A Generalized Proximal Point Algorithm and its Convergence Rate

Etienne Corman * **and Xiaoming Yuan**†

July 4, 2013

Abstract

We propose a generalized proximal point algorithm (PPA), in the generic setting of finding a zero point of a maximal monotone operator. In addition to the classical PPA, a number of benchmark operator splitting methods in PDE and optimization literatures such as the Douglas-Rachford splitting method, Peaceman-Rachford splitting method, alternating direction method of multipliers, generalized alternating direction method of multipliers and split inexact Uzawa method can be retrieved by this generalized PPA scheme. We establish the convergence rate of this generalized PPA scheme under different conditions, including estimating the worst-case iteration complexity under mild assumptions and deriving the 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

*Department of Mathematics, University of Toulouse, 31042 Toulouse cedex 9, France. Email: corman@etud.insa-toulouse.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: 203712.

1 Introduction

Let $\mathbb{H}$ be a Hilbert space with the scalar product $\langle\cdot\rangle$ and norm $\|\cdot\|$. Let $T:\mathbb{H}\to 2^{\mathbb{H}}$ be a point-to-set 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 [7, 20, 21, 26] is a classical scheme. Starting from $v^0\in\mathbb{H}$, the iterative scheme of PPA reads

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

where $\lambda>0$ is a proximal parameter. In fact, PPA plays a significant theoretical and algorithmic role in many areas of scientific computing such as optimization, PDE and image processing; and a number of celebrated algorithms turn out to be specific cases of PPA when the operator T is specified accordingly. Such examples include the augmented Lagrangian method [17, 24] (see [26]), the Douglas-Rachford splitting method in [5, 19] (see [6]), the split inexact Uzawa method [29], and so on. To illustrate the role of the proximal parameter λ, let us take the example of convex minimization problem in $\mathbb{R}^n$, i.e., T represents the subdifferential of a nonsmooth convex function $\theta(v)$:

$$\arg\min_{v\in R^n}\{\theta(v)\}. \tag{3}$$

Then, PPA approaches to a solution of (3) by solving the following subproblems recursively:

$$\arg\min_{v\in R^n}\{\theta(v)+\frac{1}{2\lambda}\|v-v^k\|\}, \tag{4}$$

where the term $\frac{1}{2\lambda}\|v-v^k\|$ often plays a regularization role for some applications in image processing or ill-posed inverse problems. If λ takes larger values, the term $\frac{1}{2\lambda}\|v-v^k\|$ tends to play a less important role in the objective and the PPA subproblem (4) tends to be more accurate to the original problem (1). This means the solution of (4) should be closer to the solution of (1)—-a faster convergence is thus implied provided that the subproblem (4) is assumed to be solved exactly. In fact, super linear convergence of PPA has been derived in [26] under the condition that $\lambda\to\infty$. On the other hand, if λ takes smaller values, the term $\frac{1}{2\lambda}\|v-v^k\|$ plays a more influential role in the objective and the solution of (4) should be easier (meaning a subproblem with a better condition number if an ill-posed problem is considered) and closer to x^k (the solution of (4) approaches to v^k if $\lambda\to 0$)—-this represents the fact that the subproblem (4) is easier while the whole sequence should converge to a solution of (3) on a slower rate if λ takes smaller values. Therefore, to implement PPA a generic (conceptual) rule for choosing λ is to make a balance between the ease of the subporblems and the speed of convergence. But we do not discuss how to choose λ in this paper.

Recall (see e.g. [25]) the resolvent operator of a set-valued monotone operator:

$$J_\lambda^T=(I+\lambda T)^{-1}. \tag{5}$$

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

$$v^{k+1} = J_\lambda^T(v^k), \tag{6}$$

i.e., at each iteration it requires an exact estimate of the resolvent operator J_λ^T [1]. Note the resolvent J_λ^T of a monotone operator is always single-valued.

The problem (1) is an abstract model in 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

$$0 \in A(v) + B(v). \tag{7}$$

A special case of (7) is the least-squares problem with l_1 regularization:

$$\arg\min_{v\in\mathbb{R}^n}\{\|v\|_1 + \frac{\tau}{2}\|Sv - t\|_2^2\}, \tag{8}$$

where $S \in \mathbb{R}^{m\times n}$ is a matrix, $t \in \mathbb{R}^n$, $\tau > 0$; $\|\cdot\|_1$ and $\|\cdot\|_2$ represent the regular l_1 and l_2 norms, respectively. To recover (8) by (7), just take $A(v) = \partial(\|v\|_1)$ and $B(v) = \tau S^T(Sv - t)$ where $\partial(\cdot)$ denotes the subdifferential of a convex but nonsmooth function. A very useful application of (8) is when $m \ll n$. Then (8) can be explained as finding a sparse vector satisfying the equations $Sv = t$ —- the well-known basis pursuit problem, LASSO problem and many others are thus recovered.

To apply PPA to solve (7), as we have analyzed, it requires to estimate exactly the resolvent operator J_λ^{A+B} at each iteration; the iterative scheme is

$$v^{k+1} = J_\lambda^{A+B}(v^k). \tag{9}$$

We, however, have many applications that the estimate of J_λ^A and J_λ^B is much easier than that of J_λ^{A+B}. In fact, note that

$$\arg\min\{\|v\|_1 + \frac{1}{2\lambda}\|v - v^k\|^2\}$$

has a closed form solution given by the soft-shrinkage operator (see e.g. [29]); then the model (8) is such an example. Thus, for solving (7), we are more interested in designing algorithms requiring only the estimate of J_λ^A and J_λ^B, rather than using the original PPA scheme (6) straightforwardly (requiring the estimate of J_λ^T) —- the so-called operator splitting methods are thus named. Two most successful methods along this line are the Douglas-Rachford splitting method (DRSM) in [19]

$$u_{n+1} \in J_\lambda^B\left(J_\lambda^A(I - \lambda B) + \lambda B\right)u_n; \tag{10}$$

[1] In practice, it is often too restrictive to estimate J_λ^T exactly. Thus, inexact versions of PPA which require only to solve (6) approximately subject to certain inexactness criteria have been intensively studied in the literature, see e.g.[26] for a foundation work. Later we will also discuss the case of solving (6) approximately in Section 3.2.

and the Peaceman-Rachford splitting method (PRSM) in [19]

$$u_{n+1} \in J_\lambda^B(I - \lambda A)J_\lambda^A(I - \lambda B)u_n. \tag{11}$$

Since A and B could be set-valued, it is necessary to explain how to read the schemes (10) and (11). For a given u_0, we choose b_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). The algorithms (10) and (11) become respectively

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

and

$$v_{n+1} = (2J_\lambda^A - I)(2J_\lambda^B - I)v_n. \tag{13}$$

Obviously, these two schemes (12) and (13) can be retrieved by the scheme

$$v_{n+1} = v_n + \gamma\left(J_\lambda^A(2J_\lambda^B - I)v_n - J_\lambda^B v_n\right) \tag{14}$$

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

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

$$v_{n+1} = \gamma J_\lambda^T(v_n) + (1 - \gamma)v_n, \tag{15}$$

where $\lambda > 0$ and $\gamma > 0$. The original PPA (6) is obviously a special case of (15) with $\gamma = 1$. One more motivation of studying this generalized PPA scheme is that the formula (14) 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 [6]),

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

we have

$$G_{\lambda,A,B} = J_1^{S_{\lambda,A,B}}.$$

Therefore, the scheme (14) is a special case of (15) with $T = S_{\lambda,A,B}$, $\gamma = 1$ and $\lambda \equiv 1$. Note that it has been studied in [6] that $S_{\lambda,A,B}$ defined above is maximal monotone when A and B are both maximal monotone [2]. Aiming at extending the scheme (14), we are thus interested in the generalized PPA scheme (15).

[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 $\langle x - \bar{x}, y - \bar{y}\rangle = \lambda\langle a - \bar{a}, v - \bar{v}\rangle + \lambda\langle b - \bar{b}, u - \bar{u}\rangle \geq 0$.

Let us specify the generalized PPA scheme (15) for the particular context of convex minimization problem:

$$\min_{x\in\mathbb{R}^n} f(x) + g(Mx), \tag{16}$$

where $f : \mathbb{R}^n \mapsto]-\infty, +\infty]$ and $g : \mathbb{R}^m \mapsto]-\infty, +\infty]$ are closed, convex and proper functions; and $M \in \mathbb{R}^{m\times n}$. Obviously, the model (8) is a special case of (16). Another very important application of (16) is the image restoration and reconstruction models with total variation regularization. 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 the discrete gradient operator (or total variation operator, see [27]). The dual of (16) is

$$\min_{p\in\mathbb{R}^m} \{f^*(-M^T p) + g^*(p)\}, \tag{17}$$

where "*" denotes the Fenchel transform, see, e.g., [25]. Thus, (17) is a special case of (7) with $A = \partial\left(f^* \circ (-M^T)\right)$ and $B = \partial(g^*)$. Applying the generalized PPA scheme (15), we thus obtain the following iterative scheme for solving (16):

$$\begin{cases} \textbf{Step 1.} & x_{k+1} = \arg\min\limits_{x\in\mathbb{R}^n} \quad f(x) + \langle p_k, Mx\rangle + \frac{\lambda}{2}||Mx - y_k||^2 \\ \textbf{Step 2.} & y_{k+1} = \arg\min\limits_{y\in\mathbb{R}^m} \quad g(y) - \langle p_k, y\rangle + \frac{\lambda}{2}||\gamma M x_{k+1} + (1-\gamma)y_k - y||^2 \\ \textbf{Step 3.} & p_{k+1} = p_k + \lambda(\gamma M x_{k+1} + (1-\gamma)y_k - y_{k+1}) \end{cases} \tag{18}$$

Note when $\gamma = 1$, the classical alternating direction method of multipliers (ADMM) in [8, 9] is recovered; when $\gamma \in (0, 2)$, the generalized ADMM in [6] is recovered; and (18) with $\gamma = 2$ reduces to the application of PRSM to (17). The sequence $(v_k)_{k\geq 0}$ generated by (14) can be linked to the sequences $(p_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ generated by (18) by the relation

$$v_k = p_k + \lambda y_k. \tag{19}$$

For more details of the relationship between DRSM and ADMM, see [6].

Now, we explain the allowable range for γ in (15). As we just show, 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). Here, we note that

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

where the inequality is due to the firm non-expansiveness of J_λ^T (see e.g. [25]) and the fact that v is a zero point of T. Therefore, a big difference between the cases $\gamma \in (0, 2)$

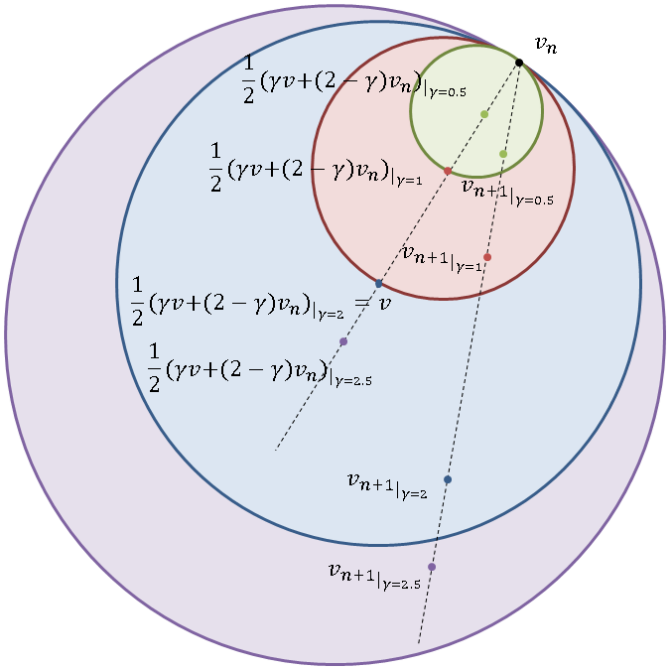

Figure 1: The firm non-expansiveness of the J_λ^T implies that the iterate v_{n+1} lies in a ball whose size depends on v and γ.

and $\gamma = 2$ occurs: If $\gamma = 2$, we only have $\|v_{n+1}-v\| \leq \|v_n-v\|$ and this might prevent the iterative sequence from being strictly contractive with respect to the set of zero points 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 (15) 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 study for this case. The necessity of studying the case where $\gamma > 2$ can be seen by 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 zero point of this* T.

Let the scheme (15) be implemented with $\lambda = 1$ and the starting point $(-2, -2)$. We plot the iterative procedure of (15) with different values of γ in Figure 2, and we can see for this example that $\gamma > 2$ do can accelerate the convergence.

Our main purpose is to analyze the convergence rate for the generalized PPA scheme (15) with a generic T and $\gamma > 0$. As we have mentioned, the value of γ results in different iterative performance of the scheme (15). We thus will discuss three cases individually: $\gamma \in (0, 2)$, $\gamma = 2$ and $\gamma > 2$. We first estimate the worst-case iteration complexity for (15). Note that as [22, 23], the convergence rate of an iterative scheme can be measured by the worst-case iteration complexity. More specifically, we shall show

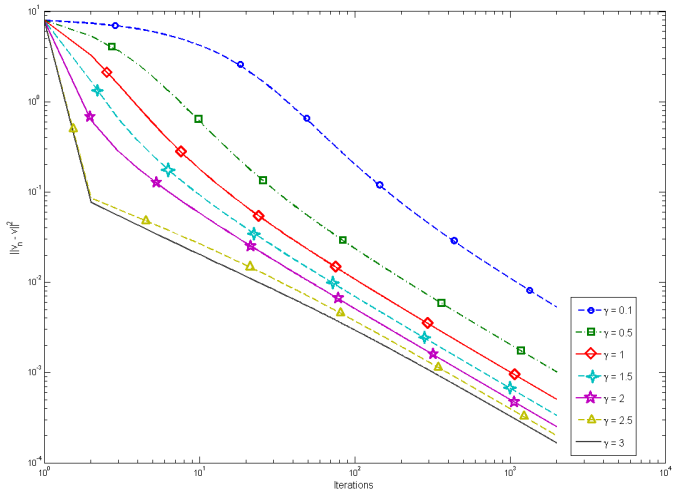

Figure 2: $||v_n - v||^2$ with regard to the iterations in log scale. We compare various choices of γ. For $\gamma > 2$ the algorithm is still convergent and more efficient than choices $\gamma \in (0, 2]$.

that after n iterations of (15), we can find an approximate root of T with the accuracy $O(1/n)$; or equivalently, the scheme (15) requires at most $O(1/\epsilon)$ iterations to achieve an approximate root of T with the accuracy ϵ. Then, we shall discuss under what kinds of conditions the scheme (15) converges to a root of T on a linear rate.

A brief review on existing convergence rate results of some special cases of the scheme (15). For the special optimization model (16), the ADMM scheme which is a special case of (18) with $\gamma = 1$ was shown to have a worst-case $O(1/n)$ convergence rate in [14] (ergodic sense) and in [15] (nonergodic sense). Recently, the linear convergence of ADMM for some special cases and under some stronger conditions has been discussed in [2, 4, 13], and the linear convergence of an extended ADMM scheme can be found in [18]. A more comprehensive convergence rate analysis for decomposition methods was presented most recently in [28]. In [11], the author established a worst-case $O(1/n)$ convergence rate of (4), i.e., the application of PPA to (3) or the special case of (15 with $T = \partial\theta$ and $\gamma = 1$), and an accelerated version with a worst-case $O(1/n^2)$ convergence rate. For the generic DRSM scheme (10), a worst-case $O(1/n)$ convergence rate can be found in [16]. The linear convergence of the DRSM scheme (10) and PRSM scheme (11) were discussed in [19] under some further conditions on A and B. We also refer to [10], where the convergence rates of DRSM and PRSM schemes were discussed for special cases of (7).

The rest of this paper is organized as follows. In Section 2, we summarize some useful preliminaries and prove some basic properties for further discussion. Then, we discuss the convergence rate of the generalized PPA scheme (15) in Sections 3-5 for different cases of γ. In Section 6, we discuss the linear convergence rate of (15). Finally, we make some conclusions in Section 7.

2 Preliminaries

In this section, we provide some preliminaries and prove some basic propositions 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 $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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 TJ_λ^T (which could be set-valued) in the following proposition.

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

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

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 $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, let J_λ^T be the resolvent operator and T_λ be the Yosida approximation operator of T. Then, $\forall v_1, v_2 \in \mathcal{H}$, 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 $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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.*

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

$$\langle T_\lambda(v_1) - T_\lambda(v_2), J_\lambda^T(v_1) - J_\lambda^T(v_2)\rangle \geq 0.$$

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. □

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

Proposition 2.4 *For a set-valued maximal monotone operator $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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

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

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 generated by PPA is to calculate $\|T(v_n)\|$. Here, we use $\|T_\lambda(v_n)\|$ rather than $\|T(v_n)\|$ as an indicator of accuracy for iterates generated by PPA. In fact, we can show that $\|T(v_n)\|$ and $\|T_\lambda(v_n)\|$ are comparable in the measurement of accuracy for PPA's iterates. 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

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

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 $\{v_n\}$ generated by the proposed generalized PPA scheme (15), and they will be used often later.

First of all, by using the Yosida approximation operator, we can rewrite the scheme (15) as

$$v_{n+1} = v_n - \gamma\lambda T_\lambda(v_n). \tag{20}$$

We first compare the difference of proximity to a root of T (denoted by v) for two consecutive iterates v_{n+1} and v_n generated by (15).

Lemma 2.5 *For a set-valued maximal monotone operator $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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 (15) 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 (20), 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, 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 $\|v_n - v\|^2$ is non-increasing if $\gamma \in (0, 2]$. Moreover, the sequence $\{v_n\}$ is contractive with respect to the set of roots of T when $\gamma \in (0, 2)$. Based on this fact, the convergence of the generalized PPA scheme (15) 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 $\{\|T_\lambda(v_n)\|^2\}$ where $\{v_n\}$ is generated by the generalized PPA scheme (15). Recall that we have shown that $\|T_\lambda(v)\|^2$ can be used to measure the accuracy of v to a root of T.

Lemma 2.6 *For a set-valued maximal monotone operator $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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 (15) 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 (15), 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 thus 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 $\gamma \in (0, 2]$.

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

Now, we start to estimate the convergence rate of the generalized PPA scheme (15). We first focus on estimating its worst-case iteration complexity without additional assumption on the mapping T. As we have mentioned, the techniques to derive the worst-case iteration complexity for different values of γ are different (e.g., it follows from Lemma 2.6 that the sequence $\{v_n\}$ generated by (15) is contractive with respect to the set of roots of T, while this property does not hold for other cases of γ). Thus we discuss the cases $\gamma \in (0, 2)$, $\gamma = 2$ and $\gamma > 2$ in this and the next two sections, respectively.

3.1 Convergence rate with exact estimate of the resolvent operator

As the starting point, we first assume that the resolvent operator J_λ^T can be estimated accurately at any point and thus the exact estimate of J_λ^T is available to implement the scheme (15). For this case, we can estimate the iteration complexity of (15) in terms of $\|T_\lambda(v_n)\|^2$, as shown in the following theorem.

Theorem 3.1 *For a set-valued maximal monotone operator $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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 (15) 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 $i = 0, 1, 2, \cdots, n-1$, we get

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

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 $\|T_\lambda(v)\|^2$ can be used to measure the accuracy of v to a root of T (see Proposition 2.4). Thus, Theorem 3.1 shows that after n iterations, the iterate generated by (15) with $\gamma \in (0,2)$ produces an approximate root of T with the accuracy of $O(1/n)$. Thus, a worst-case $O(1/n)$ iteration complexity is established for (15) with $\gamma \in (0,2)$. Note this is an extended result of the work [15] which is only for the special case of the scheme (18) with $\gamma = 1$.

3.2 Convergence rate with an inexact estimate of the resolvent operator

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 inspired the seminal work of approximate PPA in [26]. Let us consider the inexact version of the generalized PPA scheme (15) with $\gamma \in (0,2)$:

$$\begin{aligned} &v_{n+1} = \gamma w_n + (1-\gamma)v_n \\ &\text{s.t} \quad ||w_n - J_\lambda^T(v_n)|| \leq \epsilon_n. \end{aligned} \tag{21}$$

In (21), w_n represents an inexact estimate of J_λ^T and ϵ_n denotes the accuracy of estimating J_λ^T at the point v_n. Choosing different ϵ_n leads to different versions of the generalized PPA, and there are of course many ways to design appropriate inexact criteria to control the accuracy ϵ_n. In fact, some well-studied criteria in PPA literature (e.g. [12, 26]) can be used here for (21). Also, there are alternative criteria which do not involve $J_\lambda^T(v_n)$ and thus can be implemented directly. However, for the purpose of succinctness and clearer exposition of our main result, we just discuss the most fundamental inexact criteria in (21) analogous to those in [26] for PPA. This is a conceptual one but also the foundation of other criteria in PPA literature.

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

$$\forall n \geq 0, \quad \epsilon_n = O\left(\frac{1}{(n+1)^\alpha}\right), \quad \alpha > 1, \tag{22}$$

and then estimate the convergence rate for the inexact version of generalized PPA (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 (22), obviously it holds that

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

Now, we derive a worst-case iteration complexity for the scheme (21) with $\gamma \in (0, 2)$ and (22) in the following theorem.

Theorem 3.2 *For a set-valued maximal monotone operator $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, 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 (21) with $\gamma \in (0, 2)$ 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 $\bar{v}_n$ the exact iterate (15)

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

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 (21) 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

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

If we sum up this inequality over $i = 0, 1, \cdots, n-1$, we get

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

We also have

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

Using the Young inequality on the last term, 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

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

Combining this equation from p to n 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

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

As ϵ_n satisfies the requirement (22), there exists a constant $K > 0$ such that

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

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 $O(1/n)$. A worst-case $O(1/n)$ convergence rate is thus established for the inexact version of generalized PPA (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 (15) admits a sublinear convergence rate.

4 Case 2: $\gamma = 2$

Now, we discuss the convergence rate of the proposed generalized PPA (15) 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 set of roots of T. This makes the analysis of convergence much more challenging. Therefore, in this section we first analyze the convergence issues for this case and then derive its convergence rate under one additional assumption on T. Note we only discuss the exact version (15) where J_λ^T is assumed to be estimated exactly, and skip the discussion on the inexact version (21) which allows for inexact estimate of J_λ^T.

4.1 Convergence issues

In Lemma 2.5, we show that the sequence generated by the generalized PPA scheme (15) is strictly contractive with respect to the set of roots of T if $\gamma \in (0, 2)$. Thus, convergence for this case can be easily established, see [6, 16] 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

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

It means that 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 (15) with $\gamma = 2$ stay the same distance away from a root of T and it never converges.

In the next, let us take a closer look at this scalar product. Below we list two useful information about this scalar product.

1. **A characterization of** $\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle = 0$.

   As $T_\lambda(v) \in T(J_\lambda^T(v))$ (see Proposition 2.1), if we set $u_n = J_\lambda^T(v_n)$, it exists $t_n \in T(u_n)$ such that $t_n = T_\lambda(v_n)$. So we can rewrite (23) as

   $$\langle t_n, u_n - v\rangle.$$

   Therefore, $\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle = 0$ when t_n is orthogonal to the vector $u_n - v$.

2. **A fact about** $J_\lambda^T(v_n)$ **when** $\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle = 0$.

   If $\langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle = 0$, then we

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

   Thus, $J_\lambda^T(v_n)$ should lie on the border of the ball with center $\frac{v_n+v}{2}$ and radius $\frac{v_n-v}{2}$.

We illustrate these two facts by Figure 3.

To compare the convergence difference of 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 (15), all iterates generated by (15) stay the same distance away from $(0, 0)$. However, if $\gamma \in (0, 2)$, the sequence generated by (15) converges to $(0, 0)$ (in fact, the convergent rate is linear).

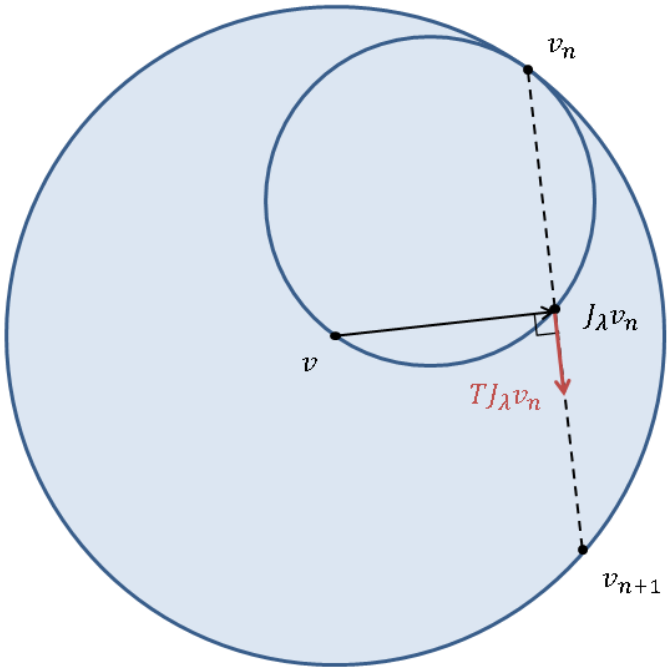

Figure 3: Illustration of a case where v_n and v_{n+1} are at the same distance to a root v and $\langle t_n, u_n - v\rangle = 0$ when $\gamma = 2$.

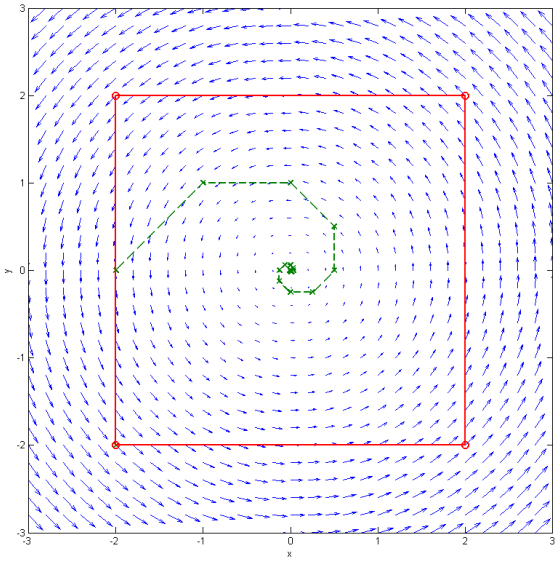

Figure 4: $T : v \in \mathbb{R}^2 \to u \in \mathbb{R}^2$ is defined as $\{y^1 = -x^2, y^2 = x^1\}$. Starting point: $(-2, -2)$; and $\lambda = 1$ in (15). The arrows represent the vector field of T . The case $\gamma = 2$ (in plain line) does not converge and the sequence has four cluster points (corner points); and the case $\gamma = 1$ (in dash line) converges to $(0, 0)$.

In Figure 4, we plot the difference of convergence for the cases where $\gamma = 2$ and $\gamma \in (0, 2)$ in (15).

Considering the analysis before, we thus need to pose certain additional assumptions on T in order to ensure the convergence of the generalized PPA (15) with $\gamma = 2$. Our assumption is as follows.

Assumption 1 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone and v be a root of T. For any bounded sequence $(v_n)_{n\geq 0}$ and $\lambda > 0$, we have*

$$\lim_{n\to+\infty} \langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle = 0 \quad \Rightarrow \quad 0 \in T\left(\lim_{n\to+\infty} J_\lambda^T.(v_n)\right)$$

Assumption 1 ensures that the sequence $(J_\lambda^T(x_n))_{n\geq 0}$ converges to a root of T. Note that it is possible that $\lim_{n\to+\infty} J_\lambda^T(v_n) = v$ while $\lim_{n\to+\infty} v_n \nrightarrow v$ when T is set-valued at the point v.

Remark We give some scenarios in the settings of (7) and (16) where Assumption 1 is satisfied. As mentioned, we are particularly interested in the case where $T = S_{\lambda,A,B}$ and the convex optimization model (16). (To understand the following, the monotonicity of $S_{\lambda,A,B}$ is helpful.)

- For all bounded sequence $(v_n)_{n\geq 0}$, T satisfies:

  $$(t_n \in T(v_n), t \in T(v)), \qquad \lim_{n\to+\infty} \langle t_n - t, v_n - v\rangle = 0 \quad \Rightarrow \quad \lim_{n\to+\infty} v_n = v. \quad (24)$$

  For the particular case where $T = S_{\lambda,A,B}$, this condition means that one of the operators A *and* B is single-valued and it satisfies (24). This is exactly the assumption made in [19] to ensure the convergence of the PRSM scheme (11).

- T is strictly monotone, i.e.

  $$\forall t_v \in T(v), t_u \in T(u), \quad \langle t_v - t_u, v - u\rangle = 0 \quad \Rightarrow \quad v = u.$$

  - **Scheme (14)**
    This case is $T = S_{\nu,A,B}$. It means that either A *or* B is single-valued and strictly monotone.
  - **Convex optimization model (16)**
    Either f *or* g is differentiable and strictly convex.

- T is firmly non-expansive i.e.

  $$\forall v, u \in \mathcal{H}, \quad \langle T(v) - T(u), x - y\rangle \geq ||T(v) - T(u)||^2.$$

  - **Scheme (14)**
    This case is $T = S_{\nu,A,B}$, and it means A and B are both firmly non-expansive; *or* A and B are both strongly monotone.
  - **Convex optimization model (16)**
    f and g are both strongly convex; *or* M has full rank, both f and g are differentiable and ∇f and ∇g are both Lipschitz.

4.2 Convergence rate

In this subsection, we estimate a worst-case convergence rate for the generalized PPA (15) with $\gamma = 2$ under Assumption 1. Different from the case where $\gamma \in (0, 2)$, the convergence rate to be derived is in ergodic sense.

We first show a characterization of Assumption 1.

Proposition 4.1 *Assumption 1 is equivalent to*

$$\forall v \in \mathcal{H}, \quad \langle T_\lambda(v), J_\lambda^T(v) - J_\lambda^T(v_0)\rangle = 0 \quad \Rightarrow \quad 0 \in T(J_\lambda^T(v_0)).$$

Proof The proof consists of two parts.

- The implication "Proposition 4.1 $\Rightarrow$ Assumption 1" is obvious.
- For "Assumption 1 $\Rightarrow$ Proposition 4.1", we prove it by contradiction. Assume that it exists $\bar{v}$ is such that $0 \notin T(J_\lambda^T(\bar{v}))$ (it implies that $J_\lambda^T(\bar{v}) \neq J_\lambda^T(v)$) and

  $$\langle T_\lambda(\bar{v}), J_\lambda^T(\bar{v}) - J_\lambda^T(v)\rangle = 0.$$

  We want to show that it exists a bounded sequence $(v_n)_{n\geq 0}$ such that

  $$0 \notin T\left(\lim_{n\to+\infty} J_\lambda^T(v_n)\right)$$

  and

  $$\lim_{n\to+\infty} \langle T_\lambda(v_n), J_\lambda^T(v_n) - J_\lambda^T(v)\rangle = 0.$$

  We set $\bar{u} = J_\lambda^T(\bar{v})$ and $\bar{t} = T_\lambda(\bar{v}) \in T(\bar{u})$ such that $\langle t, \bar{u} - v\rangle = 0$, and set $u = kv + (1-k)\bar{u}$ for $0 < k < 1$ and $t \in Tu$. Since T is monotone, we have

  $$\begin{aligned} 0 &\leq \langle t - \bar{t}, u - \bar{u}\rangle; \\ \langle t, \bar{u} - u\rangle &\leq \langle \bar{t}, \bar{u} - u\rangle; \\ 0 \leq \langle t, u - v\rangle \frac{k}{1-k} &\leq k\langle \bar{t}, \bar{u} - v\rangle = 0. \end{aligned}$$

  Therefore, for any $k \in (0, 1)$ we have $\langle t, u - v\rangle = 0$. As $0 \notin T(u)$ and the set of roots of T is closed (since T is maximal monotone), it exists $k_0 \in (0, 1)$ such that $0 \notin T(k_0 v + (1-k_0)\bar{u})$. Hence, we can define our sequence $(v_n)_{n\geq 0}$ as

  $$J_\lambda^T(v_n) = \begin{cases} \bar{u} & \text{if } n \text{ even} \\ k_0 v + (1-k_0)\bar{u} & \text{if } n \text{ odd} \end{cases}$$

  This sequence is bounded and not convergent. This contradiction verifies "Assumption 1 $\Rightarrow$ Proposition 4.1".

The proof is complete. □

Proposition 4.1 tells us that 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 iterations. That is, let

$$\delta_n := \frac{1}{n}\sum_{i=0}^{n-1}\langle T_\lambda(v_i), J_\lambda^T(v_i) - J_\lambda^T(v)\rangle, \tag{25}$$

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

Theorem 4.2 *For a set-valued maximal monotone operator $T :\in \mathcal{H} \to 2^{\mathcal{H}}$, let δ_n be defined in (25). Let $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (15) with $\gamma = 2$ and v be a root of T. For any $\lambda > 0$, we have*

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

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

Theorem 4.2 shows a worst-case $O(1/n)$ convergence rate for the generalized PPA scheme (15) with $\gamma = 2$ in ergodic sense.

5 Case 3: $\gamma \geq 2$

As we have shown, for some cases the generalized PPA scheme (15) with $\gamma > 2$ can converge faster. It is thus necessary to discuss the convergence rate of (15) which allows γ to be greater than 2. Again, we skip the discussion for the inexact version (21). To the best of our knowledge, even for the special cases (18) for solving the convex optimization model (16), there is no rigorous analysis of its convergence rate when $\gamma > 2$.

Due to the more relaxed restriction onto the allowable 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 $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone and L-firmly non-expansive, i.e.,*

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

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

- **Scheme (14)**
  In this case, $T = S_{\lambda,A,B}$. Then, $S_{\lambda,A,B}$ is L-firmly (with $L = \frac{1}{2\lambda}\min(\alpha, \beta)$) non-expansive when one of the following conditions is true:

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

- **Convex optimization model (16)**
  For (16), Assumption 2 is satisfied when one of the following conditions is met:

  1. f and g are strongly convex; or
  2. M has full rank, ∇f and ∇g are Lipschitz.

Under Assumption 2, we can estimate a worst-case iteration complexity for (15) where γ could be greater than 2.

Theorem 5.1 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone and L-firmly non-expansive, $(v_n)_{n\geq 0}$ be the sequence generated by the generalized PPA scheme (15) with $\gamma \in (0, 2 + \frac{2L}{\lambda})$ and v be a root of T. Then we have*

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

Proof Combining Assumption 2 with Lemma 2.5, we have

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

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

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

Moreover, since $||T_\lambda(v_n)||^2$ is nonincreasing when $0 < \gamma < 2 + \frac{2L}{\lambda}$, by using Assumption 2 in Proposition 2.6 we conclude

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

The proof is complete. □

Theorem 5.1 indicates that the generalized PPA (15) still holds a worst-case $O(1/n)$ convergence rate (in term of $\|T_\lambda(v_n)\|^2$) even if $\gamma \in (2, 2 + \frac{2L}{\lambda})$. Moreover, the bound in Theorem 5.1 is minimized when $\gamma = 1 + \frac{L}{\lambda}$. This fact provides a useful strategy of choosing an appropriate γ provides that L is known when implementing the scheme (15). To see if $\gamma = 1 + \frac{L}{\lambda}$ can accelerate convergence, we take again **Example 1**. For this example, $L = \frac{8}{9}$. In Figure 5, we implement the scheme (15) with the initial iterate $(-2, -2)$ and $\lambda = 1$, and compare the convergence with different values of γ. We can see that the choice $\gamma_{opt} = 1 + \frac{L}{\lambda} = 2.12$ outperforms other choices such as 0.5, 1, or 2.

Finally, we would mention that $\gamma \in (0, 2 + \frac{2L}{\lambda})$ is just a sufficient condition to ensure the worst-case $O(1/n)$ convergence rare of (15) in Theorem 5.1. For some applications,

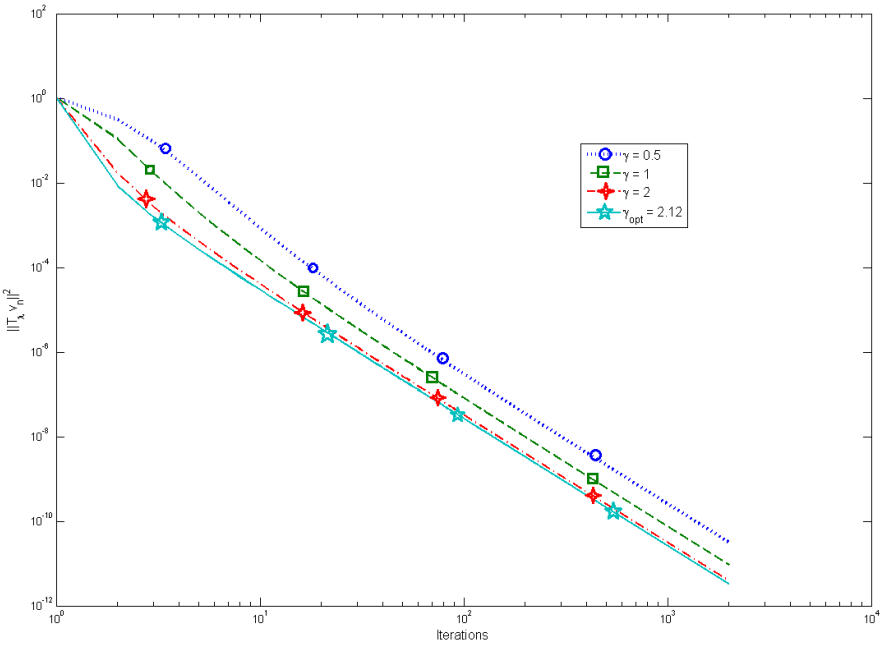

Figure 5: $||T_\lambda(v_n)||^2$ with regard to the iterations in log scale. We compare the choices of $\gamma = 0.5, 1, 2$ with $\gamma_{opt} = 1 + \frac{L}{\lambda}$ to minimize the bound obtained in Theorem 5.1.

the scheme (15) with $\gamma > 2 + \frac{2L}{\lambda}$ 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 $2 + \frac{2L}{\lambda} = 4.25$. But the scheme (15) converges even for some values of $\gamma > 4.25$, and sometimes values larger than 4.25 are even faster. In Figure 6, we plot the convergence performance for some cases.

6 Linear convergence

In Sections 3-5, we have analyzed the convergence rate for the generalized PPA (15) with various choices of γ in terms of the worst-case iteration complexity under mild conditions. When the operator T has special properties, we expect that the scheme (15) has sharper convergence rate. In this section, we discuss the linear convergence rate of (15) under certain additional assumptions on T. We split the discussion into two cases $\gamma \in (0, 2)$ and $\gamma \geq 2$. Note we combine the cases $\gamma = 2$ and $\gamma > 2$ 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 (7) and (16).

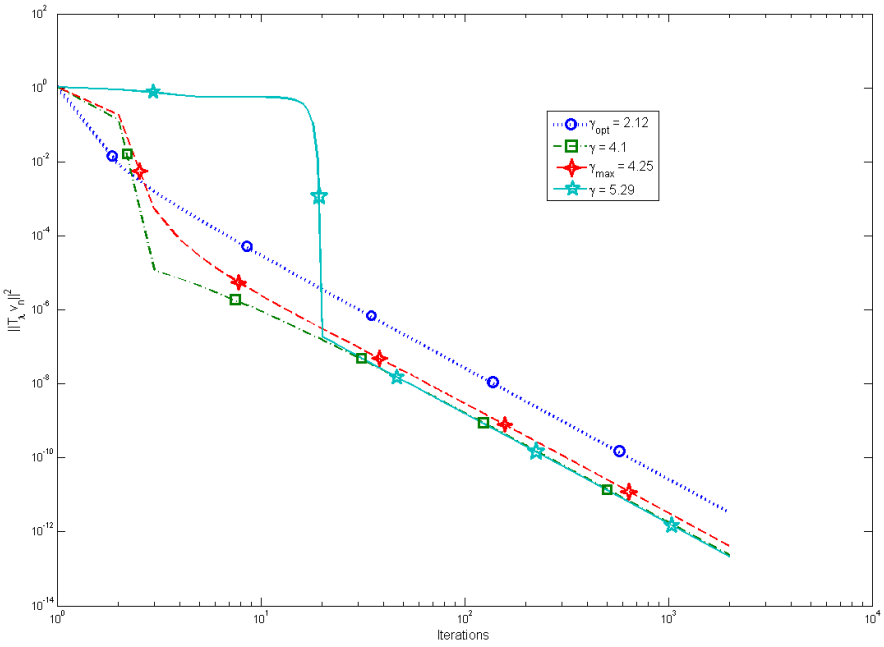

Figure 6: $||T_\lambda(v_n)||^2$ with regard to the iterations in log scale. We compare $\gamma_{opt} = 1 + \frac{L}{\lambda} = 2.12$, $\gamma = 4.1$, $\gamma_{max} = 2 + \frac{2L}{\lambda} = 4.25$ and $\gamma = 5.29$.

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.

Assumption 3 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ 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 when T is α-strongly monotone, the linear convergence of PPA (6), i.e., the special case of (15) with $\gamma = 1$, has been shown in [26]. Here, we shall show the same convergence rate under the same assumption, but for the generalized PPA scheme (15) with $\gamma \in (0, 2)$. We first prove a proposition of J_λ^T under Assumption 3.

Proposition 6.1 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone and α-strongly monotone. Then we have*

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

Proof Let be $u := J_\lambda^T(v)$, then

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

The proof is complete. □

Now, we are ready to prove the linear convergence rate for the scheme (15) with $\gamma \in (0,2)$ under Assumption 3.

Theorem 6.2 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone and α-strongly monotone. Let the sequence $(v_n)_{n\geq 0}$ be generated by (15) 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 (20). We have

$$\begin{aligned}||v_{n+1}-v||^2 =& (1-\gamma)^2||v_n - v||^2 + \gamma^2||J_\lambda^T(v_n) - J_\lambda^T(v)||^2 \quad (26)\\ &+ 2\gamma(1-\gamma)\langle J_\lambda^T(v_n) - J_\lambda^T(v), v_n - v\rangle \quad (27)\end{aligned}$$

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 $0 < \gamma < 1$.

- $1 \leq \gamma < 2$
  We can rewrite (26) 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

$$||v_{n+1}-v||^2 \leq (1-\gamma)^2||v_n - v||^2 + (\gamma^2 + 2\gamma(1-\gamma)(1+\lambda\alpha))||J_\lambda^T(v_n) - J_\lambda^T(v)||^2. \quad (28)$$

  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.

- $\gamma \leq 1$
  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 is ensured by implementing the generalized PPA (15).

- **Scheme (14)**
  In this case, $T = S_{\nu,A,B}$. Then $S_{\nu,A,B}$ 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; or
  2. B is ν-strongly monotone and A is β-firmly non-expansive.

  Note the linear convergence of the special DRSM and PRSM schemes has been shown when B is both Lipschitz and strongly monotone in [19]. Here, in order to show the linear convergence for the general case (15), we need the firm non-expansiveness of at least one operator; this is an assumption stronger than Lipschitz monotonicity.

- **Convex optimization model (16)**
  For the model (16), 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 strongly convex; or
  2. f is strongly convex, g is convex and smooth and ∇g Lipschitz continuous.

6.2 The case $\gamma \geq 2$

In this subsection, we discuss the linear convergence of the generalized PPA (15) where γ is allowed to be greater than 2. Since γ is allowed to be in a wider range, the conditions to ensure linear convergence of (15) is expected to be stronger. First, we would show that Assumption 3 is not sufficient to ensure the linear convergence of (15) when $\gamma \geq 2$. In fact, recall the inequality (28). 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 (28) does not give us any information about the reduction of proximity to v, and thus we cannot establish the linear convergence rate for (15) in this case. Recall also **Example 2**, which shows that the generalized PPA (15) is divergent with $\gamma = 2$ while linearly convergent with $\gamma \in (0, 2)$ (see Figure 4).

We need the following assumption.

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

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

with $M > 0$.

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

Proposition 6.3 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone and M-Lipschitz continuous. Then, we have*

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

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

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

The proof is complete. □

Now, under Assumptions 3 and 4 we are ready to establish the linear convergence rate for the generalized PPA (15) where γ could be greater than 2.

Theorem 6.4 *Let $T : \mathcal{H} \to 2^{\mathcal{H}}$ be set-valued maximal monotone, α-strongly monotone and M-Lipschitz continuous. Then, the sequence $(v_n)_{n\geq 0}$ generated by he generalized PPA (15) converges to a root of T on a linear rate when $\gamma \in (0, 2 + \frac{2\alpha}{M(2+\lambda M)-2\alpha})$. 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 $1 + \frac{1}{1+2\alpha\lambda} \leq \gamma < 2 + \frac{2\alpha}{M(2+\lambda M)-2\alpha}$, then*

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

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

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 (28) 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:

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

Note $K < 1$ whenever $\gamma < 2 + \frac{2\alpha}{M(2+\lambda M)-2\alpha}$. 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

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

in order to minimize the basis K in the bounds derived. For the case where $M > \sqrt{2}\alpha$, we know that

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

Thus it is suggested to choose $\gamma = 1 + \frac{1}{1+2\alpha\lambda}$ which is independent of the Lipschitiz continuous constant M. Nevertheless, this is just a general suggestion to choosing γ. At the same time, if we know enough information of M 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 M 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

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

with $M = \sqrt{\alpha^2 + \beta^2}$. For this example, the generalized PPA (15) reduces to

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

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

We first choose $\alpha = 1$ and $\beta = 0.5$. Then, $\alpha \approx M$. To implement (15), we choose $\lambda = 1$ and $(1, 1)$ as the starting point. In this case, the choice $\gamma_a := 1 + \frac{\alpha}{2(M-\alpha)+M^2\lambda} = 1.67$ suggested by Theorem 6.4 works very well. In fact, it works almost the same as the real optimal choice $\gamma_{opt} := 1 + \frac{\alpha}{\lambda M^2} = 1.8$. In Figure 7, we plot the convergence performance of the scheme (15) with γ_a and γ_{opt}. For comparison, we also plot some naive 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 $M = \sqrt{10} > \alpha = 1$ (but the difference is not too much). To implement (15), we choose $\lambda = X$ and (X,X) 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 $\gamma_{opt} := 1 + \frac{\alpha}{\lambda M^2} = 1.1$. In Figure 8, we plot the convergence performance of the scheme (15) with γ_a and γ_{opt}. For comparison, we also plot some naive 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 M, 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 (15) where γ is allowed to be greater than 2 to the specific settings (7) and (16).

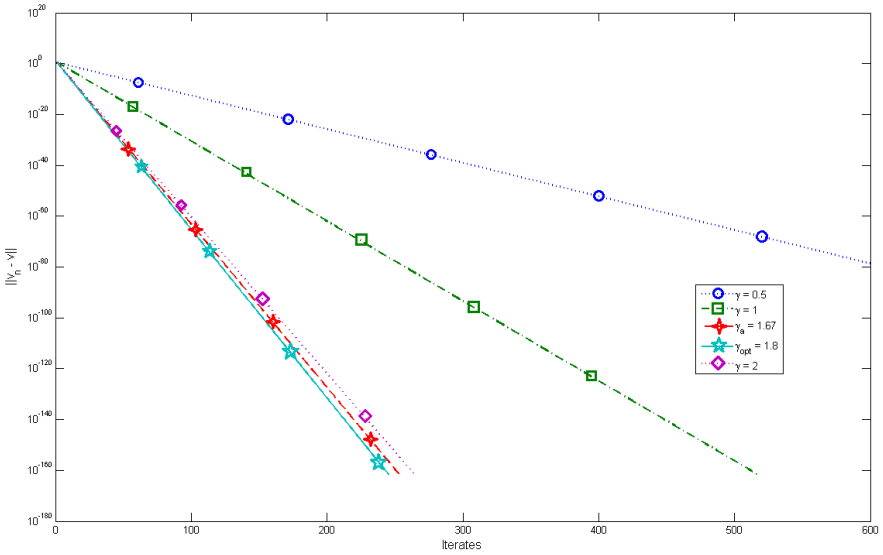

Figure 7: $||v_n - v||$ with regard to the iterations, coordinate in log scale; $\alpha = 1$ and $\beta = 0.5$ for Example 3; $\gamma_a = 1.67$; $\gamma_{opt} = 1.8$; and $\gamma = 0.5, 1, 2$.

- **Scheme (14)**
  In this case, $T = S_{\nu,A,B}$. 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; and
  4. A is strongly monotone and firmly non-expansive, B is firmly non-expansive.

- **Convex optimization model (16)**
  Assumptions 3 and 4 are satisfied if one of the following conditions holds:

  1. f strongly convex, g strongly convex and ∇g Lipschitz;
  2. M full rank, f strongly convex and ∇f Lipschitz, g strongly convex;
  3. M full rank, f strongly convex and ∇f Lipschitz, ∇g Lipschitz; and
  4. M full rank, ∇f Lipschitz, g strongly convex and ∇g Lipschitz.

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

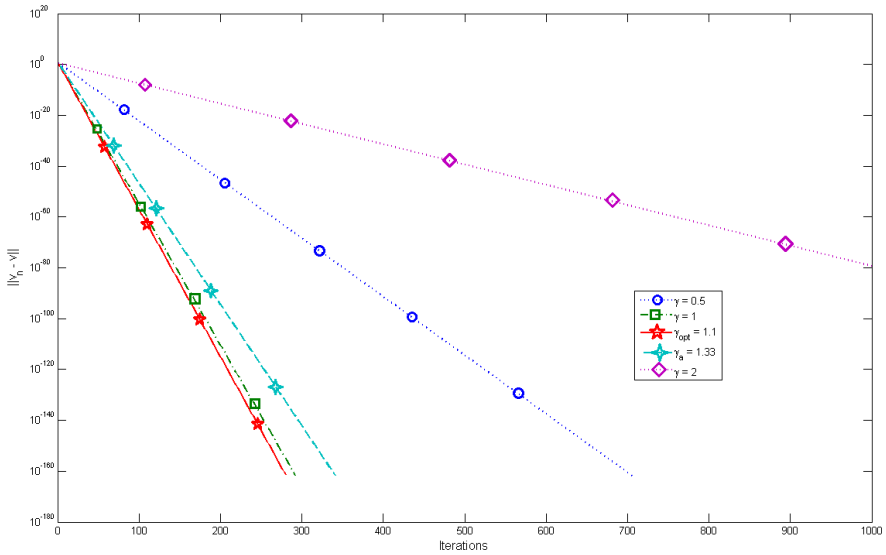

Figure 8: $||v_n - v||$ with regard to the iterations, coordinate in log scale; $\alpha = 1$ and $\beta = 3$ for Example 3; $\gamma_a = 1.33$; $\gamma_{opt} = 1.1$; and $\gamma = 0.5, 1, 2$.

benchmark algorithms in PDE and optimization literatures are included by this generalized PPA scheme. Our main result is to analyze the convergence rate of this generalized PPA scheme—-estimating its worst-case iteration complexity under mild assumptions and linear convergence rate under stronger assumptions. Operator splitting methods in PDE and optimization literatures are particularly treated in our analysis, and some existing results of convergence rate in these areas fall into the general result established by this paper. The use of Yosida approximation operator turns out to be 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 convergence rate for relevant problems. For the challenging case where $\gamma = 2$ in the generalized PPA scheme, we can only derive a worst-case iteration complexity in ergodic sense. Whether or not it is possible to derive such a convergence rate in nonergodic sense deserves further research in the future.

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*, manuscript, 2012.

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

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

[5] 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.

[6] 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.

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

[8] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite-element approximations*, Comput. Math. Appli., 2 (1976), pp. 17-40.

[9] 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.

[10] 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.

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

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

[13] D. R. Han and X. M. Yuan, *Local linear convergence of the alternating direction method of multipliers for quadratic programs*, SIAM J. Num. Anal., under revision.

[14] 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.

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

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

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

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

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

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

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

[22] 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.

[23] 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.

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

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

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

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

[28] R. Shefi, and M. Teboulle, *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization, manuscript, 2013.*

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