A Generalized Proximal Point Algorithm and its Convergence Rate

Etienne Corman * **and Xiaoming Yuan**†

March 13, 2014

Abstract

We propose a generalized proximal point algorithm (PPA), in the generic setting of finding a root of a maximal monotone operator. In addition to the classical PPA, a number of benchmark operator splitting methods in the PDE and optimization literatures can be retrieved by this generalized PPA scheme. We establish the convergence rate of this generalized PPA scheme under different conditions, including estimating its worst-case convergence rate measured by the iteration complexity under mild assumptions and deriving its linear convergence rate under certain stronger conditions. Throughout our discussion, we pay particular attention to the special case where the operator is the sum of two maximal monotone operators, and specify our theoretical results in generic setting to this special case. Our result turns out to be a general and unified study on the convergence rate of a number of existing methods, and subsumes some existing results in the literature.

Key words: Convex Optimization, Proximal Point Algorithm, Operator Splitting Methods, Convergence Rate

1 Introduction

Let $\mathbb{H}$ be a Hilbert space with the scalar product $\langle\cdot\rangle$ and norm $\|\cdot\|$; $T : \mathbb{H} \to 2^{\mathbb{H}}$ be a set-valued maximal monotone operator. A fundamental mathematical problem is to find a root of T:

$$0 \in T(v). \tag{1}$$

To solve (1), the proximal point algorithm (PPA) tracking back to [10, 23, 24, 30] is a classical scheme. Starting from $v^0 \in \mathbb{H}$, the iterative scheme of PPA reads as

$$0 \in T(v) + \frac{1}{\lambda}(v - v_n), \tag{2}$$

*CMAP, Ecole Polytechnique, 91128 Palaiseau, France. Email: etienne.corman@cmap.polytechnique.fr

†Corresponding author. Department of Mathematics, Hong Kong Baptist University, Hong Kong, P.R. China. Email: xmyuan@hkbu.edu.hk. This author was supported by the General Research Fund from Hong Kong Research Grants Council: 203613.

where $\lambda > 0$ is a proximal parameter. In fact, the PPA plays a significant theoretical and algorithmic role in many areas such as optimization, PDE and image processing; and a number of celebrated algorithms turn out to be specific cases of the PPA when the operator T is specified accordingly. Such examples include the augmented Lagrangian method [20, 28] (see [30]), the Douglas-Rachford splitting method in [7, 22] (see [9]), the split inexact Uzawa method [33], and so on.

Recall (see e.g. [29]) the resolvent operator of a set-valued monotone operator is

$$J_\lambda^T = (I + \lambda T)^{-1}. \tag{3}$$

Then, the PPA scheme for solving (1) can be written as

$$v_{n+1} = J_\lambda^T(v_n). \tag{4}$$

That is, at each iteration it requires an exact evaluation of the resolvent operator J_λ^T [1]. Note that the resolvent operator of a set-valued monotone operator is always single-valued.

The problem (1) is an abstract model in the generic setting, and it can be specified as various concrete forms with special structures for different applications. For example, a representative case is that the operator T in (1) is the sum of two maximal monotone operators A and B. In this case, the problem (1) becomes

$$0 \in A(v) + B(v). \tag{5}$$

A special case of (5) is the least-squares problem with the l_1 regularization:

$$\arg\min_{v\in\mathbb{R}^l}\{\|v\|_1 + \frac{1}{2\tau}\|Sv - t\|_2^2\}, \tag{6}$$

where $S \in \mathbb{R}^{m\times l}$ is a matrix, $t \in \mathbb{R}^m$, $\tau > 0$; $\|\cdot\|_1$ and $\|\cdot\|_2$ represent the standard l_1 and l_2 norms, respectively. To recover (6) by (5), just take $A(v) = \partial(\|v\|_1)$ and $B(v) = \frac{1}{\tau}S^T(Sv - t)$ where $\partial(\cdot)$ denotes the subdifferential of a convex but nonsmooth function. A very useful application of (6) is when $m \ll l$. For this case, (6) can be explained as finding a sparse vector satisfying the equations $Sv = t$.

Applying the PPA to solve (5) results in the iterative scheme

$$v_{n+1} = J_\lambda^{A+B}(v_n). \tag{7}$$

As we have analyzed, the scheme (7) requires evaluating the resolvent operator J_λ^{A+B} at each iteration. For many applications, however, evaluating J_λ^{A+B} is much harder than evaluating J_λ^A and J_λ^B individually. In fact, notice that the minimization problem

$$\arg\min_{v\in\mathbb{R}^l}\{\|v\|_1 + \frac{1}{2\tau}\|v - v_n\|^2\}$$

with $\tau > 0$ has a closed-form solution given by the soft-shrinkage operator (see e.g. [4]). Then obviously the model (6) is such an example. Thus, for solving (5), we are

[1] In practice, it is often too restrictive to estimate J_λ^T exactly. Thus, inexact versions of the PPA which require only solving (4) approximately subject to certain inexactness criteria have been intensively studied in the literature, see e.g.[30] for a seminal work. Later we will also discuss inexact versions of the generalized PPA to be proposed in Section 3.2.

more interested in designing an algorithm that requires only evaluating J_λ^A and J_λ^B, than just using the original PPA scheme (4) straightforwardly which needs to estimate J_λ^{A+B} —- the so-called operator splitting methods are thus named. Two influential operator splitting methods are the Douglas-Rachford splitting method (DRSM) in [7, 22]

$$u_{n+1} \in J_\lambda^B \left(J_\lambda^A(I-\lambda B)+\lambda B\right) u_n; \tag{8}$$

and the Peaceman-Rachford splitting method (PRSM) in [22, 27]

$$u_{n+1} \in J_\lambda^B(I-\lambda A)J_\lambda^A(I-\lambda B)u_n. \tag{9}$$

Since A and B could be set-valued, it is necessary to explain how to read the schemes (8) and (9). For a given u_0, we choose $b_0 \in Bu_0$ and denote $v_0 = u_0 + \lambda b_0$ such that $u_0 = J_\lambda^B v_0$ (the existence of such a pair is unique by the Representation Lemma, see [9]). The algorithms (8) and (9) become respectively

$$v_{n+1} = J_\lambda^A(2J_\lambda^B - I)v_n + (I - J_\lambda^B)v_n \tag{10}$$

and

$$v_{n+1} = (2J_\lambda^A - I)(2J_\lambda^B - I)v_n. \tag{11}$$

Obviously, these two schemes (10) and (11) can be retrieved by the scheme

$$v_{n+1} = v_n + \gamma\left(J_\lambda^A(2J_\lambda^B - I)v_n - J_\lambda^B v_n\right) \tag{12}$$

with $\gamma = 1$ and $\gamma = 2$, respectively.

In this paper, we propose the following generalized PPA scheme for solving (1):

$$v_{n+1} = \gamma J_\lambda^T(v_n) + (1-\gamma)v_n, \tag{13}$$

where $\lambda > 0$ and $\gamma > 0$. The original PPA (4) is obviously a special case of (13) with $\gamma = 1$. One more motivation of studying this generalized PPA scheme is that the formula (12) can be further written as

$$v_{n+1} = \gamma G_{\lambda,A,B} v_n + (1-\gamma)v_n$$

with

$$G_{\lambda,A,B} = J_\lambda^A(2J_\lambda^B - I) + I - J_\lambda^B.$$

Thus, let

$$S_{\lambda,A,B} := G_{\lambda,A,B}^{-1} - I,$$

or, more precisely (see [9]),

$$S_{\lambda,A,B} = (G_{\lambda,A,B})^{-1} - I = \{(v+\lambda b, u-v) | (u,b) \in B, (v,a) \in A, v+\lambda a = u - \lambda b\},$$

we have

$$G_{\lambda,A,B} = J_\lambda^{\frac{1}{\lambda}S_{\lambda,A,B}}.$$

Therefore, the scheme (12) is a special case of (13) with $T = \frac{1}{\lambda}S_{\lambda,A,B}$. Note that it has been studied in [9] that $S_{\lambda,A,B}$ defined above is maximal monotone when A and B are

both maximal monotone [2]. Aiming at extending the scheme (12), we are thus interested in the generalized PPA scheme (13).

Let us take a look at a particular convex minimization problem:

$$\min_{x\in\mathbb{R}^l} f(x) + g(Mx), \tag{14}$$

where $f : \mathbb{R}^l \mapsto]-\infty, +\infty]$ and $g : \mathbb{R}^m \mapsto]-\infty, +\infty]$ are closed, convex and proper functions; and $M \in \mathbb{R}^{m\times l}$. Obviously, the model (6) is a special case of (14). Applications of the model (14) include a range of image restoration models with the total variation regularization in [31]. For such an application, f denotes a data-fidelity term (e.g., the least-squares term), g represents a regularization term (e.g., the l_1-norm term to induce sparsity) and M is the matrix representation of a discrete gradient operator (e.g. the total variation operator in [31]). Introducing an auxiliary variable $y = Mx$, the model (14) can be reformulated as

$$\min_{x\in\mathbb{R}^l, y\in\mathbb{R}^m} \{f(x) + g(y) \mid ; Mx - y = 0\}. \tag{15}$$

A benchmark solver for (15) is the alternating direction method of multipliers (ADMM) proposed originally in [11]. Its iterative scheme reads as

$$\begin{cases} x_{n+1} = \arg\min_{x\in\mathbb{R}^l} \big\{ f(x) + \langle p_n, Mx\rangle + \dfrac{\lambda}{2}||Mx - y_n||^2 \big\}, \\ y_{n+1} = \arg\min_{y\in\mathbb{R}^m} \big\{ g(y) - \langle p_n, y\rangle + \dfrac{\lambda}{2}||Mx_{n+1} - y||^2 \big\}, \\ p_{n+1} = p_n + \lambda(Mx_{n+1} - y_{n+1}), \end{cases} \tag{16}$$

where p_n is the Lagrange multiplier of (15) and λ plays the role of a penalty parameter. Then, the generalized ADMM scheme was proposed in [9]

$$\begin{cases} x_{n+1} = \arg\min_{x\in\mathbb{R}^l} \big\{ f(x) + \langle p_n, Mx\rangle + \dfrac{\lambda}{2}||Mx - y_n||^2 \big\}, \\ y_{n+1} = \arg\min_{y\in\mathbb{R}^m} \big\{ g(y) - \langle p_n, y\rangle + \dfrac{\lambda}{2}||\gamma Mx_{n+1} + (1-\gamma)y_n - y||^2 \big\}, \\ p_{n+1} = p_n + \lambda(\gamma Mx_{n+1} + (1-\gamma)y_n - y_{n+1}), \end{cases} \tag{17}$$

with $\gamma \in (0, 2)$. Obviously, (17) includes (16) as a special case with $\gamma = 1$. In [8], it was shown that the generalized ADMM scheme (17) can be obtained by applying the scheme (12) with $\gamma \in (0, 2)$ to the dual of (14):

$$\min_{p\in\mathbb{R}^m} \{f^*(-M^T p) + g^*(p)\}, \tag{18}$$

where "$*$" denotes the Fenchel transform, see, e.g., [29]. Note that (18) is a special case of (5) with $A = \partial\left(f^* \circ (-M^T)\right)$ and $B = \partial(g^*)$. Hence, both the schemes (16) and (17) are special cases of the generalized PPA scheme (13) under consideration.

[2] If $(x, y), (\bar{x}, \bar{y}) \in S_{\lambda,A,B}$, then it exists $(u, b), (\bar{u}, \bar{b}) \in B$, $(v, a), (\bar{v}, \bar{a}) \in A$ such that $v + \lambda a = u - \lambda b$ and $\bar{v} + \lambda\bar{a} = \bar{u} - \nu\bar{b}$. We thus have $\langle x - \bar{x}, y - \bar{y}\rangle = \lambda\langle a - \bar{a}, v - \bar{v}\rangle + \lambda\langle b - \bar{b}, u - \bar{u}\rangle \geq 0$.

Now, we explain the allowable range for γ in (13). As we just showed, in the literature it is often required to choose $\gamma \in (0, 2]$ and the case with $\gamma > 2$ is seldom addressed (to the best of our knowledge). Note that for a root v of T, we have

$$\begin{aligned}||v_{n+1} - \frac{1}{2}(\gamma v + (2-\gamma)v_n)||^2 &= ||\gamma(J_\lambda^T(v_n) - J_\lambda^T(v)) + \frac{\gamma}{2}(v - v_n)||^2 \\ &= ||\frac{\gamma}{2}(v - v_n)||^2 \\ &\quad + \gamma^2 \left(||J_\lambda^T(v_n) - J_\lambda^T(v)||^2 - \langle J_\lambda^T(v) - J_\lambda^T(v_n), v - v_n\rangle\right) \\ &\leq ||\frac{\gamma}{2}(v - v_n)||^2,\end{aligned}$$

where the inequality is due to the firm non-expansiveness of J_λ^T (see e.g. [29]) and the fact that v is a root of T. Therefore, a big difference between the cases $\gamma \in (0, 2)$ and $\gamma = 2$ occurs: If $\gamma = 2$, we only have $\|v_{n+1} - v\| \leq \|v - v_n\|$ and thus the sequence $(v_n)_{n\geq 0}$ might not be strictly contractive with respect to the root set of T. When $\gamma \in (0, 2)$, the above fact illuminates that v_{n+1} lies in the ball centered at $\frac{1}{2}(\gamma v + (2-\gamma)v_n)$ with the radius $||\frac{\gamma}{2}(v - v_n)||$. This fact thus raises the difference in analyzing the convergence rate of (13) for the cases $\gamma \in (0, 2)$ and $\gamma = 2$. We use Figure 1 to illustrate this fact. Finally, we notice that the case $\gamma > 2$ is also worth investigation although in the literature, to the best of our knowledge, there is no rigorous convergence study for this case. The necessity of studying the case where $\gamma > 2$ can be seen from the following example.

Example 1: *Let* $T : x \in \mathbb{R}^2 \to y \in \mathbb{R}^2$ *be defined as* $\{y_1 = \frac{x_1^3}{1+x_1^2},\ y_2 = \frac{x_2^3}{1+|x_2|^3}\}$. *Then* $(0, 0)$ *is a root of this* T.

Let the scheme (13) be implemented with $\lambda = 1$ and the starting point $(-2, -2)$. We plot the iterative procedure of (13) with different values of γ in Figure 2, and we can see for this example that $\gamma > 2$ can accelerate the convergence.

Our main purpose is to analyze the convergence rate for the generalized PPA scheme (13) with a generic T and $\gamma > 0$. As we have mentioned, the value of γ results in different iterative performance of the scheme (13). We thus will discuss three cases individually: $\gamma \in (0, 2)$, $\gamma = 2$ and $\gamma \in (0, \nu)$ with $\nu > 2$. We first estimate a worst-case convergence rate measured by the iteration complexity for (13). Note that as [25, 26], the worst-case convergence rate of an iterative scheme can be measured by the iteration complexity, which means precisely that we can find an approximate root of T with an accuracy of $O\left(\frac{1}{n}\right)$ after n iterations of (13). Then, we shall discuss under which conditions the scheme (13) converges to a root of T on a linear rate.

We briefly review existing convergence rate results for some special cases of the scheme (13). For some special optimization models, the ADMM which is a special case of (13) with $\gamma = 1$ was shown to have a worst-case $O\left(\frac{1}{n}\right)$ convergence rate in [17] (the ergodic sense) and [18] (a nonergodic sense). Later, the $O\left(\frac{1}{n}\right)$ convergence rate of the ADMM was improved in [5] to an order of $o\left(\frac{1}{n}\right)$. Recently, the linear convergence of ADMM for some special cases and under some stronger conditions has been discussed in [2, 6, 15], and the linear convergence of some extended versions of the ADMM scheme can be found in [16, 21]. A more comprehensive convergence rate analysis for operator splitting methods was presented most recently in [32]. In [13], the author established a worst-case $O\left(\frac{1}{n}\right)$ convergence rate for the application of the classical PPA to a convex minimization model, and an accelerated version with a worst-case $O\left(\frac{1}{n^2}\right)$ convergence rate. For the generic DRSM scheme (8), a worst-case $O\left(\frac{1}{n}\right)$ convergence rate can be

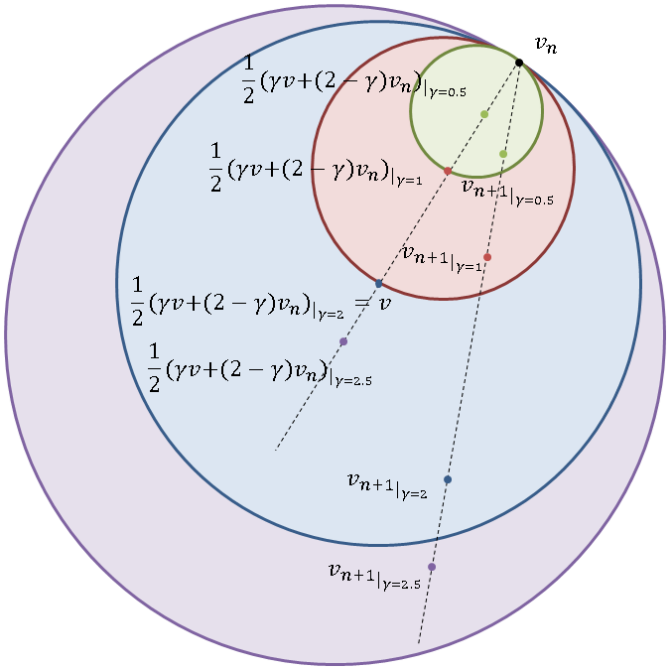

Figure 1: The firm non-expansiveness of the J_λ^T implies that the iterate v_{n+1} lies in a ball whose size depends on v and γ.

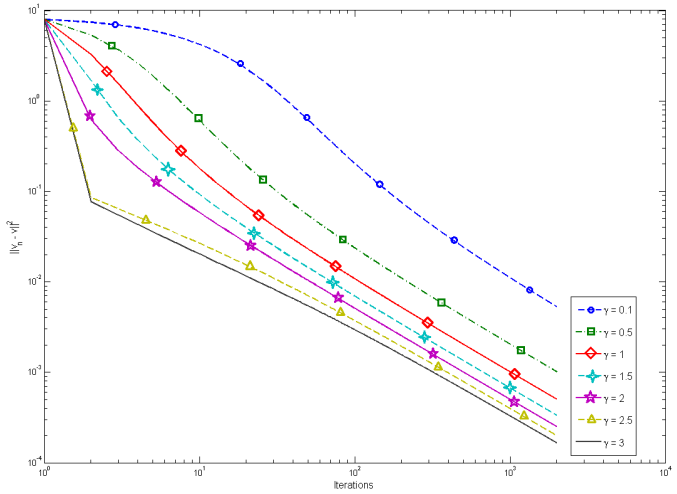

Figure 2: $||v_n - v||^2$ with regard to the iterations in log scale. We compare various choices of γ. For $\gamma > 2$ the algorithm is still convergent and more efficient than some choices $\gamma \in (0, 2]$.

found in [19]. The linear convergence of the DRSM scheme (8) and PRSM scheme (9) were discussed in [22] under some additional conditions on A and B. We also refer to [12], where the convergence rates of the DRSM and PRSM schemes were discussed for some special cases of (5).

The rest of this paper is organized as follows. In Section 2, some preliminaries are presented. Then, we discuss the convergence rate of (13) in Sections 3-5 for different cases of γ. In Section 6, we discuss the linear convergence rate of (13). Finally, we make some conclusions in Section 7.

2 Preliminaries

In this section, we summarize some known results and then prove some basic propositions which are useful for further discussion.

2.1 Yosida Approximation

We first recall the Yosida approximation operator and some of its properties. All results in this subsection can be found in the literature, e.g., [3]. Since the proofs of the properties to be stated are very short, we include them for completeness.

For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, the Yoshida approximation operator (with parameter $\lambda > 0$) is defined as

$$T_\lambda = \frac{1}{\lambda}(I - J_\lambda^T),$$

where J_λ^T is the resolvent operator of T. The Yosida approximation operator T_λ is single-valued, and it is related to the operator $T(J_\lambda^T)$ (which could be set-valued) in the following proposition.

Proposition 2.1 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let J_λ^T be the resolvent operator and T_λ be the Yosida approximation operator of T. Then, we have*

$$\forall v \in \mathbb{H}, \quad T_\lambda(v) \in T(J_\lambda^T(v)).$$

Proof According to the definitions of J_λ^T and T_λ, we have

$$T_\lambda(v) = \frac{1}{\lambda}(v - J_\lambda^T(v)) \in \frac{1}{\lambda}((I + \lambda T)J_\lambda^T(v) - J_\lambda^T(v)) = T(J_\lambda^T(v)).$$

This completes the proof. □

The following identity is very useful in our analysis.

Proposition 2.2 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let J_λ^T be the resolvent operator and T_λ be the Yosida approximation operator of T. Then, $\forall v_1, v_2 \in \mathbb{H}$, we have*

$$\langle T_\lambda(v_1) - T_\lambda(v_2), v_1 - v_2\rangle = \lambda||T_\lambda(v_1) - T_\lambda(v_2)||^2 + \langle T_\lambda(v_1) - T_\lambda(v_2), J_\lambda^T(v_1) - J_\lambda^T(v_2)\rangle.$$

Proof Using the definition of J_λ^T, we have

$$\begin{aligned}\langle T_\lambda(v_1) - T_\lambda(v_2), v_1 - v_2\rangle &= \langle T_\lambda(v_1) - T_\lambda(v_2), \lambda T_\lambda(v_1) - \lambda T_\lambda(v_2)\rangle \\ &\quad + \langle T_\lambda(v_1) - T_\lambda(v_2), J_\lambda^T(v_1) - J_\lambda^T(v_2)\rangle \\ &= \lambda||T_\lambda(v_1) - T_\lambda(v_2)||^2 + \langle T_\lambda(v_1) - T_\lambda(v_2), J_\lambda^T(v_1) - J_\lambda^T(v_2)\rangle.\end{aligned}$$

The assertion is proved. □

Based on Propositions 2.1 and 2.2, we immediately have the following proposition.

Proposition 2.3 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let J_λ^T be the resolvent operator and T_λ be the Yosida approximation operator of T. Then, T_λ is λ-firmly non-expansive and $\frac{1}{\lambda}$-Lipschitz continuous.*

Proof It follows from Proposition 2.1 that $T_\lambda(v) \in T(J_\lambda^T(v))$. We thus have

$$\forall v_1, v_2 \in \mathbb{H}, \quad \langle T_\lambda(v_1) - T_\lambda(v_2), J_\lambda^T(v_1) - J_\lambda^T(v_2)\rangle \geq 0.$$

Then, substituting this inequality into the assertion of Proposition 2.2, we conclude immediately that T_λ is λ-firmly non-expansive and $\frac{1}{\lambda}$-Lipschitz continuous. □

The following proposition inspires us to measure the accuracy of an iterate to a root of T by $\|T_\lambda(v)\|^2$.

Proposition 2.4 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let T_λ be the Yosida approximation operator of T. Then we have*

$$\forall \lambda > 0, \qquad 0 \in T(v) \quad \Leftrightarrow \quad T_\lambda(v) = 0.$$

Proof Because of the definition of T_λ, we know that $u \in T_\lambda(v) \Leftrightarrow u \in T(v - \lambda u)$. Hence, we have

$$0 \in T(v) \quad \Leftrightarrow \quad 0 \in T(v - \lambda 0) \quad \Leftrightarrow \quad 0 \in T_\lambda(v).$$

But T_λ is indeed single-valued. Thus, we have $T_\lambda(v) = 0$. The proof is complete. □

Remark A natural way to measure the accuracy of an iterate v_n is to calculate $\|T(v_n)\|$. Here, we use $\|T_\lambda(v_n)\|$, rather than $\|T(v_n)\|$, as the measurement of accuracy for the iterate v_n to a root of T. In fact, we can show that $\|T(v_n)\|$ and $\|T_\lambda(v_n)\|$ are comparable in measuring the accuracy for an iterate. First we use Proposition 2.1:

$$T_\lambda(v_n) \in T(J_\lambda^T(v_n)) = T(v_n + (v_{n+1} - v_n)/\gamma)$$

and thus have

$$\min_{t_n \in T(v_n + (v_{n+1} - v_n)/\gamma)} \|t_n\| \leq \|T_\lambda(v_n)\|.$$

Moreover, Proposition 2.1 implies that

$$\langle t_n - T_\lambda(v_n), v_n - J_\lambda^T(v_n)\rangle \geq 0, \quad \forall t_n \in T(v_n),$$

which leads to

$$\|T_\lambda(v_n)\| \leq \|t_n\|, \quad \forall t_n \in T(v_n).$$

Hence, we have

$$\min_{t \in T(v_n + (v_{n+1} - v_n)/\gamma)} \|t\| \leq \|T_\lambda(v_n)\| \leq \min_{t \in T(v_n)} \|t\|$$

which shows that the accuracy of v_n to a root of T can be measured by either $\|T_\lambda(v_n)\|^2$ or $\|T(v_n)\|^2$.

2.2 Some preliminary properties

In this subsection, we prove some properties of the sequence $(v_n)_{n\geq 0}$ generated by the proposed generalized PPA scheme (13), and they will be used often later.

First of all, by using the Yosida approximation operator, we can rewrite the scheme (13) as

$$v_{n+1} = v_n - \gamma\lambda T_\lambda(v_n). \tag{19}$$

We first compare the difference of the proximity to a root of T (denoted by v) for two consecutive iterates v_{n+1} and v_n generated by (13).

Lemma 2.5 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let J_λ^T be the resolvent operator and T_λ be the Yosida approximation operator of T. Let $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (13) and v be a root of T. For any $\lambda > 0$, we have*

$$||v_{n+1} - v||^2 = ||v_n - v||^2 - \gamma(2-\gamma)\lambda^2||T_\lambda(v_n)||^2 - 2\gamma\lambda\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle.$$

Proof Using the expression (19), we have

$$\begin{aligned} ||v_{n+1} - v||^2 &= ||v_n - v - \gamma\lambda T_\lambda(v_n)||^2 \\ &= ||v_n - v||^2 + \gamma^2\lambda^2||T_\lambda(v_n)||^2 - 2\gamma\lambda\langle T_\lambda(v_n), v_n - v\rangle. \end{aligned}$$

Then, applying the assertion of Proposition 2.2 and using the fact that $T_\lambda(v) = 0$, we get

$$||v_{n+1} - v||^2 = ||v_n - v||^2 - \gamma(2-\gamma)\lambda^2||T_\lambda(v_n)||^2 - 2\gamma\lambda\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle.$$

The proof is complete. □

Remark Since T is maximal monotone and $T_\lambda(v) \in T(J_\lambda^T(v))$, we have

$$\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle \geq 0.$$

Therefore, the assertion of Lemma 2.5 implies that the sequence $(\|v_n - v\|^2)_{n\geq 0}$ is non-increasing if $\gamma \in (0, 2]$. Moreover, the sequence $(v_n)_{n\geq 0}$ is strictly contractive with respect to the root set of T when $\gamma \in (0, 2)$. Based on this fact, the convergence of the generalized PPA scheme (13) with $\gamma \in (0, 2)$ can be readily derived by standard techniques of contraction methods, see. e.g. [1].

In the following, we study the monotonicity of the sequence $(\|T_\lambda(v_n)\|^2)_{n\geq 0}$ where $(v_n)_{n\geq 0}$ is generated by the generalized PPA scheme (13). Recall we have shown that $\|T_\lambda(v_n)\|^2$ can be used to measure the accuracy of v_n to a root of T.

Lemma 2.6 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let J_λ^T be the resolvent operator and T_λ be the Yosida approximation operator of T. Let $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (13) and v be a root of T. For any $\lambda > 0$, we have*

$$\begin{aligned} ||T_\lambda(v_{n+1})||^2 &= ||T_\lambda(v_n)||^2 - \frac{2-\gamma}{\gamma}||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 \\ &\quad -\frac{2}{\gamma\lambda}\langle T_\lambda(v_{n+1}) - T_\lambda(v_n), J_\lambda^T(v_{n+1}) - J_\lambda^T(v_n)\rangle. \end{aligned}$$

Proof Using the formula (13), we have

$$\begin{aligned}||T_\lambda(v_{n+1})||^2 &= ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 + ||T_\lambda(v_n)||^2 + 2\langle T_\lambda(v_{n+1}) - T_\lambda(v_n), T_\lambda(v_n)\rangle \\ &= ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 + ||T_\lambda(v_n)||^2 - \frac{2}{\gamma\lambda}\langle T_\lambda(v_{n+1}) - T_\lambda(v_n), v_{n+1} - v_n\rangle.\end{aligned}$$

Then, applying the assertion of Proposition 2.2, we get

$$\begin{aligned}||T_\lambda(v_{n+1})||^2 &= ||T_\lambda(v_n)||^2 - \frac{2-\gamma}{\gamma}||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 \\ &\quad -\frac{2}{\gamma\lambda}\langle T_\lambda(v_{n+1}) - T_\lambda(v_n), J_\lambda^T(v_{n+1}) - J_\lambda^T(v_n)\rangle.\end{aligned}$$

The proof is complete. □

Remark Recall Proposition 2.1 and the monotonicity of T. We know that

$$\langle T_\lambda(v_{n+1}) - T_\lambda(v_n), J_\lambda^T(v_{n+1}) - J_\lambda^T(v_n)\rangle \geq 0.$$

Hence, Lemma 2.6 shows that the sequence $\{||T_\lambda(v_n)||\}^2$ is non-increasing when $\gamma \in (0, 2]$.

Finally, we recall Lemma 2.1 in [5] which is useful for refining the $O\left(\frac{1}{n}\right)$ convergence rate to be established to an order of $o\left(\frac{1}{n}\right)$. We omit the proof which can be found in [5].

Lemma 2.7 *Let $(u_n)_n$ be a non-negative, monotonically non-increasing and summable sequence ($\sum_{n=1}^{\infty} u_n < +\infty$). Then we have $u_n = o\left(\frac{1}{n}\right)$.*

3 Case 1: $\gamma \in (0, 2)$

Now, we start to estimate the convergence rate of the generalized PPA scheme (13). We first focus on estimating its worst-case convergence rate measured by the iteration complexity without additional assumptions on the operator T. As we have mentioned, the techniques to derive the worst-case convergence rate for different values of γ are different (e.g., it follows from Lemma 2.6 that the sequence $(v_n)_{n\geq 0}$ generated by (13) is contractive with respect to the root set of T when $\gamma \in (0, 2]$, but it does not hold when $\gamma > 2$). Thus we discuss the cases $\gamma \in (0, 2)$, $\gamma = 2$ and $\gamma \in (0, \nu)$ with $\nu > 2$ individually in the coming sections.

3.1 Convergence rate with the exact evaluation of J_λ^T

We first assume that the resolvent operator J_λ^T can be evaluated accurately at any point for implementing the generalized PPA scheme (13). For this case, we can estimate the worst-case convergence rate of (13) in terms of $\|T_\lambda(v_n)\|^2$, as shown in the following theorem.

Theorem 3.1 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let T_λ be the Yosida approximation operator of T. Let $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (13) with $\gamma \in (0,2)$ and v be a root of T. For any $\lambda > 0$, we have*

$$||T_\lambda(v_n)||^2 \leq \frac{||v_0 - v||^2}{\gamma(2-\gamma)\lambda^2(n+1)}.$$

Proof It follows from Lemma 2.5 and its remark that

$$||v_{n+1} - v||^2 \leq ||v_n - v||^2 - \gamma(2-\gamma)\lambda^2||T_\lambda(v_n)||^2.$$

Summing this inequality over $i = 0, 1, 2, \cdots, n$, we get

$$\begin{aligned}\gamma(2-\gamma)\lambda^2 \sum_{i=0}^{n} ||T_\lambda(v_i)||^2 &\leq ||v_0 - v||^2 - ||v_{n+1} - v||^2 \\ &\leq ||v_0 - v||^2.\end{aligned}$$

Then, it follows from Lemma 2.6 that

$$||T_\lambda(v_{i+1})||^2 \leq ||T_\lambda(v_i)||^2$$

when $\gamma \in (0,2)$. Thus, we have

$$||T_\lambda(v_n)||^2 \leq \frac{||v_0 - v||^2}{\gamma(2-\gamma)\lambda^2(n+1)}.$$

The proof is complete. □

Recall that $\|T_\lambda(v_n)\|^2$ can be used to measure the accuracy of v_n to a root of T (see Proposition 2.4). Thus, Theorem 3.1 shows that after n iterations, the iterate generated by (13) with $\gamma \in (0,2)$ produces an approximate root of T with an accuracy of $O\left(\frac{1}{n}\right)$. Thus, a worst-case $O\left(\frac{1}{n}\right)$ convergence rate measured by the iteration complexity is established for (13) with $\gamma \in (0,2)$. This is an extended result of the work [18] which is only for the special scheme (16). Moreover, note that the sequence $(||T_\lambda(v_n)||^2)_{n\geq 0}$ fulfills all the requirements of Lemma 2.7. Therefore we can refine the assertion in Theorem 3.1 as

$$||T_\lambda(v_n)||^2 = o\left(\frac{1}{n}\right),$$

which means a worst-case $o\left(\frac{1}{n}\right)$ convergence rate of (13) with $\gamma \in (0,2)$.

3.2 Convergence rate with an estimate of J_λ^T

We then discuss the case where the resolvent operator J_λ^T can only be estimated approximately. This consideration makes senses for many applications, and it has inspired

the seminal work of approximate PPA in [30]. Let us consider an inexact version of the generalized PPA scheme (13) with $\gamma \in (0, 2)$ as following:

$$\begin{array}{l} v_{n+1} = \gamma w_n + (1-\gamma)v_n \\ \text{s.t} \quad ||w_n - J_\lambda^T(v_n)|| \leq \epsilon_n. \end{array} \tag{20}$$

In (20), w_n represents an estimate of J_λ^T at the point v_n and ϵ_n denotes the accuracy of this estimate. Choosing different ϵ_n leads to different inexact versions of the generalized PPA (13), and there are many ways to design appropriate inexact criteria to control the accuracy ϵ_n. In fact, some well-studied criteria in the PPA literature (e.g. [14, 30]) can be used here for (20). Also, there are alternative criteria which do not involve $J_\lambda^T(v_n)$ and thus can be implemented directly. However, for the succinctness and a clearer exposition of our main result, we just discuss an inexact criterion analogous to that in [30] for the classical PPA (4). This is a conceptual criterion but also a fundamental one scalable to other existing criteria in the PPA literature.

A necessary rule of choosing ϵ_n is that $\epsilon_n \to 0$ when $n \to +\infty$, i.e., the accuracy of solving the subproblems should tend to more and more accurate as the iteration goes on. We choose ϵ_n as

$$\forall n \geq 0, \quad \epsilon_n = O\left(\frac{1}{(n+1)^\alpha}\right), \quad \alpha > 1, \tag{21}$$

and then estimate the convergence rate for the inexact version of generalized PPA (20) with the criterion (21).

For notational simplicity, we denote

$$E_1 = \sum_{i=0}^{\infty} \epsilon_i \ \text{ and } \ E_2 = \sum_{i=0}^{\infty} \epsilon_i^2.$$

With the choice (21), obviously it holds that

$$E_1 < +\infty \ \text{ and } \ E_2 < +\infty.$$

Now, we derive a worst-case convergence rate for the scheme (20) with $\gamma \in (0, 2)$ and the criterion (21) in the following theorem.

Theorem 3.2 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let T_λ be the Yosida approximation of T. Let $(v_n)_{n\geq 0}$ be the sequence generated by the inexact version of generalized PPA scheme (20) with $\gamma \in (0, 2)$ and the criterion (21), and v be a root of T. For any $\lambda > 0$, we have*

$$||T_\lambda(v_n)||^2 = O\left(\frac{1}{n+1}\right), \quad \forall n \geq 0.$$

Proof We denote by $\bar{v}_{n+1}$ the iterate generated by the generalized PPA (13). That is,

$$\bar{v}_{n+1} = \gamma J_\lambda^T(v_n) + (1-\gamma)v_n.$$

By Lemma 2.5 we have

$$||\bar{v}_{n+1} - v||^2 \leq ||v_n - v||^2 - \gamma(2-\gamma)\lambda^2||T_\lambda(v_n)||^2.$$

Now we find a bound for the proximity of v_{n+1} to v. In fact, the scheme (20) shows $||v_{n+1} - \bar{v}_{n+1}|| \leq \gamma\epsilon_n$. So, we have

$$\begin{aligned} ||v_{n+1} - v|| &\leq ||\bar{v}_{n+1} - v|| + ||v_{n+1} - \bar{v}_{n+1}|| \\ &\leq ||\bar{v}_{n+1} - v|| + \gamma\epsilon_n \\ &\leq ||v_0 - v|| + \sum_{i=0}^{n} \gamma\epsilon_i \\ &\leq ||v_0 - v|| + \gamma E_1. \end{aligned}$$

Moreover, simple manipulation gives us

$$\begin{aligned} ||v_{n+1} - v||^2 &= ||\bar{v}_{n+1} - v||^2 + ||v_{n+1} - \bar{v}_{n+1}||^2 + 2\langle \bar{v}_{n+1} - v, v_{n+1} - \bar{v}_{n+1} \rangle \\ &\leq ||\bar{v}_{n+1} - v||^2 + ||v_{n+1} - \bar{v}_{n+1}||^2 + 2||\bar{v}_{n+1} - v||\ ||v_{n+1} - \bar{v}_{n+1}|| \\ &\leq ||v_n - v||^2 - \gamma(2-\gamma)\lambda^2 ||T_\lambda(v_n)||^2 + \gamma^2\epsilon_n^2 + 2\gamma\epsilon_n(||v_0 - v|| + \gamma E_1). \end{aligned}$$

Summarizing this inequality over $i = 0, 1, \cdots, n$, we get

$$\begin{aligned} \gamma(2-\gamma)\lambda^2 \sum_{i=0}^{n} ||T_\lambda(v_i)||^2 &\leq ||v_0 - v||^2 - ||v_{n+1} - v||^2 + \sum_{i=0}^{n} \left(\gamma^2\epsilon_i^2 + 2\gamma\epsilon_i(||v_0 - v|| + \gamma E_1)\right) \\ &\leq (||v_0 - v|| + \gamma E_1)^2 + \gamma^2(E_2 + E_1^2). \end{aligned}$$

We also have

$$\begin{aligned} ||T_\lambda(v_{n+1})||^2 &= ||T_\lambda(v_n)||^2 + ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 + 2\langle T_\lambda(v_{n+1}) - T_\lambda(v_n), T_\lambda(v_n) \rangle \\ &\leq ||T_\lambda(v_n)||^2 - \frac{2-\gamma}{\gamma} ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 \\ &+ \frac{2}{\lambda} ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||\ ||w_n - J_\lambda^T(v_n)||. \end{aligned}$$

Using the Young inequality on the last term above, we get

$$\frac{2}{\lambda} ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||\ ||w_n - J_\lambda^T(v_n)|| \leq \frac{2-\gamma}{\gamma} ||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2 + \frac{\gamma}{2-\gamma} \frac{\epsilon_n^2}{\lambda^2}.$$

Therefore, we have

$$||T_\lambda(v_{n+1})||^2 \leq ||T_\lambda(v_n)||^2 + \frac{\gamma}{2-\gamma} \frac{\epsilon_n^2}{\lambda^2}. \tag{22}$$

Combining this equation from p to $n-1$ yields

$$||T_\lambda(v_n)||^2 \leq ||T_\lambda(v_p)||^2 + \frac{\gamma}{2-\gamma} \sum_{j=p}^{n-1} \frac{\epsilon_j^2}{\lambda^2}.$$

Hence, we have

$$||T_\lambda(v_n)||^2 \leq \frac{(||v_0 - v|| + \gamma E_1)^2 + \gamma^2(E_2 + E_1^2)}{\gamma(2-\gamma)\lambda^2(n+1)} + \gamma \frac{\sum_{i=0}^{n} \sum_{j=i}^{n-1} \epsilon_j^2}{(2-\gamma)\lambda^2(n+1)}.$$

As ϵ_n satisfies the requirement (21), there exists a constant $K > 0$ such that

$$\sum_{i=0}^{n}\sum_{j=i}^{n-1}\epsilon_j^2 \le K\int_0^{n+1}\int_y^{n+1}(x+1)^{-2\alpha}\,dx\,dy$$
$$\le \frac{K}{2(2\alpha-1)(\alpha-1)}, \quad \text{since } \alpha > 1.$$

Finally, we have

$$||T_\lambda(v_n)||^2 = O\left(\frac{1}{n+1}\right),$$

and the proof is complete. □

Theorem 3.2 thus shows that the accuracy of v_n to a root of T (measured by $\|T_\lambda(v_n)\|^2$) is in order of $O\left(\frac{1}{n}\right)$. A worst-case $O\left(\frac{1}{n}\right)$ convergence rate is thus established for the inexact version of generalized PPA (20). We can also refine the result in Theorem 3.2 to an order of $o\left(\frac{1}{n}\right)$. In fact, Lemma 2.7 is not applicable for this purpose. But, using (22), we have

$$n||T_\lambda(v_{2n})||^2 \le \sum_{i=n}^{2n}||T_\lambda(v_i)||^2 + \frac{\gamma}{(2-\gamma)\lambda^2}\sum_{i=n}^{2n}\sum_{j=i}^{2n-1}\epsilon_j^2.$$

As ϵ_n satisfies the requirement (21), the right-hand term above goes to zero as n goes to infinity. Thus, we have

$$||T_\lambda(v_n)||^2 = o\left(\frac{1}{n}\right).$$

A worst-case $o\left(\frac{1}{n}\right)$ convergence rate is thus established for the inexact version of generalized PPA (20) with the criterion (21).

Remark The analysis in Theorem 3.2 also shows an interesting fact: If the accuracy ϵ_n is increased rapidly enough, e.g., α is increased rapidly enough, the inexact version of generalized PPA (13) admits a sublinear convergence rate.

4 Case 2: $\gamma = 2$

Now, we discuss the convergence rate of the generalized PPA (13) with $\gamma = 2$. As we have shown in the introduction and Lemma 2.5, this case differs from the case $\gamma \in (0, 2)$ significantly in that its sequence might not be strictly contractive with respect to the root set of T. This makes the convergence analysis much more challenging. Therefore, in this section we first analyze some convergence issues for this case and then derive its convergence rate under one additional assumption on T. Note that we only discuss the exact version (13) where J_λ^T is assumed to be evaluated exactly, and skip the discussion on inexact versions of (13) with estimates of J_λ^T.

4.1 Convergence issues

In Lemma 2.5, we show that the sequence generated by the generalized PPA scheme (13) is strictly contractive with respect to the root set of T if $\gamma \in (0, 2)$. Thus, the convergence for this case can be easily established, see [9, 19] for more details. Let us now explain more why the case with $\gamma = 2$ deserves special consideration. In fact, by Lemma 2.5, we know that if $\gamma = 2$, we have

$$||v_{n+1} - v||^2 = ||v_n - v||^2 - 4\lambda\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle.$$

Hence, whether or not the new iterate v_{n+1} is closer to a root of T than the previous iterate v_n is determined by the scalar product

$$\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle. \tag{23}$$

If it happens that this scalar product remains zero during the iteration, then the sequence $(v_n)_{n\geq 0}$ generated by (13) with $\gamma = 2$ maintains a constant distance from a root of T and it never converges.

To compare the difference of convergence for the cases where $\gamma \in (0, 2)$ and $\gamma = 2$, let us consider the following example.

Example 2 *Let $T : v \in \mathbb{R}^2 \to u \in \mathbb{R}^2$ be defined as $\{y^1 = -x^2, y^2 = x^1\}$. The root of this T is $(0, 0)$.*

It is easy to verify that $\langle T(v), v\rangle = 0$ for all $v \in \mathbb{R}^2$. Thus, if γ takes 2 in (13), all the iterates generated by (13) maintain a constant distance from $(0, 0)$. However, if $\gamma \in (0, 2)$, the sequence generated by (13) converges to $(0, 0)$ (in fact, the convergent rate is linear). In Figure 3, we plot the difference of convergence for the cases where $\gamma = 1$ and $\gamma = 2$ in (13).

Considering the analysis before, we thus need to pose certain additional assumptions on T in order to ensure the convergence of the generalized PPA (13) with $\gamma = 2$. Our assumption is as follows.

Assumption 1 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone; J_λ^T be the resolvent operator, T_λ be the Yosida approximation operator, and v be a root of T. The following property is assumed to hold for T:*

$$\forall u \in \mathbb{H}, \quad \langle T_\lambda(u), J_\lambda^T(u) - J_\lambda^T(v)\rangle = 0 \quad \Rightarrow \quad 0 \in T(J_\lambda^T(u)).$$

4.2 Convergence rate

In this subsection, we derive a worst-case convergence rate for the generalized PPA (13) with $\gamma = 2$ under Assumption 1. Different from the case where $\gamma \in (0, 2)$, the convergence rate to be derived is in the ergodic sense.

Under Assumption 1, the scalar product $\langle T_\lambda(v_n), J_\lambda(v_n) - J_\lambda(v)\rangle$ can be used to measure the accuracy of $J_\lambda^T(v_n)$ to a root of T. We are interested in the average of $\langle T_\lambda(v_n), J_\lambda(v_n) - J_\lambda(v)\rangle$ over all the first $n + 1$ iterations. That is, let

$$\delta_n := \frac{1}{n+1}\sum_{i=0}^{n}\langle T_\lambda(v_i), J_\lambda^T(v_i) - J_\lambda^T(v)\rangle, \tag{24}$$

we will find a bound of δ_n in the following theorem.

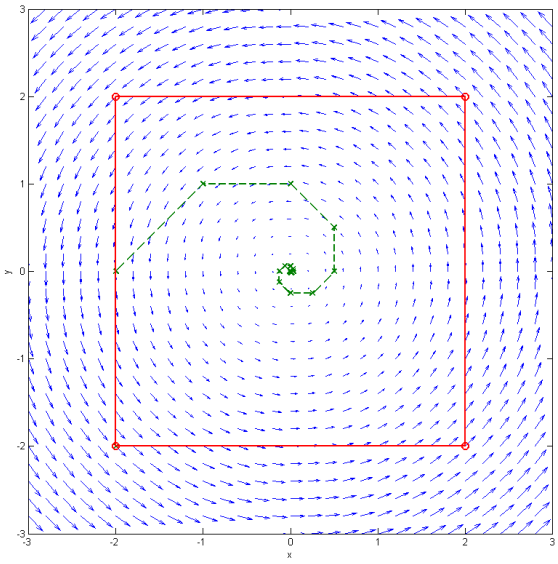

Figure 3: $T : v \in \mathbb{R}^2 \to u \in \mathbb{R}^2$ is defined as $\{y^1 = -x^2, y^2 = x^1\}$. Starting point: $(-2, -2)$; and $\lambda = 1$ in (13). The arrows represent the vector field of T . The case $\gamma = 2$ (in plain line) does not converge and the sequence has four cluster points (corner points); and the case $\gamma = 1$ (in dash line) converges to $(0, 0)$.

Theorem 4.1 *For a set-valued maximal monotone operator $T :\in \mathbb{H} \to 2^{\mathbb{H}}$, let δ_n be defined in (24). Let $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (13) with $\gamma = 2$ and v be a root of T. For any $\lambda > 0$, we have*

$$\forall n \geq 1, \qquad \delta_n \leq \frac{||v_0 - v||^2}{4\lambda(n+1)}.$$

Proof The assertion is obtained by setting $\gamma = 2$ in Lemma 2.5 and taking the average over all the first $n + 1$ iterates. The proof is complete. □

Theorem 4.1 shows a worst-case $O\left(\frac{1}{n}\right)$ convergence rate for the generalized PPA scheme (13) with $\gamma = 2$ in the ergodic sense. On the other hand, notice that the sequence $(\langle T_\lambda(v_n), J^T_\lambda(v_n) - J^T_\lambda(v)\rangle)_{n\geq 0}$ is in general not monotone. Thus the conclusion in Lemma 2.7 is not applicable. However, if we take the minimum over the set $\{1...n\}$, and uses Lemmas 2.5 and 2.7, then we obtain

$$\min_{i\in\{1...n\}} \langle T_\lambda(v_i), J^T_\lambda(v_i) - J^T_\lambda(v)\rangle = o\left(\frac{1}{n}\right),$$

which implies a worst-case $o\left(\frac{1}{n}\right)$ convergence rate of (13) with $\gamma = 2$.

5 Case 3: $\gamma \in (0, \nu)$ with $\nu > 2$

As we have shown, for some cases the generalized PPA scheme (13) with $\gamma > 2$ can converge faster. It is thus necessary to discuss the convergence rate of (13) which allows

γ to be greater than 2. Again, we skip the discussion for inexact versions of (13). To the best of our knowledge, even for the special cases (17) for solving the convex optimization model (14), there is no rigorous convergence analysis when $\gamma > 2$.

Due to the more aggressive range of γ, it is expected that certain additional assumptions onto T should be posed in order to derive the same worst-case convergence rate as the cases with narrower ranges of γ. Our analysis is conducted under the following assumption.

Assumption 2 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and F-firmly non-expansive, i.e.,*

$$\langle T(v_1) - T(v_2), v_1 - v_2 \rangle \geq F||T(v_1) - T(v_2)||^2, \ \forall v_1, v_2 \in \mathbb{H}.$$

Remark Let us specify Assumption 2 to some special cases of T.

- **Scheme (12)**
  In this case, $T = \frac{1}{\lambda} S_{\lambda,A,B}$. Then, $S_{\lambda,A,B}$ is F-firmly (with $F = \frac{1}{2\lambda}\min(\alpha, \beta)$) non-expansive when one of the following conditions is true:
  1. A is α-firmly non-expansive and B is β-firmly non-expansive;
  2. A is α-strongly monotone and B is β-strongly monotone.
- **Convex optimization model (14)**
  For (14), Assumption 2 is satisfied when one of the following conditions is met:
  1. f and g are strongly convex;
  2. M is full rank, ∇f and ∇g are Lipschitz continuous.

Under Assumption 2, we can estimate a worst-case $O\left(\frac{1}{n}\right)$ convergence rate measured by the iteration complexity for (13) where γ could be greater than 2.

Theorem 5.1 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and Assumption 2 hold, $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (13) with $\gamma \in (0, 2+\frac{2F}{\lambda})$ and v be a root of T. Then we have*

$$||T_\lambda(v_n)||^2 \leq \frac{||v_0 - v||^2}{(2(\lambda + F) - \lambda\gamma)\lambda\gamma(n+1)}.$$

Proof Combining Assumption 2 with Lemma 2.5, we have

$$(2(\lambda + F) - \lambda\gamma)\lambda\gamma||T_\lambda(v_n)||^2 \leq ||v_n - v||^2 - ||v_{n+1} - v||^2.$$

Summing all the inequalities for $i = 0, 1, \cdots, n$, we get

$$\sum_{i=0}^{n} ||T_\lambda(v_i)||^2 \leq \frac{||v_0 - v||^2}{(2(\lambda + F) - \lambda\gamma)\lambda\gamma}.$$

Moreover, it follows from Assumption 2 and Lemma 2.6 that

$$||T_\lambda(v_{n+1})||^2 \leq ||T_\lambda(v_n)||^2 - \left(\frac{2-\gamma}{\gamma} + \frac{2F}{\gamma\lambda}\right)||T_\lambda(v_{n+1}) - T_\lambda(v_n)||^2,$$

so $(||T_\lambda(v_n)||^2)_{n\geq 0}$ is non-increasing when $0 < \gamma < 2 + \frac{2F}{\lambda}$. Finally, we have

$$(n+1)||T_\lambda(v_n)||^2 \leq \sum_{i=0}^{n} ||T_\lambda(v_i)||^2 \leq \frac{||v_0 - v||^2}{(2(\lambda+F) - \lambda\gamma)\lambda\gamma}.$$

The proof is complete. □

Theorem 5.1 indicates that the generalized PPA (13) still holds a worst-case $O\left(\frac{1}{n}\right)$ convergence rate (in term of $\|T_\lambda(v_n)\|^2$) even if $\gamma \in (2, 2 + \frac{2F}{\lambda})$. Furthermore, it is easy to check that the sequence $(||T_\lambda(v_n)||^2)_{n\geq 0}$ fulfils all the requirements of Lemma 2.7. Therefore, we can refine the result in Theorem 5.1 to

$$||T_\lambda(v_n)||^2 = o\left(\frac{1}{n}\right),$$

which means a worst-case $o\left(\frac{1}{n}\right)$ convergence rate of (13) with $\gamma \in (0, 2 + \frac{2F}{\lambda})$.

Moreover, the bound in Theorem 5.1 is minimized when $\gamma = 1 + \frac{F}{\lambda}$. This fact provides a useful strategy of choosing an appropriate γ provides that F is known when implementing the scheme (13). To see if $\gamma = 1 + \frac{F}{\lambda}$ can accelerate convergence, we use **Example 1** again. For this example, $F = \frac{9}{8}$. In Figure 4, we implement the scheme (13) with the initial iterate $(-2, -2)$ and $\lambda = 1$, and compare the convergence with different values of γ. We can see that the choice $\gamma_{opt} = 1 + \frac{F}{\lambda} = 2.125$ outperforms other choices such as 0.5, 1, or 2.

Finally, we would mention that $\gamma \in (0, 2 + \frac{2F}{\lambda})$ is just a sufficient condition to ensure the worst-case $O\left(\frac{1}{n}\right)$ convergence rare of (13) in Theorem 5.1. For some applications, the scheme (13) with $\gamma > 2 + \frac{2F}{\lambda}$ also works very well, even though its convergence rate is not yet provable. Nevertheless, we illustrate this fact by the same example just mentioned. For this example, we have $2 + \frac{2F}{\lambda} = 4.25$. In fact, the scheme (13) converges even for some values of $\gamma > 4.25$, and sometimes values larger than 4.25 are even faster. In Figure 5, we plot the convergence performance for some cases.

6 Linear convergence

In Sections 3-5, we have analyzed worst-case convergence rates for the generalized PPA (13) with various choices of γ under mild conditions. When the operator T has special properties, we expect that the scheme (13) has sharper convergence rates. In this section, we discuss the linear convergence rate of (13) under certain additional assumptions on T. We split the discussion into two cases $\gamma \in (0, 2)$ and $\gamma \in (0, \nu)$ with $\nu \geq 2$. Note that we combine the cases $\gamma = 2$ and $\gamma \in (0, \nu)$ with $\nu > 2$ in the discussion of linear convergence, as they share the same analysis. Again, throughout our discussion we specify the conditions on T in the generic setting (1) to the specific settings (5) and (14).

6.1 Case 1: $\gamma \in (0, 2)$

In this subsection, we focus on the case where $\gamma \in (0, 2)$. Let us make the following assumption.

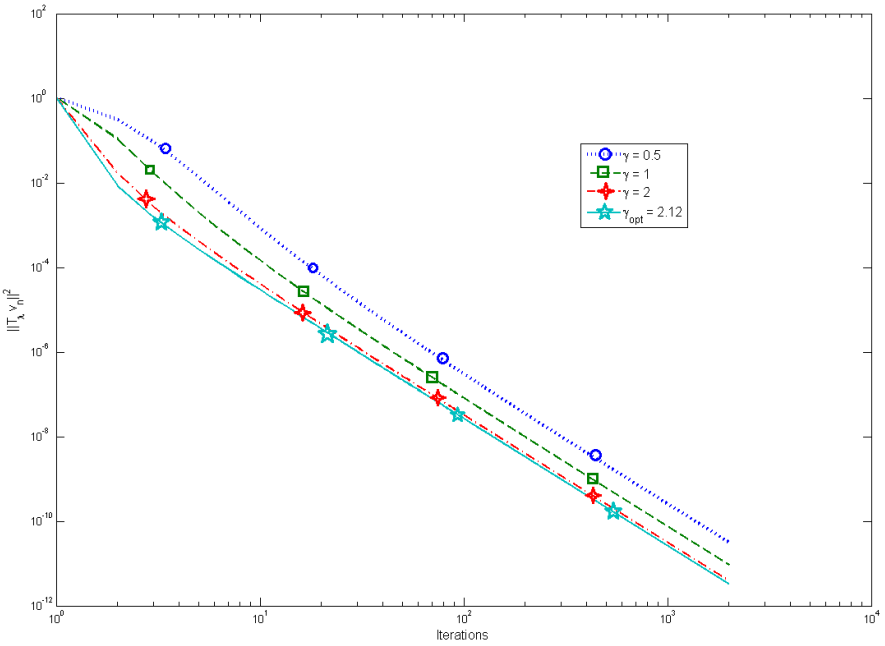

Figure 4: $||T_\lambda(v_n)||^2$ with regard to the iterations in log scale. We compare the choices of $\gamma = 0.5, 1, 2$ with $\gamma_{opt} = 1 + \frac{F}{\lambda}$ to minimize the bound obtained in Theorem 5.1.

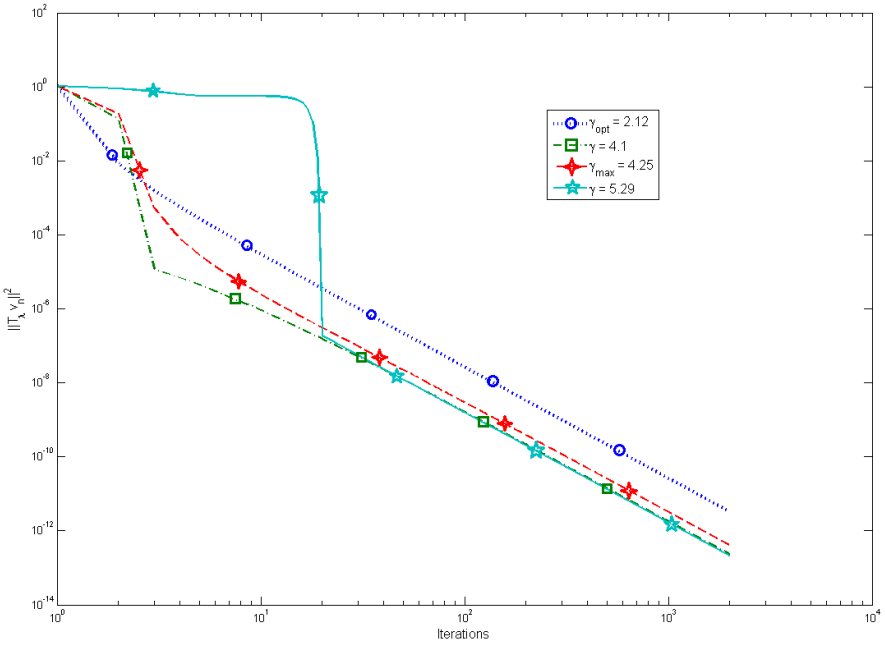

Figure 5: $||T_\lambda(v_n)||^2$ with regard to the iterations in log scale. We compare $\gamma_{opt} = 1 + \frac{F}{\lambda} = 2.125$, $\gamma = 4.1$, $\gamma_{max} = 2 + \frac{2F}{\lambda} = 4.25$ and $\gamma = 5.29$.

Assumption 3 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and α-strongly monotone, i.e.*

$$\langle T(v_1) - T(v_2), v_1 - v_2 \rangle \geq \alpha ||v_1 - v_2||^2,$$

with $\alpha > 0$.

Note that when T is α-strongly monotone, the linear convergence of PPA (4), i.e., the special case of (13) with $\gamma = 1$, has been shown in [30]. Here, we shall show the same convergence rate under the same assumption, but for the generalized PPA scheme (13) with $\gamma \in (0, 2)$. We first prove a proposition of J_λ^T under Assumption 3.

Proposition 6.1 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and Assumption 3 hold. Then we have*

$$||J_\lambda^T(v_n) - J_\lambda^T(v)|| \leq (1 + \alpha\lambda)^{-1} ||v_n - v||.$$

Proof Let $u := J_\lambda^T(v)$. Then we have

$$||v - \bar{v}||^2 = ||u - \bar{u}||^2 + \lambda^2 ||T(u) - T(\bar{u})||^2 + 2\lambda \langle T(u) - T(\bar{u}), u - \bar{u} \rangle.$$

Using Assumption 3, we have

$$\langle T(u) - T(\bar{u}), u - \bar{u} \rangle \geq \alpha \|u - \bar{u}\|^2.$$

So, together with $||T(u) - T(\bar{u})|| \geq \alpha ||u - \bar{u}||$, we obtain

$$\begin{aligned} ||v - \bar{v}||^2 &\geq (1 + \alpha\lambda)^2 ||u - \bar{u}||^2 \\ &= (1 + \alpha\lambda)^2 ||J_\lambda^T(v) - J_\lambda^T(\bar{v})||^2. \end{aligned}$$

The proof is complete. □

Now, we are ready to prove the linear convergence rate for the scheme (13) with $\gamma \in (0, 2)$ under Assumption 3.

Theorem 6.2 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and Assumption 3 hold. Let the sequence $(v_n)_{n \geq 0}$ be generated by (13) with $\gamma \in (0, 2)$. Then, $(v_n)_{n \geq 0}$ converges to a root of T on a linear rate. More specifically, we have*

- *If $0 < \gamma \leq 1 + \frac{1}{1+2\alpha\lambda}$, then*

  $$||v_n - v|| \leq K^n ||v_0 - v||$$

  where $K = \left|1 - \frac{\gamma\lambda\alpha}{1+\lambda\alpha}\right|$; and

- *If $1 + \frac{1}{1+2\alpha\lambda} \leq \gamma < 2$ then*

  $$||v_n - v|| \leq |1 - \gamma|^n ||v_0 - v||.$$

Proof Recall the expression (13). We have

$$\begin{aligned}||v_{n+1}-v||^2 =&(1-\gamma)^2||v_n-v||^2+\gamma^2||J_\lambda^T(v_n)-J_\lambda^T(v)||^2\\ &+2\gamma(1-\gamma)\langle J_\lambda^T(v_n)-J_\lambda^T(v),v_n-v\rangle.\end{aligned}\tag{25}$$

This identity is our basis of proof. To know the sign of the last scalar product, we need to consider two cases separately: $1\leq\gamma<2$ and $0<\gamma<1$.

- The case $1\leq\gamma<2$. For this case, we can rewrite (25) as

$$\begin{aligned}||v_{n+1}-v||^2 =&(1-\gamma)^2||v_n-v||^2+(\gamma^2+2\gamma(1-\gamma))||J_\lambda^T(v_n)-J_\lambda^T(v)||^2\\ &+2\lambda\gamma(1-\gamma)\langle J_\lambda^T(v_n)-J_\lambda^T(v),T_\lambda(v_n)\rangle.\end{aligned}$$

  Since $1-\gamma\leq 0$, using the strong monotonicity of T and Proposition 2.1, we obtain

$$||v_{n+1}-v||^2\leq(1-\gamma)^2||v_n-v||^2+(\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha))||J_\lambda^T(v_n)-J_\lambda^T(v)||^2.\tag{26}$$

  Now we should consider the sign of $(\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha))$ and discuss the following cases individually.

  - If $\gamma\geq 1+\frac{1}{1+2\alpha\lambda}$, then $(\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha))$ is negative. So, we have

$$||v_{n+1}-v||^2\leq(1-\gamma)^2||v_n-v||^2,$$

    which ensures a linear convergence rate (recall that $\gamma<2$).
  - If $\gamma\leq 1+\frac{1}{1+2\alpha\lambda}$, then using Proposition 6.1 we have

$$||v_{n+1}-v||^2\leq K^2||v_n-v||^2,$$

    with $K^2=\left(1-\frac{\gamma\lambda\alpha}{1+\lambda\alpha}\right)^2$. Obviously, when $\gamma<1+\frac{1}{1+2\alpha\lambda}$, it is ensured that $K^2<1$, and a linear convergence rate is ensured.

- The case $0<\gamma\leq 1$. For this case, we have

$$2\lambda\gamma(1-\gamma)\langle J_\lambda^T(v_n)-J_\lambda^T(v),T_\lambda(v_n)\rangle\geq 0.$$

  Using Cauchy-Schwarz inequality, we can show that

$$\begin{aligned}||v_{n+1}-v||^2\leq&(1-\gamma)^2||v_n-v||^2+\gamma^2||J_\lambda^T(v_n)-J_\lambda^T(v)||^2\\ &+2\gamma(1-\gamma)||J_\lambda^T(v_n)-J_\lambda^T(v)||\ ||v_n-v||\\ \leq&K^2||v_n-v||^2\end{aligned}$$

  where $K=\left|1-\frac{\gamma\lambda\alpha}{1+\lambda\alpha}\right|$.

The proof is complete. □

Finally, we specify some interesting cases where Assumption 3 is satisfied and thus the linear convergence of (13) is ensured.

- **Scheme (12)**
  In this case, $T = \frac{1}{\lambda} S_{\lambda,A,B}$. Then $S_{\lambda,A,B}$ is α-strongly monotone (with $\alpha = \frac{1}{2}\min(\lambda\nu, \frac{\beta}{\lambda})$) when one of the following conditions is satisfied:

  1. A is ν-strongly monotone and B is β-firmly non-expansive;
  2. B is ν-strongly monotone and A is β-firmly non-expansive.

  Note that the linear convergence of the special DRSM and PRSM schemes has been shown when B is both Lipschitz and strongly monotone in [22]. Here, in order to show the linear convergence for the general case (13), we need the firm non-expansiveness of at least one operator; this is an assumption stronger than the Lipschitz continuity.

- **Convex optimization model (14)**
  For the model (14), Assumption 3 is satisfied if one of the following conditions is satisfied:

  1. M is full rank, f is convex and smooth and ∇f is Lipschitz continuous, and g is strongly convex;
  2. f is strongly convex, g is convex and smooth and ∇g is Lipschitz continuous.

6.2 Case 2: $\gamma \in (0, \nu)$ with $\nu \geq 2$

In this subsection, we discuss the linear convergence of the generalized PPA (13) where γ is allowed to be greater than 2. Since γ is allowed to be in a wider range, the conditions to ensure the linear convergence of (13) is expected to be stronger. First, we would show that Assumption 3 is not sufficient to ensure the linear convergence of (13) when $\gamma \geq 2$. In fact, recall the inequality (26). If $\gamma \geq 1 + \frac{1}{1+2\alpha\lambda}$, then we have

$$\gamma^2 + 2\gamma(1-\gamma)(1+\lambda\alpha) \leq 0.$$

Then, the inequality (26) does not give us any information about the reduction of the proximity to v, and thus we cannot establish the linear convergence rate for (13) in this case. Recall **Example 2**, which shows that the generalized PPA (13) is divergent with $\gamma = 2$ while linearly convergent with $\gamma \in (0, 2)$ (see Figure 3).

We need the following assumption.

Assumption 4 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and L-Lipschitz continuous, i.e.,*

$$\forall (v_1, v_2) \in \mathbb{H} \times \mathbb{H}, \quad ||T(v_1) - T(v_2)|| \leq L||v_1 - v_2||,$$

with $L > 0$.

With Assumption 4, we can show a useful proposition of J_λ^T.

Proposition 6.3 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone and Assumption 4 hold. Then, we have*

$$||J_\lambda^T(v_n) - J_\lambda^T(v)|| \geq (1 + L\lambda)^{-1}||v_n - v||.$$

Proof Let $u := J_\lambda^T(v)$. Then, it is easy to derive that

$$\begin{aligned}||v-\bar{v}||^2 &= ||u-\bar{u}||^2 + \lambda^2||T(u)-T(\bar{u})||^2 + 2\lambda\langle T(u)-T(\bar{u}), u-\bar{u}\rangle \\ &\leq (1+L\lambda)^2||u-\bar{u}||^2 \\ &= (1+L\lambda)^2||J_\lambda^T(v)-J_\lambda^T(\bar{v})||^2.\end{aligned}$$

The proof is complete. □

Now, under Assumptions 3 and 4 we are ready to establish the linear convergence rate for the generalized PPA (13) where γ could be greater than 2.

Theorem 6.4 *Let $T : \mathbb{H} \to 2^{\mathbb{H}}$ be set-valued maximal monotone, Assumptions 3 and 4 hold. Then, the sequence $(v_n)_{n\geq 0}$ generated by he generalized PPA (13) converges to a root of T on a linear rate when $\gamma \in (0, 2+\frac{2\alpha}{L(2+\lambda L)-2\alpha})$. More specifically, we have*

- *If $0 < \gamma \leq 1+\frac{1}{1+2\alpha\lambda}$, then*

$$||v_n - v|| \leq K^n||v_0 - v||,$$

  where $K = \left|1-\frac{\gamma\lambda\alpha}{1+\lambda\alpha}\right|$.

- *If $1+\frac{1}{1+2\alpha\lambda} \leq \gamma < 2+\frac{2\alpha}{L(2+\lambda L)-2\alpha}$, then*

$$||v_n - v|| \leq K^n||v_0 - v||,$$

  where $K = \left((1-\gamma)^2 + \frac{\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha)}{(1+L\lambda)^2}\right)^{\frac{1}{2}} \in (0,1)$.

Proof The proof of the first case is the same as that of Theorem 6.2. Now, we prove the second case. Since $\gamma > 1$, the inequality (26) holds:

$$||v_{n+1}-v||^2 \leq (1-\gamma)^2||v_n-v||^2 + (\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha))||J_\lambda(v_n)-J_\lambda(v)||^2.$$

When $\gamma \geq 1+\frac{1}{1+2\alpha\lambda}$, then $(\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha))$ is negative. So we use Proposition 6.3 and obtain the following:

$$\begin{aligned}||v_{n+1}-v||^2 &\leq \left((1-\gamma)^2 + \frac{\gamma^2+2\gamma(1-\gamma)(1+\lambda\alpha)}{(1+L\lambda)^2}\right)||v_n-v||^2 \\ &= K^2||v_n-v||^2.\end{aligned}$$

Note $K < 1$ whenever $\gamma < 2+\frac{2\alpha}{L(2+\lambda L)-2\alpha}$. The proof is complete. □

Remark The bound given in Theorem 6.4 is actually tight. This can be easily checked if we take $T = \alpha I$ where α is the strong monotone modulus of T.

Theorem 6.4 also indicates that an informative choice of γ is

$$\gamma = \max(1 + \frac{1}{1+2\alpha\lambda}, 1 + \frac{\alpha}{2(L-\alpha)+L^2\lambda})$$

in order to minimize the basis K in the bounds derived. For the case where $L > 1.5\alpha$, we know that

$$1 + \frac{1}{1+2\alpha\lambda} > 1 + \frac{\alpha}{2(L-\alpha)+L^2\lambda}.$$

Thus it is suggested to choose $\gamma = 1 + \frac{1}{1+2\alpha\lambda}$ which is independent of the Lipschitiz continuous constant L. Nevertheless, this is just a general suggestion to choosing γ. At the same time, if we know enough information of L and α (which does not take place often in practice), the optimal choice of γ might not coincide with this general rule. For instance, if we know that the difference of these two constants L and α is very big, then $\gamma = 1$ is already a good choice. We use the following example for illustration.

Example 3 *Let T be a linear mapping defined on $\mathbb{R}^2$, i.e., $T(v) = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} v$ where α and β being real positive numbers.*

Then Assumptions 3 and 4 are both satisfied. In fact, we have

$$\langle T(v_1) - T(v_2), v_1 - v_2 \rangle = \alpha ||v_1 - v_2||^2$$

and

$$||T(v_1) - T(v_2)|| = L||v_1 - v_2||,$$

with $L = \sqrt{\alpha^2 + \beta^2}$. For this example, the generalized PPA (13) reduces to

$$||v_n|| = \left| (1-\gamma)^2 + \frac{\gamma^2 + 2\gamma(1-\gamma)(1+\lambda\alpha)}{1 + 2\lambda\alpha + \lambda^2 L^2} \right|^n ||v_0||.$$

Then, the optimal choice of γ is obviously $\gamma_{opt} := 1 + \frac{\alpha}{\lambda L^2}$. Therefore, if $\alpha \ll L$, the optimal choice of γ should be very close 1, which is different from the general rule suggested by Theorem 6.4: $1 + \frac{1}{1+2\alpha\lambda}$.

We first choose $\alpha = 1$ and $\beta = 0.5$. Then, $\alpha \approx L$. To implement (13), we choose $\lambda = 1$ and $(1, 1)$ as the starting point. In this case, the choice $\gamma_a := 1 + \frac{\alpha}{2(L-\alpha)+L^2\lambda} = 1.67$ suggested by Theorem 6.4 works very well. In fact, it works almost the same as the real optimal choice $\gamma_{opt} := 1 + \frac{\alpha}{\lambda L^2} = 1.8$. In Figure 6, we plot the convergence performance of the scheme (13) with γ_a and γ_{opt}. For comparison, we also plot the simple choices: $\gamma = 0.5, 1, 2$. All cases exhibit linear convergence. Moreover, we see from this figure that γ_a and γ_{opt} lead to much better numerical performance. Thus, it is verified that the bound given by Theorem 6.4 is useful for us to choose a more suitable γ for the scenario where the strongly monotone modulus and Lipschitz continuous constants are known.

Then, we choose $\alpha = 1$ and $\beta = 3$. For this case, we have $L = \sqrt{10} > \alpha = 1$ (but the difference is not too much). To implement (13), we choose $\lambda = 1$ and (1,1) as the starting point. In this case, the choice $\gamma_a := 1 + \frac{1}{1+2\alpha\lambda} = 1.33$ suggested by Theorem 6.4 works less efficiently than the real optimal choice $\gamma_{opt} := 1 + \frac{\alpha}{\lambda L^2} = 1.1$. In Figure 7, we plot the convergence performance of the scheme (13) with γ_a and γ_{opt}. For comparison, we also plot the simple choices: $\gamma = 0.5, 1, 2$. All cases exhibit linear convergence. Moreover, we

see from this figure that $\gamma = 1$ works almost the same as γ_{opt}. Thus, for the case where α differs significantly from L, the bound given by Theorem 6.4 does not necessarily make us find the optimal choice of γ. But for this case, simply taking $\gamma = 1$ is already good enough.

Finally, we specify the requirements on T in the generic setting (1) to ensure the linear convergence of (13) where γ is allowed to be greater than 2 to the specific settings (5) and (14).

- **Scheme (12)**
  In this case, $T = \frac{1}{\lambda} S_{\lambda,A,B}$. Then, Assumptions 3 and 4 are satisfied if one of the following conditions holds:
  1. A is strongly monotone, B is strongly monotone and firmly non-expansive;
  2. A is strongly monotone and firmly non-expansive, B is strongly monotone;
  3. A is firmly non-expansive, B is strongly monotone and firmly non-expansive;
  4. A is strongly monotone and firmly non-expansive, B is firmly non-expansive.
- **Convex optimization model (14)**
  Assumptions 3 and 4 are satisfied if one of the following conditions holds:
  1. f is strongly convex, g is strongly convex and ∇g is Lipschitz continuous;
  2. M is full rank, f is strongly convex and ∇f is Lipschitz continuous, g is strongly convex;
  3. M is full rank, f is strongly convex and ∇f is Lipschitz continuous, ∇g is Lipschitz continuous;
  4. M is full rank, ∇f is Lipschitz continuous, g is strongly convex and ∇g is Lipschitz continuous.

7 Conclusions

We propose a generalized proximal point algorithm (PPA), in the generic setting of fining a root of a set-valued maximal monotone operator in a Hilbert space. A number of benchmark algorithms in the PDE and optimization literatures are special cases of this generalized PPA scheme. Our main result is to analyze the convergence rate of this generalized PPA scheme—-estimating its worst-case convergence rate measured by the iteration complexity under mild assumptions and its linear convergence rate under stronger assumptions. Some operator splitting methods in the PDE and optimization literatures are particularly treated in our analysis, and some existing convergence rate results in these areas fall into the general result established by this paper. Using the Yosida approximation operator is critical in our analysis. With it, it becomes convenient to measure the accuracy of an iterate to a root of the operator under consideration and thus the analysis for deriving convergence rates in the generic setting becomes doable. This may shed some light on deriving sharper results of the convergence rate for relevant problems.

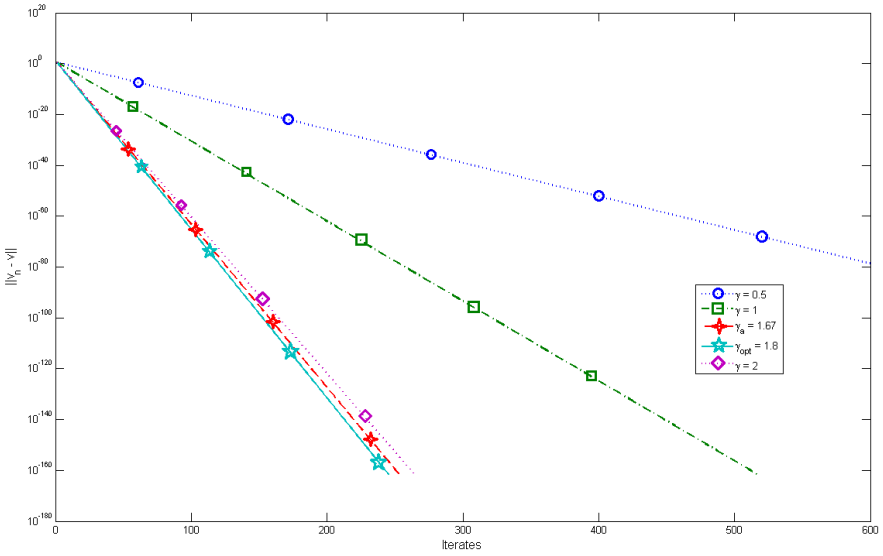

Figure 6: $||v_n - v||$ with regard to the iterations, coordinate in log scale; $\alpha = 1$ and $\beta = 0.5$ for Example 3; $\gamma_a = 1.67$; $\gamma_{opt} = 1.8$; and $\gamma = 0.5, 1, 2$.

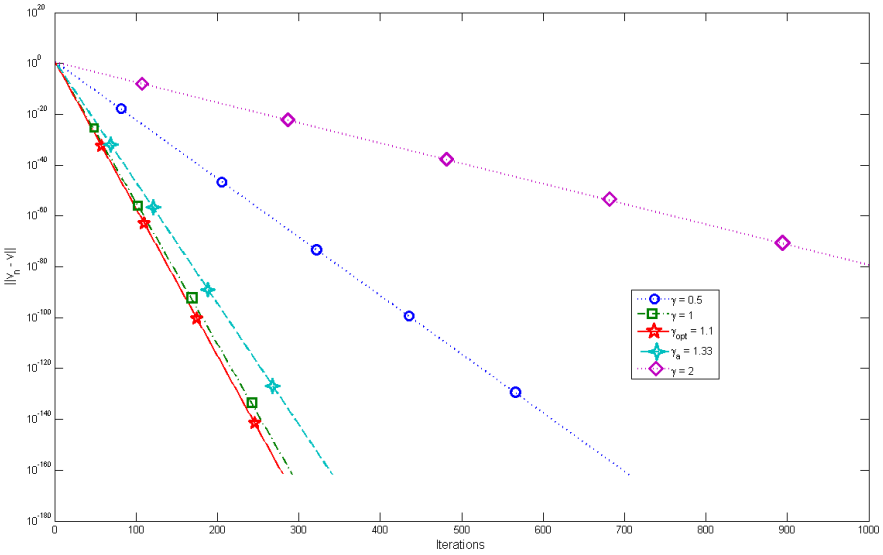

Figure 7: $||v_n - v||$ with regard to the iterations, coordinate in log scale; $\alpha = 1$ and $\beta = 3$ for Example 3; $\gamma_a = 1.33$; $\gamma_{opt} = 1.1$; and $\gamma = 0.5, 1, 2$.

References

[1] E. Blum and W. Oettli, Mathematische Optimierung. Grundlagen und Verfahren. Ökonometrie und Unternehmensforschung, Springer-Verlag, Berlin-Heidelberg-New York, 1975.

[2] D. Boley, *Local linear convergence of ADMM on quadratic or linear programs*, SIAM J. Optim, 23(4) (2013), pp. 2183-2207.

[3] H. Brezis., Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert, North Holland, 1973.

[4] S. Chen, D. Donoho, and M. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33-61.

[5] W. Deng, M. Lai and W. Yin, *On the $o(1/k)$ convergence and parallelization of the alternating direction method of multipliers*, manuscript, 2013.

[6] W. Deng and W. Yin, *On the global and linear convergence of the generalized alternating direction method of multipliers*, manuscript, 2012.

[7] J. Douglas and H. H. Rachford, *On the numerical solution of the heat conduction problem in* 2 *and* 3 *space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421-439.

[8] J. Eckstein, *Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results*, manuscript, 2012.

[9] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal points algorithm for maximal monotone operators*, Math. Program., 55 (1992), pp. 293–318.

[10] M. Fukushima and H. Mine, *A generalized proximal point algorithm for certain nonconvex minimization problems*, Intern. J. Sys. Sci. 12 (1981), pp. 989-1000.

[11] R. Glowinski and A. Marrocco, *Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problèmes non linéaires*, R.A.I.R.O., R2 (1975), pp. 41-76.

[12] R. Glowinski, T. Kärkkäinen and K. Majava, *On the convergence of operator-splitting methods*, in Numerical Methods for Scienfic computing, Variational Problems and Applications, edited by Y. Kuznetsov, P. Neittanmaki and O. Pironneau, Barcelona, 2003.

[13] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Optim., 29(2) (1991), pp. 403-419.

[14] D. R. Han and B. S. He, *A new accuracy criterion for approxiamte poximal point algorithms*, J. Math. Anal. Appl., 263 (2001), 343–354.

[15] D. R. Han and X. M. Yuan, *Local linear convergence of the alternating direction method of multipliers for quadratic programs*, SIAM J. Num. Anal., 51(6) (2013), pp. 3446-3457.

[16] B. S. He, M. Tao and X. M. Yuan, *Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming*, Math. Oper. Res., under revision.

[17] B. S. He and X. M. Yuan, *On the O(1/n) convergence rate of Douglas-Rachford*

alternating direction method, SIAM J. Num. Anal., 50 (2012), pp. 700-709.

[18] B. S. He and X. M. Yuan, *On nonergodic convergence rate of Douglas-Rachford alternating direction method of multipliers*, submission, 2012.

[19] B. S. He and X. M. Yuan, *On convergence rate of the Douglas-Rachford operator splitting method*, Math. Program, under revision.

[20] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 302-320.

[21] M. Hong and Z. Luo, *On the linear convergence of the alternating direction method of multipliers*, manuscript, 2012.

[22] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Num. Anal., 16 (1979), pp. 964-979.

[23] B. Martinet, *Regularisation, d'inéquations variationelles par approximations successives*, Rev. Francaise d'Inform. Recherche Oper., 4 (1970), pp. 154-159.

[24] J. J. Moreau, *Proximité et dualit 'e dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273-299.

[25] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, New York, 1983.

[26] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$,Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543-547.

[27] D. H. Peaceman and H. H. Rachford, *The numerical solution of parabolic elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28-41.

[28] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, In R. Fletcher, editor, Optimization. Academic Press, 1969.

[29] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1970.

[30] R.T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Con. Optim., 14 (1976), pp. 877-898.

[31] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.

[32] R. Shefi and M. Teboulle, *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization*, SIAM J. Optim, to appear.

[33] X. Q. Zhang, M. Burger, X. Bresson and S. Osher, *Bregmanized nonlocal regularization for deconvolution and sparse reconstruction*, SIAM J. Imaging Sci., 3(3) (2010), pp. 253-276.