Exact Optimal Accelerated Complexity for Fixed-Point Iterations

Jisun Park [1] **Ernest K. Ryu** [1]

Abstract

Despite the broad use of fixed-point iterations throughout applied mathematics, the optimal convergence rate of general fixed-point problems with nonexpansive nonlinear operators has not been established. This work presents an acceleration mechanism for fixed-point iterations with nonexpansive operators, contractive operators, and nonexpansive operators satisfying a Hölder-type growth condition. We then provide matching complexity lower bounds to establish the exact optimality of the acceleration mechanisms in the nonexpansive and contractive setups. Finally, we provide experiments with CT imaging, optimal transport, and decentralized optimization to demonstrate the practical effectiveness of the acceleration mechanism.

1. Introduction

The fixed-point iteration with $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ computes

$$x_{k+1} = \mathbb{T}x_k$$

for $k = 0, 1, \dots$ with some starting point $x_0 \in \mathbb{R}^n$. The general rubric of formulating solutions of a problem at hand as fixed points of an operator and then performing the fixed-point iterations is ubiquitous throughout applied mathematics, science, engineering, and machine learning.

Surprisingly, however, the iteration complexity of the abstract fixed-point iteration has not been thoroughly studied. This stands in sharp contrast with the literature on convex optimization algorithms, where convergence rates and matching lower bounds are carefully studied.

In this paper, we establish the exact optimal complexity of fixed-point iterations by providing an accelerated method and a matching complexity lower bound. The acceleration is based on a Halpern mechanism, which follows the footsteps of Lieder (2021); Kim (2021); Yoon & Ryu (2021), and is distinct from Nesterov's acceleration.

[1]Department of Mathematical Sciences, Seoul National University. Correspondence to: Ernest K. Ryu <ernestryu@snu.ac.kr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1.1. Preliminaries and notations

We review standard definitions and set up the notation.

Monotone and set-valued operators. We follow standard notation of Bauschke & Combettes (2017); Ryu & Yin (2020). For the underlying space, consider $\mathbb{R}^n$ with standard inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, although our results can be extended to infinite-dimensional Hilbert spaces.

We say $\mathbb{A}$ is an operator on $\mathbb{R}^n$ and write $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ if $\mathbb{A}$ maps a point in $\mathbb{R}^n$ to a subset of $\mathbb{R}^n$. For notational simplicity, also write $\mathbb{A}x = \mathbb{A}(x)$. Write $\operatorname{Gra}\mathbb{A} = \{(x,u) \mid u \in \mathbb{A}x\}$ for the graph of $\mathbb{A}$. Write $\mathbb{I}\colon \mathbb{R}^n \to \mathbb{R}^n$ for the identity operator. We say $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is monotone if

$$\langle \mathbb{A}x - \mathbb{A}y, x - y\rangle \geq 0, \qquad \forall x, y \in \mathbb{R}^n,$$

i.e., if $\langle u - v, x - y\rangle \geq 0$ for all $u \in \mathbb{A}x$ and $v \in \mathbb{A}y$. For $\mu \in (0, \infty)$, say $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is μ-strongly monotone if

$$\langle \mathbb{A}x - \mathbb{A}y, x - y\rangle \geq \mu\|x - y\|^2, \qquad \forall x, y \in \mathbb{R}^n.$$

An operator $\mathbb{A}$ is maximally monotone if there is no other monotone $\mathbb{B}$ such that $\operatorname{Gra}\mathbb{A} \subset \operatorname{Gra}\mathbb{B}$ properly, and is maximally μ-strongly monotone if there is no other μ-strongly monotone $\mathbb{B}$ such that $\operatorname{Gra}\mathbb{A} \subset \operatorname{Gra}\mathbb{B}$ properly.

For $L \in (0, \infty)$, single-valued operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz if

$$\|\mathbb{T}x - \mathbb{T}y\| \leq L\|x - y\|, \qquad \forall x, y \in \mathbb{R}^n.$$

$\mathbb{T}$ is contractive if it is L-Lipschitz with $L < 1$ and nonexpansive if it is 1-Lipschitz. For $\theta \in (0, 1)$, an operator $\mathbb{S}\colon \mathbb{R}^n \to \mathbb{R}^n$ is θ-averaged if $\mathbb{S} = (1-\theta)\mathbb{I} + \theta\mathbb{T}$ and a nonexpansive operator $\mathbb{T}$.

Write $\mathbb{J}_{\mathbb{A}} = (\mathbb{I} + \mathbb{A})^{-1}$ for the resolvent of $\mathbb{A}$, and $\mathbb{R}_{\mathbb{A}} = 2\mathbb{J}_{\mathbb{A}} - \mathbb{I}$ for the reflected resolvent of $\mathbb{A}$. When $\mathbb{A}$ is maximal monotone, it is well known that $\mathbb{J}_{\mathbb{A}}$ is single-valued with $\operatorname{dom}\mathbb{J}_{\mathbb{A}} = \mathbb{R}^n$, $\mathbb{R}_{\mathbb{A}}$ is a nonexpansive operator, and $\mathbb{J}_{\mathbb{A}} = \frac{1}{2}\mathbb{I} + \frac{1}{2}\mathbb{R}_{\mathbb{A}}$ is $1/2$-averaged.

We say $x_\star \in \mathbb{R}^n$ is a zero of $\mathbb{A}$ if $0 \in \mathbb{A}x_\star$. We say $y_\star$ is a fixed-point of $\mathbb{T}$ if $\mathbb{T}y_\star = y_\star$. Write $\operatorname{Zer}\mathbb{A}$ for the set of zeros of $\mathbb{A}$ and $\operatorname{Fix}\mathbb{T}$ for the set of all fixed-points of $\mathbb{T}$. For any $x \in \mathbb{R}^n$ such that $x = \mathbb{J}_{\mathbb{A}}y$ for some $y \in \mathbb{R}^n$, define $\tilde{\mathbb{A}}x = y - \mathbb{J}_{\mathbb{A}}y$ as the resolvent residual of $\mathbb{A}$ at x. Note that $\tilde{\mathbb{A}}x \in \mathbb{A}x$. For any $y \in \mathbb{R}^n$, define $y - \mathbb{T}y$ as the fixed-point residual of $\mathbb{T}$ at y.

Fixed-point iterations. There is a long and rich history of iterative methods for finding a fixed point of an operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ (Rhoades, 1991; Brezinski, 2000; Rhoades & Saliga, 2001; Berinde & Takens, 2007). In this work, we consider the following three: the *Picard iteration*

$$y_{k+1} = \mathbb{T}y_k,$$

the *Krasnosel'skiĭ–Mann iteration (KM iteration)*

$$y_{k+1} = \lambda_{k+1}y_k + (1-\lambda_{k+1})\mathbb{T}y_k,$$

and the *Halpern iteration*

$$y_{k+1} = \lambda_{k+1}y_0 + (1-\lambda_{k+1})\mathbb{T}y_k,$$

where $y_0 \in \mathbb{R}^n$ is an initial point and $\{\lambda_k\}_{k\in\mathbb{N}} \subset (0,1)$. Under suitable assumptions, the $\{y_k\}_{k\in\mathbb{N}}$ sequence of these iterations converges to a fixed point of $\mathbb{T}$.

1.2. Prior work

Fixed-point iterations. Picard iteration's convergence with a contractive operator was established by Banach's fixed-point theorem (Banach, 1922). What we refer to as the Krasnosel'skiĭ–Mann iteration is a generalization of the setups by Krasnosel'skiĭ (1955) and Mann (1953). Its convergence with general nonexpansive operators is due to Martinet (1972). The iteration of Halpern (1967) converges for the wider choice of parameter λ_k (including $\lambda_k = \frac{1}{k+1}$) due to Wittmann (1992). Halpern iteration is later generalized to the sequential averaging method (Xu, 2004). Ishikawa iteration (Ishikawa, 1976) is an iteration with two sequences updated in an alternating manner. Anderson acceleration (Anderson, 1965) is another acceleration scheme for fixed-point iterations, and it has recently attracted significant interest (Walker & Ni, 2011; Scieur et al., 2020; Barré et al., 2020; Zhang et al., 2020; Bertrand & Massias, 2021). A number of inertial fixed-point iterations have also been proposed to accelerate fixed-point iterations (Maingé, 2008; Dong et al., 2018; Shehu, 2018; Reich et al., 2021). Our presented method is optimal (in the sense made precise by the theorems) when compared these prior non-stochastic fixed-point iterations.

Convergence rates of fixed-point iterations. The squared fixed-point residual $\|y_k - \mathbb{T}y_k\|^2$ is the error measure for fixed-point problems that we focus on. Its convergence to 0 (without a specified rate) is referred to as asymptotic regularity (Browder & Petryshyn, 1966), and it has been established for KM (Ishikawa, 1976; Borwein et al., 1992) and Halpern (Wittmann, 1992; Xu, 2002).

The convergence rate of the KM iteration in terms of $\|y_k - \mathbb{T}y_k\|^2$ was shown to exhibit $\mathcal{O}(1/k)$-rate (Cominetti et al., 2014; Liang et al., 2016; Bravo & Cominetti, 2018) and $o(1/k)$-rate (Baillon & Bruck, 1992; Davis & Yin, 2016; Matsushita, 2017) under various setups. In addition, Borwein et al. (2017); Lin & Xu (2021) studied the convergence rate of the distance to solution under additional bounded Hölder regularity assumption.

For the convergence rate of the Halpern iteration in terms of $\|y_k - \mathbb{T}y_k\|^2$, Leustean (2007) proved a $\mathcal{O}(1/(\log k)^2)$-rate and later Kohlenbach (2011) improved this to a $\mathcal{O}(1/k)$-rate. Sabach & Shtern (2017) first proved the $\mathcal{O}(1/k^2)$-rate of Halpern iteration, and this rate has been improved in its constant by a factor of 16 by Lieder (2021).

Monotone inclusions and splitting methods. As we soon establish in Section 2, monotone operators are intimately connected to fixed-point iterations. Splitting methods such as forward-backward splitting (FBS) (Bruck Jr, 1977; Passty, 1979), augmented Lagrangian method (Hestenes, 1969; Powell, 1969), Douglas–Rachford splitting (DRS) (Peaceman & Rachford, 1955; Douglas & Rachford, 1956; Lions & Mercier, 1979), alternating direction method of multiplier (ADMM) (Gabay & Mercier, 1976), Davis–Yin splitting (DYS) (Davis & Yin, 2017), (PDHG) (Chambolle & Pock, 2011), and Condat–Vũ (Condat, 2013; Vũ, 2013) are all fixed-point iterations with respect to specific nonexpansive operators. Therefore, an acceleration of the abstract fixed-point iteration is applicable to the broad range of splitting methods for monotone inclusions.

Acceleration. Since the seminal work by Nesterov (1983) on accelerating gradient methods convex minimization problems, much work as been dedicated to algorithms with faster accelerated rates. Gradient descent (Cauchy, 1847) can be accelerated in terms of function value suboptimality for smooth convex minimization problems (Nesterov, 1983; Kim & Fessler, 2016a), smooth strongly convex minimization problems (Nesterov, 2004; Van Scoy et al., 2018; Park et al., 2021; Taylor & Drori, 2021; Salim et al., 2022), and convex composite minimization problems (Güler, 1992; Beck & Teboulle, 2009). Recently, accelerated methods for reducing the squared gradient magnitude for smooth convex minimization (Kim & Fessler, 2021; Lee et al., 2021) and smooth convex-concave minimax optimization (Diakonikolas & Wang, 2021; Yoon & Ryu, 2021) were presented.

Recently, it was discovered that acceleration is also possible in solving monotone inclusions. The accelerated proximal point method (APPM) (Kim, 2021) provides an accelerated $\mathcal{O}(1/k^2)$-rate of $\|\tilde{\mathbb{A}}x_k\|^2$ compared to the $\mathcal{O}(1/k)$-rate of proximal point method (PPM) (Martinet, 1970; Gu & Yang, 2020) for monotone inclusions. Maingé (2021) improved this rate to $o(1/k^2)$ rate with another accelerated variant of proximal point method called CRIPA-S.

Complexity lower bound. Under the information-based complexity framework (Nemirovski, 1992), complexity lower bound on first-order methods for convex optimization has been thoroughly studied (Nesterov, 2004; Drori, 2017; Drori & Shamir, 2020; Carmon et al., 2020; 2021; Drori & Taylor, 2022). When a complexity lower bound matches an algorithms' guarantee, it establishes optimality of the algorithm (Nemirovski, 1992; Drori & Teboulle, 2016; Kim & Fessler, 2016a; Taylor & Drori, 2021; Yoon & Ryu, 2021; Salim et al., 2022). In the fixed-point theory literature, Diakonikolas (2020) provided the lower bound result for the rate of $\langle \mathbb{A}x_k, x_k - x_\star\rangle$ for variational inequalities with Lipschitz, monotone operator. Colao & Marino (2021) showed $\Omega\left(1/k^{2-\frac{2}{q}}\right)$ lower bound on $\|y_k - y_\star\|^2$ for Halpern iterations in q-uniformly smooth Banach spaces. Recently, there has been work establishing complexity lower bounds for the more restrictive "1-SCLI" class of algorithms (Arjevani et al., 2016). The class of 1-SCLI fixed-point iterations includes the KM iteration but not Halpern. Up-to-constant optimality of the KM iteration among 1-SCLI algorithms was proved with the $\Omega(1/k)$ lower bound by Diakonikolas & Wang (2021).

There also has been recent work on lower bounds for the general class of algorithms (not just 1-SCLI) for fixed-point problems. Contreras & Cominetti (2021) established a $\Omega(1/k^2)$ lower bound on the fixed-point residual for the general Mann iteration, which includes the KM and Halpern iterations, in Banach spaces. Our $\Omega(1/k^2)$ lower bound of Section 4 is more general than the result of Contreras & Cominetti (2021) as it applies to all deterministic algorithms, not just Mann iterations. Diakonikolas & Wang (2021) established a $\Omega(1/k^2)$ lower bound on the squared operator norm for algorithms finding zeros of cocoercive operators, which are equivalent to methods finding fixed points of nonexpansive operators. Our lower bound of Section 4 improves upon this result (by a constant of about 80) and establishes exact optimality of the methods in Section 3.

Acceleration with restart. Restarting is a technique that allows one to render a standard accelerated method to be adaptive to the local structure (Nemirovski & Nesterov, 1985; Nesterov, 2013; Lin & Xiao, 2014; O'Donoghue & Candes, 2015; Kim & Fessler, 2018; Fercoq & Qu, 2019; Roulet & d'Aspremont, 2020; Ito & Fukuda, 2021). Our method of Section 5 was inspired specifically by the restarting scheme of Roulet & d'Aspremont (2020).

Performance estimation problem. The discovery of the main algorithm of Section 3 heavily relied on the use of the performance estimation problem (PEP) technique (Drori & Teboulle, 2014). Loosely speaking, the PEP is a computer-assisted methodology for finding optimal methods by numerically solving semidefinite programs (Drori & Teboulle, 2014; Kim & Fessler, 2016a; Taylor et al., 2018; Drori & Taylor, 2020; Kim & Fessler, 2021). We discuss the details of our use of the PEP in Section C of the appendix.

1.3. Contributions

We summarize the contribution of this work as follows. First, we present novel accelerated fixed-point iteration (OC-Halpern) and its equivalent form (OS-PPM) for monotone inclusions. Second, we present exact matching complexity lower bounds and thereby establish the exact optimality of our presented methods. Third, using a restarting mechanism, we extend the acceleration to a broader setup with operators satisfying a Hölder-type growth condition. Finally, we demonstrate the effectiveness of the proposed acceleration mechanism through extensive experiments.

2. Equivalence of nonexpansive operators and monotone operators

Before presenting the main content, we quickly establish the equivalence between the fixed-point problem

$$\underset{y\in\mathbb{R}^n}{\text{find}} \quad y = \mathbb{T}y$$

and the monotone inclusion

$$\underset{x\in\mathbb{R}^n}{\text{find}} \quad 0 \in \mathbb{A}x,$$

where $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ is $1/\gamma$-Lipschitz with $\gamma \geq 1$ and $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is maximal μ-strongly monotone with $\mu \geq 0$.

Lemma 2.1. *Let $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. If $\mathbb{T}$ is $1/\gamma$-Lipschitz with $\gamma \geq 1$, then*

$$\mathbb{A} = \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(1 + \frac{1}{\gamma}\right) - \mathbb{I}$$

is maximal $\frac{\gamma-1}{2}$-strongly monotone. Likewise, If $\mathbb{A}$ is maximal μ-strongly monotone with $\mu \geq 0$, then

$$\mathbb{T} = \left(1 + \frac{1}{1+2\mu}\right)\mathbb{J}_{\mathbb{A}} - \frac{1}{1+2\mu}\mathbb{I}$$

is $\frac{1}{1+2\mu}$-Lipschitz. Under these transformations, $x_\star$ is a zero of $\mathbb{A}$ if and only if it is a fixed point of $\mathbb{T}$, i.e., $\operatorname{Zer}\mathbb{A} = \operatorname{Fix}\mathbb{T}$.

The equivalence in case of $\gamma = 1$ and $\mu = 0$ is well known in optimization literature (Bauschke & Combettes, 2017, Theorem 23.8) (Bauschke et al., 2012; Combettes, 2018). This lemma generalizes the equivalence to $\gamma \geq 1$ and $\mu \geq 0$. As we see in Appendix A of the appendix, the equivalence is straightforwardly established using the scaled relative graph (SRG) (Ryu et al., 2021), but we also provide a classical proof based on inequalities without using the SRG.

Remark. Since $\mathbb{I} - \mathbb{T} = (1 + \frac{1}{1+2\mu})(\mathbb{I} - \mathbb{J}_{\mathbb{A}})$, finding an algorithm that effectively reduces $\|y_{N-1} - \mathbb{T}y_{N-1}\|^2$ for fixed-point problem is equivalent to finding an algorithm that effectively reduces $\|\tilde{\mathbb{A}}x_N\|^2$ for monotone inclusions.

3. Exact optimal methods

We now present our methods and their accelerated rates.

For a $1/\gamma$-contractive operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$, the *Optimal Contractive Halpern* (OC-Halpern) is

$$y_k = \left(1 - \frac{1}{\varphi_k}\right)\mathbb{T}y_{k-1} + \frac{1}{\varphi_k}y_0 \qquad \text{(OC-Halpern)}$$

for $k = 1, 2, \dots$, where $\varphi_k = \sum_{i=0}^{k} \gamma^{2i}$ and $y_0 \in \mathbb{R}^n$ is a starting point. For a maximal μ-strongly monotone operator $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, the *Optimal Strongly-monotone Proximal Point Method* (OS-PPM) is

$$\begin{aligned} x_k &= \mathbb{J}_{\mathbb{A}} y_{k-1} \qquad \text{(OS-PPM)}\\ y_k &= x_k + \frac{\varphi_{k-1} - 1}{\varphi_k}(x_k - x_{k-1}) - \frac{2\mu\varphi_{k-1}}{\varphi_k}(y_{k-1} - x_k)\\ &\quad + \frac{(1+2\mu)\varphi_{k-2}}{\varphi_k}(y_{k-2} - x_{k-1}) \end{aligned}$$

for $k = 1, 2, \dots$, where $\varphi_k = \sum_{i=0}^{k}(1+2\mu)^{2i}$, $\varphi_{-1} = 0$, and $x_0 = y_0 = y_{-1} \in \mathbb{R}^n$ is a starting point. These two methods are equivalent.

Lemma 3.1. *Suppose $\gamma = 1 + 2\mu$. Let $\mathbb{A} = \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(1 + \frac{1}{\gamma}\right) - \mathbb{I}$ given $\mathbb{T}$, or equivalently let $\mathbb{T} = \left(1 + \frac{1}{1+2\mu}\right)\mathbb{J}_{\mathbb{A}} - \frac{1}{1+2\mu}\mathbb{I}$ given $\mathbb{A}$. Then the y_k-iterates of (OC-Halpern) and (OS-PPM) are identical provided they start from the same initial point $y_0 = \tilde{y}_0$.*

We now state the convergence rates.

Theorem 3.2. *Let $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be maximal μ-strongly monotone with $\mu \geq 0$. Assume $\mathbb{A}$ has a zero and let $x_\star \in \operatorname{Zer}\mathbb{A}$. For $N = 1, 2, \dots$, (OS-PPM) exhibits the rate*

$$\|\tilde{\mathbb{A}}x_N\|^2 \leq \left(\frac{1}{\sum_{k=0}^{N-1}(1+2\mu)^k}\right)^2 \|y_0 - x_\star\|^2.$$

Corollary 3.3. *Let $\mathbb{T}\colon \mathbb{R} \to \mathbb{R}$ be γ^{-1}-contractive with $\gamma \geq 1$. Assume $\mathbb{T}$ has a fixed point and let $y_\star \in \operatorname{Fix}\mathbb{T}$. For $N = 0, 1, \dots$, (OC-Halpern) exhibits the rate*

$$\|y_N - \mathbb{T}y_N\|^2 \leq \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^{N}\gamma^k}\right)^2 \|y_0 - y_\star\|^2.$$

When $\mathbb{A}$ is strongly monotone ($\mu > 0$), (OS-PPM) exhibits an accelerated $\mathcal{O}(e^{-4\mu N})$-rate compared to the $\mathcal{O}(e^{-2\mu N})$-rate of the proximal point method (PPM) (Rockafellar, 1976; Bauschke & Combettes, 2017). When $\mathbb{T}$ is contractive ($\gamma < 1$), both (OC-Halpern) and the Picard iteration exhibit $\mathcal{O}(\gamma^{-2N})$-rates on the squared fixed-point residual. In fact, the Picard iteration with the $\mathbb{T}$ of Lemma 2.1 instead of $\mathbb{J}_{\mathbb{A}}$ is faster than the regular PPM and achieves a $\mathcal{O}(e^{-4\mu N})$ rate. (OC-Halpern) is exactly optimal and is faster than Picard in higher order terms hidden in the big-$\mathcal{O}$ notation. To clarify, the $\mathcal{O}$ considers the regime $\mu \to 0$.

When $\mathbb{A}$ is not strongly monotone ($\mu = 0$) or $\mathbb{T}$ is not contractive ($\gamma = 1$), (OS-PPM) and (OC-Halpern) respectively reduces to accelerated PPM (APPM) of Kim (2021) and Halpern iteration of Lieder (2021), sharing the same $\mathcal{O}(1/N^2)$-rate. In this paper, we refer to the method of Lieder (2021) as the optimized Halpern method (OHM).

The discovery of (OC-Halpern) and (OS-PPM) was assisted by the performance estimation problem (Drori & Teboulle, 2014; Kim & Fessler, 2016b; Taylor et al., 2017; Drori & Taylor, 2020; Ryu et al., 2020; Kim & Fessler, 2021; Park & Ryu, 2021) The details are discussed in Section C of the appendix.

3.1. Proof outline of Theorem 3.2

Here, we quickly outline the proof of Theorem 3.2 while deferring the full proof to Section B of the appendix.

Define the Lyapunov function

$$\begin{aligned} V^k &= (1 + \gamma^{-k})\Bigg[\left(\sum_{n=0}^{k-1}\gamma^n\right)^2 \|\tilde{\mathbb{A}}x_k\|^2\\ &+ 2\left(\sum_{n=0}^{k-1}\gamma^n\right)\langle \tilde{\mathbb{A}}x_k - \mu(x_k - x_\star), x_k - x_\star\rangle\\ &+ \gamma^{-k}\left\|\left(\sum_{n=0}^{k-1}\gamma^n\right)\tilde{\mathbb{A}}x_k - \gamma^k(x_k - x_\star) + (x_k - y_0)\right\|^2\Bigg]\\ &\quad + (1 - \gamma^{-k})\|y_0 - x_\star\|^2 \qquad \text{(OS-PPM-Lyapunov)} \end{aligned}$$

for $k = 0, 1, \dots$, where $\gamma = 1 + 2\mu$ and $\tilde{\mathbb{A}}x_k = y_{k-1} - x_k \in \mathbb{A}x_k$. After some calculations (deferred to the appendix), we use μ-strong monotonicity of $\mathbb{A}$ to conclude

$$\begin{aligned} V^{k+1} - V^k &= -2\gamma^{-2k}(1+\gamma)\varphi_k\varphi_{k-1}\\ &\quad \langle \tilde{\mathbb{A}}x_{k+1} - \tilde{\mathbb{A}}x_k - \mu(x_{k+1} - x_k), x_{k+1} - x_k\rangle\\ &\leq 0. \end{aligned}$$

Therefore,

$$V^N \leq V^{N-1} \leq \cdots \leq V^0 = 2\|y_0 - x_\star\|^2$$

and we conclude

$$\|\tilde{\mathbb{A}}x_N\|^2 \leq \left(\frac{1}{\sum_{k=0}^{N-1}\gamma^k}\right)^2 \|y_0 - x_\star\|^2.$$

4. Complexity lower bound

We now establish *exact* optimality of (OC-Halpern) and (OS-PPM) through matching complexity lower bound. By exact, we mean that the lower bound is exactly equal to upper bounds of Theorem 3.2 and Corollary 3.3.

Theorem 4.1. *For $n \geq N+1$ and any initial point $y_0 \in \mathbb{R}^n$, there exists an $1/\gamma$-Lipschitz operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ with a fixed point $y_\star \in \operatorname{Fix}\mathbb{T}$ such that*

$$\|y_N - \mathbb{T}y_N\|^2 \geq \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^{N}\gamma^k}\right)^2 \|y_0 - y_\star\|^2$$

for any iterates $\{y_k\}_{k=0}^N$ satisfying

$$y_k \in y_0 + \operatorname{span}\{y_0 - \mathbb{T}y_0, y_1 - \mathbb{T}y_1, \dots, y_{k-1} - \mathbb{T}y_{k-1}\}$$

for $k = 1, \dots, N$.

The following corollary translates Theorem 4.1 to an equivalent complexity lower bound for proximal point methods in monotone inclusions.

Corollary 4.2. *For $n \geq N$ and any initial point $x_0 = y_0 \in \mathbb{R}^n$, there exists a maximal μ-strongly monotone operator $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with a zero $x_\star \in \operatorname{Zer}\mathbb{A}$ such that*

$$\|\tilde{\mathbb{A}}x_N\|^2 \geq \left(\frac{1}{\sum_{k=0}^{N-1}(1+2\mu)^k}\right)^2 \|y_0 - x_\star\|^2$$

for any iterates $\{x_k\}_{k=0,1,\dots}$ and $\{y_k\}_{k=0,1,\dots}$ satisfying

$$\begin{aligned} x_k &= \mathbb{J}_{\mathbb{A}} y_{k-1} \\ y_k &\in y_0 + \operatorname{span}\{\tilde{\mathbb{A}}x_1, \tilde{\mathbb{A}}x_2, \dots, \tilde{\mathbb{A}}x_k\} \end{aligned}$$

for $k = 1, \dots, N$, where $\tilde{\mathbb{A}}x_k = y_{k-1} - x_k$.

(OC-Halpern) and (OS-PPM) satisfy the span assumptions stated in Theorem 4.1 and Corollary 4.2, respectively. Therefore, the rates of (OC-Halpern) and (OS-PPM) are exactly optimal. The lower bounds in the cases where $\gamma = 1$ and $\mu = 0$ establish that the prior rates of OHM (Lieder, 2021) and APPM (Kim, 2021) are exactly optimal. To clarify, the lower bound is novel even for the case $\gamma = 1$ and $\mu = 0$.

4.1. Construction of the worst-case operator

We now describe the construction of the worst-case operator, while deferring the proofs to Section D of the appendix. Let e_k be the canonical basis vector with 1 at the k-th entry and 0 at remaining entries.

Lemma 4.3. *$\mathbb{T}$ is $\frac{1}{\gamma}$-contractive if and only if $\mathbb{G} = \frac{\gamma}{1+\gamma}(\mathbb{I} - \mathbb{T})$ is $\frac{1}{1+\gamma}$-averaged.*

By Lemma 4.3, finding the worst-case $\frac{1}{\gamma}$-contractive operator $\mathbb{T}$ is equivalent to finding the worst-case $\frac{1}{1+\gamma}$-averaged operator $\mathbb{G}$, which we define in the following lemma.

Lemma 4.4. *Let $R > 0$. Define $\mathbb{N}, \mathbb{G}\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ as*

$$\mathbb{N}(x_1, x_2, \dots, x_N, x_{N+1}) = (x_{N+1}, -x_1, -x_2, \dots, -x_N) - \frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}} R e_1$$

and

$$\mathbb{G} = \frac{1}{1+\gamma}\mathbb{N} + \frac{\gamma}{1+\gamma}\mathbb{I}.$$

That is,

$$\mathbb{G}x = \underbrace{\frac{1}{1+\gamma}\begin{bmatrix} \gamma & 0 & \cdots & 0 & 1 \\ -1 & \gamma & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma & 0 \\ 0 & 0 & \cdots & -1 & \gamma \end{bmatrix}}_{=:H} x \underbrace{- \frac{1}{1+\gamma}\frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}} R e_1}_{=:b}.$$

Then $\mathbb{N}$ is nonexpansive, and $\mathbb{G}$ is $\frac{1}{1+\gamma}$-averaged.

Following lemma states the property of iterations $\{y_k\}_{k=0}^N$ with respect to $\mathbb{G}$, that proper span condition results in gradually expanding support of y_k.

Lemma 4.5. *Let $\mathbb{G}\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ be defined as in Lemma 4.4. For any $\{y_k\}_{k=0}^N$ with $y_0 = 0$ satisfying*

$$y_k \in y_0 + \operatorname{span}\{\mathbb{G}y_0, \mathbb{G}y_1, \dots, \mathbb{G}y_{k-1}\}, \quad k = 1, \dots, N,$$

we have

$$\begin{aligned} y_k &\in \operatorname{span}\{e_1, e_2, \dots, e_k\} \\ \mathbb{G}y_k &\in \operatorname{span}\{e_1, e_2, \dots, e_{k+1}\}, \quad k = 0, \dots, N. \end{aligned}$$

4.2. Proof outline of Theorem 4.1

Let $\mathbb{T}_0\colon \mathbb{R}^n \to \mathbb{R}^n$ be the worst-case $\frac{1}{\gamma}$-contraction for initial point 0. For any given $y_0 \in \mathbb{R}^n$, we show in section D of the appendix that $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ defined as $\mathbb{T}(\cdot) = \mathbb{T}_0(\cdot - y_0) + y_0$ becomes the worst-case $\frac{1}{\gamma}$-contraction with initial point $y_0 \in \mathbb{R}^n$. Therefore, it suffices to consider the case $y_0 = 0$.

Define $\mathbb{G}$, H, and b as in Lemma 4.4. By Lemma 4.3, $\mathbb{T} = \mathbb{I} - \frac{1+\gamma}{\gamma}\mathbb{G}$ is a $1/\gamma$-contraction. Note that H is invertible, as we can use Gaussian elimination on H to obtain an upper triangular matrix with nonzero diagonals. This makes $\mathbb{G}$ an invertible affine operator with the unique zero

$$y_\star = \frac{R}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}}\begin{bmatrix}\gamma^N & \gamma^{N-1} & \cdots & \gamma & 1\end{bmatrix}^\intercal.$$

So $\operatorname{Fix}\mathbb{T} = \operatorname{Zer}\mathbb{G} = \{y_\star\}$ and $\|y_0 - y_\star\| = \|y_\star\| = R$.

Let the iterates $\{y_k\}_{k=0}^N$ satisfy the span condition of Theorem 4.1, which is equivalent to

$$y_k \in y_0 + \operatorname{span}\{\mathbb{G}y_0, \mathbb{G}y_1, \dots, \mathbb{G}y_{k-1}\}, \quad k = 1, \dots, N.$$

By Lemma 4.5, $y_N \in \operatorname{span}\{e_1, \dots, e_N\}$. Therefoere

$$\mathbb{G}y_N = Hy_N - b \in \operatorname{span}\{He_1, \dots, He_N\} - b.$$

and

$$\|\mathbb{G}y_N\|^2 \geq \left\|\mathcal{P}_{\operatorname{span}\{He_1,\dots,He_N\}^\perp}(b)\right\|^2,$$

where $\mathcal{P}_V$ is the orthogonal projection onto the subspace V. As $\operatorname{span}\{He_1, \dots, He_N\}^\perp = \operatorname{span}\{v\}$ with

$$v = \begin{bmatrix}1 & \gamma & \cdots & \gamma^{N-1} & \gamma^N\end{bmatrix}^\intercal,$$

we get

$$\|\mathbb{G}y_N\|^2 \geq \left\|\mathcal{P}_{\operatorname{span}\{v\}}(b)\right\|^2 = \left\|\frac{\langle b, v\rangle}{\langle v, v\rangle}v\right\|^2 \overset{(*)}{=} \left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 R^2,$$

where $(*)$ is established in the Section D of the appendix. Finally,

$$\begin{aligned}\|y_N - \mathbb{T}y_N\|^2 &= \left\|\left(1+\frac{1}{\gamma}\right)\mathbb{G}y_N\right\|^2 \\ &\geq \left(1+\frac{1}{\gamma}\right)^2\left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 R^2 \\ &= \left(1+\frac{1}{\gamma}\right)^2\left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 \|y_0 - y_\star\|^2.\end{aligned}$$

4.3. Generalized complexity lower bound result

In order to extend the lower bound results of Theorem 4.1 and Corollary 4.2 to general deterministic fixed-point iterations and proximal point methods (which do not necessarily satisfy the span condition), we use the resisting oracle technique of Nemirovski & Yudin (1983). Here, we quickly state the result while deferring the proofs to the Section D of the appendix.

Theorem 4.6. *Let $n \geq 2N$ for $N \in \mathbb{N}$. For any deterministic fixed-point iteration $\mathbf{A}$ and any initial point $y_0 \in \mathbb{R}^n$, there exists a $\frac{1}{\gamma}$-Lipschitz operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ with a fixed point $y_\star \in \operatorname{Fix}\mathbb{T}$ such that*

$$\|y_N - \mathbb{T}y_N\|^2 \geq \left(1+\frac{1}{\gamma}\right)^2\left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 \|y_0 - y_\star\|^2$$

where $\{y_t\}_{t\in\mathbb{N}} = \mathbf{A}[y_0; \mathbb{T}]$.

5. Acceleration under Hölder-type growth condition

While (OS-PPM) provides an accelerated rate when the underlying operator is monotone or strongly monotone, many operators encountered in practice have a structure lying between these two assumptions. For (OC-Halpern), this corresponds to a fixed-point operator that is not strictly contractive but has structure stronger than nonexpansiveness. In this section, we accelerate the proximal point method when the underlying operator is uniformly monotone, an assumption weaker than strong monotonicity but stronger than monotonicity.

We say an operator $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *uniformly monotone with parameters $\mu > 0$ and $\alpha > 1$* if it is monotone and

$$\langle \mathbb{A}x, x - x_\star\rangle \geq \mu\|x - x_\star\|^{\alpha+1}$$

for any $x \in \mathbb{R}^n$ and $x_\star \in \operatorname{Zer}\mathbb{A}$. This is a special case of uniform monotonicity in Bauschke & Combettes (2017, Definition 22.1). We also refer to this as a Hölder-type growth condition, as it resembles the Hölderian error bound condition with function-value suboptimality replaced by $\langle \mathbb{A}x, x - x_\star\rangle$ (Lojasiewicz, 1963; Bolte et al., 2017).

The following theorem establishes a convergence rate of the (unaccelerated) proximal point method. This rate serves as a baseline to improve upon with acceleration.

Theorem 5.1. *Let $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be uniformly monotone with parameters $\mu > 0$ and $\alpha > 1$. Let $x_\star \in \operatorname{Zer}\mathbb{A}$. Then the iterates $\{x_k\}_{k=0}^N$ generated by the proximal point method $x_{k+1} = \mathbb{J}_{\mathbb{A}}x_k$ starting from $x_0 \in \mathbb{R}^n$ satisfy*

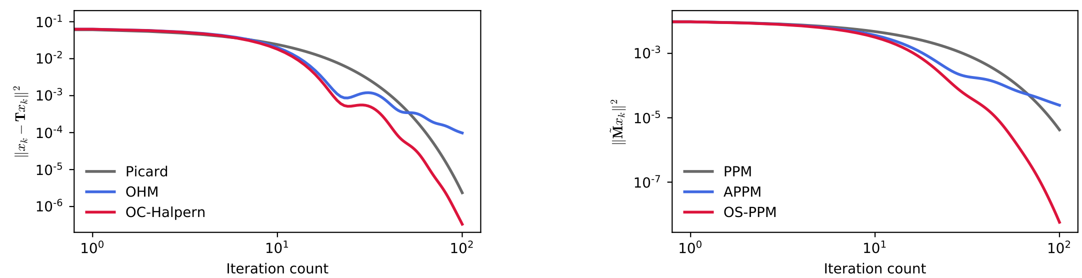
$$\|\tilde{\mathbb{A}}x_N\|^2 \leq \frac{2^{\frac{\alpha+3}{\alpha-1}}\max\left\{\left(\frac{2^{\frac{\alpha}{\alpha-1}}-2}{\mu}\right)^{\frac{2}{\alpha-1}}, \|x_0 - x_\star\|^2\right\}}{N^{\frac{\alpha+1}{\alpha-1}}}$$

for $N \in \mathbb{N}$ where $\tilde{\mathbb{A}}x_N = x_{N-1} - x_N$.

We now present an accelerated method based on (OS-PPM) and restarting (Nesterov, 2013; Roulet & d'Aspremont, 2020). Given a uniformly monotone operator $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with $\mu > 0$ and $\alpha > 1$, $x_\star \in \operatorname{Zer}\mathbb{A}$, and an initial point $x_0 \in \mathbb{R}^n$, *Restarted OS-PPM* is:

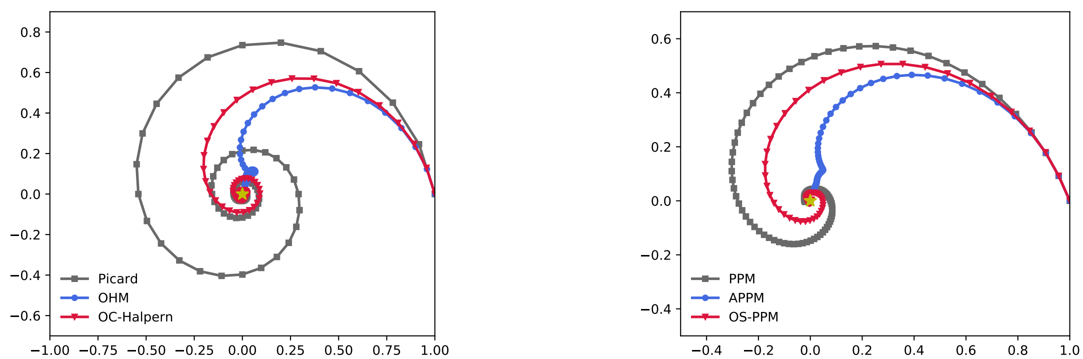
$$\begin{aligned}&\tilde{x}_0 = \mathbb{J}_{\mathbb{A}}x_0 && (\text{OS-PPM}_0^{\text{res}})\\ &\tilde{x}_k \leftarrow \textbf{OS-PPM}_0(\tilde{x}_{k-1}, t_k), && k = 1, \dots, R,\end{aligned}$$

where $\textbf{OS-PPM}_0(\tilde{x}_{k-1}, t_k)$ is the execution of t_k iterations of (OS-PPM) with $\mu = 0$ starting from $\tilde{x}_{k-1}$. The following theorem provides a restarting schedule, i.e., specified values of $t_1, \dots, t_R$, and an accelerated rate.



(a) Fixed-point residual of $\mathbb{T}_\theta$ (b) Resolvent residual norm of $\mathbb{M}$

Figure 1. Fixed-point and resolvent residuals versus iteration count for the 2D toy example of Section 6.1. Here, $\gamma = 1/0.95 = 1.0526$, $\mu = 0.035$, $\theta = 15^\circ$ and $N = 101$. Indeed, (OC-Halpern) and (OS-PPM) exhibit the fastest rates.



(a) Trajectory of $\mathbb{T}_\theta$ (b) Trajectory of $\mathbb{M}$

Figure 2. Trajectories of iterates for the 2D toy example of Section 6.1. Here, $\gamma = 1/0.95 = 1.0526$, $\mu = 0.035$, $\theta = 15^\circ$ and $N = 101$. A marker is placed at every iterate. Picard and PPM are slowed down by the cyclic behavior. Halpern and APPM dampens the cycling behavior, but does so too aggressively. The fastest rate is achieved by (OC-Halpern) and (OS-PPM), which appears to be due to the cycling behavior being optimally dampened.

Theorem 5.2. *Let $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be uniformly monotone with parameters $\mu > 0$ and $\alpha > 1$, $x_\star \in \operatorname{Zer} \mathbb{A}$, and N be the total number of iterations. Define*

$$\lambda = \left(\frac{e}{\mu}\right)^{\frac{1}{\alpha}} \|x_0 - x_\star\|^{1-\frac{1}{\alpha}}, \qquad \beta = 1 - \frac{1}{\alpha}.$$

Let $R \in \mathbb{N}$ be an integer satisfying

$$\sum_{k=1}^{R} \lceil \lambda e^{\beta k} \rceil \le N - 1 < \sum_{k=1}^{R+1} \lceil \lambda e^{\beta k} \rceil,$$

and let t_k be defined as

$$t_k = \begin{cases} \lceil \lambda e^{\beta k} \rceil & \text{for } k = 1, \dots, R-1 \\ N - 1 - \sum_{k=1}^{R-1} t_k. & \text{for } k = R. \end{cases}$$

Then (OS-PPM$_0^{\text{res}}$) exhibits the rate

$$\begin{aligned} &\|\tilde{\mathbb{A}} x_N\|^2 \\ &\le \left\{ \tfrac{e^{\beta}-1}{\lambda e^{2\beta}} \left(N - 2 - \tfrac{1}{\beta} \log\left(\tfrac{e^{\beta}-1}{\lambda e^{\beta}} (N-1) + 1 \right) \right) + \tfrac{1}{e^{\beta}} \right\}^{-\frac{2\alpha}{\alpha - 1}} \\ &\quad \times \|x_0 - x_\star\|^2 \\ &= \mathcal{O}\left(N^{-\frac{2\alpha}{\alpha-1}} \right). \end{aligned}$$

The proofs of Theorems 5.1 and 5.2 are presented in Section E of the appendix. When the values of α, μ, and $\|x_0 - x_\star\|^2$ are unknown, as in the case in most practical setups, one can use a grid search as in Roulet & d'Aspremont (2020) and retain the $\mathcal{O}\left(N^{-\frac{2\alpha}{\alpha-1}} (\log N)^2\right)$-rate. Using Lemma 4.3, (OS-PPM$_0^{\text{res}}$) can be translated into a restarted OC-Halpern method. The experiments of Section 6 indicate that (OS-PPM$_0^{\text{res}}$) does provide an acceleration in cases where (OS-PPM) by itself does not.

6. Experiments

We now present experiments with illustrative toy examples and real-world problems in medical imaging, optimal transport, and decentralized compressed sensing. Further experimental details are provided in Section F of the appendix.

6.1. Illustrative 2D toy examples

Consider a $\frac{1}{\gamma}$-contractive operator $\mathbb{T}_\theta \colon \mathbb{R}^2 \to \mathbb{R}^2$

$$\mathbb{T}_\theta \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\gamma} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and a maximal μ-strongly monotone operator $\mathbb{M} \colon \mathbb{R}^2 \to \mathbb{R}^2$

$$\mathbb{M} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \left(\frac{1}{N-1} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + \begin{bmatrix} \mu & 0 \\ 0 & \mu \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

$\mathbb{T}_\theta$ is a counterclockwise θ-rotation followed by $\frac{1}{\gamma}$-scaling on 2D plane, and $\mathbb{M}$ is a linear combination of the worst-case instances of the proximal point method applied to monotone operators (Gu & Yang, 2020) and μ-strongly monotone operators (Rockafellar, 1976). The results of Figure 1 indicate that (OC-Halpern) and (OS-PPM) indeed provide acceleration.

6.2. Computed tomography (CT) imaging

Consider the medical imaging application of total variation regularized computed tomography (CT), which solves

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|Ex - b\|^2 + \lambda \|Dx\|_1,$$

where $x \in \mathbb{R}^n$ is a vectorized image, $E \in \mathbb{R}^{m \times n}$ is the discrete Radon transform, $b = Ex$ is the measurement, and D is the finite difference operator. We use primal-dual hybrid gradient (PDHG) (Zhu & Chan, 2008; Pock et al., 2009; Esser et al., 2010; Chambolle & Pock, 2011), an instance of a nonexpansive fixed-point iteration via variable metric PPM (He & Yuan, 2012). The results of Figure 3(a) indicate that restarted OC-Halpern ($\text{OS-PPM}_0^{\text{res}}$) provides an acceleration.

6.3. Earth mover's distance

Consider the earth mover's distance between two probability measures, also referred to as the Wasserstein distance or the optimal transport problem. The distance is defined through the discretized optimization problem

$$\begin{aligned} \underset{m_x, m_y}{\text{minimize}} \quad & \|\mathbf{m}\|_{1,1} = \sum_{i=1}^n \sum_{j=1}^n |m_{x,ij}| + |m_{y,ij}| \\ \text{subject to} \quad & \operatorname{div}(\mathbf{m}) + \rho_1 - \rho_0 = 0, \end{aligned}$$

where ρ_0, ρ_1 are probability measures on $\mathbb{R}^{n \times n}$, div is a discrete divergence operator, and $\mathbf{m} = (m_x, m_y) \in \mathbb{R}^{(n-1) \times n} \times \mathbb{R}^{n \times (n-1)}$ is the optimization variable. We use the algorithm of Li et al. (2018), an instance of a nonexpansive fixed-point iteration via PDHG. The results of Figure 3(b) indicate that restarted OC-Halpern ($\text{OS-PPM}_0^{\text{res}}$) provides an acceleration.

6.4. Decentralized optimization with PG-EXTRA

Consider a decentralized optimization setting where each agent $i \in \{1, 2, \ldots, n\}$ has access to the sensing matrix $A_{(i)} \in \mathbb{R}^{m_i \times n}$ and the noisy measurement $b_{(i)} \approx A_{(i)}x$. The goal is to recover the sparse signal $x \in \mathbb{R}^n$ by solving the following compressed sensing problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n \|A_{(i)}x - b_{(i)}\|^2 + \lambda \|x\|_1.$$

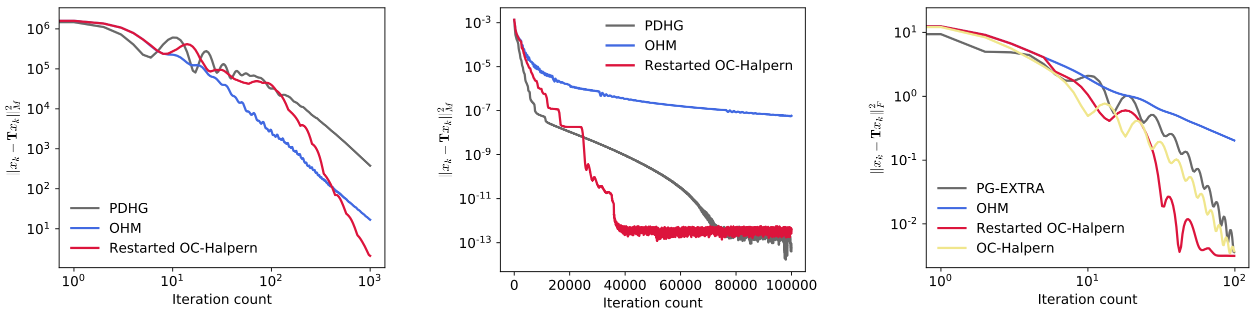
We use PG-EXTRA (Shi et al., 2015), which is an instance of a nonexpansive fixed-point iteration via the Condat–Vũ (Condat, 2013; Vũ, 2013) splitting method (Wu et al., 2018). The results of Figure 3(c) indicate that restarted OC-Halpern ($\text{OS-PPM}_0^{\text{res}}$) provides an acceleration.

7. Conclusion

This work presents an acceleration mechanism for fixed-point iterations and provides an exact matching complexity lower bound. The acceleration mechanism is an instance of Halpern's method, also referred to as anchoring, and the complexity lower bound is based on an explicit construction satisfying the zero-chain condition.

In this work, we measure the suboptimality of iterates with the fixed-point residual. However, the fixed-point iteration is a *meta*-algorithm, and almost all instances of it have further specific structure and suboptimality measures that are better suited for the particular problem of interest, such as function-value suboptimality, infeasibility for constrained problems, and primal-dual gap for minimax problems. Therefore, the fact that our proposed method accelerates the reduction of the fixed-point residual does not necessarily imply that it accelerates the reduction of the problem-specific suboptimality measure of practical interest.

Interestingly, the experimental results of Sections 6 and F indicate that our proposed acceleration does indeed provide a benefit in practice. This raises the following question: Under what setups can we expect anchoring-based acceleration to theoretically provide a benefit in terms of other suboptimality measures? Investigating this question would be an interesting direction of future work.



(a) CT imaging (b) Earth mover's distance (c) Decentralized compressed sensing

Figure 3. Reduction of fixed-point residuals. The norms are the norms under which the fixed-point operator $\mathbb{T}$ is nonexpansive.

Acknowledgements

JP and EKR were supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2022R1C1C1010010], the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2022R1A5A6000840], and the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02). We thank TaeHo Yoon for providing careful reviews and valuable feedback. We thank Jelena Diakonikolas for the discussion on the prior work on complexity lower bounds of the fixed-point iterations. Finally, we thank the anonymous reviewers for their thoughtful comments.

References

Anderson, D. G. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4):547–560, 1965.

Arjevani, Y., Shalev-Shwartz, S., and Shamir, O. On lower and upper bounds in smooth and strongly convex optimization. *The Journal of Machine Learning Research*, 17 (1):4303–4353, 2016.

Baillon, J.-B. and Bruck, R. E. Optimal rates of asymptotic regularity for averaged nonexpansive mappings. *Fixed Point Theory and Applications*, 128:27–66, 1992.

Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.

Barré, M., Taylor, A., and d'Aspremont, A. Convergence of constrained anderson acceleration. *arXiv preprint arXiv:2010.15482*, 2020.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, second edition, 2017.

Bauschke, H. H., Moffat, S. M., and Wang, X. Firmly nonexpansive mappings and maximally monotone operators: correspondence and duality. *Set-Valued and Variational Analysis*, 20:131–153, 2012.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Berinde, V. and Takens, F. *Iterative Approximation of Fixed Points*. Springer, 2007.

Bertrand, Q. and Massias, M. Anderson acceleration of coordinate descent. *AISTATS*, pp. 1288–1296, 2021.

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.

Borwein, J., Reich, S., and Shafrir, I. Krasnoselski-Mann iterations in normed spaces. *Canadian Mathematical Bulletin*, 35(1):21–28, 1992.

Borwein, J. M., Li, G., and Tam, M. K. Convergence rate analysis for averaged fixed point iterations in common fixed point problems. *SIAM Journal on Optimization*, 27 (1):1–33, 2017.

Bravo, M. and Cominetti, R. Sharp convergence rates for averaged nonexpansive maps. *Israel Journal of Mathematics*, 227:163–188, 2018.

Brezinski, C. Convergence acceleration during the 20th century. *Journal of Computational and Applied Mathematics*, 122(1-2):1–21, 2000.

Browder, F. E. and Petryshyn, W. The solution by iteration of nonlinear functional equations in Banach spaces. *Bulletin of the American Mathematical Society*, 72(3): 571–575, 1966.

Bruck Jr, R. E. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185:315–355, 2021.

Cauchy, A.-L. Méthode générale pour la résolution des systemes d'équations simultanées. *Comptes rendus de l'Académie des Sciences*, 25:536–538, 1847.

Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

Colao, V. and Marino, G. On the rate of convergence of Halpern iterations. *Journal of Nonlinear and Convex Analysis*, 22(12):2639–2646, 2021.

Combettes, P. L. Monotone operator theory in convex optimization. *Mathematical Programming*, 170(1):177–206, 2018.

Cominetti, R., Soto, J. A., and Vaisman, J. On the rate of convergence of Krasnosel'skiĭ-Mann iterations and their connection with sums of Bernoullis. *Israel Journal of Mathematics*, 199(2):757–772, 2014.

Condat, L. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

Contreras, J. P. and Cominetti, R. Optimal error bounds for nonexpansive fixed-point iterations in normed spaces. *arXiv preprint arXiv:2108.10969*, 2021.

Davis, D. and Yin, W. Convergence rate analysis of several splitting schemes. In Glowinski, R., Osher, S. J., and Yin, W. (eds.), *Splitting Methods in Communication, Imaging, Science, and Engineering*, pp. 115–163. Springer, 2016.

Davis, D. and Yin, W. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017.

Diakonikolas, J. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. *COLT*, 2020.

Diakonikolas, J. and Wang, P. Potential function-based framework for making the gradients small in convex and min-max optimization. *arXiv preprint arXiv:2101.12101*, 2021.

Dong, Q., Yuan, H., Cho, Y., and Rassias, T. M. Modified inertial Mann algorithm and inertial CQ-algorithm for nonexpansive mappings. *Optimization Letters*, 12(1): 87–102, 2018.

Douglas, J. and Rachford, H. H. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82(2):421–439, 1956.

Drori, Y. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

Drori, Y. and Shamir, O. The complexity of finding stationary points with stochastic gradient descent. *ICML*, 2020.

Drori, Y. and Taylor, A. On the oracle complexity of smooth strongly convex minimization. *Journal of Complexity*, 68, 2022.

Drori, Y. and Taylor, A. B. Efficient first-order methods for convex minimization: a constructive approach. *Mathematical Programming*, 184(1–2):183–220, 2020.

Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1–2):451–482, 2014.

Drori, Y. and Teboulle, M. An optimal variant of Kelley's cutting-plane method. *Mathematical Programming*, 160 (1–2):321–351, 2016.

Esser, E., Zhang, X., and Chan, T. F. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

Fercoq, O. and Qu, Z. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019.

Gabay, D. and Mercier, B. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Gu, G. and Yang, J. Tight sublinear convergence rate of the proximal point algorithm for maximal monotone inclusion problems. *SIAM Journal on Optimization*, 30(3): 1905–1921, 2020.

Güler, O. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Halpern, B. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

He, B. and Yuan, X. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1): 119–149, 2012.

Hestenes, M. R. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

Ishikawa, S. Fixed points and iteration of a nonexpansive mapping in a Banach space. *Proceedings of the American Mathematical Society*, 59(1):65–71, 1976.

Ito, M. and Fukuda, M. Nearly optimal first-order methods for convex optimization under gradient norm measure: An adaptive regularization approach. *Journal of Optimization Theory and Applications*, 188(3):770–804, 2021.

Kim, D. Accelerated proximal point method for maximally monotone operators. *Mathematical Programming*, 190 (1–2):57–87, 2021.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1–2):81–107, 2016a.

Kim, D. and Fessler, J. A. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016b.

Kim, D. and Fessler, J. A. Adaptive restart of the optimized gradient method for convex optimization. *Journal of Optimization Theory and Applications*, 178(1):240–263, 2018.

Kim, D. and Fessler, J. A. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of Optimization Theory and Applications*, 188(1):192–219, 2021.

Kohlenbach, U. On quantitative versions of theorems due to F. E. Browder and R. Wittmann. *Advances in Mathematics*, 226(3):2764–2795, 2011.

Krasnosel'skiĭ, M. A. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10:123–127, 1955.

Lee, J., Park, C., and Ryu, E. K. A geometric structure of acceleration and its role in making gradients small fast. *NeurIPS*, 2021.

Leustean, L. Rates of asymptotic regularity for Halpern iterations of nonexpansive mappings. *Journal of Universal Computer Science*, 13(11):1680–1691, 2007.

Li, W., Ryu, E. K., Osher, S., Yin, W., and Gangbo, W. A parallel method for earth mover's distance. *Journal of Scientific Computing*, 75(1):182–197, 2018.

Liang, J., Fadili, J., and Peyré, G. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1–2):403–434, 2016.

Lieder, F. On the convergence rate of the Halpern-iteration. *Optimization Letters*, 15(2):405–418, 2021.

Lin, Q. and Xiao, L. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *ICML*, 2014.

Lin, Y. and Xu, Y. Convergence rate analysis for fixed-point iterations of generalized averaged nonexpansive operators. *arXiv preprint arXiv:2108.06714*, 2021.

Lions, P.-L. and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

Lojasiewicz, S. Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aux Dérivées Partielles*, 117:87–89, 1963.

Maingé, P.-E. Convergence theorems for inertial KM-type algorithms. *Journal of Computational and Applied Mathematics*, 219(1):223–236, 2008.

Maingé, P.-E. Accelerated proximal algorithms with a correction term for monotone inclusions. *Applied Mathematics & Optimization*, 84(2):2027–2061, 2021.

Mann, W. R. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

Martinet, B. Régularisation d'inéquations variationnelles par approximations successives. *Revue Française de Informatique et Recherche Opérationnelle*, 4(R3):154–158, 1970.

Martinet, B. *Algorithmes pour la résolution de problèmes d'optimisation et de minimax*. PhD thesis, Université Joseph-Fourier-Grenoble I, 1972.

Matsushita, S.-Y. On the convergence rate of the Krasnosel'skiĭ–Mann iteration. *Bulletin of the Australian Mathematical Society*, 96(1):162–170, 2017.

Nemirovski, A. S. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

Nemirovski, A. S. and Nesterov, Y. E. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(3–4):21–30, 1985.

Nemirovski, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269:543–547, 1983.

O'Donoghue, B. and Candes, E. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

Park, C. and Ryu, E. K. Optimal first-order algorithms as a function of inequalities. *arXiv preprint arXiv:2110.11035*, 2021.

Park, C., Park, J., and Ryu, E. K. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *arXiv preprint arXiv:2102.07366*, 2021.

Passty, G. B. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

Peaceman, D. W. and Rachford, Jr, H. H. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics*, 3(1):28–41, 1955.

Pock, T., Cremers, D., Bischof, H., and Chambolle, A. An algorithm for minimizing the Mumford-Shah functional. *ICCV*, 2009.

Powell, M. J. A method for nonlinear constraints in minimization problems. *Optimization*, pp. 283–298, 1969.

Reich, S., Thong, D. V., Cholamjiak, P., and Van Long, L. Inertial projection-type methods for solving pseudomonotone variational inequality problems in Hilbert space. *Numerical Algorithms*, 88(2):813–835, 2021.

Rhoades, B. Some fixed point iteration procedures. *International Journal of Mathematics and Mathematical Sciences*, 14(1):1–16, 1991.

Rhoades, B. and Saliga, L. Some fixed point iteration procedures. II. *Nonlinear Analysis Forum*, 6(1):193–217, 2001.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

Roulet, V. and d'Aspremont, A. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

Ryu, E. and Yin, W. *Large-scale Convex Optimization via Monotone Operators*. Draft, 2020.

Ryu, E. K., Taylor, A. B., Bergeling, C., and Giselsson, P. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.

Ryu, E. K., Hannah, R., and Yin, W. Scaled relative graphs: Nonexpansive operators via 2d Euclidean geometry. *Mathematical Programming*, 2021.

Sabach, S. and Shtern, S. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Salim, A., Condat, L., Kovalev, D., and Richtárik, P. An optimal algorithm for strongly convex minimization under affine constraints. *AISTATS*, 2022.

Scieur, D., d'Aspremont, A., and Bach, F. Regularized nonlinear acceleration. *Mathematical Programming*, 179 (1–2):47–83, 2020.

Shehu, Y. Convergence rate analysis of inertial Krasnoselskii–Mann type iteration with applications. *Numerical Functional Analysis and Optimization*, 39(10): 1077–1091, 2018.

Shi, W., Ling, Q., Wu, G., and Yin, W. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, 2015.

Taylor, A. and Drori, Y. An optimal gradient method for smooth (possibly strongly) convex minimization. *arXiv preprint arXiv:2101.09741*, 2021.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1–2):307–345, 2017.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018.

Van Scoy, B., Freeman, R. A., and Lynch, K. M. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.

Vũ, B. C. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Walker, H. F. and Ni, P. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

Wittmann, R. Approximation of fixed points of nonexpansive mappings. *Archiv der Mathematik*, 58(5):486–491, 1992.

Wu, T., Yuan, K., Ling, Q., Yin, W., and Sayed, A. H. Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):293–307, 2018.

Xu, H.-K. Iterative algorithms for nonlinear operators. *Journal of the London Mathematical Society*, 66(1):240–256, 2002.

Xu, H.-K. Viscosity approximation methods for nonexpansive mappings. *Journal of Mathematical Analysis and Applications*, 298(1):279–291, 2004.

Yoon, T. and Ryu, E. K. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. *ICML*, 2021.

Zhang, J., O'Donoghue, B., and Boyd, S. Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization*, 30(4): 3170–3197, 2020.

Zhu, M. and Chan, T. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report 08-34*, 2008.

A. Omitted proofs of Section 2

Proof of Lemma 2.1 with inequalities. Suppose $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ is $\frac{1}{\gamma}$-Lipschitz for $\gamma \geq 1$. Define $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as

$$\mathbb{A} = \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(1 + \frac{1}{\gamma}\right) - \mathbb{I}.$$

For any $x, y \in \mathbb{R}^n$, let $u \in \mathbb{A}x$ and $v \in \mathbb{A}y$. Then

$$\begin{aligned} u \in \mathbb{A}x \implies & u \in \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(1 + \frac{1}{\gamma}\right)x - x \\ \iff & x + u \in \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(1 + \frac{1}{\gamma}\right)x \\ \iff & \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)(x+u) = x + \frac{1}{\gamma}x \\ \iff & \mathbb{T}(x+u) = x - \frac{1}{\gamma}u \end{aligned}$$

Likewise,

$$\mathbb{T}(y+v) = y - \frac{1}{\gamma}v.$$

From the $\frac{1}{\gamma}$-Lipschitzness of $\mathbb{T}$,

$$\begin{aligned} \|\mathbb{T}(x+u) - \mathbb{T}(y+v)\| \leq \frac{1}{\gamma}\|(x+u)-(y+v)\| \iff & \left\|\left(x - \frac{1}{\gamma}u\right) - \left(y - \frac{1}{\gamma}v\right)\right\| \leq \frac{1}{\gamma}\|(x+u)-(y+v)\| \\ \iff & \left\|(x-y) - \frac{1}{\gamma}(u-v)\right\|^2 \leq \frac{1}{\gamma^2}\|(x-y)+(u-v)\|^2 \\ \iff & \left(1 - \frac{1}{\gamma^2}\right)\|x-y\|^2 \leq \left(\frac{2}{\gamma^2} + \frac{2}{\gamma}\right)\langle u-v, x-y\rangle \\ \iff & \langle u-v, x-y\rangle \geq \frac{\gamma-1}{2}\|x-y\|^2. \end{aligned}$$

This holds for any $u \in \mathbb{A}x$ and $v \in \mathbb{A}y$ for any $x, y \in \mathbb{R}^n$, so $\mathbb{A}$ is $\frac{\gamma-1}{2}$-strongly monotone.

We can further prove that

$$\begin{aligned} x_\star \in \operatorname{Zer}\mathbb{A} \iff & 0 \in \mathbb{A}x_\star = \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(x_\star + \frac{1}{\gamma}x_\star\right) - x_\star \\ \iff & x_\star \in \left(\mathbb{T} + \frac{1}{\gamma}\mathbb{I}\right)^{-1}\left(x_\star + \frac{1}{\gamma}x_\star\right) \\ \iff & \mathbb{T}x_\star + \frac{1}{\gamma}x_\star = x_\star + \frac{1}{\gamma}x_\star \\ \iff & x_\star = \mathbb{T}x_\star \\ \iff & x_\star \in \operatorname{Fix}\mathbb{T}. \end{aligned}$$

Suppose $\mathbb{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is μ-strongly monotone for $\mu \geq 0$. Define $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ as

$$\mathbb{T} = \left(1 + \frac{1}{1+2\mu}\right)\mathbb{J}_{\mathbb{A}} - \frac{1}{1+2\mu}\mathbb{I}.$$

For any $x, y \in \mathbb{R}^n$, let $u = \mathbb{T}x$ and $v = \mathbb{T}y$. Then

$$
\begin{aligned}
& u = \mathbb{T}x \\
\iff\ & u = \left(1 + \frac{1}{1+2\mu}\right) \mathbb{J}_{\mathbb{A}} x - \frac{1}{1+2\mu} x \\
\iff\ & \frac{1}{1+2\mu} x + u = \left(1 + \frac{1}{1+2\mu}\right) \mathbb{J}_{\mathbb{A}} x \\
\iff\ & \frac{1+2\mu}{2+2\mu} \left(\frac{1}{1+2\mu} x + u\right) = \frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u = \mathbb{J}_{\mathbb{A}} x \\
\iff\ & x \in (\mathbb{I} + \mathbb{A}) \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right) \\
\iff\ & \frac{1+2\mu}{2+2\mu}(x - u) \in \mathbb{A} \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right).
\end{aligned}
$$

Likewise,

$$
\frac{1+2\mu}{2+2\mu}(y - v) \in \mathbb{A} \left(\frac{1}{2+2\mu} y + \frac{1+2\mu}{2+2\mu} v\right).
$$

From the μ-strong monotonicity of $\mathbb{A}$,

$$
\begin{aligned}
& \left\langle \mathbb{A} \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right) - \mathbb{A} \left(\frac{1}{2+2\mu} y + \frac{1+2\mu}{2+2\mu} v\right), \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right) - \left(\frac{1}{2+2\mu} y + \frac{1+2\mu}{2+2\mu} v\right) \right\rangle \\
& \geq \mu \left\| \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right) - \left(\frac{1}{2+2\mu} y + \frac{1+2\mu}{2+2\mu} v\right) \right\|^2 \\
& \implies \left\langle \frac{1+2\mu}{2+2\mu}(x-u) - \frac{1+2\mu}{2+2\mu}(y-v), \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right) - \left(\frac{1}{2+2\mu} y + \frac{1+2\mu}{2+2\mu} v\right) \right\rangle \\
& \qquad \geq \mu \left\| \left(\frac{1}{2+2\mu} x + \frac{1+2\mu}{2+2\mu} u\right) - \left(\frac{1}{2+2\mu} y + \frac{1+2\mu}{2+2\mu} v\right) \right\|^2 \\
& \iff \left\langle \frac{1+2\mu}{2+2\mu}(x-y) - \frac{1+2\mu}{2+2\mu}(u-v), \frac{1}{2+2\mu}(x-y) + \frac{1+2\mu}{2+2\mu}(u-v) \right\rangle \\
& \qquad \geq \mu \left\| \frac{1}{2+2\mu}(x-y) + \frac{1+2\mu}{2+2\mu}(u-v) \right\|^2 \\
& \iff \langle (1+2\mu)(x-y) - (1+2\mu)(u-v), (x-y) + (1+2\mu)(u-v) \rangle \geq \mu \|(x-y) + (1+2\mu)(u-v)\|^2 \\
& \iff (1+\mu)\|x-y\|^2 \geq (1+\mu)(1+2\mu)^2 \|u-v\|^2 \\
& \iff \|u-v\|^2 \leq \frac{1}{(1+2\mu)^2} \|x-y\|^2.
\end{aligned}
$$

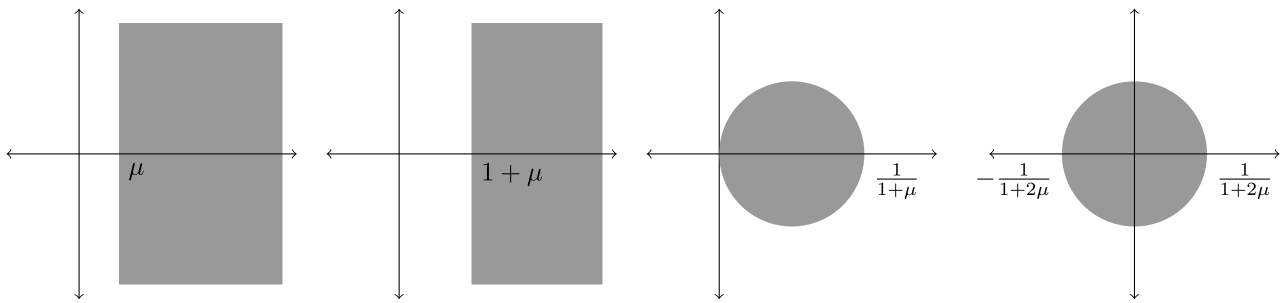
This holds for any $u = \mathbb{T}x$ and $v = \mathbb{T}y$ for any $x, y \in \mathbb{R}^n$, so $\mathbb{T}$ is $\frac{1}{1+2\mu}$-Lipschitz.

Finally, we can also prove that

$$
\begin{aligned}
x_\star \in \operatorname{Fix} \mathbb{T} \iff\ & x_\star = \mathbb{T} x_\star = \left(1 + \frac{1}{1+2\mu}\right) \mathbb{J}_{\mathbb{A}} x_\star - \frac{1}{1+2\mu} x_\star \\
\iff\ & \frac{2+2\mu}{1+2\mu} x_\star = \frac{2+2\mu}{1+2\mu} \mathbb{J}_{\mathbb{A}} x_\star \\
\iff\ & x_\star = \mathbb{J}_{\mathbb{A}} x_\star = (\mathbb{I} + \mathbb{A})^{-1} x_\star \\
\iff\ & x_\star \in x_\star + \mathbb{A} x_\star \\
\iff\ & 0 \in \mathbb{A} x_\star \\
\iff\ & x_\star \in \operatorname{Zer} \mathbb{A}.
\end{aligned}
$$

$\square$

Proof of Lemma 2.1 with scaled relative graph. In this proof, we use the notations of Ryu et al. (2021) for the operator classes, which we list below. Consider a class of operators $\mathcal{M}_\mu$ of μ-strongly monotone operators and $\mathcal{L}_{1/\gamma}$ of $\frac{1}{\gamma}$-contractions. As $\mathcal{M}_\mu$, $\mathcal{L}_{1/\gamma}$ are SRG-full classes, which means that the inclusion of the SRG of some operator to the SRG of an operator class is equivalent to membership of that operator to the given operator class (Ryu et al., 2021, Section 3.3). Instead of showing that the operators satisfy the equivalent inequality condition to the membership, we show the membership in terms of the SRGs.



(a) SRG of $\mathbb{A}$ (b) SRG of $\mathbb{I}+\mathbb{A}$ (c) SRG of $\mathbb{J}_{\mathbb{A}}$ (d) SRG of $\mathbb{T}$

Figure 4. SRG changing with invertible transformation F.

Consider an invertible transformation $F : \mathbb{C}\cup\{\infty\} \to \mathbb{C}\cup\{\infty\}$ defined as

$$F(z) = \left(1 + \frac{1}{1+2\mu}\right)(1+z)^{-1} - \frac{1}{1+2\mu}.$$

F is a composition of only scalar addition/subtraction/multiplication and inversion, therefore preserves the SRG of $\mathcal{M}_\mu$ and $\mathcal{L}_{1/\gamma}$. SRG of $F(\mathcal{M}_\mu)$ and $\mathcal{L}_{1/\gamma}$ match, and the SRG of $F^{-1}(\mathcal{L}_{1/\gamma})$ and $\mathcal{M}_\mu$ match. $\square$

B. Omitted proofs of Section 3

B.1. Proof of Lemma 3.1

Lemma B.1. *The y_k-update in algorithm (OS-PPM) is equivalent to*

$$\begin{aligned} x_k &= \mathbb{J}_{\mathbb{A}} y_{k-1} \\ y_k &= \left(1 - \frac{1}{\varphi_k}\right)\left\{\left(1+\frac{1}{\gamma}\right)x_k - \frac{1}{\gamma}y_{k-1}\right\} + \frac{1}{\varphi_k}y_0 \end{aligned}$$

where $\gamma = 1+2\mu$.

Proof. It suffices to show the equivalence of y_k-iterates. For $k=1$, from (OS-PPM) update,

$$\begin{aligned} y_1 &= x_1 + \frac{\varphi_0 - 1}{\varphi_1}(x_1 - y_0) - \frac{(\gamma-1)\varphi_0}{\varphi_1}(y_0 - x_1) \\ &= x_1 - \frac{\gamma-1}{\gamma^2+1}(y_0 - x_1) && (\varphi_1 = 1+\gamma^2) \\ &= \left(1 - \frac{1}{\varphi_1}\right)\left\{\left(1+\frac{1}{\gamma}\right)x_1 - \frac{1}{\gamma}y_0\right\} + \frac{1}{\varphi_1}y_0. \end{aligned}$$

Assume that the equivalence of the iterates holds for $k = 1, 2, \dots, l$. From the (OS-PPM) update,

$$\begin{aligned}
y_{l+1} &= x_{l+1} + \frac{\varphi_l - 1}{\varphi_{l+1}}(x_{l+1} - x_l) - \frac{(\gamma-1)\varphi_l}{\varphi_{l+1}}(y_l - x_{l+1}) + \frac{\gamma\varphi_{l-1}}{\varphi_{l+1}}(y_{l-1} - x_l) \\
&= \left\{1 + \frac{\varphi_l - 1}{\varphi_{l+1}} + \frac{(\gamma-1)\varphi_l}{\varphi_{l+1}}\right\} x_{l+1} - \left(\frac{\varphi_l - 1}{\varphi_{l+1}} + \frac{\gamma\varphi_{l-1}}{\varphi_{l+1}}\right) x_l - \frac{(\gamma-1)\varphi_l}{\varphi_{l+1}} y_l + \frac{\gamma\varphi_{l-1}}{\varphi_{l+1}} y_{l-1} \\
&= \gamma(\gamma+1)\frac{\varphi_l}{\varphi_{l+1}} x_{l+1} - \gamma(\gamma+1)\frac{\varphi_{l-1}}{\varphi_{l+1}} x_l - \frac{(\gamma-1)\varphi_l}{\varphi_{l+1}} y_l + \frac{\gamma\varphi_{l-1}}{\varphi_{l+1}} y_{l-1}.
\end{aligned}$$

From the inductive hypothesis, we have

$$y_l = \left(1 - \frac{1}{\varphi_l}\right)\left\{\left(1 + \frac{1}{\gamma}\right) x_l - \frac{1}{\gamma} y_{l-1}\right\} + \frac{1}{\varphi_l} y_0,$$

or

$$\gamma\varphi_{l-1} y_{l-1} = \gamma(\gamma+1)\varphi_{l-1} x_l - \varphi_l y_l + y_0.$$

Plugging this into the $\gamma\varphi_{l-1}y_{l-1}$-term in y_{l+1}, we get

$$\begin{aligned}
y_{l+1} &= \gamma(\gamma+1)\frac{\varphi_l}{\varphi_{l+1}} x_{l+1} - \gamma(\gamma+1)\frac{\varphi_{l-1}}{\varphi_{l+1}} x_l - \frac{(\gamma-1)\varphi_l}{\varphi_{l+1}} y_l + \frac{1}{\varphi_{l+1}}\left\{\gamma(\gamma+1)\varphi_{l-1} x_l - \varphi_l y_l + y_0\right\} \\
&= \gamma(\gamma+1)\frac{\varphi_l}{\varphi_{l+1}} x_{l+1} - \gamma\frac{\varphi_l}{\varphi_{l+1}} y_l + \frac{1}{\varphi_{l+1}} y_0 \\
&= \frac{\gamma^2\varphi_l}{\varphi_{l+1}}\left\{\left(1 + \frac{1}{\gamma}\right) x_{l+1} - \frac{1}{\gamma} y_l\right\} + \frac{1}{\varphi_{l+1}} y_0 \\
&= \left(1 - \frac{1}{\varphi_{l+1}}\right)\left\{\left(1 + \frac{1}{\gamma}\right) x_{l+1} - \frac{1}{\gamma} y_l\right\} + \frac{1}{\varphi_{l+1}} y_0.
\end{aligned}$$

The same equivalence holds for y_{l+1}, so we are done. □

Proof of Lemma 3.1. Start from the same initial iterate $y_0 = \tilde{y}_0$. Suppose $y_k = \tilde{y}_k$ for some $k \geq 0$. Then,

$$\begin{aligned}
y_{k+1} &= \left(1 - \frac{1}{\varphi_{k+1}}\right)\left\{\left(1 + \frac{1}{1+2\mu}\right) x_{k+1} - \frac{1}{1+2\mu} y_k\right\} + \frac{1}{\varphi_{k+1}} y_0 && \text{(Lemma B.1)} \\
&= \left(1 - \frac{1}{\varphi_{k+1}}\right)\left\{\left(1 + \frac{1}{1+2\mu}\right)\mathbb{J}_{\mathbb{A}} - \frac{1}{1+2\mu}\mathbb{I}\right\} y_k + \frac{1}{\varphi_{k+1}} y_0 \\
&= \left(1 - \frac{1}{\varphi_{k+1}}\right)\mathbb{T} y_k + \frac{1}{\varphi_{k+1}} y_0 \\
&= \left(1 - \frac{1}{\varphi_{k+1}}\right)\mathbb{T} \tilde{y}_k + \frac{1}{\varphi_{k+1}} y_0 = \tilde{y}_{k+1}. && (y_k = \tilde{y}_k \text{ by induction hypothesis})
\end{aligned}$$

□

B.2. Proof of Theorem 3.2

Recall that

$$\begin{aligned}
V^k = (1 + \gamma^{-k})\Bigg[&\left(\sum_{n=0}^{k-1}\gamma^n\right)^2 \|\tilde{\mathbb{A}} x_k\|^2 + 2\left(\sum_{n=0}^{k-1}\gamma^n\right)\langle \tilde{\mathbb{A}} x_k - \mu(x_k - x_\star), x_k - x_\star\rangle \\
&+ \gamma^{-k}\left\|\left(\sum_{n=0}^{k-1}\gamma^n\right)\tilde{\mathbb{A}} x_k - \gamma^k(x_k - x_\star) + (x_k - y_0)\right\|^2\Bigg] + (1 - \gamma^{-k})\|y_0 - x_\star\|^2. \qquad \text{(OS-PPM-Lyapunov)}
\end{aligned}$$

for $k = 1, 2, \dots, N$ and $V^0 = 2\|y_0 - x_\star\|^2$, where $\gamma = 1 + 2\mu$, $\varphi_k = \sum_{n=0}^{k} \gamma^{2n}$ and $\tilde{\mathbb{A}}x_k = y_{k-1} - x_k \in \mathbb{A}x_k$. We will often use the following identity.

$$(1+\gamma)\varphi_k = (1+\gamma)\sum_{n=0}^{k}\gamma^{2n} = (1+\gamma^{k+1})\sum_{n=0}^{k}\gamma^n.$$

First, we show that V^k has an alternate form as below. This form is useful in proving the monotone decreasing property of V^k in k.

Lemma B.2. *V^k defined in* (OS-PPM-Lyapunov) *can be equivalently written as*

$$V^k = \gamma^{-2k}(1+\gamma)^2\varphi_{k-1}^2\|\tilde{\mathbb{A}}x_k\|^2 + 2\gamma^{-2k}(1+\gamma)\varphi_{k-1}\langle \tilde{\mathbb{A}}x_k - \mu(x_k - y_0), x_k - y_0\rangle + 2\|y_0 - x_\star\|^2.$$

Proof. Expanding the square term,

$$\begin{aligned}
&\left\|\left(\sum_{n=0}^{k-1}\gamma^n\right)\tilde{\mathbb{A}}x_k - \gamma^k(x_k - x_\star) + (x_k - y_0)\right\|^2 \\
&= \left\|\left(\sum_{n=0}^{k-1}\gamma^n\right)\tilde{\mathbb{A}}x_k - (\gamma^k - 1)(x_k - y_0) - \gamma^k(y_0 - x_\star)\right\|^2 \\
&= \left(\sum_{n=0}^{k-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_k\|^2 - 2\left(\sum_{n=0}^{k-1}\gamma^n\right)(\gamma^k-1)\langle\tilde{\mathbb{A}}x_k, x_k - y_0\rangle - 2\left(\sum_{n=0}^{k-1}\gamma^n\right)\gamma^k\langle\tilde{\mathbb{A}}x_k, y_0 - x_\star\rangle \\
&\quad + (\gamma^k-1)^2\|x_k - y_0\|^2 - 2\gamma^k(\gamma^k-1)\langle x_k - y_0, y_0 - x_\star\rangle + \gamma^{2k}\|y_0 - x_\star\|^2.
\end{aligned}$$

Also, we have

$$\begin{aligned}
&\langle\tilde{\mathbb{A}}x_k - \mu(x_k - x_\star), x_k - x_\star\rangle \\
&= \langle\tilde{\mathbb{A}}x_k - \mu(x_k - y_0) - \mu(y_0 - x_\star), (x_k - y_0) + (y_0 - x_\star)\rangle \\
&= \langle\tilde{\mathbb{A}}x_k - \mu(x_k - y_0), x_k - y_0\rangle + \langle\tilde{\mathbb{A}}x_k, y_0 - x_\star\rangle - 2\mu\langle x_k - y_0, y_0 - x_\star\rangle.
\end{aligned}$$

Then V^k is expressed as

$$\begin{aligned}
V^k &= 2(1+\gamma^{-k})\left(\sum_{n=0}^{k-1}\gamma^n\right)\langle\big\{\langle\tilde{\mathbb{A}}x_k - \mu(x_k - y_0), x_k - y_0\rangle + \langle\tilde{\mathbb{A}}x_k, y_0 - x_\star\rangle - 2\mu\langle x_k - y_0, y_0 - x_\star\rangle\big\} \\
&\quad + (1+\gamma^{-k})\gamma^{-k}\Bigg\{\left(\sum_{n=0}^{k-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_k\|^2 - 2\left(\sum_{n=0}^{k-1}\gamma^n\right)(\gamma^k-1)\langle\tilde{\mathbb{A}}x_k, x_k - y_0\rangle + (\gamma^k-1)^2\|x_k - y_0\|^2 \\
&\qquad\qquad - 2\left(\sum_{n=0}^{k-1}\gamma^n\right)\gamma^k\langle\tilde{\mathbb{A}}x_k, y_0 - x_\star\rangle - 2\gamma^k(\gamma^k-1)\langle x_k - y_0, y_0 - x_\star\rangle + \gamma^{2k}\|y_0 - x_\star\|^2\Bigg\} \\
&\quad + (1+\gamma^{-k})\left(\sum_{n=0}^{k-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_k\|^2 + (1-\gamma^{-k})\|y_0 - x_\star\|^2 \\
&= (1+\gamma^{-k})^2\left(\sum_{n=0}^{k-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_k\|^2 + 2(1+\gamma^{-k})\left(\sum_{n=0}^{k-1}\gamma^n\right)\langle\tilde{\mathbb{A}}x_k - \mu(x_k - y_0), x_k - y_0\rangle \\
&\quad - 2\gamma^{-k}(1+\gamma^{-k})(\gamma^k-1)\left(\sum_{n=0}^{k-1}\gamma^n\right)\langle\tilde{\mathbb{A}}x_k, x_k - y_0\rangle + \gamma^{-k}(1+\gamma^{-k})(\gamma^k-1)^2\|x_k - y_0\|^2 + 2\|y_0 - x_\star\|^2 \\
&= (1+\gamma^{-k})^2\left(\sum_{n=0}^{k-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_k\|^2 + 2\gamma^{-k}(1+\gamma^{-k})\left(\sum_{n=0}^{k-1}\gamma^n\right)\langle\tilde{\mathbb{A}}x_k - \mu(x_k - y_0), x_k - y_0\rangle + 2\|y_0 - x_\star\|^2.
\end{aligned}$$

As

$$(1+\gamma^{-k})\left(\sum_{n=0}^{k-1}\gamma^n\right) = \frac{1+\gamma^k}{\gamma^k}\left(\sum_{n=0}^{k-1}\gamma^n\right) = \frac{1+\gamma}{\gamma^k}\varphi_{k-1},$$

we have

$$V^k = \gamma^{-2k}(1+\gamma)^2\varphi_{k-1}^2\|\tilde{\mathbf{A}}x_k\|^2 + 2\gamma^{-2k}(1+\gamma)\varphi_{k-1}\langle\tilde{\mathbf{A}}x_k - \mu(x_k-y_0), x_k-y_0\rangle + 2\|y_0-x_\star\|^2.$$

□

Next, we prove that $\{V^k\}_{k=0}^N$ is monotonically decreasing in k.

Lemma B.3. *For $k=0,1,\dots,N$ with V^k defined as (OS-PPM-Lyapunov), we have*

$$V^N \le V^{N-1} \le \cdots \le V^1 \le V^0.$$

Proof. We use the form of V^k as in Lemma B.2.

$$\begin{aligned}
V^1 - V^0 &= \gamma^{-2}(1+\gamma)^2\|\tilde{\mathbf{A}}x_1\|^2 + 2\gamma^{-2}(1+\gamma)\langle\tilde{\mathbf{A}}x_1 - \mu(x_1-y_0), x_1-y_0\rangle \\
&= \gamma^{-2}(1+\gamma)\left\{(1+\gamma)\|\tilde{\mathbf{A}}x_1\|^2 + 2\langle\tilde{\mathbf{A}}x_1 - \mu(x_1-y_0), x_1-y_0\rangle\right\} \\
&= \gamma^{-2}(1+\gamma)\left\{(1+\gamma)\|\tilde{\mathbf{A}}x_1\|^2 - 2(1+\mu)\|\tilde{\mathbf{A}}x_1\|^2\right\} && (x_1 - y_0 = -\tilde{\mathbf{A}}x_1) \\
&= 0. && (1+\gamma = 2(1+\mu))
\end{aligned}$$

Now, consider $k \ge 1$. Then,

$$\begin{aligned}
V^{k+1} - V^k = \gamma^{-2(k+1)}&(1+\gamma)^2\varphi_k^2\|\tilde{\mathbf{A}}x_{k+1}\|^2 - \gamma^{-2k}(1+\gamma)^2\varphi_{k-1}^2\|\tilde{\mathbf{A}}x_k\|^2 \\
&+ 2\gamma^{-2(k+1)}(1+\gamma)\varphi_k\langle\tilde{\mathbf{A}}x_{k+1} - \mu(x_{k+1}-y_0), x_{k+1}-y_0\rangle \\
&- 2\gamma^{-2k}(1+\gamma)\varphi_{k-1}\langle\tilde{\mathbf{A}}x_k - \mu(x_k-y_0), x_k-y_0\rangle.
\end{aligned}$$

Now, we claim that

$$V^{k+1} - V^k + 2\gamma^{-2k}(1+\gamma)\varphi_k\varphi_{k-1}\langle\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k - \mu(x_{k+1}-x_k), x_{k+1}-x_k\rangle = 0.$$

First,

$$\begin{aligned}
&V^{k+1} - V^k + 2\gamma^{-2k}(1+\gamma)\varphi_k\varphi_{k-1}\langle\tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k - \mu(x_{k+1}-x_k), x_{k+1}-x_k\rangle \\
&= V^{k+1} - V^k + 2\gamma^{-2k}(1+\gamma)\varphi_k\varphi_{k-1}\langle\tilde{\mathbf{A}}x_{k+1} - \mu(x_{k+1}-y_0), x_{k+1}-x_k\rangle \\
&\quad - 2\gamma^{-2k}(1+\gamma)\varphi_k\varphi_{k-1}\langle\tilde{\mathbf{A}}x_k - \mu(x_k-y_0), x_{k+1}-x_k\rangle \\
&= \gamma^{-2(k+1)}(1+\gamma)^2\langle\varphi_k\tilde{\mathbf{A}}x_{k+1} - \gamma\varphi_{k-1}\tilde{\mathbf{A}}x_k, \varphi_k\tilde{\mathbf{A}}x_{k+1} + \gamma\varphi_{k-1}\tilde{\mathbf{A}}x_k\rangle \\
&\quad + 2\gamma^{-2(k+1)}(1+\gamma)\varphi_k\langle\tilde{\mathbf{A}}x_{k+1} - \mu(x_{k+1}-y_0), \gamma^2\varphi_{k-1}(x_{k+1}-x_k) + (x_{k+1}-y_0)\rangle \\
&\quad - 2\gamma^{-2k}(1+\gamma)\varphi_{k-1}\langle\tilde{\mathbf{A}}x_k - \mu(x_k-y_0), \varphi_k(x_{k+1}-x_k) + (x_k-y_0)\rangle.
\end{aligned}$$

From Lemma B.1, we have

$$y_k = \left(1-\frac{1}{\varphi_k}\right)\left\{\left(1+\frac{1}{\gamma}\right)x_k - \frac{1}{\gamma}y_{k-1}\right\} + \frac{1}{\varphi_k}y_0.$$

Using the fact that $y_{k-1} = x_k + \tilde{\mathbf{A}}x_k$, $y_k = x_{k+1} + \tilde{\mathbf{A}}x_{k+1}$, and $\varphi_k = \gamma^2\varphi_{k-1} + 1$, we obtain

$$\varphi_k(x_{k+1}-y_0) + \varphi_k\tilde{\mathbf{A}}x_{k+1} = \gamma^2\varphi_{k-1}(x_k-y_0) - \gamma\varphi_{k-1}\tilde{\mathbf{A}}x_k.$$

Letting $U^k = \varphi_k(x_{k+1} - y_0) - \gamma^2\varphi_{k-1}(x_k - y_0) = -\varphi_k\tilde{\mathbb{A}}x_{k+1} - \gamma\varphi_{k-1}\tilde{\mathbb{A}}x_k$, above formula is simplified as

$$
\begin{aligned}
&V^{k+1} - V^k + 2\gamma^{-2k}(1+\gamma)\varphi_k\varphi_{k-1}\langle \tilde{\mathbb{A}}x_{k+1} - \tilde{\mathbb{A}}x_k - \mu(x_{k+1} - x_k), x_{k+1} - x_k\rangle \\
&= -\gamma^{-2(k+1)}(1+\gamma)^2\langle \varphi_k\tilde{\mathbb{A}}x_{k+1} - \gamma\varphi_{k-1}\tilde{\mathbb{A}}x_k, U_k\rangle \\
&\quad + 2\gamma^{-2(k+1)}(1+\gamma)\varphi_k\langle \tilde{\mathbb{A}}x_{k+1} - \mu(x_{k+1} - y_0), U_k\rangle \\
&\quad - 2\gamma^{-2k}(1+\gamma)\varphi_{k-1}\langle \tilde{\mathbb{A}}x_k - \mu(x_k - y_0), U_k\rangle \\
&= \gamma^{-2(k+1)}(1+\gamma)\Big\langle -(1+\gamma)(\varphi_k\tilde{\mathbb{A}}x_{k+1} - \gamma\varphi_{k-1}\tilde{\mathbb{A}}x_k) + 2\varphi_k\{\tilde{\mathbb{A}}x_{k+1} - \mu(x_{k+1} - y_0)\} \\
&\qquad\qquad - 2\gamma^2\varphi_{k-1}\{\tilde{\mathbb{A}}x_k - \mu(x_k - y_0)\}, U_k\Big\rangle \\
&= \gamma^{-2(k+1)}(1+\gamma)\Big\langle (1-\gamma)(\varphi_k\tilde{\mathbb{A}}x_{k+1} + \gamma\varphi_{k-1}\tilde{\mathbb{A}}x_k) - 2\mu\{\varphi_k(x_{k+1} - y_0) - \gamma^2\varphi_{k-1}(x_k - y_0)\}, U_k\Big\rangle \\
&= \gamma^{-2(k+1)}(1+\gamma)\langle (\gamma - 1)U_k - 2\mu U_k, U_k\rangle = 0 \qquad (\gamma - 1 = 2\mu)
\end{aligned}
$$

□

We now prove Theorem 3.2.

Proof of Theorem 3.2. According to Lemma B.3, we have $V^N \le V^{N-1} \le \cdots \le V^0 = 2\|y_0 - x_\star\|^2$. Therefore,

$$
\begin{aligned}
2\|y_0 - x_\star\|^2 &\ge V^N \\
&= (1+\gamma^{-N})\left(\sum_{n=0}^{N-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_N\|^2 + 2(1+\gamma^{-N})\left(\sum_{n=0}^{N-1}\gamma^n\right)\langle \tilde{\mathbb{A}}x_N - \mu(x_N - x_\star), x_N - x_\star\rangle \\
&\quad + \gamma^{-N}(1+\gamma^{-N})\left\|\left(\sum_{n=0}^{N-1}\gamma^n\right)\tilde{\mathbb{A}}x_N - \gamma^N(x_N - x_\star) + (x_N - y_0)\right\|^2 + (1-\gamma^{-N})\|y_0 - x_\star\|^2 \\
&\ge (1+\gamma^{-N})\left(\sum_{n=0}^{N-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_N\|^2 + (1-\gamma^{-N})\|y_0 - x_\star\|^2,
\end{aligned}
$$

which can be simplified as

$$
(1+\gamma^{-N})\|y_0 - x_\star\|^2 \ge (1+\gamma^{-N})\left(\sum_{n=0}^{N-1}\gamma^n\right)^2\|\tilde{\mathbb{A}}x_N\|^2,
$$

or equivalently,

$$
\|\tilde{\mathbb{A}}x_N\|^2 \le \left(\frac{1}{\sum_{n=0}^{N-1}\gamma^n}\right)^2\|y_0 - x_\star\|^2.
$$

□

Proof of Corollary 3.3. This immediately follows from Theorem 3.2 and Lemma 3.1 by

$$
\tilde{\mathbb{A}}x_N = y_{N-1} - x_N = \left(1 + \frac{1}{\gamma}\right)^{-1}(y_{N-1} - \mathbb{T}y_{N-1}) \in \mathbb{A}x_N.
$$

□

C. Details on the formulation of performance estimation problem for (OS-PPM)

In order to obtain an estimate on the worst-case complexity of the algorithm, performance estimation problem (PEP) technique solves a certain form of semidefinite problem (SDP). This SDP holds positive semidefinite matrix as an optimization variable, and solves the problem under constraints formulated from the interpolation condition of an operator in hand.

When discovering (OS-PPM), we used maximal monotonicity as our interpolation condition, just as in Ryu et al. (2020); Kim (2021). We further extended this to cover the case of maximal strongly-monotone operators, in a slightly different way with Taylor & Drori (2021) who considered strongly convex interpolation. The optimization variable is a positive semidefinite matrix, and this is of a Gram matrix form which stores information on the iterates of algorithms. Usual choice of basis vectors for the gram matrix in PEP is usually $\nabla f(x)$ for convex minimization setup (Kim & Fessler, 2016b; Taylor et al., 2018; Taylor & Drori, 2021), or $\tilde{\mathbb{A}}x$ for operator setup (Kim, 2021). Here, we used x-iterates to form the gram matrix of SDP.

This basic SDP is a primal problem of the PEP (Primal-PEP), and solving this returns an estimate to the worst-case complexity of given algorithm. If we form a dual problem (dual-PEP) and minimize the optimal value of dual-PEP over possible choices of stepsizes as in Kim & Fessler (2016b); Taylor et al. (2018); Kim (2021); Taylor & Drori (2021), this provides possibly the fastest rate, and solution to this minimization problem gives possibly optimal algorithms. We considered a class of algorithms satisfying the span assumption in Corollary 4.2, and obtained (OS-PPM).

D. Omitted proofs of Section 4

D.1. Proving complexity lower bound with span condition

Proof of Lemma 4.3 with inequalities. From (Bauschke & Combettes, 2017, Proposition 4.35), $\mathbb{G}$ is $\frac{1}{1+\gamma}$-averaged if and only if

$$\|\mathbb{G}x - \mathbb{G}y\|^2 + \frac{\gamma-1}{\gamma+1}\|x-y\|^2 \le \frac{2\gamma}{1+\gamma}\langle \mathbb{G}x - \mathbb{G}y, x-y\rangle, \qquad \forall x, y \in \mathbb{R}^n.$$

Then for any $x, y \in \mathbb{R}^n$, we get the chain of equivalences as follows.

$$\begin{aligned}
\|\mathbb{T}x - \mathbb{T}y\|^2 \le \frac{1}{\gamma^2}\|x-y\|^2 &\iff \|\gamma\mathbb{T}x - \gamma\mathbb{T}y\|^2 \le \|x-y\|^2 \\
&\iff \|\{(1+\gamma)\mathbb{G}x - \gamma x\} - \{(1+\gamma)\mathbb{G}y - \gamma y\}\|^2 \le \|x-y\|^2 \\
&\iff (1+\gamma)^2\|\mathbb{G}x - \mathbb{G}y\|^2 - 2\gamma(1+\gamma)\langle \mathbb{G}x - \mathbb{G}y, x-y\rangle + \gamma^2\|x-y\|^2 \le \|x-y\|^2 \\
&\iff (1+\gamma)^2\|\mathbb{G}x - \mathbb{G}y\|^2 + (\gamma^2-1)\|x-y\|^2 \le 2\gamma(1+\gamma)\langle \mathbb{G}x - \mathbb{G}y, x-y\rangle \\
&\iff \|\mathbb{G}x - \mathbb{G}y\|^2 + \frac{\gamma-1}{\gamma+1}\|x-y\|^2 \le \frac{2\gamma}{\gamma+1}\langle \mathbb{G}x - \mathbb{G}y, x-y\rangle. \qquad (\because 1+\gamma > 0)
\end{aligned}$$

Therefore, $\mathbb{T}$ is $\frac{1}{\gamma}$-contractive if and only if $\mathbb{G}$ is $\frac{1}{1+\gamma}$-averaged. □

Proof of Lemma 4.3 with scaled relative graph. Using the notion of SRG (Ryu et al., 2021), we get the following equivalence of SRGs. Here, $\mathcal{N}_{\frac{1}{1+\gamma}}$ is a class of $\frac{1}{1+\gamma}$-averaged operators. Therefore, we get the chain of equivalences

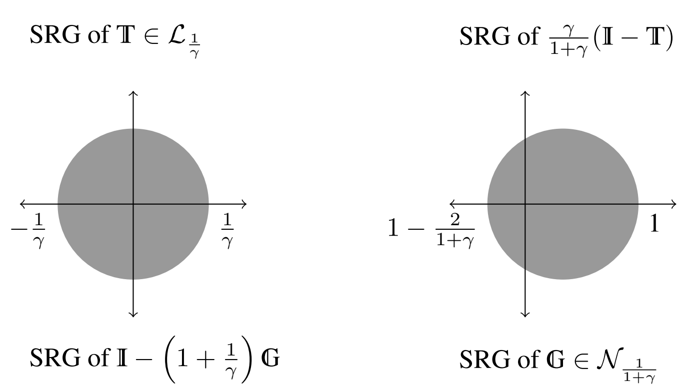


Figure 5. SRG of $\mathbb{T}$ and $\mathbb{G}$

$$\begin{aligned}\mathbb{T} \in \mathcal{L}_{1/\gamma} \iff \gamma\mathbb{T} \in \mathcal{L}_1 \iff -\gamma\mathbb{T} \in \mathcal{L}_1 \\ \iff \mathbb{G} = \frac{\gamma}{1+\gamma}\mathbb{I} + \frac{1}{1+\gamma}(-\gamma\mathbb{T}) \in \mathcal{N}_{\frac{1}{1+\gamma}},\end{aligned}$$

and conclude that $\mathbb{T}$ is $\frac{1}{\gamma}$-Lipschitz if and only if $\mathbb{G}$ is $\frac{1}{1+\gamma}$-averaged. □

Proof of Lemma 4.4. We restate the definition of $\mathbb{N}\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$.

$$\mathbb{N}x = \mathbb{N}(x_1, x_2, \dots, x_{N+1}) = (x_{N+1}, -x_1, \dots, -x_N) - \frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}}Re_1, \qquad x \in \mathbb{R}^{N+1}.$$

For any $x, y \in \mathbb{R}^{N+1}$ such that

$$x = (x_1, x_2, \dots, x_{N+1}), \qquad y = (y_1, y_2, \dots, y_{N+1}),$$

we have

$$\begin{aligned}\|\mathbb{N}x - \mathbb{N}y\|^2 &= \|(x_{N+1}, -x_1, \dots, -x_N) - (y_{N+1}, -y_1, \dots, -y_N)\|^2 \\ &= (x_{N+1} - y_{N+1})^2 + (x_1 - y_1)^2 + \cdots + (x_N - y_N)^2 \\ &= \|x - y\|^2.\end{aligned}$$

Then $\mathbb{N}$ is nonexpansive, and by definition, $\mathbb{G} = \frac{1}{1+\gamma}\mathbb{N} + \frac{\gamma}{1+\gamma}\mathbb{I}$ is a $\frac{1}{1+\gamma}$-averaged operator. □

Proof of Lemma 4.5. By the definition of $\mathbb{G}\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$, for any $x \in \mathbb{R}^{N+1}$,

$$\begin{aligned}\mathbb{G}x &= \frac{1}{1+\gamma}\mathbb{N}x + \frac{\gamma}{1+\gamma}x \\ &= \underbrace{\frac{1}{1+\gamma}\begin{bmatrix} \gamma & 0 & 0 & \dots & 0 & 1 \\ -1 & \gamma & 0 & \dots & 0 & 0 \\ 0 & -1 & \gamma & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \gamma & 0 \\ 0 & 0 & 0 & \dots & -1 & \gamma \end{bmatrix}}_{=H} x - \underbrace{\frac{1}{1+\gamma}\frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}}Re_1}_{=b}\end{aligned}$$

where $\gamma = 1 + 2\mu$. Observe that $\mathbb{G}e_k \in \operatorname{span}\{e_1, e_k, e_{k+1}\}$ for $k = 1, \dots, N$.

We use induction on k to prove the Lemma. The claim holds for $k = 0$ from

$$\mathbb{G}y_0 = \mathbb{G}\mathbf{0} = -\frac{1}{1+\gamma}\frac{1+\gamma^{N+1}}{\sqrt{1+\gamma^2+\cdots+\gamma^{2N}}}Re_1 \in \operatorname{span}\{e_1\}.$$

Now, suppose that the claim holds for $k < N$, i.e.,

$$\begin{aligned}y_k &\in \operatorname{span}\{e_1, e_2, \dots, e_k\} \\ \mathbb{G}y_k &\in \operatorname{span}\{e_1, e_2, \dots, e_{k+1}\}.\end{aligned}$$

Then

$$\begin{aligned}y_{k+1} &\in y_0 + \operatorname{span}\{\mathbb{G}y_0, \mathbb{G}y_1, \dots, \mathbb{G}y_k\} \\ &\subseteq \operatorname{span}\{e_1, e_2, \dots, e_{k+1}\} \\ \mathbb{G}y_{k+1} &= Hy_{k+1} - b \\ &\in H\operatorname{span}\{e_1, e_2, \dots, e_{k+1}\} - b \\ &\subseteq \operatorname{span}\{e_1, e_2, \dots, e_{k+2}\}.\end{aligned}$$

□

Proof of Theorem 4.1. The proof outline of Theorem 4.1 in Section 4.2 is complete except for the part that the identity $(*)$ holds, and that Theorem 4.1 holds for any initial point $y_0 \in \mathbb{R}^n$ which is not necessarily zero.

First, we show that for any initial point $y_0 \in \mathbb{R}^n$, there exists an worst-case operator $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ which cannot exhibit better than the desired rate. Denote by $\mathbb{T}_0\colon \mathbb{R}^n \to \mathbb{R}^n$ the worst-case operator constructed in the proof of Theorem 4.1 for $y_0 = 0$. Define $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ as

$$\mathbb{T}y = \mathbb{T}_0(y - y_0) + y_0$$

given $y_0 \in \mathbb{R}^n$. Then, first of all, the fixed point of $\mathbb{T}$ is $y_\star = \tilde{y}_\star + y_0$ where $\tilde{y}_\star$ is the unique solution of $\mathbb{T}_0$. Also, if $\{y_k\}_{k=0}^N$ satisfies the span condition

$$y_k \in y_0 + \operatorname{span}\{y_0 - \mathbb{T}y_0, \dots, y_{k-1} - \mathbb{T}y_{k-1}\}, \qquad k = 1, \dots, N,$$

then $\tilde{y}_k = y_k - y_0$ forms a sequence satisfying

$$\tilde{y}_k \in \underbrace{\tilde{y}_0}_{=0} + \operatorname{span}\{\tilde{y}_0 - \mathbb{T}_0\tilde{y}_0, \dots, \tilde{y}_{k-1} - \mathbb{T}_0\tilde{y}_{k-1}\}, \qquad k = 1, \dots, N,$$

which is the same span condition in Theorem 4.1 with respect to $\mathbb{T}_0$. This is true from the fact that

$$y_k - \mathbb{T}y_k = \underbrace{y_k - y_0}_{=\tilde{y}_k} + \mathbb{T}_0(\underbrace{y_k - y_0}_{\tilde{y}_k}) = \tilde{y}_k - \mathbb{T}_0\tilde{y}_k$$

for $k = 1, \dots, N$.

Now, $\{\tilde{y}_k\}_{k=0}^N$ is a sequence starting from $\tilde{y}_0 = 0$ satisfying the span condition for $\mathbb{T}_0$. This implies that,

$$\begin{aligned}
\|y_N - \mathbb{T}y_N\|^2 &= \|\tilde{y}_N - \mathbb{T}_0\tilde{y}_N\|^2 \\
&\geq \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 \|\tilde{y}_0 - \tilde{y}_\star\|^2 \\
&= \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 \|y_0 - y_\star\|^2.
\end{aligned}$$

$\mathbb{T}$ is our desired worst-case $\frac{1}{\gamma}$-contraction on $\mathbb{R}^n$.

It remains to show that

$$\|\mathbb{G}y_N\|^2 \geq \left\|\mathcal{P}_{\operatorname{span}\{v\}}(b)\right\|^2 = \left\|\frac{\langle b, v\rangle}{\langle v, v\rangle} v\right\|^2 \overset{(*)}{=} \left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 R^2$$

where

$$v = \begin{bmatrix} 1 & \gamma & \gamma^2 & \dots & \gamma^N \end{bmatrix}^\intercal,$$

especially the identity $(*)$.

$$\begin{aligned}
\left\|\frac{\langle b, v\rangle}{\langle v, v\rangle} v\right\|^2 &= \frac{|\langle b, v\rangle|^2}{\|v\|^2} \\
&= \left(\frac{R}{1+\gamma} \times \frac{1 + \gamma^{N+1}}{\sqrt{1 + \gamma^2 + \gamma^4 + \dots + \gamma^{2N}}}\right)^2 \times \frac{1}{1 + \gamma^2 + \gamma^4 + \dots + \gamma^{2N}} \\
&= \left(\frac{R}{1+\gamma} \times \frac{1 + \gamma^{N+1}}{1 + \gamma^2 + \gamma^4 + \dots + \gamma^{2N}}\right)^2 \\
&= \left(\frac{R}{1 + \gamma + \gamma^2 + \dots + \gamma^N}\right)^2 \\
&= \left(\frac{1}{\sum_{k=0}^N \gamma^k}\right)^2 R^2.
\end{aligned}$$

$\square$

Proof of Corollary 4.2. According to Lemma 2.1, $\mathbb{T}$ is $1/\gamma$-contractive if and only if $\mathbb{A} = (\mathbb{T} + 1/\gamma\mathbb{I})^{-1} - \mathbb{I}$ is $\frac{\gamma-1}{2}$-strongly monotone. For any $y \in \mathbb{R}^n$, if $x = \mathbb{J}_{\mathbb{A}} y$, then

$$y - \mathbb{T}y = y - \left\{ \left(1 + \frac{1}{\gamma}\right) \mathbb{J}_{\mathbb{A}} y - \frac{1}{\gamma} y \right\} = \left(1 + \frac{1}{\gamma}\right)(y - x) = \left(1 + \frac{1}{\gamma}\right) \tilde{\mathbb{A}} x.$$

This implies that

$$y_k \in y_0 + \operatorname{span}\{y_0 - \mathbb{T}y_0, y_1 - \mathbb{T}y_1, \dots, y_{k-1} - \mathbb{T}y_{k-1}\}, \qquad k = 1, \dots, N,$$

if and only if

$$\begin{aligned} &x_k = \mathbb{J}_{\mathbb{A}} y_{k-1} \\ &y_k \in y_0 + \operatorname{span}\{\tilde{\mathbb{A}} x_1, \dots, \tilde{\mathbb{A}} x_k\}, \qquad k = 1, \dots, N \end{aligned}$$

where $x_k = \mathbb{J}_{\mathbb{A}} y_{k-1}$. Span conditions in the statements of Theorem 4.1 and Corollary 4.2 are equivalent under the transformation $\mathbb{A} = (\mathbb{T} + 1/\gamma\mathbb{I})^{-1} - \mathbb{I}$. Therefore, the lower bound result of this corollary can be derived from the lower bound result of Theorem 4.1. $\square$

D.2. Deterministic algorithm classes

In this section, we provide basic terminologies and necessary concepts in proving the complexity lower bound result for general algorithms. We follow the *information-based complexity* framework developed by Nemirovski & Yudin (1983), and use the resisting oracle technique to extend the results of Theorem 4.1 and Corollary 4.2 to general fixed-point iterations and general proximal point methods. The proof itself is motivated by the works of Carmon et al. (2020; 2021), and large portion of the definitions and notations are due to their work.

In the information-based complexity framework, every iterate $\{y_k\}_{k \in \mathbb{N}}$ is a query from an *information oracle*, which returns restrictive information on a given function or operator. Then, assumptions on the algorithm, such as linear span condition, illustrates how it uses such information. For instance, provided with a gradient oracle $\mathcal{O}_f(x) = \nabla f(x)$ of convex function f to be minimized, usually the first-order algorithms search within the span of previous gradients to reach the next iterate.

A *deterministic fixed-point iteration* $\mathbf{A}$ is a mapping of an initial point y_0 and an operator $\mathbb{T}$ to a sequence of iterates $\{y_t\}_{t \in \mathbb{N}}$ and $\{\bar{y}_t\}_{t \in \mathbb{N}}$, such that the output depends on $\mathbb{T}$ only through the *fixed-point residual oracle* $\mathcal{O}_{\mathbb{T}}(y) = y - \mathbb{T}y$. Here, 'deterministic' means that given the same initial point y_0 and the sequence of oracle evaluations $\{\mathcal{O}_{\mathbb{T}}(y_t)\}_{t \in \mathbb{N}}$, the algorithm yields the same sequence of iterates $\{(y_t, \bar{y}_t)\}_{t \in \mathbb{N}}$. More precisely, we define $\mathbf{A}$ per iteration by setting $\mathbf{A} = \{\mathbf{A}_t\}_{t \in \mathbb{N}}$ with

$$(y_t, \bar{y}_t) = \mathbf{A}_t[y_0; \mathbb{T}] = \mathbf{A}_t[y_0, \mathcal{O}_{\mathbb{T}}(y_0), \dots, \mathcal{O}_{\mathbb{T}}(y_{t-1})],$$

where y_t is the t-th *query point* and $\bar{y}_t$ is the t-th *approximate solution* produced by $\mathbf{A}_t$. Here, we consider the algorithms whose query points and approximate solutions are identical ($y_t = \bar{y}_t$).

Even though the $\mathbf{A}$ is defined to produce infinitely many y_t- and $\bar{y}_t$-iterates, the definition includes the case where algorithm terminates at a predetermined total iteration count N, i.e., the algorithm may have a predetermined iteration count N and the behavior may depend on the specified value of N. In such cases, $y_N = \bar{y}_N = y_{N+1} = \bar{y}_{N+1} = \cdots$.

Similarly, a *deterministic proximal point method* $\mathbf{A}$ is a mapping of an initial point y_0 and a maximal monotone operator $\mathbb{A}$ to a sequence of query points $\{y_t\}_{t \in \mathbb{N}}$ and approximate solutions $\{\bar{y}_t\}_{t \in \mathbb{N}}$, such that the output depends on $\mathbb{A}$ only through the resolvent residual oracle $\mathcal{O}_{\mathbb{A}}(y) = y - \mathbb{J}_{\mathbb{A}} y = \tilde{\mathbb{A}} x \in \mathbb{A} x$ where $x = \mathbb{J}_{\mathbb{A}} y$. Indeed, this method $\mathbf{A}$ yields the same sequence of iterates given the same initial point y_0 and oracle evaluations $\{\mathcal{O}_{\mathbb{A}}(y_t)\}_{t \in \mathbb{N}}$.

D.3. Generalized complexity lower bound

As mentioned earlier, the general deterministic fixed-point iterations have no accounts for the span condition. We use the resisting oracle technique (Nemirovski & Yudin, 1983) to prove the lower bound result for general deterministic fixed-point iterations. Recall that Theorem 4.6 is

Theorem 4.6 (Complexity lower bound of general deterministic fixed-point iterations)**.** *Let $n \geq 2N$ for $N \in \mathbb{N}$. For any deterministic fixed-point iteration $\mathbf{A}$ and any initial point $y_0 \in \mathbb{R}^n$, there exists a $\frac{1}{\gamma}$-Lipschitz operator $\mathbb{T} \colon \mathbb{R}^n \to \mathbb{R}^n$ with a* *fixed point* $y_\star \in \operatorname{Fix}\mathbb{T}$ *such that*

$$\|y_N - \mathbb{T}y_N\|^2 \geq \left(1+\frac{1}{\gamma}\right)^2 \left(\frac{1}{\sum_{k=0}^{N}\gamma^k}\right)^2 \|y_0 - y_\star\|^2$$

where $\{y_t\}_{t\in\mathbb{N}} = \mathbf{A}[y_0;\mathbb{T}]$.

By the equivalency of the optimization problems and algorithms stated in Lemma 2.1 and Lemma 3.1, Theorem 4.6 also generalizes Corollary 4.2 to general proximal point methods.

Corollary D.1 (Complexity lower bound of general proximal point methods)**.** *Let* $n \geq 2N-2$ *for* $N \in \mathbb{N}$*. For any deterministic proximal point method* $\mathbf{A}$ *and arbitrary initial point* $y_0 \in \mathbb{R}^n$*, there exists a* μ*-strongly monotone operator* $\mathbb{A}\colon \mathbb{R}^n \to \mathbb{R}^n$ *with a zero* $x_\star \in \operatorname{Zer}\mathbb{A}$ *such that*

$$\|\tilde{\mathbb{A}}x_N\|^2 \geq \left(\frac{1}{1+\gamma+\cdots+\gamma^{N-1}}\right)^2 \|y_0 - x_\star\|^2$$

where $\{y_t\}_{t\in\mathbb{N}} = \mathbf{A}[y_0;\mathbb{T}]$.

D.4. Proof of Theorem 4.6

In order to prove Theorem 4.6, we first extend the result of Theorem 4.1 to the *zero-respecting* sequences, which is a requirement slightly more general than the span assumption. The worst-case operator of Theorem 4.1 covers the case of zero-respecting sequences, and this result will be successfully extended to general deterministic fixed-point iterations.

We say that a sequence $\{z_t\}_{t\in\mathbb{N}\cup\{0\}} \subseteq \mathbb{R}^d$ is *zero-respecting with respect to* $\mathbb{T}$ if

$$\operatorname{supp}\{z_t\} \subseteq \cup_{s<t}\operatorname{supp}\{z_s - \mathbb{T}z_s\}$$

for every $t \in \mathbb{N}\cup\{0\}$, where $\operatorname{supp}\{z\} := \{i \in [d] \mid \langle z, e_i\rangle \neq 0\}$. An deterministic fixed-point iteration $\mathbf{A}$ is called *zero-respecting* if $\mathbf{A}$ generates a sequence $\{z_t\}_{t\in\mathbb{N}\cup\{0\}}$ which is zero-respecting with respect to $\mathbb{T}$ for any nonexpansive $\mathbb{T}\colon \mathbb{R}^d \to \mathbb{R}^d$. Note that by definition, $z_0 = 0$. And for notational simplicity, define $\operatorname{supp}V = \bigcup_{z\in V}\operatorname{supp}\{z\}$.

This property serves as an important intermediate step to the generalization of Theorem 4.1, where its similar form called 'zero-chain' has numerously appeared on the relevant references in convex optimization (Nesterov, 2004; Drori, 2017; Carmon et al., 2020; Drori & Taylor, 2022). The worst-case operator found in the proof of Theorem 4.1 still performs the best among all the zero-respecting query points with respect to $\mathbb{T}$, according to the following lemma.

Lemma D.2. *Let* $\mathbb{T}\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ *be the worst-case operator defined in the proof of Theorem 4.1. If the iterates* $\{z_t\}_{t=0}^N$ *are zero-respecting with respect to* $\mathbb{T}$*,*

$$\|z_N - \mathbb{T}z_N\|^2 \geq \left(1+\frac{1}{\gamma}\right)^2 \left(\frac{1}{1+\gamma+\cdots+\gamma^N}\right)^2 \|z_0 - z_\star\|^2$$

for $z_\star \in \operatorname{Fix}\mathbb{T}$.

Proof. Let $\mathbb{G}$ be defined as in the proof of Theorem 4.1. Then we have

$$z \in \operatorname{span}\{e_1, e_2, \ldots, e_k\} \implies \mathbb{G}z \in \operatorname{span}\{e_1, e_2, \ldots, e_{k+1}\}.$$

We claim that any zero-respecting sequence $\{z_k\}_{k=0,1,\ldots,N}$ satisfies

$$\begin{aligned} z_k &\in \operatorname{span}\big\{e_1, e_2, \ldots, e_k\big\} \\ \mathbb{G}z_k &= \frac{\gamma}{1+\gamma}(z_k - \mathbb{T}z_k) \in \operatorname{span}\big\{e_1, e_2, \ldots, e_{k+1}\big\} \end{aligned}$$

for $k = 0, 1, \ldots, N$, so that the lower bound result of Theorem 4.1 is applicable.

If $k = 0$, then $y_0 = 0$ and from this, $\mathbb{G}0 \in \operatorname{span}\{e_1\}$. So the case of $k = 0$ holds. Now, suppose that $0 < k \le N$ and the claim holds for all $n < k$. Then $\mathbb{G}z_n \in \operatorname{span}\{e_1, \dots, e_{n+1}\} \subseteq \operatorname{span}\{e_1, \dots, e_k\}$ for $0 \le k < n$. $\{z_k\}_{k=0}^N$ is zero-respecting with respect to $\mathbb{T}$, so

$$\begin{aligned}\operatorname{supp}\{z_k\} &\subseteq \bigcup_{n<k} \operatorname{supp}\{z_n - \mathbb{T}z_n\} \\ &= \operatorname{supp}\big\{\mathbb{G}z_0, \mathbb{G}z_1, \dots, \mathbb{G}z_{k-1}\big\} \\ &\subseteq \operatorname{supp}\{e_1, e_2, \dots, e_k\}.\end{aligned}$$

Therefore, $z_k \in \operatorname{span}\{e_1, e_2, \dots, e_k\}$, and $\mathbb{G}z_k \in \operatorname{span}\{e_1, e_2, \dots, e_{k+1}\}$. The claim holds for $k = 1, \dots, N$.

According to the proof of Theorem 4.1,

$$\|z_N - \mathbb{T}z_N\|^2 \ge \left(1 + \frac{1}{\gamma}\right)^2 \left(\frac{1}{1 + \gamma + \dots + \gamma^N}\right)^2 \|z_0 - z_\star\|^2$$

for any zero-respecting iterates $\{z_k\}_{k=0}^N$ with respect to $\mathbb{T}$. $\square$

We say that a matrix $U \in \mathbb{R}^{m\times n}$ with $m \ge n$ is *orthogonal*, if each columns $\{u_i\}_{i=1}^n \subseteq \mathbb{R}^m$ of U as in

$$U = \begin{bmatrix} | & \dots & | \\ u_1 & \dots & u_n \\ | & \dots & | \end{bmatrix}$$

are orthonormal to each other, or in other words, $U^\intercal U = I_n$. It directly follows that $UU^\intercal$ is an orthogonal projection from $\mathbb{R}^m$ to the range $\mathcal{R}(U)$ of U.

Lemma D.3. *For any orthogonal matrix $U \in \mathbb{R}^{m\times n}$ with $m \ge n$ and any arbitrary vector $y_0 \in \mathbb{R}^m$, if $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ is a $\frac{1}{\gamma}$-contractive operator with $\gamma \ge 1$, then $\mathbb{T}_U : \mathbb{R}^m \to \mathbb{R}^m$ defined as*

$$\mathbb{T}_U(y) := U\mathbb{T}U^\intercal(y - y_0) + y_0, \qquad \forall y \in \mathbb{R}^m$$

is also a $\frac{1}{\gamma}$-contractive operator. Furthermore, $z_\star \in \operatorname{Fix}\mathbb{T}$ if and only if $y_\star = y_0 + Uz_\star \in \operatorname{Fix}\mathbb{T}_U$.

Proof. For any $x, z \in \mathbb{R}^m$,

$$\begin{aligned}\|\mathbb{T}_U x - \mathbb{T}_U z\| &= \|U\mathbb{T}U^\intercal(x - y_0) - U\mathbb{T}U^\intercal(z - y_0)\| \\ &= \|\mathbb{T}U^\intercal(x - y_0) - \mathbb{T}U^\intercal(z - y_0)\| && (U \text{ is an orthogonal matrix}) \\ &\le \frac{1}{\gamma}\|U^\intercal(x - y_0) - U^\intercal(z - y_0)\| && (\mathbb{T} \text{ is } \tfrac{1}{\gamma}\text{-contractive}) \\ &= \frac{1}{\gamma}\|UU^\intercal(x - z)\| \\ &\le \frac{1}{\gamma}\|x - z\|. && (UU^\intercal \text{ is an orthogonal projection onto } \mathcal{R}(U))\end{aligned}$$

Now, suppose $z_\star$ is a fixed point of $\mathbb{T}$. Then

$$\begin{aligned}\mathbb{T}_U(y_\star) &= U\mathbb{T}U^\intercal Uz_\star + y_0 = U\mathbb{T}z_\star + y_0 \\ &= Uz_\star + y_0 = y_\star\end{aligned}$$

so $y_\star$ is a fixed point of $\mathbb{T}_U$. On the other hand, if $y_\star$ is a fixed point of $\mathbb{T}_U$, then $z_\star = U^\intercal(y_\star - y_0)$ satisfies

$$\begin{aligned}\mathbb{T}(z_\star) &= \mathbb{T}U^\intercal(y_\star - y_0) \\ &= U^\intercal U\mathbb{T}U^\intercal(y_\star - y_0) && (U^\intercal U = I_n) \\ &= U^\intercal(\mathbb{T}_U y_\star - y_0) \\ &= U^\intercal(y_\star - y_0) = z_\star && (y_\star \in \operatorname{Fix}\mathbb{T}_U)\end{aligned}$$

so it is a fixed point of $\mathbb{T}$. $\square$

Lemma D.4. *Let $\mathbf{A}$ be a general deterministic fixed-point iteration, and $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ be a $1/\gamma$-contractive operator. For $m \geq n+N-1$ and any arbitrary point $y_0 \in \mathbb{R}^m$, there exists an orthogonal matrix $U \in \mathbb{R}^{m\times n}$ and the iterates $\{y_t\}_{t=1}^N = \mathbf{A}[y_0; \mathbb{T}_U]$ with the following properties.*

(i) Let $z^{(t)} := U^\intercal(y_t - y_0)$ for $t = 0, 1, \dots, N$. Then $\{z^{(t)}\}_{t=0}^N$ is zero-respecting with respect to $\mathbb{T}$.

(ii) $\{z^{(t)}\}_{t=0}^N$ satisfies

$$\|z^{(t)} - \mathbb{T}z^{(t)}\| \leq \|y_t - \mathbb{T}_U y_t\|, \qquad t = 0, \dots, N.$$

Proof. We first show that (i) implies (ii). From (i), we know that $z^{(t)} = U^\intercal(y_t - y_0)$ for $t = 0, 1, \dots, N$. Therefore,

$$\begin{aligned}
\|z^{(t)} - \mathbb{T}z^{(t)}\| &= \|U^\intercal(y_t - y_0) - \mathbb{T}U^\intercal(y_t - y_0)\| \\
&= \|UU^\intercal\{(y_t - y_0) - U\mathbb{T}U^\intercal(y_t - y_0)\}\| && (U \text{ is orthogonal}) \\
&= \|UU^\intercal(y_t - y_0) - UU^\intercal U\mathbb{T}U^\intercal(y_t - y_0)\| && (U^\intercal U = I_n) \\
&= \|UU^\intercal\{(y_t - y_0) - U\mathbb{T}U^\intercal(y_t - y_0)\}\| \\
&\leq \|(y_t - y_0) - U\mathbb{T}U^\intercal(y_t - y_0)\| && (UU^\intercal \text{ is an orthogonal projection}) \\
&= \|y_t - \mathbb{T}_U y_t\|. && (\text{Definition of } \mathbb{T}_U)
\end{aligned}$$

Now we prove the existence of orthogonal $U \in \mathbb{R}^{m\times n}$ with $\{y_t\}_{t=0}^N = \mathbf{A}[y_0; \mathbb{T}_U]$ and (i) holds. In order to show the existence of such orthogonal matrix U as in (i), we provide the inductive scheme that finds the columns of U at each iteration. Before describing the actual scheme, we first provide some observations useful to deriving the necessary conditions for the columns $\{u_i\}_{i=1}^n$ of U to satisfy.

Let $t \in \{1, \dots, N\}$, and define the set of indices S_t as

$$S_t = \cup_{s<t} \mathrm{supp}\{z^{(s)} - \mathbb{T}z^{(s)}\}.$$

For $\{z^{(t)}\}_{t=0}^N$ to satisfy the zero-respecting property with respect to $\mathbb{T}$, $z^{(t)}$ is required to satisfy

$$\mathrm{supp}\{z^{(t)}\} \subseteq S_t$$

for $t = 1, \dots, N$. This requirement is fulfilled when

$$y_t - y_0 \in \mathrm{span}\{u_i\}_{i\in S_t}$$

or equivalently,

$$\langle u_i, y_t - y_0\rangle = 0$$

for every $i \notin S_t$. Note that $z^{(0)} = U^\intercal(y_0 - y_0) = 0$ is trivial.

We now construct $U \in \mathbb{R}^{m\times n}$. Note that $S_0 = \emptyset \subseteq S_1 \subseteq \cdots \subseteq S_t$. $\{u_i\}_{i\in S_t\setminus S_{t-1}}$ is chosen inductively starting from $t = 1$. Suppose we have already chosen $\{u_i\}_{i\in S_{t-1}}$. Choose $\{u_i\}_{i\in S_t\setminus S_{t-1}}$ from the orthogonal complement of

$$W_t := \mathrm{span}\left(\{y_1 - y_0, \cdots, y_{t-1} - y_0\} \cup \{u_i\}_{i\in S_{t-1}}\right)$$

and let them be orthogonal to each other. In case of $S_N \neq \emptyset$, for $i \notin S_N$, choose proper vectors u_i so that U becomes an orthogonal matrix. This is possible when the dimension of $W_t^\perp$ is large enough to draw $|S_t \setminus S_{t-1}|$-many orthogonal vectors, or in other words,

$$\dim W_t^\perp \geq |S_t \setminus S_{t-1}|.$$

From the assumption, $m - t + 1 \geq m - N + 1 \geq n$, so we have a guarantee that

$$\dim W_t^\perp = m - \dim W_t \geq m - \{(t-1) + |S_{t-1}|\} \geq |S_{t-1}^c| = n - |S_{t-1}| \geq |S_t \setminus S_{t-1}|.$$

The columns $\{u_i\}_{i=1}^n$ of constructed U satisfies $\langle u_i, y_t - y_0\rangle = 0$ if $i \notin S_t$, for $t = 1, \dots, N$. Therefore,

$$z^{(t)} = U^\intercal(y_t - y_0) \in \mathrm{span}\{e_i\}_{i\in S_t}$$

which leads to $\mathrm{supp}\{z^{(t)}\} \subseteq S_t$. ∎

We now prove the complexity lower bound result for general fixed-point iterations.

Proof of Theorem 4.6. For any deterministic fixed-point iteration $\mathbf{A}$ and initial point $y_0 \in \mathbb{R}^n$, consider a worst-case operator $\mathbb{T}\colon \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ defined in the proof of Theorem 4.1. According to Lemma D.4, there exists an orthogonal $U \in \mathbb{R}^{n\times(N+1)}$ with $n \geq (N+1)+(N-1) = 2N$ such that $z^{(k)} = U^\intercal(y_k - y_0)$ for $k = 0,\dots,N$,

$$\|z^{(k)} - \mathbb{T}z^{(k)}\| \leq \|y_k - \mathbb{T}_U y_k\|, \quad k = 0,\dots,N$$

where the query points $\{y_k\}_{k=0}^N$ are generated from applying $\mathbf{A}$ to $\mathbb{T}_U$ given initial point y_0, and $\{z^{(k)}\}_{k=0}^N$ is a zero-respecting sequence with respect to $\mathbb{T}$. According to Lemma D.2,

$$\|z^{(N)} - \mathbb{T}z^{(N)}\|^2 \geq \left(1+\frac{1}{\gamma}\right)^2\left(\frac{1}{1+\gamma+\cdots+\gamma^N}\right)^2\|z^{(0)} - z_\star\|^2.$$

According to Lemma D.3, $y_\star = y_0 + Uz_\star \in \operatorname{Fix}\mathbb{T}_U$ for $z_\star \in \operatorname{Fix}\mathbb{T}$, so

$$\|y_0 - y_\star\|^2 = \|U(z^{(0)} - z_\star)\|^2 = \|z^{(0)} - z_\star\|^2$$

where the second identity comes from orthogonality of U. We may conclude that

$$\|y_N - \mathbb{T}_U y_N\|^2 \geq \left(1+\frac{1}{\gamma}\right)^2\left(\frac{1}{1+\gamma+\cdots+\gamma^N}\right)^2\|y_0 - y_\star\|^2$$

and that $\mathbb{T}_U : \mathbb{R}^n \to \mathbb{R}^n$ is the desired worst-case $\frac{1}{\gamma}$-contraction with $n \geq 2N$. $\square$

E. Omitted proofs of Section 5

E.1. Convergence rate of proximal point method

Lemma E.1. *Let $\{x_k\}_{k\in\mathbb{N}}$ be the iterates generated by applying PPM $x_{k+1} = \mathbb{J}_{\mathbb{A}}x_k$ starting from $x_0 \in \mathbb{R}^n$, given a uniformly monotone operator $\mathbb{A}$ with parameters $\mu > 0$ and $\alpha > 1$. Now let $A_k := \|x_k - x_\star\|^2$ and $B_k := \|\tilde{\mathbb{A}}x_{k+1}\|^2$. Then for any $k \in \mathbb{N}\cup\{0\}$,*

$$\begin{aligned}
A_k &\geq A_{k+1}\left(1+\mu A_{k+1}^{\frac{\alpha-1}{2}}\right)^2\\
B_k &\geq B_{k+1}.
\end{aligned}$$

Proof. Note that PPM update $x_{k+1} = \mathbb{J}_{\mathbb{A}}x_k$ is equivalent to $x_k = x_{k+1} + \tilde{\mathbb{A}}x_{k+1}$ where $\tilde{\mathbb{A}}x_{k+1} \in \mathbb{A}x_{k+1}$.

$$x_k - x_\star = (x_{k+1} + \tilde{\mathbb{A}}x_{k+1}) - x_\star = (x_{k+1} - x_\star) + \tilde{\mathbb{A}}x_{k+1}.$$

Then

$$\begin{aligned}
A_k &= A_{k+1} + B_k + 2\langle \tilde{\mathbb{A}}x_{k+1}, x_{k+1} - x_\star\rangle\\
&\geq A_{k+1} + B_k + 2\mu\|x_{k+1} - x_\star\|^{\alpha+1}\\
&= A_{k+1} + B_k + 2\mu A_{k+1}^{\frac{\alpha+1}{2}}\\
&\geq A_{k+1} + \mu^2 A_{k+1}^{\alpha} + 2\mu A_{k+1}^{\frac{\alpha+1}{2}}\\
&\geq A_{k+1}\left(1+\mu A_{k+1}^{\frac{\alpha-1}{2}}\right)^2
\end{aligned}$$

where the second inequality follows from

$$\|\tilde{\mathbb{A}}x_{k+1}\|\|x_{k+1} - x_\star\| \geq \langle \tilde{\mathbb{A}}x_{k+1}, x_{k+1} - x_\star\rangle \geq \mu\|x_{k+1} - x_\star\|^{\alpha+1}.$$

Also, from

$$\begin{aligned} B_k - B_{k+1} &= \|\tilde{\mathbf{A}}x_k\|^2 - \|\tilde{\mathbf{A}}x_{k+1}\|^2 \\ &= (\|\tilde{\mathbf{A}}x_k\|^2 + \|\tilde{\mathbf{A}}x_{k+1}\|^2) - 2\|\tilde{\mathbf{A}}x_{k+1}\|^2 \\ &\geq 2\langle \tilde{\mathbf{A}}x_k, \tilde{\mathbf{A}}x_{k+1}\rangle - 2\|\tilde{\mathbf{A}}x_{k+1}\|^2 && \text{(Young's inequality)} \\ &= -2\langle \tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k, \tilde{\mathbf{A}}x_{k+1}\rangle \\ &= 2\langle \tilde{\mathbf{A}}x_{k+1} - \tilde{\mathbf{A}}x_k, x_{k+1} - x_k\rangle \geq 0, && \text{(Monotonicity of } \mathbb{A}\text{)} \end{aligned}$$

we get $B_k \geq B_{k+1}$. □

Theorem E.2. *If $\mathbb{A} \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a uniformly monotone operator with parameters $\mu > 0$ and $\alpha > 1$, there exists $C > 0$ such that the iterates $\{x_k\}_{k\in\mathbb{N}}$ generated by PPM exhibits the rate*

$$\|x_k - x_\star\|^2 \leq \frac{C}{k^{\frac{2}{\alpha-1}}}$$

for any $k \in \mathbb{N}$.

Proof. We use the induction on k to show the convergence rate, and find the necessary conditions for $C > 0$ to satisfy.

In case of $k = 1$, $\|x_1 - x_\star\|^2 \leq C$ must be satisfied. Lemma E.1 implies the monotonicity of A_k, so C with $C \geq \|x_0 - x_\star\|^2$ is a suitable choice.

Now, suppose that $A_k \leq Ck^{-\frac{2}{\alpha-1}}$ and $k \geq 1$. We claim that $A_{k+1} \leq C(k+1)^{-\frac{2}{\alpha-1}}$ for the same $C > 0$. Define $f_\mu^\alpha \colon [0,\infty) \to [0,\infty)$ as

$$f_\mu^\alpha(t) := t\left(1 + \mu t^{\frac{\alpha-1}{2}}\right)^2.$$

Then $f_\mu^\alpha(A_{k+1}) \leq A_k$ from Lemma E.1. If $f_\mu^\alpha(A_{k+1}) \leq f_\mu^\alpha\left(C(k+1)^{-\frac{2}{\alpha-1}}\right)$, since f_μ^α is a monotonically increasing function over $[0,\infty)$, we are done. Define

$$a_n := (n+1)\left\{\left(1 + \frac{1}{n}\right)^{\frac{1}{\alpha-1}} - 1\right\},$$

and function $g \colon (0,\infty) \to \mathbb{R}$ as

$$g(x) = \left(1 + \frac{1}{x}\right)\left\{(1+x)^{\frac{1}{\alpha-1}} - 1\right\}$$

so that $a_n = g\left(\frac{1}{n}\right)$ for $n \in \mathbb{N}$. Then

$$\begin{aligned} g'(x) &= -\frac{1}{x^2}\left\{(1+x)^{\frac{1}{\alpha-1}} - 1\right\} + \frac{1}{\alpha-1}\left(1 + \frac{1}{x}\right)(1+x)^{\frac{1}{\alpha-1}-1} \\ &= -\frac{(1+x)^{\frac{1}{\alpha-1}}}{x^2} + \frac{1}{x^2} + \frac{1}{\alpha-1}\frac{(1+x)(1+x)^{\frac{1}{\alpha-1}-1}}{x} \\ &= \frac{-(1+x)^{\frac{1}{\alpha-1}} + 1 + \frac{x}{\alpha-1}(1+x)^{\frac{1}{\alpha-1}}}{x^2} \\ &= \frac{(1+x)^{\frac{1}{\alpha-1}}}{x^2}\left\{(1+x)^{-\frac{1}{\alpha-1}} - \left(1 - \tfrac{1}{\alpha-1}x\right)\right\}. \end{aligned}$$

As $x \mapsto (1+x)^{-\frac{1}{\alpha-1}}$ is a convex function on $[0,\infty)$ and $x \mapsto 1 - \frac{1}{\alpha-1}x$ is a first-order approximation at 0 of it, $g'(x) \geq 0$ for $x > 0$. g is a monotonically increasing function, so g obtains its maximum in $(0,1]$ at $x = 1$, and we have

$$\sup_{n\in\mathbb{N}} a_n = \sup_{n\in\mathbb{N}} g\left(\frac{1}{n}\right) = g(1) = 2(2^{\frac{1}{\alpha-1}} - 1) = 2^{\frac{\alpha}{\alpha-1}} - 2.$$

The boundedness of a_n leads to the equivalency as

$$
\begin{aligned}
a_k \leq 2^{\frac{\alpha}{\alpha-1}} - 2 &\iff \left(1+\frac{1}{k}\right)^{\frac{2}{\alpha-1}} \leq \left(1+\frac{2^{\frac{\alpha}{\alpha-1}}-2}{k+1}\right)^2 \\
&\iff \frac{C}{k^{\frac{2}{\alpha-1}}} \leq \frac{C}{(k+1)^{\frac{2}{\alpha-1}}} \left\{1+\frac{2^{\frac{\alpha}{\alpha-1}}-2}{C^{\frac{\alpha-1}{2}}}\left(\frac{C}{(k+1)^{\frac{2}{\alpha-1}}}\right)^{\frac{\alpha-1}{2}}\right\}^2,
\end{aligned}
$$

for any choice of $C > 0$. Choosing $C \geq \mu^{-\frac{2}{\alpha-1}}(2^{\frac{\alpha}{\alpha-1}}-2)^{\frac{2}{\alpha-1}}$ which is equivalent to

$$
\frac{2^{\frac{\alpha}{\alpha-1}}-2}{C^{\frac{\alpha-1}{2}}} \leq \mu,
$$

we get

$$
\begin{aligned}
\frac{C}{(k+1)^{\frac{2}{\alpha-1}}} \left\{1+\frac{2^{\frac{\alpha}{\alpha-1}}-2}{C^{\frac{\alpha-1}{2}}}\left(\frac{C}{(k+1)^{\frac{2}{\alpha-1}}}\right)^{\frac{\alpha-1}{2}}\right\}^2 &\leq \frac{C}{(k+1)^{\frac{2}{\alpha-1}}} \left\{1+\mu\left(\frac{C}{(k+1)^{\frac{2}{\alpha-1}}}\right)^{\frac{\alpha-1}{2}}\right\}^2 \\
&= f_\mu^\alpha\left(\frac{C}{(k+1)^{\frac{2}{\alpha-1}}}\right).
\end{aligned}
$$

Gathering all the inequalities above, if $C \geq \mu^{-\frac{2}{\alpha-1}}(2^{\frac{\alpha}{\alpha-1}}-2)^{\frac{2}{\alpha-1}}$, then

$$
f_\mu^\alpha(A_{k+1}) \leq A_k \leq \frac{C}{k^{\frac{2}{\alpha-1}}} \leq f_\mu^\alpha\left(\frac{C}{(k+1)^{\frac{2}{\alpha-1}}}\right)
$$

so we get

$$
A_k \leq \frac{C}{k^{\frac{2}{\alpha-1}}} \implies A_{k+1} \leq \frac{C}{(k+1)^{\frac{2}{\alpha-1}}}
$$

for $k = 1, 2, \ldots$.

Therefore,

$$
\|x_k - x_\star\|^2 \leq \frac{C}{k^{\frac{2}{\alpha-1}}} = \frac{\max\left\{\left(\frac{2^{\frac{\alpha}{\alpha-1}}-2}{\mu}\right)^{\frac{2}{\alpha-1}}, \|x_0 - x_\star\|^2\right\}}{k^{\frac{2}{\alpha-1}}}.
$$

□

We now prove the convergence rate of PPM in terms of $B_k = \|\tilde{\mathbb{A}} x_{k+1}\|^2$.

Proof of Theorem 5.1. We claim the convergence rate of $B_{k-1} = \|\tilde{\mathbb{A}} x_k\|^2$ to be as above. From the proof of Lemma E.1, we have

$$
B_k \leq A_k - A_{k+1} - 2\mu A_{k+1}^{\frac{\alpha+1}{2}} \leq A_k - A_{k+1}.
$$

If $N = 1$, then

$$
B_0 \leq A_0 - A_1 \leq A_0 = \|x_0 - x_\star\|^2.
$$

Suppose $N \geq 2$. Let $n := \lfloor \frac{N}{2} \rfloor$ where $\lfloor x \rfloor$ is the largest integer not exceeding x. Summing up the above inequality from $k = n$ to $k = N-1$ and using the monotonicity of B_k, we have

$$
\frac{N}{2} B_{N-1} \leq \sum_{k=n}^{N-1} B_{N-1} \leq \sum_{k=n}^{N-1} B_k \leq \sum_{k=n}^{N-1} (A_k - A_{k+1}) = A_n - A_N \leq A_n.
$$

Note that from the convergence analysis of A_k, or Theorem E.2, we have

$$A_n \leq \frac{C}{n^{\frac{2}{\alpha-1}}}$$

where $C = \max\left\{ \left(\frac{2^{\frac{\alpha}{\alpha-1}}-2}{\mu}\right)^{\frac{2}{\alpha-1}}, \|x_0 - x_\star\|^2 \right\}$. Therefore,

$$\frac{N}{2} B_{N-1} \leq \frac{C}{n^{\frac{2}{\alpha-1}}} \leq \frac{C}{\left(\frac{N-1}{2}\right)^{\frac{2}{\alpha-1}}},$$

so we may conclude that, for any $N \geq 2$,

$$\begin{aligned}
B_{N-1} &\leq \frac{2^{\frac{\alpha+1}{\alpha-1}} C}{(N-1)^{\frac{2}{\alpha-1}} N} \\
&= \frac{2^{\frac{\alpha+1}{\alpha-1}} \max\left\{ \left(\frac{2^{\frac{\alpha}{\alpha-1}}-2}{\mu}\right)^{\frac{2}{\alpha-1}}, \|x_0 - x_\star\|^2 \right\}}{(N-1)^{\frac{2}{\alpha-1}} N} \\
&\leq \frac{2^{\frac{\alpha+3}{\alpha-1}} \max\left\{ \left(\frac{2^{\frac{\alpha}{\alpha-1}}-2}{\mu}\right)^{\frac{2}{\alpha-1}}, \|x_0 - x_\star\|^2 \right\}}{N^{\frac{\alpha+1}{\alpha-1}}} \\
&= \mathcal{O}\left(N^{-\frac{\alpha+1}{\alpha-1}}\right)
\end{aligned}$$

where the second inequality follows from $2(N-1) \geq N$. Since this bound also holds for the case of $N = 1$ from $B_0 \leq \|x_0 - x_\star\|^2$, we are done. $\square$

E.2. Convergence rate of restarted OS-PPM (OS-PPM$_0^{\text{res}}$)

Roulet & d'Aspremont (2020) showed that if the objective function f of a smooth convex minimization problem satisfies a *Hölderian error bound condition*

$$\frac{\mu}{r}\|x - x_\star\|^r \leq f(x) - f^\star, \qquad \forall\, x \in K \subset \mathbb{R}^n$$

where $x_\star \in K$ is a minimizer of f and K is a given set, then the unaccelerated base algorithm can be accelerated with a restarting scheme. The restarting schedule uses t_k iterations for each k-th outer loop recursively satisfying

$$f(x_k) - f^\star \leq e^{-\eta k}(f(x_0) - f^\star), \qquad k = 1, 2, \ldots$$

for some $\eta > 0$, where $x_k = \mathcal{A}(x_{k-1}, t_k)$ is the output of k-th outer loop, which applies t_k iterations of the base algorithm $\mathcal{A}$ starting from x_{k-1}. If an objective function is strongly convex near the solution ($r = 2$), a constant restarting schedule $t_k = \lambda$ provides a faster rate compared to an unaccelerated base algorithm (Nemirovski & Nesterov, 1985). If an objective function satisfies a Hölderian error bound condition but it is not strongly convex ($r > 2$), then an exponentially-growing schedule $t_k = \lambda e^{\beta k}$ for some $\lambda > 0$ and $\beta > 0$ results in a faster sublinear convergence rate.

As notable prior work, Kim (2021) studied APPM with a constant restarting schedule in the strongly monotone setup but was not able to obtain a rate faster than plain PPM. We show that restarting with an exponentially increasing schedule accelerates (OS-PPM) under uniform monotonicity, as for the case of $r > 2$ in Roulet & d'Aspremont (2020).

Proof of Theorem 5.2. Suppose that given an initial point $x_0 \in \mathbb{R}^n$, let $\tilde{x}_0$ be an iterate generated by applying APPM on x_0 only once. Then

$$\tilde{x}_0 = \frac{1}{2}(2\mathbb{J}_{\mathbb{A}} x_0 - x_0) + \frac{1}{2}x_0 = \mathbb{J}_{\mathbb{A}} x_0,$$

so we get

$$\begin{aligned}
\|x_0 - x_\star\|^2 &= \|\tilde{\mathbb{A}}\tilde{x}_0 + (\tilde{x}_0 - x_\star)\|^2 \\
&= \|\tilde{\mathbb{A}}\tilde{x}_0\|^2 + 2\langle \tilde{\mathbb{A}}\tilde{x}_0, \tilde{x}_0 - x_\star\rangle + \|\tilde{x}_0 - x_\star\|^2.
\end{aligned}$$

From the monotonicity of $\mathbb{A}$, $\langle \tilde{\mathbb{A}}\tilde{x}_0, \tilde{x}_0 - x_\star \rangle \ge 0$, so we may conclude that

$$\|\tilde{\mathbb{A}}\tilde{x}_0\|^2 \le \|x_0 - x_\star\|^2.$$

Now we describe the restarting scheme of APPM. Let t_k be the number of inner iterations applying APPM for the kth outer iteration. This iteration starts from $\tilde{x}_{k-1}$ and outputs $\tilde{x}_k$ after applying t_k iterations of APPM. Then the kth outer iteration results in

$$\|\tilde{\mathbb{A}}\tilde{x}_k\|^2 \le \frac{1}{(t_k+1)^2}\|\tilde{x}_{k-1} - x_\star\|^2 \le \frac{1}{t_k^2}\|\tilde{x}_{k-1} - x_\star\|^2 \le \frac{1}{\mu^{2/\alpha}t_k^2}\|\tilde{\mathbb{A}}\tilde{x}_{k-1}\|^{2/\alpha},$$

where the last inequality follows from

$$\|\tilde{\mathbb{A}}\tilde{x}_{k-1}\|\|\tilde{x}_{k-1} - x_\star\| \ge \langle \tilde{\mathbb{A}}\tilde{x}_{k-1}, \tilde{x}_{k-1} - x_\star \rangle \ge \mu\|\tilde{x}_{k-1} - x_\star\|^{\alpha+1}.$$

In order to find a possible choice of restart schedule, we will iteratively find the number t_k of inner iterations for kth outer iteration which satisfies

$$\|\tilde{\mathbb{A}}\tilde{x}_k\|^2 \le e^{-\eta k}\|x_0 - x_\star\|^2$$

for some $\eta > 0$. The case of $k = 0$ holds automatically. Suppose $k \ge 1$, and $t_1, \dots, t_{k-1}$ are already chosen to satisfy

$$\|\tilde{\mathbb{A}}\tilde{x}_{k-1}\|^2 \le e^{-\eta(k-1)}\|x_0 - x_\star\|^2$$

for $k \ge 1$. Then

$$\|\tilde{\mathbb{A}}\tilde{x}_k\|^2 \le \frac{1}{\mu^{2/\alpha}t_k^2}\|\tilde{\mathbb{A}}\tilde{x}_{k-1}\|^{2/\alpha} \le \frac{1}{\mu^{2/\alpha}t_k^2}e^{-\frac{\eta(k-1)}{\alpha}}\|x_0 - x_\star\|^{\frac{2}{\alpha}},$$

so that the claimed convergence rate is guaranteed if

$$\frac{1}{\mu^{2/\alpha}t_k^2}e^{-\frac{\eta(k-1)}{\alpha}}\|x_0 - x_\star\|^{\frac{2}{\alpha}} \le e^{-\eta k}\|x_0 - x_\star\|^2.$$

This is equivalent to

$$t_k \ge \underbrace{\mu^{-\frac{1}{\alpha}}e^{\frac{\eta}{2\alpha}}\|x_0 - x_\star\|^{\frac{1}{\alpha}-1}}_{:=\lambda}\exp\left\{\underbrace{\frac{\eta}{2}\left(1 - \frac{1}{\alpha}\right)}_{:=\beta}k\right\},$$

so if $t_k \ge \lambda e^{\beta k}$ for $k = 1, \dots, R$, then $\|\tilde{\mathbb{A}}\tilde{x}_k\|^2 \le e^{-\eta k}\|x_0 - x_\star\|^2$ for $k = 1, \dots, R$.

Now we prove that the choice of

$$t_k = \begin{cases} \lceil \lambda e^{\beta k} \rceil & (k = 1, \dots, R-1) \\ N - 1 - \sum_{k=1}^{R-1} t_k & (k = R) \end{cases}$$

for integer R satisfying

$$\sum_{k=1}^{R}\lceil \lambda e^{\beta k} \rceil \le N - 1 < \sum_{k=1}^{R+1}\lceil \lambda e^{\beta k} \rceil$$

results in $\mathcal{O}\left(N^{-\frac{2\alpha}{\alpha-1}}\right)$-rate of $\|\tilde{\mathbb{A}}\hat{x}\|^2$ for restarted OS-PPM (OS-PPM$_0^{\mathrm{res}}$).

For $k = 1, \dots, R-1$, $t_k \ge \lambda e^{\beta k}$ by definition of t_k. If $k = R$, from

$$N - 1 = \sum_{k=1}^{R-1} t_k + t_R = \sum_{k=1}^{R-1}\lceil \lambda e^{\beta k} \rceil + t_R,$$

we have

$$t_R = N - 1 - \sum_{k=1}^{R-1}\lceil \lambda e^{\beta k} \rceil \ge \lceil \lambda e^{\beta R} \rceil \ge \lambda e^{\beta R}.$$

Therefore, $t_k \geq \lambda e^{\beta k}$ for $k = 1, 2, \ldots, R$, and we get

$$\|\tilde{\mathbb{A}}\tilde{x}_R\|^2 \leq e^{-\eta R}\|x_0 - x_\star\|^2.$$

To find the upper bound to $\|\tilde{\mathbb{A}}\tilde{x}_R\|^2$ using the inequality above, we obtain a lower bound to R. From $\lambda e^{\beta k} \leq \lceil \lambda e^{\beta k} \rceil \leq \lambda e^{\beta k} + 1$ and $\lceil \lambda e^{\beta R} \rceil \leq t_R < \lceil \lambda e^{\beta R} \rceil + \lceil \lambda e^{\beta(R+1)} \rceil$, we have

$$\sum_{k=1}^{R} \lambda e^{\beta k} \leq N - 1 = \sum_{k=1}^{R} t_k = \sum_{k=1}^{R-1} \lceil \lambda e^{\beta k} \rceil + t_R \leq \sum_{k=1}^{R+1} \lambda e^{\beta k} + R + 1. \tag{1}$$

Using the first inequality in (1), we have

$$\lambda e^{\beta} \frac{e^{\beta R} - 1}{e^{\beta} - 1} \leq N - 1$$

or equivalently,

$$R \leq \frac{1}{\beta} \log\left(\frac{N-1}{\lambda}\frac{e^{\beta}-1}{e^{\beta}} + 1\right).$$

Plugging this upper bound of R to the second inequality of (1), we get

$$N - 1 \leq \lambda e^{\beta} \frac{e^{\beta(R+1)} - 1}{e^{\beta} - 1} + \frac{1}{\beta} \log\left(\frac{N-1}{\lambda}\frac{e^{\beta}-1}{e^{\beta}} + 1\right) + 1.$$

Simplifying this to obtain a lower bound on R, we get

$$e^{-\beta}\left\{\frac{e^{\beta}-1}{\lambda e^{\beta}}\left(N - 2 - \frac{1}{\beta}\log\left(\frac{N-1}{\lambda}\frac{e^{\beta}-1}{e^{\beta}} + 1\right)\right) + 1\right\} \leq e^{\beta R}.$$

Therefore,

$$\begin{aligned}
\|\tilde{\mathbb{A}}\tilde{x}_R\|^2 &\leq e^{-\eta R}\|x_0 - x_\star\|^2 \\
&\leq e^{\eta}\left\{\frac{e^{\beta}-1}{\lambda e^{\beta}}\left(N - 2 - \frac{1}{\beta}\log\left(\frac{N-1}{\lambda}\frac{e^{\beta}-1}{e^{\beta}} + 1\right)\right) + 1\right\}^{-\frac{\eta}{\beta}} \|x_0 - x_\star\|^2 \\
&= \left\{\frac{e^{\beta}-1}{\lambda e^{2\beta}}\left(N - 2 - \frac{1}{\beta}\log\left(\frac{e^{\beta}-1}{\lambda e^{\beta}}(N-1) + 1\right)\right) + \frac{1}{e^{\beta}}\right\}^{-\frac{2\alpha}{\alpha-1}} \|x_0 - x_\star\|^2 \qquad \text{(Choose } \eta = 2\text{)} \\
&= \mathcal{O}\left(N^{-\frac{2\alpha}{\alpha-1}}\right)
\end{aligned}$$

where $\lambda = \left(\frac{e}{\mu}\right)^{\frac{1}{\alpha}} \|x_0 - x_\star\|^{-\left(1-\frac{1}{\alpha}\right)}$. $\square$

This is a rate faster than $\mathcal{O}(N^{-\frac{\alpha+1}{\alpha-1}})$-rate of PPM. Although the monotonicity parameter $\mu > 0$ and $\alpha > 1$ are unknown, one can obtain a suboptimal restart schedule with additional cost for the grid search as in Roulet & d'Aspremont (2020), where the total cost for the algorithm is of $\mathcal{O}\left(N^{-\frac{2\alpha}{\alpha-1}}(\log N)^2\right)$.

F. Experiment details

We now describe the experiments of Section 6 in further detail.

F.1. Experiment details of Section 6.1

In the first example, $\mathbb{T}_\theta$ is constructed with $\theta = 15°$ and $\gamma = \frac{1}{0.95}$, and (OC-Halpern) is applied on $\mathbb{T}_\theta$ with the same $\gamma = \frac{1}{0.95}$. In the second example, $\mathbb{M}$ is constructed with $\mu = 0.035$, and (OS-PPM) is applied on $\mathbb{M}$ with the same $\mu = 0.035$. For both experiments, we use $N = 101$ total iterations. The plots of both experiments display the position of every iterate with markers, when methods started from initial point $y_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\intercal \in \mathbb{R}^2$.

F.2. Experiment details of Section 6.2

X-ray CT reconstructs the image from the received from a number of detectors. Reconstruction of the original image is often formulated as a least-squares problem with total variation regularization

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \tfrac{1}{2}\|Ex-b\|^2+\lambda\|Dx\|_1, \tag{2}$$

where $x\in\mathbb{R}^n$ is a vectorized image, $E\in\mathbb{R}^{m\times n}$ is the discrete Radon transform, $b=Ex$ is the measurement, and D is the finite difference operator. This regularized least-squares problem can be solved using PDHG, also known as the Chambolle–Pock method (Chambolle & Pock, 2011). PDHG can be interpreted as an instance of variable metric PPM (He & Yuan, 2012); it is a nonexpansive fixed-point iteration $(x^{k+1},u^{k+1},v^{k+1})=\mathbb{T}(x^k,u^k,v^k)$ defined as

$$\begin{aligned}
x^{k+1} &= x^k-\alpha E^\intercal u^k-\beta D^\intercal v^k\\
u^{k+1} &= \frac{1}{1+\alpha}\left(u^k+\alpha E(2x^{k+1}-x^k)-\alpha b\right)\\
v^{k+1} &= \Pi_{[-\lambda\alpha/\beta,\lambda\alpha/\beta]}\left(v^k+\beta D(2x^{k+1}-x^k)\right)
\end{aligned}$$

with respect to the metric matrix

$$M=\begin{bmatrix}(1/\alpha)I & -E^\intercal & -(\beta/\alpha)D^\intercal\\ -E & (1/\beta)I & 0\\ -(\beta/\alpha)D & 0 & (1/\beta)I\end{bmatrix}.$$

Therefore, we apply OHM on $\mathbb{T}$ as

$$(x^{k+1},u^{k+1},v^{k+1})=\left(1-\frac{1}{k+2}\right)\mathbb{T}(x^k,u^k,v^k)+\frac{1}{k+2}(x^0,u^0,v^0) \tag{PDHG with OHM}$$

and use additional restarting strategy to yield a faster convergence.

In our experiment, we use the a Modified Shepp-Logan phantom image. We applied PDHG, PDHG combined with OHM, and PDHG combined with restarted OC-Halpern (OS-PPM$_0^{\text{res}}$), where the parameters are given as $\alpha=0.01$, $\beta=0.03$ and $\lambda=1.0$. We applied restarting with the schedule illustrated in Theorem 5.2, with properly chosen $\lambda>0$ and $\beta>0$.

Figure 6. Images reconstructed by applying PDHG, PDHG with OHM, and PDHG with restarted OC-Halpern for 1000 iterations.

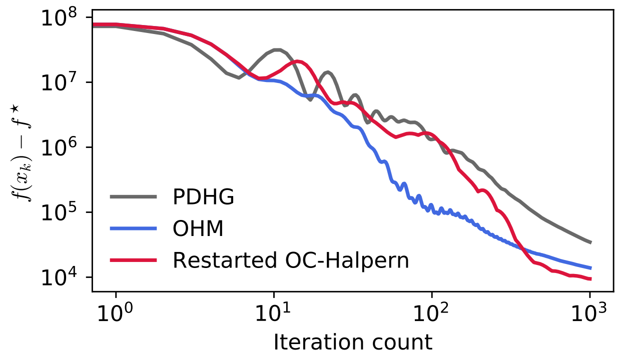


Figure 7. Function value suboptimality $f(x_k) - f^\star$ plot of PDHG, PDHG with OHM, and PDHG with restarted OC-Halpern (OS-PPM$_0^{\rm res}$) in CT image reconstruction.

Figure 6 shows the reconstructed images after 1000 iterations. Restarted OC-Halpern (OS-PPM$_0^{\rm res}$) can effectively recover the original image, in a faster rate. Figure 7 shows that even without theoretical guarantee, the function value suboptimality decreases in a faster rate for OHM and restarted OC-Halpern.

F.3. Experiment details of Section 6.3

In this section, we approximated the Wasserstein distance (or Earth mover's distance) of two different probability distributions by solving the following discretized problem

$$\begin{array}{ll} \underset{m_x, m_y}{\text{minimize}} & \|\mathbf{m}\|_{1,1} = \sum_{i=1}^n \sum_{j=1}^n |m_{x,ij}| + |m_{y,ij}| \\ \text{subject to} & \mathrm{div}(m) + \rho_1 - \rho_0 = 0. \end{array}$$

To solve this problem Li et al. (2018) used PDHG (Chambolle & Pock, 2011)

$$\begin{aligned} \tilde{m}_{x,ij}^{k+1} &= \frac{1}{1+\varepsilon\mu}\mathrm{shrink}_1\left(\tilde{m}_{x,ij}^k + \mu(\nabla\Phi^k)_{x,ij}, \mu\right) \\ \tilde{m}_{y,ij}^{k+1} &= \frac{1}{1+\varepsilon\mu}\mathrm{shrink}_1\left(\tilde{m}_{y,ij}^k + \mu(\nabla\Phi^k)_{y,ij}, \mu\right) \\ \Phi_{ij}^{k+1} &= \Phi_{ij}^k + \tau\Big((\mathrm{div}(2\mathbf{m}^{k+1} - \mathbf{m}^k))_{ij} + \rho_{ij}^1 - \rho_{ij}^0\Big) \end{aligned} \tag{Primal-dual method for EMD-L_1}$$

for $k = 1, 2, \ldots$ where $\tilde{\mathbf{m}} = (\tilde{m}_x, \tilde{m}_y)$ is $\mathbf{m} = (m_x, m_y)$ with zero padding on their last row and last column, respectively, hence making $\tilde{m}_x, \tilde{m}_y \in \mathbb{R}^{n\times n}$. We denote this fixed-point iteration by $\mathbb{T}$, so that $(\tilde{m}_x^{k+1}, \tilde{m}_y^{k+1}, \Phi^{k+1}) = \mathbb{T}(\tilde{m}_x^k, \tilde{m}_y^k, \Phi^k)$. Combining OHM on this fixed-point iteration yields the iteration

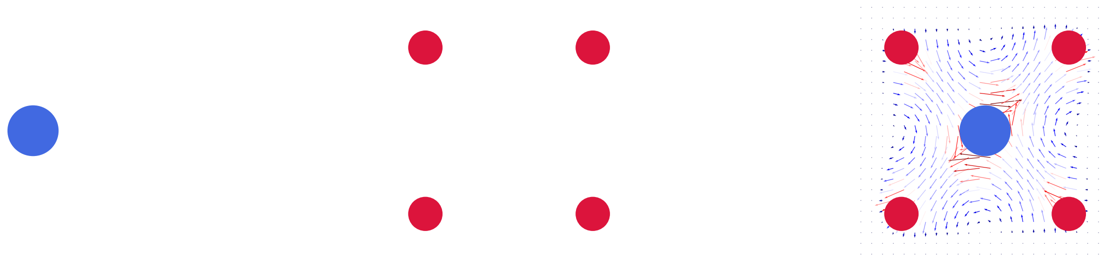
$$(\tilde{m}_x^{k+1}, \tilde{m}_y^{k+1}, \Phi^{k+1}) = \left(1 - \frac{1}{k+2}\right)\mathbb{T}(\tilde{m}_x^k, \tilde{m}_y^k, \Phi^k) + \frac{1}{k+2}(\tilde{m}_x^0, \tilde{m}_y^0, \Phi^0)$$

for $k = 1, 2, \ldots$, and we also combine restarting technique with exponential schedule to hope for further acceleration.

This experiment calculated an approximation of Wasserstein distance between the two probability distributions as in Figure 8. We applied 3 different algorithms for $N = 100,000$ iterations with algorithm parameters $\mu = 1.0 \times 10^{-6}$ and $\varepsilon = 1.0$. Restarting the algorithm with Halpern scheme every $10,000$ iterations provided the accelerated rate, but we chose better exponential schedule for the plot in Figure 9.

F.4. Experiment details of Section 6.4

We follow the settings of decentralized compressed sensing experiment in section IV of (Shi et al., 2015). The underlying network has 10 nodes and 18 edges, and these edges connect the nodes as in Figure 11.



(a) Probabilistic distribution ρ_0

(b) Probabilistic distribution of ρ_1

(c) Solution of discretized problem in Section 6.3

Figure 8. Probabilistic distribution of ρ_0 and ρ_1. This distribution is expressed as the colored parts in 256×256 grid, there ρ_0 contains the part of $x^2 + y^2 \le (0.3)^2$ and ρ_1 contains 4 identical circles with radius 0.2, centered at $(\pm 1, \pm 1)$.

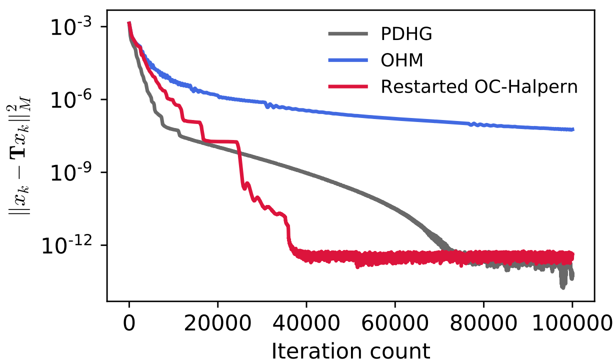


Figure 9. Fixed-point residual of $\mathbb{T}$ versus iteration count plot for approximating Wasserstein distance.

Experiment considered the regularized least-squares problem on $\mathbb{R}^{50}$, where the sparse signal $x_\star$ has 10 nonzero entries.

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \|A_{(i)}x - b_{(i)}\|^2 + \lambda \|x\|_1.$$

Each node i maintains its local estimate x_i of $x \in \mathbb{R}^n$, and have access to sensing matrix $A_{(i)} \in \mathbb{R}^{m_i \times n}$, where m_i is the number of accessible sensors. Here, we assume to have $m_i = 3$ many sensors for each node and has total $m = 30$ sensors.

We applied PG-EXTRA, PG-EXTRA combined with OHM, PG-EXTRA with (OC-Halpern), and PG-EXTRA with Restarted OC-Halpern ($\text{OS-PPM}_0^{\text{res}}$), since PG-EXTRA can be understood as a fixed-point iteration (Wu et al., 2018). Let $\mathbf{x}^k \in \mathbb{R}^{n \times 10}$ be a vertical stack of $\mathbb{R}^n$ vectors, where each i-th row vector $\mathbf{x}_i^k$ is a local copy of x stored in node i. The vectors in node i only interact with other vectors in close neighborhood of node i. The fixed-point iteration $(\mathbf{x}^{k+1}, \mathbf{w}^{k+1}) = \mathbb{T}(\mathbf{x}^k, \mathbf{w}^k)$ is

$$\mathbf{x}_i^{k+1} = \text{Prox}_{\alpha\lambda\|\cdot\|_1}\left(\sum_j W_{i,j}\mathbf{x}_j^k - \alpha A_{(i)}^\intercal (A_{(i)}\mathbf{x}_i^k - b_{(i)}) - \mathbf{w}_i^k \right)$$
$$\mathbf{w}^{k+1} = \mathbf{w}^k + \frac{1}{2}(I - W)\mathbf{x}^k$$

and PG-EXTRA combined with OHM is

$$(\mathbf{x}^{k+1}, \mathbf{w}^{k+1}) = \left(1 - \frac{1}{k+2}\right) \mathbb{T}(\mathbf{x}^k, \mathbf{w}^k) + \frac{1}{k+2}(\mathbf{x}^0, \mathbf{w}^0) \qquad \text{(PG-EXTRA with OHM)}$$

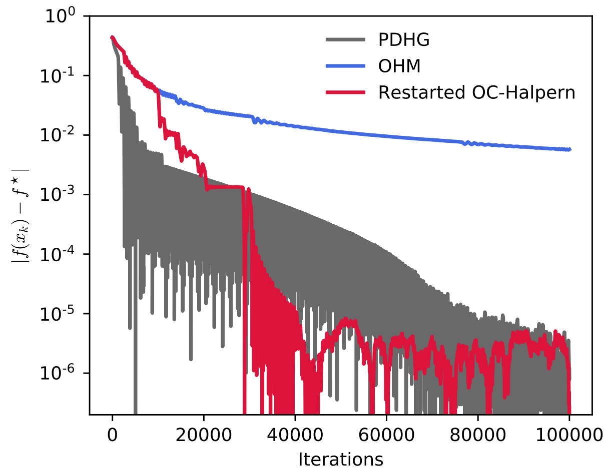


Figure 10. Absolute function-value suboptimality $|f(x_k) - f^\star|$ versus iteration count plot for approximating Wasserstein distance.

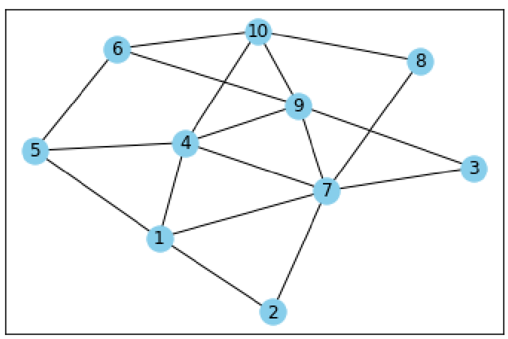


Figure 11. The network underlying the setting of Section 6.4.

for $k = 0, 1, \ldots$. For all these methods, we chose the mixing matrix $W \in \mathbb{R}^{10\times 10}$ to be Metropolis-Hastings weight with each (i, j)-entry $W_{i,j}$ being

$$W_{i,j} = \begin{cases} \frac{1}{\max\{\deg(i),\deg(j)\}} & (i \neq j) \\ 1 - \sum_{j\neq i} W_{i,j} & (i = j) \end{cases}$$

where $\deg(i)$ is the number of edges connected to node i. We applied each methods (PG-EXTRA, PG-EXTRA with OHM, PG-EXTRA with OC-Halpern, and PG-EXTRA with restarted OC-Halpern (OS-PPM$_0^{\mathrm{res}}$)) with stepsize $\alpha = 0.005$ and regularization parameter $\lambda = 0.002$ for 100 iterations.

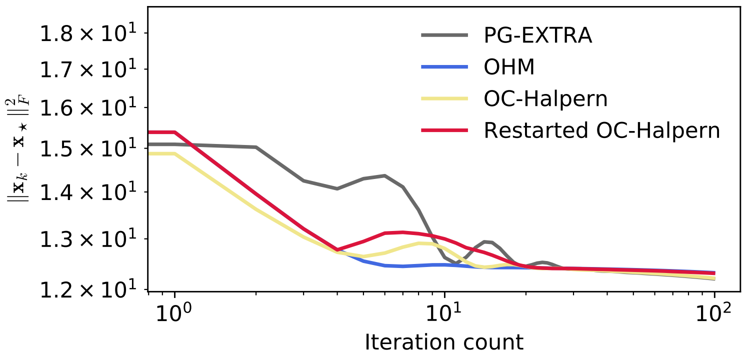


Figure 12. Distance to solution $\|\mathbf{x}_k - \mathbf{x}_\star\|_F^2$ versus iteration count plot for PG-EXTRA, PG-EXTRA with OHM, PG-EXTRA with (OC-Halpern), and PG-EXTRA with Restarted OC-Halpern ($\text{OS-PPM}_0^{\text{res}}$).