CONVERGENCE RATE ANALYSIS OF THE FORWARD-DOUGLAS-RACHFORD SPLITTING SCHEME*

DAMEK DAVIS†

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

In this paper, we analyze the convergence rate of the forward-Douglas-Rachford splitting (FDRS) algorithm, which is a generalization of the forward-backward splitting (FBS) and Douglas-Rachford splitting (DRS) algorithms. Under general convexity assumptions, we derive the ergodic and nonergodic convergence rates of the FDRS algorithm, and show that these rates are the best possible. Under Lipschitz differentiability assumptions, we show that the best iterate of FDRS converges as quickly as the last iterate of the FBS algorithm. Under strong convexity assumptions, we derive convergence rates for a sequence that strongly converges to a minimizer. Under strong convexity and Lipschitz differentiability assumptions, we show that FDRS converges linearly. We also provide examples where the objective is strongly convex, yet FDRS converges arbitrarily slowly. Finally, we relate the FDRS algorithm to a primal-dual forward-backward splitting scheme and clarify its place among existing splitting methods. Our results show that the FDRS algorithm automatically adapts to the regularity of the objective functions and achieves rates that improve upon the sharp worst case rates that hold in the absence of smoothness and strong convexity.

Key words. forward-Douglas-Rachford splitting, Douglas-Rachford splitting, forward-backward splitting, generalized forward-backward splitting, fixed-point algorithm, primal-dual algorithm

AMS subject classifications. 47H05, 65K05, 65K15, 90C25

1. Introduction.

Operator-splitting schemes are algorithms for splitting complicated problems arising in PDE, monotone inclusions, optimization, and control into many simpler subproblems. The achieved decomposition can give rise to inherently parallel and, in some cases, distributed algorithms. These characteristics are particularly desirable for large-scale problems that arise in machine learning, finance, control, image processing, and PDE [5].

In optimization, the Douglas-Rachford splitting (DRS) algorithm [21] minimizes sums of (possibly) nonsmooth functions $f, g : \mathcal{H} \to (-\infty, \infty]$ on a Hilbert space $\mathcal{H}$:

$$\underset{x\in\mathcal{H}}{\text{minimize}}\, f(x) + g(x). \tag{1.1}$$

During each step of the algorithm, DRS applies the proximal operator, which is the basic subproblem in nonsmooth minimization, to f and g individually rather than to the sum $f+g$. Thus, the key assumption in DRS is that f and g are easy to minimize *independently*, but the sum $f+g$ is difficult to minimize. We note that many complex objectives arising in machine learning [5] and signal processing [11] are the sum of nonsmooth terms with simple or closed-form proximal operators.

The forward-backward splitting (FBS) algorithm [23] is another technique for solving (1.1) when g is known to be *smooth*. In this case, the proximal operator of g is never evaluated. Instead, FBS combines gradient (forward) steps with respect to g

*This work is partially supported by grants NSF DGE-0707424 (graduate research fellowship program) and NSF DMS-1317602.

†Department of Mathematics, University of California, Los Angeles Los Angeles, CA 90025, USA damek@math.ucla.edu

and proximal (backward) steps with respect to f. FBS is especially useful when the proximal operator of g is complex and its gradient is simple to compute.

Recently, the forward-Douglas-Rachford splitting (FDRS) algorithm [7] was proposed to combine DRS and FBS and extend their applicability (see Algorithm 1). More specifically, let $V \subseteq \mathcal{H}$ be a *closed vector space* and suppose g is smooth. Then FDRS applies to the following constrained problem:

$$\underset{x \in V}{\text{minimize}}\, f(x) + g(x). \tag{1.2}$$

Throughout the course of the algorithm, the proximal operator of f, the gradient of g, and the projection operator onto V are all employed separately.

The FDRS algorithm can also apply to affinely constrained problems. Indeed, if $V = V_0 + b$ for a closed vector subspace $V_0 \subseteq \mathcal{H}$ and a vector $b \in \mathcal{H}$, then Problem (1.2) can be reformulated as

$$\underset{x \in V_0}{\text{minimize}}\, f(x+b) + g(x+b). \tag{1.3}$$

For simplicity, we only consider linearly constrained problems.

The FDRS algorithm is a generalization of the generalized forward-backward splitting (GFBS) algorithm [24], which solves the problem $\text{minimize}_{x \in \mathcal{H}} \sum_{i=1}^{n} f_i(x) + g(x)$ where $f_i : \mathcal{H} \to (-\infty, \infty]$ are closed, proper, convex and (possibly) nonsmooth. In the GFBS algorithm, the proximal mapping of each function f_i is evaluated *in parallel*. We note that GFBS can be derived as an application of FDRS to the equivalent problem:

$$\min_{\substack{(x_1, x_2, \ldots, x_n) \in \mathcal{H}^n \\ x_1 = x_2 = \cdots = x_n}} \sum_{i=1}^{n} f_i(x_i) + g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right). \tag{1.4}$$

In this case, the vector space $V = \{(x, \ldots, x) \in \mathcal{H}^n \mid x \in \mathcal{H}\}$ is the diagonal set of $\mathcal{H}^n$ and the function f is separable in the components of $(x_1, \cdots, x_n)$.

The FDRS algorithm is the only primal operator-splitting method capable of using all structure in Equation (1.2). In order to achieve good practical performance, the other primal splitting methods require stringent assumptions on f, g, and V. Primal DRS cannot use the smooth structure of g, so the proximal operator of g must be simple. On the other hand, primal FBS and forward-backward-forward splitting (FBFS) [25] cannot separate the coupled nonsmooth structure of f and V, so minimizing $f(x)$ subject to $x \in V$ must be simple. In contrast, FDRS achieves good practical performance if it is simple to minimize f, evaluate ∇g, and project onto V.

Modern primal-dual splitting methods [8, 18, 13, 26, 6, 19] can also decompose problem (1.2), but they introduce extra variables and are, thus, less memory efficient. It is unclear whether FDRS will perform better than primal-dual methods when memory is not a concern. However, it is easier to choose algorithm parameters for FDRS and, hence, it can be more convenient to use in practice.

Application: constrained quadratic programming and support vector machines. Let d and m be natural numbers. Suppose that $Q \in \mathbf{R}^{d \times d}$ is a symmetric positive semi-definite matrix, $c \in \mathbf{R}^d$ is a vector, $C \subseteq \mathbf{R}^d$ is a constraint set, $A \in \mathbf{R}^{m \times d}$

is a linear map, and $b \in \mathbf{R}^m$ is a vector. Consider the problem:

$$\begin{aligned} \underset{x\in\mathbf{R}^d}{\text{minimize}}\ & \frac{1}{2}\langle Qx, x\rangle + \langle c, x\rangle \\ \text{subject to: } & x \in C \\ & Ax = b. \end{aligned} \tag{1.5}$$

Problem (1.5) arises in the dual form soft-margin kernelized support vector machine classifier [14] in which C is a box constraint, b is 0, and A has rank one. Note that by the argument in (1.3), we can always assume that $b = 0$.

Define the smooth function $g(x) := (1/2)\langle Qx, x\rangle + \langle c, x\rangle$, the indicator function $f(x) := \chi_C(x)$ (which is 0 on C and ∞ elsewhere), and the vector space $V := \{x \in \mathbf{R}^d \mid Ax = 0\}$. With this notation, (1.5) is in the form (1.2) and, thus, FDRS can be applied. This splitting is nice because $\nabla g(x) = Qx + c$ is simple whereas the proximal operator of g requires a matrix inversion $\mathbf{prox}_{\gamma g} = (I_{\mathbf{R}^d} + \gamma Q)^{-1} \circ (I_{\mathbf{R}^d} - \gamma c)$, which is expensive for large-scale problems.

1.1. Goals, challenges, and approaches. This work seeks to characterize the convergence rate of the FDRS algorithm applied to Problem (1.2). Recently, [16] has shown that the sharp convergence rate of the fixed-point residual (FPR) (see Equation (1.21)) of the FDRS algorithm is $o(1/(k+1))$. To the best of our knowledge, nothing is else is known about the convergence rate of FDRS. Furthermore, it is unclear how the FDRS algorithm relates to other algorithms. We seek to fill this gap.

The techniques used in this paper are based on [15, 16, 17]. These techniques are quite different from those used in classical objective error convergence rate analysis. The classical techniques do not apply because the FDRS algorithm is driven by the fixed-point iteration of a nonexpansive operator, not by the minimization of a model function. Thus, we must explicitly use the properties of nonexpansive operators in order to derive convergence rates for the objective error.

We summarize our contributions and techniques as follows:

(i) We analyze the objective error convergence rates (Theorems 3.1 and 3.4) of the FDRS algorithm under general convexity assumptions. We show that FDRS is, in the worst case, *nearly as slow as the subgradient method* yet *nearly as fast as the proximal point algorithm (PPA) in the ergodic sense.* Our nonergodic rates are shown by relating the objective error to the FPR through a *fundamental inequality.* We also show that the derived rates are sharp through counterexamples (Remarks 4 and 5).

(ii) We show that if f or g is strongly convex, then a natural sequence of points converges strongly to a minimizer. Furthermore, the *best iterate* converges with rate $o(1/(k+1))$, the *ergodic iterate* converges with rate $O(1/(k+1))$, and the *nonergodic iterate* converges with rate $o(1/\sqrt{k+1})$. The results follow by showing that a certain sequence of squared norms is summable. We also show that some of the derived rates are sharp by constructing a novel counterexample (Theorem 6.6).

(iii) We show that if f is differentiable and ∇f is Lipschitz, then the *best iterate* of the FDRS algorithm has objective error of order $o(1/(k+1))$ (Theorem 5.2). This rate is an improvement over the sharp $o(1/\sqrt{k+1})$ convergence rate for nonsmooth f. The result follows by showing that the objective error is summable.

(iv) We establish scenarios under which FDRS converges linearly (Theorem 6.1) and show that linear convergence is impossible under other scenarios (Theorem 6.6).

(v) We show that even if f and g are strongly convex, the FDRS algorithm can converge *arbitrarily slowly* (Theorem 6.5).

(vi) We show that the FDRS algorithm is the limiting case of a recently developed primal-dual forward-backward splitting algorithm (Section 7) and, thus, clarify how FDRS relates to existing algorithms.

Our analysis builds on the techniques and results of [7, 16, 17]. The rest of this section contains a brief review of these results.

1.2. Notation and facts. Most of the definitions and notation that we use in this paper are standard and can be found in [3]. Throughout this paper, we use $\mathcal{H}$ to denote (a possibly infinite dimensional) Hilbert space. In fixed-point iterations, $(\lambda_j)_{j\geq 0} \subset \mathbf{R}_+$ will denote a sequence of relaxation parameters, and

$$\Lambda_k := \sum_{i=0}^{k} \lambda_i \tag{1.6}$$

is its kth partial sum.

For any subset $C \subseteq \mathcal{H}$, we define the distance function:

$$d_C(x) := \inf_{y \in C} \|x - y\|. \tag{1.7}$$

In addition, we define the indicator function $\chi_C : \mathcal{H} \to \{0, \infty\}$ of C: for all $x \in C$ and $y \in \mathcal{H}\backslash C$, we have $\chi_C(x) = 0$ and $\chi_C(y) = \infty$.

Given a closed, proper, and convex function $f : \mathcal{H} \to (-\infty, \infty]$, the set $\partial f(x) = \{p \in \mathcal{H} \mid \text{for all } y \in \mathcal{H}, f(y) \geq f(x) + \langle y - x, p\rangle\}$ denotes its subdifferential at x and

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

denotes a subgradient. (This notation was used in [4, Eq. (1.10)].) If f is Gâteaux differentiable at $x \in \mathcal{H}$, we have $\partial f(x) = \{\nabla f(x)\}$ [3, Proposition 17.26].

Let $I_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ be the identity map on $\mathcal{H}$. For any $x \in \mathcal{H}$ and $\gamma \in \mathbf{R}_{++}$, we let

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

which are known as the *proximal* and *reflection* operators, respectively.

The subdifferential of the indicator function χ_V where $V \subseteq \mathcal{H}$ is a closed vector subspace is defined as follows: for all $x \in \mathcal{H}$,

$$\partial \chi_V(x) = \begin{cases} V^\perp & \text{if } x \in \mathcal{H}; \\ \emptyset & \text{otherwise} \end{cases} \tag{1.8}$$

where $V^\perp$ is the orthogonal complement of V. Evidently, if $P_V(\cdot) = \arg\min_{y \in V} \|y - \cdot\|^2$ is the projection onto V, then

$$\mathbf{prox}_{\gamma\chi_V} = P_V \qquad \text{and} \qquad \mathbf{refl}_{\gamma\chi_V} = 2P_V - I_{\mathcal{H}} = P_V - P_{V^\perp},$$

and these operators are independent of γ.

Let $\lambda > 0$, let $L \geq 0$, and let $T : \mathcal{H} \to \mathcal{H}$ be a map. The map T is called *L-Lipschitz* continuous if $\|Tx - Ty\| \leq L\|x - y\|$ for all $x, y \in \mathcal{H}$. The map T is called *nonexpansive* if it is 1-Lipschitz. We also use the notation:

$$T_\lambda := (1 - \lambda) I_{\mathcal{H}} + \lambda T. \tag{1.9}$$

If $\lambda \in (0,1)$ and T is nonexpansive, then T_λ is called λ-*averaged* [3, Definition 4.23].

We call the following identity the *cosine rule*:

$$\|y-z\|^2 + 2\langle y-x, z-x\rangle = \|y-x\|^2 + \|z-x\|^2, \quad \forall x,y,z \in \mathcal{H}. \tag{1.10}$$

Young's inequality is the following: for all $a, b \geq 0$ and $\varepsilon > 0$, we have

$$ab \leq a^2/(2\varepsilon) + \varepsilon b^2/2. \tag{1.11}$$

1.3. Assumptions.

ASSUMPTION 1 (Convexity). *f and g are closed, proper, and convex.*

We also assume the existence of a particular solution to (1.2)

ASSUMPTION 2 (Solution existence). $\operatorname{zer}(\partial f + \nabla g + \partial \chi_V) \neq \emptyset$

Finally we assume that ∇g is sufficiently nice.

ASSUMPTION 3 (Differentiability). *The function g is differentiable, ∇g is $(1/\beta)$-Lipschitz, and $P_V \circ \nabla g \circ P_V$ is $(1/\beta_V)$-Lipschitz.*

1.4. The FDRS algorithm. FDRS is summarized in Algorithm 1.

Algorithm 1: Relaxed Forward-Douglas-Rachford splitting (relaxed FDRS)

input : $z^0 \in \mathcal{H},\ \gamma \in (0,\infty),\ (\lambda_j)_{j\geq 0} \in (0,\infty)$
for $k = 0,\ 1, \ldots$ **do**
$\quad z^{k+1} = (1-\lambda_k)z^k + \lambda_k \left(\frac{1}{2}I_{\mathcal{H}} + \frac{1}{2}\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V}\right) \circ (I - \gamma P_V \circ \nabla g \circ P_V)(z^k);$

For now, we do not specify the stepsize parameters. See section 1.6 for choices that ensure convergence and, see Lemma 2.1 and Figure 2.1 for intuition.

Evidently, Algorithm 1 has the form: for all $k \geq 0$, $z^{k+1} = (T_{\mathrm{FDRS}})_{\lambda_k}(z^k)$ where

$$T_{\mathrm{FDRS}} := \left(\frac{1}{2}I_{\mathcal{H}} + \frac{1}{2}\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V}\right) \circ (I_{\mathcal{H}} - \gamma P_V \circ \nabla g \circ P_V). \tag{1.12}$$

Because T_{FDRS} is nonexpansive (Part 7 of Proposition 1.1), it follows that the FDRS algorithm is a special case of the Krasnosel'skiĭ-Mann (KM) iteration [20, 22, 10].

By choosing particular f, g and V, we recover several other splitting algorithms:

$$\begin{aligned}
\text{DRS: } (g \equiv 0) \quad & z^{k+1} = (1-\lambda_k)z^k + \lambda_k \left(\frac{1}{2}I_{\mathcal{H}} + \frac{1}{2}\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V}\right)(z^k); \\
\text{FBS: } (V = \mathcal{H}) \quad & z^{k+1} = (1-\lambda_k)z^k + \lambda_k \mathbf{prox}_{\gamma f} \circ (I_{\mathcal{H}} - \gamma \nabla g)(z^k); \\
\text{FBS: } (f \equiv 0) \quad & z^{k+1} = (1-\lambda_k)z^k + \lambda_k P_V \circ (z - \gamma P_V \circ \nabla g \circ P_V)(z^k).
\end{aligned}$$

For general f, g and V, the primal DRS and FBS algorithms are not capable splitting Problem (1.2) in the same way as (1.12). Indeed, the DRS algorithm cannot use the smooth structure of g, and the FBS algorithm requires the evaluation of $\mathbf{prox}_{\gamma(f+\chi_V)}(\cdot) = \arg\min_{x\in V}\left(f(x) + (1/2\gamma)\|x - \cdot\|^2\right)$. The FDRS algorithm eliminates these difficult problems and replaces them with (possibly) more tractable ones.

1.5. Proximal, averaged, and FDRS operators. We briefly review some operator-theoretic properties.

PROPOSITION 1.1. *Let $\lambda > 0$, let $\gamma > 0$, let $\alpha > 0$, and let $f : \mathcal{H} \to (-\infty, \infty]$ be closed, proper, and convex.*

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

2. Optimality conditions of $\mathbf{prox}_{\chi_V}$: *Let $x \in \mathcal{H}$. Then $x^+ = \mathbf{prox}_{\gamma\chi_V}(x)$ if, and only if, $\widetilde{\nabla}\chi_V(x^+) := (1/\gamma)(x - x^+) \in \partial\chi_V(x^+)$. Also, $\gamma\widetilde{\nabla}\chi_V(x^+) = P_{V^\perp}x \in V^\perp$.*

3. Averaged operator contraction property: *A map $T : \mathcal{H} \to \mathcal{H}$ is α-averaged (see (1.9)) if, and only if, for all $x, y \in \mathcal{H}$,*

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1-\alpha}{\alpha}\|(I_{\mathcal{H}} - T)x - (I_{\mathcal{H}} - T)y\|^2. \tag{1.13}$$

4. Composition of averaged operators: *Let $\alpha_1, \alpha_2 \in (0, 1)$. Suppose $T_1 : \mathcal{H} \to \mathcal{H}$ and $T_2 : \mathcal{H} \to \mathcal{H}$ are α_1 and α_2-averaged operators, respectively. Then for all $x, y \in \mathcal{H}$, the map $T_1 \circ T_2 : \mathcal{H} \to \mathcal{H}$ is averaged with parameter*

$$\alpha_{1,2} := \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2} \in (0, 1) \tag{1.14}$$

5. Wider relaxations: *A map $T : \mathcal{H} \to \mathcal{H}$ is α-averaged if, and only if, T_λ (see (1.9)) is $\lambda\alpha$-averaged for all $\lambda \in (0, 1/\alpha)$.*

6. Proximal operators are (1/2)-averaged: *The operator $\mathbf{prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ is (1/2)-averaged and, hence, the operator $\mathbf{refl}_{\gamma f} = 2\mathbf{prox}_{\gamma f} - I_{\mathcal{H}}$ is nonexpansive.*

7. Averaged property of the FDRS operator: *Suppose that $\gamma \in (0, 2\beta)$. Then the operator T_{FDRS} (see (1.12)) is $\alpha_{\mathrm{FDRS}} := 2\beta/(4\beta - \gamma)$ averaged.*

Proof. Parts 1, 2, 3, 5, and 6 can be found in [3]. Part 4 can be found in [12]. Part 7 follows from two facts: The operator $((1/2)I_{\mathcal{H}} + (1/2)\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V})$ is (1/2)-averaged by Part 6, and $I - \gamma P_V \circ \nabla g \circ P_V$ is $(\gamma/2\beta)$-averaged by [7, Proposition 4.1 (ii)]. Thus, Part 4 proves Part 7. ∎

Remark 1. *Later we require $(\lambda_j)_{j\geq 0} \subseteq (0, 1/\alpha_{\mathrm{FDRS}})$ so we hope that α_{FDRS} is small. Note that the expression for α_{FDRS} is new and improves upon the previous constant:* $\max\{2/3, 2\gamma/(\gamma + 2\beta)\}$. *See also [12, Remark 2.7 (i)].*

The proof of the following Proposition is essentially contained in [12, Theorem 2.4]. We reproduce it in Appendix B.1 in order to derive a bound. The reader should note the following inequality before reading the proof.

Remark 2. *Let $\varepsilon \in (0, 1)$. Then it is easy to show that*

$$\lambda \leq \frac{(1-\varepsilon)(1+\varepsilon\alpha_{1,2})}{\alpha_{1,2}} \implies \lambda \leq 1/\alpha_{1,2} - \varepsilon^2 \text{ and } \lambda - 1 \leq \frac{1 - \alpha_{1,2}\lambda}{\alpha_{1,2}\varepsilon}. \tag{1.15}$$

Proposition 1.2. *Let $\alpha_1, \alpha_2 \in (0, 1)$. Suppose that $T_1 : \mathcal{H} \to \mathcal{H}$ and $T_2 : \mathcal{H} \to \mathcal{H}$ are α_1 and α_2-averaged operators, respectively, and that z^* is a fixed-point of $T_1 \circ T_2$. Define $\alpha_{1,2} \in (0, 1)$ as in (1.14). Let $z^0 \in \mathcal{H}$, let $\varepsilon \in (0, 1)$, and consider a sequence $(\lambda_j)_{j\geq 0} \subseteq (0, (1-\varepsilon)(1+\varepsilon\alpha_{1,2})/\alpha_{1,2})$. Let $(z^j)_{j\geq 0}$ be generated by the following iteration: for all $k \geq 0$, let $z^{k+1} = (T_1 \circ T_2)_{\lambda_k}(z^k)$. Then*

$$\sum_{i=0}^{\infty} \lambda_i \|(I_{\mathcal{H}} - T_2)(z^i) - (I_{\mathcal{H}} - T_2)(z^*)\|^2 \leq \frac{\alpha_2(1 + 1/\varepsilon)\|z^0 - z^*\|^2}{1 - \alpha_2}.$$

1.6. Convergence properties of FDRS. The paper [7] assumed the stepsize constraint $\gamma \in (0, 2\beta)$ in order to guarantee convergence of Algorithm 1. We now show that the parameter γ can (possibly) be increased beyond 2β, which can result in faster practical performance. The proof follows by constructing a new Lipschitz differentiable function h so that the triple (f, h, V) generates the same FDRS operator, T_{FDRS}, as (f, g, V). This result was not included in [7].

LEMMA 1.3. *Define a function*

$$h := g \circ P_V. \tag{1.16}$$

Then the FDRS operator associated to (f, g, V) is identical to the FDRS operator associated to (f, h, V). Let $1/\beta_V$ be the Lipschitz constant of ∇h. Then $\beta_V \geq \beta$. In addition, let $\gamma \in (0, 2\beta_V)$. Then T_{FDRS} is α_{FDRS}^V-averaged where

$$\alpha_{\mathrm{FDRS}}^V := \frac{2\beta_V}{4\beta_V - \gamma}. \tag{1.17}$$

Proof. The averaged property of T_{FDRS} and the equivalence of FDRS operators follows from Part 7 of Proposition 1.1. The bound $\beta_V \geq \beta$ follows because for all $x, y \in \mathcal{H}$,

$$\begin{aligned}\|\nabla h(x) - \nabla h(y)\| = \|P_V \circ g \circ P_V(x) - P_V \circ g \circ P_V(y)\| &\leq \|\nabla g \circ P_V(x) - \nabla g \circ P_V(y)\| \\ &\leq (1/\beta)\|P_V(x) - P_V(y)\| \leq (1/\beta)\|x - y\| \qquad \square\end{aligned}$$

There are cases where β_V is significantly larger than β. For instance, in the quadratic programming example in (1.5), β is the reciprocal of the Lipschitz constant of Q, which is the maximal eigenvalue $\lambda_{\max}(Q)$ of Q. On the other hand, the gradient $\nabla h = P_V \circ Q \circ P_V$ has rank at most $d - \mathrm{rank}(A)$. Thus, unless the eigenvectors of Q with eigenvalue $\lambda_{\max}(Q)$ lie in the $(d - \mathrm{rank}(A))$-dimensional space V, the constant $\beta_V = 1/\lambda_{\max}(P_V \circ Q \circ P_V)$ is larger than $\beta = 1/\lambda_{\max}(Q)$. See Appendix A for experimental evidence.

Most of our results do not require that $(z^j)_{j\geq 0}$ converges. However, for completeness we include the following weak convergence result.

PROPOSITION 1.4. *Let $\gamma \in (0, 2\beta_V)$, let $(\lambda_j)_{j\geq 0} \subseteq (0, 1/\alpha_{\mathrm{FDRS}}^V)$, and suppose that $\sum_{i=0}^{\infty} \lambda_i(1 - \lambda_i \alpha_{\mathrm{FDRS}}^V) = \infty$. Then $(z^j)_{j\geq 0}$ (from Algorithm 1) weakly converges to a fixed-point of T_{FDRS}.*

Proof. Apply [7, Proposition 3.1] with the new averaged parameter α_{FDRS}^V. $\square$

The following theorem recalls several results on convergence rates for the iteration of averaged operators [16]. In addition, we show that $(\lambda_j \|\nabla h(z^j) - \nabla h(z^*)\|^2)_{j\geq 0}$ is a summable sequence [7] whenever $(\lambda_j)_{j\geq 0}$ is chosen properly.

THEOREM 1.5. *Suppose that $(z^j)_{j\geq 0}$ is generated by Algorithm 1 with $\gamma \in (0, 2\beta_V)$ and $(\lambda_j)_{j\geq 0} \subseteq (0, 1/\alpha_{\mathrm{FDRS}}^V)$, and let z^* be a fixed-point of T_{FDRS}. Then*

1. Fejér monotonicity: *the sequence $(\|z^j - z^*\|^2)_{j\geq 0}$ is nonincreasing. In addition, for all $z \in \mathcal{H}$ and $\lambda \in (0, 1/\alpha_{\mathrm{FDRS}}^V)$, we have $\|(T_{\mathrm{FDRS}})_\lambda z - z^*\| \leq \|z - z^*\|$.*
2. Summable fixed-point residual: *The sum is finite:*

$$\sum_{i=0}^{\infty} \frac{1 - \lambda_i \alpha_{\mathrm{FDRS}}^V}{\lambda_i \alpha_{\mathrm{FDRS}}^V} \|z^{i+1} - z^i\|^2 \leq \|z^0 - z^*\|^2.$$

3. Convergence rates of fixed-point residual: *For all $k \geq 0$, let $\tau_k := (1 - \lambda_k \alpha_{\mathrm{FDRS}}^V)\lambda_k/\alpha_{\mathrm{FDRS}}^V$. Suppose that $\underline{\tau} := \inf_{j\geq 0} \tau_j > 0$. Then for $\lambda > 0$ and $k \geq 0$,*

$$\|(T_{\mathrm{FDRS}})_\lambda(z^k) - z^k\|^2 \leq \frac{\lambda^2 \|z^0 - z^*\|^2}{\underline{\tau}(k+1)} \quad \text{and} \quad \|(T_{\mathrm{FDRS}})_\lambda(z^k) - z^k\|^2 = o\left(\frac{1}{k+1}\right). \tag{1.18}$$

4. Gradient summability: *Let $\varepsilon \in (0,1)$ and suppose that*

$$(\lambda_j)_{j\geq 0} \subseteq \left(0, \frac{(1-\varepsilon)(1+\varepsilon\alpha_{\mathrm{FDRS}}^V)}{\alpha_{\mathrm{FDRS}}^V}\right). \tag{1.19}$$

Then the following gradient sum is finite:

$$\sum_{i=0}^{\infty} \lambda_i \|\nabla h(z^i) - \nabla h(z^*)\|^2 \leq \frac{(1+\varepsilon)}{\gamma\varepsilon(2\beta_V - \gamma)}\|z^0 - z^*\|^2. \tag{1.20}$$

Proof. Parts 1, 2, and 3 are a direct consequence of [16, Theorem 1] applied to the α_{FDRS}^V-averaged operator T_{FDRS}. Part 4 is a direct consequence of Proposition 1.2 applied to the $(1/2)$-averaged operator $T_1 := ((1/2)I_{\mathcal{H}} + (1/2)\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V})$ (see Part 6 of Proposition 1.1) and the $(\gamma/(2\beta_V))$-averaged operator $T_2 := I_{\mathcal{H}} - \gamma\nabla h$ (from the Baillon-Haddad Theorem [1] and [3, Proposition 4.33]). □

We call the following term the fixed-point residual (FPR):

$$\|T_{\mathrm{FDRS}} z^k - z^k\|^2 = \frac{1}{\lambda_k^2}\|z^{k+1} - z^k\|^2 \tag{1.21}$$

Remark 3. *Note that the convergence rate proved for $\|T_{\mathrm{FDRS}} z^k - z^k\|^2$ in* (1.18) *is sharp for the T_{FDRS} operator [16, Section 6.1.1].*

2. Subgradients and fundamental inequalities. In this section, we prove several algebraic identities of the FDRS algorithm. In addition, we prove a relationship between the FPR and the objective error (Propositions 2.4 and 2.5).

In first-order optimization algorithms, we only have access to (sub)gradients and function values. Consequently, the FPR is usually the squared norm of a linear combination of (sub)gradients of the objective functions. For example, the gradient descent algorithm for a smooth function f generates a sequence of iterates by using forward gradient steps: $z^{k+1} := z^k - \nabla f(z^k)$; the FPR is $\|z^{k+1} - z^k\|^2 = \|\nabla f(z^k)\|^2$.

In splitting algorithms, the FPR is more complex because the subgradients are generated via forward-gradient or proximal (backward) steps (see Part 1 of Proposition 1.1) at different points. Thus, unlike the gradient descent algorithm where the objective error $f(z^k) - f(x^*) \leq \langle z^k - x^*, \nabla f(x^k)\rangle$ can be bounded with the subgradient inequality, splitting algorithms for two or more functions can only bound the objective error when some or all of the functions are evaluated at separate points — unless a Lipschitz assumption is imposed. In order to use this Lipschitz assumption, we enforce consensus among the variables, which is why the FPR rate is useful.

2.1. A subgradient representation of FDRS. Figure 2.1 pictures one iteration of Algorithm 1: FDRS projects z onto V to get $x_h = z - \gamma\widetilde{\nabla}\chi_V(x_h)$. The reflection of z across V is $x_h - \gamma\widetilde{\nabla}\chi_V(x_h) = z - 2\gamma\widetilde{\nabla}\chi_V(x_h)$. Then FDRS takes a forward-gradient with respect to $\nabla h(x_h)$ from the reflected point $x_h - \gamma\widetilde{\nabla}\chi_V(x_h)$ and a proximal (backward) step with respect to f to get x_f. Finally, we move from x_f to $T_{\mathrm{FDRS}} z$ by traveling along the positive subgradient $\gamma\widetilde{\nabla}\chi_V(x_h)$.

The following lemma is proved in Appendix B.2.

Lemma 2.1 (FDRS identities). *Let $z \in \mathcal{H}$. Define points x_h and x_f:*

$$x_h := P_V z \qquad \text{and} \qquad x_f := \mathbf{prox}_{\gamma f} \circ \mathbf{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma\nabla h)(z). \tag{2.1}$$

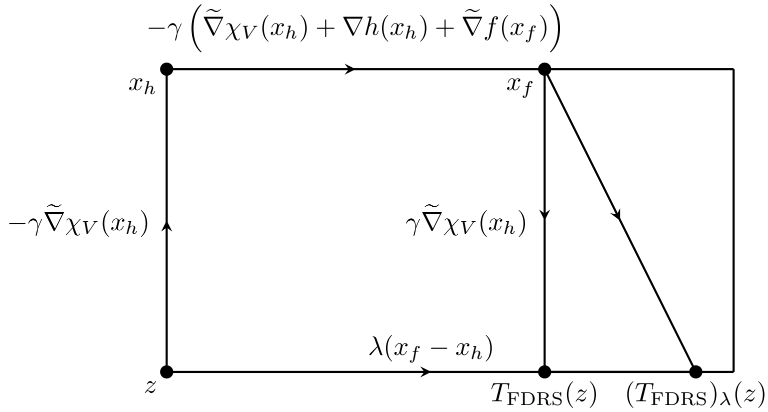

FIG. 2.1. *A single FDRS iteration, from z to $(T_{\mathrm{FDRS}})_\lambda(z)$ (see Lemma 2.1). Both occurrences of $\widetilde{\nabla}\chi_V(x_h)$ represent the same subgradient $(1/\gamma)P_{V^\perp}z = (1/\gamma)(z - x_h) \in V^\perp$.*

Then the identities hold

$$x_h = z - \gamma\widetilde{\nabla}\chi_V(x_h) \quad \text{and} \quad x_f = x_h - \gamma\left(\widetilde{\nabla}\chi_V(x_h) + \nabla h(x_h) + \widetilde{\nabla} f(x_f)\right) \tag{2.2}$$

where $\widetilde{\nabla}\chi_V(x_h) = (1/\gamma)P_{V^\perp}(z)$ and $\widetilde{\nabla} f(x_f)$ is uniquely defined by Part 1 of Proposition 1.1. In addition, each FDRS step has the following form:

$$(T_{\mathrm{FDRS}})_\lambda(z) - z = \lambda(x_f - x_h) = -\gamma\lambda\left(\widetilde{\nabla}\chi_V(x_h) + \nabla h(x_h) + \widetilde{\nabla} f(x_f)\right). \tag{2.3}$$

In particular, $T_{\mathrm{FDRS}}(z) = x_f + \gamma\widetilde{\nabla}\chi_V(x_h)$.

DEFINITION 2.2 (Ergodic iterates). *Let $(z^j)_{j\geq 0}$ be generated by Algorithm 1 and define $(x_h^j)_{j\geq 0}$ and $(x_f^j)_{j\geq 0}$ as in* (2.1) *(with $z = z^j$). Then define* ergodic iterates:

$$\overline{x}_h^k := \frac{1}{\Lambda_k}\sum_{i=0}^k \lambda_i x_h^i \qquad \text{and} \qquad \overline{x}_f^k := \frac{1}{\Lambda_k}\sum_{i=0}^k \lambda_i x_f^i \tag{2.4}$$

2.2. Optimality conditions of FDRS. The following lemma characterizes the zeros of $\partial f + \nabla h + \partial\chi_V$ in terms of the fixed-points of the FDRS operator. The intuition is the following: If z^* is a fixed-point of T_{FDRS}, then the base of the rectangle in Figure 2.1 has length zero. Thus, $x^* := x_h^* = x_f^*$, and if we travel around the perimeter of the rectangle, we will start and begin at z^*. This argument shows that $\gamma\widetilde{\nabla} f(x^*) + \gamma\nabla h(x^*) + \gamma\widetilde{\nabla}\chi_V(x^*) = 0$, i.e., $x^* \in \mathrm{zer}(\partial f + \nabla h + \partial\chi_V)$.

The following lemma is proved in Appendix B.3.

LEMMA 2.3 (FDRS optimality conditions). *The following set equality holds:*

$$\mathrm{zer}(\partial f + \nabla h + \partial\chi_V) = \{P_V z \mid z \in \mathcal{H}, T_{\mathrm{FDRS}} z = z\}$$

That is, if z^* is a fixed-point of T_{FDRS}, then $x^* := P_V z^* = x_h^* = x_f^*$ is a minimizer of (1.2), *and $z^* - x^* = P_{V^\perp}(z^*) = \gamma\widetilde{\nabla}\chi_V(x_h^*) \in \partial\chi_V(x^*)$.*

2.3. Fundamental inequalities. In this section, we prove two fundamental inequalities that relate the FPR (see (1.21)) to the objective error.

Throughout the rest of the paper, we use the following notation: The functions f and g are μ_f and μ_g-strongly convex, respectively, where we allow μ_f or μ_g to be zero (i.e., no strong convexity). In addition, we assume that f is $(1/\beta_f)$-Lipschitz differentiable, where we allow $\beta_f = 0$. If $\beta_f > 0$, then $\widetilde{\nabla} f = \nabla f$. With these assumptions, we get the following lower bounds [3, Theorem 18.15]:

$$\forall x, y \in \mathrm{dom}(\partial f) \qquad f(x) \geq f(y) + \langle x - y, \widetilde{\nabla} f(y)\rangle + S_f(x,y); \tag{2.5}$$

$$\forall x, y \in \mathcal{H} \qquad h(x) \geq h(y) + \langle x - y, \nabla h(y)\rangle + S_h(x,y); \tag{2.6}$$

where $\widetilde{\nabla} f(y) \in \partial f(y)$, and for any $x, y \in \mathcal{H}$,

$$S_f(x,y) := \begin{cases} \max\left\{\frac{\mu_f}{2}\|x-y\|^2, \frac{\beta_f}{2}\|\nabla f(x) - \nabla f(y)\|^2\right\} & \text{if } \beta_f > 0; \\ \frac{\mu_f}{2}\|x-y\|^2 & \text{otherwise;} \end{cases} \tag{2.7}$$

$$S_h(x,y) := \max\left\{\frac{\mu_g}{2}\|P_V x - P_V y\|^2, \frac{\beta_V}{2}\|\nabla h(x) - \nabla h(y)\|^2\right\}. \tag{2.8}$$

See Appendices B.4, B.5, and B.6 for the proofs of the following inequalities:

Proposition 2.4 (Upper fundamental inequality). *Let $z \in \mathcal{H}$, let $\lambda > 0$, and let $z^+ := (T_{\mathrm{FDRS}})_\lambda(z)$. Then for all $x \in V \cap \mathrm{dom}(\partial f)$, we have the following inequality:*

$$\begin{aligned} &2\gamma\lambda\left(f(x_f) + h(x_h) - f(x) - h(x) + S_f(x_f, x) + S_h(x_h, x)\right) \\ &\qquad \leq \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 + 2\gamma\langle \nabla h(x_h), z - z^+\rangle \end{aligned} \tag{2.9}$$

where x_f and x_h are defined as in Lemma 2.1.

Proposition 2.5 (Lower fundamental inequality). *Let $z^* \in \mathcal{H}$ be a fixed-point of T_{FDRS}, and let $x^* := P_V z^*$. Choose subgradients $\widetilde{\nabla} f(x^*) \in \partial f(x^*)$ and $\widetilde{\nabla}\chi_V(x^*) \in \partial\chi_V(x^*)$ with $\widetilde{\nabla} f(x^*) + \nabla h(x^*) + \widetilde{\nabla}\chi_V(x^*) = 0$ (see Lemma 2.3). Then for all $x_f \in \mathrm{dom}(f)$ and $x_h \in V$, we have*

$$f(x_f) + h(x_h) - f(x^*) - g(x^*) \geq \langle x_f - x_h, \widetilde{\nabla} f(x^*)\rangle + S_f(x_f, x^*) + S_h(x_h, x^*). \tag{2.10}$$

Corollary 2.6. *Let $z \in \mathcal{H}$, let $\lambda > 0$, and let $z^+ := (T_{\mathrm{FDRS}})_\lambda(z)$. Let $z^* \in \mathcal{H}$ be a fixed-point of T_{FDRS}, and let $x^* := P_V z^*$. Then with x_f and x_h from Lemma 2.1,*

$$\begin{aligned} 4\gamma\lambda(S_f(x_f, x^*) + S_h(x_h, x^*)) \leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 \\ + 2\gamma\langle \nabla h(x_h) - \nabla h(x^*), z - z^+\rangle. \end{aligned} \tag{2.11}$$

3. Objective convergence rates. In this section, we analyze the ergodic and nonergodic convergence rates of the FDRS algorithm applied to (1.2).

Throughout the rest of the paper, z^* will denote an arbitrary fixed-point of T_{FDRS}, and we define a minimizer of (1.2) using Lemma 2.3: $x^* := P_V z^*$.

All of our bounds will be produced on objective errors of the form:

$$f(x_f^k) + h(x_h^k) - f(x^*) - g(x^*) \qquad \text{and} \qquad f(x_h^k) + h(x_h^k) - f(x^*) - g(x^*). \tag{3.1}$$

The objective error on the left hand side of (3.1) can be negative. Thus, we bound its absolute value. In addition, we bound $\|x_f^k - x_h^k\|$. Because $x_h^k \in V$, the objective error on the right hand size of (3.1) is positive. Consequently, x_h^k is the natural point at which to measure the convergence rate. To derive such a bound, we assume f is Lipschitz. Note that in both cases, we have the identity $h(x_h^k) = (g \circ P_V)(x_h^k) = g(x_h^k)$.

3.1. Ergodic convergence rates. In this section, we analyze the ergodic convergence rate of the FDRS algorithm. The key idea is to use the telescoping property of the upper and lower fundamental inequalities, together with the summability of the difference of gradients shown in Part 4 of Theorem 1.5. See Section 1.2 for the distinction between ergodic and nonergodic convergence rates.

THEOREM 3.1 (Ergodic convergence of FDRS). *Let $\gamma \in (0, 2\beta_V)$, let $\varepsilon \in (0,1)$, and suppose that $(\lambda_j)_{j\geq 0}$ satisfies (1.19). Define $(\overline{x}_f^j)_{j\geq 0}$ and $(\overline{x}_h^j)_{j\geq 0}$ as in (2.4). Then we have the following convergence rate: for all $k \geq 0$,*

$$\begin{aligned}\frac{-2\|z^0 - z^*\|\|\widetilde{\nabla} f(x^*)\|}{\Lambda_k} &\leq f(\overline{x}_f^k) + h(\overline{x}_h^k) - f(x^*) - h(x^*)\\ &\leq \frac{\left(\|z^0 - z^*\| + 4\gamma\|\nabla h(x^*)\| + \frac{(1+\varepsilon)\gamma\|z^0 - z^*\|}{\varepsilon^3(2\beta_V - \gamma)}\right)\|z^0 - z^*\|}{2\gamma\Lambda_k}.\end{aligned}$$

In addition the following feasibility bound holds: $\|\overline{x}_f^k - \overline{x}_h^k\| \leq (2/\Lambda_k)\|z^0 - z^*\|$.

Proof. Fix $k \geq 0$. The feasibility bound follows from Part 1 of Theorem 1.5:

$$\begin{aligned}\|\overline{x}_f^k - \overline{x}_h^k\| = \left\|\frac{1}{\Lambda_k}\sum_{i=0}^{k}\left(z^{i+1} - z^i\right)\right\| = \frac{1}{\Lambda_k}\|z^0 - z^{k+1}\| &\leq \frac{1}{\Lambda_k}\left(\|z^0 - z^*\| + \|z^* - z^{k+1}\|\right)\\ &\leq \frac{2}{\Lambda_k}\|z^0 - z^*\|. \qquad (3.2)\end{aligned}$$

Now we prove the objective convergence rates. For all $k \geq 0$, let $\eta_k := 2/\lambda_k - 1$. Note that $\eta_k > 0$ by (1.15) because we have $\lambda_k < 1/\alpha_{\mathrm{FDRS}}^V - \varepsilon^2 \leq 2 - \varepsilon^2$ and $1/\eta_k = \lambda_k/(2-\lambda_k) \leq \lambda_k/\varepsilon^2$. Thus, by Cauchy-Schwarz and (1.11), we have

$$\begin{aligned}&2\gamma\langle\nabla h(x_h^k), z^k - z^{k+1}\rangle = 2\gamma\langle\nabla h(x^*), z^k - z^{k+1}\rangle + 2\gamma\langle\nabla h(x_h^k) - \nabla h(x^*), z^k - z^{k+1}\rangle\\ &\leq 2\gamma\langle\nabla h(x^*), z^k - z^{k+1}\rangle + \frac{\gamma^2}{\eta_k}\|\nabla h(x_h^k) - \nabla h(x^*)\|^2 + \eta_k\|z^k - z^{k+1}\|^2. \qquad (3.3)\end{aligned}$$

Therefore, by Jensen's inequality, the Cauchy-Schwarz inequality, (2.9), and the bound $\|z^0 - z^{k+1}\| \leq 2\|z^0 - z^*\|$ (see (3.2)), we have

$$\begin{aligned}&f(\overline{x}_f^k) + h(\overline{x}_h^k) - f(x^*) - h(x^*) \leq \frac{1}{\Lambda_k}\sum_{i=0}^{k}\lambda_i\left(f(x_f^i) + h(x_h^i) - f(x^*) - h(x^*)\right)\\ &\overset{(2.9)}{\leq} \frac{1}{2\gamma\Lambda_k}\sum_{i=0}^{k}\left(\|z^i - x^*\|^2 - \|z^{i+1} - x^*\|^2 - \eta_i\|z^{i+1} - z^i\|^2 + 2\gamma\langle\nabla h(x_h^i), z^i - z^{i+1}\rangle\right)\\ &\overset{(3.3)}{\leq} \frac{1}{2\gamma\Lambda_k}\left(\|z^0 - x^*\|^2 + 2\gamma\langle\nabla h(x^*), z^0 - z^{k+1}\rangle + (\gamma^2/\varepsilon^2)\sum_{i=0}^{\infty}\lambda_i\|\nabla h(x_h^i) - \nabla h(x^*)\|^2\right)\\ &\overset{(1.20)}{\leq} \frac{\|z^0 - x^*\|^2 + 4\gamma\|\nabla h(x^*)\|\|z^0 - z^*\| + (1+\varepsilon)\gamma\|z^0 - z^*\|^2/(\varepsilon^3(2\beta_V - \gamma))}{2\gamma\Lambda_k}.\end{aligned}$$

The lower bound in Proposition 2.5 and the Cauchy-Schwarz inequality show that

$$\begin{aligned}f(\overline{x}_f^k) + h(\overline{x}_h^k) - f(x^*) - h(x^*) \geq \langle\overline{x}_f^k - \overline{x}_h^k, \widetilde{\nabla} f(x^*)\rangle &\geq -\|\overline{x}_f^k - \overline{x}_h^k\|\|\widetilde{\nabla} f(x^*)\|\\ &\geq \frac{-2\|z^0 - z^*\|\|\widetilde{\nabla} f(x^*)\|}{\Lambda_k}. \qquad \square\end{aligned}$$

In general, $\overline{x}_h^k$ and $\overline{x}_h^k$ are not in $\mathrm{dom}(f)$. However, the conclusion of Theorem 3.1 can be improved if f is Lipschitz continuous. The following proposition gives a sufficient condition for Lipschitz continuity on a ball.

PROPOSITION 3.2 (Lipschitz continuity on a ball [3, Proposition 8.28]). *Suppose that $f : \mathcal{H} \to (-\infty, \infty]$ is proper and convex. Let $\rho > 0$, and let $x_0 \in \mathrm{dom}(f)$. If $\delta = \sup_{x,y \in B(x_0, 2\rho)} |f(x) - f(y)| < \infty$, then f is (δ/ρ)-Lipschitz on $B(x_0, \rho)$.*

To use this fact, we need to show that the sequences $(x_f^j)_{j\geq 0}$, and $(x_h^j)_{j\geq 0}$ are bounded. Recall that $x_h^s = P_V(z^s)$ and $x_f^s = \mathbf{prox}_{\gamma f} \circ \mathbf{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma \nabla h)(z^s)$ for $s \in \{*, k\}$. Proximal, reflection, and forward-gradient maps are nonexpansive (see Proposition 1.1, the Baillon-Haddad Theorem [1], and [3, Proposition 4.33]), so we have $\max\{\|x_f^k - x^*\|, \|x_h^k - x^*\|\} \leq \|z^k - z^*\| \leq \|z^0 - z^*\|$ for all $k \geq 0$. Thus, $(x_f^j)_{j\geq 0}, (x_h^j)_{j\geq 0} \subseteq \overline{B(x^*, \|z^0 - z^*\|)}$. The ball is convex, so $(\overline{x}_f^j)_{j\geq 0}, (\overline{x}_h^j)_{j\geq 0} \subseteq \overline{B(x^*, \|z^0 - z^*\|)}$.

COROLLARY 3.3 (Ergodic convergence with Lipschitz f). *Let the notation be as in Theorem 3.1. Let $L \geq 0$ and suppose f is L-Lipschitz on $\overline{B(x^*, \|z^0 - z^*\|)}$. Then*

$$
\begin{aligned}
0 &\leq f(\overline{x}_h^k) + h(\overline{x}_h^k) - f(x^*) - h(x^*) \\
&\leq \frac{\left(\|z^0 - z^*\| + 4\gamma \|\nabla h(x^*)\| + \frac{(1+\varepsilon)\gamma \|z^0 - z^*\|}{\varepsilon^3 (2\beta_V - \gamma)}\right) \|z^0 - z^*\|}{2\gamma \Lambda_k} + \frac{2L\|z^0 - z^*\|}{\Lambda_k}.
\end{aligned}
$$

Proof. The proof follows from by combining the upper bound in Theorem 3.1 with the following bound: $f(\overline{x}_h^k) \leq f(\overline{x}_f^k) + L\|\overline{x}_f^k - \overline{x}_h^k\| \leq f(\overline{x}_f^k) + 2L\|z^0 - z^*\|/\Lambda_k$. ∎

REMARK 4. *Corollary 3.3 is sharp [16, Proposition 8].*

3.2. Nonergodic convergence rates. In this section, we analyze the nonergodic convergence rate of FDRS when $(\lambda_j)_{j\geq 0}$ is bounded away from 0 and $1/\alpha_{\mathrm{FDRS}}^V$. The proof bounds the inequalities in Propositions 2.4 and 2.5 with Theorem 1.5.

THEOREM 3.4 (Nonergodic convergence of FDRS). *For all $k \geq 0$, let $\lambda_k \in (0, 1/\alpha_{\mathrm{FDRS}}^V)$. Suppose that $\underline{\tau} := \inf_{j\geq 0}(1 - \alpha_{\mathrm{FDRS}}^V \lambda_j)\lambda_j/\alpha_{\mathrm{FDRS}}^V > 0$. Then*

$$
\|x_f^k - x_h^k\| \leq \frac{\|z^0 - z^*\|}{\sqrt{\underline{\tau}(k+1)}}, \qquad \|x_f^k - x_h^k\| = o\left(\frac{1}{\sqrt{k+1}}\right),
$$

and

$$
\begin{aligned}
-\frac{\|z^0 - z^*\| \|\widetilde{\nabla} f(x^*)\|}{\sqrt{\underline{\tau}(k+1)}} &\leq f(x_f^k) + h(x_h^k) - f(x^*) - g(x^*) \\
&\leq \frac{\left(\|z^* - x^*\| + (1 + \gamma/\beta_V)\|z^0 - z^*\| + \gamma \|\nabla h(x^*)\|\right) \|z^0 - z^*\|}{\gamma \sqrt{\underline{\tau}(k+1)}},
\end{aligned}
$$

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

Proof. First we note that $\left(\|\nabla h(x_h^j)\|\right)_{j\geq 0}$ is bounded: for all $k \geq 0$,

$$
\begin{aligned}
\|\nabla h(x_h^k)\| &\leq \|\nabla h(x_h^k) - \nabla h(x^*)\| + \|\nabla h(x^*)\| = \|\nabla h(z^k) - \nabla h(z^*)\| + \|\nabla h(x^*)\| \\
&\leq \frac{1}{\beta_V}\|z^k - z^*\| + \|\nabla h(x^*)\| \leq \frac{1}{\beta_V}\|z^0 - z^*\| + \|\nabla h(x^*)\| \qquad (3.4)
\end{aligned}
$$

because $(\|z^j - z^*\|)_{j\geq 0}$ is decreasing (see Part 1 of Theorem 1.5).

Next fix $k \geq 0$. For any $\lambda > 0$, define $z_\lambda := (T_{\mathrm{FDRS}})_\lambda(z^k)$. Observe that x_f^k and x_h^k do not depend on the value of λ_k. Therefore, by Proposition 2.4 and Lemma 2.1,

$$
\begin{aligned}
&f(x_f^k) + h(x_h^k) - f(x^*) - g(x^*) \\
&\leq \inf_{\lambda \in [0, 1/\alpha_{\mathrm{FDRS}}^V)} \frac{1}{2\gamma\lambda}\left(\|z^k - x^*\|^2 - \|z_\lambda - x^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z_\lambda - z^k\|^2 \right.\\
&\qquad\qquad \left. + 2\gamma\langle \nabla h(x_h^k), z^k - z_\lambda\rangle\right) \\
&\overset{(1.10)}{=} \inf_{\lambda \in [0, 1/\alpha_{\mathrm{FDRS}}^V)} \frac{1}{2\gamma\lambda}\left(2\langle z_\lambda - x^*, z^k - z_\lambda\rangle + 2\left(1 - \frac{1}{\lambda}\right)\|z_\lambda - z^k\|^2 \right.\\
&\qquad\qquad \left. + 2\gamma\langle \nabla h(x_h^k), z^k - z_\lambda\rangle\right) \\
&\overset{(3.4)}{\leq} \frac{1}{2\gamma}\left(2\langle z_1 - x^*, z^k - z_1\rangle + 2\gamma\left(\frac{1}{\beta_V}\|z^0 - z^*\| + \|\nabla h(x^*)\|\right)\|z_1 - z^k\|\right) \qquad (3.5) \\
&\overset{(1.18)}{\leq} \frac{\left(\|z_1 - x^*\| + (\gamma/\beta_V)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|\right)\|z^0 - z^*\|}{\gamma\sqrt{\underline{\tau}(k+1)}} \\
&\leq \frac{\left(\|z^* - x^*\| + (1 + \gamma/\beta_V)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|\right)\|z^0 - z^*\|}{\gamma\sqrt{\underline{\tau}(k+1)}}
\end{aligned}
$$

where we use $\|z_1 - x^*\| \leq \|z_1 - z^*\| + \|z^* - x^*\| \leq \|z^0 - z^*\| + \|z^* - x^*\|$ (Theorem 1.5).

The lower bound follows from (2.10) and Part 3 of Theorem 1.5:

$$
\begin{aligned}
f(x_f^k) + h(x_h^k) - f(x^*) - g(x^*) \geq \langle x_f^k - x_h^k, \widetilde{\nabla} f(x^*)\rangle &= \frac{1}{\lambda_k}\langle z^{k+1} - z^k, \widetilde{\nabla} f(x^*)\rangle \qquad (3.6) \\
&\overset{(1.18)}{\geq} -\frac{\|z^0 - z^*\|\|\widetilde{\nabla} f(x^*)\|}{\sqrt{\underline{\tau}(k+1)}}.
\end{aligned}
$$

The $o(1/\sqrt{k+1})$ rates follow from (3.5) and (3.6), and the corresponding rates for the FPR in (1.18). The bounds on $x_f^k - x_h^k$ follow from $x_f^k - x_h^k = T_{\mathrm{FDRS}} z^k - z^k$. ∎

If f is Lipschitz continuous, we can evaluate the entire objective function at x_h^k. The proof of the following corollary is analogous to Corollary 3.3. We ask the reader to recall from Section 3.1 that $(x_f^j)_{j\geq 0}, (x_h^j)_{j \geq 0} \subseteq \overline{B(x^*, \|z^0 - z^*\|)}$.

Corollary 3.5 (Nonergodic convergence with Lipschitz f). *Let the notation be as in Theorem 3.4. Let $L \geq 0$ and suppose f is L-Lipschitz on $\overline{B(x^*, \|z^0 - z^*\|)}$. Then*

$$
\begin{aligned}
0 &\leq f(x_h^k) + h(x_h^k) - f(x^*) - h(x^*) \\
&\leq \frac{\left(\|z^* - x^*\| + (1 + \gamma/\beta_V)\|z^0 - z^*\| + \gamma\|\nabla h(x^*)\|\right)\|z^0 - z^*\|}{\gamma\sqrt{\underline{\tau}(k+1)}} + \frac{L\|z^0 - z^*\|}{\sqrt{\underline{\tau}(k+1)}},
\end{aligned}
$$

and $f(x_h^k) + h(x_h^k) - f(x^*) - h(x^*) = o(1/\sqrt{k+1})$.

Proof. Combine the upper bound in Theorem 3.4 with the following bound: $f(x_h^k) \leq f(x_f^k) + L\|x_f^k - x_h^k\| \leq f(x_f^k) + L\|z^0 - z^*\|/\sqrt{\underline{\tau}(k+1)}$. The $o(1/\sqrt{k+1})$ rate follows because $\|x_f^k - x_h^k\| = \|T_{\mathrm{FDRS}} z^k - z^k\| = o(1/\sqrt{k+1})$ (see (2.3) and (1.18)) and $|f(x_f^k) + h(x_h^k) - f(x^*) - h(x^*)| = o(1/\sqrt{k+1})$ (see Theorem 3.4). ∎

Remark 5. *Corollary 3.5 is sharp [16, Theorem 11].*

4. Strong convexity. In this section, we show that $(x_f^j)_{j\geq 0}$, $(x_h^j)_{j\geq 0}$, and their ergodic variants converge strongly whenever f or g is strongly convex. The techniques in this section are similar to those in Section 3, so we defer the proof to Appedix B.7

THEOREM 4.1 (Auxiliary term bound). *Let $\gamma \in (0, 2\beta_V)$, let $(\lambda_j)_{j\geq 0} \subseteq (0, 1/\alpha_{\mathrm{FDRS}}^V)$, let $z^0 \in \mathcal{H}$, and suppose that $(z^j)_{j\geq 0}$ is generated by Algorithm 1. Then*

1. "Best" iterate convergence: *Let $\varepsilon \in (0,1)$ and suppose that $(\lambda_j)_{j\geq 0}$ satisfies* (1.19)*. If $\underline{\lambda} := \inf_{j\geq 0} \lambda_j > 0$, then*

$$\min_{0\leq j\leq k} \left(S_f(x_f^j, x^*) + S_h(x_h^j, x^*)\right) \leq \frac{\left(1 + \frac{(1+\varepsilon)\gamma}{\varepsilon^3(2\beta_V - \gamma)}\right)\|z^0 - z^*\|^2}{4\gamma\underline{\lambda}(k+1)}.$$

and $\min_{0\leq j\leq k} S_f(x_f^j, x^*) = o(1/(k+1))$ *and* $\min_{0\leq j\leq k} S_h(x_h^j, x^*) = o(1/(k+1))$.

2. Ergodic convergence: *If $\varepsilon \in (0,1)$, and $(\lambda_j)_{j\geq 0}$ satisfies* (1.19)*, then*

$$\frac{\mu_f}{2}\|\overline{x}_f^k - x^*\|^2 + \frac{\mu_h}{2}\|\overline{x}_h^k - x^*\|^2 \leq \frac{\left(1 + \frac{(1+\varepsilon)\gamma}{\varepsilon^3(2\beta_V - \gamma)}\right)\|z^0 - z^*\|^2}{4\gamma\Lambda_k}.$$

3. Nonergodic convergence: *If $\underline{\tau} := \inf_{j\geq 0}(1 - \alpha_{\mathrm{FDRS}}^V\lambda_j)\lambda_j/\alpha_{\mathrm{FDRS}}^V > 0$, then $S_f(x_f^k, x^*) + S_h(x_h^k, x^*) = o(1/\sqrt{k+1})$ and*

$$S_f(x_f^k, x^*) + S_h(x_h^k, x^*) \leq \frac{(1 + \gamma/\beta_V)\|z^0 - z^*\|^2}{2\gamma\sqrt{\underline{\tau}(k+1)}},$$

REMARK 6. *See Section 6.1 for a proof that the nonergodic "best" rates are sharp. It is not clear if we can improve the general nonergodic rates to $o(1/(k+1))$.*

5. Lipschitz differentiability. In this section, we assume f is smooth:

ASSUMPTION 4. *f is differentiable and ∇f is $(1/\beta_f)$-Lipschitz where $\beta_f > 0$.*

Under Assumption 4, we will show that the objective value

$$f(x_h^k) + h(x_h^k) - f(x^*) - h(x^*) = f(x_h^k) + g(x_h^k) - f(x^*) - g(x^*)$$

is summable. Therefore, by [16, Lemma 3] the minimal objective error after k iterations is of order $o(1/(k+1))$. We will need the following upper bound to prove this. See Appendix B.8 for the proof.

PROPOSITION 5.1 (Fundamental inequality under Assumption 4). *If $\gamma \in (0, 2\beta_V)$, $\lambda > 0$, $z \in \mathcal{H}$, $z^+ := (T_{\mathrm{FDRS}})_\lambda(z)$, z^* is a fixed-point of T_{FDRS}, and $x^* = P_V z^*$, then*

$$\begin{aligned}
&2\gamma\lambda(f(x_h) + h(x_h) - f(x^*) - g(x^*)) \\
&\leq \begin{cases} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 + \frac{\gamma - \beta_f}{\beta_f\lambda}\right)\|z - z^+\|^2 & \\ +2\gamma\langle \nabla h(x_h) - \nabla h(x^*), z - z^+\rangle & \text{if } \gamma \leq \beta_f \\ \left(1 + \frac{\gamma - \beta_f}{2\beta_f}\right)(\|z - z^*\|^2 - \|z^+ - z^*\|^2 + \|z - z^+\|^2) & \\ +2\gamma\left(1 + \frac{\gamma - \beta_f}{2\beta_f}\right)\langle \nabla h(x_h) - \nabla h(x^*), z - z^+\rangle & \text{if } \gamma > \beta_f. \end{cases}
\end{aligned} \tag{5.1}$$

The next theorem shows that the upper bound in Proposition 5.1 is summable and, as a consequence, we will have $o(1/(k+1))$ convergence.

THEOREM 5.2 (Convergence rates under Assumption 4). *Let $\gamma \in (0, 2\beta_V)$, let $\varepsilon \in (0,1)$, and suppose $(\lambda_j)_{j\geq 0}$ satisfies* (1.19)*. Suppose that $\underline{\tau} := \inf_{j\geq 0}\{(1 -$*

$\alpha^V_{\mathrm{FDRS}}\lambda_j)\lambda_j/\alpha^V_{\mathrm{FDRS}}\} > 0$ *and let* $\underline{\lambda} := \inf_{j\geq 0}\lambda_j > 0$*. Let* $z^0 \in \mathcal{H}$*, let* z^* *be a fixed-point of* T_{FDRS}*, and let* $x^* := P_V z^*$*. Then*

$$\min_{0\leq j\leq k}\left(f(x_h^j) + h(x_h^j) - f(x^*) - h(x^*)\right) = o\left(\frac{1}{k+1}\right).$$

Proof. Let $\delta := \inf_{j\geq 0}\left\{(1-\lambda_j\alpha^V_{\mathrm{FDRS}})/(\lambda_j\alpha^V_{\mathrm{FDRS}})\right\}$. Note that $0 < \delta < \infty$ because $\underline{\tau} > 0$. Now, recall that, by Part 2 of Theorem 1.5, we have

$$\sum_{i=0}^{\infty}\|z^{i+1}-z^i\|^2 \leq \frac{1}{\delta}\sum_{i=0}^{\infty}\frac{1-\lambda_i\alpha^V_{\mathrm{FDRS}}}{\lambda_i\alpha^V_{\mathrm{FDRS}}}\|z^{i+1}-z^i\|^2 \leq \frac{1}{\delta}\|z^0-z^*\|^2.$$

Next, we use the Cauchy-Schwarz inequality and (1.11) to show that

$$\begin{aligned}\sum_{i=0}^{\infty}2\gamma\langle\nabla h(x_h^i)-\nabla h(x^*), z^i-z^{i+1}\rangle &\leq \sum_{i=0}^{\infty}\left(\lambda_i\gamma^2\|\nabla h(x_h^i)-\nabla h(x^*)\|^2 + \frac{1}{\lambda_i}\|z^i-z^{i+1}\|^2\right)\\ &\overset{(1.20)}{\leq}\left(\frac{(1+\varepsilon)\gamma}{\varepsilon(2\beta_V-\gamma)}+\frac{1}{\underline{\lambda}\delta}\right)\|z^0-z^*\|^2.\end{aligned}$$

If we combine the previous two sum bounds with (5.1), we get

$$\begin{aligned}&\sum_{i=0}^{\infty}(f(x_h^i)+h(x_h^i)-f(x^*)-h(x^*))\\ &\leq\frac{\left(1+\frac{1}{\delta}+\frac{(1+\varepsilon)\gamma}{\varepsilon(2\beta_V-\gamma)}+\frac{1}{\underline{\lambda}\delta}\right)\|z^0-z^*\|^2}{2\gamma\underline{\lambda}}\times\begin{cases}1 & \text{if } \gamma\leq\beta_f;\\ \left(1+\frac{\gamma-\beta_f}{2\beta_f}\right) & \text{if } \gamma>\beta_f.\end{cases}\end{aligned}$$

The convergence rate now follows from [16, Lemma 3]. □

REMARK 7. *Theorem 5.2 is sharp under Assumption 4 [16, Theorem 12].*

6. Linear convergence.

In this section, we prove FDRS converges linearly when $\beta_f(\mu_g+\mu_f) > 0$.

THEOREM 6.1 (Linear convergence). *Let* $\gamma\in(0,2\beta_V)$*, let* $(\lambda_j)_{j\geq 0}\subseteq(0,1/\alpha^V_{\mathrm{FDRS}})$*, let* $z^0\in\mathcal{H}$*, let* z^* *be a fixed-point of* T_{FDRS}*, and let* $x^* := P_V z^*$*. Let* $c > 1/2$*, let* $\gamma < \beta_V/c$*, and let* $(\lambda_j)_{j\geq 0}\subseteq(0,(2c-1)/c)$*. For all* $\lambda\in(0,(2c-1)/c)$*, define*

$$\begin{aligned}C_1(\lambda) &:= \left(1-\frac{\lambda}{3}\min\left\{\frac{\gamma\mu_g}{(1+\gamma/\beta_V)^2},\frac{\beta_f}{\gamma},\frac{2c-1}{c}-\lambda\right\}\right)^{1/2};\\ C_2(\lambda) &:= \left(1-\frac{\lambda}{3}\min\left\{\frac{\gamma\mu_f}{(1+\gamma/\beta_f)^2},\frac{\beta_V-c\gamma}{\gamma},\frac{1}{4}\left(\frac{2c-1}{c}-\lambda\right)\right\}\right)^{1/2}.\end{aligned}$$

Then for all $k\geq 0$*, we have*

$$\|z^{k+1}-z^*\|\leq\|z^k-z^*\|\times\begin{cases}C_1(\lambda_k) & \text{if } \mu_g\beta_f>0;\\ C_2(\lambda_k) & \text{if } \mu_f\beta_f>0;\end{cases}\tag{6.1}$$

$$\|z^{k+1}-z^*\|\leq\|z^0-z^*\|\times\begin{cases}\prod_{i=0}^k C_1(\lambda_i) & \text{if } \mu_g\beta_f>0;\\ \prod_{i=0}^k C_2(\lambda_i) & \text{if } \mu_f\beta_f>0.\end{cases}$$

Proof. (2.11) shows that for all $k \geq 0$, we have

$$\begin{aligned}
&\gamma\lambda_k\mu_f\|x_f^k - x^*\|^2 + \gamma\lambda_k\beta_f\|\nabla f(x_f^k) - \nabla f(x^*)\|^2 \\
&+ \gamma\lambda_k\mu_g\|x_h^k - x^*\|^2 + \gamma\lambda_k\beta_V\|\nabla h(x_h^k) - \nabla h(x^*)\|^2 \\
&\leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left(1 - \frac{2}{\lambda_k}\right)\|z^{k+1} - z^k\|^2 \\
&+ 2\gamma\langle \nabla h(x_h^k) - \nabla h(x^*), z^k - z^{k+1}\rangle.
\end{aligned}$$

In addition, by the Cauchy-Schwarz inequality and (1.11), we have

$$2\gamma\langle \nabla h(x_h^k) - \nabla h(x^*), z^k - z^{k+1}\rangle \leq c\gamma^2\lambda_k\|\nabla h(x_h^k) - \nabla h(x^*)\|^2 + \frac{1}{c\lambda_k}\|z^k - z^{k+1}\|^2.$$

Therefore, for all $k \geq 0$,

$$\begin{aligned}
&\gamma\lambda_k\mu_f\|x_f^k - x^*\|^2 + \gamma\lambda_k\beta_f\|\nabla f(x_f^k) - \nabla f(x^*)\|^2 \\
&+ \gamma\lambda_k\mu_g\|x_h^k - x^*\|^2 + \gamma\lambda_k(\beta_V - c\gamma)\|\nabla h(x_h^k) - \nabla h(x^*)\|^2 \\
&\leq \|z^k - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \left(1 - \frac{2c-1}{c\lambda_k}\right)\|z^{k+1} - z^k\|^2.
\end{aligned}$$

Recall that we assume $1 - (2c-1)/(c\lambda_k) < 0$ and $\beta_V - c\gamma > 0$.

Now suppose that $\beta_f\mu_g > 0$. The following identity follows from from Lemma 2.1:

$$z^k = T_{\mathrm{FDRS}}(z^k) + (z^k - T_{\mathrm{FDRS}}(z^k)) = x_h^k - \gamma\nabla h(x_h^k) - \gamma\nabla f(x_f^k) + \frac{1}{\lambda_k}(z^k - z^{k+1}).$$

This identity results from tracing the perimeter of Figure 2.1 from x_h to x_f to $T_{\mathrm{FDRS}}z^k$ to z^k. Likewise, we have $z^* = x^* - \gamma\nabla h(x^*) - \gamma\nabla f(x^*)$.

Note that

$$\begin{aligned}
\|(x_h^k - \gamma\nabla h(x_h^k)) - (x^* - \gamma\nabla h(x^*))\| &\leq \|x_h^k - x^*\| + \gamma\|\nabla h(x_h^k) - \nabla h(x^*)\| \\
&\leq (1 + \gamma/\beta_V)\|x_h^k - x^*\|. \qquad (6.2)
\end{aligned}$$

Now, fix $k \geq 0$, and let $C_1' := 3\max\left\{(1+\gamma/\beta_V)^2/(\gamma\lambda_k\mu_g), \gamma^2/(\gamma\lambda_k\beta_f), (1/\lambda_k^2)\left(\frac{2c-1}{c\lambda_k} - 1\right)^{-1}\right\}$.
By the convexity of $\|\cdot\|^2$, we have

$$\begin{aligned}
&\|z^k - z^*\|^2 \leq 3(1+\gamma/\beta_V)^2\|x_h^k - x^*\|^2 + 3\gamma^2\|\nabla f(x_f^k) - \nabla f(x^*)\|^2 + \frac{3}{\lambda_k^2}\|z^{k+1} - z^k\|^2 \\
&\leq C_1'\left(\gamma\lambda_k\mu_g\|x_h^k - x^*\|^2 + \gamma\lambda_k\beta_f\|\nabla f(x_f^k) - \nabla f(x^*)\|^2 + \left(\frac{2c-1}{c\lambda_k} - 1\right)\|z^{k+1} - z^k\|^2\right) \\
&\leq C_1'\|z^k - z^*\|^2 - C_1'\|z^{k+1} - z^*\|^2.
\end{aligned}$$

Therefore, $\|z^{k+1} - z^*\| \leq (1 - (1/C_1'))^{1/2}\|z^k - z^*\|$.

Now assume that $\beta_f\mu_f > 0$. Observe that:

$$\begin{aligned}
z^k &= x_h^k - \gamma\nabla h(x_h^k) - \gamma\nabla f(x_f^k) + \frac{1}{\lambda_k}(z^k - z^{k+1}) \\
&= x_f^k - \gamma\nabla h(x_h^k) - \gamma\nabla f(x_f^k) + \frac{2}{\lambda_k}(z^k - z^{k+1})
\end{aligned}$$

where we use the identity $x_h^k - x_f^k = (1/\lambda_k)(z^k - z^{k+1})$ (see (2.3)). The proof of this case is similar to the case $\beta_f \mu_h > 0$ except that we use the above identity for z^k, the bound $\|(x_f^k - \gamma\nabla f(x_f^k)) - (x^* - \gamma \nabla f(x^*))\|^2 \leq (1+\gamma/\beta_f)^2\|x_f^k - x^*\|^2$, and the constant $C_2' := 3\max\left\{(1+\gamma/\beta_f)^2/(\gamma\lambda_k\mu_f), \gamma^2/(\gamma\lambda_k(\beta_V - c\gamma)), (4/\lambda_k^2)\left(\frac{2c-1}{c\lambda_k} - 1\right)^{-1}\right\}$ in place of C_1'. Then the contraction $\|z^{k+1} - z^*\| \leq (1-1/C_2')^{1/2}\|z^k - z^*\|$ follows.

In both cases, the linear rate for $(z^j)_{j\geq 0}$ follows by unfolding (6.1). □

REMARK 8. *Note that smaller c lead to larger γ and smaller $(\lambda_j)_{j\geq 0}$, while larger c lead to smaller γ and larger $(\lambda_j)_{j\geq 0}$.*

6.1. Arbitrarily slow convergence for strongly convex problems. In general, we cannot expect linear convergence of FDRS when f is not differentiable—even if f and g are strongly convex. In this section, we construct an example to prove this claim. The following example is based on [2, Section 7] and [16, Example 1].

A family of slow examples. Let $\mathcal{H} := \ell_2^2(\mathbf{N}) = \mathbf{R}^2 \oplus \mathbf{R}^2 \oplus \cdots$. 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 $e_\theta := R_\theta e_0$. Let $(\theta_j)_{j\geq 0}$be 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

$$V := \mathbf{R}e_0 \oplus \mathbf{R}e_0 \oplus \cdots \qquad \text{and} \qquad U := \mathbf{R}e_{\theta_0} \oplus \mathbf{R}e_{\theta_1} \oplus \cdots. \tag{6.3}$$

Note that [2, Section 7] proves the projection identities

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

We now begin our extension of this example. Choose $a \geq 0$ and set $f := \chi_U + (a/2)\|\cdot\|^2$ and $g := (1/2)\|\cdot\|^2$. Note that $\mu_g = 1$ and $\mu_f = a$. In addition, for $h := g \circ P_V$, we have $(\nabla h(x))_i = (P_V \circ I_{\mathcal{H}} \circ P_V)_i = (P_V)_i$. Thus, ∇h is 1-Lipschitz, and, hence, $\beta_V = 1$ and we can choose $\gamma = 1 < 2\beta_V$. Therefore, $\alpha_{\text{FDRS}}^V = 2\beta_V/(4\beta_V - \gamma) = 2/3$, so we can choose $\lambda_k \equiv 1 < 1/\alpha_{\text{FDRS}}^V$. We also note that $\mathbf{prox}_{\gamma f} = (1/(1+a))P_U$.

Define $N : \mathcal{H} \to \mathcal{H}$ on each 2-dimensional component of $\mathcal{H}$ as follows: for all $i \geq 0$,

$$\begin{aligned}(N)_i &:= \left(\frac{1}{2}I_{\mathcal{H}} + \frac{1}{2}\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V}\right)_i = \frac{1}{a+1}(P_U)_i(2(P_V)_i - I_{\mathbf{R}^2}) + I_{\mathbf{R}^2} - (P_V)_i \\ &= \frac{1}{a+1}(P_U)_i\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{a+1}\begin{bmatrix} \cos^2(\theta_i) & -\sin(\theta_i)\cos(\theta_i) \\ \sin(\theta_i)\cos(\theta_i) & \cos^2(\theta_i) + a \end{bmatrix}\end{aligned}$$

where the second equality follows by direct expansion. Therefore, we have

$$T_{\text{FDRS}} = N \circ (I - P_V) = \bigoplus_{i\geq 0} \frac{1}{a+1}\begin{bmatrix} 0 & -\sin(\theta_i)\cos(\theta_i) \\ 0 & \cos^2(\theta_i) + a \end{bmatrix}. \tag{6.4}$$

Note that for all $i \geq 0$, the operator $(T_{\text{FDRS}})_i$ has eigenvector

$$z_i := \left(-\frac{\cos(\theta_i)\sin(\theta_i)}{a + \cos^2(\theta_i)}, 1\right) \tag{6.5}$$

with eigenvalue $b_i := (a + c_i^2)/(a+1) < 1$. Each component also has the eigenvector $(1,0)$ with eigenvalue 0. Thus, the only fixed-point of T_{FDRS} is $0 \in \mathcal{H}$. Finally,

$$\|z_i\|^2 = \frac{c_i^2(1-c_i^2)}{(a+c_i^2)^2} + 1 \qquad \text{and} \qquad \|(P_V)_i z_i\|^2 = \frac{c_i^2(1-c_i^2)}{(a+c_i^2)^2}. \tag{6.6}$$

Slow convergence proofs. We know that $z^{k+1}-z^k \to 0$ from (1.18). Therefore, because T_{FDRS} is linear, [3, Proposition 5.27] proves the following lemma.

LEMMA 6.2 (Strong convergence for linear operators). *Any sequence $(z^j)_{j\geq 0} \subseteq \mathcal{H}$ generated by the T_{FDRS} operator in (6.4) converges strongly to 0. Consequently, the sequences $(x_h^j)_{j\geq 0} = (P_V z^j)_{j\geq 0}$ and $(x_f^j)_{j\geq 0}$ converge strongly to zero.*

LEMMA 6.3 (Slow sequences [16, Lemma 6]). *Suppose that $F : \mathbf{R}_+ \to (0,1)$ is a function that is strictly decreasing to zero such that $\{1/(j+1) \mid j \in \mathbf{N}\backslash\{0\}\} \subseteq$* range$(F)$ *Then there exists a monotonic sequence $(b_j)_{j\geq 0} \subseteq (0,1)$ such that $b_k \to 1^-$ as $k \to \infty$ and an increasing sequence $(n_j)_{j\geq 0} \subseteq \mathbf{N} \cup \{0\}$ such that for all $k \geq 0$,*

$$\frac{b_{n_k}^{k+1}}{(n_k+1)} > e^{-1}F(k+1).$$

The following is a simple corollary of Lemma 6.3.

COROLLARY 6.4. *Let the notation be as in Lemma 6.3. Then for all $\eta \in (0,1)$, we can find a sequence $(b_j)_{j\geq 0} \subseteq (\eta, 1)$ that satisfies the conditions of the lemma.*

Proof. For any $\varepsilon \in (0, 1-\eta)$, replace the sequence $(b_j)_{j\geq 0}$ in Lemma 6.3 with $(\max\{b_j, \eta+\varepsilon\})_{j\geq 0}$. $\square$

We are now ready to show that FDRS can converge arbitrarily slowly.

THEOREM 6.5 (Arbitrarily slow FDRS). *For every function $F : \mathbf{R}_+ \to (0,1)$ that strictly decreases to zero and satisfies $\{1/(j+1) \mid j \in \mathbf{N}\backslash\{0\}\} \subseteq$* range$(F)$, *there is a point $z^0 \in \ell_2^2(\mathbf{N})$ and two closed subspaces U and V with zero intersection, $U \cap V = \{0\}$, such that the FDRS sequence $(z^j)_{j\geq 0}$ generated with the functions $f := \chi_U + (a/2)\|\cdot\|^2$ and $g := (1/2)\|\cdot\|^2$ and parameters $\lambda_k \equiv 1$ and $\gamma = 1$ strongly converges to zero, but for all $k \geq 1$, we have*

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

Proof. For all $i \geq 0$, define $z_i^0 = (\|z_i\|^{-1}/(i+1))z_i$ with z_i as in (6.5). Then $\|z_i^0\| = 1/(i+1)$ and z_i^0 is an eigenvector of $(T_{\mathrm{FDRS}})_i$ with eigenvalue $b_i := (a + c_i^2)/(a+1)$. Define the concatenated vector $z^0 := (z_i^0)_{i\geq 0}$. Note that $z^0 \in \mathcal{H}$ because $\|z^0\|^2 = \sum_{i=0}^\infty 1/(i+1)^2 < \infty$. Thus, for all $k \geq 0$, we let $z^{k+1} := T_{\mathrm{FDRS}} z^k$.

Now, recall that $z^* = 0$. Thus, for all $n \geq 0$ and $k \geq 0$, we have

$$\|z^{k+1} - z^*\|^2 = \|T_{\mathrm{FDRS}}^{k+1} z^0\|^2 = \sum_{i=0}^\infty b_i^{2(k+1)}\|z_i^0\|^2 = \sum_{i=0}^\infty \frac{b_i^{2(k+1)}}{(i+1)^2} \geq \frac{b_n^{2(k+1)}}{(n+1)^2}.$$

Thus, $\|z^{k+1} - z^*\| \geq b_n^{k+1}/(n+1)$. Choose b_n and the sequence $(n_j)_{j\geq 0}$ using Corollary 6.4 with $\eta \in (a/(a+1), 1)$. Then solve $c_n = \sqrt{b_n(1+a) - a} > 0$. $\square$

REMARK 9. *Theorems 6.5 and 4.1 show that the sequence $(z^j)_{j\geq 0}$ can converge* arbitrarily slowly *even if $(x_f^j)_{j\geq 0}$ and $(x_h^j)_{j\geq 0}$ converge with rate $o(1/\sqrt{k+1})$.*

The following theorem shows that $(x_f^j)_{j\geq 0}$ and $(x_h^j)_{j\geq 0}$ do not converge linearly. See Appendix B.9 for the proof.

THEOREM 6.6. *There exists a sequence $(c_i)_{i\geq 0}$ so that $(x_h^j)_{j\geq 0}$ and $(x_f^j)_{j\geq 0}$ converge strongly, but not linearly. In particular, for any $\alpha > 1/2$, there is an initial point $z^0 \in \mathcal{H}$ so that for all $k \geq 1$,*

$$\|x_h^k - x^*\|^2 \geq \frac{1}{(k+1)^{2\alpha}} \qquad \text{and} \qquad \|x_f^k - x^*\|^2 \geq \frac{(a+1/2)^2}{(a+1)^2(k+1)^{2\alpha}}.$$

Thus, the nonergodic "best" convergence rates in Part 3 of Theorem 4.1 are sharp.

7. Primal-dual splittings. In this section, we reformulate FDRS as a primal-dual algorithm applied to the dual of the following problem: $\text{minimize}_{x\in V} f(x)+h(x)$.

LEMMA 7.1 (FDRS is a primal-dual algorithm). *Let $\tau := 1/\gamma$, and suppose that $(z^j)_{j\geq 0}$ is generated by the FDRS algorithm with $\lambda_k \equiv 1$. For all $k \geq 0$, let $y^k := -\widetilde{\nabla}\chi_V(x_h^k)$. Then for all $k \geq 0$, we have the recursive update rule:*

$$\begin{cases} y^{k+1} &= P_{V^\perp}(y^k - \tau x_f^k); \\ x_f^{k+1} &= \mathbf{prox}_{\gamma f}\left(x_f^k - \gamma\nabla h(x_f^k) + \gamma(2y^{k+1} - y^k)\right). \end{cases} \tag{7.1}$$

Proof. Fix $k \geq 0$. By Lemma 2.1, $z^{k+1} = x_f^k - \gamma y^k$, so $(-1/\gamma)z^{k+1} = y^k - \tau x_f^k$. Thus, the formula for $(y^j)_{j\geq 0}$ follows from $y^{k+1} = -\widetilde{\nabla}\chi_V(x_h^{k+1}) = -(1/\gamma)P_{V^\perp}z^{k+1}$.

Now observe that

$$x_f^k = P_V x_f^k + P_{V^\perp}x_f^k = P_V(z^{k+1}+\gamma y^k) + P_{V^\perp}(z^{k+1}+\gamma y^k) = x_h^{k+1} + \gamma(y^k - y^{k+1}).$$

Furthermore, $\nabla h(x_f^k) = \nabla h(P_V x_f^k) = \nabla h(P_V(z^{k+1}+\gamma y^k)) = \nabla h(x_h^{k+1})$. Thus,

$$\begin{aligned} x_f^{k+1} &\overset{(2.3)}{=} x_h^{k+1} - \gamma\left(\widetilde{\nabla}\chi_V(x_h^{k+1}) + \nabla h(x_h^{k+1}) + \widetilde{\nabla} f(x_f^{k+1})\right) \\ &= \mathbf{prox}_{\gamma f}(x_h^{k+1} - \gamma\nabla h(x_h^{k+1}) + \gamma y^{k+1}) \\ &= \mathbf{prox}_{\gamma f}(x_f^k - \gamma\nabla h(x_f^k) + \gamma(2y^{k+1} - y^k)). \qquad \square \end{aligned}$$

The algorithm in (7.1) is the primal-dual forward-backward algorithm of Vũ and Condat [26, 13] applied to the following dual problem: $\text{minimize}_{x\in V^\perp}\ (f+h)^*(x)$ where $(f+h)^*(\cdot) = \sup_{x\in\mathcal{H}}\langle x, \cdot\rangle - (f+h)(x)$ is the Legendre-Fenchel transform of $f+h$ [3, Definition 13.1]. For convergence, [26, Theorem 3.1] requires $\gamma\tau < 1$ and $2\beta_V > \left(\min\{1/\gamma, 1/\tau\}\left(1-\sqrt{\gamma\tau}\right)\right)^{-1}$ whereas FDRS requires $\gamma < 2\beta_V$ (and $\tau = 1/\gamma$).

Thus, the FDRS algorithm is a limiting case of Vũ and Condat's algorithm, much like the DRS algorithm [21] is a limiting case of Chambolle and Pock's primal-dual algorithm [8]. In addition, the convergence rate analysis in Section 3 cannot be subsumed by the recent convergence rate analysis of the primal-dual gap of Vũ and Condat's algorithm [15], which only applies when $\gamma\tau < 1$. The original FDRS paper did not show this connection [7, Remark 6.3 (iii)].

8. Conclusion. In this paper, we provided a comprehensive convergence rate analysis of the FDRS algorithm under general convexity, strong convexity, and Lipschitz differentiability assumptions. In almost all cases, the derived convergence rates are shown to be sharp. In addition, we showed that the FDRS algorithm is the limiting case of a recently developed primal-dual forward-backward operator splitting algorithm and, thus, clarify how it relates to existing algorithms. Future work on FDRS might evaluate the performance of the algorithm on realistic problems.

Acknowledgement. We thank Prof. Wotao Yin and the anonymous reviewers for helpful comments. We also thank the two anonymous referees for their insightful and detailed comments.

Appendix A. Performance improvement: β_V versus β.

In this section, we briefly illustrate the benefits of using β_V in place of β on a Kernelized SVM problem, which is discussed in Section 1; see (1.5) for notation. In Figure A.1 we plot the FPR associated to the FDRS algorithm applied to a 1000-dimensional quadratic program. To generate the quadratic program, we use

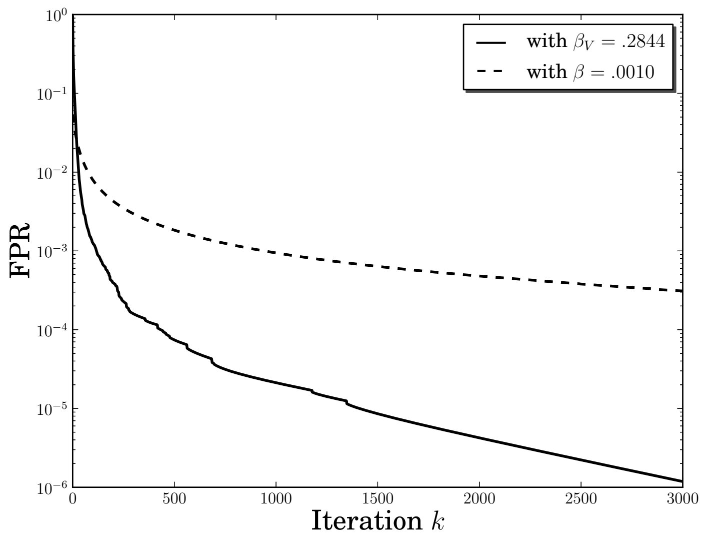

FIG. A.1. *We plot the normalized FPR, $\|T_{\mathrm{FDRS}} z^k - z^k\|/(1 + \|T_{\mathrm{FDRS}} z^k\|)$, in a dual SVM example. See Appendix A for the details.*

a random 1000-element subset of the the "a7a" dataset (available from the LIBSVM website [9]) denoted by $X = \{(x_1, y_1)^T, \cdots, (x_{1000}, y_{1000})^T\} \subseteq \mathbf{R}^{123}$ where for each $i = 1, \cdots, 1000$, $x_i \in \mathbf{R}^{122}$ is a data point and $y_i \in \{-1, 1\}$ is a class label. We use the matrix $Q \in \mathbf{R}^{1000\times 1000}$ with i, j entry given by the formula $Q_{i,j} = y_i y_j \exp(-2^{-3}\|x_i - x_j\|^2)$ for $i, j \in \{1, \cdots, 1000\}$ (i.e., we use the radial basis function kernel). The matrix A is the row vector $(y_1, \cdots, y_{1000}) \in \mathbf{R}^{1\times 1000}$, and the set C is the box $[0, 10]^{1000} \subseteq \mathbf{R}^{1000}$. In this case, P_V has rank 999, but the maximal eigenvalue ($1/\beta_V \approx 3.5159$) of $P_V \circ Q \circ P_V$ is approximately 275.8248 times smaller than the maximal eigenvalue ($1/\beta \approx 969.7836$) of Q. Figure A.1 shows that choosing $\gamma = 1.99\beta_V$ results in a tremendous speedup. (In both examples, we chose $\lambda_k \equiv 1$.)

Appendix B. Proofs of technical results.

B.1. Proof of Proposition 1.2. For the proof, we ask the reader to recall (1.15). For all $k \geq 0$, set

$$p^k := \frac{1-\alpha_1}{\alpha_1}\|(I_{\mathcal{H}} - T_1) \circ T_2(z^k) - (I_{\mathcal{H}} - T_1) \circ T_2(z^*)\|^2 + \frac{1-\alpha_2}{\alpha_2}\|(I_{\mathcal{H}} - T_2)(z^k) - (I_{\mathcal{H}} - T_2)(z^*)\|^2.$$

By applying (1.13) twice, we get $\|T_1 \circ T_2(z^k) - T_1 \circ T_2(z^*)\|^2 \leq \|z^k - z^*\|^2 - p^k$.

Part 5 of Proposition 1.1 shows that $(T_1 \circ T_2)_{\lambda_k}$ is $(\alpha_{1,2}\lambda_k)$-averaged. Thus,

$$\|z^{k+1} - z^*\|^2 \overset{(1.13)}{\leq} \|z^k - z^*\|^2 - \frac{\lambda_k(1 - \lambda_k\alpha_{1,2})}{\alpha_{1,2}}\|T_1 \circ T_2(z^k) - z^k\|^2.$$

Therefore, $\sum_{i=0}^{\infty} \frac{\lambda_i(1-\alpha_{1,2}\lambda_i)}{\alpha_{1,2}}\|T_1 \circ T_2(z^i) - z^i\|^2 \leq \|z^0 - z^*\|^2$.

By [3, Corollary 2.14], the following holds: for all $x, y \in \mathcal{H}$ and all $\lambda \in \mathbf{R}$, we have $\|\lambda x + (1-\lambda)y\|^2 = \lambda\|x\|^2 + (1-\lambda)\|y\|^2 - \lambda(1-\lambda)\|x-y\|^2$. Therefore, we have

$$
\begin{aligned}
&\|z^{k+1} - z^*\|^2 \\
&= (1-\lambda_k)\|z^k - z^*\|^2 + \lambda_k\|T_1 \circ T_2(z^k) - T_1 \circ T_2(z^*)\|^2 - \lambda_k(1-\lambda_k)\|z^k - T_1 \circ T_2(z^k)\|^2 \\
&\leq \|z^k - z^*\|^2 - \lambda_k p^k + \lambda_k(\lambda_k - 1)\|z^k - T_1 \circ T_2(z^k)\|^2 \\
&\leq \|z^k - z^*\|^2 - \lambda_k p^k + \frac{\lambda_k(1-\alpha_{1,2}\lambda_k)}{\alpha_{1,2}\varepsilon}\|z^k - T_1 \circ T_2(z^k)\|^2.
\end{aligned}
$$

Thus, take $k \to \infty$ in the following inequality to get the result:

$$
\begin{aligned}
&\sum_{i=0}^{k} \lambda_i \|(I_{\mathcal{H}} - T_2)(z^i) - (I_{\mathcal{H}} - T_2)(z^*)\|^2 \leq \frac{\alpha_2}{1-\alpha_2}\sum_{i=0}^{k}\lambda_i p^i \\
&\leq \frac{\alpha_2}{1-\alpha_2}\sum_{i=0}^{k}\left(\|z^i - z^*\|^2 - \|z^{i+1} - z^*\|^2 + \frac{\lambda_i(1-\alpha_{1,2}\lambda_i)}{\alpha_{1,2}\varepsilon}\|z^i - T_1 \circ T_2(z^i)\|^2\right) \\
&\leq \frac{\alpha_2(1+1/\varepsilon)\|z^0 - z^*\|^2}{1-\alpha_2}. \qquad \square
\end{aligned}
$$

B.2. Proof of Lemma 2.1. The identity for $x_h = z - \gamma\widetilde{\nabla}\chi_V(x_h)$ follows from Part 1 of Proposition 1.1. Note that by the Moreau identity $P_{V^\perp} = I - P_V$, we have $\gamma\widetilde{\nabla}\chi_V(x_h) = P_{V^\perp}z$. Note that by definition, $\nabla h(z) = P_V \circ \nabla g \circ P_V(z) = P_V \circ \nabla g(x_h) = \nabla h(x_h)$ and $\nabla h(z) \in V$. Thus, we get the identity for x_f:

$$
\begin{aligned}
&\mathbf{prox}_{\gamma f} \circ \mathbf{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma\nabla h)(z) = \mathbf{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma\nabla h)(z) - \gamma\widetilde{\nabla} f(x_f) \\
&= x_h - \gamma\nabla h(z) - P_{V^\perp}z - \gamma\widetilde{\nabla} f(x_f) = x_h - \gamma\left(\widetilde{\nabla}\chi_V(x_h) + \nabla h(x_h) + \widetilde{\nabla} f(x_f)\right).
\end{aligned}
$$

Finally, given the identity $(T_{\mathrm{FDRS}})_\lambda(z) - z = \lambda(T_{\mathrm{FDRS}}(z) - z)$, (2.3) will follow as soon as we show $T_{\mathrm{FDRS}}(z) = x_f + z - x_h = x_f + \gamma\widetilde{\nabla}\chi_V(x_h)$:

$$
\begin{aligned}
\left(\frac{1}{2}I_{\mathcal{H}} + \frac{1}{2}\mathbf{refl}_{\gamma f} \circ \mathbf{refl}_{\chi_V}\right)(z - \gamma\nabla h(z)) &= \left(\mathbf{prox}_{\gamma f} \circ \mathbf{refl}_{\chi_V} + I_{\mathcal{H}} - P_V\right)(z - \gamma\nabla h(z)) \\
&= x_f + P_{V^\perp}(z - \gamma\nabla h(z)) = x_f + \gamma\widetilde{\nabla}\chi_V(x_h). \qquad \square
\end{aligned}
$$

B.3. Proof of Lemma 2.3. Let $x \in \mathrm{zer}(\partial f + \nabla h + \partial\chi_V)$. Choose subgradients $\widetilde{\nabla} f(x) \in \partial f(x)$ and $\widetilde{\nabla}\chi_V(x) \in \partial\chi_V(x) = V^\perp$ (by (1.8)) such that $\widetilde{\nabla} f(x) + \nabla h(x) + \widetilde{\nabla}\chi_V(x) = 0$ and set $z := x + \gamma\widetilde{\nabla}\chi_V(x)$. We claim that z is a fixed-point of T_{FDRS}. From Lemma 2.1, we get the points: $x_h := P_V(z) = x$ and $x_f := \mathbf{prox}_{\gamma f} \circ \mathbf{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma\nabla h)(z)$. But $\widetilde{\nabla}\chi_V(x_h) + \nabla h(x_h) \in -\partial f(x)$, and

$$
\begin{aligned}
&\mathbf{refl}_{\chi_V} \circ (I_{\mathcal{H}} - \gamma\nabla h)(z) = P_V(z - \gamma\nabla h(z)) + (P_V - I_{\mathcal{H}})(z - \gamma\nabla h(z)) \\
&= x - \gamma\nabla h(x) - P_{V^\perp}z = x - \gamma\nabla h(x) - \gamma\widetilde{\nabla}\chi_V(x) = x + \gamma\widetilde{\nabla} f(x).
\end{aligned}
$$

Therefore, $x_f = \mathbf{prox}_{\gamma f}(x + \gamma\widetilde{\nabla} f(x)) = x = x_h$ (see Part 1 of Proposition 1.1). Thus, by Lemma 2.1, $T_{\mathrm{FDRS}}z = z + x_f - x_h = z$. We have proved the first inclusion.

On the other hand, suppose that $z \in \mathcal{H}$ and $T_{\mathrm{FDRS}}z = z$. Then $x := x_h = P_V z$, and $0 = T_{\mathrm{FDRS}}z - z = x_f - x_h = -\gamma\left(\widetilde{\nabla}\chi_V(x_h) + \nabla h(x_h) + \widetilde{\nabla} f(x_f)\right)$. Because $x_f = x_h$, we get $x \in \mathrm{zer}(\partial f + \nabla h + \partial\chi_V)$. $\square$

B.4. Proof of Proposition 2.4. In the following derivation, we use (2.5) and (2.6), Lemma 2.1, the cosine rule, and the inclusion $\widetilde{\nabla}\chi_V(x_h) \in V^\perp$:

$$\begin{aligned}
&2\gamma\lambda\left(f(x_f) + h(x_h) - f(x) - h(x) + S_f(x_f, x) + S_h(x_h, x)\right)\\
&\leq 2\gamma\lambda\left(\langle\widetilde{\nabla} f(x_f), x_f - x\rangle + \langle\nabla h(x_h), x_h - x\rangle + \langle\widetilde{\nabla}\chi_V(x_h), x_h - x\rangle\right)\\
&= 2\gamma\lambda\left(\langle\widetilde{\nabla} f(x_f) + \nabla h(x_h) + \widetilde{\nabla}\chi_V(x_h), x_f - x\rangle + \langle\nabla h(x_h) + \widetilde{\nabla}\chi_V(x_h), x_h - x_f\rangle\right)\\
&= 2\langle z - z^+, x_f - x\rangle + 2\langle\gamma\nabla h(x_h) + \gamma\widetilde{\nabla}\chi_V(x_h), z - z^+\rangle\\
&= 2\langle z - z^+, x_f + \gamma\widetilde{\nabla}\chi_V(x_h) - x\rangle + 2\gamma\langle\nabla h(x_h), z - z^+\rangle\\
&= 2\langle z - z^+, T_{\mathrm{FDRS}} z - x\rangle + 2\gamma\langle\nabla h(x_h), z - z^+\rangle\\
&= 2\langle z - z^+, z - x\rangle + \frac{2}{\lambda}\langle z - z^+, z^+ - z\rangle + 2\gamma\langle\nabla h(x_h), z - z^+\rangle\\
&\overset{(1.10)}{=} \|z - x\|^2 - \|z^+ - x\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 + 2\gamma\langle\nabla h(x_h), z - z^+\rangle. \qquad \square
\end{aligned}$$

B.5. Proof of Proposition 2.5. By (2.5) and (2.6) and because $\widetilde{\nabla}\chi_V(x^*) \in V^\perp$, we have

$$\begin{aligned}
f(x_f) + h(x_h) - f(x^*) - g(x^*) &\geq \langle x_h - x^*, \widetilde{\nabla} f(x^*) + \nabla h(x^*) + \widetilde{\nabla}\chi_V(x^*)\rangle\\
&\quad + \langle x_f - x_h, \widetilde{\nabla} f(x^*)\rangle + S_f(x_f, x^*) + S_h(x_h, x^*)\\
&= \langle x_f - x_h, \widetilde{\nabla} f(x^*)\rangle + S_f(x_f, x^*) + S_h(x_h, x^*). \qquad \square
\end{aligned}$$

B.6. Proof of Corollary 2.6. By (1.10), we have $\|z - x^*\|^2 - \|z^+ - x^*\|^2 = \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^*\rangle$. Therefore, by Proposition 2.4,

$$\begin{aligned}
&2\gamma\lambda\left(f(x_f) + h(x_h) - f(x^*) - h(x^*) + S_f(x_f, x^*) + S_h(x_h, x^*)\right)\\
&\quad\leq \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^*\rangle\\
&\quad + \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 + 2\gamma\langle\nabla h(x_h), z - z^+\rangle. \qquad \text{(B.1)}
\end{aligned}$$

Equation (2.11) now follows from (B.1) and (2.10):

$$\begin{aligned}
&4\gamma\lambda(S_f(x_f, x^*) + S_h(x_h, x^*)) \overset{(2.10)}{\leq} -2\gamma\lambda\langle x_f - x_h, \widetilde{\nabla} f(x^*)\rangle\\
&+ 2\gamma\lambda(f(x_f) + h(x_h) - f(x^*) - h(x^*) + S_f(x_f, x^*) + S_h(x_h, x^*))\\
&\overset{(B.1)}{\leq} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^*\rangle - 2\gamma\lambda\langle x_f - x_h, \widetilde{\nabla} f(x^*)\rangle\\
&+ \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 + 2\gamma\langle\nabla h(x_h), z - z^+\rangle\\
&\overset{(2.3)}{=} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 + 2\gamma\langle\nabla h(x_h) - \nabla h(x^*), z - z^+\rangle. \qquad \square
\end{aligned}$$

B.7. Proof of Theorem 4.1. Let $\eta_k = 2/\lambda_k - 1$. By (3.3), we have

$$2\gamma\langle\nabla h(x_h^k) - \nabla h(x^*), z^k - z^{k+1}\rangle \leq \frac{\gamma^2}{\eta_k}\|\nabla h(x_h^k) - \nabla h(x^*)\|^2 + \eta_k\|z^k - z^{k+1}\|^2. \qquad \text{(B.2)}$$

Hence, for all $k \geq 0$, we have (using $1/\eta_k \leq \lambda_k/\varepsilon^2$ as in (3.3) and (1.15))

$$
\begin{aligned}
4\gamma\underline{\lambda}\sum_{i=0}^{k}(S_f(x_f^i, x^*) + S_h(x_h^i, x^*)) &\leq \sum_{i=0}^{k} 4\gamma\lambda_i(S_f(x_f^i, x^*) + S_h(x_h^i, x^*)) \\
&\overset{(2.11)}{\leq} \sum_{i=0}^{k}\Big(\|z^i - z^*\|^2 - \|z^{i+1} - z^*\|^2 - \eta_i\|z^{i+1} - z^i\|^2 \\
&\qquad + 2\gamma\langle\nabla h(x_h^i) - \nabla h(x^*), z^i - z^{i+1}\rangle\Big) \\
&\overset{(B.2)}{\leq} \sum_{i=0}^{k}\left(\|z^i - z^*\|^2 - \|z^{i+1} - z^*\|^2 + (\gamma^2\lambda_i/\varepsilon^2)\|\nabla h(x_h^i) - \nabla h(x^*)\|^2\right) \\
&\overset{(1.20)}{\leq} \|z^0 - z^*\|^2 - \|z^{k+1} - z^*\|^2 + \frac{(1+\varepsilon)\gamma}{\varepsilon^3(2\beta_V - \gamma)}\|z^0 - z^*\|^2.
\end{aligned}
$$

The "best" convergence rates now follow by taking $k \to \infty$ and using [16, Lemma 3]. In addition, we apply Jensen's inequality to $\|\cdot\|^2$ in the first term to get

$$
\frac{\mu_f}{2}\|\overline{x}_f^k - x^*\|^2 + \frac{\mu_h}{2}\|\overline{x}_h^k - x^*\|^2 \leq \frac{\left(1 + \frac{(1+\varepsilon)\gamma}{\varepsilon^3(2\beta_V - \gamma)}\right)\|z^0 - z^*\|^2}{4\gamma\Lambda_k}.
$$

We now fix $k \geq 0$. For all $\lambda > 0$, define $z_\lambda := (T_{\mathrm{FDRS}})_\lambda(z^k)$. Observe that $S_f(x_f^k, x^*)$ and $S_h(x_h^k, x^*)$ do not depend on the value of λ_k. Therefore, we use (2.11) to get

$$
\begin{aligned}
S_f(x_f^k, x^*) + S_h(x_h^k, x^*) &\leq \inf_{\lambda\in[0,1/\alpha_{\mathrm{FDRS}}^V)} \frac{1}{4\gamma\lambda}\Big(2\gamma\langle\nabla h(x_h^k) - \nabla h(x^*), z^k - z_\lambda\rangle \\
&\quad + \|z^k - z^*\|^2 - \|z_\lambda - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z_\lambda - z^k\|^2\Big) \\
&\overset{(1.10)}{=} \inf_{\lambda\in[0,1/\alpha_{\mathrm{FDRS}}^V)} \frac{1}{4\gamma\lambda}\Big(2\gamma\langle\nabla h(x_h^k) - \nabla h(x^*), z^k - z_\lambda\rangle \\
&\quad + 2\langle z_\lambda - z^*, z^k - z_\lambda\rangle + 2\left(1 - \frac{1}{\lambda}\right)\|z_\lambda - z^k\|^2\Big) \\
&\leq \frac{1}{4\gamma}\left(2\langle z_1 - z^*, z^k - z_1\rangle + \frac{2\gamma}{\beta_V}\|z^k - z^*\|\|z_1 - z^k\|\right) \qquad \text{(B.3)} \\
&\overset{(1.18)}{\leq} \frac{(1+\gamma/\beta_V)\|z^0 - z^*\|^2}{2\gamma\sqrt{\underline{\tau}(k+1)}}
\end{aligned}
$$

where (B.3) uses the $(1/\beta_V)$-Lipschitz continuity of ∇h and the identity $\nabla h(x_h^k) - \nabla h(x^*) = \nabla h(z^k) - \nabla h(z^*)$, and the last line uses the Fejér property $\|z_1 - z^*\| \leq \|z^k - z^*\| \leq \|z^0 - z^*\|$ (see Part 1 of Theorem 1.5). The $o(1/\sqrt{k+1})$ rates follow from (B.3) and the corresponding rates for the FPR in (1.18). □

B.8. Proof of Proposition 5.1. Because ∇f is $(1/\beta_f)$-Lipschitz, we have

$$
f(x_h) \leq f(x_f) + \langle x_h - x_f, \nabla f(x_f)\rangle + \frac{1}{2\beta_f}\|x_h - x_f\|^2; \tag{B.4}
$$

$$
S_f(x_f, x^*) \overset{(2.7)}{\geq} \frac{\beta_f}{2}\|\nabla f(x_f) - \nabla f(x^*)\|^2. \tag{B.5}
$$

where the first inequality follows from [3, Theorem 18.15(iii)]. By applying the identity $z^* - x^* = \gamma\widetilde{\nabla}\chi_V(x^*) = -\gamma\nabla f(x^*) - \gamma\nabla h(x^*)$, the cosine rule (1.10), and the identity $z - z^+ = \lambda(x_h - x_f)$ (see (2.3)) multiple times, we have

$$\begin{aligned}
&2\langle z - z^+, z^* - x^*\rangle + 2\gamma\lambda\langle x_h - x_f, \nabla f(x_f)\rangle = 2\lambda\langle x_h - x_f, \gamma\widetilde{\nabla}\chi_V(x^*) + \gamma\nabla f(x_f)\rangle \\
&= 2\lambda\langle \gamma\widetilde{\nabla}\chi_V(x_h) + \gamma\nabla h(x_h) + \gamma\nabla f(x_f), \gamma\nabla f(x_f) - \gamma\nabla f(x^*)\rangle - 2\langle z - z^+, \gamma\nabla h(x^*)\rangle \\
&= \lambda\Big(\|\gamma\nabla f(x_f) - \gamma\nabla f(x^*)\|^2 + \|x_h - x_f\|^2 \\
&- \gamma^2\|\widetilde{\nabla}\chi_V(x_h) + \nabla h(x_h) - \widetilde{\nabla}\chi_V(x^*) - \nabla h(x^*)\|^2\Big) - 2\langle z - z^+, \gamma\nabla h(x^*)\rangle. \qquad \text{(B.6)}
\end{aligned}$$

By (2.3) (i.e., $z - z^+ = \lambda(x_h - x_f)$), we have

$$\left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + \lambda\left(\frac{\gamma}{\beta_f} + 1\right)\|x_h - x_f\|^2 = \left(1 + \frac{\gamma - \beta_f}{\beta_f\lambda}\right)\|z - z^+\|^2.$$

Therefore,

$$\begin{aligned}
&2\gamma\lambda(f(x_h) + h(x_h) - f(x^*) - h(x^*)) \\
&\overset{\text{(B.4)}}{\leq} 2\gamma\lambda(f(x_f) + h(x_h) - f(x^*) - h(x^*)) + 2\gamma\lambda\langle x_h - x_f, \nabla f(x_f)\rangle + \frac{\gamma\lambda}{\beta_f}\|x_h - x_f\|^2 \\
&\overset{\text{(B.1)}}{\leq} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + 2\langle z - z^+, z^* - x^*\rangle + 2\gamma\lambda\langle x_h - x_f, \nabla f(x_f)\rangle \\
&+ \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2 + 2\gamma\langle\nabla h(x_h), z - z^+\rangle + \frac{\gamma\lambda}{\beta_f}\|x_h - x_f\|^2 - 2\gamma\lambda S_f(x_f, x^*) \\
&\overset{\text{(B.6)}}{\leq} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 + \lambda\left(\frac{\gamma}{\beta_f} + 1\right)\|x_h - x_f\|^2 \\
&+ \lambda\|\gamma\nabla f(x_f) - \gamma\nabla f(x^*)\|^2 + 2\gamma\langle\nabla h(x_h) - \nabla h(x^*), z - z^+\rangle - 2\gamma\lambda S_f(x_f, x^*) \\
&\overset{\text{(B.5)}}{\leq} \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 + \frac{\gamma - \beta_f}{\beta_f\lambda}\right)\|z - z^+\|^2 \\
&+ 2\gamma\langle\nabla h(x_h) - \nabla h(x^*), z - z^+\rangle + \gamma\lambda(\gamma - \beta_f)\|\nabla f(x_f) - \nabla f(x^*)\|^2. \qquad \text{(B.7)}
\end{aligned}$$

If $\gamma \leq \beta_f$, then we can drop the last term. If $\gamma > \beta_f$, then use (2.11) to get

$$\begin{aligned}
\gamma\lambda(\gamma - \beta_f)\|\nabla f(x_f) - \nabla f(x^*)\|^2 \leq \frac{\gamma - \beta_f}{2\beta_f}\Big(&2\gamma\langle\nabla h(x_h) - \nabla h(x^*), z - z^+\rangle \\
&+ \|z - z^*\|^2 - \|z^+ - z^*\|^2 + \left(1 - \frac{2}{\lambda}\right)\|z^+ - z\|^2\Big)
\end{aligned}$$

The result follows by (B.7) and

$$\left(1 + \frac{\gamma - \beta_f}{\beta_f\lambda}\right)\|z - z^+\|^2 + \frac{\gamma - \beta_f}{2\beta_f}\left(1 - \frac{2}{\lambda}\right)\|z - z^+\|^2 = \left(1 + \frac{\gamma - \beta_f}{2\beta_f}\right)\|z - z^+\|^2. \qquad \square$$

B.9. Proof of Theorem 6.6. For all $i \geq 0$, let $c_i := (i/(i+1))^{1/2}$. Let $\kappa_a := (1/2) + 2(a+1)^2$, and let $z^0 := \sqrt{2\alpha\kappa_a}e^{(1/(a+1))} \times \left((\|z_i\|^{-1}/(i+1)^\alpha)z_i\right)_{i\geq 0}$. Then

$\|z^0\|^2 = 2\alpha\kappa_a e^{2/(a+1)} \sum_{i=0}^{\infty}(1/(i+1)^{2\alpha}) < \infty$ and, hence, $z^0 \in \mathcal{H}$. Now for all $i \geq 1$, we have

$$\frac{\|z_i\|^2(a+c_i^2)^2}{c_i^2} \overset{(6.6)}{=} (1-c_i^2) + \frac{(a+c_i^2)^2}{c_i^2} \leq \kappa_a \tag{B.8}$$

because $c_i^2 \in [1/2, 1)$. In addition, for all $i \geq 1$, we have

$$\begin{aligned}\|(P_V)_i z_i^0\|^2 &= \frac{2\alpha\kappa_a e^{2/(a+1)}}{\|z_i\|^2(i+1)^{2\alpha}}\|(P_V)_i z_i\|^2 \overset{(6.6)}{=} \frac{2\alpha\kappa_a e^{2/(a+1)} c_i^2(1-c_i^2)}{\|z_i\|^2(a+c_i^2)^2(i+1)^{2\alpha}} \\ &= \frac{2\alpha\kappa_a e^{2/(a+1)} c_i^2}{\|z_i\|^2(a+c_i^2)^2(i+1)^{1+2\alpha}} \overset{(B.8)}{\geq} \frac{2\alpha e^{2/(a+1)}}{(i+1)^{1+2\alpha}}\end{aligned}$$

where the third equality follows because $1 - c_i^2 = 1 - i/(i+1) = 1/(i+1)$.

Now, for all $k \geq 0$, let $z^{k+1} := T_{\mathrm{FDRS}} z^k$. Again, for all $i \geq 0$, let $b_i := (a + c_i^2)/(a+1) = 1 - (i+1)^{-1}(a+1)^{-1}$ be the eigenvalue of $(T_{\mathrm{FDRS}})_i$ associated to z_i. Note that $b_i^{2k} \geq e^{-2/(1+a)}$ whenever $i \geq k \geq 0$ (hint: use the bound $e^{-1/(a+1)} \leq (1-(i+1)^{-1}(a+1)^{-1})^i = b_i^i$, and note that b_i^{2k} is increasing in i for fixed k). Therefore, for all $k \geq 1$, we have

$$\begin{aligned}\|x_h^k - x^*\|^2 &= \|P_V T_{\mathrm{FDRS}}^k z^0\|^2 = \sum_{i=0}^{\infty} b_i^{2k}\|(P_V)_i z_i^0\|^2 \geq \sum_{i=k}^{\infty} b_i^{2k}\frac{2\alpha e^{2/(a+1)}}{(i+1)^{1+2\alpha}} \\ &\geq \sum_{i=k}^{\infty}\frac{2\alpha}{(i+1)^{1+2\alpha}} \geq \frac{1}{(k+1)^{2\alpha}}.\end{aligned} \tag{B.9}$$

where we use $x^* = 0$ and the lower integral approximation of the sum.

Now we prove the bound for $(x_f^j)_{j\geq 0}$. For all $k \geq 0$, $x_f^k = T_{\mathrm{FDRS}} z^k - \gamma\widetilde{\nabla}\chi_V(x_h^k) = T_{\mathrm{FDRS}} z^k - P_{V^\perp} z^k = (T_{\mathrm{FDRS}} - P_{V^\perp})T_{\mathrm{FDRS}}^k z^0$ (see (2.1)). In addition, for all $i \geq 0$,

$$(T_{\mathrm{FDRS}} - P_{V^\perp})_i = \frac{1}{(a+1)}\begin{bmatrix}0 & -\cos(\theta_i)\sin(\theta_i)\\ 0 & \cos^2(\theta_i) + a - (a+1)\end{bmatrix} = -\frac{\sin(\theta_i)}{(a+1)}\begin{bmatrix}0 & \cos(\theta_i)\\ 0 & \sin(\theta_i)\end{bmatrix}.$$

Thus, for all $i \geq 0$, we have

$$\begin{aligned}\|(T_{\mathrm{FDRS}} - P_{V^\perp})_i z_i^0\|^2 &= \frac{2\alpha\kappa_a e^{2/(a+1)}\sin^2(\theta_i)(\cos^2(\theta_i)+\sin^2(\theta_i))}{\|z_i\|^2(a+1)^2(i+1)^{2\alpha}} = \frac{2\alpha\kappa_a e^{2/(a+1)}(1-c_i^2)}{\|z_i\|^2(a+1)^2(1+i)^{2\alpha}} \\ &\overset{(B.8)}{\geq} \frac{2\alpha e^{2/(a+1)}(a+c_i^2)^2}{c_i^2(a+1)^2(1+i)^{1+2\alpha}}.\end{aligned}$$

where the last inequality follows because $1 - c_i^2 = 1 - i/(i+1) = 1/(i+1)$ and $\kappa_a/\|z_i\|^2 \geq (a+c_i^2)^2/c_i^2$. Note that for all $i \geq 1$, we have $(a+c_i^2)^2/c_i^2 \geq (a+1/2)^2$ because $c_i^2 \in [1/2, 1)$. Therefore, for all $k \geq 1$, we have

$$\begin{aligned}\|x_f^k - x^*\|^2 = \|(T_{\mathrm{FDRS}} - P_{V^\perp})T_{\mathrm{FDRS}}^k z^0\|^2 &\geq \sum_{i=k}^{\infty} b_i^{2k}\frac{2\alpha e^{2/(a+1)}(a+c_i^2)^2}{c_i^2(a+1)^2(1+i)^{1+2\alpha}} \\ &\geq \frac{(a+1/2)^2}{(a+1)^2(k+1)^{2\alpha}}\end{aligned}$$

where we use similar arguments to those used in (B.9). $\square$

REFERENCES

[1] J.-B. Baillon and G. Haddad, *Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones*, Israel Journal of Mathematics, 26 (1977), pp. 137–150.
[2] H. H. Bauschke, J. Y. Bello Cruz, T. T. A. Nghia, H. M. Phan, and X. Wang, *The rate of linear convergence of the Douglas-Rachford algorithm for subspaces is the cosine of the friedrichs angle*, Journal of Approximation Theory, 185 (2014), pp. 63–79.
[3] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer, 2011.
[4] D. P. Bertsekas, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, Optimization for Machine Learning, (2010), pp. 1–38.
[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.
[6] L. M. Briceño-Arias and P. L. Combettes, *A monotone+skew splitting model for composite monotone inclusions in duality*, SIAM Journal on Optimization, 21 (2011), pp. 1230–1250.
[7] L. M. Briceño-Arias, *Forward-Douglas-Rachford splitting and forward-partial inverse method for solving monotone inclusions*, Optimization, 64 (2015), pp. 1239–1261.
[8] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145.
[9] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2 (2011), pp. 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
[10] P. L. Combettes, *Solving monotone inclusions via compositions of nonexpansive averaged operators*, Optimization, 53 (2004), pp. 475–504.
[11] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-point algorithms for inverse problems in science and engineering, Springer, 2011, pp. 185–212.
[12] P. L. Combettes and I. Yamada, *Compositions and convex combinations of averaged nonexpansive operators*, Journal of Mathematical Analysis and Applications, 425 (2015), pp. 55 – 70.
[13] L. Condat, *A Primal–Dual Splitting Method for Convex Optimization Involving Lipschitzian, Proximable and Linear Composite Terms*, Journal of Optimization Theory and Applications, 158 (2013), pp. 460–479.
[14] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, 20 (1995), pp. 273–297.
[15] D. Davis, *Convergence rate analysis of primal-dual splitting schemes*, arXiv preprint arXiv:1408.4419v2, (2014).
[16] D. Davis and W. Yin, *Convergence rate analysis of several splitting schemes*, arXiv preprint arXiv:1406.4834v2, (2014).
[17] ———, *Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions*, arXiv preprint arXiv:1407.5210v2, (2014).
[18] E. Esser, X. Zhang, and T. Chan, *A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046.
[19] N. Komodakis and J.-C. Pesquet, *Playing with Duality: An Overview of Recent Primal-Dual Approaches for Solving Large-Scale Optimization Problems*, arXiv preprint arXiv:1406.5429v2, (2014).
[20] M. Krasnosel'skiĭ, *Zwei Bemerkungen über die Methode der sukzessiven Approximationen.*, Usp. Mat. Nauk, 10 (1955), pp. 123–127.
[21] P. Lions and B. Mercier, *Splitting Algorithms for the Sum of Two Nonlinear Operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.
[22] W. R. Mann, *Mean Value Methods in Iteration*, Proceedings of the American Mathematical Society, 4 (1953), pp. pp. 506–510.
[23] G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, Journal of Mathematical Analysis and Applications, 72 (1979), pp. 383 – 390.
[24] H. Raguet, J. Fadili, and G. Peyré, *A Generalized Forward-Backward Splitting*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1199–1226.
[25] P. Tseng, *A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings*, SIAM Journal on Control and Optimization, 38 (2000), pp. 431–446.
[26] B. C. Vũ, *A splitting algorithm for dual monotone inclusions involving cocoercive operators*, Advances in Computational Mathematics, 38 (2013), pp. 667–681.