $z \in Z$ with $||z - \bar{z}^0|| < \rho$, we will have it holding at $\bar{z}$, since $||\bar{z} - \bar{z}^0|| < \rho$ by Theorem 2.1. There may be traction in that, because the property can be automatic for some classes of functions. For instance, as seen through (3.15), it is universal for convex functions f that are piecewise linear-quadratic [26].

It might be wondered whether the property (3.14) is "stable" in the sense of holding for z in a neighborhood of $\bar{z}$ in Z when it holds at $\bar{z}$, but no. This is shown by the following example based on Theorem 3.2, which will help also to bring out the difference between the two conditions (1.13) and (1.14) employed in Theorems 2.2 and 2.3. In $I\!R^2$ with the Euclidean norm, let

$$f(z_1, z_2) = \frac{1}{4}\,\mathrm{dist}^4\left((z_1, z_2), I\!B\right) + \frac{1}{2}\max{}^2\{0, z_1 - 1\}, \tag{3.19}$$

where $I\!B = I\!B_1((0,0))$ is the unit disk. This is a differentiable convex function with minimum value 0 attained on $Z = I\!B$. At the point $\bar{z} = (1,0) \in Z$, the normal cone $N_Z(\bar{z})$ is $\{\,(\zeta_1, 0)\,|\,\zeta_1 \geq 0\}$, and f is twice epi-differentiable with $d^2 f(\bar{z}\,|\,0) = \max^2\{0, \zeta_1\}$. We have the condition in Theorem 3.3 fulfilled with $\hat{a} = \frac{1}{2}$, hence (1.14) holding for $a = \frac{1}{2}$. But the quadratic growth condition in Theorem 3.2 fails for all values of a, and indeed condition (1.13) doesn't hold in this case. In fact, at every boundary point z of Z other than $\bar{z}$, one has $d^2 f(\bar{z})(\zeta) = 0$ for all $\zeta \in N_Z(z)$.

4 Generalization to a variable-metric implementation

The task now is extending our convergence results to the variable metric version of the proximal point algorithm that was outlined at the end of Section 1. Instead of (1.3), the iterations will have

$$\begin{aligned}&z^{k+1} \approx P'_k(z^k) \text{ for } P'_k = (I + c_k B_k^{-1} T)^{-1}, \text{ where } 1 \leq c_k \to c_\infty \in (0, \infty] \\ &\text{and the linear mappings } B_k : \mathcal{H} \to \mathcal{H} \text{ are self-adjoint, positive-definite.}\end{aligned} \tag{4.1}$$

Fixed points of the modified mappings P'_k are the same as for P_k:

$$z = P'_k(z) \iff z \in Z = T^{-1}(0), \tag{4.2}$$

Conditions on B_k will need to be joined under which the z^k sequence generated by iterations is sure to tend to such a point $\bar{z}$. However, the eventual goal isn't just getting convergence but enabling effects that the more flexible scheme might bestow on rates of convergence.

The inner products $\langle \cdot, \cdot \rangle_{B_k}$ and norms $|| \cdot ||_{B_k}$ in (1.18) will have a role alongside of the given inner product and norm in $\mathcal{H}$. The key observation is that our assumed monotonicity of T in $\mathcal{Z} \times \mathcal{W}$ translates into such monotonicity of $T'_k = B_k^{-1} T$ in $\mathcal{Z} \times B_k^{-1}\mathcal{W}$ with respect to $\langle \cdot, \cdot \rangle_{B_k}$ instead of $\langle \cdot, \cdot \rangle$:

$$\begin{aligned}&\text{if } w_i \in B_k^{-1} T(z_i) \text{ for i=0,1, so that } B_k w_i \in T(z_i), \text{ then for } \mathcal{W}_k = B_k^{-1}\mathcal{W}, \\ &0 \leq \langle z_1 - z_0, B_k w_1 - B_k w_0 \rangle = \langle z_1 - z_0, w_1 - w_0 \rangle_{B_k} \text{ when } (z_i, w_i) \in \mathcal{Z} \times \mathcal{W}_k.\end{aligned} \tag{4.3}$$

Maximal monotonicity translates in the same way and applies not just to T'_k but also to $I + c_k T'_k = I + c_k B_k^{-1} T$ (inasmuch as I is trivially maximal monotone with respect to $\langle \cdot, \cdot \rangle_{B_k}$ as well as $\langle, \cdot, \cdot \rangle$, even strongly monotone). That leads to P'_k being a single-valued mapping which is firmly nonexpansive with respect to $|| \cdot ||_{B_k}$, i.e.,

$$||P'_k(z_1) - P'_k(z_0)||^2_{B_k} + ||Q'_k(z_1) - Q_k(z_0)||^2_{B_k} \leq ||z_1 - z_0||^2_{B_k} \text{ for } Q'_k = I - P'_k. \tag{4.4}$$

(This just follows from expressing $z_i = P'_k(z_i) + Q'_k(z_i)$ in expanding $||z_1 - z_0||^2_{B_k}$.) Here

$$(P'_k(z), c_k^{-1} B_k Q'_k(z)) \in \operatorname{gph} T, \tag{4.5}$$

and the domain of the firm nonexpansiveness property (4.4) consists therefore of the points $z \in \mathcal{Z}$ such that $c_k^{-1} Q_k(z) \in \mathcal{W}_k$ for the open neighborhood $\mathcal{W}_k$ of 0 in (4.3).

If B_k converges as $k \to \infty$ to some mapping B_∞, likewise positive-definite, then $\mathcal{W}_k \to \mathcal{W}_\infty = B_\infty^{-1} \mathcal{W}$, so that ultimately the domains of firm nonexpansiveness will stabilize and rate-of-convergence properties will be dictated by B_∞ in its relationship to T. The prospect, however, is that we maybe don't have knowledge of B_∞ in advance and instead contemplate creating it by successive modifications from B_k to B_{k+1} as computations proceed. That could be aimed at a sort of Jacobian approximation to T in approaching a solution point, but more simply it could just be matter of separate adjustments in proximal parameters attached to different components of the solution in the case of block-coordinate structure (1.15).

Although convergence of B_k to some B_∞ won't be essential to convergence of z^k, assumptions will be needed that keep these linear mappings and their inverses from blowing up. The passage from B_k to B_{k+1} requires scrutiny from that perspective.

In exact execution, the algorithm with iterations (4.1) would have $z^{k+1} = (P'_k \circ \cdots \circ P'_1 \circ P'_0)(z^0)$. But while each P'_k is nonexpansive in its own way, little can be said in that direction about the product $P'_k \circ \cdots \circ P'_1 \circ P'_0$, because the standard for nonexpansiveness changes from one factor to the next. Even the boundedness of the generated sequence $\{z^k\}$ could then be in question, in contrast with the situation in exact execution of the original proximal point algorithm and its iterations (1.3). To prevent that, we assume

$$||z||_{B_0} \leq \alpha_0 ||z|| \text{ and } ||z||_{B_k} \leq \alpha_k ||z||_{B_{k-1}} \text{ with } \alpha_k \geq 1 \text{ such that } \infty > \beta = \Pi_{k=0}^\infty \alpha_k. \tag{4.6}$$

In a similar context in [13], Parente, Lotito and Solodov introduce bounds that here would come out as having $\alpha_k = 1 + \alpha'_k$ and $\sum_{k=1}^\infty \alpha'_k < \infty$. This is a slightly stricter condition, as can be seen through identifying (4.6) with $\sum_{k=1}^\infty \log(1 + \alpha'_k) < \infty$ and invoking the inequality $\log(1 + \alpha'_k) \leq \alpha'_k$.

As a byproduct, the bounds in (4.6) put a ceiling β over the norms $||B_k||$, since they lead to

$$\beta_k^{-1} ||z||_{B_k} \leq \beta_{k-1}^{-1} ||z||_{B_{k-1}} \leq \cdots \leq \beta_0^{-1} ||z||_{B_0} \leq ||z|| \text{ for } \beta_k = \alpha_k \alpha_{k-1} \cdots \alpha_0 \in [1, \beta], \tag{4.7}$$

hence $\langle z, B_k z \rangle \leq \beta_k^2 ||z||^2$ in particular, so that $||B_k|| \leq \beta_k^2$.[7]

To rein in the inverses B_k^{-1}, we suppose that

$$0 < \gamma_k \leq \gamma \in [1, \infty) \text{ with } \gamma_k^{-1} ||z|| \leq ||z||_{B_k} \text{ or equivalently } ||B_k^{-1}|| \leq \gamma_k^2, \tag{4.8}$$

where the equivalence is seen from writing $\gamma_k^{-1} ||z|| \leq ||z||_{B_k}$ as $\gamma_k^{-2} ||z||^2 \leq \langle z, B_k z \rangle$ and identifying that with γ_k^{-2} being a lower bound to the eigenvalues of B_k. That makes γ_k^2 be an upper bound to the eigenvalues of B_k^{-1}, the max of which gives the matrix norm $||B_k^{-1}||$.

[7] The printed version of the paper mistakenly had β in place of β^2 here.

For inexact iterations, within the indicated domains where the mappings P'_k are nonexpansive with respect to the norms $||\cdot||_{B_k}$, an appropriate substitute for the rules in (1.6) must be chosen. We take it to be

$$||z^{k+1} - P_k(z^k)||_{B_k} \leq \begin{cases} \text{(a)} & \varepsilon_k \\ \text{(b)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||_{B_k}\} \\ \text{(c)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||^2_{B_k}\}, \end{cases} \tag{4.9}$$

with the conditions on ε_k still being those in (1.5). Alternatively, errors can be guided by

$$c_k\gamma_k^2 \,\text{dist}\,(0, S'_k(z^{k+1})) \leq \begin{cases} \text{(a)} & \varepsilon_k \\ \text{(b)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||_{B_k}\} \\ \text{(c)} & \varepsilon_k \min\{1, ||z^{k+1} - z^k||^2_{B_k}\} \end{cases} \quad \text{where } \; S'_k(z) = T(z) + c_k^{-1}B_k[z - z^k], \tag{4.9′}$$

through the estimate

$$||z^{k+1} - P_k(z^k)||_{B_k} \;\leq\; c_k \,\text{dist}_{B_k}(0, B_k^{-1}S'_k(z^{k+1})) \;\leq\; c_k\gamma_k^2 \,\text{dist}\,(0, S'_k(z^{k+1})) \tag{4.10}$$

in which dist_{B_k} gives distances with respect to the norm $||\cdot||_{B_k}$. That estimate comes from the observation that having $w \in S'_k(z^{k+1})$ is equivalent to having $z^k + c_kB_k^{-1}w \in [I + c_kB_k^{-1}T](z^{k+1})$, hence $z^{k+1} = P'_k(z^k + c_kB_k^{-1}w)$. The B_k-nonexpansiveness of P'_k implies that $||z^{k+1} - P'_k(z^k)||_{B_k} \leq ||(z^k + c_kB_k^{-1}w) - z^k||_{B_k} = c_k||B_k^{-1}w||_{B_k}$ and thereby yields the first inequality in (4.10). The second inequality then follows via (4.8) from $||B_k^{-1}w||^2_{B_k} = \langle B_k^{-1}w, B_kB_k^{-1}w\rangle = \langle w, B_k^{-1}w\rangle \leq ||B_k^{-1}||{\cdot}||w||^2$.

To accommodate the extra flexibility in the algorithm, we strengthen our earlier assumption (2.1) on the closeness of z^0 to Z and the size of the errors ε_k in (1.5) to[8]

$$\exists\, \rho > \gamma\beta^3[\text{dist}(z^0, Z) + \sigma] \;\text{ such that }\; (z, w) \in \mathcal{Z} \times \mathcal{W} \;\text{ if }\; ||z - z^0|| < 3\rho,\; ||w|| < 2\rho. \tag{4.11}$$

We adapt the "trust region" condition (2.2) on the admissibility of z^{k+1} to

$$||z^{k+1} - z^k||_{B_k} < \rho. \tag{4.12}$$

Theorem 4.1 (local convergence of the variable metric version)**.** *Suppose the modified algorithm in (4.1) under (4.6)–(4.7) is initiated at a point z^0 satisfying (2.1) with approximations controlled by (4.9a) (as could be guaranteed by (4.9′a)), along with (4.12). Let $\bar{z}^0$ be the point of Z in (2.3).*

Then the iterations will be feasible and, without needing to consider any elements of gph T *outside of $\mathcal{Z} \times \mathcal{W}$, will generate a sequence of points z^k that converges to a point $\bar{z}$ as in (2.4), while also ensuring that $P'_k(z^k) \to \bar{z}$ and $Q'_k(z^k) \to 0$.*

Proof. As in the proof of Theorem 2.1, we can suppose at first that T is maximal monotone globally and later claim that this doesn't matter because the procedure will automatically be limited to $\mathcal{Z}\times\mathcal{W}$, where maximal monotonicity of T prevails locally. Because P'_k is nonexpansive with respect to $||\cdot||_{B_k}$ and has $\bar{z}^0$ as a fixed point, the estimate $||z^{k+1} - \bar{z}^0||_{B_k} \leq ||P'_k(z^k) - \bar{z}^0||_{B_k} + ||z^{k+1} - P'_k(z^k)||_{B_k}$ in combination with the inexactness rule in (4.9a) gives us

$$||z^{k+1} - \bar{z}^0||_{B_k} \leq ||z^k - \bar{z}^0||_{B_k} + \varepsilon_k. \tag{4.13}$$

The challenge is extracting from this a connection from one iteration to the next that shows an overall boundedness of the sequence of points z^k.

[8]The printed version had $\beta * 2$ here instead of β^3

Multiplying (4.13) by the factor $\beta_k^{-1} \in (0,1]$, and invoking the relationship between $||\cdot||_{B_{k+1}}$ and $||\cdot||_{B_k}$ in (4.7), we obtain

$$\beta_{k+1}^{-1}||z^{k+1} - \bar{z}^0||_{B_{k+1}} \leq \beta_k^{-1}||z^k - \bar{z}^0||_{B_k} + \varepsilon_k. \tag{4.14}$$

When applied iteratively starting at $k = 0$, this reveals through (4.7) that

$$\beta_k^{-1}||z^k - \bar{z}^0||_{B_k} \leq \beta_0^{-1}||z^0 - \bar{z}_0||_{B_0} + \sigma_k \leq ||z^0 - \bar{z}^0|| + \sigma_k \text{ for } \sigma_k = \sum\nolimits_{j=0}^{k-1} \varepsilon_j,$$

where $||z^0 - \bar{z}^0|| = \text{dist}(z^0, Z)$. Therefore

$$||z^k - \bar{z}^0||_{B_k} \leq \beta_k[\text{dist}(z^0, Z) + \sigma_k] < \beta[\text{dist}(z^0, Z) + \sigma]. \tag{4.15}$$

Because $||P_k'(z^k) - \bar{z}^0||_{B_k}^2 + ||Q_k'(z^k)||_{B_k}^2 \leq ||z^k - \bar{z}^0||_{B_k}^2$ as the case of the firm nonexpansivity relation (4.4) for $z_1 = z^k$ and $z_0 = \bar{z}^0$, we then actually have

$$\left.\begin{matrix} ||z^k - \bar{z}^0||_{B_k} \\ ||P_k'(z^k) - \bar{z}^0||_{B_k} \\ ||Q_k'(z^k)||_{B_k} \end{matrix}\right\} \leq \beta_k[\text{dist}(z^0, Z) + \sigma_k] < \beta[\text{dist}(z^0, Z) + \sigma]. \tag{4.16}$$

as well as, by (4.8) and (4.11),

$$\left.\begin{matrix} ||z^k - z^0|| \\ ||P_k'(z^k) - z^0|| \end{matrix}\right\} < 2\gamma\beta[\text{dist}(z^0, Z) + \sigma] < 2\beta^{-1}\rho \leq 2\rho. \tag{4.17}$$

Thus, the sequences of points z^k and $P_k'(z^k)$ are bounded and contained in $\mathcal{Z}$.

The points $c_k^{-1}B_kQ_k'(z^k)$ are also of interest in connection with (4.5). Since $c_k \geq 1$, we have for them that $||c_k^{-1}B_kQ_k'(z^k)|| \leq ||B_k||\cdot||Q_k'(z^k)||$, where $||B_k|| \leq \beta_k^2$,[9] but on the other hand, $||Q_k'(z^k)|| \leq \gamma_k||Q_k'(z^k)||_{B_k}$ by (4.8). From (4.16) that furnishes us with[10]

$$||c_k^{-1}B_kQ_k'(z^k)|| \leq \gamma\beta^3[\text{dist}(z^0, z) + \sigma] < \rho, \tag{4.18}$$

so $c_k^{-1}B_kQ_k'(z^k) \in \mathcal{W}$ by (4.11). Thus, the elements of $\text{gph}\, T$ in (4.5) all lie in $\mathcal{Z} \times \mathcal{W}$. A further observation, utilizing (4.9a) and $z^k = P_k'(z^k) + Q_k'(z^k)$, is that

$$||z^{k+1} - z^k||_{B_k} \leq ||Q_k'(z^k)||_{B_k} + ||z^{k+1} - P_k'(z^k)||_{B_k} \leq ||Q_k'(z^k)||_{B_k} + \varepsilon_k,$$

hence in applying (4.16),

$$\begin{aligned} ||z^{k+1} - z^k||_{B_k} \leq \beta_k[\text{dist}(z^0, Z) + \sigma_k] + \varepsilon_k &= \beta_k[\text{dist}(z^0, Z) + \sigma_{k+1}] \\ &< \beta[\text{dist}(z^0, Z) + \sigma] < \rho/\gamma\beta \leq \rho. \end{aligned} \tag{4.19}$$

Thus, the restriction on z^{k+1} in (4.12) won't come into play when T is maximal monotone globally and the algorithm is initiated in the manner specified. Indeed, with z^k belonging to an open ball around z^0 of radius 2ρ by (4.17), the "trust region" for z^{k+1} in (4.12) will, by (4.19), lie within such a ball of radius 3ρ and therefore entirely within $\mathcal{Z}$ in each iteration. The addition of (4.12) to the iteration rule (4.1) with (4.9a), while superfluous under global maximal monotonicity, safeguards the algorithm under local maximal monotonicity with respect to $\mathcal{Z} \times \mathcal{W}$ to operate ain the same way.

[9] The printed version had β_k instead of β_k^2.

[10] The printed version has β^2 instead of β^3.

Substituting (4.9′a) for (4.9a) in this would make it necessary to consider vectors $w \in S'_k(z^{k+1})$ having $||w|| < \varepsilon_k$. Could that involve parts of gph T outside of $\mathcal{Z} \times \mathcal{W}$, which has so far been avoided? No, because $w \in S'_k(z^{k+1})$ corresponds to[11] $(z^{k+1}, c_k^{-1}B_k[z^k - z^{k+1}] + w) \in \operatorname{gph} T$, where we already know $z^{k+1} \in \mathcal{Z}$ and on the other hand have $||c_k^{-1}B_k[z^k - z^{k+1}] + w|| \leq c_k^{-1}||B_k||\cdot||z^k - z^{k+1}|| + ||w||$ with $c_k \geq 1$, $||B_k|| \leq \beta_k$, and $||z^k - z^{k+1}|| \leq \gamma_k||z^{k+1} - z^k||_{B_k}$ by (4.8) but $||z^{k+1} - z^k||_{B_k} < \rho/\beta\gamma$ by (4.19). Therefore, with ρ as in (4.11) and $||w|| < \varepsilon_k$, we have $||c_k^{-1}B_k[z^k - z^{k+1}] + w|| < \rho + \varepsilon_k \leq 2\rho$, which ensures that $c_k^{-1}B_k[z^k - z^{k+1}] + w \in \mathcal{W}$.

Now let Z^∞ denote the nonempty set of all cluster points of the bounded z^k sequence, which by (4.17) and (4.11) lies within $\mathcal{Z}$. To complete the proof of the theorem, we must show that $Z^\infty \subset Z$, that Z^∞ can't contain more than one element $\bar{z}$, and that such $\bar{z}$ satisfies (2.4), i.e., $||\bar{z} - \bar{z}^0|| < \rho$. The latter is easy: any $\bar{z} \in Z$ must by (4.15) and (4.8) satisfy $||\bar{z} - \bar{z}^0|| \leq \gamma\beta[\operatorname{dist}(z^0, z) + \sigma]$ but from (4.11) that is $< \rho$.

It will help in the rest that Z^∞ is also the set of all cluster points of the $P'_k(z^k)$ sequence, because

$$||z^{k+1} - P'_k(z^k)|| \leq \gamma_k||z^{k+1} - P'_k(z^k)||_{B_k} \leq \gamma_k\varepsilon_k \leq \gamma\varepsilon_k \to 0$$

through (4.8) and (4.9a). It will also help that, in deriving (4.13) and then (4.14), all we used about $\bar{z}^0$ was that it was a fixed point of the mappings P'_k, an element of Z. The same would work for any $z \in Z$, yielding $\beta_{k+1}^{-1}||z^{k+1} - z||_{B_{k+1}} \leq \beta_k^{-1}||z^k - z||_{B_k} + \varepsilon_k$. Since $\beta_k \nearrow \beta$ as $k \to \infty$, and $\sum_{k=0}^\infty \varepsilon_k < \infty$, this implies the existence of a finite limit value

$$\mu(z) = \lim_{k\to\infty} ||z^k - z||_{B_k} = \lim_{k\to\infty} ||P'_k(z^k) - z||_{B_k} \text{ for any } z \in Z. \tag{4.20}$$

By posing the firm nonexpansiveness relation (4.4) in the case of $z_0 = z \in Z$ and $z_1 = z^k$ as

$$||Q'_k(z^k)||^2_{B_k} \leq ||z^k - z||^2_{B_k} - ||P'_k(z^k) - z||^2_{B_k}$$

and taking the limit as $k \to \infty$, in which both terms on the right approach $\mu(z)$ by (4.20), we see that $||Q'_k(z^k)||^2_{B_k} \to 0$. But $||Q'_k(z^k)||^2_{B_k} \geq \gamma^{-1}||Q'_k(z^k)||$ by (4.8), so this implies $Q'_k(z^k) \to 0$.

Consider next, along with the elements $(P'_k(z^k), c_k^{-1}B_kQ'_k(z^k))$ of gph T in (4.5), any other pair $(z, w) \in [\mathcal{Z} \cap \mathcal{W}] \cap \operatorname{gph} T$. The monotonicity of T gives us $0 \leq \langle z - P'_k(z^k), w - c_k^{-1}B_kQ'_k(z^k)\rangle$, where $c_k^{-1}B_kQ'_k(z^k) \to 0$ because $Q'_k(z^k) \to 0$, $c_k \geq 1$, and the mappings B_k are known from (4.7) to be bounded in norm. In consequence of this, any $\bar{z} \in Z^\infty$, as a cluster point of the $P'_k(z^k)$ sequence, must satisfy $0 \leq \langle z - \bar{z}, w\rangle$. This being true for arbitrary $(z, w) \in \operatorname{gph} T$ in $\mathcal{Z} \times \mathcal{W}$, the maximality of the monotonicity of T with respect to $\mathcal{Z} \times \mathcal{W}$ allows us to conclude that $(\bar{z}, 0) \in \operatorname{gph} T$. Thus, $Z^\infty \subset Z$, and accordingly from (4.20) we have

$$\mu(\bar{z}) = \lim_{k\to\infty} ||z^k - \bar{z}||_{B_k} \text{ for any } \bar{z} \in Z^\infty. \tag{4.21}$$

As a cluster point of $\{z^k\}$, any $\bar{z} \in Z^\infty$ is the limit of a subsequence of $\{z^k\}$, and $||z^k - \bar{z}||$ then tends to 0 for that subsequence. Since $||z^k - \bar{z}||_{B_k} \leq \gamma||z^k - \bar{z}||$ by (4.7), it follows from (4.21) that $\mu(\bar{z}) = 0$, but then further that $||z^k - \bar{z}|| \to 0$, because $||z^k - \bar{z}||_{B_k} \geq \alpha^{-1}||z^k - \bar{z}||$ by (4.8). In other words $z^k \to \bar{z}$, this therefore being the unique element of Z^∞. □

Our aim now is to gain an understanding of linear convergence of the proximal point algorithm in the variable metric framework by generalizing Theorem 2.2. Clearly this should involve distances $\operatorname{dist}_{B_k}(z, Z)$ with respect to the norm $||\cdot||_{B_k}$ instead of just $\operatorname{dist}(z, Z)$, but the simple idea that $||w||_{B_k}$

[11] The arguments of this pair were reversed in the printed version

should likewise replace $||w||$ in assumption (1.13) turns out to be wrong. The right kind of condition is that

$$\exists\,\delta>0,\ a_k\geq 0\ \text{ such that }\ z\in T^{-1}(w),\ ||z-\bar z||<\delta,\ ||w||<\delta\ \Rightarrow\ \mathrm{dist}_{B_k}(z,Z)\leq a_k||w||_{B_k^{-1}},\quad (4.22)$$

where $||w||_{B_k^{-1}}$ is the norm associated with the inner product in $\mathcal{H}$ induced by B_k^{-1} rather than B_k. We know of course from (4.6) and (4.8) that

$$\gamma^{-1}||z||\leq\gamma_k^{-1}||z||\leq||z||_{B_k}\leq\beta_k||z||\leq\beta||z||,\tag{4.23}$$

with γ_k^{-1} and β_k being lower and upper bounds, respectively, to the eigenvalues of B_k, so that also

$$\beta^{-1}||w||\leq\beta_k^{-1}||w||\leq||w||_{B_k^{-1}}\leq\gamma_k||w||\leq\gamma||w||.\tag{4.24}$$

In fact, the adapted conditions in (4.22) are thereby *equivalent* to the earlier condition (1.13) and each other as *existence* statements. But the progression of norms and their constants a_k may be able to unlock secrets about a rate of linear convergence that would otherwise not come to light.

The reader may wonder why the δ bounds in (4.22) only refer to the original norm and don't vary with k. That's because we'll need a core neighborhood of $(\bar z,0)$ that is stable amid the changing metrics.

Theorem 4.2 (linear convergence of the variable metric version)**.** *Tighten the stopping criterion in Theorem 4.1 to (4.9b) or (4.9′b), and suppose that (4.22) holds for T at the limit point $\bar z$ of the sequence of points z^k, moreover with $B_k\to B_\infty$ and $a_k\to a_\infty<\infty$. Then*

$$\mathrm{dist}_{B_\infty}(z^k,Z)\to 0\ \ \textit{at a linear rate bounded by}\ \ \frac{a_\infty}{\sqrt{a_\infty^2+c_\infty^2}},\tag{4.25}$$

this being superlinear convergence when $c_\infty=\infty$. Moreover, (4.25) is sure to hold at $\bar z$ if actually

$$\begin{aligned}&\exists\,\delta>0,\ a_k\geq 0,\ \ \textit{such that}\\ &\qquad z\in T^{-1}(w),\ ||w||<\delta,\ ||z-\bar z^0||<\rho\implies \mathrm{dist}_{B_k}(z,Z)\leq a_k||w||_{B_k^{-1}}.\end{aligned}\tag{4.26}$$

Proof. The argument of Luque [11] that we relied on in proving Theorem 2.2 will be extended to this new situation. We know from Theorem 4.1 that all action in the iterations takes place within the region $\mathcal{Z}\times\mathcal{W}$ in which T is assumed to be maximal monotone locally and the nearest point of z^k to the convex set $Z\cap\mathcal{Z}$ is well defined and unique, being also the nearest point of z^k to Z. The same must eventually be true then also for the norms $||\cdot||_{B_k}$ by (4.23), so we can pose

$$\mathrm{dist}_{B_k}(z^k,Z)=||z^k-\bar z^k||_{B_k}\ \ \text{for}\ \ \bar z^k=\mathrm{proj}_{B_k}(z^k,Z)\tag{4.27}$$

and similarly, since also $P_k'(z^k)\to\bar z$ by Theorem 4.1, express

$$\mathrm{dist}_{B_k}(P_k'(z^k),Z)=||p^k-\bar z^k||_{B_k}\ \ \text{for}\ \ p^k=\mathrm{proj}_{B_k}(P_k'(z^k),Z),\tag{4.28}$$

as will be needed a bit later. From the firm nonexpansiveness rule in (4.4) in the case of the points $z_1=z^k$ and $z_0=\bar z^k\in Z$, we have $||P_k'(z^k)-\bar z^k||_{B_k}^2+||Q_k'(z^k)||_{B_k}^2\leq||z^k-\bar z^k||_{B_k}^2=\mathrm{dist}_{B_k}^2(z^k,Z)$, where $||P_k'(z^k)-\bar z^k||_{B_k}\geq\mathrm{dist}_{B_k}(P_k'(z^k),Z)$. Therefore

$$||Q_k'(z^k)||_{B_k}^2\leq\mathrm{dist}_{B_k}^2(z^k,Z)-\mathrm{dist}_{B_k}^2(P_k'(z^k),Z).\tag{4.29}$$

On the other hand, from (4.22) we know for the elements in (4.5) that

$$\begin{aligned}\operatorname{dist}^2_{B_k}(P'_k(z^k), Z) &\le a_k^2 ||c_k^{-1} B_k Q'_k(z^k)||^2_{B_k^{-1}} \\ &= a_k^2 c_k^{-2} \langle B_k Q'_k(z^k), B_k^{-1} B_k Q'_k(z^k)\rangle = a_k^2 c_k^{-2} ||Q'_k(z^k)||^2_{B_k}.\end{aligned}$$

Combining this with (4.29) yields $\operatorname{dist}^2_{B_k}(P'_k(z^k), Z) \le a_k^2 c_k^{-2}[\operatorname{dist}^2_{B_k}(z^k, Z) - \operatorname{dist}^2_{B_k}(P'_k(z^k), Z)]$, which can be written as

$$\operatorname{dist}_{B_k}(P'_k(z^k), Z) \le \mu_k \operatorname{dist}_{B_k}(z^k, Z) \quad \text{for} \quad \mu_k = a_k \Big/ \sqrt{a_k^2 + c_k^2}. \tag{4.30}$$

We have to translate (4.30) into a condition on $\operatorname{dist}_{B_k}(z^{k+1}, Z)$ instead of $\operatorname{dist}_{B_k}(P'_k(z^k), Z)$, and this is where the stopping rule (4.9b) will come in. Since $z^{k+1} - z^k \to 0$ and $\varepsilon_k \searrow 0$, we can suppose in (4.9b) that $||z^{k+1} - z^k||_{B_k} < 1$ and $\varepsilon_k < 1$. In terms of (4.28) we then have

$$\operatorname{dist}_{B_k}(z^{k+1}, Z) \le ||z^{k+1} - p^k||_{B_k}. \tag{4.31}$$

Using $||z^{k+1} - z^k||_{B_k} \le ||z^{k+1} - p^k||_{B_k} + ||z^k - p^k||_{B_k}$, we can estimate through (4.9b) that

$$\begin{aligned}||z^{k+1} - p^k||_{B_k} &\le ||z^{k+1} - P'_k(z^k)||_{B_K} + ||P'_k(z^k) - p^k||_{B_k} \\ &\le \varepsilon_k ||z^{k+1} - p^k||_{B_k} + \varepsilon_k ||z^k - p^k||_{B_k} + \operatorname{dist}_{B_k}(P'_k(z^k), Z).\end{aligned} \tag{4.32}$$

Also $||z^k - p^k||_{B_k} \le ||z^k - \bar{z}^k||_{B_k} + ||\bar{z}^k - p^k||_{B_k}$ and

$$||\bar{z}^k - p^k||_{B_k} = ||\operatorname{proj}_{B_k}(z^k, Z) - \operatorname{proj}_{B_k}(P'_k(z^k), Z)||_{B_k} \le ||z^k - P'_k(z^k)||_{B_k} = ||Q'_k(z^k)||_{B_k},$$

where $||Q'_k(z^k)||_{B_k} \le \operatorname{dist}_{B_k}(z^k, Z)$ by (4.29). Through this we get from (4.32) that

$$(1 - \varepsilon_k)||z^{k+1} - p^k||_{B_k} \le (1 + 2\varepsilon_k) \operatorname{dist}_{B_k}(z^k, Z)$$

and then from (4.30) vua (4.31) that

$$\operatorname{dist}_{B_k}(z^{k+1}, Z) \le \mu'_k \operatorname{dist}_{B_k}(z^k, Z) \quad \text{for} \quad \mu'_k = \mu_k \frac{1 + 2\varepsilon_k}{1 - \varepsilon_k}. \tag{4.33}$$

We can now bring in the assumption that $B_k \to B_\infty$, which implies the existence of $\theta_k \ge 1$ such that $\theta_k^{-1} ||z||_{B_\infty} \le ||z||_{B_k} \le \theta_k ||z||_{B_\infty}$ with $\theta_k \to 1$. For (4.33) this says $\operatorname{dist}_{B_k}(z^{k+1}, Z) \ge \theta_k^{-1} \operatorname{dist}(z^{k+1}, Z)$ and $\operatorname{dist}_{B_k}(z^k, Z) \le \theta_k \operatorname{dist}(z^{k+1}, Z)$, so that

$$\operatorname{dist}_{B_\infty}(z^{k+1}, Z) \le \mu''_k \operatorname{dist}_{B_\infty}(z^k, Z) \quad \text{for} \quad \mu''_k = \theta_k^2 \mu_k \frac{1 + 2\varepsilon_k}{1 - \varepsilon_k}.$$

Then in the limit as $a_k \to a_\infty$ and $c_k \to c_\infty$, we have $\mu''_k \to \mu_\infty = a_\infty / \sqrt{a_\infty^2 + c_\infty^2}$. Hence the rate of linear convergence in (4.25) is correct.

Finally, the condition in (4.25) guarantees that (4.22) will be available regardless of the particular limit point $\bar{z}$ of the sequence generated by the algorithm, since that point is known through (2.4) to have $||\bar{z} - \bar{z}^0|| < \rho$. □

Theorem 4.2 can be compared to the linear convergence result of Parente, Lotito and Solodov [13] for their variable metric version of the proximal point algorithm. A key distinction is that their focus was on global convergence under global monotonicity, whereas ours has been on local convergence under local monotonicity and the mathematical maneuvers entailed by that. From the angle of computational methology, their developments are more sophisticated and moreover also allow an additional kind of approximation of the mapping T to be brought in. However, the condition they require is the "upper Lipshitz continuity" of T^{-1} at 0 used by Robinson [17]. That is a global property of the possibly unbounded set $Z = T^{-1}(0)$ which is akin to Luque's condition (1.12) and subject to the same troubles as Luque's, as we noted in Section 1. Here we get by with something weaker.

How does this all play out in application to problems of minimization, in which $T = \partial f$ as in Section 3 with variational convexity coming in as there? Parallel to (3.3), we have the mappings P'_k described in this case by

$$P'_k(z^k) = \{\, z \,|\, 0 \in \partial f^k(z)\}, \;\text{ where }\; f^k(z) = f(z) + \frac{1}{2c_k}||z - z^k||^2_{B_k}. \tag{4.34}$$

Then, because $\partial f^k(z) = \partial f(z) + c_k^{-1}B_k[z - z^k]$, the stopping rules in (4.9') have

$$\mathrm{dist}(0, S'_k(z^{k+1})) = \mathrm{dist}(0, \partial f^k(z)). \tag{4.35}$$

This explains in particular why those rules have been formulated with dist rather than dist_{B_k}. The mode of excecution of the algorithm is now to

$$\text{get } z^{k+1} \text{ from } z^k \text{ by approximately minimizing } f^k(z) \text{ subject to } ||z - z^k|| < \rho, \tag{4.36}$$

and, for that, (4.35) is an appropriate quanitity to monitor.

In this setting, the growth condition that will replace the one in (3.6) for generalizing Theorem 3.3 is as follows:

$$\exists\, a_k \geq 0,\; \lambda > 0, \;\text{ such that }\; f(z) \geq \mu + \frac{1}{a_k}\,\mathrm{dist}^2_{B_k}(z, Z) \;\text{ when }\; ||z - \bar{z}|| < \lambda, \tag{4.37}$$

where μ is the minimum value of f on $\mathcal{Z}$ as in (3.7).

Theorem 4.3 (variable metric growth condition)**.** *In the local minimization version of the proximal point algorithm as extended from Theorems 3.1 and 3.2 in the manner of (4.36), the condition (4.22) posed at $\bar{z}$ for the linear convergence result in Theorem 4.2 will hold under (4.37). Moreover, in the specified circumstances of initiation at z^0, (4.22) is guaranteed at $\bar{z}$ under the broader growth condition that*

$$\exists\, a_k \geq 0,\; \lambda > 0, \;\textit{ such that }\; f(z) \geq \mu + \frac{1}{a_k}\,\mathrm{dist}^2_{B_k}(z, Z) \;\textit{ when }\; ||z - \bar{z}^0|| < \rho. \tag{4.38}$$

Proof. The proof of Theorem 3.2 can be imitated with minor changes. In considering $w \in \partial f(z)$ with z near enough to $\bar{z}$ and $||z|| < \delta$, the projection $\mathrm{proj}_{B_k}(z, Z)$ is taken instead of $\mathrm{proj}(z, Z)$ to give z'. Then $\mathrm{dist}_{B_k}(z, Z) = ||z - z'||_{B_k}$. The argument based on the variational convexity of f that previously showed $a^{-1}||z - z'||^2 \leq \langle w, z - z'\rangle$ shows instead, with (4.37) replacing (3.6), that $a_k^{-1}||z - z'||^2_{B_k} \leq \langle w, z - z'\rangle$. The key difference comes then in recognizing that

$$\langle w, z - z'\rangle = \langle B_k^{-1}w, B_k(z - z')\rangle = \langle B_k^{-1}w, z - z'\rangle_{B_k} \leq ||B_k^{-1}w||_{B_k}{\cdot}||z - z'||_{B_k}$$

and moreover that $||B_k^{-1}w||^2_{B_k} = \langle B_k^{-1}w, B_kB_k^{-1}w\rangle = ||w||^{-1}_{B_k}$. When that is put together with the earlier inequality $a_k^{-1}||z - z'||^2_{B_k} \leq \langle w, z - z'\rangle$, the result is $a_k^{-1}\,\mathrm{dist}^2_{B_k}(z, Z) \leq ||w||_{B_k^{-1}}{\cdot}\,\mathrm{dist}_{B_k}(z, Z)$. This confirms (4.22). $\square$

Our results in Theorems 2.3, 2.4 and 3.3 could be targets also for generalization to a variable metric setting, but we stop short of that. Issues are raised that need more space to be worked out than is available here.

References

[1] Aragon, F.J., Geoffroy, M.F., "Characterization of metric regularity of subdifferential mappings," *J. Convex Analysis* 15 (2008), 365–380.

[2] Bonnans, J.F., Gilbert, J.C., Lemarechal,C., Sagastizabal, C., "A family of variable metric proximal point methods," *Mathematical Programming* 68 (1995), 15–47.

[3] Burke, J.V., Qian, M., "A variable metric proximal point algorithm for monotone operators," *SIAM J. Control Opt.* 37 (1999), 25–44.

[4] Burke, J.V., Qian, M., "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, Semismooth and smoothing Methods*, M. Fukushima and L. Qi (eds.), Kluer Academic Publishers, 1999, 317–334.

[5] Burke, J.V., Qian, M., "On the superlinear convergence of the variable metric proximal point algorithm using Broyden and BFGS matrix secant updating," *Mathematical Programming* 88 (2000), 157–181.

[6] Chen, X., Fukushima, M., "Proximal quasi-Newton methods for nondifferentiable convex optimization," *Mathematical Programming* 85 (1999), 313–334.

[7] Dontchev, A.D., Rockafellar, R.T., *Implicit Functions and Solution Mappings,* Springer Verlag, second edition: 2014.

[8] Ecstein, J., Bertsekas, D.P., "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming* 55 (1992), 293–318.

[9] Lemarechal, C., Sagastizabal, C., "Variable metric bundle methods: from conceptual to implementable forms," *Mathematical Programming* 76 (1997), 393-410.

[10] Liu, H., Mairal, J., Harchaoui, Z., "An inexact variable metric proximal point algorithm for generic quasi-Newton acceleration," preprint 2019.

[11] Luque, F.J., "Asymptotic convergence analysis of the proximal point algorithm," *SIAM J. Control Opt.* 22 (1984), 277–293.

[12] Mordukhovich, B.M., Yuan, X., Zeng, S., Zhang, J., "A globally convergent proximal point Newton-type method in nonsmooth convex optimization," preprint 2020.

[13] Parente, L.A., Lotito, P.A., Solodov, M.V., "A class of inexact variable metric proximal point algorithms," *SIAM J. Optimization* 19 (2008), 240–260.

[14] Pennanen, T., "Local convergence of the proximal point algorithm and multiplier methods without monotonicity." *Mathematics of Operations Research* 27 (2002), 170–191.

[15] Poliquin, R. A., "Subgradient monotonicity and convex functions," *Nonlinear Analysis: Theory, Methods and Applications* 14 (1990), 385–398

[16] Qi, L., Chen, X., "A preconditioning proximal Newton's method for nondifferentiable convex optimization," *Mathematical Programming* 76 (1995), 411–430.

[17] Robinson, S. M., "Linear convergence of epsilon-subgradient descent methods for a class of convex functions," *Mathematical Programming* 86 (1999), 41–50.

[18] Rockafellar, R. T., "The multiplier method of Hestenes and Powell applied to convex programming," *J. Optimization Theorm Appl.* 12 (1973), 555–562.

[19] Rockafellar, R. T., "Augmented Lagrange muliplier functions and duality in nonconvex programming." *SIAM J. Control* 12 (1974), 268–285.

[20] Rockafellar, R. T., "Monotone operators and the proximal point algorithm." *SIAM J. Control Opt.* 14 (1976), 877–898.

[21] Rockafellar, R. T., "Augmented Lagrangians and applications of the proximal point algorithm in convex programming." *Math. of Operations Research* 1 (1976), 97–116.

[22] Rockafellar, R. T., "Progressive decoupling of linkages in monotone variational inequalities and convex optimization," *Proceedings of the Conference on Nonlinear Analysis and Convex Analysis, Chitose, Japan, 2017*, submitted.

[23] Rockafellar, R. T., "Variational convexity and local monotonicity of subgradient mappings," *Vietnam Journal of Mathematics*, *Vietnam J. Math.* 47 (2019), 547–561: dx.doi.org/10.1007/s10013-019-00339-5.

[24] Rockafellar, R. T., "Progressive decoupling of linkages in optimization and variational inequalities with elicitable convexity or monotonicity," *Set-Valued and Variational Analysis* 27 (2019), 863–893: dx.doi.org/10.1007/s1128-018-0496-1.

[25] Rockafellar, R. T., "Augmented Lagrangians and hidden convexity in sufficient conditions for local optimality," submitted to Mathematical Programming, 2020.

[26] Rockafellar, R. T., and Wets, R. J-B, *Variational Analysis*, No. 317 in the series *Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, 1997.