Noname manuscript No.
(will be inserted by the editor)

Convergence rate analysis of several splitting schemes

Damek Davis · Wotao Yin

Received: date / Accepted: date

Abstract Splitting schemes are a class of powerful algorithms that solve complicated monotone inclusions and convex optimization problems that are built from many simpler pieces. They give rise to algorithms in which the simple pieces of the decomposition are processed individually. This leads to easily implementable and highly parallelizable algorithms, which often obtain nearly state-of-the-art performance.

In the first part of this paper, we analyze the convergence rates of several general splitting algorithms and provide examples to prove the tightness of our results. The most general rates are proved for the *fixed-point residual* (FPR) of the Krasnosel'skiĭ-Mann (KM) iteration of nonexpansive operators, where we improve the known big-O rate to little-o. We show the tightness of this result and improve it in several special cases. In the second part of this paper, we use the convergence rates derived for the KM iteration to analyze the *objective error* convergence rates for the Douglas-Rachford (DRS), Peaceman-Rachford (PRS), and ADMM splitting algorithms under general convexity assumptions. We show, by way of example, that the rates obtained for these algorithms are tight in all cases and obtain the surprising statement: The DRS algorithm is nearly as fast as the proximal point algorithm (PPA) in the ergodic sense and nearly as slow as the subgradient method in the nonergodic sense. Finally, we provide several applications of our result to feasibility problems, model fitting, and distributed optimization. Our analysis is self-contained, and most results are deduced from a basic lemma that derives convergence rates for summable sequences, a simple diagram that decomposes each relaxed PRS iteration, and fundamental inequalities that relate the FPR to objective error.

Keywords Krasnosel'skiĭ-Mann algorithm · Douglas-Rachford Splitting · Peaceman-Rachford Splitting · Alternating Direction Method of Multipliers · nonexpansive operator · averaged operator · fixed-point algorithm · little-o convergence

Mathematics Subject Classification (2000) 47H05 · 65K05 · 65K15 · 90C25

D. Davis · W. Yin
Department of Mathematics, University of California, Los Angeles
Los Angeles, CA 90025, USA
E-mail: damek / wotaoyin@ucla.edu

1 Introduction

Operator-splitting and alternating-direction methods have a long history, and they have been, and still are, some of the most useful methods in scientific computing. These algorithms solve problems composed of several competing structures, such as finding a point in the intersection of two sets, minimizing the sum of two functions, and, more generally, finding a zero of the sum of two monotone operators. They give rise to algorithms that are simple to implement and converge quickly in practice. Since the 1950s, operator-splitting methods have been applied to solving partial differential equations (PDEs) and feasibility problems. Recently, certain operator-splitting methods such as ADMM (for alternating direction methods of multipliers) [23,24] and Split Bregman [25] have found new applications in (PDE and non-PDE related) image processing, statistical and machine learning, compressive sensing, matrix completion, finance, and control. They have also been extended to handle distributed and decentralized optimization (see [8,36,38]).

In convex optimization, operator-splitting methods split constraint sets and objective functions into subproblems that are easier to solve than the original problem. Throughout this paper, we will consider two prototype optimization problems: We analyze the unconstrained problem

$$\underset{x\in\mathcal{H}}{\text{minimize}} \;\; f(x)+g(x) \tag{1}$$

where $\mathcal{H}$ is a Hilbert space. In addition, we analyze the linearly constrained variant

$$\begin{aligned} &\underset{x\in\mathcal{H}_1,\; y\in\mathcal{H}_2}{\text{minimize}} \;\; f(x)+g(y) \\ &\text{subject to} \;\; Ax+By=b \end{aligned} \tag{2}$$

where $\mathcal{H}_1, \mathcal{H}_2$, and $\mathcal{G}$ are Hilbert spaces, the vector b is an element of $\mathcal{G}$, and $A:\mathcal{H}_1\to\mathcal{G}$ and $B:\mathcal{H}_2\to\mathcal{G}$ are bounded linear operators. Our working assumption throughout the paper is that the subproblems involving f and g separately are much simpler to solve than the joint minimization problem.

Problem (1) is often used to model tasks in signal recovery that enforce prior knowledge of the form of the solution, such as sparsity, low rank, and smoothness [16]. The knowledge-enforcing function, also known as the regularizer, often has properties that make it difficult to jointly optimize with the remaining parts of the problem. Therefore, operator-splitting methods become the natural choice.

Problem (2) is often used to model tasks in machine learning, image processing, and distributed optimization. For special choices of A and B, operator-splitting schemes naturally give rise to algorithms with parallel or distributed implementations [6,8]. Because of the flexibility of operator-splitting algorithms, they have become a standard tool that addresses the emerging need for computational approaches to analyze a massive amount of data in a fast, parallel, distributed, or even real-time manner.

1.1 Goals, challenges, and approaches

This work seeks to improve the theoretical understanding of the most well-known operator-splitting algorithms including Peaceman-Rachford splitting (PRS), Douglas-Rachford splitting (DRS), the alternating direction method of multipliers (ADMM), as well as their relaxed versions. (The proximal point algorithm (PPA) and forward-backward splitting (FBS) are covered in some limited aspects too.) When applied to convex optimization problems, they are known to converge under rather general conditions. However, their convergence rates are largely unknown with only a few exceptions [4,38]. Among the few rates known in the

literature, the majority are given in terms of somewhat awkward quantities, such as the *fixed-point residual* (FPR) (the squared distance between two consecutive iterates) and the violation to a Lagrangian-type optimality condition (variational inequalities/duality gaps) [12,7,19,32,26,17,30]. Moreover, unlike the well-developed complexity estimates for (sub)gradient methods [33], there are no known *lower complexity* results for most *strictly primal* or *strictly dual* splitting algorithms. (As an exception, lower complexity results and optimal algorithms are known for the *primal-dual* case [13].) This paper attempts to close this gap. The convergence rates in terms of FPR and objective errors are derived for operator-splitting algorithms applied to Problem (1). Convergence rates for constraint violations, the primal objective error, and the dual objective error are derived for ADMM, which applies to Problem (2). Some of the derived rates are convenient to use, for example, to determine how many iterations are needed to reach a certain accuracy, to decide when to stop an algorithm, and to compare an algorithm to others in terms of their worst-case complexities.

The techniques we develop in this paper are quite different from those used in classical optimization convergence analysis, mainly because splitting algorithms are driven by fixed-point operators instead of driven by minimizing objectives. (An exception is FBS, which is driven by both.) Splitting algorithms are fixed-point iterations derived from certain optimality conditions, and they converge due to the contraction of the fixed-point operators. Some of them do not even reduce objectives monotonically. Thus, objective convergence is a consequence of operator convergence rather than the cause of it. Therefore, we first perform an operator-theoretic analysis and then, based on these results, derive optimization related rates.

We now describe our contributions and our techniques as follows:

- We show that the FPR of the fixed-point iterations of nonexpansive operators converge with rate $o(1/(k+1))$ (Theorem 1). This rate is optimal and improves on the known big-O rate [17,30]. For the special cases of FBS applied to Problem (1) and one-dimensional DRS, we improve this rate to $o(1/(k+1)^2)$. In addition, we provide examples (Section 6) to show that all of the rates we derive are tight. Specifically, for each rate $o(1/(k+1)^p)$, we give an example with rate $\Omega(1/(k+1)^{p+\varepsilon})$ for any $\varepsilon > 0$. A detailed list of our contributions and a comparison with existing results appear in Section 6.1.1. The analysis is based on establishing summable and, in many cases, monotonic sequences, whose convergence rates are summarized in Lemma 3.
- We demonstrate that even when the DRS algorithm converges in norm to a solution, it may do so *arbitrarily slowly* (Theorem 9).
- We give the objective convergence rates of the relaxed PRS algorithm and show that it is, in the worst case, nearly *as slow as the subgradient method* (Theorems 7 and 11), yet *nearly as fast as PPA* in the ergodic sense (Theorems 6 and 12). The rates are obtained by relating the objective error to the aforementioned *FPR rates* through a *fundamental inequality* (Proposition 4). These rates are also optimal through several examples (Section 7).
- We give the convergence rates of the primal objective and feasibility of the current iterates generated by ADMM. Our analysis follows by a simple application of the Fenchel-Young inequality (Section 8)

1.2 Notation

In what follows, $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2, \mathcal{G}$ denote (possibly infinite dimensional) Hilbert spaces. In fixed-point iterations, $(\lambda_j)_{j\geq 0} \subset \mathbf{R}_+$ will denote a sequence of relaxation parameters and

$$\Lambda_k := \sum_{i=0}^{k} \lambda_i$$

is its kth partial sum. To ease notational memory, the reader may assume that $\lambda_k \equiv (1/2)$ and $\Lambda_k = (k+1)/2$ in the DRS algorithm or that $\lambda_k \equiv 1$ and $\Lambda_k = (k+1)$ in the PRS algorithm. Given the sequence $(x^j)_{j\geq 0} \subset \mathcal{H}$, we let $\overline{x}^k = (1/\Lambda_k)\sum_{i=0}^k \lambda_i x^i$ denote its kth average with respect to the sequence $(\lambda_j)_{j\geq 0}$.

We call a convergence result *ergodic* if it is in terms of the sequence $(\overline{x}^j)_{j\geq 0}$, and *nonergodic* if it is in terms of $(x^j)_{j\geq 0}$.

Given a closed, proper, convex function $f : \mathcal{H} \to (-\infty, \infty]$, $\partial f(x)$ denotes its subdifferential at x and

$$\widetilde{\nabla} f(x) \in \partial f(x), \tag{3}$$

denotes a subgradient. (This notation was used in [5, Eq. (1.10)].)

The convex conjugate of a proper, closed, and convex function f is

$$f^*(y) := \sup_{x\in\mathcal{H}} \langle y, x\rangle - f(x). \tag{4}$$

Let $I_{\mathcal{H}}$ denote the identity map. Finally, for any $x \in \mathcal{H}$ and scalar $\gamma \in \mathbf{R}_{++}$, we let

$$\mathbf{prox}_{\gamma f}(x) := \operatorname*{arg\,min}_{y\in\mathcal{H}} f(y) + \frac{1}{2\gamma}\|y - x\|^2 \quad \text{and} \quad \mathbf{refl}_{\gamma f} := 2\mathbf{prox}_{\gamma f} - I_{\mathcal{H}}, \tag{5}$$

which are known as the *proximal* and *reflection* operators, and we define the PRS operator:

$$T_{\mathrm{PRS}} := \mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\gamma g}. \tag{6}$$

1.3 Assumptions

We list the the assumptions used throughout this papers as follows.

Assumption 1 *Every function we consider is closed, proper, and convex.*

Unless otherwise stated, a function is not necessarily differentiable.

Assumption 2 (Differentiability) *Every differentiable function we consider is Fréchet differentiable [2, Definition 2.45].*

Assumption 3 (Solution existence) *Functions $f, g : \mathcal{H} \to (-\infty, \infty]$ satisfy*

$$\operatorname{zer}(\partial f + \partial g) \neq \emptyset. \tag{7}$$

Note that this assumption is slightly stronger than the existence of a minimizer, because $\operatorname{zer}(\partial f + \partial g) \neq \operatorname{zer}(\partial(f+g))$, in general [2, Remark 16.7]. Nevertheless, this assumption is standard.

1.4 The Algorithms

This paper covers several operator-splitting algorithms that are all based on the atomic evaluation of the *proximal* and *gradient* operators. By default, all algorithms start from an arbitrary $z^0 \in \mathcal{H}$. To minimize a function f, the *proximal point algorithm* (PPA) iteratively applies the proximal operator of f as follows:

$$z^{k+1} = \mathbf{prox}_{\gamma f}(z^k), \quad k = 0, 1, \dots \tag{8}$$

where $\gamma > 0$ is a tuning parameter. Another equivalent form of the iteration, which is often used in this paper, is

$$z^{k+1} = z^k - \gamma \widetilde{\nabla} f(z^{k+1}) \tag{9}$$

where $\widetilde{\nabla} f(z^{k+1}) \in \partial f(z^{k+1})$. Given z^k, the point z^{k+1} is unique and so is the subgradient $\widetilde{\nabla} f(z^{k+1})$ (Lemma 1). The iteration resembles the (sub)gradient descent iteration, which uses a (sub)gradient of f at z^k instead of its (sub)gradient at z^{k+1}.

In the literature, (9) is referred to as the *backward* iteration, where the (sub)gradient is drawn at the destination z^{k+1}. On the contrary, a *forward* iteration draws the (sub)gradient at the start z^k, resulting in the update rule: $z^{k+1} = z^k - \gamma \widetilde{\nabla} f(z^k)$. Most of the splitting schemes in this paper are built from forward, backward, and reflection operators.

In problem (1), let g be a C^1 function with Lipschitz derivative. The *forward-backward splitting* (FBS) algorithm is the iteration:

$$z^{k+1} = \mathbf{prox}_{\gamma f}(z^k - \gamma \nabla g(z^k)), \quad k = 0, 1, \dots. \tag{10}$$

The FBS algorithm directly generalizes PPA and has the following subgradient representation:

$$z^{k+1} = z^k - \gamma \widetilde{\nabla} f(z^{k+1}) - \gamma \nabla g(z^k) \tag{11}$$

where $\widetilde{\nabla} f(z^{k+1}) \in \partial f(z^{k+1})$, and z^{k+1} and $\widetilde{\nabla} f(z^{k+1})$ are unique (Lemma 1) given z^k and $\gamma > 0$.

A direct application of the PPA algorithm (8) to minimizing $f + g$ would require computing the operator $\mathbf{prox}_{\gamma(f+g)}$, which can be difficult to evaluate. The Douglas-Rachford splitting (DRS) algorithm eliminates this difficulty by separately evaluating the proximal operators of f and g as follows:

$$\begin{cases} x_g^k = \mathbf{prox}_{\gamma g}(z^k); \\ x_f^k = \mathbf{prox}_{\gamma f}(2x_g^k - z^k); \qquad k = 0, 1, \dots, \\ z^{k+1} = z^k + (x_f^k - x_g^k). \end{cases}$$

which has the equivalent operator-theoretic and subgradient form (Lemma 4):

$$z^{k+1} = \frac{1}{2}(I_{\mathcal{H}} + T_{\mathrm{PRS}})(z^k) = z^k - \gamma(\widetilde{\nabla} f(x_f^k) + \widetilde{\nabla} g(x_g^k)), \quad k = 0, 1, \dots$$

where $\widetilde{\nabla} f(x_f^k) \in \partial f(x_f^k)$ and $\widetilde{\nabla} g(x_g^k) \in \partial g(x_g^k)$. In the above algorithm, we can replace the $1/2$ average of $I_{\mathcal{H}}$ and T_{PRS} with any other weight, so in this paper we study the *relaxed PRS* algorithm:

Algorithm 1: Relaxed Peaceman-Rachford splitting (relaxed PRS)

input : $z^0 \in \mathcal{H},\ \gamma > 0,\ (\lambda_j)_{j \geq 0} \subseteq (0, 1]$
for $k = 0,\ 1, \dots$ **do**
  $z^{k+1} = (1 - \lambda_k) z^k + \lambda_k \mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\gamma g}(z^k)$;

The special cases $\lambda_k \equiv 1/2$ and $\lambda_k \equiv 1$ are called the DRS and PRS algorithms, respectively.

The relaxed PRS algorithm can be applied to problem (2). To this end we define the Lagrangian:

$$\mathcal{L}(x, y; w) := f(x) + g(y) - \langle w, Ax + By - b\rangle.$$

Section 8 presents Algorithm 1 applied to the Lagrange dual of (2), which reduces to the following algorithm:

Algorithm 2: Relaxed alternating direction method of multipliers (relaxed ADMM)

input : $w^{-1} \in \mathcal{H}, x^{-1} = 0, y^{-1} = 0, \lambda_{-1} = 1/2, \gamma > 0, (\lambda_j)_{j\geq 0} \subseteq (0, 1]$
for $k = -1,\ 0, \ldots$ **do**
$y^{k+1} = \arg\min_y \mathcal{L}(x^k, y; w^k) + \gamma(2\lambda_k - 1)\langle By, (Ax^k + By^k - b)\rangle;$
$w^{k+1} = w^k - \gamma(Ax^k + By^{k+1} - b) - \gamma(2\lambda_k - 1)(Ax^k + By^k - b);$
$x^{k+1} = \arg\min_x \mathcal{L}(x, y^{k+1}; w^{k+1});$

If $\lambda_k \equiv 1/2$, Algorithm 2 recovers the standard ADMM.

Each of the above algorithms is a special case of the Krasnosel'skiĭ-Mann (KM) iteration ([29,31]). An averaged operator is the average of a nonexpansive operator $T : \mathcal{H} \to \mathcal{H}$ and the identity mapping $I_{\mathcal{H}}$. For all $\lambda \in (0, 1]$, define

$$T_\lambda := (1 - \lambda)I_{\mathcal{H}} + \lambda T. \tag{12}$$

The fixed-point iteration of T_λ is the KM algorithm:

Algorithm 3: Krasnosel'skiĭ-Mann (KM)

input : $z^0 \in \mathcal{H}, (\lambda_j)_{j\geq 0} \subseteq (0, 1]$
for $k = 0,\ 1, \ldots$ **do**
$z^{k+1} = T_{\lambda_k}(z^k);$

1.5 Basic properties of averaged operators

This section describes the basic properties of proximal, reflection, nonexpansive, and averaged operators. We demonstrate that proximal and reflection operators are nonexpansive maps, and that averaged operators have a contractive property. These properties are included in textbooks such as [2].

Lemma 1 (Optimality conditions of prox) *Let $x \in \mathcal{H}$. Then $x^+ = \mathbf{prox}_{\gamma f}(x)$ if, and only if, $(1/\gamma)(x - x^+) \in \partial f(x^+)$.*

It is straightforward to use Lemma 1 to deduce the firm nonexpansiveness of the proximal operator.

Proposition 1 (Firm nonexpansiveness of prox) *Let $x, y \in \mathcal{H}$, let $x^+ := \mathbf{prox}_{\gamma f}(x)$, and let $y^+ := \mathbf{prox}_{\gamma f}(y)$. Then*

$$\|x^+ - y^+\|^2 \leq \langle x^+ - y^+, x - y\rangle. \tag{13}$$

In particular, $\mathbf{prox}_{\gamma f}$ is nonexpansive.

The next proposition introduces the most important operator in this paper.

Proposition 2 (Nonexpansiveness of the PRS operator) *The operator* $\mathbf{refl}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ *is nonexpansive. Therefore, the composition is nonexpansive:*

$$T_{\mathrm{PRS}} := \mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\gamma g} \tag{14}$$

The next proposition shows that averaged operators have a nice contraction property.

Proposition 3 (Contraction property of averaged operator) *Let* $T : \mathcal{H} \to \mathcal{H}$ *be a nonexpansive operator. Then for all* $\lambda \in (0, 1]$ *and* $(x, y) \in \mathcal{H} \times \mathcal{H}$*, the averaged operator* T_λ *defined in* (12) *satisfies*

$$\|T_\lambda x - T_\lambda y\|^2 \leq \|x - y\|^2 - \frac{1-\lambda}{\lambda}\|(I_{\mathcal{H}} - T_\lambda)x - (I_{\mathcal{H}} - T_\lambda)y\|^2. \tag{15}$$

If $\lambda = 1/2$, then T_λ is called firmly nonexpansive. Rearranging Equation (15) shows that a nonexpansive operator T is firmly nonexpansive, if, and only if, for all $x, y \in \mathcal{H}$, the inequality holds:

$$\|Tx - Ty\|^2 \leq \langle Tx - Ty, x - y\rangle.$$

The next corollary applies Proposition 3 to $\mathbf{prox}_{\gamma f}$.

Corollary 1 (Proximal operators are $1/2$**-averaged)** *The operator* $\mathbf{prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ *is* $1/2$*-averaged and satisfies the following contraction property:*

$$\|\mathbf{prox}_{\gamma f}(x) - \mathbf{prox}_{\gamma f}(y)\|^2 \leq \|x - y\|^2 - \|(x - \mathbf{prox}_{\gamma f}(x)) - (y - \mathbf{prox}_{\gamma f}(y))\|^2. \tag{16}$$

The following lemma relates the fixed points of T_λ to those of T.

Lemma 2 *Let* $T : \mathcal{H} \to \mathcal{H}$ *be nonexpansive and* $\lambda > 0$*. Then,* T_λ *and* T *have the same set of fixed points.*

Finally, we note that the forward and forward-backward operators are averaged whenever the implicit stepsize parameter γ is small enough. See Section 3.3 for more details.

2 Summable sequence convergence lemma

This section presents a lemma on the convergence rates of nonnegative summable sequences. Such sequences are constructed throughout this paper to establish various rates.

Lemma 3 (Summable sequence convergence rates) *Suppose that the nonnegative scalar sequences* $(\lambda_j)_{j\geq 0}$ *and* $(a_j)_{j\geq 0}$ *satisfy* $\sum_{i=0}^{\infty} \lambda_i a_i < \infty$*. Let* $\Lambda_k := \sum_{i=0}^{k} \lambda_i$ *for* $k \geq 0$*.*

1. **Monotonicity:** *If* $(a_j)_{j\geq 0}$ *is* monotonically nonincreasing*, then*

$$a_k \leq \frac{1}{\Lambda_k}\left(\sum_{i=0}^{\infty} \lambda_i a_i\right) \qquad \text{and} \qquad a_k = o\left(\frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}}\right). \tag{17}$$

In particular,
(a) if $(\lambda_j)_{j\geq 0}$ *is bounded away from* 0 *and* ∞*, then* $a_k = o(1/(k+1))$*;*
(b) if $\lambda_k = (k+1)^p$ *for some* $p \geq 0$ *and all* $k \geq 1$*, then* $a_k = o(1/(k+1)^{p+1})$*;*
(c) as a special case, if $\lambda_k = (k+1)$ *for all* $k \geq 0$*, then* $a_k = o(1/(k+1)^2)$*.*

2. **Monotonicity up to errors:** *Let $(e_j)_{j\geq 0}$ be a sequence of scalars. Suppose that $a_{k+1} \leq a_k + e_k$ for all k (where e_k represents an error) and that $\sum_{i=0}^{\infty} \Lambda_i e_i < \infty$. Then*

$$a_k \leq \frac{1}{\Lambda_k}\left(\sum_{i=0}^{\infty} \lambda_i a_i + \sum_{i=0}^{\infty} \Lambda_i e_i\right) \qquad \text{and} \qquad a_k = o\left(\frac{1}{\Lambda_k - \Lambda_{\lceil k/2 \rceil}}\right).$$

In particular, the results of Parts 1a, 1b, and 1c continue to hold if $e_k = O(1/(k+1)^q)$ for some $q > 2$, $q > p+2$, and $q > 3$, respectively.

3. **Faster rates:** *Suppose $(b_j)_{j\geq 0}$ and $(e_j)_{j\geq 0}$ are nonnegative scalar sequences. Suppose that $\sum_{i=0}^{\infty} b_j < \infty$ and $\sum_{i=0}^{\infty}(i+1)e_i < \infty$. Finally, suppose that for all $k \geq 0$ we have $\lambda_k a_k \leq b_k - b_{k+1} + e_k$. Then the following sum is finite:*

$$\sum_{i=0}^{\infty}(i+1)\lambda_i a_i \leq \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty}(i+1)e_i < \infty. \tag{18}$$

4. **No monotonicity:** *For all $k \geq 0$, define the sequence of indices*

$$k_{\text{best}} := \arg\min_i \{a_i | i = 0, \cdots, k\}.$$

Then $(a_{j_{\text{best}}})_{j\geq 0}$ is monotonically nonincreasing and the above bounds continue to hold when a_k is replaced with $a_{k_{\text{best}}}$.

Proof Part 1. Because $a_k \leq a_i$ whenever $k \geq i$ and the inequality holds $\lambda_i a_i \geq 0$, we get the upper bound $\Lambda_k a_k \leq \sum_{i=0}^{k} \lambda_i a_i \leq \sum_{i=0}^{\infty} \lambda_i a_i$. This proves the first part of (17). To prove the second part of (17), observe that

$$(\Lambda_{2k} - \Lambda_k)a_{2k} \leq \lambda_{2k}a_{2k} + \lambda_{2k-1}a_{2k-1} + \cdots \lambda_{k+1}a_{k+1} = \sum_{i=k+1}^{2k} \lambda_i a_i \overset{k\to\infty}{\to} 0.$$

Part 1a. Let $\underline{\lambda} := \inf_{j\geq 0} \lambda_j > 0$ and $\overline{\lambda} := \sup_{j\geq 0} \lambda_j < \infty$. Then $(k+1)\underline{\lambda} \leq \Lambda_k \leq (k+1)\overline{\lambda}$, and we have the lower bound, $\Lambda_k - \Lambda_{\lceil k/2 \rceil} \geq \underline{\lambda}\Omega(k) + \Lambda_{\lceil k/2 \rceil} - \Lambda_{\lceil k/2 \rceil} = \Omega(\underline{\lambda}k)$. Therefore, $a_k = o(1/(\Lambda_k - \Lambda_{\lceil k/2 \rceil})) = o(1/(k+1))$.

Part 1b. Because $\Lambda_{\lceil k/2 \rceil} \leq 1 + \int_0^{\lceil k/2 \rceil}(t+1)^p \mathrm{d}t = (p+1)^{-1}(\lceil k/2 \rceil + 1)^{p+1} + 1 - (p+1)^{-1}$ and $\Lambda_k \geq \int_0^k (t+1)^p \mathrm{d}t = (p+1)^{-1}(k+1)^{p+1} - (p+1)^{-1}$, we have $\Lambda_k - \Lambda_{\lceil k/2 \rceil} \geq (p+1)^{-1}((k+1)^{p+1} - (\lceil k/2 \rceil + 1)^{p+1}) - 1 = \Omega((k+1)^{p+1})$. Therefore, $a_k = o(1/(\Lambda_k - \Lambda_{\lceil k/2 \rceil})) = o(1/(k+1)^{p+1})$.

Part 1c directly follows from Part 1b.

Part 2 is a straightforward extension of Part 1.

Part 3. Note that

$$\lambda_k(k+1)a_k \leq (k+1)b_k - (k+1)b_{k+1} + (k+1)e_k = b_{k+1} + ((k+1)b_k - (k+2)b_k) + (k+1)e_k.$$

Thus, because the upper bound on $(k+1)\lambda_k a_k$ is the sum of a telescoping term and a summable term, we have $\sum_{i=0}^{\infty}(i+1)\lambda_i a_i \leq \sum_{i=0}^{\infty} b_i + \sum_{i=0}^{\infty}(i+1)e_i < \infty$.

Part 4 is straightforward, so we omit its proof. □

Part 1 of Lemma 3 is a generalization of [28, Theorem 3.3.1] and [19, Lemma 1.2], which state that a nonnegative, summable, monotonic sequence converges at the rate of $o(1/(k+1))$. This result is key for deducing the convergence rates of several quantities in this paper.

Without sequence monotonicity, we cannot establish such a strong rate for the entire sequence, $(a_j)_{j\geq 0}$. Furthermore, the size of the "best iterate" does not predict the actual behavior of the entire sequence. Indeed, let $h : \mathbf{R}_+ \to \mathbf{R}_+$ be a monotonic function increasing to ∞. Define a sequence: for all $k \geq 0$, let

$$a_k = \begin{cases} \frac{1}{h^{-1}(k)} & \text{if } k = \lceil h(n^2) \rceil \text{ for some } n \in \mathbf{N}; \\ 0 & \text{otherwise.} \end{cases}$$

Then $h^{-1}(\lceil h(x) \rceil) \geq h^{-1}(h(x)) = x$ and, hence, $a_{\lceil h(n^2) \rceil} \leq 1/n^2$ and $\sum_{i=0}^{\infty} a_i \leq \sum_{i=1}^{\infty}(1/i^2) < \infty$. For example, if $h(x) = e^x$, then we must wait until $k \geq \lceil e^{n^2} \rceil$ to guarantee that $a_k \leq 1/n^2$, whereas the best iterate must have order $o(1/\lceil e^{n^2} \rceil)$.

3 Iterative fixed-point residual analysis

In this section we establish the convergence rate of the *fixed-point residual* (FPR), $\|Tz^k - z^k\|^2$, at the kth iteration of Algorithm 3.

The convergence of Algorithm 3 is well-studied [15,17,30]. In particular, weak convergence of $(z^j)_{j\geq 0}$ to a fixed point of T holds under mild conditions on the sequence $(\lambda_j)_{j\geq 0}$ [15, Theorem 3.1]. Because strong convergence of Algorithm 3 may fail (in the infinite dimensional setting), the quantity $\|z^k - z^*\|$ where z^* is a fixed point of T may be bounded above zero for all $k \geq 0$. However, the property $\lim_{k\to\infty} \|Tz^k - z^k\| = 0$, known as *asymptotic regularity* [11], always holds when a fixed point of T exists. Thus, we can always measure the convergence rate of the FPR.

We measure $\|Tz^k - z^k\|^2$ when we could just as well measure $\|Tz^k - z^k\|$. We choose to measure the squared norm because it naturally appears in our analysis. In addition, it is summable and monotonic, which is naturally analyzable by Lemma 3.

In first-order optimization algorithms, the FPR typically relates to the size of objective gradient. For example, in the unit-step gradient descent algorithm, $z^{k+1} = z^k - \nabla f(z^k)$, the FPR is given by $\|\nabla f(z^k)\|^2$. In the proximal point algorithm, the FPR is given by $\|\nabla f(z^{k+1})\|^2$. When the objective is the sum of multiple functions, the FPR is a combination of the (sub)gradients of those functions in the objective. Using the subgradient inequality, we will derive a rate on $f(z^k) - f(x^*)$ from a rate on the FPR.

3.1 $o(1/(k+1))$ FPR of averaged operators

We now prove the main result of this section. We do not include the known weak convergence result [15, Theorem 3.1], but we deduce a convergence rate for the FPR.

Theorem 1 (Convergence rate of averaged operators) *Let $T : \mathcal{H} \to \mathcal{H}$ be a nonexpansive operator, let z^* be a fixed point of T, let $(\lambda_j)_{j\geq 0} \subseteq (0, 1]$ be a sequence of positive numbers, let $\tau_k := \lambda_k(1-\lambda_k)$, and let $z^0 \in \mathcal{H}$. Suppose that $(z^j)_{j\geq 0} \subseteq \mathcal{H}$ is generated by Algorithm 3: for all $k \geq 0$, let*

$$z^{k+1} = T_{\lambda_k}(z^k). \tag{19}$$

Then, the following results hold

1. $\|z^k - z^*\|^2$ *is monotonically nonincreasing;*
2. $\|Tz^k - z^k\|^2$ *is monotonically nonincreasing;*
3. $\tau_k\|Tz^k - z^k\|^2$ *is summable:*

$$\sum_{i=0}^{\infty} \tau_i \|Tz^i - z^i\|^2 \leq \|z^0 - z^*\|^2; \tag{20}$$

4. *if* $\tau_k > 0$ *for all* $k \geq 0$*, then the convergence estimates hold:*

$$\|Tz^k - z^k\|^2 \leq \frac{\|z^0 - z^*\|^2}{\sum_{i=0}^k \tau_i} \qquad \text{and} \qquad \|Tz^k - z^k\|^2 = o\left(\frac{1}{\sum_{i=\lceil \frac{k}{2} \rceil + 1}^k \tau_i}\right). \tag{21}$$

In particular, if $(\tau_j)_{j\geq 0} \subseteq (\varepsilon, \infty)$ *for some* $\varepsilon > 0$*, then* $\|Tz^k - z^k\|^2 = o(1/(k+1))$.

Proof Part 1 follows from the Fejér–type inequality:

$$\begin{aligned}
\|z^{k+1} - z^*\|^2 &= \|T_{\lambda_k} z^k - T_{\lambda_k} z^*\|^2 \\
&\overset{(15)}{\leq} \|z^k - z^*\|^2 - \frac{1-\lambda_k}{\lambda_k}\|(z^k - T_{\lambda_k}(z^k)) - (z^* - T_{\lambda_k} z^*)\|^2 \\
&= \|z^k - z^*\|^2 - \frac{1-\lambda_k}{\lambda_k}\|z^{k+1} - z^k\|^2.
\end{aligned} \tag{22}$$

Part 2. Recall that

$$z^{k+1} - z^k = \lambda_k(Tz^k - z^k). \tag{23}$$

By the triangle inequality and the nonexpansiveness of T, the FPR is monotonic:

$$\begin{aligned}
\|Tz^{k+1} - z^{k+1}\| &\leq \|Tz^{k+1} - ((1-\lambda_k)z^k + \lambda_k Tz^k)\| \\
&\leq \|Tz^{k+1} - Tz^k\| + (1-\lambda_k)\|Tz^k - z^k\| \\
&\leq \|z^{k+1} - z^k\| + (1-\lambda_k)\|Tz^k - z^k\| \\
&\overset{(23)}{=} \|Tz^k - z^k\|.
\end{aligned}$$

We get Part 3 by $((1-\lambda_k)/\lambda_k)\|z^{k+1} - z^k\|^2 \overset{(23)}{=} \tau_k\|Tz^k - z^k\|^2$ and summing Equation (22) over all $k \geq 0$.

Part 4 is a direct application of Lemma 3 to the sequences τ_k and $\|Tz^k - z^k\|^2$.

□

3.1.1 Notes on Theorem 1

The FPR, $\|Tz^k - z^k\|^2$, is a normalized version of the successive iterate differences $z^{k+1} - z^k = \lambda_k(Tz^k - z^k)$. Thus, the convergence rates of $\|Tz^k - z^k\|^2$ naturally induce convergence rates of $\|z^{k+1} - z^k\|^2$.

Note that $o(1/(k+1))$ is the optimal convergence rate for the class of nonexpansive operators [9, Remarque 4]. In the special case that $T = \mathbf{prox}_{\gamma f}$ for some closed, proper, and convex function f, the rate of $\|Tz^k - z^k\|^2$ improves to $O(1/(k+1)^2)$ [9, Théorème 9]. See section 6 for more optimality results.

In general, it is possible that the nonexpansive operator, $T : \mathcal{H} \to \mathcal{H}$, is already averaged, i.e. there exists a nonexpansive operator $N : \mathcal{H} \to \mathcal{H}$ and a positive constant $\alpha \in (0, 1]$ such that $T = (1 - \alpha)I_{\mathcal{H}} + \alpha N$. In this case, Lemma 2 shows that T and N share the same fixed point set. Thus, we can apply Theorem 1 to $N = (1 - (1/\alpha))I_{\mathcal{H}} + (1/\alpha)T$. Furthermore, $N_\lambda = (1 - \lambda/\alpha)I_{\mathcal{H}} + (\lambda/\alpha)T$. Thus, when we translate this back to an iteration on T, it enlarges the region of relaxation parameters to $\lambda_k \in (0, 1/\alpha)$ and modifies τ_k accordingly to $\tau_k = \lambda_k(1 - \alpha\lambda_k)/\alpha$, and the same convergence results continue to hold.

To the best of our knowledge, The little-o rates produced in Theorem 1 have never been established for the KM iteration. It is also possible to extend our results to the inexact version of Equation (19),

$$z^{k+1} = z^k + \lambda_k(Tz^k - z^k + e^k), \tag{24}$$

and maintain little-o convergence rates as long as $\lambda_k\|e^k\| = O(1/(k + 1)^q)$ for any $q > 1$, by using Part 2 of Lemma 3. See [17,30] for similar big-O results. Note that in the Banach space case, we cannot improve the big-O rates to little-o [17, Section 2.4].

3.2 $o(1/(k + 1))$ FPR of relaxed PRS

In this section, we apply Theorem 1 to the T_{PRS} operator defined in Proposition 2. For the special case of DRS ((1/2)-averaged PRS), it is straightforward to establish the rate of the FPR

$$\|(T_{\text{PRS}})_{1/2}z^k - z^k\|^2 = O\left(\frac{1}{k+1}\right)$$

from two existing results: (i) the DRS iteration is a proximal iteration applied to a certain monotone operator [20, Section 4]; (ii) the convergence rate of the FPR for proximal iterations is $O(1/(k+1))$ [9, Proposition 8] whenever a fixed point exists. Our results below are established for general averaged PRS operators and the rate is improved to $o(1/(k + 1))$.

The following corollary is an immediate consequence of Theorem 1.

Corollary 2 (Convergence rate of relaxed PRS) *Let z^* be a fixed point of T_{PRS}, let $(\lambda_j)_{j\geq 0} \subseteq (0, 1]$ be a sequence of positive numbers, let $\tau_k := \lambda_k(1 - \lambda_k)$ for all $k \geq 0$, and let $z^0 \in \mathcal{H}$. Suppose that $(z^j)_{j\geq 0} \subseteq \mathcal{H}$ is generated by Algorithm 1. Then the sequence $\|z^k - z^*\|^2$ is monotonically nonincreasing and the following inequality holds:*

$$\sum_{i=0}^{\infty} \tau_i\|T_{\text{PRS}}z^i - z^i\|^2 \leq \|z^0 - z^*\|^2. \tag{25}$$

Furthermore, if $\underline{\tau} := \inf_{j\geq 0} \tau_j > 0$, then the following convergence rates hold:

$$\|T_{\text{PRS}}z^k - z^k\|^2 \leq \frac{\|z^0 - z^*\|^2}{\underline{\tau}(k+1)} \qquad \textit{and} \qquad \|T_{\text{PRS}}z^k - z^k\|^2 = o\left(\frac{1}{\underline{\tau}(k+1)}\right). \tag{26}$$

3.3 $o(1/(k+1)^2)$ FPR of FBS and PPA

In this section, we assume that ∇g is $(1/\beta)$-Lipschitz, and we analyze the convergence rate of FBS algorithm given in Equations (10) and (11). If $g = 0$, FBS reduces to PPA and $\beta = \infty$. If $f = 0$, FBS reduces to gradient descent. The FBS algorithm can be written in the following operator form:

$$T_{\text{FBS}} := \mathbf{prox}_{\gamma f} \circ (I - \gamma \nabla g).$$

Because $\mathbf{prox}_{\gamma f}$ is $(1/2)$-averaged and $I - \gamma \nabla g$ is $\gamma/(2\beta)$-averaged [2, Proposition 4.33], it follows that T_{FBS} is α_{FBS}-averaged for

$$\alpha_{\text{FBS}} := \frac{\max\{1, \frac{\gamma}{\beta}\}}{\max\{\frac{1}{2}, \frac{\gamma}{2\beta}\} + 1}$$

whenever $\gamma < 2\beta$ [2, Proposition 4.32]. Thus, we have $T_{\text{FBS}} = (1-\alpha_{\text{FBS}})I + \alpha_{\text{FBS}} T$ for a certain nonexpansive operator T, and $T_{\text{FBS}}(z^k) - z^k = \alpha_{\text{FBS}}(Tz^k - z^k)$. In particular, for all $\gamma < 2\beta$ the following sum is finite:

$$\sum_{i=0}^{\infty} \|T_{\text{FBS}}(z^k) - z^k\|^2 \overset{(20)}{\leq} \frac{\alpha_{\text{FBS}} \|z^0 - z^*\|^2}{(1 - \alpha_{\text{FBS}})}.$$

To analyze the FBS algorithm we need to derive a joint subgradient inequality for $f + g$. First, we recall the following sufficient descent property for Lipschitz differentiable functions.

Theorem 2 (Descent theorem) *If g is differentiable and ∇g is $(1/\beta)$-Lipschitz, then for all $x, y \in \text{dom}(g)$ we have the upper bound*

$$g(x) \leq g(y) + \langle x - y, \nabla g(y) \rangle + \frac{1}{2\beta} \|x - y\|^2. \tag{27}$$

Proof See [2, Theorem 18.15(iii)]. □

Corollary 3 (Joint descent theorem) *If g is differentiable and ∇g is $(1/\beta)$-Lipschitz, then for all points $x, y \in \text{dom}(g) \cap \text{dom}(f)$ and $z \in \text{dom}(g)$, and subgradients $\widetilde{\nabla} f(x) \in \partial f(x)$, we have*

$$f(x) + g(x) \leq f(y) + g(y) + \langle x - y, \nabla g(z) + \widetilde{\nabla} f(x) \rangle + \frac{1}{2\beta} \|z - x\|^2. \tag{28}$$

Proof Inequality (28) follows from adding the upper bound

$$g(x) - g(y) \leq g(z) - g(y) + \langle x - z, \nabla g(z) \rangle + \frac{1}{2\beta} \|z - x\|^2 \leq \langle x - y, \nabla g(z) \rangle + \frac{1}{2\beta} \|z - x\|^2$$

with the subgradient inequality:

$$f(x) \leq f(y) + \langle x - y, \widetilde{\nabla} f(x) \rangle. \tag{29}$$

□

We now improve the $O(1/(k+1)^2)$ FPR rate for PPA in [9, Théorème 9] by showing that the FPR rate of FBS is actually $o(1/(k+1)^2)$.

Theorem 3 (Objective and FPR convergence of FBS) *Let $z^0 \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$. Suppose that $(z^j)_{j\geq 0}$ is generated by FBS (iteration (10)) where ∇g is $(1/\beta)$-Lipschitz and $\gamma < 2\beta$. Then for all $k \geq 0$,*

$$f(z^{k+1}) + g(z^{k+1}) - f(x^*) - g(x^*) \leq \frac{\|z^0 - x^*\|^2}{k+1} \times \begin{cases} \frac{1}{2\gamma} & \text{if } \gamma \leq \beta; \\ \left(\frac{1}{2\gamma} + \left(\frac{1}{2\beta} - \frac{1}{2\gamma}\right)\frac{\alpha_{\mathrm{FBS}}}{(1-\alpha_{\mathrm{FBS}})}\right) & \text{otherwise.} \end{cases},$$

and

$$f(z^{k+1}) + g(z^{k+1}) - f(x^*) - g(x^*) = o(1/(k+1)).$$

In addition, for all $k \geq 0$, we have

$$\|T_{\mathrm{FBS}} z^{k+1} - z^{k+1}\|^2 \leq \frac{\|z^0 - x^*\|^2}{\left(\frac{1}{\gamma} - \frac{1}{2\beta}\right)(k+1)^2} \times \begin{cases} \frac{1}{2\gamma} & \text{if } \gamma \leq \beta; \\ \left(\frac{1}{2\gamma} + \left(\frac{1}{2\beta} - \frac{1}{2\gamma}\right)\frac{\alpha_{\mathrm{FBS}}}{(1-\alpha_{\mathrm{FBS}})}\right) & \text{otherwise.} \end{cases},$$

and

$$\|T_{\mathrm{FBS}} z^{k+1} - z^{k+1}\|^2 = o(1/(k+1)^2).$$

Proof Recall that $z^k - z^{k+1} = \gamma\widetilde{\nabla} f(z^{k+1}) + \gamma\nabla g(z^k)$ for all $k \geq 0$. Thus, the joint descent theorem shows that for all $x \in \mathrm{dom}(f)$, we have

$$\begin{aligned} f(z^{k+1}) + g(z^{k+1}) - f(x) - g(x) &\overset{(28)}{\leq} \frac{1}{\gamma}\langle z^{k+1} - x, z^k - z^{k+1}\rangle + \frac{1}{2\beta}\|z^k - z^{k+1}\|^2 \\ &= \frac{1}{2\gamma}\left(\|z^k - x\|^2 - \|z^{k+1} - x\|^2\right) + \left(\frac{1}{2\beta} - \frac{1}{2\gamma}\right)\|z^{k+1} - z^k\|^2. \end{aligned} \tag{30}$$

Let $h := f + g$. If we set $x = x^*$ in Equation (30), we see that $(h(z^{j+1}) - h(x^*))_{j\geq 0}$ is positive, summable, and

$$\sum_{i=0}^{\infty}\left(h(x^{i+1}) - h(x^*)\right) \leq \begin{cases} \frac{1}{2\gamma}\|z^0 - x^*\|^2 & \text{if } \gamma \leq \beta; \\ \left(\frac{1}{2\gamma} + \left(\frac{1}{2\beta} - \frac{1}{2\gamma}\right)\frac{\alpha_{\mathrm{FBS}}}{(1-\alpha_{\mathrm{FBS}})}\right)\|z^0 - x^*\|^2 & \text{otherwise.} \end{cases} \tag{31}$$

In addition, if we set $x = z^k$ in Equation (30), then we see that $(h(z^{j+1}) - h(x^*))_{j\geq 0}$ is decreasing:

$$\left(\frac{1}{\gamma} - \frac{1}{2\beta}\right)\|z^{k+1} - z^k\|^2 \leq h(z^k) - h(z^{k+1}) = (h(z^k) - h(x^*)) - (h(z^{k+1}) - h(x^*)).$$

Therefore, the rates for $f(z^{k+1}) + g(z^{k+1}) - f(x^*) - g(x^*)$ follow by Lemma 3 Part 1a, with $a_k = h(z^{k+1}) - h(x^*)$ and $\lambda_k \equiv 1$.

Now we prove the rates for $\|T_{\mathrm{FBS}} z^{k+1} - z^{k+1}\|^2$. We apply Part 3 of Lemma 3 with $a_k = (1/\gamma - 1/(2\beta))\,\|z^{k+2} - z^{k+1}\|^2$, $\lambda_k \equiv 1$, $e_k = 0$, and $b_k = h(z^{k+1}) - h(x^*)$ for all $k \geq 0$, to show that $\sum_{i=0}^{\infty}(i+1)a_i$ is less than the sum in Equation (31). Part 2 of Theorem 1 shows that $(a_j)_{j\geq 0}$ is monotonically nonincreasing. Therefore, the convergence rate of $(a_j)_{j\geq 0}$ follows from Part 1b of Lemma 3. □

When $f = 0$, the objective error upper bound in Theorem 3 is strictly better than the bound provided in [34, Corollary 2.1.2]. In FBS, the objective error rate is the same as the one derived in [4, Theorem 3.1], when $\gamma \in (0, \beta]$, and is new in the case that $\gamma \in (\beta, 2\beta)$. The little-$o$ FPR rate is new in all cases, and the upper bound for the FPR of PPA is tighter (by a factor of $\sqrt{2}$) than the one given in [9, Théorème 9]

3.4 $o(1/(k+1)^2)$ FPR of one dimensional DRS

Whenever the operator $(T_{\mathrm{PRS}})_{1/2}$ is applied in $\mathbf{R}$, the convergence rate of the FPR improves to $o(1/(k+1)^2)$.

Theorem 4 *Suppose that $\mathcal{H} = \mathbf{R}$, and suppose that $(z^j)_{j\geq 0}$ is generated by the DRS algorithm, i.e. Algorithm 1 with $\lambda_k \equiv 1/2$. Then for all $k \geq 0$,*

$$\|(T_{\mathrm{PRS}})_{1/2}z^{k+1} - z^{k+1}\|^2 = \frac{\|z^0 - z^*\|^2}{2(k+1)^2} \qquad \text{and} \qquad \|(T_{\mathrm{PRS}})_{1/2}z^{k+1} - z^{k+1}\|^2 = o\left(\frac{1}{(k+1)^2}\right).$$

Proof Note that $(T_{\mathrm{PRS}})_{1/2}$ is $(1/2)$-averaged, and, hence, it is the resolvent of some maximal monotone operator on $\mathbf{R}$ [2, Corollary 23.8]. Furthermore, every maximal monotone operator on $\mathbf{R}$ is the subdifferential operator of a closed, proper, and convex function [2, Corollary 22.19]. Therefore, DRS is equivalent to the proximal point algorithm applied to a certain convex function on $\mathbf{R}$. Thus, the result follows by Theorem 3 applied to this function. □

3.5 $O(1/\Lambda_k^2)$ ergodic FPR of Fejér monotone sequences

The following definition has proved to be quite useful in the analysis of optimization algorithms [14].

Definition 1 (Fejér monotone sequences) A sequence $(z^j)_{j\geq 0} \subseteq \mathcal{H}$ is *Fejér monotone* with respect to a nonempty set $C \subseteq \mathcal{H}$ if for all $z \in C$,

$$\|z^{k+1} - z\|^2 \leq \|z^k - z\|^2.$$

The following fact is trivial, but allows us to deduce ergodic convergence rates of many algorithms.

Theorem 5 *Let $(z^j)_{j\geq 0}$ be a Fejér monotone sequence with respect to a nonempty set $C \subseteq \mathcal{H}$. Suppose that $z^{k+1} - z^k = \lambda_k(x^k - y^k)$ for a sequence $((x^j, y^j))_{j\geq 0} \subseteq \mathcal{H}^2$, and a sequence of positive real numbers $(\lambda_j)_{j\geq 0}$. For all $k \geq 0$, let $\overline{z}^k := (1/\Lambda_k)\sum_{i=0}^k \lambda_k z^k$, let $\overline{x}^k := (1/\Lambda_k)\sum_{i=0}^k \lambda_k x^k$, and let $\overline{y}^k := (1/\Lambda_k)\sum_{i=0}^k \lambda_k y^k$. Then we get the following bound for all $z \in C$:*

$$\|\overline{x}^k - \overline{y}^k\|^2 \leq \frac{4\|z^0 - z\|^2}{\Lambda_k^2}.$$

Proof It follows directly from the inequality:

$$\|\overline{x}^k - \overline{y}^k\| = \frac{\left\|\sum_{i=0}^k \left(z^{k+1} - z^k\right)\right\|}{\Lambda_k} = \frac{\left\|z^{k+1} - z^0\right\|}{\Lambda_k} \leq \frac{2\left\|z^0 - z\right\|}{\Lambda_k}.$$

□

In view of Part 1 of Theorem 1, we see that any sequence $(z^j)_{j\geq 0}$ generated by Algorithm 3 is Fejér monotone to the set of fixed points of T. Therefore, Theorem 5 directly applies to the KM iteration in Equation (19) with the choice $x^k = Tz^k$ and $y^k = z^k$ for all $k \geq 0$.

The interested reader can proceed to Section 6 for several examples that show the optimality of the rates predicted in this section.

4 Subgradients and fundamental inequalities

We now shift the focus from operator-theoretic analysis to function minimization. This section establishes fundamental inequalities that connect the *FPR* in Section 3 to the *objective error* of the relaxed PRS algorithm.

In first-order optimization algorithms, we only have access to (sub)gradients and function values. Consequently, the FPR at each iteration is usually some linear combination of (sub)gradients. In simple first-order algorithms, for example the (sub)gradient method, a (sub)gradient is drawn from a single point at each iteration. In splitting algorithms for problems with multiple convex functions, each function draws at subgradient at a different point. There is no natural point at which we can evaluate the entire objective function; this complicates the analysis of the relaxed PRS algorithm.

In the relaxed PRS algorithm, there are two objective functions f and g, and the two operators $\mathbf{refl}_{\gamma f}$ and $\mathbf{refl}_{\gamma g}$ are calculated one after another at different points, neither of which equals z^k or z^{k+1}. Consequently, the expression $z^k - z^{k+1}$ is more complicated, and the analysis for standard (sub)gradient iteration does not carry through.

We **let x_f and x_g be the points where subgradients of f and g are drawn, respectively**, and introduce a triangle diagram in Figure 1 for deducing the algebraic relations among points z, x_f and x_g. These relations will be used frequently in our analysis. Propositions 4 and 5 use this diagram to bound the objective error in terms of the FPR. In these bounds, the objective errors of f and g are measured at two points x_f and x_g such that $x_f \neq x_g$. Later we will assume that one of the objectives is Lipschitz continuous and evaluate both functions at the same point (See Corollaries 4 and 5).

We conclude this introduction by combining the subgradient notation in Equation (3) and Lemma 1 to arrive at the expressions

$$\mathbf{prox}_{\gamma f}(x) = x - \gamma \widetilde{\nabla} f(\mathbf{prox}_{\gamma f}(x)) \qquad \text{and} \qquad \mathbf{refl}_{\gamma f}(x) = x - 2\gamma \widetilde{\nabla} f(\mathbf{prox}_{\gamma f}(x)). \tag{32}$$

With this notation, we can decompose the FPR at each iteration of the relaxed PRS algorithm in terms of subgradients drawn at certain points.

4.1 A subgradient representation of relaxed PRS

In this section we write the relaxed PRS algorithm in terms of subgradients. Lemma 4, Table 1, and Figure 1 summarize a single iteration of relaxed PRS.

The way to read Figure 1 is the following: Given input z, relaxed PRS takes a *backward–forward* step with respect to g, then takes a *backward–forward* step with respect to f, resulting in the point $T_{\mathrm{PRS}}(z)$. (Refer to the discussion below (9) for the concepts of "backward" and "forward.") Finally, it averages the input and output: $(T_{\mathrm{PRS}})_\lambda(z) = (1-\lambda)z + \lambda T_{\mathrm{PRS}}(z)$.

Lemma 4 summarizes and proves the identities depicted in Figure 1.

Lemma 4 *Let $z \in \mathcal{H}$. Define auxiliary points $x_g := \mathbf{prox}_{\gamma g}(z)$ and $x_f := \mathbf{prox}_{\gamma f}(\mathbf{refl}_{\gamma g}(z))$. Then the identities hold:*

$$x_g = z - \gamma \widetilde{\nabla} g(x_g) \qquad \text{and} \qquad x_f = x_g - \gamma \widetilde{\nabla} g(x_g) - \gamma \widetilde{\nabla} f(x_f). \tag{33}$$

In addition, each relaxed PRS step has the following representation:

$$(T_{\mathrm{PRS}})_\lambda(z) - z = 2\lambda(x_f - x_g) = -2\lambda\gamma(\widetilde{\nabla} g(x_g) + \widetilde{\nabla} f(x_f)). \tag{34}$$

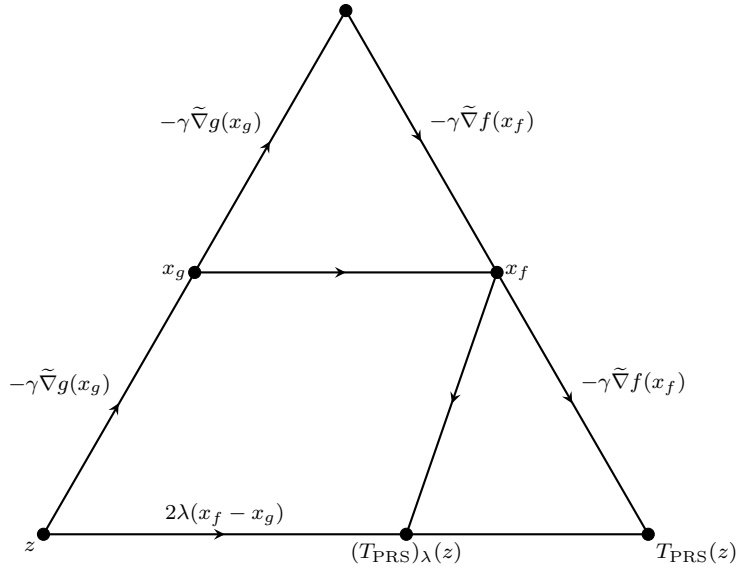

Fig. 1 A single relaxed PRS iteration, from z to $(T_{\mathrm{PRS}})_\lambda(z)$.

Point	Operator identity	Subgradient identity
x_g^s	$= \mathbf{prox}_{\gamma g}(z^s)$	$= z^s - \gamma\widetilde{\nabla}g(x_g^s)$
x_f^s	$= \mathbf{prox}_{\gamma f}(\mathbf{refl}_{\gamma g}(z^s))$	$= x_g^s - \gamma(\widetilde{\nabla}g(x_g^s) + \widetilde{\nabla}f(x_f^s))$
$(T_{\mathrm{PRS}})_\lambda(z^s)$	$= (1-\lambda)z^s + \lambda T_{\mathrm{PRS}}(z^s)$	$= z^s - 2\gamma\lambda(\widetilde{\nabla}g(x_g^s) + \widetilde{\nabla}f(x_f^s))$

Table 1 Overview of the main identities used throughout the paper. The letter s denotes a superscript (e.g. $s = k$ or $s = *$). The vector $z^s \in \mathcal{H}$ is an arbitrary input point. See Lemma 4 for a proof.

Proof Figure 1 provides an illustration of the identities. Equation (33) follows from $\mathbf{refl}_{\gamma g}(z) = 2x_g - z = x_g - \gamma\widetilde{\nabla}g(x_g)$ and Equation (32). Now, we can compute $T_{\mathrm{PRS}}(z) - z$:

$$T_{\mathrm{PRS}}(z) - z \overset{(14)}{=} \mathbf{refl}_{\gamma f}(\mathbf{refl}_{\gamma g}(z)) - z = 2x_f - \mathbf{refl}_{\gamma g}(z) - z = 2x_f - (2x_g - z) - z = 2(x_f - x_g).$$

The subgradient identity in (34) follows from (33). Finally, Equation (34) follows from $(T_{\mathrm{PRS}})_\lambda(z) - z = (1-\lambda)z + \lambda T_{\mathrm{PRS}}(z) - z = \lambda(T_{\mathrm{PRS}}(z) - z)$. □

4.2 Optimality conditions of relaxed PRS

The following lemma characterizes the zeros of $\partial f + \partial g$ in terms of the fixed points of the PRS operator. The intuition is the following: If z^* is a fixed point of T_{PRS}, then the base of the triangle in Figure 1 has length zero. Thus, $x^* := x_g^* = x_f^*$, and if we travel around the perimeter of the triangle, we will start and begin at z^*. This shows that $-2\gamma\widetilde{\nabla}g(x^*) = 2\gamma\widetilde{\nabla}f(x^*)$, i.e. $x^* \in \mathrm{zer}(\partial f + \partial g)$.

Lemma 5 (Optimality conditions of T_{PRS}) *The following identity holds:*

$$\mathrm{zer}(\partial f + \partial g) = \{\mathbf{prox}_{\gamma g}(z) \mid z \in \mathcal{H}, T_{\mathrm{PRS}} z = z\}. \tag{35}$$

That is, if z^* is a fixed point of T_{PRS}, then $x^* = x_g^* = x_f^*$ is a solution to Problem 1 and

$$z^* - x^* = \gamma \widetilde{\nabla} g(x^*) \in \gamma \partial g(x^*). \tag{36}$$

Proof See [2, Proposition 25.1] for the proof of Equation (35). Equation (36) follows because $x^* = \mathbf{prox}_{\gamma g}(z^*)$ if, and only if, $z^* - x^* \in \gamma \partial g(x^*)$.

4.3 Fundamental inequalities

We now deduce inequalities on the objective function $f + g$. In particular, we compute upper and lower bounds of the quantities $f(x_f^k) + g(x_g^k) - g(x^*) - f(x^*)$. Note that x_f^k and x_g^k are not necessarily equal, so this quantity can be negative.

The most important properties of the inequalities we establish below are:

1. The upper fundamental inequality has a telescoping structure in z^k and z^{k+1}.
2. They can be bounded in terms of $\|z^{k+1} - z^k\|^2$.

Properties 1 and 2 will be used to deduce ergodic and nonergodic rates, respectively.

Proposition 4 (Upper fundamental inequality) *Let $z \in \mathcal{H}$, let $z^+ := (T_{\mathrm{PRS}})_\lambda(z)$, and let x_f and x_g be defined as in Lemma 4. Then for all $x \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$*

$$4\gamma\lambda(f(x_f) + g(x_g) - f(x) - g(x)) \leq \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{1}{\lambda}\right)\|z^+ - z\|^2. \tag{37}$$

Proof We use the subgradient inequality and (34) multiple times in the following derivation:

$$\begin{aligned}
4\gamma\lambda(f(x_f) + g(x_g) - f(x) - g(x)) &\leq 4\lambda\gamma\left(\langle x_f - x, \widetilde{\nabla} f(x_f)\rangle + \langle x_g - x, \widetilde{\nabla} g(x_g)\rangle\right) \\
&= 4\lambda\gamma\left(\langle x_f - x_g, \widetilde{\nabla} f(x_f)\rangle + \langle x_g - x, \widetilde{\nabla} f(x_f) + \widetilde{\nabla} g(x_g)\rangle\right) \\
&= 2\left(\langle z^+ - z, \gamma\widetilde{\nabla} f(x_f)\rangle + \langle x - x_g, z^+ - z\rangle\right) \\
&= 2\langle z^+ - z, x + (z - x_g + \gamma\widetilde{\nabla} f(x_f)) - z\rangle \\
&= 2\langle z^+ - z, x + \gamma(\widetilde{\nabla} g(x_g) + \widetilde{\nabla} f(x_f)) - z\rangle \\
&= 2\langle z^+ - z, x - \frac{1}{2\lambda}(z^+ - z) - z\rangle \\
&= \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{1}{\lambda}\right)\|z^+ - z\|^2.
\end{aligned}$$

□

Proposition 5 (Lower fundamental inequality) *Let z^* be a fixed point of T_{PRS} and let $x^* := \mathbf{prox}_{\gamma g}(z^*)$. Then for all $x_f \in \mathrm{dom}(f)$ and $x_g \in \mathrm{dom}(g)$, the lower bound holds:*

$$f(x_f) + g(x_g) - f(x^*) - g(x^*) \geq \frac{1}{\gamma}\langle x_g - x_f, z^* - x^*\rangle. \tag{38}$$

Proof This proof essentially follows from the subgradient inequality. Indeed, let $\widetilde{\nabla} g(x^*) = (z^* - x^*)/\gamma \in \partial g(x^*)$ and let $\widetilde{\nabla} f(x^*) = -\widetilde{\nabla} g(x^*) \in \partial f(x^*)$. Then the result follows by adding the following equations:

$$\begin{aligned} f(x_f) - f(x^*) &\geq \langle x_f - x^*, \widetilde{\nabla} f(x^*)\rangle, \\ g(x_g) - g(x^*) &= \langle x_g - x_f, \widetilde{\nabla} g(x^*)\rangle + \langle x_f - x^*, \widetilde{\nabla} g(x^*)\rangle. \end{aligned}$$

□

5 Objective convergence rates

In this section we will prove ergodic and nonergodic convergence rates of relaxed PRS when f and g are closed, proper, and convex functions that are possibly nonsmooth.

To ease notational memory, we note that the reader may assume that $\lambda_k = (1/2)$ for all $k \geq 0$. This simplification implies that $\Lambda_k = (1/2)(k+1)$, and $\tau_k = \lambda_k(1-\lambda_k) = (1/4)$ for all $k \geq 0$.

Throughout this section the point z^* denotes an arbitrary fixed point of T_{PRS}, and we define a minimizer of $f + g$ by the formula (Lemma 5):

$$x^* = \mathbf{prox}_{\gamma g}(z^*).$$

The constant $(1/\gamma)\|z^* - x^*\|$ appears in the bounds of this section. This term is independent of γ: For any fixed point z^* of T_{PRS}, the point $x^* = \mathbf{prox}_{\gamma g}(z^*)$ is a minimizer and $z^* - \mathbf{prox}_{\gamma g}(z^*) = \gamma\widetilde{\nabla} g(x^*) \in \gamma\partial g(x^*)$. Conversely, if $x^* \in \mathrm{zer}(\partial f + \partial g)$ and $\widetilde{\nabla} g(x^*) \in (-\partial f(x^*)) \cap \partial g(x^*)$, then $z^* = x^* + \gamma\widetilde{\nabla} g(x^*)$ is a fixed point. Thus, we always assume that $(1/\gamma)\|z^* - x^*\| = \|\widetilde{\nabla} g(x^*)\|$ is minimal.

5.1 Ergodic convergence rates

In this section, we analyze the ergodic convergence of relaxed PRS. The proof follows the telescoping property of the upper and lower fundamental inequalities and an application of Jensen's inequality.

Theorem 6 (Ergodic convergence of relaxed PRS) *For all $k \geq 0$, let $\lambda_k \in (0, 1]$. Then we have the following convergence rate*

$$-\frac{2\|z^0 - z^*\|\|z^* - x^*\|}{\gamma\Lambda_k} \leq f(\overline{x}_f^k) + g(\overline{x}_g^k) - f(x^*) - g(x^*) \leq \frac{1}{4\gamma\Lambda_k}\|z^0 - x^*\|^2.$$

In addition, the following feasibility bound holds:

$$\|\overline{x}_g^k - \overline{x}_f^k\| \leq \frac{2\|z^0 - z^*\|}{\Lambda_k}. \tag{39}$$

Proof Equation (39) follows directly from Theorem 5.

Recall the upper fundamental inequality from Proposition 4 :

$$4\gamma\lambda_k(f(x_f^k)+g(x_g^k)-f(x^*)-g(x^*)) \leq \|z^k-x^*\|^2-\|z^{k+1}-x^*\|^2+\left(1-\frac{1}{\lambda_k}\right)\|z^{k+1}-z^k\|^2. \tag{40}$$

Because $\lambda_k \leq 1$, it follows that $(1-(1/\lambda_k)) \leq 0$. Thus, we sum Equation (40) from $i=0$ to k, divide by Λ_k, and apply Jensen's inequality to get

$$\begin{aligned}\frac{1}{4\gamma\Lambda_k}(\|z^0-x^*\|^2-\|z^{k+1}-x^*\|^2) &\geq \frac{1}{\Lambda_k}\sum_{i=0}^{k}\lambda_i(f(x_f^i)+g(x_g^i)-f(x^*)-g(x^*)) \\ &\geq f(\overline{x}_f^k)+g(\overline{x}_g^k)-f(x^*)-g(x^*).\end{aligned}$$

The lower bound is a consequence of the fundamental lower inequality and Equation (39)

$$f(\overline{x}_f^k)+g(\overline{x}_g^k)-f(x^*)-g(x^*) \overset{(38)}{=} \frac{1}{\gamma}\langle \overline{x}_g^k-\overline{x}_f^k, z^*-x^*\rangle \overset{(39)}{\geq} -\frac{2\|z^0-z^*\|\|z^*-x^*\|}{\gamma\Lambda_k}. \tag{41}$$

□

In general, $x_f^k \notin \mathrm{dom}(g)$ and $x_g^k \notin \mathrm{dom}(f)$, so we cannot evaluate g at x_f^k or f at x_g^k. However, the conclusion of Theorem 6 can be improved if f or g is Lipschitz continuous. The following proposition gives a sufficient condition for Lipschitz continuity on a ball:

Proposition 6 (Lipschitz continuity on a ball) *Suppose that $f : \mathcal{H} \to (-\infty,\infty]$ is proper and convex. Let $\rho > 0$ and let $x_0 \in \mathcal{H}$. If $\delta = \sup_{x,y\in B(x_0,2\rho)}|f(x)-f(y)| < \infty$, then f is (δ/ρ)-Lipschitz on $B(x_0,\rho)$.*

Proof See [2, Proposition 8.28]. □

To use this fact, we need to show that the sequences $(x_f^j)_{j\geq 0}$, and $(x_g^j)_{j\geq 0}$ are bounded. Recall that $x_g^s = \mathbf{prox}_{\gamma g}(z^s)$ and $x_f^s = \mathbf{prox}_{\gamma f}(\mathbf{refl}_{\gamma g}(z^s))$, for $s \in \{*,k\}$. Proximal and reflection maps are nonexpansive, so we have the following simple bound:

$$\max\{\|x_f^k-x^*\|, \|x_g^k-x^*\|\} \leq \|z^k-z^*\| \leq \|z^0-z^*\|.$$

Thus, $(x_f^j)_{j\geq 0}, (x_g^j)_{j\geq 0} \subseteq \overline{B(x^*, \|z^0-z^*\|)}$.

Corollary 4 (Ergodic convergence with single Lipschitz function) *Let the notation be as in Theorem 6. Suppose that f (respectively g) is L-Lipschitz continuous on $\overline{B(x^*,\|z^0-z^*\|)}$, and let $x^k = x_g^k$ (respectively $x^k = x_f^k$). Then the following convergence rate holds*

$$0 \leq f(\overline{x}^k)+g(\overline{x}^k)-f(x^*)-g(x^*) \leq \frac{1}{4\gamma\Lambda_k}\|z^0-x^*\|^2+\frac{2L\|z^0-z^*\|}{\Lambda_k}.$$

Proof From Equation (39), we have $\|\overline{x}_g^k-\overline{x}_f^k\| \leq (2/\Lambda_k)\|z^0-z^*\|$. In addition, $(x_f^j)_{j\geq 0}, (x_g^j)_{j\geq 0} \subseteq \overline{B(x^*,\|z^0-z^*\|)}$. Thus, it follows that

$$\begin{aligned}0 \leq f(\overline{x}^k)+g(\overline{x}^k)-f(x^*)-g(x^*) &\leq f(\overline{x}_f^k)+g(\overline{x}_g^k)-f(x^*)-g(x^*)+L\|\overline{x}_f^k-\overline{x}_g^k\| \\ &= f(\overline{x}_f^k)+g(\overline{x}_g^k)-f(x^*)-g(x^*)+\frac{2L\|z^0-z^*\|}{\Lambda_k}.\end{aligned}$$

The upper bound follows from this equation and Theorem 6. □

5.2 Nonergodic convergence rates

In this section, we prove the nonergodic convergence rate of the Algorithm 1 whenever $\underline{\tau} := \inf_{j\geq 0} \tau_j > 0$. The proof uses Theorem 1 to bound the fundamental inequalities in Propositions 4 and 5.

Theorem 7 (Nonergodic convergence of relaxed PRS) *For all $k \geq 0$, let $\lambda_k \in (0,1)$. Suppose that $\underline{\tau} := \inf_{j\geq 0} \lambda_k(1-\lambda_k) > 0$. Then we have the convergence rates:*

1. *In general, we have the bounds:*

$$-\frac{\|z^0 - z^*\|\|z^* - x^*\|}{2\gamma\sqrt{\underline{\tau}(k+1)}} \leq f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) \leq \frac{(\|z^0 - z^*\| + \|z^* - x^*\|)\|z^0 - z^*\|}{2\gamma\sqrt{\underline{\tau}(k+1)}}$$

   and $|f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*)| = o\left(1/\sqrt{k+1}\right)$.
2. *If $\mathcal{H} = \mathbf{R}$ and $\lambda_k \equiv 1/2$, then for all $k \geq 0$,*

$$\frac{\|z^0 - z^*\|\|z^* - x^*\|}{\sqrt{2}\gamma(k+1)} \leq f(x_f^{k+1}) + g(x_g^{k+1}) - f(x^*) - g(x^*) \leq \frac{(\|z^0 - z^*\| + \|z^* - x^*\|)\|z^0 - z^*\|}{\sqrt{2}\gamma(k+1)}$$

   and $|f(x_f^{k+1}) + g(x_g^{k+1}) - f(x^*) - g(x^*)| = o\left(1/(k+1)\right)$.

Proof We prove Part 1 first. For all $\lambda \in [0,1]$, let $z_\lambda = (T_{\mathrm{PRS}})_\lambda(z^k)$. Evaluate the upper inequality in Equation (37) at $x = x^*$ to get

$$4\gamma\lambda(f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*)) \leq \|z^k - x^*\|^2 - \|z_\lambda - x^*\|^2 + \left(1 - \frac{1}{\lambda}\right)\|z_\lambda - z^k\|^2.$$

Recall the following identity:

$$\|z^k - x^*\|^2 - \|z_\lambda - x^*\|^2 - \|z_\lambda - z^k\|^2 = 2\langle z_\lambda - x^*, z^k - z_\lambda\rangle.$$

By the triangle inequality, because $\|z_\lambda - z^*\| \leq \|z^k - z^*\|$, and because $(\|z^j - z^*\|)_{j\geq 0}$ is monotonically nonincreasing (Corollary 2), it follows that

$$\|z_\lambda - x^*\| \leq \|z_\lambda - z^*\| + \|z^* - x^*\| \leq \|z^0 - z^*\| + \|z^* - x^*\|. \tag{42}$$

Thus, we have the bound:

$$\begin{aligned} f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) & \\ &\leq \inf_{\lambda\in[0,1]} \frac{1}{4\gamma\lambda}\left(2\langle z_\lambda - x^*, z^k - z_\lambda\rangle + 2\left(1 - \frac{1}{2\lambda}\right)\|z_\lambda - z^k\|^2\right) \\ &\leq \frac{1}{\gamma}\|z_{1/2} - x^*\|\|z^k - z_{1/2}\| \\ &\overset{(42)}{\leq} \frac{1}{\gamma}\left(\|z^0 - z^*\| + \|z^* - x^*\|\right)\|z^k - z_{1/2}\| \\ &\overset{(26)}{\leq} \frac{(\|z^0 - z^*\| + \|z^* - x^*\|)\|z^0 - z^*\|}{2\gamma\sqrt{\underline{\tau}(k+1)}}. \end{aligned} \tag{43}$$

The lower bound follows from the identity $x_g^k - x_f^k = (1/2\lambda_k)(z^k - z^{k+1})$ and the fundamental lower inequality in Equation (38):

$$f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) \geq \frac{1}{2\gamma\lambda_k}\langle z^k - z^{k+1}, z^* - x^*\rangle \geq -\frac{\|z^{k+1} - z^k\|\|z^* - x^*\|}{2\gamma\lambda_k} \tag{44}$$
$$\overset{(26)}{\geq} -\frac{\|z^0 - z^*\|\|z^* - x^*\|}{2\gamma\sqrt{\underline{\tau}(k+1)}}.$$

Finally, the $o(1/\sqrt{k+1})$ convergence rate follows from Equations (43) and (44) combined with Corollary 2.

Part 2 follows by the same analysis but uses Theorem 3 to estimate the FPR convergence rate. □

Whenever f or g is Lipschitz, we can compute the convergence rate of $f + g$ evaluated at the same point. The following theorem is analogous to Corollary 4 in the ergodic case. The proof essentially follows by combining the nonergodic convergence rate in Theorem 7 with the convergence rate of $\|x_f^k - x_g^k\| = (1/\lambda_k)\|z^{k+1} - z^k\|$ deduced in Corollary 2.

Corollary 5 (Nonergodic convergence with Lipschitz assumption) *Let the notation be as in Theorem 7. Suppose that f (respectively g) is L-Lipschitz continuous on $\overline{B(x^*, \|z^0 - z^*\|)}$, and let $x^k = x_g^k$ (respectively $x^k = x_f^k$). Then we have the convergence rates of the nonnegative term:*

1. *In general, we have the bounds:*

$$0 \leq f(x^k) + g(x^k) - f(x^*) - g(x^*) \leq \frac{\left(\|z^0 - z^*\| + \|z^* - x^*\| + \gamma L\right)\|z^0 - z^*\|}{2\gamma\sqrt{\underline{\tau}(k+1)}}$$

   and $f(x^k) + g(x^k) - f(x^*) - g(x^*) = o\left(1/\sqrt{k+1}\right)$:
2. *If $\mathcal{H} = \mathbf{R}$ and $\lambda_k \equiv 1/2$, then for all $k \geq 0$,*

$$0 \leq f(x^{k+1}) + g(x^{k+1}) - f(x^*) - g(x^*) \leq \frac{\left(\|z^0 - z^*\| + \|z^* - x^*\| + \gamma L\right)\|z^0 - z^*\|}{\sqrt{2}\gamma(k+1)}$$

   and $f(x^{k+1}) + g(x^{k+1}) - f(x^*) - g(x^*) = o\left(1/(k+1)\right)$.

Proof We prove Part 1 first. First recall that $\|x_g^k - x_f^k\| = (1/(2\lambda_k))\|z^{k+1} - z^k\|$. In addition, $(x_f^j)_{j\geq 0}, (x_g^j)_{j\geq 0} \subseteq \overline{B(x^*, \|z^0 - z^*\|)}$ (See Section (5.1)). Thus, it follows that

$$f(x^k) + g(x^k) - f(x^*) - g(x^*) \leq f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) + L\|x_f^k - x_g^k\|$$
$$= f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) + \frac{L\|z^{k+1} - z^k\|}{2\lambda_k} \tag{45}$$
$$\overset{(26)}{\leq} f(x_f^k) + g(x_g^k) - f(x^*) - g(x^*) + \frac{\gamma L\|z^0 - z^*\|}{2\gamma\sqrt{\underline{\tau}(k+1)}}. \tag{46}$$

Therefore, the upper bound follows from Theorem 7 and Equation (46). In addition, the $o(1/\sqrt{k+1})$ bound follows from Theorem 7 combined with Equation (45) and Corollary 2.

Part 2 follows by the same analysis, but uses Theorem 3 to estimate the FPR convergence rate. □

6 Optimal FPR rate and arbitrarily slow convergence

In this section, we provide two examples where the DRS algorithm converges slowly. Both examples are a special cases of the following example, which originally appeared in [1, Section 7].

Example 1 (DRS applied to two subspaces) Let $\mathcal{H} = \ell_2^2(\mathbf{N}) = \mathbf{R}^2 \times \mathbf{R}^2 \times \cdots$. Let R_θ denote counterclockwise rotation in $\mathbf{R}^2$ by θ degrees. Let $e_0 := (1, 0)$ denote the standard unit vector, and let $e_\theta := R_\theta e_0$. Suppose that $(\theta_j)_{j\geq 0}$ is a sequence of angles in $(0, \pi/2]$ such that $\theta_i \to 0$ as $i \to \infty$. For all $i \geq 0$, let $c_i := \cos(\theta_i)$. We let

$$U := \mathbf{R}^2 e_0 \times \mathbf{R}^2 e_0 \times \cdots \qquad \text{and} \qquad V := \mathbf{R}^2 e_{\theta_0} \times \mathbf{R}^2 e_{\theta_1} \times \cdots . \tag{47}$$

See Figure 2 for an illustration.

Note that [1, Section 7] shows the projection identities

$$(P_V)_i = \begin{bmatrix} \cos^2(\theta_i) & \sin(\theta_i)\cos(\theta_i) \\ \sin(\theta_i)\cos(\theta_i) & \sin^2(\theta_i) \end{bmatrix} \qquad \text{and} \qquad (P_U)_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

the DRS operator identity

$$T := (T_{\mathrm{PRS}})_{1/2} = c_0 R_{\theta_0} \oplus c_1 R_{\theta_1} \oplus \cdots , \tag{48}$$

and that $(z^j)_{j\geq 0}$ converges in norm to $z^* = 0$ for any initial point z^0. □

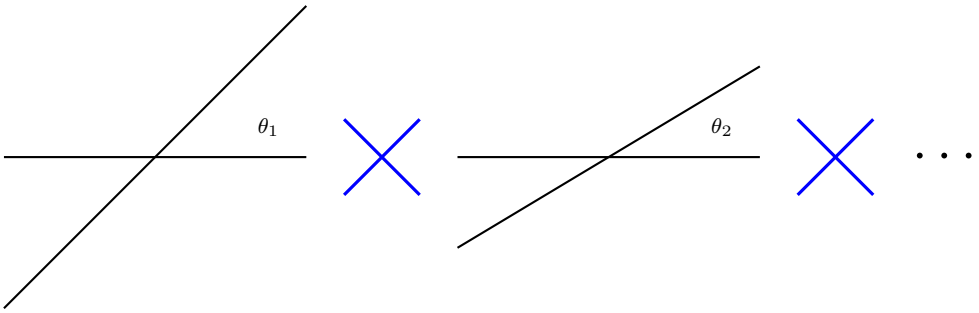

Fig. 2 Illustration of Example 1. Each pair of lines represents a 2-dimensional component of $U \cup V$. The angles θ_k are converging to 0.

6.1 Optimal FPR rates

The following theorem shows that the FPR estimates derived in Corollary 2 are essentially optimal. We note that this is the first optimality result for the FPR of the DRS iteration in the case of variational problems.

Theorem 8 (Lower FPR complexity of DRS) *There exists a Hilbert space $\mathcal{H}$ and two closed subspaces U and V with zero intersection, $U \cap V = \{0\}$, such that for every $\alpha > 1/2$, there exists $z^0 \in \mathcal{H}$ such that if $(z^j)_{j\geq 0}$ is generated by $T = (T_{\mathrm{PRS}})_{1/2}$ applied to $f = \chi_V$ and $g = \chi_U$, then for all $k \geq 1$, we have the bound:*

$$\|Tz^k - z^k\|^2 \geq \frac{1}{(k+1)^{2\alpha}}.$$

Proof We assume the setting of Example 1. For all $i \geq 0$ set $c_i = (i/(i+1))^{1/2}$, and let $w^0 = (w_j^0)_{j\geq 0} \in \mathcal{H}$, where each $w_i^0 \in \mathbf{R}^2$ satisfies $\|w_i^0\| = \sqrt{2\alpha e}(i+1)^{-(1+2\alpha)/2}$. Then for all $k \geq 1$,

$$\|T^k w^0\|^2 = \sum_{i=0}^{\infty} c_i^{2k}\|w_i^0\|^2 \geq \sum_{i=k}^{\infty}\left(\frac{i}{i+1}\right)^k \frac{2\alpha e}{(i+1)^{1+2\alpha}} \geq \frac{1}{(k+1)^{2\alpha}}$$

where we have used the inequality $(i/(i+1))^k \geq e^{-1}$, for $i \geq k$ and the lower integral approximation of the sum.

Now we will show that $\sqrt{2\alpha e}w^0$ is in the range of $I - T$. Indeed, for all $i \geq 1$ each block of $I - T$ is of the form

$$I_{\mathbf{R}^2} - \cos(\theta_i)R_{\theta_i} = \begin{bmatrix} \sin^2(\theta_i) & \sin(\theta_i)\cos(\theta_i) \\ -\sin(\theta_i)\cos(\theta_i) & \sin^2(\theta_i) \end{bmatrix} = \begin{bmatrix} \frac{1}{i+1} & \frac{\sqrt{i}}{i+1} \\ -\frac{\sqrt{i}}{i+1} & \frac{1}{i+1} \end{bmatrix}. \tag{49}$$

Therefore, the point $z^0 = (\sqrt{2\alpha e}((1/(j+1)^\alpha, 0))_{j\geq 0} \in \mathcal{H}$ has image

$$w^0 = (I - T)z^0 = \left(\sqrt{2\alpha e}\left(\frac{1}{(j+1)^{\alpha+1}}, \frac{-\sqrt{j}}{(j+1)^{\alpha+1}}\right)\right)_{j\geq 0}.$$

In addition, for all $i \geq 1$, we have $\|w_i^0\| = \sqrt{2\alpha e}(i+1)^{-(1+2\alpha)/2}$, and the inequality follows.

□

Remark 1 The proof of Theorem 8 crucially relies on the strictness of inequality, $\alpha > 1/2$: if $\alpha = 1/2$, then $\|z^0\| = \infty$.

6.1.1 Notes on Theorem 8

With this new optimality result in hand, we can make the following list of optimal FPR rates, not to be confused with optimal rates in objective error, for a few standard splitting schemes:

Proximal point algorithm (PPA): For the general class of monotone operators, the counterexample furnished in [9, Remarque 4] shows that there exists a maximal monotone operator A such that when iteration (19) is applied to the resolvent $J_{\gamma A}$, the rate $o(1/(k+1))$ is tight. In addition, if $A = \partial f$ for some closed, proper, and convex function f, then the FPR rate satisfies improves to $O(1/(k+1)^2)$ [9, Théorème 9]. We improve this result to $o(1/(k+1)^2)$ in Lemma 3. This result appears to be new and is optimal by [9, Remarque 6].

Forward backward splitting (FBS): The FBS method reduces to the proximal point algorithm when the differentiable (or single valued operator) term is trivial. Thus, for the general class of monotone operators, the $o(1/(k+1))$ FPR rate is optimal by [9, Remarque 4]. We improve this rate to $o(1/(k+1)^2)$ in Lemma 3. This result appears to be new, and is optimal by [9, Remarque 6].

Douglas-Rachford splitting/ADMM: Theorem 8 shows that the optimal FPR rate is $o(1/(k+1))$. Because the DRS iteration is equivalent to a proximal point iteration applied to a special monotone operator [20, Section 4], Theorem 8 provides an alternative counterexample to [9, Remarque 4]. In particular, Theorem 8 shows that, in general, there is no closed, proper, and convex function f such that $(T_{\mathrm{PRS}})_{1/2} = \mathbf{prox}_{\gamma f}$. In the one dimensional case, we improve the FPR to $o(1/(k+1)^2)$ in Lemma 3.

Miscellaneous methods: By similar arguments we can deduce iteration complexity for the following methods, each of which at least has rate $o(1/(k+1))$ by Theorem 1: *Standard Gradient descent* $o(1/(k+1)^2)$: (the rate follows from Theorem 3. Optimality follows from the fact that PPA $\equiv$ to gradient descent on Moreau envelope [2, Proposition 12.29] and [9, Remarque 4]); *Forward-Douglas Rachford splitting* [10]: $o(1/(k+1))$ (choose a trivial cocoercive operator and use Theorem 8); *Chambolle and Pock's primal-dual algorithm* [12] $o(1/(k+1))$: (reduce to DRS ($\sigma = \tau = 1$) [12, Section 4.2] and apply Theorem 8 using the transformation $z^k = \mathrm{primal}_k + \mathrm{dual}_k$ [12, Equation (24)] and the lower bound

$$\|z^{k+1} - z^k\|^2 \leq 2\|\mathrm{primal}_{k+1} - \mathrm{primal}_k\|^2 + 2\|\mathrm{dual}_{k+1} - \mathrm{dual}_k\|^2;$$

Vũ/Condat's primal-dual algorithm [37,18] $o(1/(k+1))$: (reduces to Chambolle and Pock's method [12]).

Note that the rate established in Theorem 1 has broad applicability, and this list is hardly extensive. For PPA, FBS, and standard gradient descent, the FPR always has rate that is the square of the objective value convergence rate. We will see that the same is true for DRS in Theorem 11.

6.2 Arbitrarily slow convergence

In [1, Section 7], the DRS setting in Example 1 is shown to converge in norm, but not linearly. We improve their result by showing that a proper choice of parameters yields arbitrarily slow convergence in norm.

The following technical lemma will help us construct a sequence that convergenes arbitrarily slowly. The idea of the proof follows directly from the proof of [21, Theorem 4.2], which shows that the alternating projection algorithm can converge arbitrarily slowly.

Lemma 6 *Suppose that $h : \mathbf{R}_+ \to (0,1)$ is a function that is monotonically decreasing to zero. Then there exists a monotonic sequence $(c_j)_{j\geq 0} \subseteq (0,1)$ such that $c_k \to 1^-$, as $k \to \infty$ and an increasing sequence of integers $(n_j)_{j\geq 0} \subseteq \mathbf{N} \cup \{0\}$ such that for all $k \geq 0$,*

$$\frac{c_{n_k}^{k+1}}{n_k + 1} > h(k+1)e^{-1}. \tag{50}$$

Proof Let h_2 be the inverse of the strictly increasing function $(1/h) - 1$, let $[x]$ denote the integer part of x, and for all $k \geq 0$ let

$$c_k = \frac{h_2(k+1)}{1 + h_2(k+1)}. \tag{51}$$

Then $(c_j)_{j\geq 0}$ is monotonic and $c_k \to 1^-$. For all $x \geq 0$, we have $h_2^{-1}(x) = 1/h(x) - 1 \leq [1/h(x)]$, thus, $x \leq h_2([1/h(x)])$. Now, choose $n_k \geq 0$ such that $n_k + 1 = [1/h(k+1)]$. Therefore,

$$\frac{c_{n_k}^{k+1}}{n_k + 1} \geq h(k+1)\left(\frac{k+1}{1+(k+1)}\right)^{k+1} \geq h(k+1)e^{-1}.$$

□

Theorem 9 (Arbitrarily slow convergence of DRS) *There is a point $z_0 \in \ell_2^2(\mathbf{N})$, such that for every function $h : \mathbf{R}_+ \to (0,1)$ that strictly decreases to zero, there exists two closed subspaces U and V with zero intersection, $U \cap V = \{0\}$, such that the relaxed PRS sequence $(z^j)_{j\geq 0}$ generated with the functions $f = \chi_V$ and $g = \chi_U$ and relaxation parameters $\lambda_k \equiv 1/2$ satisfies the bound*

$$\|z^k - z^*\| \geq e^{-1}h(k)$$

but $(\|z^j - z^*\|)_{j\geq 0}$ converges to 0.

Proof We assume the setting of Example 1. Suppose that $z^0 = (z_j^0)_{j\geq 0}$, where for all $k \geq 0$, $z_k^0 \in \mathbf{R}^2$, and $\|z_k^0\| = 1/(k+1)$. Then it follows that $\|z^0\|_{\mathcal{H}}^2 = \sum_{i=0}^{\infty} 1/(k+1)^2 < \infty$ and so $z^0 \in \mathcal{H}$. Thus, for all $k, n \geq 0$

$$\|T^{k+1}x\| \geq c_n^{k+1}\|z_n^0\| = \frac{1}{n+1}c_n^{k+1}. \tag{52}$$

Therefore, we can achieve arbitrarily slow convergence by picking $(c_j)_{j\geq 0}$, and a subsequence $(n_j)_{j\geq 0} \subseteq \mathbf{N}$ using Lemma 6. □

7 Optimal objective rates

In this section we construct four examples that show the nonergodic and ergodic convergence rates in Corollary 5 and Theorem 6 are optimal up to constant factors. In particular, we provide examples of optimal ergodic convergence in the minimization case and in the feasibility case, where no objective is driving the minimization.

7.1 Ergodic convergence of feasibility problems

Proposition 7 *The ergodic feasibility convergence rate in Equation* (39) *is optimal up to a factor of two.*

Proof Figure 3 shows Algorithm 1 with $\lambda_k = 1$ for all $k \geq 0$ (i.e. PRS) applied to the functions $f = \chi_{\{(x_1,x_2)\in\mathbf{R}^2 | x_1=0\}}$ and $g = \chi_{\{(x_1,x_2)\in\mathbf{R}^2|x_2=0\}}$ with the initial iterate $z^0 = (1,1) \in \mathbf{R}^2$. Because $T_{\mathrm{PRS}} = -I_{\mathcal{H}}$, it is easy to see that the only fixed point of T_{PRS} is $z^* = (0,0)$. In addition, the following identities are satisfied:

$$x_g^k = \begin{cases} (1,0) & \text{even } k; \\ (-1,0) & \text{odd } k. \end{cases} \qquad z^k = \begin{cases} (1,1) & \text{even } k; \\ (-1,-1) & \text{odd } k. \end{cases} \qquad x_f^k = \begin{cases} (0,-1) & \text{even } k; \\ (0,1) & \text{odd } k. \end{cases}$$

Thus, the PRS algorithm oscillates around the solution $x^* = (0,0)$. However, note that the averaged iterates satisfy:

$$\overline{x}_g^k = \begin{cases} (\frac{1}{k+1}, 0) & \text{even } k; \\ (0,0) & \text{odd } k. \end{cases} \qquad \text{and} \qquad \overline{x}_f^k = \begin{cases} (0, \frac{-1}{k+1}) & \text{even } k; \\ (0,0) & \text{odd } k. \end{cases}$$

It follows that $\|\overline{x}_g^k - \overline{x}_f^k\| = (1/(k+1))\|(1,-1)\| = (1/(k+1))\|z^0 - z^*\|$, for all $k \geq 0$.

□

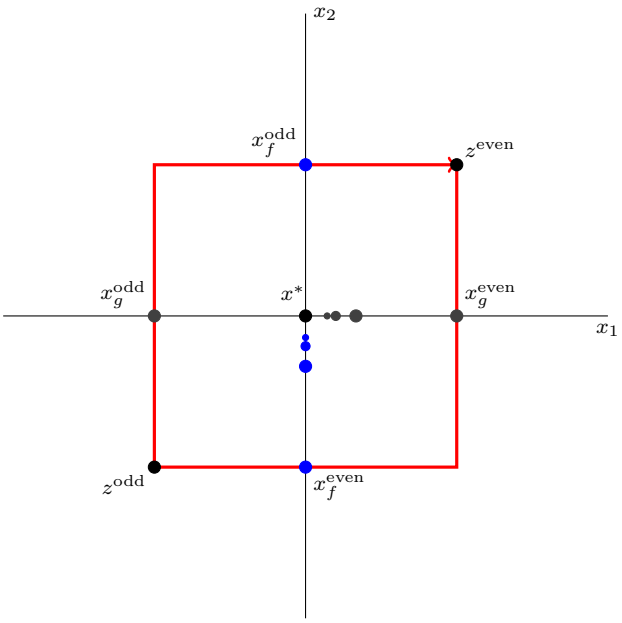

Fig. 3 Example 7.1 of PRS. z^k hops between $(1,1)$ and $(-1,-1)$ while the ergodic iterates $\overline{x}_g^k$ and $\overline{x}_f^k$ (dots of decreasing size) approach x^*.

7.2 Ergodic convergence of minimization problems

In this section, we will construct an example where the ergodic rates of convergence in Section 5.1 are optimal up to constant factors. In addition, the example we construct only converges in the ergodic sense and diverges otherwise. Throughout this section, we let $\gamma = 1$ and $\lambda_k \equiv 1$, we work in the Hilbert space $\mathcal{H} = \mathbf{R}$, and we consider the following objective functions: for all $x \in \mathbf{R}$, let

$$g(x) = 0, \qquad \text{and} \qquad f(x) = |x|. \tag{53}$$

Recall that for all $x \in \mathbf{R}$

$$\mathbf{prox}_g(x) = x, \qquad \text{and} \qquad \mathbf{prox}_f(x) = \max\left(|x| - 1, 0\right)\operatorname{sign}(x). \tag{54}$$

The following lemma characterizes the minimizer of $f + g$ and the fixed points of T_{PRS}. The proof is simple so we omit it.

Lemma 7 *The minimizer of $f + g$ is unique and equal to $0 \in \mathbf{R}$. Furthermore, 0 is the unique fixed point of T_{PRS}.*

Because of Lemma 7, we will use the notation:

$$z^* = 0 \qquad \text{and} \qquad x^* = 0. \tag{55}$$

We are ready to prove our main optimality result.

Proposition 8 (Optimality of ergodic convergence rates) *Suppose that $z^0 = 2-\varepsilon$ for some $\varepsilon \in (0,1)$. Then the PRS algorithm applied to f and g with initial point z^0 does not converge.*

Furthermore, as ε goes to 0, the ergodic objective convergence rate in Theorem 6 is tight, and the ergodic objective convergence rate in Corollary 4 is tight up to a factor of 5/2. In addition, the feasibility convergence rate of Theorem 6 is tight up to a factor of 4.

Proof We will now compute the sequences $(z^j)_{j\geq 0}$, $(x_g^j)_{j\geq 0}$, and $(x_f^j)_{j\geq 0}$. We proceed by induction: First $x_g^0 = \mathbf{prox}_{\gamma g}(z^0) = z^0$ and $x_f^0 = \mathbf{prox}_{\gamma f}\left(2x_g^0 - z^0\right) = \max\left(|z^0|-1, 0\right)\operatorname{sign}(z^0) = 1-\varepsilon$. Thus, it follows that $z^1 = z^0 + 2(x_f^0 - x_g^0) = 2-\varepsilon+2(1-\varepsilon-(2-\varepsilon)) = z^0 = -\varepsilon$. Similarly, $x_g^1 = z^1 = -\varepsilon$. Finally, $x_f^1 = \max(\varepsilon-1,0)(-\varepsilon) = 0$ and $z^2 = z^1 + 2(x_f^1 - x_g^1) = z^1 + 2(\varepsilon) = \varepsilon$. Thus, by induction we have the following identities: For all $k \geq 1$,

$$z^k = (-1)^k\varepsilon, \qquad x_g^k = (-1)^k\varepsilon, \qquad x_f^k = 0. \tag{56}$$

Notice that that $(z^j)_{j\geq 0}$ and $(x_g^j)_{j\geq 0}$ do not converge, but they oscillate around the fixed point of T_{PRS}.

We will now compute the ergodic iterates:

$$\overline{x}_g^k = \frac{1}{k+1}\sum_{i=0}^{k} x_g^i \overset{(56)}{=} \begin{cases} \frac{2-\varepsilon}{k+1} & \text{if } k \text{ is even;} \\ \frac{2-2\varepsilon}{k+1} & \text{otherwise.} \end{cases} \qquad \text{and} \qquad \overline{x}_f^k = \frac{1}{k+1}\sum_{i=0}^{k} x_f^i \overset{(56)}{=} \frac{1-\varepsilon}{k+1}. \tag{57}$$

Let us use these formulas to compute the objective values:

$$f(\overline{x}_f^k) + g(\overline{x}_f^k) - f(0) - g(0) \overset{(57)}{=} \frac{1-\varepsilon}{k+1} \quad \text{and} \quad f(\overline{x}_g^k) + g(\overline{x}_g^k) - f(0) - g(0) \overset{(57)}{=} \begin{cases} \frac{2-\varepsilon}{k+1} & \text{if } k \text{ is even;} \\ \frac{2-2\varepsilon}{k+1} & \text{otherwise.} \end{cases} \tag{58}$$

We will now compare the theoretical bounds from Theorem 6 and Corollary 4 with the rates we observed in Equation (58). Theorem 6 bounds the objective error at $\overline{x}_f^k$ by

$$\frac{|z^0 - x^*|^2}{4(k+1)} = \frac{4-4\varepsilon}{4(k+1)} + \frac{\varepsilon^2}{4(k+1)} = \frac{1-\varepsilon}{k+1} + \frac{\varepsilon^2}{4(k+1)}. \tag{59}$$

By taking ε to 0, we see that this bound is tight.

Because f is 1-Lipschitz continuous, Corollary 4 bounds the objective error at $\overline{x}_g^k$ with

$$\frac{|z^0 - x^*|^2}{4(k+1)} + \frac{2|z^0 - z^*|}{(k+1)} \overset{(59)}{=} \frac{1-\varepsilon}{k+1} + \frac{\varepsilon^2}{4(k+1)} + 2\frac{2-\varepsilon}{k+1} = \frac{5-3\varepsilon}{k+1} + \frac{\varepsilon^2}{4(k+1)}. \tag{60}$$

As we take ε to 0, we see that this bound it tight up to a factor of 5/2.

Finally, consider the feasibility convergence rate:

$$|\overline{x}_g^k - \overline{x}_f^k| \overset{(56)}{=} \begin{cases} \frac{1}{k+1} & \text{if } k \text{ is even;} \\ \frac{1-\varepsilon}{k+1} & \text{otherwise.} \end{cases}. \tag{61}$$

Theorem 6 predicts the following upper bound for Equation (61):

$$\frac{2|z^0 - z^*|}{k+1} = 2\frac{2-\varepsilon}{k+1} = \frac{4-2\varepsilon}{k+1}. \tag{62}$$

By taking ε to 0, we see that this bound is tight up to a factor of 4. □

7.3 Optimal nonergodic objective rates

Our aim in this section is to show that if $\lambda_k \equiv 1/2$, then the non-ergodic convergence rate of $o(1/\sqrt{k+1})$ in Corollary 5 is essentially tight. In particular, for every $\alpha > 1/2$, we provide examples of f and g such that f is 1-Lipschitz and

$$f(x_g^k) + g(x_g^k) - f(x^*) - g(x^*) = \Omega\left(\frac{1}{(k+1)^\alpha}\right).$$

Throughout this section, we will be working with the proximal operator of a distance functions.

Proposition 9 *Let C be a closed, convex subset of $\mathcal{H}$ and let $d_C(x) = \min_{y\in C}\|x-y\|$. Then $d_C(x)$ is 1-Lipschitz and for all $x \in \mathcal{H}$*

$$\mathbf{prox}_{\gamma d_C}(x) = \theta P_C(x) + (1-\theta)x \qquad \text{where} \qquad \theta = \begin{cases} \frac{\gamma}{d_C(x)} & \text{if } \gamma \le d_C(x); \\ 1 & \text{otherwise.} \end{cases} \tag{63}$$

Proof Follows directly from the formula for the subgradient of d_C [2, Example 16.49]. □

Proposition 9 says that $\mathbf{prox}_{\gamma d_C}(x)$ reduces to a projection map whenever x is close enough to C. Proposition 10 constructs a family of examples such that if γ is chosen large enough, then DRS does not distinguish between characteristic functions and distance functions.

Proposition 10 *Suppose that V and U are linear subspaces of $\mathcal{H}$ and $U \cap V = \{0\}$. If $\gamma > \|z^0\|$ and $\lambda_k = 1/2$ for all $k \ge 0$, then Algorithm 1 applied to the either pair of objective functions $(f = \chi_V, g = \chi_U)$ and $(f = d_V, g = \chi_U)$ produces the same sequence $(z^j)_{j\ge 0}$*

Proof Let $(z^j)_{j\ge 0}$ be the sequence generated by the functions $(f = \chi_V, g = \chi_U)$. Observe that $x^* = 0$ is a minimizer of both functions pairs and $z^* = 0$ is a fixed point of $(T_{\mathrm{PRS}})_{1/2}$. In particular, we set $\widetilde{\nabla}\chi_V(x^*) = P_V(\mathbf{refl}_g(z^*)) - x^* = 0$. Therefore, we just need to show that $\mathbf{prox}_{\gamma d_V}(\mathbf{refl}_g(z^k)) = P_V(\mathbf{refl}_g(z^k))$ for all $k \ge 0$. Note that by definition, $x_{\chi V}^k = P_V(\mathbf{refl}_g(z^k))$ and $\widetilde{\nabla}\chi_V(x_{\chi V}^k) = \mathbf{refl}_g(z^k) - P_V(\mathbf{refl}_g(z^k)) \in \partial\chi_V(x_{\chi V}^k)$. In view of Proposition 9, the identity will follow if

$$\gamma \ge d_V(\mathbf{refl}_g(z^k)) = \|\mathbf{refl}_g(z^k) - P_V(\mathbf{refl}_g(z^k))\| = \|\widetilde{\nabla}\chi_V(x_{\chi V}^k)\| = \|\widetilde{\nabla}\chi_V(x_{\chi V}^k) - \widetilde{\nabla}\chi_V(x^*)\|.$$

However, this is always the case because

$$\|\widetilde{\nabla}\chi_V(x_{\chi V}^k) - \widetilde{\nabla}\chi_V(x^*)\|^2 + \|x_{\chi V}^k - x^*\|^2 \overset{(16)}{\le} \|\mathbf{refl}_g(z^k) - \mathbf{refl}_g(z^*)\|^2 \le \|z^k - z^*\|^2 \le \|z^0 - z^*\|^2 = \|z^0\|^2 \le \gamma^2.$$

□

Theorem 10 *Assume the notation of Theorem 8. Then for all $\alpha > 1/2$, there exists a point $z^0 \in \mathcal{H}$ such that if $\gamma \ge \|z^0\|$ and $(z^j)_{j\ge 0}$ is generated by DRS applied to the functions $(f = d_V, g = \chi_U)$, then $d_V(x^*) = 0$ and*

$$d_V(x_g^k) = \Omega\left(\frac{1}{(k+1)^\alpha}\right). \tag{64}$$

Proof Let $z^0 = ((1/(j+1)^\alpha, 0))_{j\geq 0} \in \mathcal{H}$. Now, choose $\gamma^2 \geq \|z^0\|^2 = \sum_{i=0}^{\infty} 1/(i+1)^{2\alpha}$. Define $w^0 \in \mathcal{H}$ using Equation (49):

$$w^0 = (I - T)z^0 = \left(\frac{1}{(j+1)^\alpha}\left(\frac{1}{j+1}, \frac{-\sqrt{j}}{j+1}\right)\right)_{j\geq 0}.$$

Then $\|w_i^0\| = 1/(1+i)^{(1+2\alpha)/2}$.

Now we will calculate $d_V(x_g^k) = \|P_V x_g^k - x_g^k\|$. First, recall that $T^k = c_0^k R_{k\theta_0} \oplus c_1^k R_{k\theta_1} \oplus \cdots$, where

$$R_\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

Thus,

$$\begin{aligned} x_g^k := P_U(z^k) &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} c_j^k R_{k\theta}\left(\frac{1}{(j+1)^\alpha}, 0\right)\right)_{j\geq 0} \\ &= \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} c_j^k \frac{1}{(j+1)^\alpha}\left(\cos(k\theta_j), \sin(k\theta_j)\right)\right)_{j\geq 0} \\ &= \left(c_j^k \frac{\cos(k\theta_j)}{(j+1)^\alpha}(1,0)\right)_{j\geq 0}. \end{aligned}$$

Furthermore, from the identity

$$(P_V)_i = \begin{bmatrix} \cos^2(\theta_i) & \sin(\theta_i)\cos(\theta_i) \\ \sin(\theta_j)\cos(\theta_i) & \sin^2(\theta_i) \end{bmatrix} = \begin{bmatrix} \frac{i}{i+1} & \frac{\sqrt{i}}{i+1} \\ \frac{\sqrt{i}}{i+1} & \frac{1}{i+1} \end{bmatrix},$$

we have

$$P_V x_g^k = \left(c_j^k \frac{\cos(k\theta_j)}{(j+1)^\alpha}\left(\frac{j}{j+1}, \frac{\sqrt{j}}{j+1}\right)\right)_{j\geq 0}.$$

Thus, the the difference has the following form:

$$x_g^k - P_V x_g^k = \left(c_j^k \frac{\cos(k\theta_j)}{(j+1)^\alpha}\left(\frac{1}{j+1}, \frac{-\sqrt{j}}{j+1}\right)\right)_{j\geq 0}.$$

Now we derive the lower bound:

$$\begin{aligned} d_V(x_g^k)^2 &= \|x_g^k - P_V x_g^k\|^2 \\ &= \sum_{i=0}^{\infty} c_i^{2k} \frac{\cos^2(k\theta_i)}{(i+1)^{2\alpha+1}} \\ &= \sum_{i=0}^{\infty} c_i^{2k} \frac{\cos^2\left(k\cos^{-1}\left(\sqrt{\frac{i}{i+1}}\right)\right)}{(i+1)^{2\alpha+1}} \\ &\geq \frac{1}{e}\sum_{i=k}^{\infty} \frac{\cos^2\left(k\cos^{-1}\left(\sqrt{\frac{i}{i+1}}\right)\right)}{(i+1)^{2\alpha+1}}. \end{aligned} \tag{65}$$

The next several lemmas will focus on estimating the order of the sum in Equation (65). After which, Equation 10 will follow from Equation (65) and Lemma 10, below. This completes the proof of Theorem 10. □

Lemma 8 *Let $h : \mathbf{R}_+ \to \mathbf{R}_+$ be a continuously differentiable function such that $h \in L_1(\mathbf{R}_+)$ and $\sum_{i=1}^{\infty} h(i) < \infty$. Then for all positive integers k,*

$$\left| \int_k^{\infty} h(y)dy - \sum_{i=k}^{\infty} h(i) \right| \leq \sum_{i=k}^{\infty} \max_{y \in [i,i+1]} |h'(y)| .$$

Proof We just apply the Mean Value Theorem and combine the integral with the sum

$$\left| \int_k^{\infty} h(y)dy - \sum_{i=k}^{\infty} h(i) \right| \leq \left| \sum_{i=k}^{\infty} \int_i^{i+1} (h(y) - h(i))dy \right| \leq \sum_{i=k}^{\infty} \int_i^{i+1} |h(y) - h(i)|dy \leq \sum_{i=k}^{\infty} \max_{y \in [i,i+1]} |h'(y)|.$$

□

The following Lemma will quantify the deviation of integral from the sum.

Lemma 9 *The following approximation bound holds:*

$$\left| \sum_{i=k}^{\infty} \frac{\cos^2\left(k \cos^{-1}\left(\sqrt{\frac{i}{i+1}}\right)\right)}{(i+1)^{2\alpha+1}} - \int_k^{\infty} \frac{\cos^2\left(k \cos^{-1}\left(\sqrt{\frac{y}{y+1}}\right)\right)}{(y+1)^{2\alpha+1}} dy \right| = O\left(\frac{1}{(k+1)^{2\alpha+1/2}}\right). \tag{66}$$

Proof We will use Lemma 8 with

$$h(y) = \frac{\cos^2\left(k \cos^{-1}\left(\sqrt{\frac{y}{y+1}}\right)\right)}{(y+1)^{2\alpha+1}}.$$

to deduce an upper bound on the absolute value. Indeed,

$$\begin{aligned} |h'(y)| &= \left| \frac{k \sin\left(k \cos^{-1}\left(\sqrt{\frac{y}{y+1}}\right)\right) \cos\left(k \cos^{-1}\left(\sqrt{\frac{y}{y+1}}\right)\right)}{\sqrt{y}(y+1)(y+1)^{2\alpha+1}} - \frac{\cos^2\left(k \cos^{-1}\left(\sqrt{\frac{y}{y+1}}\right)\right)}{(y+1)^{2\alpha+2}} \right| \\ &= O\left(\frac{k}{(y+1)^{2\alpha+1+3/2}} + \frac{1}{(y+1)^{2\alpha+2}}\right). \end{aligned}$$

Therefore, we can bound Equation (66) by the following sum:

$$\sum_{i=k}^{\infty} \max_{y \in [i,i+1]} |h'(y)| = O\left(\frac{k}{(k+1)^{2\alpha+3/2}} + \frac{1}{(k+1)^{2\alpha+1}}\right) = O\left(\frac{1}{(k+1)^{2\alpha+1/2}}\right).$$

□

In the following Lemma, we estimate the order of the oscillatory integral approximation to the sum in Equation (65). The proof follows by a change of variables and an integration by parts.

Lemma 10 *The following bound holds:*

$$\sum_{i=k}^{\infty} \frac{\cos^2\left(k\cos^{-1}\left(\sqrt{\frac{i}{i+1}}\right)\right)}{(i+1)^{2\alpha+1}} dy = \Omega\left(\frac{1}{(k+1)^{2\alpha}}\right). \tag{67}$$

Proof Fix $k \geq 1$. We first perform a change of variables $u = \cos^{-1}(\sqrt{y/(y+1)})$ on the integral approximation of the sum:

$$\int_k^{\infty} \frac{\cos^2\left(k\cos^{-1}\left(\sqrt{\frac{y}{y+1}}\right)\right)}{(y+1)^{2\alpha+1}} dy = 2\int_0^{\cos^{-1}\left(\sqrt{k/(k+1)}\right)} \cos^2(ku)\cos(u)\sin^{4\alpha-1}(u)du. \tag{68}$$

We will show that the right hand side of Equation (68) is of order $\Omega\left(1/(k+1)^{2\alpha}\right)$. Then Equation (67) will follow by Lemma 9.

Let $\rho := \cos^{-1}(\sqrt{k/(k+1)})$. We have

$$\begin{aligned} 2\int_0^{\rho} \cos^2(ku)\cos(u)\sin^{4\alpha-1}(u)du &= \int_0^{\rho} (1+\cos(2ku))\cos(u)\sin^{4\alpha-1}(u)du \\ &= p_1 + p_2 + p_3 \end{aligned}$$

where

$$\begin{aligned} p_1 &= \int_0^{\rho} 1\cdot\cos(u)\sin^{4\alpha-1}(u)du = \frac{1}{4\alpha}\sin^{4\alpha}(\rho), \\ p_2 &= \frac{1}{2k}\sin(2k\rho)\cos(\rho)\sin^{4\alpha-1}(\rho), \\ p_3 &= -\frac{1}{2k}\int_0^{\rho} \sin(2ku)d(\cos(u)\sin^{4\alpha-1}(u)), \end{aligned}$$

and we have applied integration by parts for $\int_0^{\rho}\cos(2ku)\cos(u)\sin^{4\alpha-1}(u)du = p_2 + p_3$.

Because $\sin(\cos^{-1}(x)) = \sqrt{1-x^2}$, for all $\eta > 0$, we get

$$\sin^{\eta}(\rho) = \sin^{\eta}\cos^{-1}\left(\sqrt{k/(k+1)}\right) = \frac{1}{(k+1)^{\eta/2}}.$$

In addition, we have $\cos(\rho) = \cos\cos^{-1}\left(\sqrt{k/(k+1)}\right) = \sqrt{k/(k+1)}$ and the trivial bounds $|\sin(2k\rho)| \leq 1$ and $|\sin(2ku)| \leq 1$.

Therefore, the following bounds hold:

$$\begin{aligned} p_1 &= \frac{1}{4\alpha(k+1)^{2\alpha}}, \\ |p_2| &\leq \frac{\sqrt{k/(k+1)}}{2k(k+1)^{2\alpha-1/2}} = O\left(\frac{1}{(k+1)^{2\alpha+1/2}}\right). \end{aligned}$$

In addition, for p_3, we have $d(\cos(u)\sin^{4\alpha-1}(u)) = \sin^{4\alpha-2}(u)((4\alpha-1)\cos(u) - \sin^2(u))du$. Furthermore, for $u \in [0,\rho]$ and $\alpha > 1/2$, we have $\sin^{4\alpha-2}(u) \in [0, 1/(k+1)^{2\alpha-1}]$ and the following lower bound: $(4\alpha -$

$1)\cos(u) - \sin^2(u) \geq (4\alpha - 1)\cos(\rho) - \sin^2(\rho) = (4\alpha - 1)\sqrt{k/(k+1)} - 1/(k+1) > 0$ as long as $k \geq 1$. Therefore, we have $\sin^{4\alpha-2}(u)((4\alpha - 1)\cos(u) - \sin^2(u)) \geq 0$ for all $u \in [0, \rho]$ and, thus,

$$|p_3| \leq \frac{1}{2k}\cos(\rho)\sin^{4\alpha-1}(\rho) = \frac{\sqrt{k/(k+1)}}{2k(k+1)^{2\alpha-1/2}} = O\left(\frac{1}{(k+1)^{2\alpha+1/2}}\right).$$

Therefore,

$$p_1 + p_2 + p_3 \geq p_1 - |p_2| - |p_3| = \Omega\left(\frac{1}{(k+1)^{2\alpha}}\right).$$

□

We deduce the following theorem from the sum estimation in Lemma 10:

Theorem 11 (Lower complexity of DRS) *There exists closed, proper, and convex functions $f, g : \mathcal{H} \to (-\infty, \infty]$ such that f is 1-Lipschitz, and for every $\alpha > \frac{1}{2}$, there is a point $z^0 \in \mathcal{H}$ and $\gamma \in \mathbf{R}_{++}$ such that if $(z^j)_{j\geq 0}$ is generated by Algorithm 1 with $\lambda_k = 1/2$ for all $k \geq 0$, then*

$$f(x_g^k) + g(x_g^k) - f(x^*) - g(x^*) = \Omega\left(\frac{1}{(k+1)^{\alpha}}\right).$$

Proof Assume the setting of Theorem 10. Then $f = d_V$ and $g = \chi_U$, and

$$f(x_g^k) + g(x_g^k) - f(x^*) - g(x^*) = d_V(x_g^k) = \Omega\left(\frac{1}{(k+1)^{\alpha}}\right)$$

by Lemma 10. □

Theorem 11 shows that the DRS algorithm *is nearly as slow* as the subgradient method. We use the word nearly because the subgradient method has complexity $O(1/\sqrt{k+1})$, while DRS has complexity $o(1/\sqrt{k+1})$. To the best of our knowledge, this is the first *lower complexity* result for DRS algorithm. Note that Theorem 11 implies the same lower complexity for the Forward Douglas Rachford splitting algorithm [10].

7.4 Optimal objective and FPR rates with Lipschitz derivative

The following examples show that the objective and FPR rates derived in Theorem 3 are essentially optimal. The setup of the following counterexample already appeared in [9, Remarque 6] but the objective function lower bounds were not shown.

Theorem 12 (Lower complexity of PPA) *There exists a Hilbert space $\mathcal{H}$, and a closed, proper, and convex function f such that for all $\alpha > 1/2$, there exists $z^0 \in \mathcal{H}$ such that if $(z^j)_{j\geq 0}$ is generated by PPA (Equation (8)), then*

$$\|\mathbf{prox}_{\gamma f}(z^k) - z^k\|^2 \geq \frac{\gamma^2}{(1+2\alpha)e^{2\gamma}(k+\gamma)^{1+2\alpha}} \qquad \text{and} \qquad f(z^{k+1}) - f(x^*) \geq \frac{1}{4\alpha e^{2\gamma}(k+1+\gamma)^{2\alpha}}.$$

Proof Let $\mathcal{H} = \ell_2(\mathbf{R})$, and define a linear map $A : \mathcal{H} \to \mathcal{H}$:

$$A\left(z_1, z_2, \cdots, z_n, \cdots\right) = \left(z_1, \frac{z_2}{2}, \cdots, \frac{z_n}{n}, \cdots\right).$$

For all $z \in \mathcal{H}$, define $f(x) = (1/2)\langle Az, z\rangle$. Thus, we have the proximal identity for f and

$$\mathbf{prox}_{\gamma f}(z) = (I + \gamma A)^{-1}(z) = \left(\frac{j}{j+\gamma} z_j\right)_{j\geq 1} \qquad \text{and} \qquad (I - \mathbf{prox}_{\gamma f})(z) = \left(\frac{\gamma}{j+\gamma} z_j\right)_{j\geq 1}.$$

Now let $z^0 = (1/(j+\gamma)^\alpha)_{j\geq 1} \in \mathcal{H}$, and set $T = \mathbf{prox}_{\gamma f}$. Then we get the following FPR lower bound:

$$\begin{aligned}\|z^{k+1} - z^k\|^2 = \|T^k(T-I)z^0\|^2 = \sum_{i=1}^{\infty}\left(\frac{i}{i+\gamma}\right)^{2k}\frac{\gamma^2}{(i+\gamma)^{2+2\alpha}} &\geq \sum_{i=k}^{\infty}\left(\frac{i}{i+\gamma}\right)^{2k}\frac{\gamma^2}{(i+\gamma)^{2+2\alpha}} \\ &\geq \frac{\gamma^2}{(1+2\alpha)e^{2\gamma}(k+\gamma)^{1+2\alpha}}.\end{aligned}$$

Furthermore, the objective lower bound holds

$$\begin{aligned}f(z^{k+1}) - f(x^*) = \frac{1}{2}\langle Az^{k+1}, z^{k+1}\rangle &= \frac{1}{2}\sum_{i=1}^{\infty}\frac{1}{i}\left(\frac{i}{i+\gamma}\right)^{2(k+1)}\frac{1}{(i+\gamma)^{2\alpha}} \\ &\geq \frac{1}{2}\sum_{i=k+1}^{\infty}\left(\frac{i}{i+\gamma}\right)^{2(k+1)}\frac{1}{(i+\gamma)^{1+2\alpha}} \\ &\geq \frac{1}{4\alpha e^{2\gamma}(k+1+\gamma)^{2\alpha}}.\end{aligned}$$

□

8 From relaxed PRS to relaxed ADMM

It is well known that ADMM is equivalent to DRS applied to the Lagrange dual of Problem (2) [22]. Thus, if we let $d_f(w) := f^*(A^*w)$ and $d_g(w) := g^*(B^*w) - \langle w, b\rangle$, then relaxed ADMM is equivalent to relaxed PRS applied to the following problem:

$$\underset{w\in\mathcal{G}}{\text{minimize}}\ d_f(w) + d_g(w). \tag{69}$$

We make two assumptions regarding d_f and d_g:

Assumption 4 (Solution existence) *Functions $f, g : \mathcal{H} \to (-\infty, \infty]$ satisfy*

$$\operatorname{zer}(\partial d_f + \partial d_g) \neq \emptyset. \tag{70}$$

This is a restatement of Assumption 4, which we in our analysis of the primal case.

Assumption 5 *The following differentiation rule holds:*

$$\partial d_f(x) = A^* \circ (\partial f^*) \circ A \qquad \text{and} \qquad \partial d_g(x) = B^* \circ (\partial g^*) \circ B.$$

See [2, Theorem 16.37] for conditions that imply this identity. We need this assumption to compute subgradients of d_f and d_g.

Given an initial vector $z^0 \in \mathcal{G}$, Lemma 4 shows that at each iteration relaxed PRS performs the following computations:

$$\begin{cases} w_{d_g}^k &= \mathbf{prox}_{\gamma d_g}(z^k); \\ w_{d_f}^k &= \mathbf{prox}_{\gamma d_f}(2w_{d_g}^k - z^k); \\ z^{k+1} &= z^k + 2\lambda_k(w_{d_f}^k - w_{d_g}^k). \end{cases} \tag{71}$$

In order to apply the relaxed PRS algorithm, we need to compute the proximal operators of the dual functions d_f and d_g.

Lemma 11 (Proximity operators on the dual) *Let $w, v \in \mathcal{G}$. Then the update formulas $w^+ = \mathbf{prox}_{\gamma d_f}(w)$ and $v^+ = \mathbf{prox}_{\gamma d_g}(v)$ are equivalent to the following computations*

$$\begin{cases} x^+ = \arg\min_{x \in \mathcal{H}_1} f(x) - \langle w, Ax\rangle + \frac{\gamma}{2}\|Ax\|^2; \\ w^+ = w - \gamma A x^+. \end{cases} \quad \text{and} \quad \begin{cases} y^+ = \arg\min_{y \in \mathcal{H}_2} g(y) - \langle v, By - b\rangle + \frac{\gamma}{2}\|By - b\|^2; \\ v^+ = v - \gamma(By^+ - b). \end{cases} \tag{72}$$

respectively. In addition, the subgradient inclusions hold: $A^* w^+ \in \partial f(x^+)$ and $B^* v^+ \in \partial g(y^+)$. Finally, w^+ and v^+ are independent of the choice of x^+ and y^+, respectively, even if they are not unique solutions to the minimization subproblems.

We can use Lemma 11 to derive the relaxed form of ADMM in Algorithm 2. Note that this form of ADMM eliminates the "hidden variable" sequence $(z^j)_{j \geq 0}$ in Equation (71). This following derivation is not new, but is included for the sake of completeness. See [22] for the original derivation.

Proposition 11 (Relaxed ADMM) *Let $z^0 \in \mathcal{G}$, and let $(z^j)_{j \geq 0}$ be generated by the relaxed PRS algorithm applied to the dual formulation in Equation (69). Choose initial points $w_{d_g}^{-1} = z^0, x^{-1} = 0$ and $y^{-1} = 0$ and initial relaxation $\lambda_{-1} = 1/2$. Then we have the following identities starting from $k = -1$:*

$$\begin{aligned} y^{k+1} &= \arg\min_{y \in \mathcal{H}_2} g(y) - \langle w_{d_g}^k, Ax^k + By - b\rangle + \frac{\gamma}{2}\|Ax^k + By - b + (2\lambda_k - 1)(Ax^k + By^k - b)\|^2 \\ w_{d_g}^{k+1} &= w_{d_g}^k - \gamma(Ax^k + By^{k+1} - b) - \gamma(2\lambda_k - 1)(Ax^k + By^k - b) \\ x^{k+1} &= \arg\min_{x \in \mathcal{H}_1} f(x) - \langle w_{d_g}^{k+1}, Ax + By^{k+1} - b\rangle + \frac{\gamma}{2}\|Ax + By^{k+1} - b\|^2 \\ w_{d_f}^{k+1} &= w_{d_g}^{k+1} - \gamma(Ax^{k+1} + By^{k+1} - b) \end{aligned}$$

Proof By Equation (71) and Lemma 11, we get the following formulation for the k-th iteration: Given $z^0 \in \mathcal{H}$

$$\begin{cases} y^k &= \arg\min_{y \in \mathcal{H}_2} g(y) - \langle z^k, By - b\rangle + \frac{\gamma}{2}\|By - b\|^2 \\ w_{d_g}^k &= z^k - \gamma(By^{k+1} - b) \\ x^k &= \arg\min_{x \in \mathcal{H}_1} f(x) - \langle 2w_{d_g}^{k+1} - z^k, Ax\rangle + \frac{\gamma}{2}\|Ax\|^2 \\ w_{d_f}^k &= 2w_{d_g}^k - z^k - \gamma A x^k \\ z^{k+1} &= z^k + 2\lambda_k(w_{d_f}^k - w_{d_g}^k) \end{cases} \tag{73}$$

We will use this form to get to the claimed iteration. First,

$$2w_{d_g}^k - z^k = w_{d_g}^k - \gamma(By^k - b) \qquad \text{and} \qquad w_{d_f}^k = w_{d_g}^k - \gamma(Ax^k + By^k - b). \tag{74}$$

Furthermore, we can simplify the definition of x^k:

$$\begin{aligned} x^k &= \operatorname*{arg\,min}_{x\in\mathcal{H}_1} f(x) - \langle 2w_{d_g}^{k+1} - z^k, Ax\rangle + \frac{\gamma}{2}\|Ax\|^2 \\ &\overset{(74)}{=} \operatorname*{arg\,min}_{x\in\mathcal{H}_1} f(x) - \langle w_{d_f}^k - \gamma(By^k - b), Ax\rangle + \frac{\gamma}{2}\|Ax\|^2 \\ &= \operatorname*{arg\,min}_{x\in\mathcal{H}_1} f(x) - \langle w_{d_f}^k, Ax + By^k - b\rangle + \frac{\gamma}{2}\|Ax + By^k - b\|^2. \end{aligned} \tag{75}$$

Note that the last two lines of Equation (75) differ by terms independent of x.

We now eliminate the z^k variable from the y^k subproblem: because $w_{d_f}^k + z^k = 2w_{d_g}^k - \gamma Ax^k$, we have

$$\begin{aligned} z^{k+1} &= z^k + 2\lambda_k(w_{d_f}^k - w_{d_g}^k) \\ &\overset{(74)}{=} z^k + w_{d_f}^k - w_{d_g}^k + \gamma(2\lambda_k - 1)(Ax^k + By^k - b) \\ &= w_{d_g}^k - \gamma Ax^k - \gamma(2\lambda_k - 1)(Ax^k + By^k - b). \end{aligned} \tag{76}$$

We can simplify the definition of y^{k+1} by with the identity in Equation (76):

$$\begin{aligned} y^{k+1} &= \operatorname*{arg\,min}_{y\in\mathcal{H}_2} g(y) - \langle z^{k+1}, By - b\rangle + \frac{\gamma}{2}\|By - b\|^2 \\ &\overset{(76)}{=} \operatorname*{arg\,min}_{y\in\mathcal{H}_2} g(y) - \langle w_{d_g}^k - \gamma Ax^k - \gamma(2\lambda_k - 1)(Ax^k + By^k - b), By - b\rangle + \frac{\gamma}{2}\|By - b\|^2 \\ &= \operatorname*{arg\,min}_{y\in\mathcal{H}_2} g(y) - \langle w_{d_g}^k, Ax^k + By - b\rangle + \frac{\gamma}{2}\|Ax^k + By - b + (2\lambda_k - 1)(Ax^k + By^k - b)\|^2. \end{aligned} \tag{77}$$

The result then follows from Equations (73), (74), (75), and (77), combined with the initial conditions listed in the statement of the proposition. In particular, note that the updates of x, y, w_{d_f}, and w_{d_g} do not explicitly depend on z □

Remark 2 Proposition 11 proves that $w_{d_f}^{k+1} = w_{d_g}^{k+1} - \gamma(Ax^{k+1} + By^{k+1} - b)$. Recall that by Equation (71), $z^{k+1} - z^k = 2\lambda_k(w_{d_f}^k - w_{d_g}^k)$. Therefore, it follows that

$$z^{k+1} - z^k = -2\gamma\lambda_k(Ax^k + By^k - b). \tag{78}$$

8.1 Dual feasibility convergence rates

We can apply the results of Section 5 to deduce convergence rates for the dual objective functions. Instead of restating those theorems, we just list the following bounds on the feasibility of the primal iterates.

Theorem 13 *Suppose that $(z^j)_{j\geq 0}$ is generated by Algorithm 2, and let $(\lambda_j)_{j\geq 0} \subseteq (0, 1]$. Then the following convergence rates hold:*

1. **Ergodic convergence:** *The feasibility convergence rate holds:*

$$\|A\overline{x}^k + B\overline{y}^k - b\|^2 = \frac{4\|z^0 - z^*\|^2}{\gamma \Lambda_k^2}. \tag{79}$$

2. **Nonergodic convergence:** *Suppose that* $\underline{\tau} = \inf_{j\geq 0} \lambda_j(1-\lambda_j) > 0$. *Then*

$$\|Ax^k + By^k - b\|^2 \leq \frac{\|z^0 - z^*\|^2}{4\gamma^2 \underline{\tau}(k+1)} \qquad \text{and} \qquad \|Ax^k + By^k - b\|^2 = o\left(\frac{1}{k+1}\right). \tag{80}$$

Proof Parts 1 and 2 are straightforward applications of the FPR identity:

$$z^k - z^{k+1} \overset{(78)}{=} 2\gamma\lambda_k(Ax^k + By^k - b).$$

and Corollary 2. □

8.2 Converting dual inequalities to primal inequalities

The ADMM algorithm generates 5 sequences of iterates:

$$(z^j)_{j\geq 0}, (w_{d_f}^j)_{j\geq 0}, \text{ and } (w_{d_g}^j)_{j\geq 0} \subseteq \mathcal{G} \qquad \text{and} \qquad (x^j)_{j\geq 0} \in \mathcal{H}_1, (y^j)_{j\geq 0} \in \mathcal{H}_2.$$

The dual variables do not necessarily have a meaningful interpretation, so it is desirable to derive convergence rates involving the primal variables. In this section we will apply the Fenchel-Young inequality [2, Proposition 16.9] to convert the dual objective into a primal expression.

The following proposition will help us derive primal fundamental inequalities akin to Proposition 4 and 5.

Proposition 12 *Suppose that* $(z^j)_{j\geq 0}$ *is generated by Algorithm 2. Let* z^* *be a fixed point of* T_{PRS} *and let* $w^* = \mathbf{prox}_{\gamma d_f}(z^*)$. *Then the following identity holds:*

$$\begin{aligned} 4\gamma\lambda_k(f(x^k) + g(y^k) - f(x^*) - g(y^*)) = &-4\gamma\lambda_k(d_f(w_{d_f}^k) + d_g(w_{d_g}^k) - d_f(w^*) - d_g(w^*)) \\ &+ \left(2\left(1 - \frac{1}{2\lambda_k}\right)\|z^k - z^{k+1}\|^2 + 2\langle z^k - z^{k+1}, z^{k+1}\rangle\right). \end{aligned} \tag{81}$$

Proof We have the following subgradient inclusions from Proposition 11: $A^* w_{d_f}^k \in \partial f(x^k)$ and $B^* w_{d_g}^k \in \partial g(y^k)$. From the Fenchel-Young inequality [2, Proposition 16.9] we have the expression for f and g:

$$d_f(w_{d_f}^k) = \langle A^* w_{d_f}^k, x^k\rangle - f(x^k) \qquad \text{and} \qquad d_f(w_{d_g}^k) = \langle B^* w_{d_g}^k, y^k\rangle - g(y^k) - \langle w_{d_g}^k, b\rangle.$$

Therefore,

$$-d_f(w_{d_f}^k) - d_g(w_{d_g}^k) = f(x^k) + g(y^k) - \langle Ax^k + By^k - b, w_{d_f}^k\rangle - \langle w_{d_g}^k - w_{d_f}^k, By^k - b\rangle.$$

Let us simplify this bound with an identity from Proposition 11: from $w_{d_f}^k - w_{d_g}^k = -\gamma(Ax^k + By^k - b)$, it follows that

$$-d_f(w_{d_f}^k) - d_g(w_{d_g}^k) = f(x^k) + g(y^k) + \frac{1}{\gamma}\langle w_{d_f}^k - w_{d_g}^k, w_{d_f}^k + \gamma(By^k - b)\rangle. \tag{82}$$

Recall that $\gamma(By^k - b) = z^k - w_{d_g}^k$. Therefore

$$w_{d_f}^k + \gamma(By^k - b) = z^k + (w_{d_f}^k - w_{d_g}^k) = z^k + \frac{1}{2\lambda_k}(z^{k+1} - z^k) = \frac{1}{2\lambda_k}(2\lambda_k - 1)(z^k - z^{k+1}) + z^{k+1},$$

and the inner product term can be simplified as follows:

$$\begin{aligned}\frac{1}{\gamma}\langle w_{d_f}^k - w_{d_g}^k, w_{d_f}^k + \gamma(By^k - b)\rangle &= \frac{1}{\gamma}\langle\frac{1}{2\lambda_k}(z^{k+1} - z^k), \frac{1}{2\lambda_k}(2\lambda_k - 1)(z^k - z^{k+1})\rangle \\ &\quad + \frac{1}{\gamma}\langle\frac{1}{2\lambda_k}(z^{k+1} - z^k), z^{k+1}\rangle \\ &= -\frac{1}{2\gamma\lambda_k}\left(1 - \frac{1}{2\lambda_k}\right)\|z^{k+1} - z^k\|^2 \\ &\quad - \frac{1}{2\gamma\lambda_k}\langle z^k - z^{k+1}, z^{k+1}\rangle. \end{aligned} \tag{83}$$

Now we derive an expression for the dual objective at a dual optimal w^*. First, if z^* is a fixed point of T_{PRS}, then $0 = T_{\mathrm{PRS}}(z^*) - z^* = 2(w_{d_g}^* - w_{d_f}^*) = -2\gamma(Ax^* + By^* - b)$. Thus, from Equation (82) with k replaced by $*$, we get

$$-d_f(w^*) - d_g(w^*) = f(x^*) + g(y^*) + \langle Ax^* + Bx^* - b, w^*\rangle = f(x^*) + g(y^*). \tag{84}$$

Therefore, Equation (81) follows by subtracting (84) from Equation (82), rearranging and using the identity in Equation (83). □

The following two propositions prove two fundamental inequalities that bound the primal objective.

Proposition 13 (ADMM primal upper fundamental inequality) *Let z^* be a fixed point of T_{PRS} and let $w^* = \mathbf{prox}_{\gamma d_g}(z^*)$. Then for all $k \geq 0$, we have the bound:*

$$\begin{aligned}&4\gamma\lambda_k(f(x^k) + g(y^k) - f(x^*) - g(y^*)) \\ &\qquad \leq \|z^k - (z^* - w^*)\|^2 - \|z^{k+1} - (z^* - w^*)\|^2 + \left(1 - \frac{1}{\lambda_k}\right)\|z^k - z^{k+1}\|^2. \end{aligned} \tag{85}$$

Proof The fundamental lower inequality in Proposition 5 applied to $d_f + d_g$ shows that

$$-4\gamma\lambda_k(d_f(w_{d_f}^k) + d_g(w_{d_g}^k) - d_f(w^*) - d_g(w^*)) \leq 2\langle z^{k+1} - z^k, z^* - w^*\rangle.$$

The proof then follows from Proposition 12, and the simplification:

$$\begin{aligned}&2\langle z^k - z^{k+1}, z^{k+1} - (z^* - w^*)\rangle + 2\left(1 - \frac{1}{2\lambda_k}\right)\|z^k - z^{k+1}\|^2 \\ &\qquad = \|z^k - (z^* - w^*)\|^2 - \|z^{k+1} - (z^* - w^*)\|^2 + \left(1 - \frac{1}{\lambda_k}\right)\|z^k - z^{k+1}\|^2.\end{aligned}$$

□

Remark 3 Note that Equation (85) is nearly identical to the upper inequality in Proposition 4, except that $z^* - w^*$ appears in the former where x^* appears in the latter.

Proposition 14 (ADMM primal lower fundamental inequality) *Let z^* be a fixed point of T_{PRS} and let $w^* = \mathbf{prox}_{\gamma d_g}(z^*)$. Then for all $x \in \mathcal{H}_1$ and $y \in \mathcal{H}_2$ we have the bound:*

$$f(x) + g(y) - f(x^*) - g(y^*) \geq \langle Ax + By - b, w^* \rangle. \tag{86}$$

Proof The lower bound follows from the subgradient inequalities:

$$f(x) - f(x^*) \geq \langle x - x^*, A^* w^* \rangle,$$
$$g(y) - g(y^*) \geq \langle y - y^*, B^* w^* \rangle.$$

We add these inequalities together and use the identity $Ax^* + By^* = b$ to get Equation (86). □

Remark 4 Inequality (86) takes a special form when $x = x^k$ and $y = y^k$

$$f(x^k) + g(y^k) - f(x^*) - g(y^*) \geq \frac{1}{\gamma} \langle w_{d_g}^k - w_{d_f}^k, w^* \rangle, \tag{87}$$

or $x = \overline{x}^k$ and $y = \overline{y}^k$

$$f(\overline{x}^k) + g(\overline{y}^k) - f(x^*) - g(y^*) \geq \frac{1}{\gamma} \langle \overline{w}_{d_g}^k - \overline{w}_{d_f}^k, w^* \rangle. \tag{88}$$

These bounds are nearly identical to the fundamental lower inequality in Proposition 5, except that w^* appears in the former, where $z^* - x^*$ appeared in the latter.

8.3 Converting dual convergence rates to primal convergence rates

In this section, we use the inequalities deduced in Section 8.2 to derive convergence rates for the primal objective values. The structure of the of the proofs of the following theorems are exactly the same as in the primal convergence case in Section 5, except that we use the upper and lower inequalities derived in the Section 8.2 instead of the fundamental upper and lower inequalities in Propositions 4 and 5. This amounts to replacing the term $z^* - x^*$ and x^* by w^* and $z^* - w^*$, respectively, in all of the inequalities from Section 5. Thus, we omit the proofs.

Theorem 14 (Ergodic primal convergence of ADMM) *Define the ergodic primal iterates by the formulas: $\overline{x}^k = (1/\Lambda_k) \sum_{i=0}^k \lambda_i x^i$ and $\overline{y}^k = (1/\Lambda_k) \sum_{i=0}^k \lambda_i y^i$. Then*

$$-\frac{2\|w^*\|\|z^0 - z^*\|}{\gamma \Lambda_k} \leq f(\overline{x}^k) + g(\overline{y}^k) - f(x^*) - g(y^*) \leq \frac{\|z^0 - (z^* - w^*)\|^2}{4\gamma \Lambda_k}. \tag{89}$$

The ergodic rate presented here is stronger and easier to interpret than the one in [27] for the ADMM algorithm ($\lambda_k \equiv 1/2$). Indeed, the rate presented in [27, Theorem 4.1] shows the following bound: for all $k \geq 1$ and for any bounded set $\mathcal{D} \subseteq \mathrm{dom}(f) \times \mathrm{dom}(g) \times \mathcal{G}$, we have the following variational inequality bound

$$\sup_{(x,y,w) \in \mathcal{D}} \left(f(\overline{x}^{k-1}) + g(\overline{y}^k) - f(x) - g(y) + \langle \overline{w}_{d_g}^k, Ax + By - b \rangle - \langle A\overline{x}^{k-1} + B\overline{y}^k - b, w \rangle \right)$$
$$\leq \frac{\sup_{(x,y,w) \in \mathcal{D}} \|(x,y,w) - (x^0, y^0, w_{d_g}^0)\|^2}{2(k+1)}.$$

If $(x^*, y^*, w^*) \in \mathcal{D}$, then the supremum is positive and bounds the deviation of the primal objective from the lower fundamental inequality.

Theorem 15 (Nonergodic primal convergence of ADMM) *For all $k \geq 0$, let $\tau_k = \lambda_k(1-\lambda_k)$. In addition, suppose that $\underline{\tau} = \inf_{j\geq 0} \tau_j > 0$. Then*

1. *In general, we have the bounds:*

$$\frac{-\|z^0 - z^*\|\|w^*\|}{2\sqrt{\underline{\tau}(k+1)}} \leq f(x^k) + g(y^k) - f(x^*) - g(y^*) \leq \frac{\|z^0 - z^*\|(\|z^0 - z^*\| + \|w^*\|)}{2\gamma\sqrt{\underline{\tau}(k+1)}} \tag{90}$$

   and $|f(x^k) + g(y^k) - f(x^*) - g(y^*)| = o(1/\sqrt{k+1})$.
2. *If $\mathcal{G} = \mathbf{R}$ and $\lambda_k \equiv 1/2$, then for all $k \geq 0$,*

$$\frac{-\|z^0 - z^*\|\|w^*\|}{\sqrt{2}(k+1)} \leq f(x^{k+1}) + g(y^{k+1}) - f(x^*) - g(x^*) \leq \frac{\|z^0 - z^*\|(\|z^0 - z^*\| + \|w^*\|)}{\sqrt{2}\gamma(k+1)}$$

   and $|f(x^{k+1}) + g(x^{k+1}) - f(x^*) - g(x^*)| = o(1/(k+1))$.

The rates presented in Theorem 15 are new and, to the best of our knowledge, they are the first nonergodic convergence rate results for ADMM primal objective error.

9 Examples

In this section, we apply relaxed PRS and relaxed ADMM to concrete problems and explicitly bound the associated objectives and FPR terms with the convergence rates we derived in the previous sections.

9.1 Feasibility problems

Suppose that C_f and C_g are closed convex subsets of $\mathcal{H}$, with nonempty intersection. The goal of the feasibility problem is the find a point in the intersection of C_f and C_g. In this section, we present one way to model this problem using convex optimization and apply the relaxed PRS algorithm to reach the minimum.

In general, we cannot expect linear convergence of relaxed PRS algorithm for the feasibility problem. We showed this in Theorem 9 by constructing an example for which the DRS iteration converges in norm but does so *arbitrarily slow*. A similar result holds for the alternating projection (AP) algorithm [3]. Thus, in this section we focus on the convergence rate of the *FPR*.

Let χ_{C_f} and χ_{C_g} be the characteristic functions of C_f and C_g. Then $x \in C_f \cap C_g$, if, and only if, $\chi_{C_f}(x) + \chi_{C_g}(x) = 0$, and the sum is infinite otherwise. Thus, a point is in the intersection of C_f and C_g if, and only if, it is the minimizer of the following problem:

$$\operatorname*{minimize}_{x\in\mathcal{H}} \chi_{C_f}(x) + \chi_{C_g}(x). \tag{91}$$

The relaxed PRS algorithm applied to this problem, with $f = \chi_{C_f}$ and $g = \chi_{C_g}$, has the following form: Given $z^0 \in \mathcal{H}$, for all $k \geq 0$, let

$$\begin{cases} x_g^k = P_{C_g}(z^k); \\ x_f^k = P_{C_f}(2x_g^k - z^k); \\ z^{k+1} = z^k + 2\lambda_k(x_f^k - x_g^k). \end{cases} \tag{92}$$

Because $f = \chi_{C_f}$ and $g = \chi_{C_g}$ only take on the values 0 and ∞, the objective value convergence rates derived earlier do not provide meaningful information, other than $x_f^k \in C_f$ and $x_g^k \in C_g$. However, from the FPR identity

$$x_f^k - x_g^k = \frac{1}{2\lambda_k}(z^{k+1} - z^k),$$

we find that after k iterations, Corollary 2 produces the bound

$$\max\{d_{C_g}^2(x_f^k), d_{C_f}^2(x_g^k)\} \leq \|x_f^k - x_g^k\|^2 = o\left(\frac{1}{k+1}\right) \tag{93}$$

whenever $(\lambda_j)_{j\geq 0}$ is bounded away from 0 and 1. Theorem 8 showed that this rate is optimal. Furthermore, if we average the iterates over all k, Theorem 6 gives the improved bound

$$\max\{d_{C_g}^2(\overline{x}_f^k), d_{C_f}^2(\overline{x}_g^k)\} \leq \|\overline{x}_f^k - \overline{x}_g^k\|^2 = O\left(\frac{1}{\Lambda_k^2}\right), \tag{94}$$

which is optimal by Proposition 7. Note that the averaged iterates satisfy $\overline{x}_f^k = (1/\Lambda_k)\sum_{i=0}^k \lambda_i x_f^i \in C_f$ and $\overline{x}_g^k = (1/\Lambda_k)\sum_{i=0}^k \lambda_i x_g^i \in C_g$, because C_f and C_g are convex. Thus, we can state the following proposition:

Proposition 15 *After k iterations the relaxed PRS algorithm produces a point in each set with distance of order $O(1/\Lambda_k)$ from each other.*

9.2 Parallelized model fitting and classification

The following general scenario appears in [8, Chapter 8]. Consider the following general convex model fitting problem: Let $M : \mathbf{R}^n \to \mathbf{R}^m$ be a *feature matrix*, let $b \in \mathbf{R}^m$ be the *output* vector, let $l : \mathbf{R}^m \to (-\infty, \infty]$ be a *loss function* and let $r : \mathbf{R}^n \to (-\infty, \infty]$ be a *regularization function*. The *model fitting problem* is formulated as the following minimization:

$$\underset{x\in\mathbf{R}^n}{\text{minimize}}\ l(Mx - b) + r(x). \tag{95}$$

The function l is used to enforce the constraint $Mx = b + \nu$ up to some noise ν in the measurement, while r enforces the *regularity* of x by incorporating *prior knowledge* of the form of the solution. The function r can also be used to enforce the uniqueness of the solution of $Mx = b$ in ill-posed problems.

We can solve Equation (95) by a direct application of relaxed PRS and obtain $O(1/\Lambda_k)$ ergodic convergence and $o\left(1/\sqrt{k+1}\right)$ nonergodic convergence rates. Note that these rates do not require differentiability of f or g. In contrast, the FBS algorithm requires differentiability of one of the objective functions and a knowledge of the Lipschitz constant of its gradient. The advantage of FBS is the $o(1/(k+1))$ convergence rate shown in Theorem 3. However, we do not necessarily assume that l is differentiable, so we may need to compute $\mathbf{prox}_{\gamma l(Mx-b)}$, which can be significantly more difficult than computing $\mathbf{prox}_{\gamma l}$. Thus, in this section we separate M from l by rephrasing Equation (95) in the form of Problem (2).

In this section, we present several different ways to split Equation (95). Each splitting gives rise to a different algorithm and can be applied to general convex l and r. Our results predict convergence rates that hold for primal objectives, dual objectives, and the primal feasibility. Note that in parallelized model fitting, it is not always desirable to take the time average of all of the iterates. Indeed, when r enforces sparsity, averaging the current r-iterate with old iterates, all of which are sparse, can produce a non-sparse iterate. This will slow down vector additions and prolong convergence.

9.2.1 Auxiliary variable

We can split Equation (95) by defining an auxiliary variable for My:

$$\begin{aligned} &\underset{x\in\mathbf{R}^m, y\in\mathbf{R}^n}{\text{minimize}} \; l(x) + r(y) \\ &\text{subject to} \;\; My - x = b. \end{aligned} \tag{96}$$

The constraint in Equation (96) reduces to $Ax + By = b$ where $B = M$ and $A = -I_{\mathbf{R}^m}$. If we set $f = l$ and $g = r$ and apply ADMM, the analysis of Section 8.3 shows that

$$|l(x^k) + r(y^k) - l(Mx^* - b) - r(x^*)| = o\left(\frac{1}{\sqrt{k+1}}\right) \qquad \text{and} \qquad \|My^k - b - x^k\|^2 = o\left(\frac{1}{k+1}\right).$$

In particular, if l is Lipschitz, then $|l(x^k) - l(My^k - b)| = o\left(1/\sqrt{k+1}\right)$. Thus, we have

$$|l(My^k - b) + r(y^k) - l(Mx^* - b) - r(x^*)| = o\left(\frac{1}{\sqrt{k+1}}\right).$$

A similar analysis shows that

$$|l(M\overline{y}^k - b) + r(\overline{y}^k) - l(Mx^* - b) - r(x^*)| = O\left(\frac{1}{\Lambda_k}\right) \qquad \text{and} \qquad \|M\overline{y}^k - b - \overline{x}^k\|^2 = O\left(\frac{1}{\Lambda_k^2}\right).$$

In the following two splittings, we leave the derivation of convergence rates to the reader.

9.2.2 Splitting across examples

We assume that l is block separable: we have $l(Mx - b) = \sum_{i=1}^R l_i(M_i x - b_i)$ where

$$M = \begin{bmatrix} M_1 \\ \vdots \\ M_R \end{bmatrix} \qquad \text{and} \qquad b = \begin{bmatrix} b_1 \\ \vdots \\ b_R \end{bmatrix}.$$

Each $M_i \in \mathbf{R}^{m_i \times n}$ is a submatrix of M, each $b_i \in \mathbf{R}^{m_i}$ is a subvector of b, and $\sum_{i=1}^R m_i = m$. Therefore, an equivalent form of Equation (95) is given by

$$\begin{aligned} &\underset{x_1,\cdots,x_R, y\in\mathbf{R}^n}{\text{minimize}} \; \sum_{i=1}^R l_i(M_i x_i - b_i) + r(y) \\ &\text{subject to} \;\; x_r - y = 0, \quad r = 1, \cdots, R. \end{aligned} \tag{97}$$

We say that Equation (97) is *split across examples*. Thus, to apply ADMM to this problem, we simply stack the vectors x_i, $r = 1, \cdots, R$ into a vector $x = (x_1, \cdots, x_R)^T \in \mathbf{R}^{nR}$. Then the constraints in Equation (97) reduce to $Ax + By = 0$ where $A = I_{\mathbf{R}^{nR}}$ and $By = (-y, \cdots, -y)^T$.

9.2.3 Splitting across features

We can also split Equation (95) *across features*, whenever r is block separable in x, in the sense that there exists $C > 0$, such that $r = \sum_{i=1}^{C} r_i(x_i)$, and $x_i \in \mathbf{R}^{n_i}$ where $\sum_{i=1}^{C} n_i = n$. This splitting corresponds to partitioning the columns of M, i.e.

$$M = \begin{bmatrix} M_1, \cdots, M_C \end{bmatrix},$$

and $M_i \in \mathbf{R}^{m \times n_i}$, for all $i = 1, \cdots, C$. Note that for all $y \in \mathbf{R}^n$, $My = \sum_{i=1}^{C} M_i y_i$. With this notation, we can derive an equivalent form of Equation (97) given by

$$\begin{aligned} \underset{x, y \in \mathbf{R}^n}{\text{minimize}}\ & l\left(\sum_{i=1}^{C} x_i - b\right) + \sum_{i=1}^{C} r_i(y_i) \\ \text{subject to}\ \ & x_i - M_i y_i = 0, \quad i = 1, \cdots, C. \end{aligned} \tag{98}$$

The constraint in Equation (98) reduces to $Ax + By = 0$ where $A = I_{\mathbf{R}^{mC}}$ and $By = -(M_1 y_1, \cdots, M_C y_C)^T \in \mathbf{R}^{nC}$.

9.3 Distributed ADMM

In this section our goal is to use Algorithm 2 to

$$\underset{x \in \mathcal{H}}{\text{minimize}} \sum_{i=1}^{m} f_i(x) \tag{99}$$

by using the splitting in [35]. Note that we could minimize this function by reformulating it in the product space $\mathcal{H}^n$ as follows:

$$\underset{\mathbf{x} \in \mathcal{H}^n}{\text{minimize}} \sum_{i=1}^{m} f_i(x_i) + \chi_D(\mathbf{x}), \tag{100}$$

where $D = \{(x, \cdots, x) \in \mathcal{H}^n \mid x \in \mathcal{H}\}$ is the diagonal set. Applying relaxed PRS to this problem results in a parallel algorithm where each function performs a local minimization step and then communicates its local variable to a *central processor*. In this section, we assign each function a local variable but we never communicate it to a central processor. Instead, each function only communicates with *neighbors*.

Formally, we assume there is a simple, connected and undirected graph $G = (V, E)$ on $|V| = m$ vertices with edges, E, that describe a neighbor relation among the different functions. We introduce a new variable $x_i \in \mathcal{H}$ for each function f_i, and, hence, we set $\mathcal{H}_1 = \mathcal{H}^m$, (see Section 8). We can encode the constraint that each node communicates with neighbors by introducing an auxiliary variable for each edge in the graph:

$$\begin{aligned} \underset{\mathbf{x} \in \mathcal{H}^m, \mathbf{y} \in \mathcal{H}^{|E|}}{\text{minimize}}\ & \sum_{i=1}^{m} f_i(x_i) \\ \text{subject to}\ \ & x_i = y_{ij}, x_j = y_{ij}, \text{ for all } (i,j) \in E. \end{aligned} \tag{101}$$

The linear constraints in Equation (101) can be written in the form of $A\mathbf{x} + B\mathbf{y} = 0$ for proper matrices A and B. Thus, we reformulate Equation (101) as

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{H}^m,\mathbf{y}\in\mathcal{H}^{|E|}}{\text{minimize}} \quad & \sum_{i=1}^{m} f_i(x_i) + g(\mathbf{y}) \\ \text{subject to} \quad & A\mathbf{x} + B\mathbf{y} = 0, \end{aligned} \tag{102}$$

where $g : \mathcal{H}^{|E|} \to \mathbf{R}$ is the zero map.

Because we only care about finding the value of the variable $\mathbf{x} \in \mathcal{H}^m$, the following simplification can be made to the sequences generated by ADMM applied to Equation (102) with $\lambda_k = 1/2$ for all $k \geq 1$ [36]: Let $\mathcal{N}_i$ denote the set of neighbors of $i \in V$ and set $x_i^0 = \alpha_i^0 = 0$ for all $i \in V$. Then for all $k \geq 0$,

$$\begin{cases} x_i^{k+1} = \arg\min_{x_i\in\mathcal{H}} f_i(x) + \frac{\gamma|\mathcal{N}_i|}{2}\|x_i - x_i^k - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} x_j^k + \frac{1}{\gamma|\mathcal{N}_i|}\alpha_i\|^2 + \frac{\gamma|\mathcal{N}_i|}{2}\|x_i\|^2 \\ \alpha_i^{k+1} = \alpha_i^k + \gamma\left(|\mathcal{N}_i|x_i^{k+1} - \sum_{j\in\mathcal{N}_i} x_j^{k+1}\right). \end{cases} \tag{103}$$

Equation (103) is truly distributed because each node $i \in V$ only requires information from its local neighbors at each iteration.

In [36], linear convergence is shown for this algorithm provided that f_i are strongly convex and ∇f_i are Lipschitz. For general convex functions, we can deduce the nonergodic rates from Theorem 15

$$\left|\sum_{i=1}^{m} f_i(x_i^k) - f(x^*)\right| = o\left(\frac{1}{\sqrt{k+1}}\right) \qquad \text{and} \qquad \sum_{\substack{i\in V\\ j\in N_i}} \|x_i^k - z_{ij}^k\|^2 + \sum_{\substack{i\in V\\ i\in N_j}} \|x_j^k - z_{ij}^k\|^2 = o\left(\frac{1}{k+1}\right),$$

and the ergodic rates from Theorem 14

$$\left|\sum_{i=1}^{m} f_i(\overline{x}_i^k) - f(x^*)\right| = O\left(\frac{1}{k+1}\right) \qquad \text{and} \qquad \sum_{\substack{i\in V\\ j\in N_i}} \|\overline{x}_i^k - \overline{z}_{ij}^k\|^2 + \sum_{\substack{i\in V\\ i\in N_j}} \|\overline{x}_j^k - \overline{z}_{ij}^k\|^2 = O\left(\frac{1}{(k+1)^2}\right).$$

These convergence rates are new and complement the linear convergence results in [36]. In addition, they complement the similar ergodic rate derived in [38] for a different distributed splitting.

10 Conclusion

In this paper, we provided a comprehensive convergence rate analysis of the FPR and objective error of several splitting algorithms under general convexity assumptions. We showed that the convergence rates are essentially optimal in all cases. All results follow from some combination of a lemma that deduces convergence rates of summable monotonic sequences (Lemma 3), a simple diagram (Figure 1), and fundamental inequalities (Propositions 4 and 5) that relate the FPR to the objective error of the relaxed PRS algorithm. The most important open question is whether and how the rates we derived will improve when we enforce stronger assumptions, such as Lipschitz differentiability and/or strong convexity, on f and g. This will be the subject of future work.

Acknowledgements D. Davis' work is partially supported by NSF GRFP grant DGE-0707424. W. Yin's work is partially supported by NSF grants DMS-0748839 and DMS-1317602.

References

1. Bauschke, H.H., Bello Cruz, J.Y., Nghia, T.T.A., Phan, H.M., Wang, X.: The rate of linear convergence of the Douglas-Rachford algorithm for subspaces is the cosine of the friedrichs angle. Journal of Approximation Theory **185**(0), 63–79 (2014). DOI http://dx.doi.org/10.1016/j.jat.2014.06.002 6, 1, 6.2
2. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. Springer (2011) 2, 1.3, 1.5, 3.3, 3.3, 3.4, 4.2, 5.1, 6.1.1, 7.3, 8, 8.2, 8.2
3. Bauschke, H.H., Deutsch, F., Hundal, H.: Characterizing arbitrarily slow convergence in the method of alternating projections. International Transactions in Operational Research **16**(4), 413–425 (2009) 9.1
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009) 1.1, 3.3
5. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Optimization for Machine Learning **2010**, 1–38 1.2
6. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and distributed computation: numerical methods. Prentice-Hall, Inc. (1989) 1
7. Boţ, R.I., Hendrich, C.: Convergence analysis for a primal-dual monotone+ skew splitting algorithm with applications to total variation minimization. Journal of Mathematical Imaging and Vision pp. 1–18 1.1
8. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**(1), 1–122 (2011) 1, 1, 9.2
9. Brézis, H., Lions, P.: Produits infinis de résolvantes. Israel Journal of Mathematics **29**(4), 329–345 (1978) 3.1.1, 3.2, 3.3, 3.3, 6.1.1, 7.4
10. Briceño-Arias, L.M.: Forward-Douglas-Rachford splitting and forward-partial inverse method for solving monotone inclusions. arXiv preprint arXiv:1212.5942 (2012) 6.1.1, 7.3
11. Browder, F.E., Petryshyn, W.: The solution by iteration of nonlinear functional equations in banach spaces. Bulletin of the American Mathematical Society **72**(3), 571–575 (1966) 3
12. Chambolle, A., 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) 1.1, 6.1.1
13. Chen, Y., Lan, G., Ouyang, Y.: Optimal primal-dual methods for a class of saddle point problems. arXiv preprint arXiv:1309.5548 (2013) 1.1
14. Combettes, P.L.: Quasi-Fejérian analysis of some optimization algorithms. Studies in Computational Mathematics **8**, 115–152 (2001) 3.5
15. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization **53**(5-6), 475–504 (2004) 3, 3.1
16. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-point algorithms for inverse problems in science and engineering, pp. 185–212. Springer (2011) 1
17. Cominetti, R., Soto, J.A., Vaisman, J.: On the rate of convergence of Krasnosel'skiĭ-Mann iterations and their connection with sums of bernoullis. Israel Journal of Mathematics pp. 1–16 1.1, 3, 3.1.1
18. 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) 6.1.1
19. Deng, W., Lai, M.J., Yin, W.: On the $o(1/k)$ convergence and parallelization of the alternating direction method of multipliers. arXiv preprint arXiv:1312.3040 (2013) 1.1, 2
20. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Mathematical Programming **55**(1-3), 293–318 (1992) 3.2, 6.1.1
21. Franchetti, C., Light, W.: On the von Neumann alternating algorithm in hilbert space. Journal of mathematical analysis and applications **114**(2), 305–314 (1986) 6.2
22. Gabay, D.: Chapter ix applications of the method of multipliers to variational inequalities. Studies in mathematics and its applications **15**, 299–331 (1983) 8, 8
23. Gabay, D., 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) 1
24. Glowinski, R., Marrocco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet nonlinéaires. Rev. Francaise dAut. Inf. Rech. Oper **R-2**, 41–76 (1975) 1
25. Goldstein, T., Osher, S.: The split Bregman method for l1-regularized problems. SIAM Journal on Imaging Sciences **2**(2), 323–343 (2009) 1
26. He, B., Yuan, X.: On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. Tech. rep. (2012) 1.1
27. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM Journal on Numerical Analysis **50**(2), 700–709 (2012) 8.3
28. Knopp, K.: Infinite sequences and series. Courier Dover Publications (1956) 2

29. Krasnosel'skiĭ, M.A.: Two remarks on the method of successive approximations. Uspekhi Matematicheskikh Nauk **10**(1), 123–127 (1955) 1.4
30. Liang, J., Fadili, J., Peyré, G.: Convergence rates with inexact nonexpansive operators. arXiv preprint arXiv:1404.4837 (2014) 1.1, 3, 3.1.1
31. Mann, W.R.: Mean value methods in iteration. Proceedings of the American Mathematical Society **4**(3), 506–510 (1953) 1.4
32. Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM Journal on Optimization **23**(1), 475–507 (2013) 1.1
33. Nemirovsky, A., Yudin, D.: Problem complexity and method efficiency in optimization. 1983 1.1
34. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer (2004) 3.3
35. Schizas, I.D., Ribeiro, A., Giannakis, G.B.: Consensus in ad hoc wsns with noisy linkspart i: Distributed estimation of deterministic signals. Signal Processing, IEEE Transactions on **56**(1), 350–364 (2008) 9.3
36. Shi, W., Ling, Q., Yuan, K., Wu, G., Yin, W.: On the linear convergence of the ADMM in decentralized consensus optimization. arXiv preprint arXiv:1307.5561 (2013) 1, 9.3, 9.3
37. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Advances in Computational Mathematics **38**(3), 667–681 (2013) 6.1.1
38. Wei, E., Ozdaglar, A.: Distributed alternating direction method of multipliers. In: Decision and Control (CDC), 2012 IEEE 51st Annual Conference on, pp. 5445–5450. IEEE (2012) 1, 1.1, 9.3