Abstract. Primal-dual splitting schemes are a class of powerful algorithms that solve complicated monotone inclusions and convex optimization problems that are built from many simpler pieces. They fully decompose problems that are built from sums, linear compositions, and infimal convolutions of simple functions so that each simple term is processed individually via proximal mappings, gradient mappings, and multiplications by the linear maps. This leads to easily implementable and highly parallelizable or distributed algorithms, which often obtain nearly state-of-the-art performance.

In this paper, we analyze a general monotone inclusion problem that captures a large class of primal-dual splittings as a special case. We introduce a unifying scheme and use some abstract analysis of the algorithm to deduce convergence rates of the proximal point algorithm, forward-backward splitting, Peaceman-Rachford splitting, and forward-backward-forward splitting applied to the model problem. Our ergodic convergence rates are deduced under variable metrics, stepsizes, and relaxation. Our nonergodic convergence rates are the first shown in the literature. Finally, we apply our results to a large class of primal-dual algorithms that arise as a special case of our scheme and deduce their convergence rates.

Key words. primal-dual algorithms, proximal point algorithm, forward-backward splitting, forward-backward-forward splitting, Douglas-Rachford splitting, Peaceman-Rachford splitting, nonexpansive operator, averaged operator, fixed-point algorithm

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

1. Introduction. Primal-dual algorithms are abstract splitting schemes that solve monotone inclusion and convex optimization problems. These schemes fully decompose problems built from sums, linear compositions, parallel sums, and infimal convolutions of simple functions so that each simple term is processed individually. This decomposition is achieved by cleverly combining primal and dual pair problems into a single inclusion problem, to which standard operator splitting algorithms can be applied. This gives rise to algorithms that are inherently parallel or distributed and in which expensive matrix inversions can be avoided. The characteristics of primal-dual algorithms are especially desirable for large-scale applications in machine learning, image processing, distributed optimization, and control.

Primal-dual methods have a long history with many contributors, and an attempt to summarize and relate all of the contributions is beyond the scope of this paper. In this paper, we are mainly concerned with the line of work that began in [38, 14, 23] and the many generalizations and enhancements of the basic framework that followed [18, 21, 41, 11, 9, 8, 16, 29, 5, 17]. Thus, we consider the following prototypical convex optimization problem as our guiding example:

\[
\min_{x \in H_0} f(x) + g(x) + \sum_{i=1}^{n} (h_i \square l_i)(B_ix) \tag{1.1}
\]

where \( \square \) denotes the infimal convolution operation (see Section 1.2), \( \mathcal{H}_i \) are Hilbert spaces for \( i = 0, \ldots, n \), the functions \( f, g : H_0 \to (-\infty, \infty] \) and \( h_i, l_i : \mathcal{H}_i \to (-\infty, \infty] \)

*This work is partially supported by NSF GRFP grant DGE-0707424.
†Department of Mathematics, University of California, Los Angeles Los Angeles, CA 90025, USA damek@math.ucla.edu
are closed, proper, and convex for \( i = 1, \ldots, n \), and \( B_i: \mathcal{H}_0 \to \mathcal{H}_i \) is a bounded linear map for \( i = 1, \ldots, n \).

All of the algorithms presented in this paper completely disentangle the structure of Problem (1.1) so that each iteration only involves the individual proximal operators of each of the nondifferentiable terms, the gradient operators of the differentiable terms, and multiplication by the linear maps. Thus, the maps \( B_i \) are never inverted, and we never compute proximal operators or gradients of sums or infimal convolutions of functions. We note that this level of separability is not achieved by classical splitting methods such as forward-backward splitting, Douglas-Rachford splitting, or the alternating direction method of multipliers (ADMM) applied to the primal Problem (1.1) \([12, 35, 24, 31]\).

In Problem (1.1), the maps \( B_i \) can be used as “data matrices,” in which case \( h_i \) and \( i \) are data fitting terms and \( f \) and \( g \) enforce prior knowledge on the structure of the solution, such as sparsity, low rank, or smoothness. In other cases, the maps \( h_i \) and \( i \) may be regularizers that emphasize many competing structures. We now present a particular example:

**Application: Constrained model fitting with group-structured regularizers.** Suppose we are given a measurement \( b \in \mathbb{R}^d \) and a dictionary \( A \in \mathbb{R}^{m \times d} \); our goal is to recover a highly structured signal \( x = (x_1, \ldots, x_m) \in \mathbb{R}^m \) such that \( Ax \approx b \). For example, in the hierarchical sparse coding problem (HSCP) \([27]\), we arrange the columns of \( A \) into a directed tree structure \( T \) and allow \( x_i = 0 \) only if \( x_j = 0 \) for all descendants \( j \) in \( T \) of node \( i \). Such a hierarchical representation is particularly useful for multi-scale data such as images and text documents. This type of regularization can be generalized to include arbitrary column groupings and complicated relationships between the elements of each group. Indeed, let \( G \) be a set of (possibly overlapping) subsets of \( \{1, \ldots, m\} \). For all \( S \subseteq G \) and \( x \in \mathbb{R}^m \), let \( B_S x = L_S(x_i)_{i \in S} \in \mathbb{R}^{m_S} \) where \( m_S > 0 \) and \( L_S: \mathbb{R}^{|S|} \to \mathbb{R}^{m_S} \) is a linear map. Let \( C \subseteq \mathbb{R}^m \) be a closed convex set, and let \( \chi_C \) be the convex indicator function of \( C \). For all \( S \in G \), let \( h_S: \mathbb{R}^{m_S} \to (-\infty, \infty] \) be a regularizer, and let \( l_S = \chi_{\{0\}} \), which implies \( h_S \cap \{0\} = h_S \). Then one special case of Problem (1.1) is the group-structured regularized model fitting problem:

\[
\text{minimize } \chi_C(x) + (1/2)\|Ax - b\|^2 + \sum_{S \in G} h_S(B_S x).
\]

In \([27]\), the authors consider the nonnegativity constraint \( C = \mathbb{R}_{\geq 0}^m \) and a grouping \( G \) which consists of overlapping sets \( S_i \) for \( i \in \{1, \ldots, m\} \) such that \( S_i \) contains \( i \) and all of the descendants of \( i \) in \( T \). Furthermore, for each \( S \in G \), they consider the map \( L_S = I_{\mathbb{R}^{|S|}} \) and the function \( h_S = w_S \| (x_i)_{i \in S} \|_p \) where \( p \in [1, \infty] \) and \( w_S > 0 \). This setup induces a mixed \( \ell_1/\ell_p \) norm on \( \mathbb{R}^m \) of the form \( \sum_{S \in G} w_S \| (x_i)_{i \in S} \|_p \), which tends to “zero out” entire groups of components. Note that the sum is also highly nonseparable in the components of \( x \), which can make the proximal operator of the regularization term difficult to evaluate. If we denote \( f(x) = \chi_C(x) \) and \( g(x) = (1/2)\|Ax - b\|^2 \), then the algorithms in this paper only utilize the projection \( P_C = \text{prox}_f \) onto \( C \), the gradient \( \nabla g(x) = A^*(Ax - b) \), and for all \( S \in G \) in parallel, multiplications by the maps \( B_S \) and \( B_S^* \), and evaluations of the proximal operator of the function \( h_S \). Not only does this make each iteration of the algorithm simple to implement and computationally inexpensive, it also provides a unified algorithmic framework for higher order regularizations of the components in each group, a task which might otherwise be intractable in large-scale applications.
Finally, we note that the use of infimal convolutions in applications is not widespread, so we list a few instances where they may be useful: Infimal convolutions are used in image recovery [13, Section 5] to remove staircasing effects in the total variation model. The infimal convolution of the indicator functions of two closed convex sets is the indicator function of their Minkowski sum, which has applications in motion planning for robotics [30, Section 4.3.2]. In convex analysis, the Moreau envelope of a function arises as an infimal convolution with a multiple of the squared norm [2, Section 12.4]. More generally, the infimal convolution of $h_i$ and $l_i$ can be interpreted as a regularization or smoothing of $h_i$ by $l_i$ and vice versa [2, Section 18.3].

1.1. Goals, challenges, and approaches. This work seeks to improve the theoretical understanding of the convergence rates of primal-dual splitting schemes. In this paper, we study primal-dual algorithms that are applications of standard operator splitting algorithms in product spaces consisting of primal and dual variables. Consequently, the convergence theory for these algorithms is well-developed, and they are known to converge (weakly) under mild conditions.

Although we understand when these algorithms converge, relatively little is known about their rate of convergence. For convex optimization algorithms, the ergodic convergence rate of the primal-dual gap has been analyzed in a few cases [14, 6, 5]. However, even in cases where convergence rates are known, variable metrics and stepsizes, which can significantly improve practical performance of the algorithms [37, 25], are not analyzed. In addition, we are not aware of any convergence rate analysis for the nonergodic (or last) iterate generated by these algorithms. It is important to understand nonergodic convergence rates because the ergodic (or time-averaged) iterates can “average out” structural properties, such as sparsity and low rank, that are shared by the solution and the nonergodic iterate.

The convergence rate analysis of nonergodic primal-dual algorithms largely follows from subgradient inequalities and an application of Jensen’s inequality. In contrast, the techniques developed in this paper exploit the properties of the nonexpansive operators driving the algorithms to deduce the nonergodic convergence rate of the primal-dual gap. Thus, our techniques are quite different from those used in classical convergence rate analysis and parallel the analysis developed in [22].

We summarize our contributions and techniques as follows:

(i) We describe a model monotone inclusion problem that generalizes many primal-dual formulations that appear in the literature. We provide a simple prototype algorithm to solve the model problem, and we deduce a fundamental inequality that bounds the primal-dual gap at each iteration of the algorithm. We then simplify the inequality in the special case of four splitting algorithms (Section 2).

(ii) We derive ergodic convergence rates of the variable metric forms of the relaxed proximal point algorithm (PPA), relaxed forward-backward splitting (FBS), and forward-backward-forward splitting as well as the fixed metric relaxed Peaceman-Rachford splitting (PRS) algorithm (Section 3). After some algebraic simplifications, our analysis essentially follows from an application of Jensen’s inequality.

(iii) We derive nonergodic convergence rates of relaxed PPA, relaxed FBS, and relaxed PRS (Section 4). All of our analysis follows by bounding the primal-dual gap function by a multiple of the fixed-point residual (FPR) of the nonexpansive mapping that drives the algorithm. Thus, we show that the size of the FPR can be used as a valid stopping criteria for these three algorithms.

(iv) We apply our results to deduce ergodic and nonergodic convergence rates for a large class of primal-dual algorithms that have appeared in the literature (Section 5).
Our analysis not only deduces the convergence rates of a large class of primal-dual algorithms found in the literature. It also serves as a resource for the analysis of future primal-dual algorithms that solve further generalizations of Problem 1.1, e.g. [3, 8].

1.2. Definitions, notation and some facts. In what follows, \( \mathcal{H}, \mathcal{G} \), and \( \mathbf{H} \) denote (possibly infinite dimensional) Hilbert spaces. The space \( \mathbf{H} \) will usually denote a product Hilbert space consisting of primal variables in \( \mathcal{H} \) and dual variables in \( \mathcal{G} \). In all of the algorithms we consider, we utilize two stepsize sequences: the implicit sequence \( (\gamma_j)_{j \geq 0} \subseteq \mathbb{R}_{++} \) and the explicit sequence \( (\lambda_j)_{j \geq 0} \subset \mathbb{R}_+ \). We define the \( k \)-th partial sum of the sequence \( (\gamma_j \lambda_j)_{j \geq 0} \) by the formula:

\[
\Sigma_k := \sum_{i=0}^{k} \gamma_i \lambda_i. \tag{1.2}
\]

Given a sequence \( (x^j)_{j \geq 0} \subset \mathcal{H} \), we let \( \mathcal{P}^k = (1/\Sigma_k) \sum_{i=0}^{k} \gamma_i \lambda_i x^i \) denote its \( k \)th average with respect to the sequence \( (\gamma_i \lambda_i)_{i \geq 0} \). We call a convergence result \textit{ergodic} if it is in terms of the sequence \( (\mathcal{P}^j)_{j \geq 0} \), and \textit{nonergodic} if it is in terms of \( (x^j)_{j \geq 0} \).

We denote the set of summmable nonnegative sequences by \( \ell^+([0, \infty)) := \{(\eta_j)_{j \geq 0} \subseteq [0, \infty) \mid \sum_{j=0}^{\infty} \eta_j < \infty \} \).

The following definitions and facts are mostly standard and can be found in [2, 20].

We let \( \mathcal{B}(\mathcal{H}, \mathcal{G}) \) denote the set of bounded linear maps from \( \mathcal{H} \) to \( \mathcal{G} \), and set \( \mathcal{B}(\mathcal{H}) := \mathcal{B}(\mathcal{H}, \mathcal{H}) \). We will use the notation \( I_{\mathcal{H}} \in \mathcal{B}(\mathcal{H}) \) to denote the identity map. Given a map \( L \in \mathcal{B}(\mathcal{H}, \mathcal{G}) \), we denote its adjoint by \( L^* \in \mathcal{B}(\mathcal{G}, \mathcal{H}) \). The operator norm on \( L \in \mathcal{B}(\mathcal{H}, \mathcal{G}) \) is defined by the following supremum: \( \| L \| = \sup_{x \in \mathcal{H}}(\| Lx \|/\| x \|) \).

We let \( \mathcal{S}_p(\mathcal{H}) \subseteq \mathcal{B}(\mathcal{H}) \) denote the set of linear \( p \)-strongly monotone self-adjoint maps:

\[
\mathcal{S}_p(\mathcal{H}) := \{ U \in \mathcal{B}(\mathcal{H}) \mid U = U^*, (\forall x \in \mathcal{H}) \langle Ux, x \rangle \geq \rho \| x \|^2 \}. 
\]

We define the norm and inner product induced by \( U \in \mathcal{S}_p(\mathcal{H}) \) on \( \mathcal{H} \) by the formulae: for all \( x, y \in \mathcal{H} \), \( \| x \|^2_U := \langle Ux, x \rangle \), and \( \langle x, y \rangle_U = \langle Ux, y \rangle \). The Loewner partial ordering on \( \mathcal{S}_p(\mathcal{H}) \) is as follows:

\[
U_1 \succeq U_2 \iff (\forall x \in \mathcal{H}) \| x \|^2_{U_1} \geq \| x \|^2_{U_2}.
\]

Let \( L \geq 0 \). A map \( N : \mathcal{H} \rightarrow \mathcal{H} \) is called \( L \)-Lipschitz if for all \( x, y \in \mathcal{H} \), we have \( \| Tx - Ty \| \leq L \| x - y \| \). In particular, \( N \) is called \textit{nonexpansive} if it is 1-Lipschitz. A nonexpansive map \( N : \mathcal{H} \rightarrow \mathcal{H} \) is called \( \lambda \)-averaged [2, Section 4.4] if

\[
N = T_\lambda := (1 - \lambda)I + \lambda T \tag{1.3}
\]

for a nonexpansive map \( T : \mathcal{H} \rightarrow \mathcal{H} \) and a real number \( \lambda \in (0, 1) \). A \((1/2)\)-averaged map is called \textit{firmly nonexpansive}. We will always use a * superscript to denote a fixed point of a nonexpansive map, e.g. \( z^* \).

Let \( 2^\mathcal{H} \) denote the power set of \( \mathcal{H} \). A set-valued operator \( A : \mathcal{H} \rightarrow 2^\mathcal{H} \) is called \textit{monotone} if for all \( x, y \in \mathcal{H}, u \in Ax \), and \( v \in Ay \), we have \( \langle x - y, u - v \rangle \geq 0 \). We denote the set of zeros of a monotone operator by \( \text{zer}(A) := \{ x \in \mathcal{H} \mid 0 \in Ax \} \). The \textit{graph} of \( A \) is denoted by \( \text{gra}(A) := \{(x, y) \mid x \in \mathcal{H}, y \in Ax \} \). Evidently, \( A \) is uniquely determined by its graph. A monotone operator \( A \) is called \textit{maximal monotone} provided that \( \text{gra}(A) \) is not contained in the graph of any other monotone set-valued operator. The \textit{inverse} of \( A \), denoted by \( A^{-1} \), is defined uniquely by its graph \( \text{gra}(A^{-1}) := \{(y, x) \mid \)}
\( x \in \mathcal{H}, y \in Ax \). The operator \( A \) is called \( \beta\)-strongly monotone provided that for all \( x, y \in \mathcal{H}, u \in Ax, \) and \( v \in Ay \), we have \( (x - y, u - v) \geq \beta\|x - y\|^2 \). A single-valued operator \( B : \mathcal{H} \rightarrow 2^\mathcal{H} \) maps each point in \( \mathcal{H} \) to a singleton and will be identified with the natural \( \mathcal{H} \)-valued map it defines. A single-valued operator \( B \) is called \( \rho \)-cocoercive provided that for all \( x, y \in \mathcal{H} \), we have \( (x - y, Bx - By) \geq \rho \|Bx - By\|^2 \).

Evidently, \( B \) is \( \rho \)-cocoercive whenever \( B^{−1} \) is \( \beta \)-strongly monotone. The parallel sum of (not necessarily single-valued) monotone operators \( A \) and \( B \) is given by \( \rho A \square B := (A^{−1} + B^{−1})^{−1} \). The resolvent of a maximally monotone operator \( A \) is defined by the inversion \( J_A := (I + A)^{−1} \). Minty’s theorem shows that \( J_A \) is single-valued and has full domain \( \mathcal{H} \) if, and only if, \( A \) is maximally monotone. Note that \( A \) is monotone if, and only if, \( J_A \) is firmly nonexpansive, i.e. 1-cocoercive. Thus, the reflection operator

\[
\operatorname{refl}_A := 2J_A - I_{\mathcal{H}}
\]  

(1.4)
is nonexpansive. If \( \rho > 0 \) and \( U \in \mathcal{S}_p(\mathcal{H}) \), we have \( U^{−1}A \) is maximal monotone in \( \mathcal{H} \), if, and only if, \( A \) is maximally monotone in \( \mathcal{H} \). Let \( \gamma \in (0, \infty) \). The resolvent of the map \( \gamma U^{−1}A \) has the special identity: \( J_{\gamma U^{−1}A} = U^{−1/2}J_{\gamma U^{−1}AU^{−1/2}}U^{1/2} \) [19, Example 3.9].

Let \( f : \mathcal{H} \rightarrow (−\infty, \infty] \) denote a closed, proper, and convex function. Let \( \operatorname{dom}(f) := \{ x \in \mathcal{H} \mid f(x) < \infty \} \). We will let \( \partial f(x) : \mathcal{H} \rightarrow 2^\mathcal{H} \) denote the subdifferential of \( f \). \( \partial f(x) := \{ \nabla f(x) \in \mathcal{H} \mid \text{for all } y \in \mathcal{H}, f(y) \geq f(x) + (y - x, \nabla f(x)) \} \). We will always let

\[
\nabla f(x) \in \partial f(x)
\]
denote a subgradient of \( f \) drawn at the point \( x \). The subdifferential operator of \( f \) is maximally monotone. The inverse of \( \partial f \) is given by \( \partial f^* \) where \( f^*(y) := \sup_{x \in \mathcal{H}} (y, x) - f(x) \) is the Fenchel conjugate of \( f \). If the function \( f \) is \( \beta \)-strongly convex, then \( \partial f \) is \( \beta \)-strongly monotone. If \( f \) is closed, proper, and convex function \( g : \mathcal{H} \rightarrow (−\infty, \infty] \) is Gâteaux differentiable at \( x \in \mathcal{H} \), then \( \partial g(x) = \{ \nabla g(x) \} \). The Baillon-Haddad theorem states that \( \nabla g \) is \( (1/\beta) \)-Lipschitz, if, and only if, \( \nabla g \) is \( \beta \)-cocoercive. The resolvent operator associated to \( \partial f \) is called the proximal operator and is uniquely defined by the following (strongly convex) minimization: \( \operatorname{prox}_f(x) := J_{\partial f}(x) = \inf_{y \in \mathcal{H}} f(y) + \frac{(1/2)}{\gamma} \|y - x\|^2 \). If \( \rho > 0 \), \( U \in \mathcal{S}_p(\mathcal{H}) \), and \( \gamma \in (0, \infty) \), the proximal operator of \( f \) in the metric induced by \( U \) is given by the formula: for all \( x \in \mathcal{H} \),

\[
\operatorname{prox}_{U f}(x) := J_{\gamma U^{−1} \partial f}(x) = \inf_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma} \|y - x\|^2. \quad (1.5)
\]
The infimal convolution of two closed, proper, convex, and not necessarily differentiable functions is denoted by \( f \square g : \mathcal{H} \rightarrow \mathcal{H}, x \mapsto \inf_{y \in \mathcal{H}} f(y) + g(x - y) \). The characteristic function of a closed, convex set \( C \subseteq \mathcal{H} \) is denoted by \( \chi_C : \mathcal{H} \rightarrow [0, \infty] \); the characteristic function is 0 on \( C \) and is \( \infty \) on \( \mathcal{H} \setminus C \).

Finally, we call the following identity the cosine rule:

\[
\|y - z\|^2 + 2(y - x, z - x) = \|y - x\|^2 + \|z - x\|^2, \quad \forall x, y, z \in \mathcal{H}. \quad (1.6)
\]

1.3. Assumptions. **Assumption 1 (Convexity).** Every function we consider is closed, proper, and convex. Unless otherwise stated, a function is not necessarily differentiable.

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

We employ other assumptions throughout the paper, but we list them closer to where they are invoked.
1.4. Basic properties of metrics. A simple proof of the following Lemma recently appeared in [20, Lemma 2.1]. It previously appeared in [28, Section VI.2.6].

**Lemma 1.1 (Metric properties).** Whenever \( U, V \in S_0(\mathcal{H}) \) satisfy the inequality \( \alpha I_\mathcal{H} \succ U \succ V \succ \beta I_\mathcal{H} \) for \( \alpha, \beta > 0 \), we have the ordering \((1/\beta)I_\mathcal{H} \succ V^{-1} \succ U^{-1} \succ (1/\alpha)I_\mathcal{H} \), the inclusion \( U^{-1} \in S_{\|U\|^{-1}}(\mathcal{H}) \), and the inequality \( \|U^{-1}\| \leq (1/\beta) \).

1.5. Basic properties of resolvents and averaged operators. **Proposition 1.2.** Let \( \rho > 0 \), let \( \lambda > 0 \), let \( \alpha \in (0, 1) \), let \( U \in S_\rho(\mathcal{H}) \), let \( A : \mathcal{H} \to 2^\mathcal{H} \) be a single-valued maximal monotone operator, and let \( f : \mathcal{H} \to (-\infty, \infty) \) be closed, proper, and convex.

1. Optimality conditions of \( J \): We have \( x^+ := J_{sU^{-1}(\partial f + A)}(x) \) if, and only if, there exists a unique subgradient \( \nabla f(x^+) \in \partial f(x^+) \), such that

\[
\nabla f(x^+) + Ax^+ \defeq \frac{1}{\gamma} U(x - x^+) \in \partial f(x^+) + Ax^+.
\]

2. Averaged operator contraction property: A map \( T : \mathcal{H} \to \mathcal{H} \) is \( \lambda \)-averaged in the metric induced by \( U \) if, and only if, for all \( x, y \in \mathcal{H} \)

\[
\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1-\lambda}{\lambda} \|(I_\mathcal{H} - T)x - (I_\mathcal{H} - T)y\|^2.
\]  

3. Wider relaxations: A map \( T : \mathcal{H} \to \mathcal{H} \) is \( \alpha \) averaged in \( \|\cdot\|_U \), if, and only if, \( T_\alpha \) (Equation (1.3)) is \( \lambda \alpha \)-averaged in \( \|\cdot\|_U \) for all \( \lambda \in (0, 1/\alpha) \).

**Proof.** Part 1 follows from the definition of the resolvent, Part 2 follows from [2, Proposition 4.25], and Part 3 follows from [2, Proposition 4.28]. \( \square \)

1.6. Variable metrics. Throughout this paper we will consider sequences of metrics \( (U_j)_{j \geq 0} \in S_\rho(\mathcal{H}) \) for some \( \rho > 0 \). In order to apply the standard convergence theory for variable metrics, we will make the following assumption:

**Assumption 3.** There exists a summable sequence \( (\eta_j)_{j \geq 0} \subseteq \ell^1_+(\mathbb{N}) \) such that for all \( k \geq 0 \)

\[
(1 + \eta_k)U_k \succ U_{k+1}.
\]

In addition \( \mu := \sup_{j \geq 0} \|U_k\| < \infty \). Assumption 3 is standard in variable metric algorithms [20, 41, 19, 34].

**Remark 1.** There is an asymmetry in our notation and the notation of [20, 41, 19, 34]. In our analysis, the map \( U \in S_\rho(\mathcal{H}) \) is a metric on \( \mathcal{H} \). In other papers, the maps \( U^{-1} \) denote the metric on \( \mathcal{H} \). We follow the former convention for convenience.

The following notation will be used throughout the rest of the paper.

**Proposition 1.3 (Metric parameters).** Suppose that \( (\eta_j)_{j \geq 0} \subseteq \ell^1_+(\mathbb{N}) \). Define

\[
\eta_p := \prod_{i=0}^{\infty} (1 + \eta_i) \quad \text{and} \quad \eta_s := \sum_{i=0}^{\infty} \eta_i.
\]

Then \( \eta_p \) and \( \eta_s \) are finite.

**Proof.** The quantity \( \eta_s \) is finite by definition. The finiteness of \( \eta_p \) follows from the bound: for all \( k \geq 0 \), \( \prod_{i=0}^{k} (1 + \eta_j) \leq \exp \left( \sum_{i=0}^{k} \eta_j \right) \). \( \square \)

The following Proposition is a consequence of the proof of [20, Theorem 5.1]. We give a short proof of the results we need.
**Proposition 1.4.** Let $\mathcal{H}$ be a Hilbert space. Let $\rho \in (0, \infty)$, let $(\eta_j)_{j \geq 0} \subseteq \ell_1^+(\mathbb{N})$, and let $(U_j)_{j \geq 0} \in \mathcal{S}(\mathcal{H})_\rho$ satisfy Assumption 3. For all $k \geq 0$, let $\alpha_k \in (0,1)$, let $\lambda_k \in (0,1/\alpha_k]$ be a relaxation parameter, and let $T_k : \mathcal{H} \to \mathcal{H}$ be $\alpha_k$-averaged in the metric $\| \cdot \|_{U_k}$. Furthermore, assume that there is a point $z^* \in \mathcal{H}$ such that $T_k z^* = z^*$ for all $k \geq 0$. Let the $(z^*)_{j \geq 0}$ be generated by the following Krasnosel’ski˘ı-Mann (KM)-type iteration (Equation (1.3)):

$$z^{k+1} = (T_k)_{\lambda_k} z^k.$$  

Then the following are true:

1. For all $k \geq 0$, $\| z^{k+1} - z^* \|_{U_{k+1}}^2 \leq (1 + \eta_k) \| z^k - z^* \|_{U_k}^2$ and, hence,

$$\| z^k - z^* \|_{U_{k+1}}^2 \leq \eta_0 \| z^0 - z^* \|_{U_0}^2.$$  

2. The following sum is finite:

$$\sum_{i=0}^{\infty} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \| z^{i+1} - z^{i} \|^2 \leq \frac{1}{\rho} (1 + \eta_0 \eta_k) \| z^0 - z^* \|_{U_0}^2.$$  

**Proof.** For all $k \geq 0$, the following Fejér-type inequality is true:

$$\| z^{k+1} - z^* \|_{U_k}^2 = \| (T_k)_{\lambda_k} z^k - (T_k)_{\lambda_k} z^* \|_{U_k}^2 \leq (1 + \eta_k) \| z^k - z^* \|_{U_k}^2 = (1 + \eta_k) \| z^k - z^* \|_{U_k}^2 + (1 + \eta_k) \frac{1 - \alpha_k \lambda_k}{\alpha_k \lambda_k} \| z^{k+1} - z^k \|_{U_k}^2.$$  

Rearrange and multiplying both sides by $(1 + \eta_k)$ to get:

$$\| z^{k+1} - z^* \|_{U_{k+1}}^2 + \frac{1 - \alpha_k \lambda_k}{\alpha_k \lambda_k} \| z^{k+1} - z^k \|_{U_k}^2 \leq (1 + \eta_k) \| z^k - z^* \|_{U_k}^2 + (1 + \eta_k) \frac{1 - \alpha_k \lambda_k}{\alpha_k \lambda_k} \| z^{k+1} - z^k \|_{U_k}^2.$$  

This proves Part 1.

Part 2 follows from the following bound: for all $k \geq 0$,

$$\rho \sum_{i=0}^{k} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \| z^{i+1} - z^{i} \|_{U_{i+1}}^2 \leq \sum_{i=0}^{k} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \| z^{i+1} - z^{i} \|_{U_{i+1}}^2 \leq (1 + \eta_0 \eta_k) \| z^0 - z^* \|_{U_0}^2 + \sum_{i=0}^{k} \eta_i \| z^{i} - z^* \|_{U_i}^2 \leq \left(1 + \prod_{i=0}^{\infty} (1 + \eta_i) \right) \sum_{i=0}^{\infty} \eta_i \| z^0 - z^* \|_{U_0}^2.$$

We will use the following proposition to select parameters in the FBS algorithm. The following proof is a modification of [41, Equation (3.5)]

**Proposition 1.5.** Let $\rho > 0$, let $B : \mathcal{H} \to \mathcal{H}$ be a maximal monotone operator, and let $U \in \mathcal{S}_\rho(\mathcal{H})$. Suppose that $B$ is $\beta$-cocoercive in $\| \cdot \|$. Then $U^{-1}B$ is $\beta$-cocoercive in the norm $\| \cdot \|_{U}$. 


Proof. Recall that \(\|U^{-1}\|^{-1} \geq \rho\) (Lemma 1.1). The proof is a consequence of the following inequalities: for all \(x, y \in \mathcal{H},\)

\[
\langle U^{-1} Bx - U^{-1} By, x - y \rangle_U = \langle Bx - By, x - y \rangle
\geq \beta \|Bx - By\|^2
\geq \beta \|U^{-1}\|^{-1} \|Bx - By\| \|Bx - By\|
\geq \beta \rho \|U^{-1} Bx - U^{-1} By\| \|Bx - By\|
\geq \beta \rho \|U^{-1} Bx - U^{-1} By\| \|Bx - By\|^2.
\]

\[
\|z^{k+1} - z^*\|_{U_{k+1}}^2 \leq (1 + \eta_k)\|z^k - z^*\|_{U_k}^2.
\]

Proof. Fix \(k \geq 0\), and let \(z^* \in \text{zer}(A + B)\). For simplicity, we drop the iteration superscripts and subscripts and denote \(z := z^k, y := y^k, x := x^k, w := w^k, z^+ := z^{k+1}, U := U_k, U_+ := U_{k+1},\) and \(\eta := \eta_k\). Note that \((x, y - x) \in \text{gra}(\gamma U^{-1} A)\) and \((z^+, -\gamma U^{-1} Bz^*) \in \text{gra}(\gamma U^{-1} A)\) and, hence, by the \(U\)-monotonicity of \(U^{-1} A\), we have

\[
\langle x - z^*, x - y - \gamma U^{-1} Bz^* \rangle_U \leq 0.
\]

Furthermore, by the \(U\)-monotonicity of \(B\), we have

\[
\langle x - z^*, \gamma U^{-1} Bz^* - \gamma U^{-1} Bx \rangle_U \leq 0.
\]

Thus,

\[
\langle x - z^*, x - y - \gamma U^{-1} Bx \rangle_U = \langle x - z^*, x - y - \gamma U^{-1} Bz^* \rangle_U + \langle x - z^*, \gamma U^{-1} Bz^* - \gamma U^{-1} Bx \rangle_U \leq 0.
\]

Therefore, we have

\[
2\gamma \langle x - z^*, U^{-1} Bz - U^{-1} Bx \rangle_U = 2\langle x - z^*, U^{-1} Bz + y - x \rangle_U
+ 2\langle x - z^*, x - y - \gamma U^{-1} Bx \rangle_U
\leq 2\langle x - z^*, \gamma U^{-1} Bz + y - x \rangle_U
= 2\langle x - z^*, z - x \rangle_U
\leq \|z - z^*\|_U^2 - \|x - z^*\|_U^2 - \|z - x\|_U^2.
\]
Now we can show the Fejér property:
\[
\|z^+ - z^*\|_U^2 = \|z - y + w - z^*\|_U^2 \\
= \|x - z^* + \gamma U^{-1} B z - \gamma U^{-1} B x\|_U^2 \\
= \|x - z^*\|_U^2 + 2\gamma \langle x - z^*, U^{-1} B z - U^{-1} B x \rangle_U + \gamma^2 \|U^{-1} B z - U^{-1} B x\|_U^2 \\
\leq \|x - z^*\|_U^2 + \|z - z^*\|_U^2 - \|x - z^* - \gamma U^{-1} B z\|_U^2
\]
\[
\leq \|z - z^*\|_U^2 - \rho\|z - x\| + (\gamma^2/\rho)\|B z - B x\|^2 \\
\leq \|z - z^*\|_U^2 - \rho (1 - \gamma^2/(\beta^2 \rho^2)) \|z - x\|^2
\]
where we use \((1/\rho)T_U \geq U^{-1}\) (Lemma 1.1). The result follows because \(\gamma^2 \leq \beta^2 \rho^2\) and, hence, \(\|z^+ - z^*\|_{U^2} \leq (1 + \eta)\|z^+ - z^*\|_U^2 \leq (1 + \eta)\|z - z^*\|_U^2\). □

2. The unifying scheme. In this section, we introduce a prototype monotone inclusion problem, which generalizes and summarizes many primal-dual problem formulations found in the literature. After we describe the problem, we will introduce an abstract unifying scheme that generalizes many existing primal-dual algorithms. We will describe how to measure convergence of the unifying scheme, and introduce a fundamental inequality that bounds our measure of convergence. Finally, we will identify the key terms in the fundamental inequality and simplify them in the case of several abstract splitting algorithms.

In Section 5, we will show that this unifying scheme relates to many existing algorithms, and extend the convergence rate results of those methods.

2.1. Problem and algorithm. Throughout this paper, we consider the following problem:

**PROBLEM 1** (Prototype primal monotone problem). Let \((H, \langle \cdot, \cdot \rangle)\) be a Hilbert space, let \(f, g : H \to (-\infty, \infty]\) be proper, closed, and convex functions and let \(S : H \to H\) be a skew symmetric map: \(S^* = -S\). Then the prototype primal-dual problem is to find \(x^* \in H\) such that
\[
0 \in \partial f(x^*) + \partial g(x^*) + Sx^*	ag{2.1}
\]

Evidently, Problem 1 is a monotone inclusion problem because \(\partial f, \partial g,\) and \(S\) are maximally monotone operators on \(H\) [2, Example 20.30].

We are now ready to define our unifying scheme.

**Algorithm 1:** Unifying scheme

```
input : z^0 \in H, (\lambda_j)_{j \geq 0} \subseteq R_+, (\gamma_j)_{j \geq 0} \subseteq R_+, \rho > 0, (U_j)_{j \geq 0} \subseteq S_\rho(H)

for k = 0, 1, \ldots do

z^{k+1} = z^k - \gamma_k \lambda_k U_k^{-1} (\tilde{\nabla} f(x^k) + \tilde{\nabla} g(x^k) + Sx^k);
```

Note that the points \(x^k, x^k_f, x^k_g\) as well as the subgradients \(\tilde{\nabla} f(x^k) \in \partial f(x^k_f)\) and \(\tilde{\nabla} g(x^k) \in \partial g(x^k_g)\) are unspecified in the description of Algorithm 1. In the algorithms we study, these points and subgradients will be generated by proximal and forward gradient operators and, thus, can be determined given \(z^k\). The positive scalar sequence \((\lambda_j)_{j \geq 0}\) consists of relaxation parameters, or explicit stepsize parameters, whereas the sequence \((\gamma_j)_{j \geq 0}\) consists of proximal parameters, or implicit stepsize parameters. The strongly monotone maps \((U_j)_{j \geq 0}\) are the metrics used in each iteration of the algorithm.
In all of our applications, $H$ will be a product space of primal and dual variables. In this setting, $f$ and $g$ will be block-separable maps, and $g$ will sometimes be differentiable. The map $S$ “mixture” the primal and dual variable sequences in the product space. Mixing is necessary, because the sequences are otherwise uncoupled.

The metrics $(U_j)_{j \geq 0}$ are employed for two purposes. First, they are used because the evaluation of the resolvent $J_{\partial f + S}$, which is a basic building block most of the algorithms we study, may not be simple. Thus, the primal-dual algorithms that we study formulate special metrics $U \in S_\rho(H)$ such that $J_{U^{-1}(\partial f + S)}$ is as easy to evaluate as prox$_f$ (see Section 5). Hence, in our analysis we must at least consider fixed metrics that are different from the standard product metric on $H$. Second, we allow the metrics to vary at each iteration because it can significantly improve the practical performance of the algorithm, e.g. by employing second order information, or even simple time-varying diagonal metrics [37, 25].

2.2. Examples of the unifying scheme. In this section we will introduce four algorithms and show that they are a special case of Algorithm 1. We will also introduce several assumptions on the algorithm parameters that ensure convergence. These assumptions will remain in effect throughout the rest of the paper. Note that for brevity, we will drop this adjective throughout the rest of the paper.

2.2.1. The relaxed variable metric proximal point algorithm (PPA).

The relaxed variable metric proximal point algorithm applies to problems in which $g \equiv 0$.

**Algorithm 2:** Relaxed variable metric proximal point algorithm (PPA)

| Algorithm 2: Relaxed variable metric proximal point algorithm (PPA) |
|---|
| **input**: $z^0 \in H, (\lambda_j)_{j \geq 0} \subseteq R_{++}, (\gamma_j)_{j \geq 0} \subseteq R_{++}, \rho > 0, (U_j)_{j \geq 0} \subseteq S_\rho(H)$ |
| **for** $k = 0, 1, \ldots$ **do** |
| $x^k_f = J_{\gamma_k U^{-1}_k(\partial f + S)}(z^k)$; |
| $z^{k+1} = (1 - \lambda_k)z^k + \lambda_k x^k_f$; |
| **end for** |

**Lemma 2.1** (PPA is a special case of the unifying scheme). Algorithm 2 is a special case of the unifying scheme where for all $k \geq 0$, we have $x^k_f := x^k$.

**Proof.** The proof follows directly from Part 1 of Proposition 1. □

We can also deduce the following boundedness result from Proposition 2.4.

**Proposition 2.2** (Boundedness of PPA). Let $U \in S(H)_\rho$. Then the map

$$J_{U^{-1}(\partial f + S)}$$

is $(1/2)$-averaged in the norm $\| \cdot \|_U$. In addition, the set of fixed points of $J_{U^{-1}(\partial f + S)}$ is equal to zer $(\partial f + S)$.

Suppose that $(z^k)_{k \geq 0}$ is generated by Algorithm 2. Let $z^* \in \text{zer } (\partial f + S)$ and let $(\lambda_j)_{j \geq 0} \subseteq (0, 2]$. Then $\|z^{k+1} - z^*\|_{U_{k+1}}^2 \leq (1 + \eta_k)\|z^k - z^*\|_{U_k}^2$ and, hence,

$$\|z^k - z^*\|_{U_k}^2 \leq \eta_0 \|z^0 - z^*\|_0^2.$$

**Proof.** The map is $(1/2)$-averaged because $U^{-1}(\partial f + S)$ is maximal monotone in the Hilbert space $(H, \langle \cdot, \cdot \rangle_U)$ [2, Theorem 23.7]. The fixed-point identity follows from a
simple modification of [2, Theorem 23.38]. Finally, the boundedness follows directly from Proposition 1.4 applied to the (1/2)-averaged operators $T_k := J_{\gamma_k U_k^{-1} (\partial f + S)}$.

Our main assumption for PPA is as follows:

**Assumption 4** (PPA assumptions). We have $(\lambda_j)_{j\geq 0} \subseteq (0,2]$.

### 2.2.2. Relaxed variable metric forward-backward splitting (FBS)

The relaxed variable metric FBS algorithm can be applied whenever $g$ is differentiable and $\nabla g$ is $(1/\beta)$-Lipschitz.

**Algorithm 3:** Relaxed variable metric forward-backward algorithm (FBS)

```plaintext
input : $z^0 \in H, (\lambda_j)_{j\geq 0} \subseteq R_+, (\gamma_j)_{j\geq 0} \subseteq R_+, \rho > 0, (U_j)_{j\geq 0} \subseteq S_p(H)$

for $k = 0, 1, \ldots$ do

\begin{align*}
    x^k_f &= J_{\gamma_k U_k^{-1} (\partial f + S)} (z^k - \gamma_k U_k^{-1} \nabla g(z^k)); \\
    z^{k+1} &= (1 - \lambda_k)z^k + \lambda_k x^k_f;
\end{align*}
```

**Lemma 2.3** (FBS is a special case of the unifying scheme). *Algorithm 3* is a special case of the unifying scheme where for all $k \geq 0$, we have $x^k_g := z^k$ and $x^k_S := x^k_f$.

**Proof.** From Part 1 of Proposition 1, we have the following identity:

$$x^k_f = z^k - \gamma_k U_k^{-1} \left( \nabla f(x^k_f) + \nabla g(x^k_g) + S x^k_S \right).$$

Thus, altogether we have:

$$z^{k+1} = z^k - \gamma_k \lambda_k U_k^{-1} \left( \nabla f(x^k_f) + \nabla g(x^k_g) + S x^k_S \right).$$

QED

The following proposition will show boundedness and summability of sequences associated to the relaxed FBS algorithm.

**Proposition 2.4** (Boundedness and summability of FBS). Let $U \in S(H)_\rho$ and let $\gamma \in (0,2\beta \rho)$. Then the composition

$$T^{U,\gamma}_{FBS} := J_{\gamma U^{-1} (\partial f + S)} \circ (I_H - \gamma U^{-1} \nabla g)$$

is $\alpha_{p,\gamma}$-averaged in the norm $\| \cdot \|_U$ where

$$\alpha_{p,\gamma} := \frac{\max \{ 1, \gamma/(\beta \rho) \} \rho^2}{\max \{ 1/2, \gamma/(2\beta \rho) \} + 1}.$$

(2.3)

In addition, the set of fixed points of $T^{U,\gamma}_{FBS}$ is equal to $\text{zer} (\partial f + \partial g + S)$.

Suppose that $(z^j)'_{j\geq 0}$ is generated by Algorithm 3. Let $z^* \in \text{zer} (\partial f + \partial g + S)$, let $\varepsilon \in (0,\beta \rho)$, let $(\gamma_j)_{j\geq 0} \subseteq [0, (2\beta \rho - \varepsilon)]$, and for all $k \geq 0$, let $\lambda_k \in [0, 1/\alpha_k]$ where $\alpha_k := \alpha_{p,\gamma k}$. Then the following are true:

1. For all $k \geq 0$, $\| z^{k+1} - z^* \|_{U_k}^2 \leq (1 + \eta_k) \| z^k - z^* \|_{U_k}^2$ and, hence,

$$\| z^k - z^* \|_{U_k}^2 \leq \eta_p \| z^0 - z^* \|_{U_0}^2.$$

2. The following sum is finite:

$$\sum_{i=0}^{\infty} \frac{1 - \alpha_i \lambda_i}{\alpha_i \lambda_i} \| z^{i+1} - z^i \|^2 \leq \frac{1}{\rho} (1 + \eta_p \eta_k) \| z^0 - z^* \|_{U_0}^2.$$

(2.4)
Proof. Note that $U^{-1}\nabla g$ is $\beta\rho$-cocoercive by Proposition 1.5 and the Baillon-Haddad theorem [1]. Thus, $I_{\tilde{H}} - \gamma U^{-1}\nabla g$ is $\gamma/(2\beta\rho)$ averaged by [2, Proposition 4.33]. In addition, $J_{\gamma U^{-1}((\partial f + \partial g)\circ S)}$ is $(1/2)$-averaged by [2, Corollary 23.8]. Therefore, the formula for $\alpha_{\rho,\gamma}$ follows from [2, Proposition 4.32].

The fixed-point identity follows from a simple modification of [2, Theorem 25.1]

The other statements follow from a direct application of Proposition 1.4 to the operator $T_k := T_{\tilde{H}}^{U, J}$.

We are now ready to state our main assumptions on the FBS algorithm:

**Assumption 5 (FBS assumptions).** We assume that the following hold for the relaxed variable metric FBS algorithm:

1. We have $(\gamma_j)_{j\geq 0} \subseteq (0, 2\beta\rho - \varepsilon]$ for some $\varepsilon \in (0, \beta\rho)$.
2. For all $k \geq 0$, we have $\lambda_k \in (0, 1/\alpha_k - \delta]$ for some positive scalar $\delta < \min\{1/\alpha_j \mid j \geq 0\}$.

Part 2 of Assumption 5 ensures that the bound in Equation (2.4) is nontrivial.

### 2.2.3. Relaxed Peaceman-Rachford splitting (PRS)

In this section we fix the metric and the implicit stepsize parameters throughout the course of the algorithm. We do this because the fixed-points of the PRS operator can be different for different choices of $\gamma$ and $U$. Thus, changing these parameters will lead to an algorithm that “chooses” a new fixed-point at each iteration.

**Algorithm 4**: Relaxed Peaceman-Rachford splitting (PRS)

**input**: $x^0 \in \mathcal{H}$, $(\lambda_j)_{j \geq 0} \subseteq \mathbb{R}^+$, $\gamma > 0$, $\rho > 0$, $U \in \mathcal{S}_\rho(\mathcal{H})$, $w \in \mathbb{R}$

for $k = 0, 1, \ldots$ do

$$z^{k+1} = (1 - \frac{\lambda_k}{\gamma})z^k + \frac{\lambda_k}{\gamma}\text{refl}_{\gamma U^{-1}(\partial f + wS)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k);$$

If $\lambda_k \equiv 2$, then Algorithm 4 is called the Peaceman-Rachford splitting algorithm, and if $\lambda_k \equiv 1$, then Algorithm 4 is called the Douglas-Rachford splitting (DRS) algorithm [31].

**Lemma 2.5** (PRS is a special case of the unifying scheme). Algorithm 4 is a special case of the unifying scheme where for all $k \geq 0$, we have

$$x^k_s := J_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k) \quad \text{and} \quad x^k_f := J_{\gamma U^{-1}(\partial f + wS)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k).$$

and $x^k := wx^k_f + (1-w)x^k_g$. Furthermore,

$$z^{k+1} - z^k = \lambda_k(x^k_f - x^k_g). \quad (2.5)$$

Proof. Part 1 of Proposition 1 shows that we have unique subgradients $\nabla g(x^k_s) \in \partial g(x^k_s)$ and $\nabla f(x^k_f) \in \partial f(x^k_f)$ such that

$$\text{refl}_{\gamma U^{-1}(\partial f + wS)} \circ \text{refl}_{\gamma U^{-1}(\partial g + (1-w)S)}(z^k)$$

$$= \text{refl}_{\gamma U^{-1}(\partial f + wS)}(z^k - 2\gamma U^{-1}(\nabla g(x^k_g) + (1-w)Sx^k_s))$$

$$= z^k - 2\gamma U^{-1}(\nabla f(x^k_f) + \nabla g(x^k_g) + S(wx^k_f + (1-w)x^k_g)).$$

Therefore, if we define $x^k_s = wx^k_f + (1-w)x^k_g$, then

$$z^{k+1} = z^k - \gamma \lambda_k U^{-1}(\nabla f(x^k_f) + \nabla g(x^k_g) + Sx^k_s).$$


**Proposition 2.6.** Let $U \in S(H)_\rho$. Define the PRS operator:

$$T_{\text{PRS}} := \text{refl}_{\gamma U^{-1}(\partial f + wS)} \circ \text{refl}_{U^{-1}(\partial g + (1-w)S)}(z^k). \tag{2.6}$$

Then $T_{\text{PRS}}$ is nonexpansive in the metric $\| \cdot \|_U$.

Suppose that $(x^j)_{j \geq 0}$ is generated by Algorithm 4. Let $z^*$ be a fixed point of $T_{\text{PRS}}$, and let $(\lambda_j)_{j \geq 0} \subseteq [0, 2]$. Then for all $k \geq 0$,

$$\|z^{k+1} - z^*\|_U^2 \leq \|z^k - z^*\|_U^2. \tag{2.7}$$

**Proof.** The operator $T_{\text{PRS}}$ is nonexpansive because it is the composition of nonexpansive reflection maps [2, Corollary 23.10].

The bound in Equation (2.7) is a consequence of Theorem 1.4 applied to the (1/2)-averaged operator $(T_{\text{PRS}})_1/2$. □

**Assumption 6 (Relaxed PRS assumptions).** We have $(\lambda_j)_{j \geq 0} \subseteq (0, 2]$.

2.2.4. Forward-backward-forward splitting (FBF). The variable metric FBF algorithm can be applied whenever $\nabla g$ is assumed to be $(1/\beta)$-Lipschitz.

**Algorithm 5:** Variable metric forward-backward-forward algorithm (FBF)

```plaintext
input : $z^0 \in H$, $(\gamma_j)_{j \geq 0} \subseteq \mathbb{R}^+$, $\rho > 0$, $(U_j)_{j \geq 0} \subseteq S_{\rho}(H)$
for $k = 0, 1, \ldots$ do
    $y^k = z^k - \gamma k U_k^{-1}(\nabla g(z^k) + Sz^k)$;
    $x^k_f = J_{\gamma U_k^{-1} \partial f}(y^k)$;
    $w^k = x^k_f - \gamma_k U_k^{-1}(\nabla g(x^k_f) + Sx^k_f)$;
    $z^{k+1} = z^k - y^k + w^k$;
```

**Lemma 2.7.** Algorithm 5 is a special case of the unifying scheme where for all $k \geq 0$, we have $x^k_{S_f} := x^k_f$, and $x^k_{S_S} := x^k_f$.

Proof. There exists a unique subgradient $U_k^{-1} \nabla f(x^k_f) = y^k - x^k_f \in U_k^{-1} \partial f(x^k_f)$, and

$$z^{k+1} - z^k = w^k - y^k = w^k - x^k_f + x^k_f - y^k = -\gamma_k U_k^{-1} \left( \nabla f(x^k_f) + \nabla g(x^k_f) + Sx^k_f \right).$$

□

**Proposition 2.8.** Suppose that $(z^j)_{j \geq 0}$ is generated by Algorithm 5 and that $(\gamma_j)_{j \geq 0} \subseteq (0, \rho/(\beta^{-1} + \|S\|))$. Then for all $k \geq 0$, we have the bound $\|z^{k+1} - z^*\|^2_{U_k} \leq (1 + \eta_k)\|z^k - z^*\|^2_{U_k}$, and, hence,

$$\|z^k - z^*\|^2_{U_k} \leq \eta_k \|z^0 - z^*\|^2_{U_0}.$$
the pre-primal-dual gap function bounds the primal and dual objective errors of the iterates generated by a class of primal-dual algorithms.

Before we introduce the gap function, we analyze the optimality conditions of Problem 1. The following lemma is well-known.

**Lemma 2.9.** Let $x^* \in H$. Suppose that $x^*$ solves Problem 1, then for all $x \in \text{dom}(f) \cap \text{dom}(g)$,

$$f(x) + g(x) + \langle Sx, -x^* \rangle - f(x^*) - g(x^*) \geq 0. \quad (2.8)$$

On the other hand, if $\partial (f + g)(x^*) = \partial f(x^*) + \partial g(x^*)$ and $x^*$ satisfies Equation (2.8) for all $x \in \text{dom}(f) \cap \text{dom}(g)$, then $x^*$ solves Problem 1.

**Proof.** If $x^*$ solves Problem 1, then $-Sx^*$ is a subgradient of $f + g$ at the point $x^*$. Thus, Equation (2.8) follows after noting that $\langle Sx, x \rangle = 0$ for all $x \in H$.

The other direction follows because Equation (2.8) characterizes the set of subgradients of the form $-Sx^* \in \partial (f + g)(x^*) = \partial f(x^*) + \partial g(x^*)$. See [2, Corollary 16.38] for conditions that guarantee the additivity of the subdifferential.

Lemma 2.9 motivates the following definition:

**Definition 2.10 (Pre-primal-dual gap).** Let the setting be as in Algorithm 1. Define the pre-primal dual gap function by the formula: for all $x, x, S$, let

$$G_{\text{pre}}(x, x, x; x) = f(x) + g(x) + \langle Sx, -x^* \rangle - f(x^*) - g(x^*). \quad (2.9)$$

We name $G_{\text{pre}}$ the pre-primal-dual-gap function after the standard primal-dual gap function that appears in [14, 6, 5]. We use the word “pre” because the standard primal-dual gap function usually involves a supremum over the last variable $x$. Note that if $\partial (f + g)(x') = \partial f(x') + \partial g(x')$ and

$$\sup_{x \in H} G_{\text{pre}}(x', x', x; x) \leq 0, \quad (2.10)$$

then $x'$ is a solution of Problem 1 (Lemma 2.9).

Our goal throughout the rest of this paper is to bound the pre-primal-dual gap when $x_f = x_e = x_S$. Because of Equation (2.10), all of our upper bounds will be a function of the norm of the last component of $G_{\text{pre}}$. In some cases, we can restrict the supremum in Equation (2.10) to a smaller subset $C \subseteq H$. This is the case if, for example, $\text{dom}(f) \cap \text{dom}(g)$ is bounded. Whenever the supremum can be restricted, we obtain a meaningful convergence rate.

Finally, Lemma 2.9 shows that for all $x \in \text{dom}(f) \cap \text{dom}(g)$,

$$G_{\text{pre}}(x, x, x; x^*) \geq 0 \quad (2.11)$$

whenever $x^*$ solves Problem 1. See Section 5.1 for other lower bounds of the pre-primal-dual gap in the context of a particular convex optimization problem.

The following proposition is the main inequality that we use to bound the pre-primal-dual gap.

**Proposition 2.11 (Upper fundamental inequality for primal dual schemes).** Suppose that $(z_j)_{j \geq 0}$ generated by Algorithm 1. Let $x \in \text{dom}(f) \cap \text{dom}(g)$. Then
the following inequality holds: for all $k \geq 0$,
\[
2\gamma_k \lambda_k \mathcal{G}^{pre}(x^k_f, x^k_g, x^k_S; x) \leq \|z^k - x\|^2_{U_k} - \|z^{k+1} - x\|^2_{U_k} - \|z^{k+1} - z^k\|^2_{U_k} \\
+ 2\gamma_k \lambda_k (x^k_f - z^{k+1}, \nabla f(x^k_f)) \\
+ 2\gamma_k \lambda_k (x^k_g - z^{k+1}, \nabla g(x^k_g)) \\
+ 2\gamma_k \lambda_k (-z^{k+1}, Sx^k_S).
\]

(2.12)

\begin{proof}
First expand the norm:
\[
\|z^{k+1} - x\|^2_{U_k} = \|z^k - x\|^2_{U_k} + 2\langle x - z^{k+1}, z^k - z^{k+1}\rangle_{U_k} - \|z^{k+1} - z^k\|^2_{U_k}.
\]
Now we expand the inner product:
\[
2\langle x - z^{k+1}, z^k - z^{k+1}\rangle_{U_k} = 2\langle x - z^{k+1}, \gamma_k \lambda_k U_k^{-1} \left(\nabla f(x^k_f) + \nabla g(x^k_g) + Sx^k_S\right)\rangle_{U_k} \\
= 2\gamma_k \lambda_k \langle x - z^{k+1}, \nabla f(x^k_f)\rangle + 2\gamma_k \lambda_k \langle x - z^{k+1}, \nabla g(x^k_g)\rangle \\
+ 2\gamma_k \lambda_k \langle -z^{k+1}, Sx^k_S\rangle.
\]
We add and subtract a point in the inner products involving $f$ and $g$ and use the subgradient inequality to get:
\[
2\gamma_k \lambda_k \langle x - z^{k+1} - \nabla f(x^k_f)\rangle \leq 2\gamma_k \lambda_k \langle x^k_f - z^{k+1} - \nabla f(x^k_f)\rangle + f(x) - f(x^k_f); \\
2\gamma_k \lambda_k \langle x - z^{k+1} - \nabla g(x^k_g)\rangle \leq 2\gamma_k \lambda_k \langle x^k_g - z^{k+1} - \nabla g(x^k_g)\rangle + g(x) - g(x^k_g).
\]
Therefore Equation (2.12) follows after rearranging. \qed
\end{proof}

The upper fundamental inequality in Proposition 2.11 bounds the pre-primal-dual gap with the sum of an alternating sequence and a key term.

**Definition 2.12 (Upper key term).** Let $(z^l)_{l \geq 0}$ be generated by Algorithm 1. For all $k \geq 0$, we define the fundamental upper key term
\[
\kappa^k_u(\lambda_k) := -\|z^{k+1} - z^k\|^2_{U_k} \\
+ 2\gamma_k \lambda_k (x^k_f - z^{k+1}, \nabla f(x^k_f)) \\
+ 2\gamma_k \lambda_k (x^k_g - z^{k+1}, \nabla g(x^k_g)) \\
+ 2\gamma_k \lambda_k (-z^{k+1}, Sx^k_S).
\]

(2.13)

Throughout the rest of the paper, we will often make the dependence of the upper key term on $\lambda_k$ implicit, and denote $\kappa^k_u := \kappa^k_u(\lambda_k)$. However, in the proof of Theorem 4 we will need to keep the dependence explicit.

**2.3.1. Computing the upper key terms.** In this paper, we will need the following inequality once.

**Theorem 2.13 (Descent theorem).** Suppose that $g : \mathcal{H} \to (-\infty, \infty]$ is closed, proper, convex, and differentiable. If $\nabla g$ is $(1/\beta)$-Lipschitz, then for all $x, y \in H$, we have the upper bound
\[
g(x) \leq g(y) + \langle x - y, \nabla g(y)\rangle + \frac{1}{2\beta} \|x - y\|^2.
\]

(2.14)
Proof. See [2, Theorem 18.15(iii)]. \[ \square \]

The following proposition will compute the upper key terms induced by the PPA, FBS, PRS, and FBF algorithms. See Section 2.2 for the definitions of the points \( x^k_f, x^k_g, \) and \( x^k_S. \)

PROPOSITION 2.14 (Computing the upper key terms). Let \((z^j)_{j \geq 0}\) be generated by Algorithm 1. Then the following simplifications of the upper key terms can be made

1. Suppose assumption 4 holds. For PPA, we have
   \[
   \kappa^k_u(\lambda_k) = \left(1 - \frac{2}{\lambda_k}\right)\|z^{k+1} - z^k\|_U^2.
   \]

2. Suppose assumption 5 holds. For FBS, we have
   \[
   \kappa^k_u(\lambda_k) \leq \left(\mu - \frac{\varepsilon}{3\lambda_k}\right)\|z^{k+1} - z^k\|_U^2 + 2\gamma_k\lambda_k g(x^k_g) - 2\gamma_k\lambda_k g(x^k_f).
   \]

3. Suppose assumption 6 holds. For PRS, we have
   \[
   \kappa^k_u(\lambda_k) = \left(1 - \frac{2}{\lambda_k}\right)\|z^{k+1} - z^k\|_U^2.
   \]

4. Suppose assumption 7 holds. For FBF, we have
   \[
   \kappa^k_u(\lambda_k) \leq 0.
   \]

Proof. To simplify notation, we drop the iteration index and denote \( z = z^k, x_f := x^k_f, x_g := x^k_g, x_S := x^k_S, z^+ := z^{k+1}, \gamma := \gamma_k, \lambda := \lambda_k, U := U_k, \) and \( \kappa_u := \kappa^k_u(\lambda_k) \) throughout this proof.

For PPA, FBS, and PRS, we note that the following identities hold:

\[
\begin{align*}
  z^+ - z &= \lambda(x_f - x_g), \tag{2.15} \\
  x_S &= w x_f + (1 - w) x_g. \tag{2.16}
\end{align*}
\]

and there exists \( w \in \mathbb{R} \) such that

\[
x_f = x_g - \gamma U^{-1}\left(\nabla f(x_f) + \nabla g(x_g) + S x_S\right)
\]

for a unique subgradient \( \nabla f(x_f) \in \partial f(x_f). \)

Now we claim that in PPA and FBS, \( w = 1 \) (see Section 2.2). In PRS, \( w \) is a parameter of the algorithm, and Equations (2.16) and (2.15) are shown in Proposition 2.5. Furthermore, Part 1 of Proposition 1.2 shows that in PPA and FBS,

\[
x_f = x_g - \gamma U^{-1}\left(\nabla f(x_f) + \nabla g(x_g) + S x_S\right)
\]

for a unique subgradient \( \nabla f(x_f) \in \partial f(x_f). \)

Now we claim that in PPA, FBS, and PRS,

\[
\kappa_u = 2(x_f + \gamma U^{-1}(\nabla g(x_g) + (1 - w)S x_S) - z^+ - z)U - \|z^+ - z\|_U^2 \tag{2.17}
\]

Because \( x_S = x_g + w(x_f - x_g) = x_g + (w/\lambda)(z^+ - z) \) and \( (Sx, x) = 0 \) for all \( x \in \mathbb{H}, \) we have the simplification:

\[
2(z - z^+, \gamma(1 - w)S x_S) = 2(z - z^+, \gamma(1 - w)S x_g). \tag{2.18}
\]
Therefore,

\[
\kappa_u = -\|z^* - z\|_U^2 + 2\gamma \lambda (x_f - z^*, \nabla f(x_f)) + 2\gamma \lambda (x_g - z^*, \nabla g(x_g)) + 2\gamma \lambda (x_s - z^*, Sx_s) = -\|z^* - z\|_U^2 + 2\gamma \lambda (x_f - z^*, \nabla f(x_f)) + 2\gamma \lambda (x_g - x_f, \nabla g(x_g)) + 2\gamma \lambda (x_s - x_f, Sx_s) + 2\gamma \lambda (x_f - z^*, Sx_s)
\]

\[
= -\|z^* - z\|_U^2 + 2\gamma \lambda (x_f - z^*, \nabla f(x_f) + \nabla g(x_g) + Sx_s)
\]

\[
= -\|z^* - z\|_U^2 + 2\gamma \lambda (x_f - z^*, \nabla f(x_f) + \nabla g(x_g) + Sx_s)
\]

\[
= 2(x_f + \gamma U^{-1}(\nabla g(x_g) + (1 - w)Sx_g) - z^+, z - z^+) - \|z^* - z\|_U^2.
\]

Now we proceed with the specific cases: In PPA and FBS, we have

\[
\kappa_u \overset{(2.17)}{=} 2\langle x_f + \gamma U^{-1}\nabla g(x_g) - z^+, z - z^*\rangle_U - \|z^* - z\|_U^2
\]

\[
= 2\langle x_f - z^+, z - z^*\rangle_U + 2\gamma \langle \nabla g(x_g), z - z^*\rangle - \|z^* - z\|_U^2
\]

\[
= 2 \left(1 - \frac{1}{\lambda}\right) \|z^* - z\|_U^2 + 2\gamma \lambda (x_g - x_f) - \|z^* - z\|_U^2
\]

\[
= 2 \left(1 - \frac{2}{\lambda}\right) \|z^* - z\|_U^2 + 2\gamma \lambda (x_g) - 2\gamma \lambda x_f + \frac{\gamma}{\lambda} \|z^* - z\|^2.
\]

In PPA $g \equiv 0$ and $1/\beta = 0$, so the result follows. The inequality for FBS follows because $\gamma \leq 2\beta - \varepsilon$ and

\[
\left(1 - \frac{2}{\lambda}\right) \|z^* - z\|_U^2 + \frac{\gamma}{\lambda\beta} \|z^* - z\|^2 \leq \left(\mu + \frac{\gamma - 2\beta}{\lambda\beta}\right) \|z^* - z\|^2.
\]

For relaxed PRS, we have

\[
z^* = z + \lambda (x_f - x_g)
\]

\[
= (1 - \lambda)z + \lambda (x_f - x_g) + z
\]

\[
= (1 - \lambda)z + \lambda (x_f + \gamma U^{-1}(\nabla g(x_g) + (1 - w)Sx_g)).
\]

Therefore, from the identity in Equation (2.17), we have

\[
\kappa_u = 2\langle x_f + \gamma U^{-1}(\nabla g(x_g) + (1 - w)Sx_g) - z^+, z - z^*\rangle_U - \|z^* - z\|_U^2
\]

\[
= 2 \left(1 - \frac{1}{\lambda}\right) \|z^* - z\|_U^2 - \|z^* - z\|_U^2
\]

\[
= \left(1 - \frac{2}{\lambda}\right) \|z^* - z\|_U^2.
\]
Finally, we prove the bound for the FBF algorithm:

\[
\kappa_u = 2(x_f - z^+, \tilde{\nabla}f(x_f) + \nabla g(x_f) + Sx_f) - \|z^+ - z\|^2_U \\
= 2(x_f - z^+, z - z^+)\|U\| - \|z^+ - z\|^2_U \tag{1.6}
\]

Furthermore, the identity holds:

\[
z^+ - x_f = z^+ - z + z - x_f = -\gamma U^{-1} (\tilde{\nabla}f(x_f) + \nabla g(x_f) + Sx) + \gamma U^{-1} (\tilde{\nabla}f(x_f) + \nabla g(z) + Sz) = \gamma U^{-1} (\nabla g(z) + Sz - \nabla g(x_f) - Sx_f). \tag{2.19}
\]

Note that the operator \(\nabla g + S\) is \((1/\beta) + \|S\|\) Lipschitz. Thus,

\[
\|x_f - z^+\|^2_U - \|x_f - z\|^2_U \leq \gamma^2 \|\nabla g(z) + Sz - \nabla g(x_f) - Sx_f\|^2_U - \|x_f - z\|^2_U \\
\leq \left(\frac{\gamma^2}{\rho} \left(\frac{1}{\beta} + \|S\|\right)^2 - \rho\right) \|x_f - z\|^2_U \\
\leq 0,
\]

where we use the following bound: for all \(x \in H\), \(\|x\|^2_U - (1/\rho)\|x\|^2\) (Lemma 1.1).

3. Ergodic convergence. In this section, we prove an ergodic convergence rate for the pre-primal-dual gap. To this end, we recall the partial sum sequence \(\Sigma_k = \sum_{i=0}^k \gamma_i \lambda_i\), and for every sequence of vectors \((x^i)_{i\geq 0} \subseteq H\), we define the ergodic sequence \(x^i = (1/\Sigma_k) \sum_{i=0}^k \gamma_i \lambda_i x^i\). For each algorithm, Theorem 3.2 (below) produces an ergodic sequence \((x^i)_{i\geq 0}\) such that for all bounded subsets \(D \subseteq H\), we have

\[
\sup_{x \in D} G^{pre}(x^i, x^i, x^i; x) = O\left(\frac{1 + \sup_{x \in D} \|x\|^2}{\Sigma_k}\right).
\]

This bound is a generalization of the primal-dual gap bounds shown in [14, 6, 5]. See Section 5.1 for several lower bounds of the pre-primal-dual gap.

Before we prove our ergodic rates, we need to prove a bound for PRS. Recall that we only analyze the PRS algorithm when the metric \(U_k \equiv U\) is fixed. The following lemma will help us deduce the convergence rate of the PRS algorithm whenever \(f\) or \(g\) is Lipschitz (Part 3 of Theorem 3.2).

**Lemma 3.1.** Suppose that \((x^i)_{i\geq 0}\) be generated by the relaxed PRS algorithm, and let \(\Lambda_k = \sum_{i=0}^k \lambda_k\). Then the following ergodic bound holds:

\[
\|x^k - x^\star\|_U \leq \frac{2\|x^0 - x^\star\|_U}{\Lambda_k}.
\]
Proof. The identity $\lambda_k (x^k_f - x^k_r) = z^{k+1} - z^k$ (Equation (2.5)) and the Fejér property in Equation (2.7) show that
\[ \|x^k_f - x^k_r\|_U = \left\| \frac{\gamma}{\sum_k} \sum_{i=0}^{k} \lambda_i (x^i_f - x^i_r) \right\|_U = \frac{\|z^{k+1} - z^0\|_U}{\Lambda_k} \leq \frac{\|z^{k+1} - z^*\|_U + \|z^0 - z^*\|_U}{\Lambda_k} \leq \frac{2\|z^0 - z^*\|_U}{\Lambda_k}. \]

Lemma 3.1 shows that the difference of splitting variables $x^k_f - x^k_r$ converges to zero with rate $O(1/\Lambda_k)$. Thus, if $f$ is Lipschitz continuous, then $|f(x^k_f) - f(x^k_r)| = O(1/\Lambda_k)$.

We are now ready to prove our main ergodic convergence results.

THEOREM 3.2 (Ergodic convergence of the unifying primal dual scheme). Suppose that the sequence $(z^j)_{j \geq 0}$ is generated by Algorithm 1, and suppose that Assumption 3 holds. Then for all $x \in H$ and all $k \geq 0$, we have the following:

1. **Ergodic convergence of PPA**: Suppose Assumption 4 holds, and let $z^*$ be a solution to the prototype problem. Then the PPA iterates satisfy
\[ G^{\text{pre}}(x^0_f, 0, x^k_f; x) \leq \frac{(1 + 2\eta_p \eta_h) \|z^0 - z^*\|_{L_0}^2 + 2\mu_h \|z^* - x\|^2}{2\Sigma_k}. \]

2. **Ergodic convergence of FBS**: Suppose Assumption 5 holds, let $z^*$ be a solution to the prototype problem, and let $\lambda = \sup_{j \geq 0} \lambda_j$. Then the FBS iterates satisfy
\[ G^{\text{pre}}(x^k_f, x^k_f, x^k_r; x) \leq \frac{\left(1 + \frac{\mu - \varepsilon/(\beta \lambda)}{\rho \inf_{j \geq 0} (1 - \alpha_j)/(\alpha_j \lambda_j)}\right) (1 + 2\eta_p \eta_h) \|z^0 - z^*\|_{L_0}^2 + 2\mu_h \|z^* - x\|^2}{2\Sigma_k}. \]

3. **Ergodic convergence of relaxed PRS**: Suppose Assumption 6 holds, and let $z^*$ be a fixed point of $T_{\text{PRS}}$. Suppose that $f$ (respectively $g$) is $L$-Lipschitz, let $x^k := x^k_r$ (respectively $x^k := x^k_f$), and let $\tilde{w} = w$ (respectively $\tilde{w} = 1 - w$). Then the PRS iterates satisfy
\[ G^{\text{pre}}(x^k, x^k, x^k; x) \leq \frac{\|z^0 - x\|_U + 4(\gamma/\rho)(L + |\tilde{w}||\|S\||\|x\||)\|z^0 - z^*\|_U}{2\Sigma_k}. \]

4. **Ergodic convergence of FBF**: Suppose Assumption 7 holds, and let $z^*$ be a solution to the prototype problem. Then the FBF iterates satisfy
\[ G^{\text{pre}}(x^k_f, x^k_f, x^k_f; x) \leq \frac{(1 + 2\eta_p \eta_h) \|z^0 - z^*\|_{L_0}^2 + 2\mu_h \|z^* - x\|^2}{2\Sigma_k}. \]

Proof. For any sequence of points $(z^j)_{j \geq 0}$ that satisfies the bound
\[ \|z^{k+1} - z^*\|_{L_{k+1}}^2 \leq (1 + \eta_k) \|z^k - z^*\|_{L_k}^2 \]
for all $k \geq 0$, we have
\[ \|z^k - z^*\|_{L_k}^2 \leq \prod_{i=0}^k (1 + \eta_i) \|z^0 - z^*\|_{L_0}^2. \]
Therefore, by the convexity of $\| \cdot \|_{U_k}^2$, we have for all $k \geq 0$,

\[
\sum_{i=0}^{k} (\| z^i - x \|_{U_i}^2 - \| z^{i+1} - x \|_{U_i}^2)
\leq \| z^0 - x \|_{U_0}^2 + \sum_{i=0}^{k} \eta_i \| z^{i+1} - x \|_{U_{i+1}}^2
\leq \| z^0 - x \|_{U_0}^2 + 2 \sum_{i=0}^{k} \eta_i \left( \| z^{i+1} - z^* \|_{U_{i+1}}^2 + \| z^* - x \|_{U_{i+1}}^2 \right)
\leq \left( 1 + 2 \left( \prod_{i=0}^{\infty} (1 + \eta_i) \right) \sum_{i=0}^{\infty} \eta_i \right) \| z^0 - z^* \|_{U_0}^2 + 2 \mu \left( \sum_{i=0}^{\infty} \eta_i \right) \| z^* - x \|_2^2
= (1 + 2 \eta_p \eta_s) \| z^0 - z^* \|_{U_0}^2 + 2 \mu \eta_s \| z^* - x \|_2^2.
\]

(3.2)

We will use Equation (3.2) to produce bounds for all of the variable metric methods.

Part 1: This follows from the Jensen’s inequality, Proposition 2.14, the fundamental inequality:

\[
G^{pre}(x^k_f, 0, x^k_f; x) \leq \frac{1}{\Sigma_k} \sum_{i=0}^{k} \gamma_i \lambda_i G^{pre}(x^i_f, 0, x^i_f; x)
\overset{(2.12)}{=} \frac{1}{2 \Sigma_k} \sum_{i=0}^{k} (\kappa_i + \| z^i - x \|_{U_i}^2 - \| z^{i+1} - x \|_{U_i}^2)\]
\overset{(3.2)}{=} \frac{1}{2 \Sigma_k} \left( (1 + 2 \eta_p \eta_s) \| z^0 - z^* \|_{U_0}^2 + 2 \mu \eta_s \| z^* - x \|_2^2 \right).
\]

Part 2: We have the following bound from Proposition 2.4:

\[
\sum_{i=0}^{k} \left( \mu - \frac{\varepsilon}{\beta \lambda_k} \right) \| z^{i+1} - z^* \|_2^2 \leq \frac{(\mu - \varepsilon / (\beta \lambda))}{\rho \inf_{j \geq 0} (1 - \alpha_j \lambda_j)/(\alpha_j \lambda_j)} \left( 1 + \eta_p \eta_s \right) \| z^0 - z^* \|_{U_0}^2.
\]

(3.3)

Thus, the bound follows from Jensen’s inequality, Proposition 2.14, and the funda-
mental inequality:
\[
G^{pre}(x_1^k, x_2^k, x_3^k; x) \leq \frac{1}{\Sigma_k} \sum_{i=0}^{k} \gamma_i \lambda_i G^{pre}(x_1^i, x_2^i, x_3^i; x)
\]
\[= \frac{1}{\Sigma_k} \sum_{i=0}^{k} \left( \gamma_i \lambda_i G^{pre}(x_1^i, x_2^i, x_3^i; x) + \gamma_i \lambda_i (x_1^i) - \gamma_i \lambda_i (x_3^i) \right)
\]
\[\leq \frac{1}{2\Sigma_k} \sum_{i=0}^{k} \left( \frac{\mu - \epsilon}{\beta \lambda_k} \|z^{i+1} - x^i\|^2 \right)
\]
\[+ \frac{1}{2\Sigma_k} \left( (1 + 2\eta_p \eta_b) \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_b \|z^* - x\|^2 \right)
\]
\[\leq \frac{1}{2\Sigma_k} \left( \frac{\mu - \epsilon/(\beta \lambda_k)}{\rho \inf_{i \geq 0} (1 - \alpha_i \lambda_i)/(\alpha_i \lambda_i)} \right) \left( (1 + 2\eta_p \eta_b) \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_b \|z^* - x\|^2 \right) / 2\Sigma_k.
\]

Part 3: This follows from the Jensen’s inequality, Proposition 2.14, and the fundamental inequality:
\[
G^{pre}(x_1^k, x_2^k, x_3^k; x) = G^{pre}(x_1^k, x_2^k, x_3^k; x) + f(x_2^k) - f(x_2^0) + (S(x_1^k - x_2^k), -x)
\]
\[\leq \frac{1}{\Sigma_k} \sum_{i=0}^{k} \gamma_i \lambda_i G^{pre}(x_1^i, x_2^i, x_3^i; x)
\]
\[+ f(x_2^k) - f(x_2^0) + (S(x_1^k - x_2^k), -x)
\]
\[\leq \frac{1}{\Sigma_k} \sum_{i=0}^{k} \left( \kappa_i^i + \|z^i - x\|^2_{U_i} - \|z^{i+1} - x\|^2_{U_i} \right)
\]
\[+ L \|x_3^k - x_1^k\| + \|S\| \|x_2^k - x_2^k\| \|x\|
\]
\[\leq \frac{\|z^0 - x\|^2_{U_0} + 4(\gamma/\rho)(L + |\mu|\|S\|\|x\|) \|z^0 - z^*\|_{U_0}}{2\Sigma_k}.
\]

Part 4: This follows from the Jensen’s inequality, Proposition 2.14 (\(\kappa_i^k \leq 0\)), and the fundamental inequality:
\[
G^{pre}(x_1^k, x_2^k, x_3^k; x) \leq \frac{1}{\Sigma_k} \sum_{i=0}^{k} \gamma_i \lambda_i G^{pre}(x_1^i, x_2^i, x_3^i; x)
\]
\[\leq \frac{1}{2\Sigma_k} \sum_{i=0}^{k} \left( \kappa_i^i + \|z^i - x\|^2_{U_i} - \|z^{i+1} - x\|^2_{U_i} \right)
\]
\[\leq \frac{1}{2\Sigma_k} \left( (1 + 2\eta_p \eta_b) \|z^0 - z^*\|_{U_0}^2 + 2\mu \eta_b \|z^* - x\|^2 \right).
\]

\[\square\]

Remark 2. The PPA, FBS, and PRS, rates in Theorem 3.2 are optimal. See [22, Proposition 8] for the result when S is the zero map.
We now deduce a nonergodic convergence rate from the proof of Theorem 3.2.

**Corollary 3.3.** Let \((x^j)_{j \geq 0}\) represent the nonergodic sequence generated by any of the algorithms in Theorem 3.2. Define a sequence \((a_j)_{j \geq 0}\) by the formula: for all \(k \geq 0\),

\[
a_k := \min \left\{ \max \{G_{pre}(x^j, x^j, x^j; x), 0\} \mid j = 0, \ldots, k \right\}.
\]

Then

\[a_k = o\left(\frac{1 + \|x\|^2}{\sum_{k=0}^{\lceil k/2 \rceil}}\right).\]

In particular, if \((\gamma_j \lambda_j)_{j \geq 0}\) is bounded away from 0, then \(a_k = o((1 + \|x\|^2)/(k + 1))\).

**Proof.** The sequence \((a_j)_{j \geq 0}\) is positive, monotonically nonincreasing, and the proof of Theorem 3.2 shows that \(\sum_{i=0}^{\infty} \gamma_i \lambda_i a_i = O(1 + \|x\|^2)\). Therefore, the proof follows from [22, Lemma 1]. \(\square\)

Corollary 3.3 only bounds the pre-primal-dual gap at the “best” iterate of each of the algorithms. This result is a useful assurance, but it is not easy to verify which iterate is minimal for all \(x\). In Section 4, we will bound the pre-primal-dual gap over the entire sequence of iterates.

**4. Nonergodic convergence.** In this section we deduce nonergodic convergence rates for PPA, FBS and PRS under the following assumption:

**Assumption 8.** For all nonergodic convergence results, we fix the metrics and the implicit stepsize parameters.

For PPA, FBS, and PRS, Theorem 4.2 (below) produces a natural sequence \((x^j)_{j \geq 0}\) such that for all bounded subsets \(D \subseteq H\), we have

\[
\sup_{x \in D} G_{pre}(x^k, x^k, x^k; x) = o\left(\frac{1 + \sup_{x \in D} \|x\|}{\sqrt{k + 1}}\right).
\]

To the best of our knowledge, the rate of convergence for the nonergodic (or last) iterate generated by the class of algorithms we study has never appeared in the literature.

Nonergodic iterates tend to share structural properties, such as sparsity or low rank, with the solution of the problem. In some cases, the ergodic iterates generated in Section 3 “average out” structural properties of the nonergodic iterates. Thus, although the ergodic iterates may be “closer” to the solution, they are often poorer partial solutions than the nonergodic iterates. The results of this section provide worst-case theoretical guarantees on the quality of the nonergodic iterates in order to justify their use in practical applications.

In our analysis, we use the following result, which follows directly from [22, Theorem 1].

**Theorem 4.1 ([22, Theorem 1]).** Suppose that \(T : H \to H\) is an \(\alpha\)-averaged operator in the norm \(\|\cdot\|_U\). Let \(z^*\) be a fixed point of \(T\), let \(\tau_k := (1 - \alpha \lambda_k) \lambda_k / \alpha\) for all \(k \geq 0\), suppose that \(\underline{\tau} := \inf_{j \geq 0} \tau_j > 0\), and suppose that \((z^j)_{j \geq 0}\) is generated by the following iteration:

\[
z^{k+1} := T_{\lambda_k}(z^k).
\]

Then

\[
\|Tz^k - z^k\|_U^2 = \frac{\|z^0 - z^*\|_U^2}{\underline{\tau}(k + 1)} \quad \text{and} \quad \|Tz^k - z^k\|_U^2 = o\left(\frac{1}{k + 1}\right).
\]

(4.2)
Throughout this section, the $T$ will always denote an $\alpha$-averaged mapping. Recall that for all $\lambda \in [0,1],$

$$\|T\lambda z^k - z^*\|_U^2 \leq \|z^k - z^*\|_U^2 - \frac{1 - \alpha\lambda}{\alpha\lambda} \|T\lambda z^k - z^k\|_U^2. \tag{4.3}$$

Equation (4.3) shows that $T\lambda z^k$ is at least as close to $z^*$ as $z^k$ is. This fact will be useful in the proof Theorem 4.2 below.

In the following theorem, we will deduce little-o and big-$O$ convergence rates. Because the pre-primal-dual gap can be negative, we slightly abuse notation: a (not necessarily positive) sequence $(a_j)_{j \geq 0}$ satisfies $a_k = o((1 + \|x\|_U)/\sqrt{k + 1})$ provided that there exists a nonnegative sequence $(b_j)_{j \geq 0}$ such that $b_k = o((1 + \|x\|_U)/\sqrt{k + 1})$ and $a_k = O(b_k).$ Note that we do not measure $|a_k|$ because our only goal is to ensure that $a_k$ is negative.

**Theorem 4.2.** Suppose that Assumption 8 holds. Then each method is a special case of Iteration (4.1). For each method, assume that $u > 0$ (see Theorem 4.1). Then for all $k \geq 0$ and all $x \in H,$ the following bounds hold:

1. **Nonergodic convergence of PPA:** Suppose assumption 4 holds, and let $z^*$ be a solution to the prototype problem. Then in PPA, $\alpha = 1/2, T = J_{U^{-1}(\partial f + S)},$

$$G^{\text{pre}}(x^k_f, 0; x^0_f; x) \leq \frac{\left(\|z^0 - z^*\|_U + \|z^* - x\|_U\right) \|z^0 - z^*\|_U}{\gamma \sqrt{k+1}},$$

and $G^{\text{pre}}(x^k_f, 0; x^0_f; x) = o\left((1 + \|x\|_U)/\sqrt{k + 1}\right).$

2. **Nonergodic convergence of FBS:** Suppose assumption 5 holds, and let $z^*$ be a solution to the prototype problem. Then in FBS, $\alpha := \alpha_{\gamma, \rho}$ (Equation (2.3)), $T = T_{FBS}^{U, \gamma}$ (Equation (2.2)),

$$G^{\text{pre}}(x^k_f, x^k_f, x^k_f; x) \leq \frac{\left(\|z^0 - z^*\|_U + \|z^* - x\|_U\right) \|z^0 - z^*\|_U}{2\gamma \sqrt{k+1}},$$

and $G^{\text{pre}}(x^k_f, x^k_f, x^k_f; x) = o\left((1 + \|x\|_U)/\sqrt{k + 1}\right).$

3. **Nonergodic convergence of PRS:** Suppose assumption 6 holds, and let $z^*$ be a fixed point of $T_{FBS}$ (Equation (2.6)). Then in PRS, $\alpha = 1/2, T = (T_{PRS})^{1/2}.$ In addition, suppose that $f$ (respectively $g$) is $L$-Lipschitz, let $x := x^k_S$ (respectively $x := x^k_f),$ and let $\hat{w} = w$ (respectively $\hat{w} = 1 - w).$ Then

$$G^{\text{pre}}(x^k_f, x^k_f, x^k_f; x) \leq \frac{\left(\|z^0 - z^*\|_U + \|z^* - x\|_U + \gamma/(\rho)(L + |\hat{w}|\|S\|\|x\|_U)\right) \|z^0 - z^*\|_U}{\gamma \sqrt{k+1}},$$

and $G^{\text{pre}}(x^k_f, x^k_f, x^k_f; x) = o\left((1 + \|x\|_U)/\sqrt{k + 1}\right).$

**Proof.** Fix $k \geq 0.$ In all of the following proofs, we will bound the pre-primal-dual gap by a quantity involving $\|Tz^k - z^k\|_U.$ Then the big-$O$ and little-o convergence rates follow directly from Theorem 4.1. In addition, we will use Equation (4.3) and the independence of $x^k_f, x^k_S,$ and $x^k_g$ from $\lambda_k$ to tighten our upper bounds. To this end, we will denote $z_\lambda := T_\lambda(z^k)$ and let $C = (0, 1/\alpha]$ denote the set of $\lambda$ for which $T_\lambda$ remains nonexpansive (see Assumptions 4, 5, and 6). Note that for $\lambda \in C,$ we have $(1/\lambda)(z_\lambda - z^*) = Tz^k - z^k$ and $\|z_\lambda - z^*\|_U \leq \|z^k - z^*\|_U$ (Equation (4.3)). Therefore,
for all $\lambda \in (0, 1/\alpha]$, we have

$$\frac{(z_{\lambda} - z^k, z_{\lambda} - x)_{\mathcal{U}}}{\lambda} \leq \|Tz^k - z^k\|_{\mathcal{U}} \|z_{\lambda} - x\|_{\mathcal{U}} \quad (4.2) \quad \frac{\|z^0 - z^*\|_U^2}{\sqrt{z(k + 1)}} \leq \frac{\|z^0 - z^*\|_U^2}{\sqrt{z(k + 1)}}. \quad (4.4)$$

Note that the upper key term identities (Proposition 2.14) and the fundamental inequality (Proposition 2.11) continue to hold when $z^{k+1}$ is replaced by $z_{\lambda}$. Thus, in each of the cases below, we will take an infimum of the fundamental inequality over all $\lambda \in C$.

**Part 1**: Proposition 2.14 proves the following identity: $\kappa^k_{\lambda}(\lambda) = (1 - 2/\lambda) \|z_{\lambda} - z^k\|_U^2$. Thus, the fundamental inequality, the cosine rule, and the identity $C = (0, 2]$, show that

$$G^{pre}(x^k_f, 0, x^k_f; x) \leq \inf_{\lambda \in C} \frac{1}{2\lambda} \left( 1 - \frac{1}{\lambda} \right) \left( \|z_{\lambda} - z^k\|_U + \|z^k - x\|_U - \|z_{\lambda} - x\|_U^2 \right) \leq \frac{1}{\gamma} \langle z_1 - z^k, z_1 - x \rangle_U \leq \frac{1}{\gamma} \left( \frac{\|z^0 - z^*\|_U + \|z^* - x\|_U}{\gamma \sqrt{z(k + 1)}} \right).$$

**Part 2**: First choose $\tilde{\lambda} \in C$ small enough that $(2\mu - \varepsilon/(\beta \tilde{\lambda})) < 0$. Now recall that Proposition 2.14 proves the following inequality: $\kappa_{\alpha}(\lambda) \leq (\mu - \varepsilon/(\beta \lambda)) \|z^{k+1} - z^k\|^2 + 2\gamma \lambda g(x^k) - 2\gamma \lambda g(x^k_f)$. Thus, the fundamental inequality, the cosine rule, and the identity $C = (0, 1/\alpha]$, show that:

$$G^{pre}(x^k_f, x^k_f, x^k_f; x) = G^{pre}(x^k_f, x^k_f, x^k_f; x) + g(x^k_f) - g(x^k) \leq \inf_{\lambda \in C} \frac{1}{2\lambda} \left( 2\gamma \lambda g(x^k_f) - 2\gamma \lambda g(x^k_f) + \kappa^k_{\lambda}(\lambda) \right) \left( \|z_\lambda - z^k\|_U + \|z^k - x\|_U - \|z_{\lambda} - x\|_U^2 \right) \leq \frac{1}{2\lambda} \left( 2\gamma \lambda - z^k, z_{\lambda} - x \right) \left( \|z_{\lambda} - z^k\|_U + \|z_{\lambda} - z^k\|_U + \left( \frac{\mu - \varepsilon}{\beta \lambda} \right) \right) \leq \frac{1}{\gamma} \left( \|z^0 - z^*\|_U + \|z^* - x\|_U \right) \frac{\|z^0 - z^*\|_U}{\gamma \sqrt{z(k + 1)}}.$$

**Part 3**: We prove the result in the case that $f$ is Lipschitz. The other case is symmetric.

Proposition 2.14 proves the following identity: $\kappa^k_{\alpha}(\lambda) = (1 - 2/\lambda) \|z_{\lambda} - z^k\|_U^2$. Thus, the fundamental inequality, the cosine rule, and the identities $x^k_{\alpha} - x^k_{f} =$
\[ (1/\lambda)(z_\lambda - z^k) = Tz^k - z^k \text{ and } C = (0, 2), \text{ show that} \]

\[
\mathcal{G}^{\text{pre}}(x^k_g, x^k_f, x^k_s; x) \leq \mathcal{G}^{\text{pre}}(x^k_f, x^k_g, x^k_s; x) + f(x^k_g) - f(x^k_f) + (S(x^k_g - x^k_s), -x)
\]
\[
\leq \inf_{\lambda \in C} \frac{1}{2 \gamma} \left( \left( 1 - \frac{2}{\lambda} \right) \| z_\lambda - z^k \|_U^2 + \| z_\lambda - z^k \|_U^2 - \| z_\lambda - x \|_U^2 \right) + L\| x^k_g - x^k_f \| + |w|\| S \|\| x^k_g - x^k_f \| \| x \|
\]
\[
\leq \inf_{\lambda \in C} \frac{1}{2 \gamma} \left( 2(z_\lambda - z^k, z_\lambda - x)_U + 2 \left( 1 - \frac{1}{\lambda} \right) \| z_\lambda - z^k \|_U^2 \right) + L\| x^k_g - x^k_f \| + |w|\| S \|\| x^k_g - x^k_f \| \| x \|
\]
\[
\leq \frac{1}{\gamma} \left( \frac{\| z_0 - z^* \|_U + \| z^* - z^0 \|_U(\| z_0 - z^* \|_U + L + |w|\| S \| \| x \|)\| z_0 - z^* \|_U}{\sqrt{\gamma(k+1)}} \right)
\]

\[ \square \]

**Remark 3.** Note that we can immediately strengthen the convergence result for PRS in Theorem 4.2. Indeed, we only need to assume that \( f \) or \( g \) is Lipschitz on the closed ball \( B(x^*, \| z^0 - z^* \|) \) of radius \( \| z^0 - z^* \| \) because

\[
\| x^k_g - x^* \|_U = \| J_{U^{-1}(\partial g + (1-w)S)}(z^k) - J_{U^{-1}(\partial g + (1-w)S)}(z^*) \|_U \leq \| z^k - z^* \|_U
\]

and, by a similar derivation, \( \| x^k_f - x^* \| \leq \| z^k - z^* \| \). Thus, the sequences lie in the ball: \( \{ x^k_j \}_{j \geq 0}, \{ x^k_j \}_{j \geq 0} \subseteq B(x^*, \| z^0 - z^* \|) \). See [2, Proposition 8.28] for conditions that ensure Lipschitz continuity convex functions on balls.

**Remark 4.** Note that the PRS rates in Theorem 4.2 above are optimal. See [22, Theorem 11] for the result when \( S \) is the zero map.

**Remark 5.** In general, it is infeasible to take the supremum over the last component of \( \mathcal{G}^{\text{pre}} \) as in Equation (2.10). Thus, in practice we cannot use the pre-primal-dual gap to measure convergence. However, Theorem 4.2 bounds the pre-primal-dual gap at the \( k \)-th iteration by a multiple of the expression \( \| Tz^k - z^k \| \| x \| \). Thus, if the supremum in Equation (2.10) can be restricted to a bounded set \( D \), then \( \| Tz^k - z^k \| \sup_{x \in D} \| x \| \) can be used as a proxy for the size of the pre-primal-dual gap. See section 5.1 for examples of such sets \( D \).

### 5. Applications

In this section we will show that the four algorithms from Section 2.2 are capable of solving highly structured optimization problems. First we introduce our model problem:

**Problem 2 (Model problem).** Let \( \mathcal{H}_0 \) be a Hilbert space, let \( f, g : \mathcal{H}_0 \to (-\infty, \infty] \) be closed, proper, and convex functions. For \( i = 1, \cdots, n \), let \( \mathcal{H}_i \) be a Hilbert space, let \( h_i, l_i : \mathcal{H}_i \to (-\infty, \infty] \) be closed, proper, and convex functions, and let \( B_i : \mathcal{H} \to \mathcal{H}_i \) be a bounded linear map. Finally, let \( \mathcal{B} : \mathcal{H}_0 \to \prod_{i=1}^n \mathcal{H}_i \) be the map \( x \mapsto (B_1x, \cdots, B_nx) \). Then our model problem is as follows:

\[
\text{minimize } f(x) + g(x) + \sum_{i=1}^n (h_i \square l_i)(B_ix). \quad (5.1)
\]
In addition, the dual problem is to

\[
\min_{y \in \prod_{i=1}^{n} H_i} (f^* \square g^*)(-B^* y) + \sum_{i=1}^{n} (h_i + l_i)(y_i).
\]

All of the algorithms we consider can directly utilize the structure of the infimal convolution in Problem 2. We note that infimal convolutions are not that wide spread in applications. Generally, we think of \( h_i \square l_i \) as a regularization of \( h_i \) by \( l_i \), or vice versa. Indeed, under mild conditions, the smoothness of at least one of \( h_i \) and \( l_i \) implies the smoothness of the infimal convolution [2, Section 18.3]. When \( l_i \) or \( h_i \) is chosen properly, this operation is sometimes called dual-smoothing [32].

However, even if both functions are not regular, the infimal convolution can be useful. For example, if \( C_1 \) and \( C_2 \) are closed, convex sets, the identity holds:

\[
\chi_{C_1 \square C_2} = \chi_{C_1 \oplus C_2}
\]

where \( C_1 \oplus C_2 = \{ x + y \mid x \in C_1, y \in C_2 \} \) is the Minkowski sum. Thus, Problem 2 contains

\[
\min_{x \in C_1 \oplus C_2} f(x) + g(x)
\]

as a special case. This optimization problem may have applications in motion planning for robotics where the Minkowski sum is used to represent obstacle sets [30, Section 4.3.2].

Finally, we note that we can remove the infimal convolution operation from Problem 2 by setting \( l_i = \chi_{\{0\}} \) because \( h_i \square l_i = h_i \) for all \( i = 1, \ldots, n \).

Our main assumption throughout this paper is the existence of a specific type of solution of Problem 2.

**Assumption 9.** We assume that there exists

\[
x^* \in \zer \left( \partial f + \partial g + B^*_i \sum_{i=1}^{n} (\partial h_i \square \partial l_i)(B_i x) \right).
\]

See [2, Proposition 16.5] or [18, Proposition 4.3] for conditions that guarantee the existence of \( x^* \). In general, the containment

\[
\zer \left( \partial f + \partial g + B^*_i \sum_{i=1}^{n} (\partial h_i \square \partial l_i)(B_i x) \right) \subseteq \zer \left( \partial \left( f + g + \sum_{i=1}^{n} (h_i \square l_i)(B_i x) \right) \right)
\]

always holds, but the sets are not necessarily equal. Nevertheless, this assumption is standard.

We now review two possible splittings of Problem 2. Both of the splittings will be designated by a “level.” The level is an indication of the number of extra dual variables that are introduced into the problem. Introducing more dual variables makes the problem further separable, and, hence, further parallelizable, but it also increases the memory footprint of the algorithm. It is unclear whether the number of dual variables affects the practical convergence speed of the algorithm in a negative way.

**Proposition 5.1** (Level 1 optimality conditions). Let \( H = \prod_{i=0}^{n} H_i \), and denote an arbitrary point \( x \in H \) by \( x = (x, y_1, \ldots, y_n) = (x, y) \). Let \( f(x) = f(x) + \sum_{i=1}^{n} h_i^*(y_i) \), let \( g(x) = g(x) + \sum_{i=1}^{n} l_i^*(y_i) \), and let \( S : H \to H \) be the skew map \( (x, y) \mapsto (B^* y, -B x) \). Then

\[
0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^{n} B^*_i (\partial h_i \square \partial l_i)(B_i x^*)
\]

(5.2)
if, and only if, there is a vector $\mathbf{y}^* \in \prod_{i=1}^n \mathcal{H}_i$ such that

$$0 \in \partial f(x^*, y^*) + \partial g(x^*, y^*) + S(x^*, y^*).$$

(5.3)

**Proof.** From the definition of the parallel sum of set-valued operators and the identity $(\partial f)^{-1} = \partial f^*$ [2, Corollary 16.24], we have $y^* \in (\partial h_i \square \partial l_i)(B_i x^*) \iff B_i x^* \in \partial h_i^*(y^*) + \partial l_i^*(y^*)$. Thus,

$$0 \in \partial f(x^*) + \partial g(x^*) + B_i^* \sum_{i=1}^n (\partial h_i \square \partial l_i)(B_i x^*)$$

$$\iff \exists \mathbf{y}^* \in \prod_{i=1}^n \mathcal{H}_i \left\{ \begin{array}{l} 0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^n B_i^* y_i^*, \\ 0 \in \partial h_i^*(y_i^*) + \partial l_i^*(y_i^*) - B_i x^*, \end{array} \right. \quad i = 1, \ldots, n.$$ 

This is exactly the inclusion in Equation (5.3). □

Notice that the functions $f$ and $g$ in Equation (5.3) are completely separable in the variables of the product space $\mathbf{H}$. Thus, evaluating the proximity operators of $f$ and $g$ can be quite simple. However, the resolvent $J_{\partial f, S}$ is not necessarily simple to evaluate. This difficulty motivates the introduction of new metrics on $\mathbf{H}$ that simplify the resolvent computation (Section 5.2).

Whenever the functions $l_i^*$ are Lipschitz differentiable, or equivalently, $l_i$ is strongly convex [2, Theorem 18.15], we can apply FBS or FBF (Algorithms 3 and 5) to the splitting in Proposition 5.1. For nonsmooth $l_i^*$, we can apply the PRS algorithm.

The proof of the following proposition is analogous to Proposition 5.1, so we omit it. It is most useful in the case that $l_i^*$ is not differentiable.

**PROPOSITION 5.2 (Level 2 optimality conditions).** Let $\mathbf{H} = \mathcal{H}_0 \times (\prod_{i=1}^n \mathcal{H}_i)^2$, and denote an arbitrary $\mathbf{x} \in \mathbf{H}$ by $\mathbf{x} = (x, y_1, \ldots, y_n, v_1, \ldots, v_n) = (x, y, v)$. Let $f(\mathbf{x}) = f(x) + \sum_{i=1}^n(h_i^*(y_i) + l_i(v_i))$, let $g(\mathbf{x}) = g(x)$, let $B : \mathcal{H}_0 \rightarrow \prod_{i=0}^n \mathcal{H}_i$ be the map $x \mapsto (B_1 x, \ldots, B_n x)$, and let $S : \mathbf{H} \rightarrow \mathbf{H}$ be the skew map $(x, y, v) \mapsto (B^* y, -B x + v, -y)$. Then

$$0 \in \partial f(x^*) + \partial g(x^*) + \sum_{i=1}^n B_i^* (\partial h_i \square \partial l_i)(B_i x^*),$$

(5.4)

if, and only if, there is a vector $(y^*, v^*) \in (\prod_{i=1}^n \mathcal{H}_i)^2$ such that

$$0 \in \partial f(x^*, y^*, v^*) + \partial g(x^*, y^*, v^*) + S(x^*, y^*, v^*).$$

(5.5)

Note that if $l_i$ is differentiable, we can “assign” it to the function $g$, instead of $f$. If $g$ is also differentiable, we can apply FBS to the resulting inclusion problem.

There are many splittings that solve Problem 2. Furthermore, the complexity of Problem 2 can be increased in various ways, e.g., by precomposing each of $h_i$ and $l_i$ with linear operators [3, 8], or by solving systems of such inclusions [16, 7]. We choose to discuss this relatively simple formulation for clarity of exposition.

The next several sections relate the results and notation of the previous sections to the level 1 and 2 splittings.
5.1. Primal-dual gap functions. In this section, we discuss the pre-primal-dual gap function in the context of the level 1 splitting in Proposition 5.1. We give sufficient conditions for the gap function (Definition 2.10) to bound the primal and dual objectives of Problem 2, and show that the pre-primal-dual gap also bounds certain squared norms that arise from the strong convexity and differentiability of the terms of the objective.

In the level 1 splitting, the pre-primal-dual gap has the following form: for all \((x, y), (x^*, y^*) \in H\), we have

\[
G^\text{pre}(x, x; x^*, y^*) = f(x) + g(x) - f(x^*) - g(x^*) + \langle x - x^*, B^* y^* \rangle \\
+ \sum_{i=1}^{n} (h_i^*(y_i) + l_i^*(y_i) - h_i^*(y_i^*) - l_i^*(y_i^*)) - \langle B x^*, y - y^* \rangle, \tag{5.6}
\]

where we used the identity \(\langle S x, -x^* \rangle = \langle S x, x - x^* \rangle\). If \(x^*\) is optimal for the inclusion \(B x^* \in \partial f(x^*) + \partial g(x^*)\) and \(B x^* \in \partial h_i^*(y_i^*) + \partial l_i^*(y_i^*)\). \(5.7\)

We will now bound several terms that arise from the strong convexity and Lipschitz differentiability of the terms in the objective function.

We follow the convention that every closed, proper, and convex function \(f\) is \(\mu_f\)-strongly convex and \(\nabla f\) is \((1/\beta_f)\)-Lipschitz. Note that we allow the constants \(\beta_f\) and \(\mu_f\) to vanish. If \(\beta_f > 0\), then \(\nabla f = \nabla f\) is Lipschitz. The following quantity is useful for summarizing the lower bounds that we derive from strong convexity and Lipschitz differentiability: for all \(x, y \in \text{dom}(f)\), if

\[
S_f(x, y) := \max \left\{ \frac{\mu_f}{2} \|x - y\|^2, \frac{\beta_f}{2} \|\nabla f(x) - \nabla f(y)\|^2 \right\}, \tag{5.8}
\]

then [2, Theorem 18.15]

\[
f(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle + S_f(x, y). \tag{5.9}
\]

We use the analogous notation for \(g\) and the conjugate functions \(h_i^*, l_i^*\) for \(i = 1, \ldots, n\). Therefore, if we apply the lower bound in Equation (5.9) to each of the functions in Equation (5.6) and use the subgradient identities in Equation (5.7) to cancel inner products, we get

\[
G^\text{pre}(x, x; x^*, y^*) \geq S_f(x, x^*) + S_g(x, x^*) + \sum_{i=1}^{n} \left( S_{h_i^*}(y_i, y_i^*) + S_{l_i^*}(y_i, y_i^*) \right). \tag{5.10}
\]

Equation (5.10) shows that convergence rates for the pre-primal-dual gap function immediately imply the same convergence rates for the \(S(\cdot, \cdot)\) functions in Equation (5.8). Note that this lower bound does not require that \(\text{dom}(f)\) or \(\text{dom}(g)\) are bounded.

The next proposition gives sufficient conditions under which the pre-primal-dual gap bounds the primal and dual objectives. In general, we cannot expect such a bound to hold, unless several terms in the objective are Lipschitz continuous or certain subdifferentials are locally bounded.

Proposition 5.3 (Level 1 gap function bounds). Let \(x^*\) be a minimizer of Problem 2. Assume the notation of Proposition 5.1. Let \(D_1 \subseteq \mathcal{H}\) and let \(D_2 \subseteq \mathcal{H}\)
\( \prod_{i=1}^{n} H_i \) be bounded sets. Then for any sequence of points \( ((x^j, y^j))_{j \geq 0} \subseteq \text{dom}(f + g) \times \prod_{i=1}^{n} \text{dom}(h_i^* + l_i^*) \), the inequality

\[
f(x^k) + g(x^k) + \sum_{i=1}^{n} (h_i^* l_i) (B_i x^k) - \left( f(x^*) + g(x^*) + \sum_{i=1}^{n} (h_i^* l_i) (B_i x^*) \right) \\
\leq \sup_{x \in \{x^*\} \times D_2} \mathcal{G}^{\text{pre}}(x^k, x^k, x^k; x)
\]

holds for all \( k \geq 0 \) provided either of the following hold:
1. \( x^* \in D_1 \) and \( \text{dom}(h_n^* + l_n^*) \times \cdots \times \text{dom}(h_1^* + l_1^*) \subseteq D_2 \);
2. \( x^* \in D_1 \) and \( \partial (h_1 \square l_1) (B_1 x^k) \times \cdots \times \partial (h_n \square l_n) (B_n x^k) \subseteq D_2 \).

Similarly, the inequality

\[
(f^* \square g^*)(-B^* y^k) + \sum_{i=1}^{n} (h_i^* l_i^*) (y_i^k) - \left( f^* \square g^*)(-B^* y^*) + \sum_{i=1}^{n} (h_i^* l_i^*) (B_i y_i^*) \right) \\
\leq \sup_{x \in D_1 \times \{y^*\}} \mathcal{G}^{\text{pre}}(x^k, x^k, x^k; x)
\]

holds for all \( k \geq 0 \) provided either of the following hold:
1. \( y^* \in D_2 \) and \( \text{dom}(f + g) \subseteq D_1 \);
2. \( y^* \in D_2 \) and \( \partial (f^* \square g^*)(-B^* y^k) \subseteq D_1 \).

**Proof.** We only consider the primal case because the dual case is similar. For all \( i \in \{1, \ldots, n\} \) and \( k \geq 0 \), the Fenchel-Moreau Theorem [2, Theorem 13.32], the identity \( h_i \square l_i = (h_i^* + l_i^*)^* \), and Conditions 1 and 2 show that we can reduce the domain of the following supremum:

\[
\sum_{i=1}^{n} (h_i \square l_i) (B_i x^k) = \sup_{y \in H} \left( \langle Bx^k, y \rangle - \sum_{i=1}^{n} (h_i^* (y_i) + l_i^* (y_i)) \right) \\
= \sup_{y \in D_2} \left( \langle Bx^k, y \rangle - \sum_{i=1}^{n} (h_i^* (y_i) + l_i^* (y_i)) \right).
\]

In addition, the Fenchel-Young inequality shows that

\[
\sum_{i=1}^{n} (h_i^* (y_i^k) + l_i^* (y_i^k)) - \langle x^*, B^* y^k \rangle \geq - \sum_{i=1}^{n} (h_i \square l_i) (B_i x^*),
\]

Therefore,

\[
\sup_{x \in \{x^*\} \times D_2} \mathcal{G}^{\text{pre}}(x^k, x^k, x^k; x) \\
= f(x^k) + g(x^k) - f(x^*) - g(x^*) + \sum_{i=1}^{n} (h_i^* (y_i^k) + l_i^* (y_i^k)) - \langle x^*, B^* y^k \rangle \\
+ \sup_{y \in D_2} \left( \langle Bx^k, y \rangle - \sum_{i=1}^{n} (h_i^* (y_i) + l_i^* (y_i)) \right) \\
\geq f(x^k) + g(x^k) + \sum_{i=1}^{n} (h_i \square l_i) (B_i x^k) - \left( f(x^*) + g(x^*) + \sum_{i=1}^{n} (h_i \square l_i) (B_i x^*) \right).
\]
The bounded domain conditions in Proposition 5.3 are related to the Lipschitz continuity of the objective functions. Indeed, if \( h_i \) is Lipschitz, it follows that \( \text{dom}(h_i^*) \) is bounded \([4, \text{Proposition 4.4.6}]\). In addition, \( \text{dom}(h_i^* + l_i^*) = \text{dom}(h_i^*) \cap \text{dom}(l_i^*) \). Thus, if \( h_i^* \) has bounded domain, so does \( h_i^* + l_i^* \).

The bounded subgradient conditions in Proposition 5.3 are satisfied for \( h_i \text{□} l_i \) if the infimal convolution has full domain and the sequence \( (B_i x^j)_{j \geq 0} \) is convergent. Indeed, in this case \( \partial(h_i \text{□} l_i) \) is locally bounded \([2, \text{Proposition 16.14(iii)}]\) and, hence, the union \( \bigcup_{j \geq 0} \partial(h_i \text{□} l_i)(B_i x^j) \) is bounded. See \([10, \text{Remark 2.2}]\) for similar remarks in the context of a primal-dual forward-backward-forward splitting algorithm.

### 5.2. Two algorithm classes

In this section, we study the algorithms that arise for different classes of metrics \((U_j)_{j \geq 0}\), and show how to compute the resolvent and forward-backward operators needed in order to apply the PPA, FBS, PRS, and FBF algorithms just as they appear in Section 2.

We fix the following notation for the rest of this section: Let \( \mu_{V_i} > 0 \) and let \( V_i \in S_{\mu_{V_i}}(H_i) \) for \( i = 0, \ldots, n \). Let \( \mu_{W_i} > 0 \) and let \( W_i \in S_{\mu_{W_i}}(H_i) \) for \( i = 1, \ldots, n \). These strongly monotone maps are metrics on the spaces \( H_i \) for \( i = 0, \ldots, n \). They can be as simple as "diagonal" metrics, but they can also incorporate second order information. A substantial discussion on the best metric choice is beyond the scope of this paper, so we just refer the reader to \([37]\) for some applications of fixed "diagonal" metrics, and \([25]\) for nonconstant "diagonal" metrics that satisfy conditions similar to Assumption 3.

Now define "block-diagonal" metrics

\[
V := V_1 \oplus \cdots \oplus V_n \in S_{\mu_V} \left( \prod_{i=1}^n H_i \right) \quad \text{and} \quad W := W_1 \oplus \cdots \oplus W_n \in S_{\mu_W} \left( \prod_{i=1}^n H_i \right).
\]

where \( \mu_V = \min\{\mu_{V_1}, \ldots, \mu_{V_n}\} \), and \( \mu_W = \min\{\mu_{W_1}, \ldots, \mu_{W_n}\} \). The rest of this section will build three types of metrics from \( V_0, V, W \).

Before we introduce the metrics on the space \( H \), we recall the following definitions of the proximal operator of \( f \) under the metric \( V_0 \):

\[
\text{prox}_{V_0}^f(x) := \arg\min_{y \in H} f(y) + \frac{1}{2\gamma} \langle V_0(y - x), y - x \rangle.
\]

In particular, \( \text{prox}_{V_0}^f = J_{V_0^{-1} \partial f} \), and the following Moreau identity holds:

\[
\text{prox}_{\gamma V_0^{-1}}^f = V_0^{-1/2} \text{prox}_{\gamma f \circ V_0^{-1/2} V_0^{1/2}} = I_H - \gamma V_0^{-1} \text{prox}_{V_0^{-1} \partial f} \circ (\gamma^{-1} V_0).
\]

Similar identities hold for \( h_i^* \) and \( l_i^* \). See \([19]\) for more background on proximal operators under different metrics.

Finally, note that Part 1 of Proposition 1.2 shows the following identity: for all \( z \in H \),

\[
z^+ = J_{U^{-1}(\partial f + S)}(z) \quad \iff \quad U(z - z^+) \in \partial f(z^+) + Sz^+.
\]

See Proposition 5.5, 5.7, and 5.8 for examples of resolvent computations.
5.2.1. First metric class. In this section, our metrics depend on a parameter $w$, which appears in the PRS algorithm 4. We only use the metric for the case that $w \in \{0, 1/2, 1\}$, but we prove all of our results for the general case $w \in \mathbb{R}$. The setting $w = 1/2$ first appeared in the proof of [9, Theorem 2.1], and the setting $w = 1$ first appeared in [26].

Proposition 5.4. Let $w \in \mathbb{R}$. Assume the setting of Proposition 5.1. For all $x = (x, y) \in H$, define the map:

$$U_w x := (V_0 x - wB^*y, -wBx + Vy).$$

(5.13)

Suppose that $w^2 \|V^{-1/2}B V_0^{-1/2}\|^2 < 1$. Then $U_w$ is self-adjoint and strongly monotone: for all $x \in H$,

$$\langle x, U_w x \rangle \geq \frac{1}{2} \left(1 - w^2 \|V^{-1/2}B V_0^{-1/2}\|^2\right) \min \{\mu_{V_0}, \mu_V\} \left(\|x\|^2 + \|y\|^2\right).$$

(5.14)

Assume the setting of Proposition 5.2. For all $x = (x, y, v) \in H$, define the map:

$$U'_w x := (V_0 x - wB^*y, Vy - wBx + wv, wy + Wv).$$

(5.15)

Suppose that $w^2 \|V^{-1/2}B V_0^{-1/2}\|^2 + w^2 \|W^{-1/2}V^{-1/2}\|^2 < 1$. Then

$$\langle x, U'_w x \rangle \geq \frac{1}{3} \left(1 - w^2 \|V^{-1/2}B V_0^{-1/2}\|^2 - w^2 \|W^{-1/2}V^{-1/2}\|^2\right) \times \min \{\mu_{V_0}, \mu_V, \mu_W\} \left(\|x\|^2 + \|y\|^2 + \|v\|^2\right).$$

(5.16)

Proof. Equation (5.14) is shown in [36, Lemma 4.3, Equation (4.14)] when $w = 1$. The extension to general $w$ is straightforward. Equation (5.16) has nearly the same proof.

Note that our conditions for ergodic convergence in Theorem 3.2 require the metrics to be almost decreasing up to a summable residual in the Lowner partial ordering $\succ$ (see Section 1.2). If $(U_j)_{j \geq 0}$ is a sequence of metrics defined as in Equation (5.13), we have

$$(U_k - U_{k+1})x = ((V_{0,k} - V_{0,k+1})x, (V_k - V_{k+1})y).$$

Thus, if $V_{0,k} \succ V_{0,k+1}$ and $V_k \succ V_{k+1}$, we can guarantee that the product metric is increasing (Lemma 1.1). A similar result holds for the level 2 metrics in Equation (5.15).

The following proposition shows how to evaluate the FBS operator under the metrics induced by $U_w$ and $U'_w$.

Proposition 5.5 (Forward-Backward operators under the first metric class). Let $w \in \mathbb{R}$. Assume the setting of Proposition 5.1, and suppose that $U_w \in S_p(H)$ (Equation (5.13)) for some $p > 0$. Let $z := (x, y) \in H$. Suppose that $g, l_1^*, \cdots, l_n^*$ are differentiable. Then $z^+ := J_{U_w^{-1}(\partial z + wS)}(z - U_w^{-1}\nabla g(z))$ has the following form:

$$x^+ = \text{prox}_{V_0^*}^W(x - V_0^{-1}(wB^*y + \nabla g(x)));
\text{for } i = 1, 2, \ldots, m, \text{ in parallel do }
\begin{align*}
y^*_i &= \text{prox}_{l^*_i}^W(y_i + V_i^{-1}(wB_i(2x^+ - x) - \nabla l_i^*(y_i));
\end{align*}
$$
Assume the setting of Proposition 5.2, and suppose that $U' \in S_p(H)$ (Equation (5.15)) for some $p > 0$. Let $z := (x, y, v) \in H$, and suppose that $g$ is differentiable. Then $z^+ := J_{(U_1')^{-1}(\partial f + wS)}(z - (U_w')^{-1} \nabla g(z))$ has the following form:

$$x^+ = \text{prox}_{V_0}^y(x - V_0^{-1}(wB^* y + \nabla g(x));$$
for $l = 1, 2, \ldots, m$, in parallel do

$$v_i^+ = \text{prox}^l_{V_i}(v_i + \partial \mu^1 w_i);$$

$$y_i^+ = \text{prox}^l_{V_i}(y_i + V_i^{-1}(wB_i(2x^+ - x) - (2v_i^+ - v_i));$$

\begin{proof}

$U_w(z - z^+) \in \partial f(z^+) + wSz^+ + \nabla g(z)$

$$\Leftrightarrow \begin{cases} V_0(x - x^+) - wB^*(y - y^+) \in \partial f(x^+) + wB^* y^+ + \nabla g(x); \\
V_i(y_i - y_i^+) - wB_i(x - x^+) \in \partial h^l_i(y_i^+) - wB_i x^+ + \nabla l^i_i(y_i); & i = 1, \ldots, n \\
V_0(x - x^+) \in \partial f(x^+) + wB y + \nabla g(x); \\
V_i(y_i - y_i^+) \in \partial h^l_i(y_i^+) - wB_i(2x^+ - x) + \nabla l^i_i(y_i); & i = 1, \ldots, n. 
\end{cases}$$

The last equations are exactly the optimality conditions of the proximal operators under the metrics $V_0$ and $V$.

The level 2 case is similar, so we omit it. \[ \square \]

**Remark 6.** The results of Proposition 5.5 are not new. The level 1 case with $w \in \{0, 1/2, 1\}$ has appeared implicitly in several papers, including [21, 41, 19]. It has also explicitly appeared in [36, Lemma 4.5]. In addition, the proof of the level 2 case has appeared in [9, Equation (2.38)].

**5.3. Second metric class.** The following computation appeared implicitly in [36, Equation 4.10] for $w = 1$. We reproduce the result for completeness.

**Proposition 5.6.** Assume the setting of Proposition 5.1. For all $x \in H$, define the map:

$$U_w x := (V_0 x, (V - w^{2} B V_0^{-1} B^*) y).$$

Suppose that $w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 < 1$. Then $U$ is self adjoint and strongly monotone: for all $x \in H$,

$$\langle x, U x \rangle \geq \min \left\{ \mu_{V_0}, \left(1 - w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 \right) \mu_V \right\} (\| x \|^2 + \| y \|^2).$$

\begin{proof}

Set $C_w = wB$. For all $y \in \prod_{i=1}^n H_i$, we have

$$\langle y, (V - CV_0^{-1} C^*) y \rangle = \langle V^{1/2} y, \left(I_{\prod_{i=1}^n H_i} - V^{-1/2} C V_0^{-1} C^* V^{-1/2} \right) V^{1/2} y \rangle$$

$$= \langle Vy, y \rangle - \langle V^{1/2} y, V^{-1/2} C V_0^{-1} C^* V^{-1/2} V^{1/2} y \rangle$$

$$\geq \left(1 - \| V^{-1/2} C V_0^{-1} C V^{-1/2} \| \right) \langle Vy, y \rangle$$

$$\geq \left(1 - w^2 \| V^{-1/2} B V_0^{-1/2} \|^2 \right) \mu_V \| y \|^2.$$
Therefore,
\[
\langle x, U_w x \rangle \geq \mu V_0 \|x\|^2 + \left(1 - w^2 \|V^{-1/2}B V_0^{-1/2}\|^2\right) \mu V \|y\|
\geq \min \left\{ \mu V_0 \left(1 - w^2 \|V^{-1/2}B V_0^{-1/2}\|^2\right) \mu V \right\} \left(\|x\|^2 + \|y\|^2\right).
\]

For simplicity, and because it has not yet found an application, we do not discuss the generalization of the Equation (5.17) to the level 2 case.

Note that our conditions for ergodic convergence in Theorem 3.2 require the metrics to be almost decreasing, up to a summable residual, in the Loewner partial ordering \( \succeq \) (see Section 1.2). If \( (U_j)_{j \geq 0} \) is a sequence of metrics defined as in Equation (5.17), we have
\[
(U_k - U_{k+1}) x = \left((V_{0,k} - V_{0,k+1}) x, (V_k - V_{k+1}) + w^2B(V_{0,k+1} - V_{0,k})B^*\right) y.
\]
Thus, if \( V_{0,k} \succeq V_{0,k+1} \) and \( V_k \succeq V_{k+1} \), the product metric is decreasing (Lemma 1.1).

The following proposition shows how to evaluate the FBS operator under the metric induced by \( U \).

**Proposition 5.7** (Forward-Backward operators under the second metric class). Assume the setting of Proposition 5.1. Suppose that \( f \equiv 0 \), and that \( U \in S_\rho(H) \) (Equation (5.17)) for some \( \rho > 0 \). Let \( x := (x, y) \in H \). Suppose that \( g, l^*_1, \cdots, l^*_n \) are differentiable. Then \( \mathbf{z}^* := J_{U^{-1}(\partial f + wS)}(z - U^{-1}\nabla g(z)) \) has the following form:

1. For \( l \in \{1, 2, \cdots, n\} \), in parallel do
   \[
   \begin{align*}
   y_i^+ &= \prox_{l^*_i}^V(y_i + V^{-1}(wB_i (x - V^{-1}_0(\nabla g(x) + wB^*y) - \nabla l^*_i(y_i)))); \\
   x^+ &= x - V^{-1}_0(\nabla g(x) + wB^*y^+);
   \end{align*}
   \]

2. Proof. The result follows from an application of Equation (5.12):

\[
\mathbf{z}^* = J_{U^{-1}(\partial f + wS)}(z - U^{-1}\nabla g(z))
\]

3. For example,
\[
\begin{align*}
V_0(x - x^+) & \in \partial f(z^+) + wS\mathbf{z}^+ + \nabla g(z) \\
V_i(y_i - y_i^+) & - w^2B_i V^{-1}_0B^* (y - y^+) \in \partial h^*_i(y_i^+) - wB_i x^+ + \nabla l^*_i(y_i); \quad i = 1, \cdots, n \\
x^+ &= x - V^{-1}_0(\nabla g(x) + wB^*y^+);
\end{align*}
\]

4. For example,
\[
\begin{align*}
V_i(y_i - y_i^+) & \in \partial h^*_i(y_i^+) - wB_i x^+ + (wB_i) V^{-1}_0(wB^*) (y - y^+) + \nabla l^*_i(y_i); \quad i = 1, \cdots, n \\
x^+ &= x - V^{-1}_0(\nabla g(x) + wB^*y^+);
\end{align*}
\]

where the last equation follows because \((wB_i) V^{-1}_0(wB^*) = wB_i(x - x^+ - V^{-1}_0(\nabla g(x)))\).

5. The last equations are exactly the optimality conditions of the proximal operators under the metric \( V \).

**Remark 7.** Proposition 5.7 appears in [36, Lemma 4.10] for \( w = 1 \).

Now consider the special case \( w = 0 \). In this case, the first and second metric classes are identical. Note that we only apply this metric class to the FBS algorithm and the PRS algorithm.

**Proposition 5.8** (Resolvents of skew operators). Assume the setting of Proposition 5.1. Suppose that \( U \in S_\rho(H) \) (Equation (5.17)) for some \( \rho > 0 \). Let \( x :=
(x, y) ∈ H. Then $z^+ := J_{\gamma U - 1_g} (z)$ has the following form:

$$ (x^+, y^+) = ((I_{\mathcal{H}_0} + \gamma^2 V_0 B^* V B)^{-1} (x - \gamma V_0 B^*) y, (I_{\prod_{i=1}^n H_i} + \gamma^2 V B V_0 B^*)^{-1} (y + \gamma V B x)). $$

Proof. The case when $U = I_H$ appears in [11, Proposition 2.7]. This generalization is straightforward, so we omit the proof.

Generalizing the resolvent operator computation in Proposition 5.8 to the level 2 case is straightforward, though slightly messy. It has not found application in the literature yet, so we omit the statement.

### 5.4. New and old convergence rates.

Table 5.4 lists the application of PPA, FBS, PRS, and FBF algorithms under the metrics introduced in Section 5.2 and indicates which convergence rates have been shown in the literature. We note that, to the best of our knowledge, for all of the methods we discuss, the nonergodic fixed metric convergence rates, the ergodic convergence rates under variable metrics, stepsizes, and relaxation, and the non-ergodic/ergodic convergence rates with nonconstant relaxation have never appeared in the literature.

Any pairing between metrics, algorithms, and splittings that does not appear in Table 5.4 is an algorithm where, to the best of our knowledge, no convergence rate has appeared in the literature.

### 6. Conclusion.

In this paper, we provided a convergence rate analysis of a general monotone inclusion problem under the application of four different algorithms. We provided ergodic convergence rates under variable metrics, stepsizes, and relaxation, and recovered several known rates in the process. In addition, for three of the algorithms we provided the first nonergodic convergence rates that have appeared in the literature. Finally, we showed how our results imply convergence rates of a large class of primal-dual splitting algorithms. The techniques developed in this paper are not limited to the four algorithms we chose to study, and the proofs of this paper can be used as a template for proving convergence rates of other special cases of the unifying scheme.

### REFERENCES
Convergence rate analysis of primal-dual splitting schemes

[1] Jean-Bernard Baillon and Georges Haddad, Quelques propriétés des opérateurs angle-bornés et cycliquement monotones, Israel Journal of Mathematics, 26 (1977), pp. 137–150.

[2] Heinz H Bauschke and Patrick L Combettes, Convex analysis and monotone operator theory in Hilbert spaces, Springer, 2011.

[3] Stephen Becker and Patrick L Combettes, An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery, arXiv preprint arXiv:1305.5828, (2013).

[4] Jonathan M Borwein and Jon D Vanderwerff, Convex functions: constructions, characterizations and countourexamples, vol. 109, Cambridge University Press, 2010.

[5] Radu Ioan Boţ and Ernö Robert Csetnek, On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems, (2014).

[6] Radu Ioan Boţ, Ernö Robert Csetnek, André Heinrich, and Christopher Hendrich, On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems, Mathematical Programming, pp. 1–29.

[7] Radu Ioan Boţ, Ernö Robert Csetnek, and Erika Nagy, Solving systems of monotone inclusions via primal-dual splitting techniques, Taiwanese Journal of Mathematics, 17 (2013), pp. pp–1983.

[8] Radu Ioan Boţ and Christopher Hendrich, An algorithm for solving monotone inclusions involving parallel sums of linearly composed maximally monotone operators, arXiv preprint arXiv:1306.3191, (2013).

[9] A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators, SIAM Journal on Optimization, 23 (2013), pp. 2541–2565.

[10] Convergence analysis for a primal-dual monotone+ skew splitting algorithm with applications to total variation minimization, Journal of Mathematical Imaging and Vision, 49 (2014), pp. 551–568.

[11] Luis M Briceno-Arias and Patrick L Combettes, A monotone+ skew splitting model for composite monotone inclusions in duality, SIAM Journal on Optimization, 21 (2011), pp. 1230–1250.

[12] Ronald E Bruck Jr., On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space, Journal of Mathematical Analysis and Applications, 61 (1977), pp. 159–164.

[13] Antonin Chambolle and Pierre-Louis Lions, Image recovery via total variation minimization and related problems, Numerische Mathematik, 76 (1997), pp. 167–188.

[14] Antonin Chambolle and Thomas Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145.

[15] Peijun Chen, Jianguo Huang, and Xiaoqun Zhang, A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration, Inverse Problems, 29 (2013), p. 025011.

[16] Patrick L Combettes, Systems of structured monotone inclusions: duality, algorithms, and applications, SIAM Journal on Optimization, 23 (2013), pp. 2420–2447.

[17] Patrick L Combettes, Laurent Condat, Jean-Christophe Pesquet, and Bằng Công Vũ, A forward-backward view of some primal-dual optimization methods in image recovery, arXiv preprint arXiv:1406.5439, (2014).

[18] Patrick L Combettes and Jean-Christophe Pesquet, Primal-dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators, Set-Valued and variational analysis, 20 (2012), pp. 307–330.

[19] Patrick L Combettes and Bằng Công Vũ, Variable metric forward–backward splitting with applications to monotone inclusions in duality, Optimization, (2012), pp. 1–30.

[20] Patrick L Combettes and Bằng Công Vũ, Variable metric quasi-fejér monotonicity, Nonlinear Analysis: Theory, Methods & Applications, 78 (2013), pp. 17–31.

[21] Laurent Condat, A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms, Journal of Optimization Theory and Applications, 158 (2013), pp. 460–479.

[22] Dámek Davis and Wotao Yin, Convergence rate analysis of several splitting schemes, arXiv preprint arXiv:1406.4834, (2014).

[23] Ernie Esser, Xiaoqun Zhang, and Tony F Chan, A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046.

[24] R Glowinski and A Marrocco, Sur l’approximation, par éléments finis d’ordre un, et la
résolution, par pénalisation-dualité d’une classe de problèmes de Dirichlet nonlinéaires, Rev. Française d’Aut. Inf. Rech. Oper, R-2 (1975), pp. 41–76.

[25] Tom Goldstein, Ernie Esser, and Richard Baraniuk, Adaptive primal-dual hybrid gradient methods for saddle-point problems, arXiv preprint arXiv:1305.0546, (2013).

[26] Bingsheng He and Xiaoming Yuan, Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective, SIAM Journal on Imaging Sciences, 5 (2012), pp. 119–149.

[27] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach, Proximal methods for hierarchical sparse coding, The Journal of Machine Learning Research, 12 (2011), pp. 2297–2334.

[28] Tosio Kato, Perturbation theory for linear operators, vol. 132, Springer, 1995.

[29] Nikos Komodakis and Jean-Christophe Pesquet, Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems, arXiv preprint arXiv:1406.5429, (2014).

[30] Steven M LaValle, Planning algorithms, Cambridge University Press, 2006.

[31] Pierre-Louis Lions and Bertrand Mercier, Splitting algorithms for the sum of two nonlinear operators, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.

[32] Yuri Nesterov, Smooth minimization of non-smooth functions, Mathematical programming, 103 (2005), pp. 127–152.

[33] Daniel O’Connor and Lieven Vandenberghe, Primal-dual decomposition by operator splitting and applications to image deblurring, (2014).

[34] Lisandro A Parente, Pablo A Lotito, and Mikhail V Solodov, A class of inexact variable metric proximal point algorithms, SIAM Journal on Optimization, 19 (2008), pp. 240–260.

[35] Gregory B Passty, Ergodic convergence to a zero of the sum of monotone operators in hilbert space, Journal of Mathematical Analysis and Applications, 72 (1979), pp. 383–390.

[36] Jean-Christophe Pesquet and Audrey Repetti, A class of randomized primal-dual algorithms for distributed optimization, arXiv preprint arXiv:1406.6404, (2014).

[37] Thomas Pock and Antonin Chambolle, Diagonal preconditioning for first order primal-dual algorithms in convex optimization, in Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 1762–1769.

[38] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle, An algorithm for minimizing the mumford-shah functional, in Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1133–1140.

[39] R Tyrrell Rockafellar, Monotone operators and the proximal point algorithm, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[40] Paul Tseng, A modified forward-backward splitting method for maximal monotone mappings, SIAM Journal on Control and Optimization, 38 (2000), pp. 431–446.

[41] Băng Công Vũ, A splitting algorithm for dual monotone inclusions involving cocoercive operators, Advances in Computational Mathematics, 38 (2013), pp. 667–681.