

# Single-Forward-Step Projective Splitting: Exploiting Cocoercivity

Patrick R. Johnstone\*      Jonathan Eckstein\*

June 19, 2019

## Abstract

This work describes a new variant of projective splitting for monotone inclusions, in which cocoercive operators can be processed with a single forward step per iteration. This result establishes a symmetry between projective splitting algorithms, the classical forward-backward splitting method (FB), and Tseng’s forward-backward-forward method (FBF). Another symmetry is that the new procedure allows for larger stepsizes for cocoercive operators: the stepsize bound is  $2\beta$  for a  $\beta$ -cocoercive operator, which is the same as for FB. To complete the connection, we show that FB corresponds to an unattainable boundary case of the parameters in the new procedure. Unlike FB, the new method allows for a backtracking procedure when the cocoercivity constant is unknown. Proving convergence of the algorithm requires some departures from the usual proof framework for projective splitting. We close with some computational tests establishing competitive performance for the method.

## 1 Introduction

For a collection of real Hilbert spaces  $\{\mathcal{H}_i\}_{i=0}^n$  consider the *finite-sum convex minimization problem*:

$$\min_{x \in \mathcal{H}_0} \sum_{i=1}^n (f_i(G_i x) + h_i(G_i x)), \quad (1)$$

where every  $f_i : \mathcal{H}_i \rightarrow (-\infty, +\infty]$  and  $h_i : \mathcal{H}_i \rightarrow \mathbb{R}$  is closed, proper, and convex, every  $h_i$  is also differentiable with  $L_i$ -Lipschitz-continuous gradients, and the operators  $G_i : \mathcal{H}_0 \rightarrow \mathcal{H}_i$  are linear and bounded. Under appropriate constraint qualifications, (1) is equivalent to the monotone inclusion problem of finding  $z \in \mathcal{H}_0$  such that

$$0 \in \sum_{i=1}^n G_i^* (A_i + B_i) G_i z \quad (2)$$

---

\*Department of Management Science and Information Systems, Rutgers Business School Newark and New Brunswick, Rutgers University. Contact: patrick.r.johnstone@gmail.com, jeckstei@business.rutgers.edu

where all  $A_i : \mathcal{H}_i \rightarrow 2^{\mathcal{H}_i}$  and  $B_i : \mathcal{H}_i \rightarrow \mathcal{H}_i$  are maximal monotone and each  $B_i$  is  $L_i^{-1}$ -cocoercive, meaning that it is single-valued and

$$L_i \langle B_i x_1 - B_i x_2, x_1 - x_2 \rangle \geq \|B_i x_1 - B_i x_2\|^2$$

for some  $L_i \geq 0$ . In particular, if we set  $A_i = \partial f_i$  and  $B_i = \nabla h_i$  then the solution sets of the two problems coincide under a special case of the constraint qualification of [9, Prop. 5.3]. When  $L_i = 0$ , then  $B_i$  must be a constant operator, that is, there is some  $v_i \in \mathcal{H}_i$  such that  $B_i x = v_i$  for all  $x \in \mathcal{H}_i$ . Defining  $T_i = A_i + B_i$ , problem (2) may be written more compactly as

$$0 \in \sum_{i=1}^n G_i^* T_i G_i z. \quad (3)$$

## 1.1 Background

Operator splitting algorithms are an effective way to solve structured convex optimization problems and monotone inclusions such as (1), (2), and (3). Their defining feature is that they break the problem up into manageable chunks. At each iteration they solve a set of tractable subproblems in such a manner as to converge to a solution of the global problem. Arguably the three most popular operator splitting algorithms are the forward-backward splitting (FB) [11], Douglas/Peaceman-Rachford splitting (DR) [24], and forward-backward-forward (FBF) [38] methods. Indeed, many algorithms in convex optimization and monotone inclusions are in fact instances of one of these methods.

A different and relatively recently proposed class of operator splitting algorithms is *projective splitting*. Projective splitting has a different convergence analysis from most operator splitting schemes: while the convergence of most schemes is obtained by viewing their updates as a fixed-point iteration, often of some firmly non-expansive operator, projective splitting is analyzed (and designed) as a way of creating a sequence of separating hyperplanes between the current iterate and the primal-dual solution set. New iterates are produced by projecting onto these hyperplanes, perhaps with some over- or under-relaxation. The class of methods originated in [18], was extended to sums of  $n \geq 1$  operators in [19], to include compositions with bounded linear maps in [1], to allow asynchronous block-iterative (*i.e.* incremental) updates in [10], and to allow forward steps for Lipschitz continuous operators in [37, 21]. Further theoretical results, including some convergence rates, have been obtained in [20, 22, 27, 26].

In the context of projective splitting, [37, 21] were the first works to move away from computational updates based solely on resolvent steps on each maximal monotone operator  $T_i$  in (3). The analysis in [21] developed a procedure that could instead use two *forward* (explicit or gradient) steps for Lipschitz continuous operators. This innovation constituted significant progress because forward steps are often computationally much cheaper and more convenient than resolvents. However, the result raised a question: if projective splitting can exploit Lipschitz continuity, can it further exploit the presence of *cocoercive* operators? Cocoercivity is in general a stronger property than Lipschitz continuity. However, when an operator is the gradient of a closed proper convex function (such as  $h_i$  in (1)), the Baillon-Haddad theorem [2, 3] establishes that the two properties are equivalent:  $\nabla h_i$  is  $L_i$ -Lipschitz continuous if and only if it is  $L_i^{-1}$ -cocoercive.

Operator splitting methods that exploit cocoercivity rather than mere Lipschitz continuity typically have lower per-iteration computational complexity and a larger range of permissible stepsizes. For example, both FBF and the extragradient (EG) method [23] only require Lipschitz continuity, but need *two* forward steps per iteration and limit the stepsize to  $L^{-1}$ , where  $L$  is the Lipschitz constant. If we strengthen the assumption to  $L^{-1}$ -cocoercivity, we can use FB, which only needs one forward step per iteration and allows stepsizes bounded away from  $2L^{-1}$ . One departure from this pattern is the recently developed method of [29], which only requires Lipschitz continuity but uses just one forward step per iteration. While this property is remarkable, it should be noted that its stepsizes must be bounded by  $(1/2)L^{-1}$ , which is half the allowable stepsize for EG or FBF.

Much like FBF and EG, the projective splitting computation in [21] requires Lipschitz continuity<sup>1</sup>, two forward steps per iteration, and limits the stepsize to be less than  $L^{-1}$  (when not using backtracking). Considering the relationship between FB and FBF/EG leads to following question: is there a variant of projective splitting which converges under the stronger assumption of  $L^{-1}$ -cocoercivity, while processing each cocoercive operator with a single forward step per iteration, allowing stepsizes bounded above by  $2L^{-1}$ ?

In this paper, we answer this question in the affirmative. Referring to (2), the new procedure requires one forward step on  $B_i$  and one resolvent for  $A_i$  at each iteration. When the resolvent is easily computable (for example, when  $A_i$  is the zero map and its resolvent is simply the identity), the new procedure can effectively halve the computation necessary to run the same number of iterations as the previous procedure of [21]. This advantage is equivalent to that of FB over FBF and EG when cocoercivity is present. Another advantage of the proposed method is that it allows for a backtracking linesearch when the cocoercivity constant is unknown, whereas no such variant is currently known for FB.

The analysis of this new method is significantly different from our previous work in [21], using a novel “ascent lemma” (Lemma 16) regarding the separating hyperplanes generated by the algorithm. The new procedure has an interesting connection to the original resolvent calculation used in the projective splitting papers [18, 19, 1, 10]: in Section 2.2 below, we show that the new procedure is equivalent to one iteration of FB applied to evaluating the resolvent of  $T_i = A_i + B_i$ . That is, we can use a single forward-backward step to approximate the operator-processing procedure in [18, 19, 1, 10], but still obtain convergence.

The new procedure has significant potential for asynchronous and incremental implementation following the ideas and techniques of previous projective splitting methods [10, 17, 21]. To keep the analysis manageable, however, we plan to develop such generalizations in a follow-up paper. Here, we will simply assume that every operator is processed once per iteration.

## 1.2 The Optimization Context

Due to its importance in many applications, we give a brief discussion of the significance of the proposed method in the optimization context, i.e. solving (1). For this problem, the new method is a first-order proximal splitting method which *fully splits* the problem: it utilizes the proximal operator for each nonsmooth function  $f_i$ , a single gradient  $\nabla h_i$  for each smooth

---

<sup>1</sup>If backtracking is used then all three of these methods can converge under weaker local continuity assumptions.

function  $h_i$ , and matrix multiplications by  $G_i$  and  $G_i^*$  (not matrix inversions). Beyond these, the only computations at each iteration are a constant number of inner products, norms, scalar multiplications, and vector additions, which can all be carried out with flop counts linear in the dimension of each Hilbert space.

There are a few other first-order proximal splitting methods which can achieve full splitting on (2). The most similar to projective splitting are those in the family of *primal-dual* (PD) splitting methods; see [13, 12, 7, 33] and references therein. In fact, projective splitting is also a kind of primal-dual method, since it produces a primal sequence *and* a dual sequence both converging to a primal-dual solution. However, the convergence mechanisms are entirely different: PD methods are usually built by applying an established operator splitting technique such as FB, FBF, or DR to the appropriately formulated primal-dual inclusion in a primal-dual product space.

Apart from being new, we list two potential advantages of our proposed method over the more established PD schemes. First, unlike the PD methods, the norms  $\|G_i\|$  do not effect the stepsize constraints of our proposed method, making such constraints easier to satisfy. Furthermore, differently from PD methods, the stepsizes may vary at each iteration and may differ for each operator. Second, projective splitting methods allow for asynchronous parallel and incremental implementations in an arguably simpler way than PD methods (although we do not develop this aspect in this paper). In projective splitting it is fairly straightforward to incorporate *deterministic* asynchronous assumptions [10, 17] with deterministic convergence guarantees, with the analysis being similar to the synchronous case. In contrast, existing asynchronous and block-coordinate analyses of PD methods have required *stochastic* assumptions which only lead to probabilistic convergence guarantees [33].

### 1.3 Notation and a Simplifying Assumption

We use the same general notation as in [21, 20, 22]. Summations of the form  $\sum_{i=1}^{n-1} a_i$  will appear throughout this paper. To deal with the case  $n = 1$ , we use the standard convention that  $\sum_{i=1}^0 a_i = 0$ .

We will use a boldface  $\mathbf{w} = (w_1, \dots, w_{n-1})$  for elements of  $\mathcal{H}_1 \times \dots \times \mathcal{H}_{n-1}$ . Let  $\mathcal{H} \triangleq \mathcal{H}_0 \times \mathcal{H}_1 \times \dots \times \mathcal{H}_{n-1}$ , which we refer to as the “collective primal-dual space”, and note that the assumption on  $G_n$  implies that  $\mathcal{H}_n = \mathcal{H}_0$ . We use  $p$  to refer to points in  $\mathcal{H}$ , so  $p \triangleq (z, \mathbf{w}) = (z, w_1, \dots, w_{n-1})$ .

Throughout, we will simply write  $\|\cdot\|_i = \|\cdot\|$  as the norm for  $\mathcal{H}_i$  and let the subscript be inferred from the argument. In the same way, we will write  $\langle \cdot, \cdot \rangle_i$  as  $\langle \cdot, \cdot \rangle$  for the inner product of  $\mathcal{H}_i$ . For the collective primal-dual space we will use a special norm and inner product with its own subscript defined in (15).

We use the standard “ $\rightharpoonup$ ” notation to denote weak convergence, which is of course equivalent to ordinary convergence in finite-dimensional settings.

For any maximal monotone operator  $A$  we will use the notation  $J_{\rho A} \triangleq (I + \rho A)^{-1}$ , for any scalar  $\rho > 0$ , to denote the *resolvent operator*, also known as the proximal, backward, or implicit step with respect to  $A$ . Thus,

$$x = J_{\rho A}(t) \iff x + \rho a = t \text{ and } a \in Ax, \quad (4)$$

and the  $x$  and  $a$  satisfying this relation are unique. Furthermore,  $J_{\rho A}$  is defined everywhere and  $\text{range}(J_A) = \text{dom}(A)$  [4, Prop. 23.2].

For the rest of the paper, we will impose the simplifying assumption

$$G_n : \mathcal{H}_n \rightarrow \mathcal{H}_n \triangleq I \quad (\text{the identity operator}).$$

As noted in [21], the requirement that  $G_n = I$  is not a very restrictive assumption. For example, one can always enlarge the original problem by one operator, setting  $A_n = B_n = 0$ .

## 2 Projective Splitting

The goal of our algorithm will be to find a point in

$$\mathcal{S} \triangleq \left\{ (z, w_1, \dots, w_{n-1}) \in \mathcal{H} \mid (\forall i \in \{1, \dots, n-1\}) w_i \in T_i G_i z, -\sum_{i=1}^{n-1} G_i^* w_i \in T_n z \right\}. \quad (5)$$

It is clear that  $z^*$  solves (2)–(3) if and only if there exist  $w_1^*, \dots, w_{n-1}^*$  such that

$$(z^*, w_1^*, \dots, w_{n-1}^*) \in \mathcal{S}.$$

Under reasonable assumptions, the set  $\mathcal{S}$  is closed and convex; see Lemma 1.  $\mathcal{S}$  is often called the *Kuhn-Tucker solution set* of problem (3).

A separator-projector algorithm for finding a point in  $\mathcal{S}$  (and hence a solution to (3)) will, at each iteration  $k$ , find a closed and convex set  $H_k$  which separates  $\mathcal{S}$  from the current point, meaning  $\mathcal{S}$  is entirely in the set (preferably, the current point is not). One can then attempt to “move closer” to the solution set by projecting the current point onto the set  $H_k$ . This general setup guarantees that the sequence generated by the method is *Fejér monotone* [8] with respect to  $\mathcal{S}$ . This alone is not sufficient to guarantee that the iterates actually converge to a point in the solution set. To establish this, one needs to show that the set  $H_k$  “sufficiently separates” the current point from the solution set, or at least does so sufficiently often. Such “sufficient separation” allows one to establish that any weakly convergent subsequence of the iterates must have its limit in the set  $\mathcal{S}$ , from which overall weak convergence follows from [8, Prop. 2].

With  $\mathcal{S}$  as in (5), the separator formulation presented in [10] constructs the halfspace  $H_k$  using the function

$$\varphi_k(z, w_1, \dots, w_{n-1}) \triangleq \sum_{i=1}^{n-1} \langle G_i z - x_i^k, y_i^k - w_i \rangle + \left\langle z - x_i^n, y_i^n + \sum_{i=1}^{n-1} G_i^* w_i \right\rangle \quad (6)$$

$$= \left\langle z, \sum_{i=1}^n G_i^* y_i^k \right\rangle + \sum_{i=1}^{n-1} \langle x_i^k - G_i x_n^k, w_i \rangle - \sum_{i=1}^n \langle x_i^k, y_i^k \rangle, \quad (7)$$

for some  $(x_i^k, y_i^k) \in \mathcal{H}_i^2$  such that  $y_i^k \in T_i x_i^k$ ,  $i \in 1, \dots, n$ . These points  $(x_i^k, y_i^k)$  will be specified later and must be updated in a specific way in order to guarantee convergence to  $\mathcal{S}$ . From its expression in (7) it is clear that  $\varphi_k$  is an affine function on  $\mathcal{H}$ . Furthermore, from (6) it may easily be verified using the monotonicity of each  $T_i$  that for any  $p = (z, w_1, \dots, w_{n-1}) \in$

$\mathcal{S}$ , one has  $\varphi_k(p) \leq 0$ , so that the separator set  $H_k$  may be taken to be the halfspace  $\{p \mid \varphi_k(p) \leq 0\}$ . Projecting onto this halfspace is a low-complexity operation involving only inner products, norms, matrix multiplication by  $G_i$ , and sums of scalars. For example, when  $\mathcal{H}_i = \mathbb{R}^d$  for  $i = 1, \dots, n$  and each  $G_i = I$ , then this projection has computational complexity  $O(nd)$ .

The key question is how to select the points  $(x_i^k, y_i^k) \in \text{gra } T_i$  so that convergence to  $\mathcal{S}$  may be established. The original approach, given the current iterate  $p^k = (z^k, w_1^k, \dots, w_{n-1}^k)$ , is to choose  $(x_i^k, y_i^k) \in \text{gra } T_i$  to be some function of  $(z^k, w_i^k)$  such that  $\varphi_k(p^k)$  is positive and “sufficiently large” whenever  $p^k \notin \mathcal{S}$ . Then, since the solution set is entirely on the nonpositive side of the hyperplane  $\{p \mid \varphi_k(p) = 0\}$ , projecting the current point onto this hyperplane makes progress toward the solution and can be shown to lead to overall convergence. In the original versions of projective splitting, the calculation of  $(x_i^k, y_i^k)$  involved (perhaps approximately) evaluating a resolvent; [21] introduced the alternative of a two-forward-step calculation for Lipschitz continuous operators. Here, we introduce a one-forward-step calculation for the case of cocoercive operators

A principal difference between the analysis here and earlier work on projective splitting is that processing all the operators  $T_1, \dots, T_n$  at iteration  $k$  need not result in  $\varphi_k(p^k)$  being positive. Instead, we establish an “ascent lemma” that relates the values  $\varphi_k(p^k)$  and  $\varphi_{k-1}(p^{k-1})$  in such a way that overall convergence may still be proved, even though it is possible that  $\varphi_k(p^k) \leq 0$ . When  $\varphi_k(p^k) \leq 0$ , projection onto  $H_k = \{p \mid \varphi_k(p) \leq 0\}$  is effectively a “no-operation” resulting in  $p^{k+1} = p^k$ . In this case, the algorithm continues to compute new points  $(x_i^{k+1}, y_i^{k+1}), (x_i^{k+2}, y_i^{k+2}), \dots$  until, for some  $\ell \geq 0$ , it constructs a hyperplane  $H_{k+\ell}$  such that the  $\varphi_{k+\ell}(p^k) > 0$  and projection results in  $p^{k+\ell+1} \neq p^{k+\ell} = p^k$ .

### Additional Notation for Projective Splitting

For an arbitrary  $(w_1, w_2, \dots, w_{n-1}) \in \mathcal{H}_1 \times \mathcal{H}_2 \times \dots \times \mathcal{H}_{n-1}$  we use the notation

$$w_n \triangleq - \sum_{i=1}^{n-1} G_i^* w_i.$$

Note that when  $n = 1$ ,  $w_1 = 0$ . Under this convention, we may write  $\varphi_k : \mathcal{H} \rightarrow \mathbb{R}$  in the simpler form

$$\varphi_k(z, w_1, \dots, w_{n-1}) = \sum_{i=1}^n \langle G_i z - x_i^k, y_i^k - w_i \rangle.$$

We also use the following notation for  $i = 1, \dots, n$ :

$$\varphi_{i,k}(z, w_i) \triangleq \langle G_i z - x_i^k, y_i^k - w_i \rangle.$$

Note that  $\varphi_k(p) = \varphi_k(z, w_1, \dots, w_{n-1}) = \sum_{i=1}^n \varphi_{i,k}(z, w_i)$ .

## 2.1 The New Procedure

Suppose  $A_i = 0$  for some  $i \in \{1, \dots, n\}$ . Since  $B_i$  is cocoercive, it is also Lipschitz continuous. In [21] we introduced the following two-forward-step update for Lipschitz continuous  $B_i$ :

$$\begin{aligned} x_i^k &= G_i z^k - \rho_i^k (B_i G_i z^k - w_i^k) \\ y_i^k &= B_i x_i^k. \end{aligned}$$

Under  $L_i$ -Lipschitz continuity and the condition  $\rho_i^k < 1/L_i$ , it is possible to show that updating  $(x_i^k, y_i^k)$  in this way leads to  $\varphi_{i,k}(z^k, w_i^k)$  being sufficiently positive to establish overall convergence. Although we did not discuss it in [21], this two-forward step procedure can be extended to handle nonzero  $A_i$  in the following manner:

$$x_i^k + \rho_i^k a_i^k = G_i z^k - \rho_i^k (B_i G_i z^k - w_i^k) : \quad a_i^k \in A_i x_i^k \quad (8)$$

$$y_i^k = a_i^k + B_i x_i^k. \quad (9)$$

Following (4), it is clear that (8) is essentially a resolvent calculation applied to its right-hand side  $G_i z^k - \rho_i^k (B_i G_i z^k - w_i^k)$ . This type of update, with forward steps and backward steps together, was introduced in [37] for a more limited form of projective splitting.

An obvious drawback of (8)–(9) is that it requires two forward steps per iteration, one to compute  $B_i G_i z^k$  and another to compute  $B_i x_i^k$ . The initial motivation for the current paper was the following question: is there a way to reuse  $B_i x_i^{k-1}$  so as to avoid computing  $B_i G_i z^k$  at each iteration, perhaps under the stronger assumption of cocoercivity? With some effort we arrived at the following update for each block  $i = 1, \dots, n$  at each iteration  $k \geq 0$ :

$$x_i^k + \rho_i^k a_i^k = (1 - \alpha_i^k) x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (b_i^{k-1} - w_i^k) : \quad a_i^k \in A_i x_i^k \quad (10)$$

$$b_i^k = B_i x_i^k \quad (11)$$

$$y_i^k = a_i^k + b_i^k, \quad (12)$$

where  $\alpha_i^k \in (0, 1)$ ,  $\rho_i^k \leq 2(1 - \alpha_i^k)/L_i$ , and  $b_i^0 = B_i x_i^0$ . Condition (10) is readily satisfied by some simple linear algebra calculations followed by a resolvent calculation involving  $A_i$ . Subsequently, (11) requires only an evaluation (forward step) on  $B_i$ , and (12) is a simple vector addition. In comparison to (8), we have replaced  $B_i G_i z^k$  with the previously computed point  $B_i x_i^{k-1}$ . However, in order to establish convergence, it turns out that we also need to replace  $G_i z^k$  with a convex combination of  $x_i^{k-1}$  and  $G_i z^k$ . Note that the stepsize constraint can now be made arbitrarily close to  $2/L_i$  by setting  $\alpha_i^k$  arbitrarily close to 0. However, in practice it may be better to use an intermediate value, such as  $\alpha_i^k = 0.1$ , since this allows the update to make significant use of the information in  $z^k$ , a point computed more recently than  $x_i^{k-1}$ .

Computing  $(x_i^k, y_i^k)$  in this way does not guarantee that  $\varphi_{i,k}(z^k, w_i^k)$  is positive. In the next section, we give some intuition as to why (10)–(12) nevertheless leads to convergence to  $\mathcal{S}$ .

## 2.2 A Connection with the Forward-Backward Method

In the projective splitting literature preceding [21], the pairs  $(x_i^k, y_i^k)$  are solutions of

$$x_i^k + \rho_i^k y_i^k = G_i z^k + \rho_i^k w_i^k : \quad y_i^k \in T_i x_i^k \quad (13)$$

for some  $\rho_i^k > 0$ , which — following (4) — is a resolvent calculation. It can be shown that the resulting  $(x_i^k, y_i^k) \in \text{gra } T_i$  are such that  $\varphi_{i,k}(z^k, w_i^k)$  is positive and sufficiently large to guarantee overall convergence to a solution of (3). Since the stepsize  $\rho_i^k$  in (13) can be any positive number, let us replace  $\rho_i^k$  with  $\rho_i^k/\alpha_i^k$  for some  $\alpha_i^k \in (0, 1)$  and rewrite (13) as

$$x_i^k + \frac{\rho_i^k}{\alpha_i^k} y_i^k = G_i z^k + \frac{\rho_i^k}{\alpha_i^k} w_i^k : \quad y_i^k \in T_i x_i^k. \quad (14)$$

The reason for this reparameterization will become apparent below.

In this paper,  $T_i = A_i + B_i$ , with  $B_i$  being cocoercive and  $A_i$  maximal monotone. For  $T_i$  in this form, computing the resolvent as in (13) exactly may be impossible, even when the resolvent of  $A_i$  is available. With this structure,  $x_i^k$  in (14) satisfies:

$$0 \in \frac{\rho_i^k}{\alpha_i^k} A_i x_i^k + \frac{\rho_i^k}{\alpha_i^k} B_i x_i^k + x_i^k - \left( G_i z^k + \frac{\rho_i^k}{\alpha_i^k} w_i^k \right)$$

which can be rearranged to  $0 \in A_i x_i^k + \tilde{B}_i x_i^k$ , where

$$\tilde{B}_i v = B_i v + \frac{\alpha_i^k}{\rho_i^k} \left( v - G_i z^k - \frac{\rho_i^k}{\alpha_i^k} w_i^k \right).$$

Since  $B_i$  is  $L_i^{-1}$ -cocoercive,  $\tilde{B}_i$  is  $(L_i + \alpha_i^k/\rho_i^k)^{-1}$ -cocoercive [4, Prop. 4.12]. Consider the generic monotone inclusion problem  $0 \in A_i x + \tilde{B}_i x$ :  $A_i$  is maximal and  $\tilde{B}_i$  is cocoercive, and thus one may solve the problem with the forward-backward (FB) method [4, Theorem 26.14]. If one applies a single iteration of FB initialized at  $x_i^{k-1}$ , with stepsize  $\rho_i^k$ , to the inclusion  $0 \in A_i x + \tilde{B}_i x$ , one obtains the calculation:

$$\begin{aligned} x_i^k &= J_{\rho_i^k A_i} \left( x_i^{k-1} - \rho_i^k \tilde{B}_i x_i^{k-1} \right) \\ &= J_{\rho_i^k A_i} \left( x_i^{k-1} - \rho_i^k \left( B_i x_i^{k-1} + \frac{\alpha_i^k}{\rho_i^k} \left( x_i^{k-1} - G_i z^k - \frac{\rho_i^k}{\alpha_i^k} w_i^k \right) \right) \right) \\ &= J_{\rho_i^k A_i} \left( (1 - \alpha_i^k) x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (B_i x_i^{k-1} - w_i^k) \right), \end{aligned}$$

which is precisely the update (10). So, our proposed calculation is equivalent to one iteration of FB initialized at the previous point  $x_i^{k-1}$ , applied to the subproblem of computing the resolvent in (14). Prior versions of projective splitting require computing this resolvent either exactly or to within a certain relative error criterion, which may be costly. Here, we simply make a single FB step toward computing the resolvent, which we will prove is sufficient for the projective splitting method to converge to  $\mathcal{S}$ . However, our stepsize restriction on  $\rho_i^k$  will be slightly stronger than the natural stepsize limit that would arise when applying FB to  $0 \in A_i x + \tilde{B}_i x$

### 3 The Algorithm: Assumptions, Definition, and Basic Properties

**Assumption 1.** *Problem (2) conforms to the following:*



1.  $\mathcal{H}_0 = \mathcal{H}_n$  and  $\mathcal{H}_1, \dots, \mathcal{H}_{n-1}$  are real Hilbert spaces.
2. For  $i = 1, \dots, n$ , the operators  $A_i : \mathcal{H}_i \rightarrow 2^{\mathcal{H}_i}$  and  $B_i : \mathcal{H}_i \rightarrow \mathcal{H}_i$  are monotone. Additionally each  $A_i$  is maximal.
3. Each operator  $B_i$  is either  $L_i^{-1}$ -cocoercive for some  $L_i > 0$  (and thus single-valued) and  $\text{dom } B_i = \mathcal{H}_i$ , or  $L_i = 0$  and  $B_i x = v_i$  for all  $x \in \mathcal{H}_i$  and some  $v_i \in \mathcal{H}_i$ .
4. Each  $G_i : \mathcal{H}_0 \rightarrow \mathcal{H}_i$  for  $i = 1, \dots, n-1$  is linear and bounded.
5. Problem (2) has a solution, so the set  $\mathcal{S}$  defined in (5) is nonempty.

In order to apply a separator-projector algorithm, the target set must be closed and convex. Establishing this for  $\mathcal{S}$  is very similar to in our previous work [21], which in turn follows many earlier results.

**Lemma 1.** *Suppose Assumption 1 holds. The set  $\mathcal{S}$  defined in (5) is closed and convex.*

*Proof.* By [4, Cor. 20.28] each  $B_i$  is maximal. Furthermore, since  $\text{dom}(B_i) = \mathcal{H}_i$ ,  $T_i = A_i + B_i$  is maximal monotone by [4, Cor. 25.5(i)]. The rest of the proof is identical to [21, Lemma 3].  $\square$

Throughout, we will use  $p = (z, \mathbf{w}) = (z, w_1, \dots, w_{n-1})$  for a generic point in  $\mathcal{H}$ , the collective primal-dual space. For  $\mathcal{H}$ , we adopt the following (standard) norm and inner product:

$$\|(z, \mathbf{w})\|^2 \triangleq \|z\|^2 + \sum_{i=1}^{n-1} \|w_i\|^2 \quad \langle (z^1, \mathbf{w}^1), (z^2, \mathbf{w}^2) \rangle \triangleq \langle z^1, z^2 \rangle + \sum_{i=1}^{n-1} \langle w_i^1, w_i^2 \rangle. \quad (15)$$

**Lemma 2.** [21, Lemma 4] *Let  $\varphi_k$  be defined as in (6). Then:*

1.  $\varphi_k$  is affine on  $\mathcal{H}$ .
2. With respect to inner product  $\langle \cdot, \cdot \rangle$  on  $\mathcal{H}$ , the gradient of  $\varphi_k$  is

$$\nabla \varphi_k = \left( \sum_{i=1}^{n-1} G_i^* y_i^k + y_n^k, x_1^k - G_1 x_n^k, x_2^k - G_2 x_n^k, \dots, x_{n-1}^k - G_{n-1} x_n^k \right).$$

We sharpen the notation for the one-forward-step update introduced in (10)–(12) as follows:

**Definition 1.** *Suppose  $\mathcal{H}$  and  $\mathcal{H}'$  are real Hilbert spaces,  $A : \mathcal{H} \rightarrow 2^{\mathcal{H}}$  is maximal-monotone with nonempty domain,  $B : \mathcal{H} \rightarrow \mathcal{H}$  is  $L^{-1}$ -cocoercive, and  $G : \mathcal{H}' \rightarrow \mathcal{H}$  is bounded and linear. For  $\alpha \in [0, 1]$  and  $\rho > 0$ , define the mapping  $\mathcal{F}_{\alpha, \rho}(z, x, w; A, B, G) : \mathcal{H}' \times \mathcal{H}^2 \rightarrow \mathcal{H}^2$ , with additional parameters  $A, B$ , and  $G$ , as*

$$\mathcal{F}_{\alpha, \rho} \left( \begin{array}{c} z, x, w; \\ A, B, G \end{array} \right) := (x^+, y^+) : \begin{cases} t & \triangleq (1 - \alpha)x + \alpha Gz - \rho(Bx - w) \\ x^+ & = J_{\rho A}(t) \\ y^+ & = \rho^{-1}(t - x^+) + Bx^+. \end{cases} \quad (16)$$

---

**Algorithm 1:** One-Forward-Step Projective Splitting with Backtracking
 

---

**Input:**  $(z^1, \mathbf{w}^1) \in \mathcal{H}$ ,  $\{\beta_k\}_{k \in \mathbb{N}}$ ,  $\mathcal{B} \subseteq \{1, \dots, n\}$  the operators requiring backtracking,  $\gamma > 0$ ,  $\delta \in (0, 1)$ . For  $i = 1, \dots, n$ :  $x_i^0 \in \mathcal{H}_i$  and  $\{\alpha_i^k\}_{k \in \mathbb{N}}$ . For  $i \in \mathcal{B}$ :  $\{\hat{\rho}_i^k\}_{k \in \mathbb{N}}$ ,  $\hat{\theta}_i \in \text{dom}(A_i)$ ,  $\hat{w}_i \in A_i \hat{\theta}_i + B_i \hat{\theta}_i$ , and  $y_i^0 \in A_i x_i^0 + B_i x_i^0$ . For  $i \notin \mathcal{B}$ :  $\{\rho_i^k\}_{k \in \mathbb{N}}$ .

- 1 **for**  $k = 1, 2, \dots$  **do**
- 2     **for**  $i \in \mathcal{B}$  **do**
- 3          $\varphi_{i,k-1}(z^k, w_i^k) \leftarrow \langle G_i z^k - x_i^{k-1}, y_i^{k-1} - w_i^k \rangle$
- 4          $(x_i^k, y_i^k, \rho_i^k) \leftarrow \text{backTrack}(z^k, w_i^k, x_i^{k-1}, y_i^{k-1}, \hat{\rho}_i^k, \varphi_{i,k-1}(z^k, w_i^k), \alpha_i^k, \hat{\theta}_i, \hat{w}_i, \delta; A_i, B_i, G_i)$
- 5     **for**  $i \notin \mathcal{B}$  **do**
- 6          $(x_i^k, y_i^k) \leftarrow \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i)$      /\*  $\mathcal{F}$  defined in (16)     \*/
- 7      $(\pi_k, z^{k+1}, \mathbf{w}^{k+1}) \leftarrow \text{projectToHplane}(z^k, \mathbf{w}^k, \{x_i^k, y_i^k\}_{i=1}^n, \beta_k, \gamma)$
- 8     **if**  $\pi_k = 0$  **then**
- 9         **return**  $z^{k+1}$

---



---

**Algorithm 2:** Backtracking procedure
 

---

**Function**  $\text{backTrack}(z, w, x, y, \hat{\rho}, \varphi, \alpha, \hat{\theta}, \hat{w}, \delta; A, B, G)$ :

- 2      $\tilde{\rho}_1 = \hat{\rho}$
- 3     **for**  $j = 1, 2, \dots$  **do**
- 4          $(\tilde{x}_j, \tilde{y}_j) = \mathcal{F}_{\alpha, \tilde{\rho}_j}(z, x, w; A, B, G)$      /\*  $\mathcal{F}$  defined in (16)     \*/
- 5          $\hat{y}_j = \tilde{\rho}_j^{-1}((1 - \alpha)x + \alpha Gz - \tilde{x}_j) + w$
- 6          $\varphi_j^+ = \langle Gz - \tilde{x}_j, \tilde{y}_j - w \rangle$
- 7         **if**  $\|\tilde{x}_j - \hat{\theta}\| \leq (1 - \alpha)\|x - \hat{\theta}\| + \alpha\|Gz - \hat{\theta}\| + \tilde{\rho}_j\|w - \hat{w}\|$
- 8             **and**  $\varphi_j^+ \geq \frac{\tilde{\rho}_j}{2\alpha}(\|\tilde{y}_j - w\|^2 + \alpha\|\hat{y}_j - w\|^2) + (1 - \alpha)\left(\varphi - \frac{\tilde{\rho}_j}{2\alpha}\|y - w\|^2\right)$  **then**
- 9             **return**  $(\tilde{x}_j, \tilde{y}_j, \tilde{\rho}_j)$
- 10          $\tilde{\rho}_{j+1} = \delta \tilde{\rho}_j$

---

With this notation, (10)–(12) can be written as

$$(x_i^k, y_i^k) = \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i).$$

Algorithms 1–3 define the main method proposed in this work. They produce a sequence of primal-dual iterates  $p^k = (z^k, w_1^k, \dots, w_{n-1}^k) \in \mathcal{H}$  and, implicitly,  $w_n^k \triangleq -\sum_{i=1}^{n-1} G_i^* w_i^k$ . Algorithm 1 gives the basic outline of our method; for each operator, it invokes either our new one-forward-step update with a user-defined stepsize (through line 6) or its backtracking variant given in Algorithm 2 (through line 4). Together, algorithms 1–2 specify how to update the points  $(x_i^k, y_i^k)$  used to define the separating affine function  $\varphi_k$  in (6). Algorithm 3, called from line 7 of Algorithm 1, defines the `projectToHplane` function that performs the projection step to obtain the next iterate.

Taken together, algorithms 1–3 are essentially the same as Algorithm 2 of [21], except that the update of  $(x_i^k, y_i^k)$  uses the new procedure given in (10)–(12). For simplicity, the

---

**Algorithm 3:** Projection Update
 

---

**Global Variables for Function:**  $G_i : \mathcal{H}_0 \rightarrow \mathcal{H}_i$  for  $i = 1, \dots, n - 1$

```

1 Function projectToHplane( $z, \mathbf{w}, \{x_i, y_i\}_{i=1}^n, \beta, \gamma$ ):
2    $u_i = x_i - G_i x_n, \quad i = 1, \dots, n - 1,$ 
3    $v = \sum_{i=1}^{n-1} G_i^* y_i + y_n$ 
4    $\pi = \|u\|^2 + \gamma^{-1} \|v\|^2$ 
5   if  $\pi > 0$  then
6      $\varphi(p) = \langle z, v \rangle + \sum_{i=1}^{n-1} \langle w_i, u_i \rangle - \sum_{i=1}^n \langle x_i, y_i \rangle$ 
7      $\tau = \frac{\beta}{\pi} \max \{0, \varphi(p)\}$ 
8   else
9     return  $(0, x_n, y_1, \dots, y_{n-1})$ 
10   $z^+ = z - \gamma^{-1} \tau v$ 
11   $w_i^+ = w_i - \tau u_i, \quad i = 1, \dots, n - 1,$ 
12   $w_n^+ = - \sum_{i=1}^{n-1} G_i^* w_i^+$ 
13  return  $(\pi, z^+, \mathbf{w}^+)$ 

```

---

algorithm also lacks the block-iterative and asynchronous features of [10, 17, 21], which we plan to combine with algorithms 1–3 in a follow-up paper.

The computations in `projectToHplane` are all straightforward and of relatively low complexity. They consist of matrix multiplies by  $G_i$ , inner products, norms, and sums of scalars. In particular, there are no potentially difficult minimization problems involved. If  $G_i = I$  and  $\mathcal{H}_i = \mathbb{R}^d$  for  $i = 1, \dots, n$ , then the computational complexity of `projectToHplane` is  $O(nd)$ .

The method allows two ways to select the stepsizes  $\{\rho_i^k\}_{k \in \mathbb{N}}$ . One may either choose them directly or invoke the `backTrack` procedure. In the backtracking case, one must supply initial trial stepsizes  $\{\hat{\rho}_i^k\}_{k \in \mathbb{N}}$ . If one decides to select the stepsizes manually, the upper bound condition  $\rho_i^k \leq 2(1 - \alpha_i^k)/L_i$  is required whenever  $L_i > 0$ . However, it may be difficult to ensure that this condition is satisfied when the cocoercivity constant is hard to estimate. The global cocoercivity constant  $L_i$  may also be conservative in parts of the domain of  $B_i$ , leading to unnecessarily small stepsizes in some cases. We developed the backtracking linesearch technique for these reasons. The set  $\mathcal{B}$  holds the indices of operators for which backtracking is to be used.

For a trial stepsize  $\tilde{\rho}_j$ , Algorithm 2 generates candidate points  $(\tilde{x}_j, \tilde{y}_j)$  using the single-forward-step procedure of (16). For these candidates, Algorithm 2 checks two conditions on lines 7–8. If both of these inequalities are satisfied, then backtracking terminates and returns the successful candidate points. If either condition is not satisfied, the stepsize is reduced by the factor  $\delta \in (0, 1)$  and the process is repeated. These two conditions arise in the analysis in Section 5.

The `backTrack` procedure computes several auxiliary quantities used to check the two backtracking termination conditions. The point  $\hat{y}_j$  is calculated to be the same as  $\hat{y}$  given in Definition 2. The quantity  $\varphi_j^+ = \langle Gz - \tilde{x}_j, \tilde{y}_j - w \rangle$  is the value of  $\varphi_{i,k}(z^k, w_i^k)$  corresponding to the candidate points  $(\tilde{x}_j, \tilde{y}_j)$ . Note that the subroutine input  $\varphi_j = \langle G_i z^k - x_i^{k-1}, y_i^{k-1} - w_i^k \rangle$

is equal to  $\varphi_{i,k-1}(z^k, w_i^k)$ . Ideally, we want  $\varphi_j^+$  to be as large as possible, but the condition checked on line 8 will ultimately suffice to prove convergence.

Algorithm 1 has the following parameters. Precise assumptions regarding these parameters are given at the end of this section and in Section 5.5.

$\{\alpha_i^k\}_{k \in \mathbb{N}}$ : as described in Section 2.2.

$\{\rho_i^k\}_{k \in \mathbb{N}}$ : Note that  $\rho_i^k$  has two meanings. For  $i \notin \mathcal{B}$ , it is the user-supplied stepsize. For  $i \in \mathcal{B}$ , it is the stepsize returned by `backTrack` on line 4.

$\{\hat{\rho}_i^k\}_{k \in \mathbb{N}}$ : the initial trial stepsizes for the backtracking procedure for  $i \in \mathcal{B}$ .

$(\hat{\theta}_i, \hat{w}_i)$  these are used in the backtracking procedure for  $i \in \mathcal{B}$ . An obvious choice which we used in our numerical experiments was  $(\hat{\theta}_i, \hat{w}_i) = (x_i^0, y_i^0)$ , i.e. the initial point.

$\{\beta_k\}_{k \in \mathbb{N}}$ : relaxation factors for the projection operation.

$\gamma > 0$ : allows for the projection to be performed using a slightly more general primal-dual metric than (15). In effect, this parameter changes the relative size of the primal and dual updates in lines 10–11 of Algorithm 3. As  $\gamma$  increases, a smaller step is taken in the primal and a larger step in the dual. As  $\gamma$  decreases, a smaller step is taken in the dual update and a larger step is taken in the primal. See [19, Sec. 5.1] and [18, Sec. 4.1] for more details.

As written, Algorithm 1 is not as efficient as it could be. On the surface, it seems that we need to recompute  $B_i x_i^{k-1}$  in order to evaluate  $\mathcal{F}$  on line 6. However,  $B_i x_i^{k-1}$  was already computed in the previous iteration and can obviously be reused, so only one evaluation of  $B_i$  is needed per iteration. Similarly, within `backTrack`, each invocation of  $\mathcal{F}$  on line 4 may reuse the quantity  $Bx = B_i x_i^{k-1}$  which was computed in the previous iteration of Algorithm 1. Thus, each iteration of the loop within `backTrack` requires one new evaluation of  $B$ , to compute  $B\tilde{x}_j$  within  $\mathcal{F}$ .

Finally, we note that  $\rho_i^k$  is returned by `backTrack` only to make the notation within the proofs more streamlined. With this notation, assuming that `backTrack` always terminates finitely (which we will show to be the case), we may write

$$(x_i^k, y_i^k) = \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i)$$

for all  $i = 1, \dots, n$ . The only difference between  $i \in \mathcal{B}$  and  $i \notin \mathcal{B}$  is that in the former, the stepsize  $\rho_i^k$  is discovered by backtracking, while in the latter it is directly user-supplied.

We now list most of our parameter assumptions; however, one additional assumption is introduced in Section 5.5, where it arises in the analysis.

**Assumption 2.** *There exist  $0 < \underline{\beta} \leq \bar{\beta} < 2$  such that  $\underline{\beta} \leq \beta_k \leq \bar{\beta}$  for all  $k \geq 1$ . The constant  $\gamma$  is positive. The constant  $\delta \in (0, 1)$ .*

**Assumption 3.** *For all  $i = 1, \dots, n$  there exists  $0 < \underline{\alpha}_i \leq \bar{\alpha}_i \leq 1$  such that  $\underline{\alpha}_i \leq \alpha_i^k \leq \bar{\alpha}_i$  for all  $k \geq 1$ . There exists  $c_\alpha \in (0, 1)$  such that for all  $k \geq 1$  and  $i = 1, \dots, n$ ,*

$$\alpha_i^k \leq \frac{c_\alpha \alpha_i^{k-1}}{1 - \alpha_i^{k-1}}. \quad (17)$$

**Assumption 4.** For  $i \notin \mathcal{B}$  and for all  $k \geq 1$ :

1. There exists  $\underline{\rho}_i$  such that  $0 < \underline{\rho}_i \leq \rho_i^k$ .
2. If  $L_i = 0$ , there exists  $\bar{\rho}_i > 0$  such that  $\rho_i^k \leq \bar{\rho}_i$ .
3. If  $L_i > 0$ ,  $\rho_i^k \leq 2(1 - \alpha_i^k)/L_i$ .

**Assumption 5.** For  $i \in \mathcal{B}$ , and for all  $k \geq 1$ , there exists  $\bar{\rho}_i > 0$  such that  $\hat{\rho}_i^k \leq \bar{\rho}_i$ . Furthermore, either there exists  $\underline{\hat{\rho}}_i > 0$  such that

$$(\forall k \geq 1) : \quad \hat{\rho}_i^k \geq \underline{\hat{\rho}}_i, \quad \text{or} \quad (18)$$

$$(\forall k \geq 2) : \quad \hat{\rho}_i^k \geq \rho_i^{k-1} \quad \text{and} \quad \hat{\rho}_i^1 > 0. \quad (19)$$

If  $\alpha_i^k = \alpha_i$  is fixed, then (17) is satisfied by any  $c_\alpha \geq 1 - \alpha_i$ . When  $L_i > 0$ , to satisfy Assumption 4(3) without forcing  $\rho_i^k = 0$  (which contradicts Assumption 4(1)) requires that  $\alpha_i^k < 1$ . When  $L_i = 0$ , then  $B_i$  must be a constant-valued operator. In this case, we allow  $\alpha_i^k \leq \bar{\alpha}_i \leq 1$ .

Lemma 3 details the key results for Algorithm 1 that stem from it being a separator-projector algorithm. While these properties alone do not guarantee convergence, they are important to all of the arguments that follow.

**Lemma 3.** Suppose that assumptions 1 and 2 hold. Then for Algorithm 1

1. The sequence  $\{p^k\} = \{(z^k, w_1^k, \dots, w_{n-1}^k)\}$  is bounded.
2. If the algorithm never terminates via line 9,  $p^k - p^{k+1} \rightarrow 0$ . Furthermore  $z^k - z^{k-1} \rightarrow 0$  and  $w_i^k - w_i^{k-1} \rightarrow 0$  for  $i = 1, \dots, n$ .
3. If the algorithm never terminates via line 9 and  $\|\nabla\varphi_k\|$  remains bounded for all  $k \geq 1$ , then  $\limsup_{k \rightarrow \infty} \varphi_k(p^k) \leq 0$ .

*Proof.* Parts 1–2 are proved in lemmas 2 and 6 of [21]. Part 3 can be found in Part 1 of the proof of Theorem 1 in [21]. The analysis in [21] uses a different procedure to construct the pairs  $(x_i^k, y_i^k)$ , but the result is generic and not dependent on that particular procedure. Note also that [21] establishes the results in a more general setting allowing asynchrony and block-iterativeness, which we do not analyze here.  $\square$

## 4 The Special Case $n = 1$

Before starting the analysis, we consider the important special case  $n = 1$ . In this case, we have by assumption that  $G_1 = I$ ,  $w_1^k = 0$ , and we are solving the problem  $0 \in Az + Bz$ , where both operators are maximal monotone and  $B$  is  $L^{-1}$ -cocoercive. In this case, Algorithm 1 reduces to a method which is similar to FB. Let  $x^k \triangleq x_1^k$ ,  $y^k \triangleq y_1^k$ ,  $\alpha^k \triangleq \alpha_1^k$ , and  $\rho^k \triangleq \rho_1^k$ .

Assuming for simplicity that  $\mathcal{B} = \{\emptyset\}$ , meaning backtracking is not being used, then the updates carried out by the algorithm are

$$\begin{aligned} x^k &= J_{\rho^k A} \left( (1 - \alpha^k)x^{k-1} + \alpha^k z^k - \rho^k Bx^{k-1} \right) \\ y^k &= Bx^k + \frac{1}{\rho^k} \left( (1 - \alpha^k)x^{k-1} + \alpha^k z^k - \rho^k Bx^{k-1} - x^k \right) \\ z^{k+1} &= z^k - \tau^k y^k, \quad \text{where } \tau^k = \frac{\max\{\langle z^k - x^k, y^k \rangle, 0\}}{\|y^k\|^2}. \end{aligned} \tag{20}$$

If  $\alpha_1^k = 0$ , then for all  $k \geq 2$ , the iterates computed in (20) reduce simply to

$$x^k = J_{\rho^k A} (x^{k-1} - \rho^k Bx^{k-1})$$

which is exactly FB. However,  $\alpha_1^k = 0$  is not allowed in our analysis. Thus, FB is a forbidden boundary case which may be approached by setting  $\alpha_1^k$  arbitrarily close to 0. As  $\alpha_1^k$  approaches 0, the stepsize constraint  $\rho_i^k \leq 2(1 - \alpha_i^k)/L$  approaches the classical stepsize constraint for FB:  $\rho^k \leq 2/L - \epsilon$  for some arbitrarily small constant  $\epsilon > 0$ . A potential benefit of Algorithm 1 over FB in the  $n = 1$  case is that it does allow for backtracking when  $L$  is unknown or only a conservative estimate is available.

## 5 Main Proof

The core of the proof strategy will be to establish (21) below. If this can be done, then weak convergence to a solution follows from part 3 of Theorem 1 in [21].

**Lemma 4.** *Suppose assumptions 1 and 2 hold and Algorithm 1 produces an infinite sequence of iterations without terminating via Line 9. If*

$$(\forall i = 1, \dots, n) : \quad y_i^k - w_i^k \rightarrow 0 \text{ and } G_i z^k - x_i^k \rightarrow 0, \tag{21}$$

*then there exists  $(\bar{z}, \bar{\mathbf{w}}) \in \mathcal{S}$  such that  $(z^k, \mathbf{w}^k) \rightharpoonup (\bar{z}, \bar{\mathbf{w}})$ . Furthermore,  $x_i^k \rightharpoonup G_i \bar{z}$  and  $y_i^k \rightharpoonup \bar{w}_i$  for all  $i = 1, \dots, n - 1$ ,  $x_n^k \rightharpoonup \bar{z}$ , and  $y_n^k \rightharpoonup -\sum_{i=1}^{n-1} G_i^* \bar{w}_i$ .*

*Proof.* Equivalent to part 3 of the proof of Theorem 1 in [21]. □

Lemma 4 can be intuitively understood as follows. If we define, for all  $k \geq 1$ ,

$$\epsilon_k = \max_{i=1, \dots, n} \left\{ \max \left\{ \|y_i^k - w_i^k\|, \|G_i z^k - x_i^k\| \right\} \right\},$$

then (21) is equivalent to saying that  $\epsilon_k \rightarrow 0$ . For all  $k \geq 1$ , we have  $(x_i^k, y_i^k) \in \text{gra } T_i$ . If  $\epsilon_k = 0$ , then  $w_i^k = y_i^k \in T_i x_i^k = T_i G_i z^k$  and since  $\sum_{i=1}^n G_i^* w_i^k = 0$ , it follows that  $(z^k, \mathbf{w}^k) \in \mathcal{S}$  and  $z^k$  solves (3). Thus  $\epsilon_k$  can be thought of as the ‘‘residual’’ of the algorithm which measures how far it is from finding a point in  $\mathcal{S}$  and a solution to (3). In finite dimension, it is straightforward to show that if  $\epsilon_k \rightarrow 0$ ,  $(z^k, \mathbf{w}^k)$  must converge to some element of  $\mathcal{S}$ . This can be done using Fejér monotonicity [4, Theorem 5.5] combined with the fact that the graph of a maximal-monotone operator in a finite-dimensional Hilbert space is closed [4,

Proposition 20.38]. However in the general Hilbert space setting the proof is more delicate, since the graph of a maximal-monotone operator is not in-general closed in the weak-to-weak topology [4, Example 20.39]. Nevertheless the overall result was established in the general Hilbert space setting in part 3 of Theorem 1 of [21], which is itself an instance of [1, Proposition 2.4] (see also [4, Proposition 26.5]). An arguably more transparent proof can be found in [16] (this proof is only for the case  $n = 2$ , but it can be extended).

In order to establish (21), we start by establishing certain contractive and “ascent” properties for the mapping  $\mathcal{F}$ , and also show that the backtracking procedure terminates finitely. Then, we prove the boundedness of  $x_i^k$  and  $y_i^k$ , in turn yielding the boundedness of the gradients  $\nabla\varphi_k$  and hence the result that  $\limsup_{k \rightarrow \infty} \{\varphi_k(p^k)\} \leq 0$  by Lemma 3. Next we establish a “Lyapunov-like” recursion for  $\varphi_{i,k}(z^k, w_i^k)$ , relating  $\varphi_{i,k}(z^k, w_i^k)$  to  $\varphi_{i,k-1}(z^{k-1}, w_i^{k-1})$ . Eventually this result will allow us to establish that  $\liminf_k \varphi_k(p^k) \geq 0$  and hence that  $\varphi_k(p^k) \rightarrow 0$ , which will in turn allow an argument that  $y_i^k - w_i^k \rightarrow 0$ . The proof that  $G_i z^k - x_i^k \rightarrow 0$  will then follow fairly elementary arguments.

The primary innovations of the upcoming proof are the ascent lemma and the way that it is used in Lemma 17 to establish  $\varphi_k(p^k) \rightarrow 0$  and  $y_i^k - w_i^k \rightarrow 0$ . This technique is a significant deviation from previous analyses in the projective splitting family. In previous work, the strategy was to show that  $\varphi_{i,k}(z^k, w_i^k) \geq C \max\{\|G_i z^k - x_i^k\|^2, \|y_i^k - w_i^k\|^2\}$  for a constant  $C > 0$ , which may be combined with  $\limsup \varphi_k(p^k) \leq 0$  to imply (21). In contrast, in the algorithm of this paper we cannot establish such a result and in fact  $\varphi_k(p^k)$  may be negative. Instead, we relate  $\varphi_k(p^k)$  to  $\varphi_{k-1}(p^{k-1})$  to show that the separation improves at each iteration in a way which still leads to overall convergence.

## 5.1 Some Basic Results

We begin by stating three elementary results on sequences, which may be found in [34], and a basic, well known nonexpansivity property for forward steps with cocoercive operators.

**Lemma 5.** [34, Lemma 1, Ch. 2] *Suppose  $a_k \geq 0$  for all  $k \geq 1$ ,  $b \geq 0$ ,  $0 \leq \tau < 1$ , and  $a_{k+1} \leq \tau a_k + b$  for all  $k \geq 1$ . Then  $\{a_k\}$  is a bounded sequence.*

**Lemma 6.** [34, Lemma 3, Ch. 2] *Suppose that  $a_k \geq 0, b_k \geq 0$  for all  $k \geq 1$ ,  $b_k \rightarrow 0$ , and there is some  $0 \leq \tau < 1$  such that  $a_{k+1} \leq \tau a_k + b_k$  for all  $k \geq 1$ . Then  $a_k \rightarrow 0$ .*

**Lemma 7.** *Suppose that  $\{r_k\}, \{b_k\}$ , and  $\{\tau_k\}$  are sequences in  $\mathbb{R}$  with  $b_k \rightarrow 0$ , and that for all  $k \geq 1$ , we have  $0 \leq \tau_k \leq \bar{\tau} < 1$  and  $r_{k+1} \geq \tau_k r_k + b_k$ . Then  $\liminf_{k \rightarrow \infty} \{r_k\} \geq 0$ .*

*Proof.* Negating the claimed inequality yields  $-r_{k+1} \leq \tau_k(-r_k) - b_k$ . Applying [34, Lemma 3, Ch. 2] then yields  $\limsup \{-r_k\} \leq 0$ .  $\square$

**Lemma 8.** *Suppose  $B$  is  $L^{-1}$ -cocoercive and  $0 \leq \rho \leq 2/L$ . Then for all  $x, y \in \text{dom}(B)$*

$$\|x - y - \rho(Bx - By)\| \leq \|x - y\|. \quad (22)$$

*Proof.* Squaring the left hand side of (22) yields

$$\begin{aligned} \|x - y - \rho(Bx - By)\|^2 &= \|x - y\|^2 - 2\rho\langle x - y, Bx - By \rangle + \rho^2 \|Bx - By\|^2 \\ &\leq \|x - y\|^2 - \frac{2\rho}{L} \|Bx - By\|^2 + \rho^2 \|Bx - By\|^2 \\ &\leq \|x - y\|^2. \end{aligned}$$

□

## 5.2 A Contractive Result

We begin the main proof with a result on the one-forward-step mapping:  $\mathcal{F}$  from Definition 1. The following lemma will ultimately be used to show that the iterates remain bounded.

**Lemma 9.** *Suppose  $(x^+, y^+) = \mathcal{F}_{\alpha, \rho}(z, x, w; A, B, G)$ , where  $\mathcal{F}_{\alpha, \rho}$  is given in Definition 1. Recall that  $B$  is  $L^{-1}$ -cocoercive. If  $L = 0$  or  $\rho \leq 2(1 - \alpha)/L$ , then*

$$\|x^+ - \hat{\theta}\| \leq (1 - \alpha)\|x - \hat{\theta}\| + \alpha\|Gz - \hat{\theta}\| + \rho\|w - \hat{w}\| \quad (23)$$

for any  $\hat{\theta} \in \text{dom}(A)$  and  $\hat{w} \in A\hat{\theta} + B\hat{\theta}$ .

*Proof.* Select any  $\hat{\theta} \in \text{dom}(A)$  and  $\hat{w} \in A\hat{\theta} + B\hat{\theta}$ . Let  $\hat{a} = \hat{w} - B\hat{\theta} \in A\hat{\theta}$ . It follows immediately from (4) that

$$\hat{\theta} = J_{\rho A}(\hat{\theta} + \rho\hat{a}). \quad (24)$$

Therefore, (16) and (24) yield

$$\begin{aligned} \|x^+ - \hat{\theta}\| &= \left\| J_{\rho A}((1 - \alpha)x + \alpha Gz - \rho(Bx - w)) - J_{\rho A}(\hat{\theta} + \rho\hat{a}) \right\| \\ &\stackrel{(a)}{\leq} \left\| (1 - \alpha)x + \alpha Gz - \rho(Bx - w) - \hat{\theta} - \rho\hat{a} \right\| \\ &\stackrel{(b)}{=} \left\| (1 - \alpha) \left( x - \hat{\theta} - \frac{\rho}{1 - \alpha} (Bx - B\hat{\theta}) \right) + \alpha(Gz - \hat{\theta}) + \rho(w - \hat{a} - B\hat{\theta}) \right\| \\ &\stackrel{(c)}{\leq} (1 - \alpha) \left\| x - \hat{\theta} - \frac{\rho}{1 - \alpha} (Bx - B\hat{\theta}) \right\| + \alpha\|Gz - \hat{\theta}\| + \rho\|w - (\hat{a} + B\hat{\theta})\| \quad (25) \\ &\stackrel{(d)}{\leq} (1 - \alpha)\|x - \hat{\theta}\| + \alpha\|Gz - \hat{\theta}\| + \rho\|w - \hat{w}\|. \end{aligned}$$

To obtain (a), one uses the nonexpansivity of the resolvent [4, Prop. 23.8(ii)]. To obtain (b), one regroups terms and adds and subtracts  $B\hat{\theta}$ . Then (c) follows from the triangle inequality. Finally we consider (d): If  $L > 0$ , apply Lemma 8 to the first term on the right-hand side of (25) with the stepsize  $\rho/(1 - \alpha)$  which by assumption satisfies

$$\frac{\rho}{1 - \alpha} \leq \frac{2}{L}$$

by Assumption 4. Alternatively, if  $L = 0$ , implying that  $B$  is a constant-valued operator, then  $Bx = B\hat{\theta}$  and (d) is just an equality. □

We now prove the key ‘‘ascent lemma’’. It shows that, while the one-forward-step update is not guaranteed to find a separating hyperplane at each iteration, it does make a certain kind of progress toward separation.



**Lemma 10.** Suppose  $(x^+, y^+) = \mathcal{F}_{\alpha, \rho}(z, x, w; A, B, G)$ , where  $\mathcal{F}_{\alpha, \rho}$  is given in Definition 1. Recall  $B$  is  $L^{-1}$ -cocoercive. Let  $y \in Ax + Bx$  and define  $\varphi \triangleq \langle Gz - x, y - w \rangle$ . Further, define  $\varphi^+ \triangleq \langle Gz - x^+, y^+ - w \rangle$ ,  $t$  as in (16), and  $\hat{y} \triangleq \rho^{-1}(t - x^+) + Bx$ . If  $\rho \leq 2(1 - \alpha)/L$  whenever  $L > 0$ , then

$$\varphi^+ \geq \frac{\rho}{2\alpha} (\|y^+ - w\|^2 + \alpha\|\hat{y} - w\|^2) + (1 - \alpha) \left( \varphi - \frac{\rho}{2\alpha} \|y - w\|^2 \right). \quad (26)$$

*Proof.* Since  $y \in Ax + Bx$ , there exists  $a \in Ax$  such that  $y = a + Bx$ . Let  $a^+ \triangleq \rho^{-1}(t - x^+)$ . Note by (4) that  $a^+ \in Ax^+$ . With this notation,  $\hat{y} = a^+ + Bx$ .

We may write the  $x^+$ -update in (16) as

$$x^+ + \rho a^+ = (1 - \alpha)x + \alpha Gz - \rho(Bx - w)$$

which rearranges to

$$x^+ = (1 - \alpha)x + \alpha Gz - \rho(\hat{y} - w) \implies -x^+ = -\alpha Gz - (1 - \alpha)x + \rho(\hat{y} - w).$$

Adding  $Gz$  to both sides yields

$$Gz - x^+ = (1 - \alpha)(Gz - x) + \rho(\hat{y} - w). \quad (27)$$

Substituting this equation into the definition of  $\varphi^+$  yields

$$\begin{aligned} \varphi^+ &= \langle Gz - x^+, y^+ - w \rangle \\ &= \langle (1 - \alpha)(Gz - x) + \rho(\hat{y} - w), y^+ - w \rangle \\ &= (1 - \alpha)\langle Gz - x, y^+ - w \rangle + \rho\langle \hat{y} - w, y^+ - w \rangle \\ &= (1 - \alpha)\langle Gz - x, y - w \rangle + (1 - \alpha)\langle Gz - x, y^+ - y \rangle + \rho\langle \hat{y} - w, y^+ - w \rangle \\ &= (1 - \alpha)\varphi + (1 - \alpha)\langle Gz - x, y^+ - y \rangle + \rho\langle \hat{y} - w, y^+ - w \rangle. \end{aligned} \quad (28)$$

We now focus on the second term in (28). Assume for now that  $L > 0$  (we will deal with the  $L = 0$  case below). We write

$$\begin{aligned} \langle Gz - x, y^+ - y \rangle &= \langle x^+ - x, y^+ - y \rangle + \langle Gz - x^+, y^+ - y \rangle \\ &= \langle x^+ - x, a^+ - a \rangle + \langle x^+ - x, Bx^+ - Bx \rangle + \langle Gz - x^+, y^+ - y \rangle \\ &\geq L^{-1}\|Bx^+ - Bx\|^2 + \langle Gz - x^+, y^+ - y \rangle \\ &= L^{-1}\|Bx^+ - Bx\|^2 + \langle Gz - x^+, y^+ - w \rangle + \langle Gz - x^+, w - y \rangle \\ &= L^{-1}\|Bx^+ - Bx\|^2 + \varphi^+ + \langle Gz - x^+, w - y \rangle, \end{aligned} \quad (29)$$

$$\quad (30)$$

where the inequality above uses the monotonicity of  $A$  and  $L^{-1}$ -cocoercivity of  $B$  (recall that  $a \in Ax$  and  $a^+ \in Ax^+$ ). Substituting the resulting inequality back into (28) and moving  $(1 - \alpha)\varphi^+$  to the other side yields

$$\alpha\varphi^+ \geq (1 - \alpha) \left( \varphi + L^{-1}\|Bx^+ - Bx\|^2 + \langle Gz - x^+, w - y \rangle \right) + \rho\langle \hat{y} - w, y^+ - w \rangle. \quad (31)$$

Using (27) once again, this time to the third term on the right-hand side, we write

$$\begin{aligned}
\langle Gz - x^+, w - y \rangle &= \langle (1 - \alpha)(Gz - x) + \rho(\hat{y} - w), w - y \rangle \\
&= (1 - \alpha)\langle Gz - x, w - y \rangle + \rho\langle \hat{y} - w, w - y \rangle \\
&= (\alpha - 1)\varphi - \rho\langle \hat{y} - w, y - w \rangle.
\end{aligned} \tag{32}$$

Substituting this equation back into (31) yields

$$\alpha\varphi^+ \geq (1 - \alpha) \left( \alpha\varphi + L^{-1}\|Bx^+ - Bx\|^2 - \rho\langle \hat{y} - w, y - w \rangle \right) + \rho\langle \hat{y} - w, y^+ - w \rangle. \tag{33}$$

We next use the identity  $\langle x_1, x_2 \rangle = \frac{1}{2}\|x_1\|^2 + \frac{1}{2}\|x_2\|^2 - \frac{1}{2}\|x_1 - x_2\|^2$  on both inner products in the above expression, as follows:

$$\begin{aligned}
\langle \hat{y} - w, y - w \rangle &= \frac{1}{2} (\|\hat{y} - w\|^2 + \|y - w\|^2 - \|\hat{y} - y\|^2) \\
&= \frac{1}{2} (\|\hat{y} - w\|^2 + \|y - w\|^2 - \|a^+ - a\|^2)
\end{aligned} \tag{34}$$

and

$$\begin{aligned}
\langle \hat{y} - w, y^+ - w \rangle &= \frac{1}{2} (\|\hat{y} - w\|^2 + \|y^+ - w\|^2 - \|\hat{y} - y^+\|^2) \\
&= \frac{1}{2} (\|\hat{y} - w\|^2 + \|y^+ - w\|^2 - \|Bx^+ - Bx\|^2).
\end{aligned} \tag{35}$$

Here we have used the identities

$$\begin{aligned}
\hat{y} - y &= a^+ + Bx - (a + Bx) = a^+ - a \\
\hat{y} - y^+ &= a^+ + Bx - (a^+ + Bx^+) = Bx - Bx^+.
\end{aligned}$$

Using (34)–(35) in (33) and simplifying yields

$$\begin{aligned}
\alpha\varphi^+ &\geq (1 - \alpha) \left( \alpha\varphi - \frac{\rho}{2}\|y - w\|^2 + \frac{\rho}{2}\|a^+ - a\|^2 \right) + \left( \frac{1 - \alpha}{L} - \frac{\rho}{2} \right) \|Bx^+ - Bx\|^2 \\
&\quad + \frac{\rho}{2} (\|y^+ - w\|^2 + \alpha\|\hat{y} - w\|^2).
\end{aligned}$$

Now, the coefficient of  $\|a^+ - a\|^2$  is nonnegative so this term can be lower-bounded by 0. Furthermore, since  $\rho \leq 2(1 - \alpha)/L$ , the coefficient of  $\|Bx^+ - Bx\|^2$  is positive as well, and this term can also be lower-bounded by 0. Using these bounds and dividing through by  $\alpha$ , we obtain (26).

Finally, we deal with the case in which  $L = 0$ , which implies that  $Bx = v$  for some  $v \in \mathcal{H}$  for all  $x \in \mathcal{H}$ . The main difference is that the  $\|Bx^+ - Bx\|^2$  terms are no longer present since  $Bx^+ = Bx$ . The analysis is the same up to (28). In this case  $Bx^+ = v$  so instead of (30) we may deduce from (29) that

$$\langle Gz - x, y^+ - y \rangle \geq \varphi^+ + \langle Gz - x^+, w - y \rangle.$$

Since  $Bx^+ = Bx = v$  is constant we also have that

$$\hat{y} = a^+ + Bx = a^+ + v = a^+ + Bx^+ = y^+$$

Thus, instead of (31) in this case we have the simpler inequality

$$\alpha\varphi^+ \geq (1 - \alpha) (\varphi + \langle Gz - x^+, w - y \rangle) + \rho\|y^+ - w\|^2. \quad (36)$$

The term  $\langle Gz - x^+, w - y \rangle$  in (36) is dealt with just as in (31), by substitution of (27). This step now leads via (32) to

$$\alpha\varphi^+ \geq \alpha(1 - \alpha)\varphi - \rho(1 - \alpha)\langle y^+ - w, y - w \rangle + \rho\|y^+ - w\|^2.$$

Once again using  $\langle x_1, x_2 \rangle = \frac{1}{2}\|x_1\|^2 + \frac{1}{2}\|x_2\|^2 - \frac{1}{2}\|x_1 - x_2\|^2$  on the second term on the r.h.s. above yields

$$\alpha\varphi^+ \geq \alpha(1 - \alpha)\varphi + \rho\|y^+ - w\|^2 - \frac{\rho(1 - \alpha)}{2} (\|y^+ - w\|^2 + \|y - w\|^2 - \|y^+ - y\|^2).$$

We can lower-bound the  $\|y^+ - y\|^2$  term by 0. Dividing through by  $\alpha$  and rearranging, we obtain

$$\varphi^+ \geq \frac{\rho(1 + \alpha)}{2\alpha}\|y^+ - w\|^2 + (1 - \alpha) \left( \varphi - \frac{\rho}{2\alpha}\|y - w\|^2 \right).$$

Since  $y^+ = \hat{y}$  in the  $L = 0$  case, this is equivalent to (26).  $\square$

### 5.3 Finite Termination of Backtracking

In all the following lemmas in sections 5.3 and 5.4 regarding algorithms 1–3, assumptions 1–5 are in effect and will not be explicitly stated in each lemma. We start by proving that **backTrack** terminates in a finite number of iterations, and that the stepsizes it returns are bounded away from 0.

**Lemma 11.** *For  $i \in \mathcal{B}$ , Algorithm 2 terminates in a finite number of iterations for all  $k \geq 1$ . There exists  $\underline{\rho}_i > 0$  such that  $\rho_i^k \geq \underline{\rho}_i$  for all  $k \geq 1$ , where  $\rho_i^k$  is the stepsize returned by Algorithm 2 on line 4.*

*Proof.* Assume we are at iteration  $k \geq 1$  in Algorithm 1 and **backTrack** has been called through line 4 for some  $i \in \mathcal{B}$ . The internal variables within **backTrack** are defined in terms of the variables passed from Algorithm 1 as follows:  $z = z^k$ ,  $w = w_i^k$ ,  $x = x_i^{k-1}$ ,  $y = y_i^{k-1}$ ,  $\tilde{\rho}_1 = \hat{\rho}_i^k$ ,  $\varphi = \varphi_{i,k-1}(z^k, w_i^k)$ ,  $\alpha = \alpha_i^k$ ,  $\hat{\theta} = \hat{\theta}_i$ ,  $\hat{w} = \hat{w}_i$ ,  $A = A_i$ ,  $B = B_i$ , and  $G = G_i$ . In the following argument, we mostly refer to the internal name of the variables within **backTrack** without explicitly making the above substitutions. With that in mind, let  $L = L_i$  be the cocoercivity constant of  $B = B_i$ .

Observe that backtracking terminates via line 9 if two conditions are met. The first condition,

$$\|\tilde{x}_j - \hat{\theta}\| \leq (1 - \alpha)\|x - \hat{\theta}\| + \alpha\|Gz - \hat{\theta}\| + \tilde{\rho}_j\|w - \hat{w}\|, \quad (37)$$

is identical to (23) of Lemma 9, with  $\tilde{x}_j$  and  $\tilde{\rho}_j$  respectively in place of  $x^+$  and  $\rho$ . The initialization step of Algorithm 2 provides us with  $\hat{w} \in A\hat{\theta} + B\hat{\theta}$  for some  $\hat{\theta} \in \text{dom}(A)$ . Furthermore, since

$$(\tilde{x}_j, \tilde{y}_j) = \mathcal{F}_{\alpha, \tilde{\rho}_j}(z, x, w; A, B, G),$$

the findings of Lemma 9 may be applied. In particular, if  $L > 0$  and  $\tilde{\rho}_j \leq 2(1 - \alpha)/L$ , then (37) will be met. Alternatively, if  $L = 0$ , (37) will hold for any value of the stepsize  $\tilde{\rho}_j > 0$ .

Next, consider the second termination condition,

$$\varphi_j^+ \geq \frac{\tilde{\rho}_j}{2\alpha} (\|\tilde{y}_j - w\|^2 + \alpha\|\hat{y}_j - w\|^2) + (1 - \alpha) \left( \varphi - \frac{\tilde{\rho}_j}{2\alpha} \|y - w\|^2 \right). \quad (38)$$

This relation is identical to (26) of Lemma 10, with  $(\tilde{y}_j, \hat{y}_j, \tilde{\rho}_j)$  in place of  $(y^+, \hat{y}, \rho)$ . However, to apply the lemma we must show that  $y = y_i^{k-1} \in Ax_i^{k-1} + Bx_i^{k-1} = Ax + Bx$ . We will do so by induction.

For  $k = 1$ ,  $y = y_i^{k-1} \in Ax_i^{k-1} + Bx_i^{k-1} = Ax + Bx$  holds by the initialization step of Algorithm 1. Now assume that it holds at iteration  $k \geq 2$ . We may then apply the findings of Lemma 10 to conclude that if  $L > 0$  and  $\tilde{\rho}_j \leq 2(1 - \alpha)/L$ , then condition (38) is satisfied. Or, if  $L = 0$ , condition (38) is satisfied for any  $\tilde{\rho}_j > 0$ .

Combining the above observations, we conclude that if  $L > 0$  and  $\tilde{\rho}_j \leq 2(1 - \alpha)/L$ , backtracking will terminate in that iteration via line 9. Or, if  $L = 0$ , it will terminate in the first iteration. The stepsize decrement condition on line 10 of the backtracking procedure implies that  $\tilde{\rho}_j \leq 2(1 - \alpha)/L$  will eventually hold for large enough  $j$ , and hence that the two backtracking termination conditions must eventually hold.

Let  $j^* \geq 1$  be the iteration at which backtracking terminates when called for operator  $i$  at iteration  $k$  of Algorithm 1. For the pair  $(x_i^k, y_i^k)$  returned by `backTrack` on line 4 of Algorithm 1, we may write

$$(x_i^k, y_i^k) = (\tilde{x}_{j^*}, \tilde{y}_{j^*}) = \mathcal{F}_{\alpha, \tilde{\rho}_{j^*}}(z, x, w; A, B, G) = \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i).$$

Thus, by the definition of  $\mathcal{F}$  in (16),  $y_i^k \in A_i x_i^k + B_i x_i^k$ . Therefore, induction establishes that  $y_i^k \in A_i x_i^k + B_i x_i^k$  holds for all  $k \geq 1$ . Inductively, we now also know that `backTrack` terminates in a finite number of iterations for all  $k \geq 1$  and  $i \in \mathcal{B}$ .

For all  $k \geq 1$  and  $i \in \mathcal{B}$ , the returned stepsize  $\rho_i^k = \tilde{\rho}_{j^*}$  must satisfy

$$\begin{aligned} (\forall i : L_i > 0) : \quad \rho_i^k &\geq \min \left\{ \hat{\rho}_i^k, \frac{2\delta(1 - \alpha_i^k)}{L_i} \right\} \\ (\forall i : L_i = 0) : \quad \rho_i^k &= \hat{\rho}_i^k. \end{aligned} \quad (39)$$

If (18) is enforced then we immediately conclude that for  $i$  with  $L_i > 0$

$$\rho_i^k \geq \min \left\{ \hat{\rho}_i^k, \frac{2\delta(1 - \alpha_i^k)}{L_i} \right\} \geq \min \left\{ \hat{\rho}_i, \frac{2\delta(1 - \bar{\alpha}_i)}{L_i} \right\} \triangleq \underline{\rho}_i > 0,$$

and similarly for  $L_i = 0$ , we have  $\rho_i^k \geq \hat{\rho}_i \triangleq \underline{\rho}_i > 0$ . On the other hand, if (19) is enforced then we may argue that for all  $k \geq 1$  and all  $i \in \mathcal{B}$  such that  $L_i > 0$ , one has

$$\begin{aligned} \rho_i^k &\geq \min \left\{ \hat{\rho}_i^k, \frac{2\delta(1 - \bar{\alpha}_i)}{L_i} \right\} \geq \min \left\{ \rho_i^{k-1}, \frac{2\delta(1 - \bar{\alpha}_i)}{L_i} \right\} \\ &\geq \min \left\{ \rho_i^1, \frac{2\delta(1 - \bar{\alpha}_i)}{L_i} \right\} \geq \min \left\{ \hat{\rho}_i^1, \frac{2\delta(1 - \bar{\alpha}_i)}{L_i} \right\} \triangleq \underline{\rho}_i > 0, \end{aligned}$$

where the second inequality uses (19), the third inequality recurses, and the final inequality is just (39) for  $k = 1$ . If  $L_i = 0$ , the argument is simply

$$\rho_i^k = \hat{\rho}_i^k \geq \rho_i^{k-1} = \hat{\rho}_i^{k-1} \geq \dots \geq \hat{\rho}_i^1 \triangleq \underline{\rho}_i > 0. \quad \square$$

## 5.4 Boundedness Results and their Direct Consequences

**Lemma 12.** For  $i = 1, \dots, n$

$$\|x_i^k - \hat{\theta}_i\| \leq (1 - \alpha_i^k) \|x_i^{k-1} - \hat{\theta}_i\| + \alpha_i^k \|G_i z^k - \hat{\theta}_i\| + \rho_i^k \|w_i^k - \hat{w}_i\| \quad (40)$$

for some  $\hat{\theta}_i \in \text{dom}(A_i)$  and  $\hat{w}_i \in A\hat{\theta}_i + B\hat{\theta}_i$ .

*Proof.* For  $i \in \mathcal{B}$ , Lemma 11 establishes that `backTrack` terminates for finite  $j \geq 1$  for all  $k \geq 1$ . For fixed  $k \geq 1$  and  $i \in \mathcal{B}$ , let  $j^* \geq 1$  be the iteration of `backTrack` that terminates. At termination, the following condition is satisfied via line 7:

$$\|\tilde{x}_{j^*} - \hat{\theta}\| \leq (1 - \alpha) \|x - \hat{\theta}\| + \alpha \|Gz - \hat{\theta}\| + \tilde{\rho}_{j^*} \|w - \hat{w}\|.$$

Into this inequality, now substitute in the following variables from Algorithm 1, as passed to and from `backTrack`:  $x_i^k = \tilde{x}_{j^*}$ ,  $\hat{\theta}_i = \hat{\theta}$ ,  $\alpha_i^k = \alpha$ ,  $x_i^{k-1} = x$ ,  $G_i = G$ ,  $z^k = z$ ,  $\rho_i^k = \tilde{\rho}_{j^*}$ ,  $w_i^k = w$ , and  $\hat{w}_i = \hat{w}$ . The result is (40).

For  $i \notin \mathcal{B}$ , we note that line 6 of Algorithm 1 reads as

$$(x_i^k, y_i^k) = \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i)$$

and Assumption 4(3) holds, so we may apply Lemma 9 to yield (40).  $\square$

**Lemma 13.** For all  $i = 1, \dots, n$ , the sequences  $\{x_i^k\}$  and  $\{y_i^k\}$  are bounded.

*Proof.* Fix any  $i \in \{1, \dots, n\}$ . From (40) and the bounds on  $\{\alpha_i^k\}$  and  $\{\rho_i^k\}$ , we have

$$(\forall k \geq 1) \quad \|x_i^k - \hat{\theta}_i\| \leq (1 - \underline{\alpha}_i) \|x_i^{k-1} - \hat{\theta}_i\| + \bar{\alpha}_i \|G_i z^k - \hat{\theta}_i\| + \bar{\rho}_i \|w_i^k - \hat{w}_i\|.$$

Since  $\{z^k\}$ , and  $\{w_i^k\}$  are bounded by Lemma 3 and  $\|G_i\|$  is bounded by Assumption 1, boundedness of  $\{x_i^k\}$  now follows by applying Lemma 5 with  $\tau = 1 - \underline{\alpha}_i < 1$ .

Next, boundedness of  $B_i x_i^k$  follows from the continuity of  $B_i$ . Since Lemma 11 established that `backTrack` terminates in a finite number of iterations we have for any  $k \geq 2$  that

$$(x_i^k, y_i^k) = \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i).$$

Expanding the  $y^+$ -update in the definition of  $\mathcal{F}$  in (16), we may write

$$y_i^k = (\rho_i^k)^{-1} \left( (1 - \alpha_i^k) x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (B_i x_i^{k-1} - w_i^k) - x_i^k \right) + B x_i^k.$$

Since  $0 \leq \alpha_i^k \leq 1$ ,  $G_i$ ,  $z^k$ , and  $w_i^k$  are bounded,  $\rho_i^k \leq \bar{\rho}_i$ , and  $\rho_i^k \geq \underline{\rho}_i$  (using Lemma 11 for  $i \in \mathcal{B}$ ), we conclude that  $y_i^k$  remains bounded.  $\square$

With  $\{x_i^k\}$  and  $\{y_i^k\}$  bounded for all  $i = 1, \dots, n$ , the boundedness of  $\nabla\varphi_k$  follows immediately:

**Lemma 14.** *The sequence  $\{\nabla\varphi_k\}$  is bounded. If Algorithm 1 never terminates via line 9,  $\limsup_{k \rightarrow \infty} \varphi_k(p^k) \leq 0$ .*

*Proof.* By Lemma 2,  $\nabla_z \varphi_k = \sum_{i=1}^n G_i^* y_i^k$ , which is bounded since each  $G_i$  is bounded by assumption and each  $\{y_i^k\}$  is bounded by Lemma 13. Furthermore,  $\nabla_{w_i} \varphi_k = x_i^k - G_i x_n^k$  is bounded using the same two lemmas. That  $\limsup_{k \rightarrow \infty} \varphi_k(p^k) \leq 0$  then immediately follows from Lemma 3(3).  $\square$

Using the boundedness of  $\{x_i^k\}$  and  $\{y_i^k\}$ , we can next derive the following simple bound relating  $\varphi_{i,k-1}(z^k, w_i^k)$  to  $\varphi_{i,k-1}(z^{k-1}, w_i^{k-1})$ :

**Lemma 15.** *There exists  $M_1, M_2 \geq 0$  such that for all  $k \geq 2$  and  $i = 1, \dots, n$ ,*

$$\varphi_{i,k-1}(z^k, w_i^k) \geq \varphi_{i,k-1}(z^{k-1}, w_i^{k-1}) - M_1 \|w_i^k - w_i^{k-1}\| - M_2 \|G_i\| \|z^k - z^{k-1}\|.$$

*Proof.* For each  $i \in \{1, \dots, n\}$ , let  $M_{1,i}, M_{2,i} \geq 0$  be respective bounds on  $\{\|G_i z^{k-1} - x_i^{k-1}\|\}$  and  $\{\|y_i^{k-1} - w_i^k\|\}$ , which must exist by Lemma 3, the boundedness of  $\{x_i^k\}$  and  $\{y_i^k\}$ , and the boundedness of  $G_i$ . Let  $M_1 = \max_{i=1, \dots, m} \{M_{1,i}\}$  and  $M_2 = \max_{i=1, \dots, m} \{M_{2,i}\}$ . Then, for any  $k \geq 2$  and  $i \in \{1, \dots, n\}$ , we may write

$$\begin{aligned} \varphi_{i,k-1}(z^k, w_i^k) &= \langle G_i z^k - x_i^{k-1}, y_i^{k-1} - w_i^k \rangle \\ &= \langle G_i z^{k-1} - x_i^{k-1}, y_i^{k-1} - w_i^k \rangle + \langle G_i z^k - G_i z^{k-1}, y_i^{k-1} - w_i^k \rangle \\ &= \langle G_i z^{k-1} - x_i^{k-1}, y_i^{k-1} - w_i^{k-1} \rangle + \langle G_i z^{k-1} - x_i^{k-1}, w_i^{k-1} - w_i^k \rangle \\ &\quad + \langle G_i z^k - G_i z^{k-1}, y_i^{k-1} - w_i^k \rangle \\ &\geq \varphi_{i,k-1}(z^{k-1}, w_i^{k-1}) - M_1 \|w_i^k - w_i^{k-1}\| - M_2 \|G_i\| \|z^k - z^{k-1}\|, \end{aligned}$$

where the last step uses the Cauchy-Schwarz inequality and the definitions of  $M_1$  and  $M_2$ .  $\square$

## 5.5 A Lyapunov-Like Recursion for the Hyperplane

We now establish a Lyapunov-like recursion for the hyperplane. For this purpose, we need one more definition and assumption.

**Definition 2.** *For all  $k \geq 1$ , since Lemma 11 establishes that Algorithm 2 terminates in a finite number of iterations, we may write for  $i = 1, \dots, n$ :*

$$(x_i^k, y_i^k) = \mathcal{F}_{\alpha_i^k, \rho_i^k}(z^k, x_i^{k-1}, w_i^k; A_i, B_i, G_i).$$

Using (4) and the  $x^+$ -update in (16), there exists  $a_i^k \in A_i x_i^k$  such that

$$x_i^k + \rho_i^k a_i^k = (1 - \alpha_i^k) x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (B_i x_i^{k-1} - w_i^k).$$

Define  $\hat{y}_i^k \triangleq a_i^k + B_i x_i^{k-1}$ .

**Assumption 6.** For all  $k \geq 1$ , and  $i \notin \mathcal{B}$ ,  $\alpha_i^k$  and  $\rho_i^k$  are chosen to satisfy

$$\frac{\rho_i^{k+1}}{\alpha_i^{k+1}} \|y_i^k - w_i^k\|^2 \leq \frac{\rho_i^k}{\alpha_i^k} (\|y_i^k - w_i^k\|^2 + \alpha_i^k \|\hat{y}_i^k - w_i^k\|^2). \quad (41)$$

On the other hand, for  $i \in \mathcal{B}$ , the initial trial stepsizes  $\hat{\rho}_i^k$  and  $\alpha_i^k$  are chosen to satisfy:

$$\frac{\hat{\rho}_i^{k+1}}{\alpha_i^{k+1}} \|y_i^k - w_i^k\|^2 \leq \frac{\rho_i^k}{\alpha_i^k} (\|y_i^k - w_i^k\|^2 + \alpha_i^k \|\hat{y}_i^k - w_i^k\|^2) \quad (42)$$

where  $\rho_i^k$  is returned on line 4.

If  $\alpha_i^k = \alpha_i$  is fixed, (42) would then be guaranteed by choosing  $\hat{\rho}_i^{k+1} \leq \rho_i^k$ . However for fixed  $\alpha_i^k$ , (42) allows the trial stepsize to increase by some factor times the previously discovered stepsize. In practice we have observed that  $\|y_i^k - w_i^k\|$  and  $\|\hat{y}_i^k - w_i^k\|$  tend to be approximately equal, so (42) allows for an increase in the trial stepsize by up to a factor of approximately  $1 + \alpha_i$ .

In addition to assumptions 1–5, Assumption 6 is now in effect and will not be explicitly stated in the upcoming lemmas of sections 5.5–5.6.

**Lemma 16.** For all  $k \geq 2$  and  $i = 1, \dots, n$ ,

$$\begin{aligned} \varphi_{i,k}(z^k, w_i^k) - \frac{\rho_i^k}{2\alpha_i^k} (\|y_i^k - w_i^k\|^2 + \alpha_i^k \|\hat{y}_i^k - w_i^k\|^2) \\ \geq (1 - \alpha_i^k) \left( \varphi_{i,k-1}(z^k, w_i^k) - \frac{\rho_i^k}{2\alpha_i^k} \|y_i^{k-1} - w_i^k\|^2 \right) \end{aligned} \quad (43)$$

and

$$\varphi_{i,k}(z^k, w_i^k) - \frac{\rho_i^{k+1}}{2\alpha_i^{k+1}} \|y_i^k - w_i^k\|^2 \geq (1 - \alpha_i^k) \left( \varphi_{i,k-1}(z^k, w_i^k) - \frac{\rho_i^k}{2\alpha_i^k} \|y_i^{k-1} - w_i^k\|^2 \right). \quad (44)$$

*Proof.* Take any  $i \in \mathcal{B}$ . Lemma 11 guarantees the finite termination of `backTrack`. Now consider the backtracking termination condition

$$\varphi_j^+ \geq \frac{\tilde{\rho}_j}{2\alpha} (\|\tilde{y}_j - w\|^2 + \alpha \|\hat{y}_j - w\|^2) + (1 - \alpha) \left( \varphi - \frac{\tilde{\rho}_j}{2\alpha} \|y - w\|^2 \right).$$

Fix some  $k \geq 2$ , and let  $j^* \geq 1$  be the iteration at which `backTrack` terminates. In the above inequality, make the following substitutions for the internal variables of `backTrack` by those passed in/out of the function:  $\varphi_{i,k}(z^k, x_i^k) = \varphi_{j^*}^+$ ,  $\rho_i^k = \tilde{\rho}_{j^*}$ ,  $\alpha_i^k = \alpha$ ,  $y_i^k = \tilde{y}_j$ ,  $w_i^k = w$ ,  $\varphi_{i,k-1}(z^k, w_i^k) = \varphi$ . Furthermore,  $\hat{y}_i^k = \hat{y}_{j^*}^k$  where  $\hat{y}_{j^*}^k$  is defined in Definition 2. Together, these substitutions yield (43). We can then apply (42), along with  $\rho_i^{k+1} \leq \hat{\rho}_i^{k+1}$ , to produce (44).

Now take any  $i \in \{1, \dots, n\} \setminus \mathcal{B}$ . From line 6 of Algorithm 1, Assumption 4(3), and Lemma 10, we directly deduce (43). Combining this relation with (41) we obtain (44).  $\square$

## 5.6 Finishing the Proof

We now work toward establishing the conditions of Lemma 4, which we then apply. Unless otherwise specified, we henceforth assume that Algorithm 1 runs indefinitely and does not terminate at line 9.

**Lemma 17.** *For all  $i = 1, \dots, n$ , we have  $y_i^k - w_i^k \rightarrow 0$  and  $\varphi_k(p^k) \rightarrow 0$ .*

*Proof.* Fix any  $i \in \{1, \dots, n\}$ . First, note that for all  $k \geq 2$ ,

$$\begin{aligned} \|y_i^{k-1} - w_i^k\|^2 &= \|y_i^{k-1} - w_i^{k-1}\|^2 + 2\langle y_i^{k-1} - w_i^{k-1}, w_i^{k-1} - w_i^k \rangle + \|w_i^{k-1} - w_i^k\|^2 \\ &\leq \|y_i^{k-1} - w_i^{k-1}\|^2 + M_3 \|w_i^k - w_i^{k-1}\| + \|w_i^k - w_i^{k-1}\|^2 \\ &= \|y_i^{k-1} - w_i^{k-1}\|^2 + d_i^k, \end{aligned} \quad (45)$$

where  $d_i^k \triangleq M_3 \|w_i^k - w_i^{k-1}\| + \|w_i^k - w_i^{k-1}\|^2$  and  $M_3 \geq 0$  is a bound on  $2\|y_i^{k-1} - w_i^{k-1}\|$ , which must exist because both  $\{y_i^k\}$  and  $\{w_i^k\}$  are bounded by lemmas 3 and 13. Note that  $d_i^k \rightarrow 0$  as a consequence of Lemma 3.

Second, recall Lemma 15, which states that there exists  $M_1, M_2 \geq 0$  such that for all  $k \geq 2$ ,

$$\varphi_{i,k-1}(z^k, w_i^k) \geq \varphi_{i,k-1}(z^{k-1}, w_i^{k-1}) - M_1 \|w_i^{k-1} - w_i^k\| - M_2 \|G_i\| \|z^k - z^{k-1}\|. \quad (46)$$

Now let, for all  $k \geq 1$ ,

$$r_i^k \triangleq \varphi_{i,k}(z^k, w_i^k) - \frac{\rho_i^{k+1}}{2\alpha_i^{k+1}} \|y_i^k - w_i^k\|^2, \quad (47)$$

so that

$$\sum_{i=1}^n r_i^k = \varphi_k(p^k) - \sum_{i=1}^n \frac{\rho_i^{k+1}}{2\alpha_i^{k+1}} \|y_i^k - w_i^k\|^2. \quad (48)$$

Using (45) and (46) in (44) yields

$$(\forall k \geq 2) : \quad r_i^k \geq (1 - \alpha_i^k) r_i^{k-1} + e_i^k \quad (49)$$

where

$$e_i^k \triangleq -(1 - \alpha_i^k) \left( \frac{\rho_i^k}{2\alpha_i^k} d_i^k + M_1 \|w_i^{k-1} - w_i^k\| + M_2 \|G_i\| \|z^k - z^{k-1}\| \right). \quad (50)$$

Note that  $\rho_i^k$  is bounded,  $0 < \alpha_i^k \leq 1$ ,  $\|G_i\|$  is finite,  $\|z^k - z^{k-1}\| \rightarrow 0$  and  $\|w_i^k - w_i^{k-1}\| \rightarrow 0$  by Lemma 3, and  $d_i^k \rightarrow 0$ . Thus  $e_i^k \rightarrow 0$ .

Since  $0 < \underline{\alpha}_i \leq \alpha_i^k \leq 1$ , we may apply Lemma 7 to (49) with  $\tau_k = 1 - \alpha_i^k$  and  $\bar{\tau} = 1 - \underline{\alpha}_i < 1$  which yields  $\liminf_{k \rightarrow \infty} \{r_i^k\} \geq 0$ . Therefore

$$\liminf_{k \rightarrow \infty} \sum_{i=1}^n r_i^k \geq \sum_{i=1}^n \liminf_{k \rightarrow \infty} r_i^k \geq 0. \quad (51)$$



On the other hand,  $\limsup_{k \rightarrow \infty} \varphi_k(p^k) \leq 0$  by Lemma 14. Therefore, using (48) and (51),

$$\begin{aligned} 0 &\leq \liminf_{k \rightarrow \infty} \sum_{i=1}^n r_i^k = \liminf_{k \rightarrow \infty} \left\{ \varphi_k(p^k) - \sum_{i=1}^n \frac{\rho_i^{k+1}}{2\alpha_i^{k+1}} \|y_i^k - w_i^k\|^2 \right\} \\ &\leq \liminf_{k \rightarrow \infty} \varphi_k(p^k) \leq \limsup_{k \rightarrow \infty} \varphi_k(p^k) \leq 0. \end{aligned}$$

Therefore  $\lim_{k \rightarrow \infty} \{\varphi_k(p^k)\} = 0$ . Consider any  $i \in \{1, \dots, n\}$ . Combining  $\lim_{k \rightarrow \infty} \{\varphi_k(p^k)\} = 0$  with  $\liminf_{k \rightarrow \infty} \sum_{i=1}^n r_i^k \geq 0$ , we have

$$\limsup_{k \rightarrow \infty} \{(\rho_i^{k+1}/\alpha_i^{k+1})\|y_i^k - w_i^k\|^2\} \leq 0 \quad \Rightarrow \quad (\rho_i^{k+1}/\alpha_i^{k+1})\|y_i^k - w_i^k\|^2 \rightarrow 0.$$

Since  $\rho_i^k \geq \underline{\rho}_i > 0$  (using Lemma 11 for  $i \in \mathcal{B}$ ) and  $\alpha_i^k \leq \bar{\alpha}_i$ , we conclude that  $y_i^k - w_i^k \rightarrow 0$ .  $\square$

We have already proved the first requirement of Lemma 4, that  $y_i^k - w_i^k \rightarrow 0$  for all  $i \in \{1, \dots, n\}$ . We now work to establish the second requirement, that  $G_i z^k - x_i^k \rightarrow 0$ . In the upcoming lemmas we continue to use the quantity  $\hat{y}_i^k$  which is given in Definition 2.

**Lemma 18.** *Recall  $\{\hat{y}_i^k\}_{k \in \mathbb{N}}$  from Definition 2. For all  $i = 1, \dots, n$ ,  $\hat{y}_i^k - w_i^k \rightarrow 0$ .*

*Proof.* Fix any  $k \geq 1$ . For all  $i = 1, \dots, n$ , repeating (43) from Lemma 16, we have

$$\begin{aligned} \varphi_{i,k}(z^k, w_i^k) &\geq (1 - \alpha_i^k) \left( \varphi_{i,k-1}(z^k, w_i^k) - \frac{\rho_i^k}{2\alpha_i^k} \|y_i^{k-1} - w_i^k\|^2 \right) \\ &\quad + \frac{\rho_i^k}{2\alpha_i^k} (\|y_i^k - w_i^k\|^2 + \alpha_i^k \|\hat{y}_i^k - w_i^k\|^2) \\ &\geq (1 - \alpha_i^k) r_i^{k-1} + \frac{\rho_i^k}{2} \|\hat{y}_i^k - w_i^k\|^2 + e_i^k \end{aligned}$$

where we have used  $r_i^k$  defined (47) along with (45)–(46) and  $e_i^k$  is defined in (50). This is the same argument used in Lemma 17, but now we apply (45)–(46) to (43), rather than (44), so that we can upper bound the  $\|\hat{y}_i^k - w_i^k\|^2$  term. Summing over  $i = 1, \dots, n$ , yields

$$\varphi_k(p^k) = \sum_{i=1}^n \varphi_{i,k}(z^k, w_i^k) \geq \sum_{i=1}^n (1 - \alpha_i^k) r_i^{k-1} + \sum_{i=1}^n \frac{\rho_i^k}{2} \|\hat{y}_i^k - w_i^k\|^2 + \sum_{i=1}^n e_i^k.$$

Since  $\varphi_k(p^k) \rightarrow 0$ ,  $e_i^k \rightarrow 0$ ,  $\liminf_{k \rightarrow \infty} \{r_i^k\} \geq 0$ ,  $\rho_i^k \geq \underline{\rho}_i > 0$  for all  $k$ , and  $\alpha_i^k \leq 1$  for all  $k$ , the above inequality implies that  $\hat{y}_i^k - w_i^k \rightarrow 0$ .  $\square$

**Lemma 19.** *For  $i = 1, \dots, n$ ,  $x_i^k - x_i^{k-1} \rightarrow 0$ .*

*Proof.* Fix  $i \in \{1, \dots, n\}$ . Using the definition of  $a_i^k$  in Definition 2, we have for  $k \geq 1$  that

$$x_i^k + \rho_i^k a_i^k = (1 - \alpha_i^k) x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (B_i x_i^{k-1} - w_i^k).$$

Using the definition of  $\hat{y}_i^k$ , also in Definition 2, this implies that

$$(\forall k \geq 1) : \quad x_i^k = (1 - \alpha_i^k) x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (\hat{y}_i^k - w_i^k), \quad (52)$$

$$(\forall k \geq 2) : \quad x_i^{k-1} = (1 - \alpha_i^{k-1}) x_i^{k-2} + \alpha_i^{k-1} G_i z^{k-1} - \rho_i^{k-1} (\hat{y}_i^{k-1} - w_i^{k-1}). \quad (53)$$

Subtracting the second of these equations from the first yields, for all  $k \geq 2$ ,

$$\begin{aligned}
x_i^k - x_i^{k-1} &= (1 - \alpha_i^k)x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (\hat{y}_i^k - w_i^k) - (1 - \alpha_i^{k-1})x_i^{k-2} \\
&\quad - \alpha_i^{k-1} G_i z^{k-1} + \rho_i^{k-1} (\hat{y}_i^{k-1} - w_i^{k-1}) \\
&= (1 - \alpha_i^k)(x_i^{k-1} - x_i^{k-2}) + (\alpha_i^{k-1} - \alpha_i^k)x_i^{k-2} + \alpha_i^k (G_i z^k - G_i z^{k-1}) \\
&\quad - (\alpha_i^{k-1} - \alpha_i^k)G_i z^{k-1} + \rho_i^{k-1} (\hat{y}_i^{k-1} - w_i^{k-1}) - \rho_i^k (\hat{y}_i^k - w_i^k). \\
&= (1 - \alpha_i^k)(x_i^{k-1} - x_i^{k-2}) + \alpha_i^k G_i (z^k - z^{k-1}) \\
&\quad + (\alpha_i^{k-1} - \alpha_i^k)(x_i^{k-2} - G_i z^{k-1}) + \rho_i^{k-1} (\hat{y}_i^{k-1} - w_i^{k-1}) - \rho_i^k (\hat{y}_i^k - w_i^k). \tag{54}
\end{aligned}$$

Next, (53) can be rearranged into

$$x_i^{k-2} - G_i z^{k-1} = -\frac{1}{\alpha_i^{k-1}} (x_i^{k-1} - x_i^{k-2} + \rho_i^{k-1} (\hat{y}_i^{k-1} - w_i^{k-1})).$$

Substituting this equation into (54) yields

$$\begin{aligned}
x_i^k - x_i^{k-1} &= \left(1 - \alpha_i^k - \frac{\alpha_i^{k-1} - \alpha_i^k}{\alpha_i^{k-1}}\right) (x_i^{k-1} - x_i^{k-2}) - \rho_i^k (\hat{y}_i^k - w_i^k) \\
&\quad + \rho_i^{k-1} \left(1 - \frac{\alpha_i^{k-1} - \alpha_i^k}{\alpha_i^{k-1}}\right) (\hat{y}_i^{k-1} - w_i^{k-1}) + \alpha_i^k G_i (z^k - z^{k-1}) \\
&= \alpha_i^k \left(\frac{1}{\alpha_i^{k-1}} - 1\right) (x_i^{k-1} - x_i^{k-2}) - \rho_i^k (\hat{y}_i^k - w_i^k) + \frac{\rho_i^{k-1} \alpha_i^k}{\alpha_i^{k-1}} (\hat{y}_i^{k-1} - w_i^{k-1}) \\
&\quad + \alpha_i^k G_i (z^k - z^{k-1}).
\end{aligned}$$

Taking norms and using the triangle inequality yields, for all  $k \geq 2$ , that

$$\|x_i^k - x_i^{k-1}\| \leq \alpha_i^k \left(\frac{1}{\alpha_i^{k-1}} - 1\right) \|x_i^{k-1} - x_i^{k-2}\| + \tilde{e}_i^k \leq c_\alpha \|x_i^{k-1} - x_i^{k-2}\| + \tilde{e}_i^k, \tag{55}$$

where

$$\tilde{e}_i^k = \frac{\rho_i^{k-1} \alpha_i^k}{\alpha_i^{k-1}} \|\hat{y}_i^{k-1} - w_i^{k-1}\| + \rho_i^k \|\hat{y}_i^k - w_i^k\| + \|G_i\| \|z_i^k - z_i^{k-1}\|$$

and  $c_\alpha$  is defined in Assumption 4. When taking norms we used that  $0 < \alpha_i^{k-1} \leq 1$  and therefore that  $1/\alpha_i^{k-1} - 1 \geq 0$ . In the second inequality in (55) we used condition (17). Since  $\rho_i^k$  is bounded from above and  $\alpha_i^k$  is bounded away from 0 and above,  $\tilde{e}_i^k \rightarrow 0$  using Lemma 18, the finiteness of  $\|G_i\|$ , and Lemma 3. Furthermore,  $c_\alpha < 1$  by Assumption 4, so we may apply Lemma 6 to (55) to conclude that  $x_i^k - x_i^{k-1} \rightarrow 0$ .  $\square$

**Lemma 20.** For  $i = 1, \dots, n$ ,  $G_i z^k - x_i^k \rightarrow 0$ .

*Proof.* Recalling (52), we first write

$$\begin{aligned}
x_i^k &= (1 - \alpha_i^k)x_i^{k-1} + \alpha_i^k G_i z^k - \rho_i^k (\hat{y}_i^k - w_i^k) \\
\Leftrightarrow \alpha_i^k (G_i z^k - x_i^k) &= (1 - \alpha_i^k)(x_i^k - x_i^{k-1}) + \rho_i^k (\hat{y}_i^k - w_i^k). \tag{56}
\end{aligned}$$

Since  $\{\alpha_i^k\}$  is bounded, Lemma 19 implies that the first term on the right-hand side of (56) converges to zero. Since  $\{\rho_i^k\}$  is bounded, Lemma 18 implies that the second term on the right-hand side also converges to zero. Therefore  $\alpha_i^k \|G_i z^k - x_i^k\| \rightarrow 0$ . The claimed result then follows because  $\{\alpha_i^k\}$  is bounded away from 0.  $\square$

Finally, we can state our convergence result for Algorithm 1:

**Theorem 1.** *Suppose that assumptions 1-6 hold. If Algorithm 1 terminates by reaching line 9, then its final iterate is a member of the extended solution set  $\mathcal{S}$ . Otherwise, the sequence  $\{(z^k, \mathbf{w}^k)\}$  generated by Algorithm 1 converges weakly to some point  $(\bar{z}, \bar{\mathbf{w}})$  in the extended solution set  $\mathcal{S}$  of (2) defined in (5). Furthermore,  $x_i^k \rightharpoonup G_i \bar{z}$  and  $y_i^k \rightharpoonup \bar{w}_i$  for all  $i = 1, \dots, n-1$ ,  $x_n^k \rightharpoonup \bar{z}$ , and  $y_n^k \rightharpoonup -\sum_{i=1}^{n-1} G_i^* \bar{w}_i$ .*

*Proof.* For the finite termination result we refer to Lemma 5 of [21]. Otherwise, lemmas 17 and 20 imply that the hypotheses of Lemma 4, hold, and the result follows.  $\square$

## 6 Numerical Experiments

All our numerical experiments were implemented in Python (using `numpy`) on an Intel Xeon workstation running Linux. We restricted our attention to algorithms with comparable features and benefits to our proposed method. Thus we only considered methods that:

1. Are first-order and “fully split” the problem (that is, separate the linear operators  $G_i$  from the resolvent calculations, and use gradient-type steps for smooth functions),
2. Do not (either approximately or exactly) solve a linear system of equations at each iteration or before the first iteration,
3. Avoid having to apply “smoothing” to nonsmooth operators,
4. Incorporate a backtracking linesearch in a manner that avoids the need for bounds on Lipschitz or cocoercivity constants, and
5. Do not use iterative approximation of resolvents.

The last property we include for reasons of simplicity, while the rest contribute to making algorithms scalable and easy to apply. For a given application, there may of course be effective algorithms which could have been considered but do not satisfy all of the above requirements. However, because of the general desirability of properties 1-4 and the relative simplicity of algorithms with property 5, we only considered methods having all of them.

We compared this paper’s backtracking one-forward-step projective splitting algorithm given in Algorithm 1 (which we call `ps1fbt`) with the following methods:

- The two-forward-step projective splitting algorithm with backtracking we developed in [21] (`ps2fbt`). This method requires only Lipschitz continuity of single-valued operators, as opposed to cocoercivity.

- The adaptive three-operator splitting algorithm of [32] (**ada3op**) (where “adaptive” is used to mean “backtracking linesearch”); this method is a backtracking adaptation of the fixed-stepsize method proposed in [14]. This method requires  $G_i = I$  in problem (2) and hence can only be readily applied to two of the three test applications described below.
- The backtracking linesearch variant of the Chambolle-Pock primal-dual splitting method [28] (**cp-bt**).
- The algorithm of [12]. This is essentially Tseng’s method applied to a product-space “monotone + skew” inclusion in the following way: Assume  $T_n$  is Lipschitz monotone, problem (3) is equivalent to finding  $p \triangleq (z, w_1, \dots, w_{n-1})$  such that  $w_i \in T_i G_i z$  (which is equivalent to  $G_i z \in T_i^{-1} w_i$ ) for  $i = 1, \dots, n-1$ , and  $\sum_{i=1}^{n-1} G_i^* w_i = -T_n z$ . In other words, we wish to solve  $0 \in \tilde{A}p + \tilde{B}p$ , where  $\tilde{A}$  and  $\tilde{B}$  are defined by

$$\tilde{A}p = \{0\} \times T_1^{-1} w_1 \times \dots \times T_{n-1}^{-1} w_{n-1} \quad (57)$$

$$\tilde{B}p = \begin{bmatrix} T_n z \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & G_1^* & G_2^* & \dots & G_{n-1}^* \\ -G_1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -G_{n-1} & 0 & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} z \\ w_1 \\ \vdots \\ w_{n-1} \end{bmatrix}. \quad (58)$$

$\tilde{A}$  is maximal monotone, while  $\tilde{B}$  is the sum of two Lipschitz monotone operators (the second being skew linear), and therefore also Lipschitz monotone. The algorithm in [12] is essentially Tseng’s forward-backward-forward method [38] applied to this inclusion, using resolvent steps for  $\tilde{A}$  and forward steps for  $\tilde{B}$ . Thus, we call this method **tseng-pd**. In order to achieve good performance with **tseng-pd** we had to incorporate a diagonal preconditioner as proposed in [39].

- The recently proposed forward-reflected-backward method [29], applied to this same primal-dual inclusion  $0 \in \tilde{A}p + \tilde{B}p$  specified by (57)-(58). We call this method **frb-pd**.

Recently there have been several stochastic extensions of **ada3op** and **cp-bt** [41, 42, 31]. The method of [42] requires estimates of the Lipschitz constants and matrix norms, and so does not satisfy our experimental requirements. Since one of our problems is not in “finite-sum” format, and another includes a matrix  $G_i$  which is not equal to the identity, the methods of [41, 31] could only be applied to one of our three test problems. Even for this problem, the number of training examples in the two datasets were 60 and 127, respectively, while the feature dimensions were 7,705 and 19,806, so finite-sum methods are not particularly suitable. For these reasons we did not include these methods in our experiments.

## 6.1 Portfolio Selection

Consider the optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x) \triangleq x^\top Q x \quad \text{s.t.} \quad m^\top x \geq r, \quad \sum_{i=1}^d x_i = 1, x_i \geq 0, \quad (59)$$

where  $Q \succeq 0$ ,  $r > 0$ , and  $m \in \mathbb{R}_+^d$ . This model arises in Markowitz portfolio theory. We chose this particular problem because it features two constraint sets (a general halfspace and a simplex) onto which it is easy to project individually, but whose intersection poses a more difficult projection problem. This property makes it difficult to apply first-order methods such as ISTA/FISTA [5] as they can only perform one projection per iteration and thus cannot fully split the problem. On the other hand, projective splitting can handle an arbitrary number of constraint sets so long as one can compute projections onto each of them. We consider a fairly large instance of this problem so that standard interior point methods (for example, those in the CVXPY [15] package) are disadvantaged by their high per-iteration complexity and thus not generally competitive with first-order methods. Furthermore, backtracking variants of first-order methods are preferable for large problems as they avoid the need to estimate the largest eigenvalue of  $Q$ .

To convert (59) to a monotone inclusion, we set  $A_1 = N_{C_1}$  where  $N_{C_1}$  is the normal cone of the simplex  $C_1 = \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 0, x_i \geq 0\}$ . We set  $B_1 = 2Qx$ , which is the gradient of the objective function and is cocoercive (and Lipschitz-continuous). Finally, we set  $A_2 = N_{C_2}$ , where  $C_2 = \{x : m^\top x \geq r\}$ , and let  $B_2$  be the zero operator. Note that the resolvents of  $N_{C_1}$  and  $N_{C_2}$  (that is, the projections onto  $C_1$  and  $C_2$ ) are easily computed in  $O(d)$  operations [30]. With this notation, one may write (59) as the the problem of finding  $z \in \mathbb{R}^d$  such that

$$0 \in A_1 z + B_1 z + A_2 z,$$

which is an instance of (2) with  $n = 2$  and  $G_1 = G_2 = I$ .

To terminate each method in our comparisons, we used the following common criterion incorporating both the objective function and the constraints of (59):

$$c(x) \triangleq \max \left\{ \frac{F(x) - F^*}{F^*}, 0 \right\} - \min\{m^\top x - r, 0\} + \left| \sum_{i=1}^d x_i - 1 \right| - \max\{0, \min_i x_i\}, \quad (60)$$

where  $F^*$  is the optimal value of the problem. Note that  $c(x) = 0$  if and only if  $x$  solves (59). To estimate  $F^*$ , we used the best feasible value returned by any method after 1000 iterations.

We generated random instances of (59) as follows: we set  $d = 10,000$  to obtain a relatively large instance of the problem. We then generated a  $d \times d$  matrix  $Q_0$  with each entry drawn from  $\mathcal{N}(0, 1)$ . The matrix  $Q$  is then formed as  $(1/d) \cdot Q_0 Q_0^\top$ , which is guaranteed to be positive semidefinite. We then generate the vector  $m \in \mathbb{R}^d$  of length  $d$  to have entries uniformly distributed between 0 and 100. The constant  $r$  is set to  $\delta_r \sum_{i=1}^d m_i / d$  for various values of  $\delta_r > 0$ . We solved the problem for  $\delta_r \in \{0.5, 0.8, 1, 1.5\}$ .

All methods were initialized at the same point  $[1 \ 1 \ \dots \ 1]^\top / d$ . For all the backtracking linesearch procedures except **cp-bt**, the initial stepsize estimate is the previously discovered stepsize; at the first iteration, the initial stepsize is 1. For **cp-bt** we allowed the stepsize to increase in accordance with [28, Algorithm 4], as performance was poor otherwise. The backtracking stepsize decrement factor ( $\delta$  in Algorithm 2) was 0.7 for all algorithms.

For **ps1fbt** and **ps2fbt**,  $\rho_1^k$  was discovered via backtracking. We also set the other stepsize  $\rho_2^k$  equal to  $\rho_1^k$  at each iteration. While this is not necessary, this heuristic performed

	$\delta_r$			
	0.5	0.8	1	1.5
<b>ps1fbt</b> ( $\gamma$ )	0.01	0.01	0.5	5
<b>ps2fbt</b> ( $\gamma$ )	0.1	0.1	10	10
<b>cp-bt</b> ( $\beta^{-1}$ )	1	1	2	2
<b>tseng-pd</b> ( $\gamma_{pd}$ )	1	1	1	10
<b>frb-pd</b> ( $\gamma_{pd}$ )	1	1	10	10

Table 1: Tuning parameters for the portfolio problem (**ada3op** does not have a tuning parameter.)

well and eliminated  $\rho_2^k$  as a separately tunable parameter. For the averaging parameters in **ps1fbt**, we used  $\alpha_1^k$  fixed to 0.1 and  $\alpha_2^k$  fixed to 1 (which is allowed because  $L_2 = 0$ ). For **ps1fbt** we set  $\hat{\theta}_1 = x_1^0$  and  $\hat{w}_1 = 2Qx_1^0$ .

For **tseng-pd** and **frb-pd**, we used the following preconditioner:

$$U = \text{diag}(I_{d \times d}, \gamma_{pd}I_{d \times d}, \gamma_{pd}I_{d \times d}) \quad (61)$$

where  $U$  is used as in [39, Eq. (3.2)] for **tseng-pd** ( $M^{-1}$  on [29, p. 7] for **frb-pd**). In this case, the “monotone + skew” primal-dual inclusion described in (57)-(58) features two  $d$ -dimensional dual variables in addition to the  $d$ -dimensional primal variable. The parameter  $\gamma_{pd}$  changes the relative size of the steps taken in the primal and dual spaces, and plays a similar role to  $\gamma$  in our algorithm (see Algorithm 3). The parameter  $\beta$  in [28, Algorithm 4] plays a similar role for **cp-bt**. For all of these methods, we have found that performance is highly sensitive to this parameter: the primal and dual stepsizes need to be balanced. The only method not requiring such tuning is **ada3op**, which is a purely primal method. With this setup, all the methods have one tuning parameter except **ada3op**, which has none. For each method, we manually tuned the parameter for each  $\delta_r$ ; Table 1 shows the final choices.

We calculated the criterion  $c(x)$  in (60) for  $x_1^k$  computed by **ps1fbt** and **ps2fbt**,  $x_t$  computed on Line 3 of [32, Algorithm 1] for **ada3op**,  $y^k$  computed in [28, Algorithm 4] for **cp-bt**, and the primal iterate for **tseng-pd** and **frb-pd**. Table 2 displays the average number iterations and running time, over 10 random trials, until  $c(x)$  falls (and stays) below  $10^{-5}$  for each method. Examining the table,

- For all four problems, **ps1fbt** outperforms **ps2fbt**. This behavior is not surprising, as **ps1fbt** only requires one forward step per iteration, rather than two. Since the matrix  $Q$  is large and dense, reducing the number of forward steps should have a sizable impact.
- For  $\delta_r < 1$ , **ps1fbt** is the best-performing method. However, for  $\delta_r \geq 1$ , **ada3op** is the quickest.

	$\delta_r$			
	0.5	0.8	1	1.5
ps1fbt	<b>3.6</b> (102)	<b>4.7</b> (102)	16.3 (583)	8.5 (255.2)
ps2fbt	5.0 (151.1)	7.9 (155)	24.3 (523.4)	9.2 (222.9)
ada3op	5.3 (180.8)	9.2 (180.8)	<b>6.8</b> (174.3)	<b>3.4</b> (89.2)
cp-bt	6.2 (136)	8.3 (134.3)	11.8 (218.4)	5.6 (113.6)
tseng-pd	15.9 (387.1)	21 (387.8)	25.7 (525.3)	11.1 (245.4)
frb-pd	10.5 (559.9)	16.4 (560.4)	22.8 (1074.8)	6.3 (350.8)

Table 2: For the portfolio problem, average running times in seconds and iterations (in parentheses) for each method until  $c(x) < 10^{-5}$  for all subsequent iterations across 10 trials. The best time in each column is in **bold**.

## 6.2 Sparse Group Logistic Regression

Consider the following problem:

$$\min_{\substack{x_0 \in \mathbb{R} \\ x \in \mathbb{R}^d}} \left\{ \sum_{i=1}^n \log\left(1 + \exp(-y_i(x_0 + a_i^\top x))\right) + \lambda_1 \|x\|_1 + \lambda_2 \sum_{g \in \mathcal{G}} \|x_g\|_2 \right\}, \quad (62)$$

where  $a_i \in \mathbb{R}^d$  and  $y_i \in \{\pm 1\}$  for  $i = 1, \dots, n$  are given data,  $\lambda_1, \lambda_2 \geq 0$  are regularization parameters, and  $\mathcal{G}$  is a set of subsets of  $\{1, \dots, d\}$  such that no element is in more than one group  $g \in \mathcal{G}$ . This is the non-overlapping group-sparse logistic regression problem, which has applications in bioinformatics, image processing, and statistics [35]. It is well understood that the  $\ell_1$  penalty encourages sparsity in the solution vector. On the other hand the group-sparse penalty encourages *group sparsity*, meaning that as  $\lambda_2$  increases more groups in the solution will be set entirely to 0. The group-sparse penalty can be used when the features/predictors can be put into correlated groups in a meaningful way. As with the portfolio experiment, this problem features two nonsmooth regularizers and so methods like FISTA cannot easily be applied.

Problem (62) may be treated as a special case of (1) with  $n = 2$ ,  $G_1 = G_2 = I$ , and

$$\begin{aligned} h_1(x_0, x) &= \sum_{i=1}^n \log\left(1 + \exp(-y_i(x_0 + a_i^\top x))\right) & h_2(x_0, x) &= 0 \\ f_1(x_0, x) &= \lambda_1 \|x\|_1 & f_2(x_0, x) &= \lambda_2 \sum_{g \in \mathcal{G}} \|x_g\|_2. \end{aligned}$$

Since the logistic regression loss has a Lipschitz-continuous gradient and the  $\ell_1$ -norm and non-overlapping group-lasso penalties both have computationally simple proximal operators, all our candidate methods may be applied.

We applied 62 to two bioinformatics classification problems with real data. Following [35], we use the breast cancer dataset of [25] and the inflammatory bowel disease (IBD)

	$\lambda$ (breast cancer)			$\lambda$ (IBD)		
	0.05	0.5	0.85	0.1	0.5	1.0
# Nonzeros	114	50	20	135	40	18
# Nonzero groups	16	7	3	13	4	2
Training error	0%	5%	35%	0%	5.5%	26.8%

Table 3: The number of nonzeros and nonzero groups in the solution, along with the training error, for each value of  $\lambda$ .

dataset of [6].<sup>2</sup> The breast cancer dataset contains gene expression levels for 60 patients with estrogen-positive breast cancer. The patients were treated with tamoxifen for 5 years and classified based on whether the cancer recurred (there were 28 recurrences). The goal is to use the gene expression values to predict recurrence. The IBD data set contains gene expression levels for 127 patients, 85 of which have IBD. The IBD data set actually features three classes: ulcerative colitis (UC), Crohn’s disease (CD), and normal, and so the most natural goal would be to perform three-way classification. For simplicity, we considered a two-way classification problem of UC/CD patients versus normal patients.

For both datasets, as in [35], the group structure  $\mathcal{G}$  was extracted from the C1 dataset [36], which groups genes based on cytogenetic position data.<sup>3</sup> Genes that are in multiple C1 groups were removed from the dataset.<sup>4</sup> We also removed genes that could not be found in the C1 dataset, although doing so was not strictly necessary. After these steps, the breast cancer data had 7,705 genes in 324 groups, with each group having an average of 23.8 genes. For the IBD data there were 19,836 genes in 325 groups, with an average of 61.0 genes per group. Let  $A$  be the data matrix with each row is equal to  $a_i^\top \in \mathbb{R}^d$  for  $i = 1, \dots, n$ ; as a final preprocessing step, we normalized the columns of  $A$  to have unit  $\ell_2$ -norm, which tended to improve the performance of the first-order methods, especially the primal-dual ones.

For simplicity we set the regularization parameters to be equal:  $\lambda_1 = \lambda_2 \triangleq \lambda$ . In practice, one would typically solve (62) for various values of  $\lambda$  and then choose the final model based on cross-validation performance combined with other criteria such as sparsity. Therefore, to give an overall sense of the performance of each algorithm, we solved (62) for three values of  $\lambda$ : large, medium, and small, corresponding to decreasing the amount of regularization and moving from a relatively sparse solution to a dense solution. For the breast cancer data, we selected  $\lambda \in \{0.05, 0.5, 0.85\}$  and for IBD we chose  $\lambda \in \{0.1, 0.5, 1\}$ . The corresponding number of non-zero entries, non-zero groups, and training error of the solution are reported in Table 3. Since the goal of these experiments is to assess the computational performance of the optimization solvers, we did not break up the data into training and test sets, instead treating the entire dataset as training data.

We initialized all the methods to the 0 vector. As in the portfolio problem, all stepsizes

<sup>2</sup>The breast cancer dataset is available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1379>. The IBD dataset is available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3365>.

<sup>3</sup>The C1 dataset is available at <http://software.broadinstitute.org/gsea/index.jsp>.

<sup>4</sup>Overlapping group norms can also be handled with our method, but using a different problem formulation than (62).



	$\lambda$ (breast cancer)			$\lambda$ (IBD)		
	0.05	0.5	0.85	0.1	0.5	1.0
<b>ps1fbt</b> ( $\gamma$ )	0.05	$10^2$	$10^2$	0.1	1	1
<b>ps2fbt</b> ( $\gamma$ )	1	$10^2$	$10^5$	1	1	1
<b>cp-bt</b> ( $\beta^{-1}$ )	10	$10^3$	$10^4$	$10^4$	$10^3$	$10^5$
<b>tseng-pd</b> ( $\gamma_{pd}$ )	$10^3$	$10^5$	$10^5$	$10^4$	$10^6$	$10^6$
<b>frb-pd</b> ( $\gamma_{pd}$ )	$10^3$	$10^5$	$10^5$	$10^4$	$10^6$	$10^6$

Table 4: Tuning parameters for sparse group LR (ada3op does not have a tuning parameter).

were initially set to 1. Since the logistic regression function does not have uniform curvature, we allowed the initial trial stepsize in the backtracking linesearch to increase by a factor of 1.1 multiplied by the previously discovered stepsize. The methods **ps1fbt**, **cp-bt**, and **ada3op** have an upper bound on the trial stepsize at each iteration, so the trial stepsize was taken to be the minimum of 1.1 multiplied by the previous stepsize and this upper bound.

Otherwise, the setup was the same as the portfolio experiment. **tseng-pd** and **frb-pd** use the same preconditioner as given in (61). For **ps1fbt** and **ps2fbt** we set  $\rho_2^k$  to be equal to the discovered backtracked stepsize  $\rho_1^k$  at each iteration. For **ps1fbt** we again set  $\hat{\theta}_1 = x_1^0$ ,  $\hat{w}_1 = \nabla h_1(x_1^0)$ , and  $\alpha_1^k$  fixed to 0.1. As such, all methods (except **ada3op**) have one tuning parameter which was hand-picked for each method; the chosen values are given in Table 4.

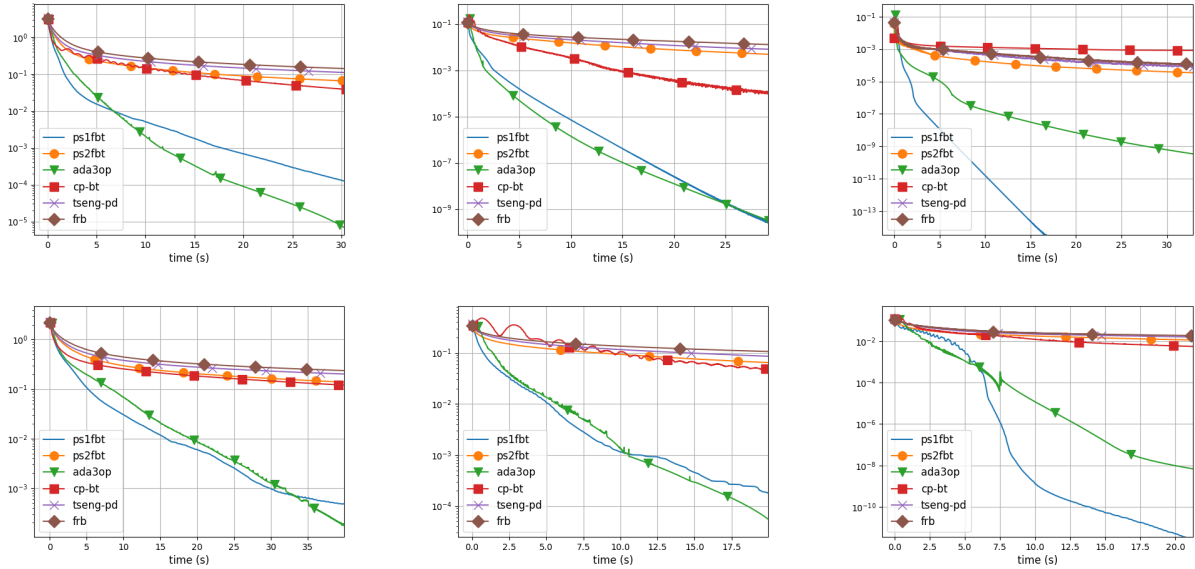


Figure 1: Results for (62) applied to bioinformatics classification problems. The top row shows breast cancer data with left:  $\lambda = 0.05$ ; middle:  $\lambda = 0.5$ ; right:  $\lambda = 0.85$ . The bottom row shows IBD data with left:  $\lambda = 0.1$ ; middle:  $\lambda = 0.5$ ; right:  $\lambda = 1.0$ . The  $y$ -axis is relative objective error:  $(F(x_0, x) - F^*)/F^*$  and the  $x$ -axis is elapsed running time in seconds.

Figure 1 shows the results of the experiments, plotting  $(F(x_0, x) - F^*)/F^*$  against time

for each algorithm, where  $F$  is the objective function in (62) and  $F^*$  is the estimated optimal value. To approximate  $F^*$ , we ran each algorithm for 4,000 iterations and took the lowest value obtained. Overall, `ps1fbt` and `ada3op` were much faster than the other methods. For the highly regularized cases (the right column of the figure), `ps1fbt` was faster than all other methods. For middle and low regularization, `ps1fbt` and `ada3op` are comparable, and for  $\lambda = 0.05$  `ada3op` is slightly faster for the breast cancer data. The methods `ps1fbt` and `ada3op` may be successful because they exploit the cocoercivity of the gradient, while `ps2fbt`, `tseng-pd`, and `frb-pd` only treat it as Lipschitz continuous. `cp-bt` also exploits cocoercivity, but its convergence was slow nonetheless. We discuss the performance of `ps1fbt` versus `ps2fbt` more in Section 6.4.

### 6.3 Rare Feature Selection

In [40], the problem of utilizing rare features in machine learning problems was studied. In many applications, certain features are rarely nonzero, making it hard to estimate their coefficients accurately. Despite this, these features can be highly informative, and the standard practice of discarding them is wasteful. The technique in [40] overcomes this difficulty by making use of an auxiliary tree data structure  $\mathcal{T}$  describing the relatedness of features. Each leaf of the tree is a feature and two features' closeness on the tree measures how "related" they are. Closely related features can then be aggregated (summed) so that more samples are captured, increasing the accuracy of the estimate for a single coefficient for the aggregated features.

To solve this aggregation and fitting problem automatically, [40] introduced the following generalized regression problem:

$$\min_{\substack{\beta_0 \in \mathbb{R} \\ \beta \in \mathbb{R}^d \\ \gamma \in \mathbb{R}^{|\mathcal{T}|}}} \left\{ \frac{1}{2n} \|\beta_0 e + X\beta - y\|_2^2 + \lambda(\alpha \|\gamma_{-r}\|_1 + (1 - \alpha) \|\beta\|_1) \mid \beta = H\gamma \right\} \quad (63)$$

where  $X \in \mathbb{R}^{n \times d}$  is the data matrix,  $y \in \mathbb{R}^n$  is the target (response) vector,  $\beta \in \mathbb{R}^d$  are the feature coefficients,  $e \in \mathbb{R}^n$  has all elements equal to 1, and  $\beta_0 \in \mathbb{R}$  is an offset. Each  $\gamma_i$  is associated with a node of the similarity tree  $\mathcal{T}$  and  $\gamma_{-r}$  means all nodes except the root. The matrix  $H \in \mathbb{R}^{d \times |\mathcal{T}|}$  contains a 1 in positions  $i, j$  for which feature  $i$  corresponds to a leaf of  $\mathcal{T}$  that is descended from node  $j$ , and elsewhere contains zeroes.  $H$  thus fuses coefficients together in the following way: if  $\gamma_i$  is nonzero for a node  $i$  and all descendants of  $\gamma_i$  in  $\mathcal{T}$  are 0, then all coefficients on the leaves which are descendant from  $\gamma_i$  are equal (see [40, Sec. 3.2] for more details). The  $\ell_1$  norm on  $\gamma$  enforces sparsity of  $\gamma$ , which in turn fuses together coefficients in  $\beta$  associated with similar features. The  $\ell_1$  norm on  $\beta$  itself additionally enforces sparsity on these coefficients, which is also desirable.

In [40], (63) is solved by a specialized application of the ADMM. The implementation involves precomputing the SVDs of  $X$  and  $H$ , and so does not fall within the scope of the methods considered in our experiments (it does not fully split the problem). Instead, we solve (63) by simply eliminating  $\beta$ , so that it becomes

$$F^* \triangleq \min_{\substack{\beta_0 \in \mathbb{R} \\ \gamma \in \mathbb{R}^{|\mathcal{T}|}}} \left\{ \frac{1}{2n} \|\beta_0 e + XH\gamma - y\|_2^2 + \lambda(\alpha \|\gamma_{-r}\|_1 + (1 - \alpha) \|H\gamma\|_1) \right\}. \quad (64)$$

This problem may be formulated as a special case of (1) with

$$\begin{aligned} f_1(t) &= \lambda(1 - \alpha)\|t\|_1 & h_1(t) &= 0 & G_1 &= H \\ f_2(\gamma, \beta_0) &= \lambda\alpha\|\gamma_{-r}\|_1 & h_2(\gamma, \beta_0) &= \frac{1}{2n}\|\beta_0 e + XH\gamma - y\|_2^2 & G_2 &= I. \end{aligned}$$

Note that  $h_2$  is Lipschitz differentiable and  $f_1$  (and  $f_2$ ) have easily-computed proximal operators. Because of the presence of the matrix  $G_1 = H$ , `ada3op` cannot easily be applied to this problem, since the proximal operator of the function  $\gamma \mapsto \|H\gamma\|_1$  cannot be computed in a simple way. All other methods, namely `ps1fbt`, `ps2fbt`, `cp-bt`, `tseng-pd`, and `frb-pd`, may still be applied.

We apply this model to a dataset of TripAdvisor reviews of hotels from [40]. The response variable was the overall review of the hotel in the set  $\{1, 2, 3, 4, 5\}$ . The features were the counts of certain adjectives in the review. Many adjectives were very rare, with 95% of the adjectives appearing in less than 5% of the reviews. The authors of [40] constructed a similarity tree using information from word embeddings and emotion lexicon labels. In the end, there were 7,573 adjectives from the 209,987 reviews and the tree  $\mathcal{T}$  had 15,145 nodes. They withheld 40,000 examples for a test set. The  $169,987 \times 7,573$  design matrix  $X$  was sparse, having only 0.32% nonzero entries. The  $7,573 \times 15,145$  matrix  $H$  arising from the similarity tree  $\mathcal{T}$  was also sparse, having 0.15% nonzero entries. In our implementation, we used the sparse matrix package `sparse` in `scipy`.

In practice, one typically would solve (64) for many values of  $(\alpha, \lambda)$  and then choose the final model based on cross validation. To assess the computational performance of the tested methods, we solve three representative examples corresponding to sparse, medium, and dense solutions. For simplicity, we fixed  $\alpha = 0.5$ . The chosen values for  $\lambda$  were  $\{10^{-5}, 10^{-2}, 10^{-1}\}$ .

The setup for the algorithms was the same as in the previous two examples, except for a few differences. For backtracking, we simply set the trial stepsize at each iteration equal to the previously discovered stepsize, as increasing it at each iteration did not provide any empirical benefit. However `cp-bt` performed better with increasing trial stepsize so we used the same scheme as before. For `ps1fbt` and `ps2fbt`, setting  $\rho_1^k$  equal to the discovered backtracking stepsize for the other operator from the previous iteration:  $\rho_2^{k-1}$ , did not work well on this model. So instead we fixed  $\rho_1^k = 1$  for `ps1fbt` and  $\rho_1^k = 10$  for `ps2fbt`, which gave the best performance across the three examples. Each tested method then has one additional tuning parameter which we hand-picked for each of the three examples. The final values are given in table 5.

The results are shown in Figure 2. For the plots, the optimal objective value  $F^*$  was estimated by running `ps1fbt` for 100,000 iterations, while the plots are shown only for the first 20,000 iterations of `ps1fbt`. The  $x$ -axis is running time excluding the time taken to actually compute the function values for the graph. Overall, there is not a large gap between the methods. However, `ps1fbt` and `cp-bt` are slightly faster for  $\lambda = 10^{-5}$ , `ps2fbt` is slightly faster for  $\lambda = 10^{-2}$ , and `ps1fbt` is slightly faster for  $\lambda = 10^{-1}$ . Since `ps1fbt` is either fastest, tied fastest, or in close second position, it is arguably the best performing algorithm overall. We suspect that the performance of `ps1fbt` (and `ps2fbt`) could greatly improve if we were to break the loss function up into blocks and use the greedy subproblem selection scheme we proposed in [21]. We plan to develop this generalization for `ps1fbt` — along with general asynchrony and block-iterativeness — in future work.

	$\lambda$		
	$10^{-5}$	$10^{-2}$	$10^{-1}$
ps1fbt ( $\gamma$ )	1	10	$10^4$
ps2fbt ( $\gamma$ )	$10^2$	10	$10^5$
cp-bt ( $\beta^{-1}$ )	10	$10^3$	$10^7$
tseng-pd ( $\gamma_{pd}$ )	1	$10^4$	$10^6$
frb-pd ( $\gamma_{pd}$ )	1	$10^4$	$10^6$

Table 5: Tuning parameters for the (64) applied to TripAdvisor data.

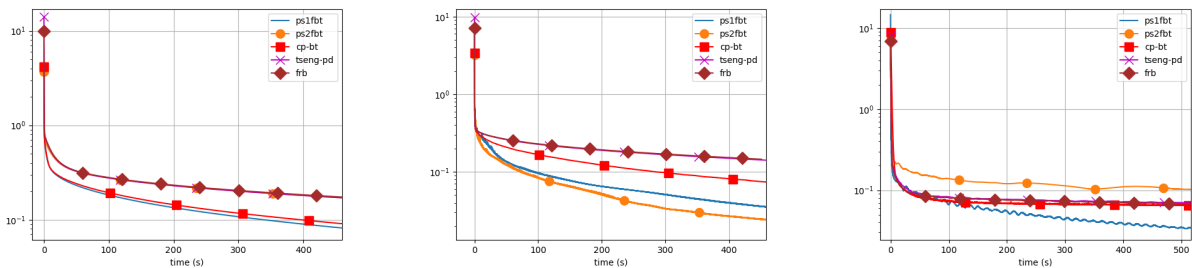


Figure 2: Results for (64) applied to tripAdvisor data. From left to right, the values of  $\lambda$  are  $\lambda = 10^{-5}$ ,  $\lambda = 10^{-2}$ , and  $\lambda = 10^{-1}$ ;  $\alpha = 0.5$  in all cases. The  $y$ -axis is relative objective error  $(F(\gamma, \beta_0) - F^*)/F^*$ , where  $F(\gamma, \beta_0)$  is the objective function in (64), and the  $x$ -axis is elapsed running time in seconds.

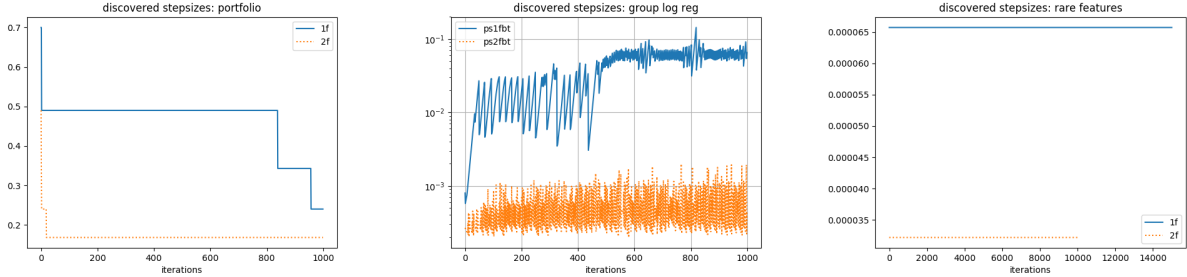


Figure 3: Discovered backtracking stepsizes for **ps1fbt** and **ps2fbt**. Left: portfolio problem with  $\delta_r = 0.5$ . Middle: group logistic regression problem applied to the IBD data with  $\lambda = 1$ . Right: rare features problem with  $\lambda = 10^{-5}$ .

## 6.4 Final Comments: **ps1fbt** versus **ps2fbt**

On the portfolio and rare feature problems, **ps1fbt** and **ps2fbt** have fairly comparable performance, with **ps1fbt** being slightly faster. However, for the group logistic regression problem, **ps1fbt** is significantly faster. Given that both methods are based on the same projective splitting framework but use different forward-step procedures to update  $(x_1^k, y_1^k)$ , this difference may be somewhat surprising. Since **ps1fbt** only requires one forward step per iteration while **ps2fbt** requires two, one might expect **ps1fbt** to be about twice as fast as **ps2fbt**. But for the group logistic regression problem, **ps1fbt** significantly outpaces this level of performance.

Examining the stepsizes returned by backtracking for both methods reveals that **ps1fbt** returns much larger stepsizes for the logistic regression problem, typically 2-3 orders of magnitude larger; see Figure 3. For the portfolio problem and the rare feature problem, where the performance of the two methods is more similar, this is not the case: the **ps1fbt** stepsizes are typically about twice as large as the **ps2fbt** stepsizes, in keeping with their theoretical upper bounds of  $1/L_i$  and  $2(1 - \alpha_i)/L_i$ , respectively.

Note that the portfolio and rare features problem both have a smooth function which is quadratic and hence has the same curvature everywhere, while group logistic regression does not. We hypothesize that the backtracking scheme in **ps1fbt** does a better job adapting to nonuniform curvature. A possible reason for this behavior is that the termination criterion for the backtracking search in **ps1fbt** may be weaker than for **ps2fbt**. For example, while **ps2fbt** requires  $\varphi_{i,k}$  to be positive at each iteration  $k$  and operator  $i$ , **ps1fbt** does not.

## Acknowledgments

We thank Xiaohan Yan and Jacob Bien for kindly sharing their data for the TripAdvisor reviews problem in Section 6.3.

This research was supported by the National Science Foundation grant CCF-1617617.

## References

- [1] Alotaibi, A., Combettes, P.L., Shahzad, N.: Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn–Tucker set. *SIAM Journal on Optimization* **24**(4), 2076–2095 (2014)
- [2] Baillon, J.B., Haddad, G.: Quelques propriétés des opérateurs angle-bornés  $n$ -cycliquement monotones. *Israel Journal of Mathematics* **26**(2), 137–150 (1977)
- [3] Bauschke, H.H., Combettes, P.L.: The Baillon-Haddad Theorem Revisited. *Journal of Convex Analysis* **17**(3-4, SI), 781–787 (2010)
- [4] Bauschke, H.H., Combettes, P.L.: *Convex analysis and monotone operator theory in Hilbert spaces*, 2nd edn. Springer (2017)
- [5] Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *Image Processing, IEEE Transactions on* **18**(11), 2419–2434 (2009)
- [6] Burczynski, M.E., Peterson, R.L., Twine, N.C., Zuberek, K.A., Brodeur, B.J., Casciotti, L., Maganti, V., Reddy, P.S., Strahs, A., Immermann, F., et al.: Molecular classification of Crohn’s disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *The Journal of Molecular Diagnostics* **8**(1), 51–61 (2006)
- [7] 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)
- [8] Combettes, P.L.: Fejér monotonicity in convex optimization. In: *Encyclopedia of optimization*, vol. 2, pp. 106–114. Springer Science & Business Media (2001)
- [9] Combettes, P.L.: Systems of structured monotone inclusions: duality, algorithms, and applications. *SIAM Journal on Optimization* **23**(4), 2420–2447 (2013)
- [10] Combettes, P.L., Eckstein, J.: Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Mathematical Programming* **168**(1-2), 645–672 (2018)
- [11] 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)
- [12] Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Valued and variational analysis* **20**(2), 307–330 (2012)
- [13] 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)

- [14] Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis* **25**(4), 829–858 (2017)
- [15] Diamond, S., Boyd, S.: CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* **17**(83), 1–5 (2016)
- [16] Dong, Y.: Weak convergence of an extended splitting method for monotone inclusions. *Optimization Online preprint* (2018)
- [17] Eckstein, J.: A simplified form of block-iterative operator splitting and an asynchronous algorithm resembling the multi-block alternating direction method of multipliers. *Journal of Optimization Theory and Applications* **173**(1), 155–182 (2017)
- [18] Eckstein, J., Svaiter, B.F.: A family of projective splitting methods for the sum of two maximal monotone operators. *Mathematical Programming* **111**(1), 173–199 (2008)
- [19] Eckstein, J., Svaiter, B.F.: General projective splitting methods for sums of maximal monotone operators. *SIAM Journal on Control and Optimization* **48**(2), 787–811 (2009)
- [20] Johnstone, P.R., Eckstein, J.: Convergence rates for projective splitting. *arXiv preprint arXiv:1806.03920* (2018)
- [21] Johnstone, P.R., Eckstein, J.: Projective splitting with forward steps: Asynchronous and block-iterative operator splitting. *arXiv preprint arXiv:1803.07043* (2018)
- [22] Johnstone, P.R., Eckstein, J.: Projective splitting with forward steps only requires continuity. *arXiv preprint arXiv:1809.07180* (2018)
- [23] Korpelevich, G.: Extragradient method for finding saddle points and other problems. *Matekon* **13**(4), 35–49 (1977)
- [24] Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis* **16**(6), 964–979 (1979)
- [25] Ma, X.J., Wang, Z., Ryan, P.D., Isakoff, S.J., Barmettler, A., Fuller, A., Muir, B., Mohapatra, G., Salunga, R., Tuggle, J.T., et al.: A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **5**(6), 607–616 (2004)
- [26] Machado, M.P.: On the complexity of the projective splitting and Spingarn’s methods for the sum of two maximal monotone operators. *Journal of Optimization Theory and Applications* **178**(1), 153–190 (2018)
- [27] Machado, M.P.: Projective method of multipliers for linearly constrained convex minimization. *Computational Optimization and Applications* **73**(1), 237–273 (2019)
- [28] Malitsky, Y., Pock, T.: A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization* **28**(1), 411–432 (2018)

- [29] Malitsky, Y., Tam, M.K.: A forward-backward splitting method for monotone inclusions without cocoercivity. arXiv preprint arXiv:1808.04162 (2018)
- [30] Michelot, C.: A finite algorithm for finding the projection of a point onto the canonical simplex of  $\mathbb{R}^n$ . *Journal of Optimization Theory and Applications* **50**(1), 195–200 (1986)
- [31] Pedregosa, F., Fatras, K., Casotto, M.: Proximal splitting meets variance reduction. In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1–10 (2019)
- [32] Pedregosa, F., Gidel, G.: Adaptive three-operator splitting. Preprint 1804.02339, ArXiv (2018)
- [33] Pesquet, J.C., Repetti, A.: A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis* **16**(12), 2453–2490 (2015)
- [34] Polyak, B.T.: *Introduction to Optimization*. Optimization Software Inc. Publications Division, New York (1987)
- [35] Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**(2), 231–245 (2013)
- [36] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43), 15,545–15,550 (2005)
- [37] Tran-Dinh, Q., Vũ, B.C.: A new splitting method for solving composite monotone inclusions involving parallel-sum operators. Preprint 1505.07946, arXiv (2015)
- [38] Tseng, P.: A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization* **38**(2), 431–446 (2000)
- [39] Vũ, B.C.: A variable metric extension of the forward–backward–forward algorithm for monotone operators. *Numerical Functional Analysis and Optimization* **34**(9), 1050–1065 (2013)
- [40] Yan, X., Bien, J.: Rare Feature Selection in High Dimensions. arXiv preprint arXiv:1803.06675 (2018)
- [41] Yurtsever, A., Vũ, B.C., Cevher, V.: Stochastic three-composite convex minimization. In: *Advances in Neural Information Processing Systems*, pp. 4329–4337 (2016)
- [42] Zhao, R., Cevher, V.: Stochastic three-composite convex minimization with a linear operator. In: A. Storkey, F. Perez-Cruz (eds.) *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, vol. 84, pp. 765–774. PMLR press, Playa Blanca, Lanzarote, Canary Islands (2018)