Primal-Dual Proximal Algorithms for Structured Convex Optimization: A Unifying Framework

Puya Latafat and Panagiotis Patrinos

Abstract We present a simple primal-dual framework for solving structured convex optimization problems involving the sum of a Lipschitz-differentiable function and two nonsmooth proximable functions, one of which is composed with a linear mapping. The framework is based on the recently proposed asymmetric forward-backward-adjoint three-term splitting (AFBA); depending on the value of two parameters, (extensions of) known algorithms as well as many new primal-dual schemes are obtained. This allows for a unified analysis that, among other things, establishes linear convergence under four different regularity assumptions for the cost functions. Most notably, linear convergence is established for the class of problems with piecewise linear-quadratic cost functions.

1 Introduction

In this chapter we consider structured convex optimization problems of the form

$$\underset{x\in\mathbf{R}^n}{\textbf{minimize}}\ f(x)+g(x)+h(Lx), \tag{1}$$

where L is a linear mapping, g and h are proper closed convex (possibly) nonsmooth functions, and f is convex, continuously differentiable with Lipschitz continuous gradient. The working assumption throughout the chapter is that one can efficiently

Puya Latafat[1,2] ✉ puya.latafat@{kuleuven.be,imtlucca.it}
Panagiotis Patrinos[1] ✉ panos.patrinos@esat.kuleuven.be
This work was supported by FWO PhD grant 1196818N; FWO research projects G086518N and G086318N: Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160: SeLMA.
[1]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium.
[2]IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100 Lucca, Italy.

evaluate the gradient of f, the proximal mapping of the nonsmooth terms g and h, the linear mapping L and its adjoint.

This model is quite rich and captures a plethora of problems arising in machine learning, signal processing and control [13, 31, 22]. As a widely popular example consider consensus or sharing type problems:

$$\underset{x_1,\ldots,x_N}{\textbf{minimize}} \sum_{i=1}^{N} f_i(x_i) \tag{2a}$$

$$\textbf{subject to } x \in C, \tag{2b}$$

where $x = (x_1,\ldots,x_N)$. In the case of consensus $C = \{(x_1,\ldots,x_N) \mid x_1 = \ldots = x_N\}$, and for sharing $C = \{(x_1,\ldots,x_N) \mid \sum_{i=1}^{N} x_i = 0\}$. The two sets are orthogonal subspaces and indeed the sharing and consensus problems are dual to each other [7]. More generally, when C is a subspace problem (2) is referred to as *extended monotropic programming* [5]. This family of problems can be written in the form of (1) with $f(x) = \sum_{i=1}^{N} f_i(x_i)$, $h \equiv 0$, and $g = \iota_C$, where ι_X denotes the indicator of the set X.

A recent trend for solving problem (1), possibly with the smooth term $f \equiv 0$ or the nonsmooth term $g \equiv 0$, is to solve the monotone inclusion defined by the primal-dual optimality conditions [10, 20, 8, 14, 15, 32, 18, 24, 9]. The popularity of this approach is mainly due to the fact that it results in *fully split* primal-dual algorithms, in which the proximal mappings of g and h, the gradient of f, the linear mapping L and its adjoint are evaluated individually. In particular, there are no matrix inversions or inner loops involved.

Different convergence analysis techniques have been proposed in the literature for primal-dual algorithms. Some can be viewed as intelligent applications of classical splitting methods such as forward-backward splitting (FBS), Douglas-Rachford splitting (DRS) and forward-backward-forward splitting (FBFS). See for example [32, 15, 8, 6, 14], while others employ different tools to show convergence [10, 21, 18, 11]. Convergence rate of primal-dual schemes has also been analyzed using different approaches, see for example [26, 16, 27, 11].

Our approach here is a systematic one and relies on solving the primal-dual optimality conditions using a new three-term splitting, *asymmetric forward-backward-adjoint splitting* (AFBA) [24]. This splitting algorithm generalizes FBS to include a third linear monotone operator. Furthermore, it includes asymmetric preconditioning that is the key for developing our unifying primal-dual framework in which one can generate a wide range of algorithms by selecting different values for the scalar parameters θ and μ (*cf.* Algorithm 1). Many of the resulting algorithms are new, while some extend previously proposed algorithms and/or result in less conservative stepsize conditions. In short, this analysis provides us with a spectrum of primal-dual algorithms each of which may have an advantage over others depending on the application. For example, in [23] a special case was exploited to develop a randomized block coordinate variant for distributed applications where the stepsizes only depend on local information. The main idea was to exploit the fact that,

for this particular primal-dual algorithm, the generated sequence is Fejér monotone with respect to $\|\cdot\|_S$, where S is a block diagonal positive definite matrix.

The convergence analysis of all the primal-dual algorithms is easily deduced from that of AFBA. It must be noted that this work complements the analysis in [24] where AFBA was introduced. The relationship between existing primal-dual algorithms is already documented in [24, Fig. 1]. In this work we simplify AFBA by considering a constant step in place of a dynamic one. This modification simplifies the analysis of the primal-dual framework. In addition, we provide a general and easy-to-check convergence condition for the stepsizes in Assumption II. Furthermore, we discuss four mild regularity assumptions on the functions involved in (1) that are sufficient for metric subregularity of the operator defining the primal-dual optimality conditions (*cf.* Lemmas 4.3 and 4.5). Linear convergence rate is then deduced based on the results developed for AFBA (*cf.* Theorem 3.6). These results do not impose additional restrictions on the stepsizes of the algorithms. It is important to note that the provided conditions are much weaker than strong convexity and in many cases do not imply a unique primal or dual solution.

It is worth mentioning that another class of primal-dual algorithms was introduced recently that rely on iterative projections onto half-spaces containing the set of solutions [1, 12]. This class of algorithms is not covered by the analysis in this work.

The paper is organized as follows. In Section 2 we present the new primal-dual framework and discuss several notable special cases. Section 3 is devoted to the introduction and analysis of a simplified variant of AFBA. In Section 4 we establish convergence for the proposed primal-dual framework based on the results developed in Section 3. In particular, linear convergence is established under four mild regularity assumptions for the cost functions.

Notation and Background

Throughout, $\mathbb{R}^n$ is the n-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|\cdot\|$. The sets of symmetric, symmetric positive semi-definite and symmetric positive definite n by n matrices are denoted by $\mathcal{S}^n$, $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$ respectively. We also write $P \succeq 0$ and $P \succ 0$ for $P \in \mathcal{S}^n_+$ and $P \in \mathcal{S}^n_{++}$ respectively. For $P \in \mathcal{S}^n_{++}$ we define the scalar product $\langle x,y\rangle_P = \langle x,Py\rangle$ and the induced norm $\|x\|_P = \sqrt{\langle x,x\rangle_P}$. For simplicity we use matrix notation for linear mappings when no ambiguity occurs.

An operator (or set-valued mapping) $A : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ maps each point $x \in \mathbb{R}^n$ to a subset Ax of $\mathbb{R}^d$. The graph of A is denoted by $\mathrm{gra}A = \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^d \mid y \in Ax\}$ and the set of its zeros by $\mathbf{zer}\,A = \{x \in \mathbb{R}^n \mid 0 \in Ax\}$. The mapping A is called monotone if $\langle x-x', y-y'\rangle \geq 0$ for all $(x,y),(x',y') \in \mathrm{gra}A$, and is said to be maximally monotone if its graph is not strictly contained in the graph of another monotone operator. The inverse of A is defined through its graph: $\mathrm{gra}A^{-1} := \{(y,x) \mid (x,y) \in$

$\mathrm{gra}A\}$. The *resolvent* of A is defined by $J_A := (\mathrm{Id}+A)^{-1}$, where Id denotes the identity operator.

For an extended-real-valued function f, we use $\mathbf{dom}\, f$ to denote its domain. Let $f:\mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$ be a proper closed, convex function. Its subdifferential is the set-valued operator $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$

$$\partial f(x) = \{y \in \mathbb{R}^n \mid \forall z \in \mathbb{R}^n,\, \langle z-x, y\rangle + f(x) \le f(z)\}.$$

The subdifferential is a maximally monotone operator. The resolvent of ∂f is called the *proximal operator* (or proximal mapping), and is single-valued. For a given $V \in \mathcal{S}^n_{++}$ the proximal mapping of f relative to $\|\cdot\|_V$ is uniquely determined by the resolvent of $V^{-1}\partial f$:

$$\mathbf{prox}^V_f(x) := (\mathrm{Id}+V^{-1}\partial f)^{-1}x = \underset{z\in\mathbf{R}^n}{\mathbf{argmin}}\{f(z)+\tfrac{1}{2}\|x-z\|^2_V\}.$$

The *Fenchel conjugate* of f, denoted f^*, is defined as

$$f^*(v) := \sup_{x\in\mathbf{R}^n}\{\langle v,x\rangle - f(x)\}.$$

The *infimal convolution* of two functions $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined as

$$(f\,\square\, g)(x) = \inf_{z\in\mathbf{R}^n}\{f(z)+g(x-z)\}.$$

Let X be a nonempty closed convex set and define the indicator of the set

$$\iota_X(x) := \begin{cases} 0 & x \in X \\ +\infty & x \notin X. \end{cases}$$

The distance to a set X with respect to $\|\cdot\|_V$ is given by $d_V(\cdot, X) = \iota_X \,\square\, \|\cdot\|_V$. We use $\Pi^V_X(\cdot)$ to denote the projection onto X with respect to $\|\cdot\|_V$.

The sequence $(x^k)_{k\in\mathbf{N}}$ is said to converge to $x^\star$ Q-linearly with Q-factor $\sigma \in (0,1)$, if there exists $\bar{k} \in \mathbb{N}$ such that for all $k \ge \bar{k}$, $\|x^{k+1}-x^\star\| \le \sigma\|x^k - x^\star\|$ holds. Furthermore, $(x^k)_{k\in\mathbf{N}}$ is said to converge to $x^\star$ R-linearly if there is a sequence of nonnegative scalars $(v_k)_{k\in\mathbf{N}}$ such that $\|x^k - x^\star\| \le v^k$ and $(v_k)_{k\in\mathbf{N}}$ converges to zero Q-linearly.

2 A Simple Framework for Primal-Dual Algorithms

In this section we present a simple framework for primal-dual algorithms. For this purpose we consider the following extension of (1)

$$\underset{x\in\mathbf{R}^n}{\mathbf{minimize}}\ f(x)+g(x)+(h\,\square\, l)(Lx), \tag{3}$$

where l is a strongly convex function. Notice that when $l = \iota_{\{0\}}$, the infimal convolution $h \,\square\, l$ reduces to h, and problem (1) is recovered.

Throughout this chapter the following assumptions hold for problem (3).

Assumption I.

(i) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $h : \mathbb{R}^r \to \overline{\mathbb{R}}$ *are proper closed convex functions, and* $L : \mathbb{R}^n \to \mathbb{R}^r$ *is a linear mapping.*

(ii) $f : \mathbb{R}^n \to \mathbb{R}$ *is convex, continuously differentiable, and for some* $\beta_f \in [0, \infty)$, ∇f *is* β_f*-Lipschitz continuous with respect to the metric induced by* $Q \succ 0$ *, i.e., for all* $x, y \in \mathbb{R}^n$*:*

$$\|\nabla f(x) - \nabla f(y)\|_{Q^{-1}} \leq \beta_f \|x - y\|_Q,$$

(iii) $l : \mathbb{R}^r \to \overline{\mathbb{R}}$ *is proper closed convex, its conjugate* l^* *is continuously differentiable, and for some* $\beta_l \in [0, \infty)$, ∇l^* *is* β_l*-Lipschitz continuous with respect to the metric induced by* $R \succ 0$*.*

(iv) The set of solutions to (3)*, denoted by* $X^\star$*, is nonempty.*

(v) (Constraint qualification) There exists $x \in \operatorname{ri} \mathbf{dom}\, g$ *such that* $Lx \in \operatorname{ri} \mathbf{dom}\, h + \operatorname{ri} \mathbf{dom}\, l$*.*

In Assumption I*(ii)* the constant β_f is not absorbed into the metric Q in order to be able to treat the case when ∇f is a constant in a uniform fashion by setting $\beta_f = 0$. The same reasoning applies to Assumption I*(iii)*.

The dual problem is given by

$$\underset{u \in \mathbb{R}^r}{\textbf{minimize}} \; (g^* \,\square\, f^*)(-L^\top u) + h^*(u) + l^*(u). \tag{4}$$

Notice the similar structure of the dual problem in which l, f and h, g have swapped roles. A well-established approach for solving (3) is to consider the associated convex-concave saddle point problem given by

$$\underset{x \in \mathbb{R}^n}{\textbf{minimize}} \; \underset{u \in \mathbb{R}^r}{\textbf{maximize}} \; \mathcal{L}(x, u) \coloneqq f(x) + g(x) + \langle Lx, u \rangle - h^*(u) - l^*(u). \tag{5}$$

The primal-dual optimality conditions are

$$\begin{cases} 0 \in \partial g(x) + \nabla f(x) + L^\top u, \\ 0 \in \partial h^*(u) + \nabla l^*(u) - Lx. \end{cases} \tag{6}$$

Under the constraint qualification condition, the set of solutions for the dual problem denoted by $U^\star$ is nonempty, a saddle point exists, and the duality gap is zero. In fact for any $x^\star \in X^\star$ and $u^\star \in U^\star$, the point $(x^\star, u^\star)$ is a primal-dual solution. See [28, Cor. 31.2.1] and [4, Thm. 19.1].

The right-hand side of the optimality conditions in (6) can be split as the sum of three operators:

$$\begin{pmatrix}0\\0\end{pmatrix} \in \underbrace{\begin{pmatrix}\partial g(x)\\ \partial h^*(u)\end{pmatrix}}_{Az} + \underbrace{\begin{pmatrix}\nabla f(x)\\ \nabla l^*(u)\end{pmatrix}}_{Cz} + \underbrace{\begin{pmatrix}0 & L^\top\\ -L & 0\end{pmatrix}\begin{pmatrix}x\\u\end{pmatrix}}_{M\quad z}. \tag{7}$$

Operator A defined above, is maximally monotone [4, Thm. 21.2, Prop. 20.33], while operator C, being the gradient of $\tilde{f}(x,u) = f(x) + l^*(u)$, is cocoercive and M is skew-symmetric and as such monotone.

Throughout this section we shall use T to denote the operator above, *i.e.*,

$$0 \in Tz \coloneqq Az + Cz + Mz. \tag{8}$$

Algorithm 1 describes the proposed primal-dual framework for solving (3). This framework is the result of solving the monotone inclusion (7) using the three-term splitting AFBA described in Section 3. We postpone the derivation and convergence analysis of Algorithm 1 to Section 4.

Notice that Algorithm 1 is not symmetric with respect to the primal and dual variables. Another variant may be obtained by switching their role. This would be equivalent to applying the algorithm to the dual problem (4).

The proposed framework involves two scalar parameters $\theta \in [0,\infty)$ and $\mu \in [0,1]$. Different primal-dual algorithms correspond to different values for these parameters. The iterates in Algorithm 1 consist of two proximal updates followed by two correction steps that may or may not be performed depending on the parameters μ and θ. Below we discuss some values for these parameters that are most interesting.

Algorithm 1 A simple framework for primal-dual algorithms

Require: $x^0 \in \mathbf{R}^n$, $u^0 \in \mathbf{R}^r$, the algorithm parameters $\mu \in [0,1]$, $\theta \in [0,\infty)$.
Initialize: Σ, Γ and λ based on Assumptions II*(ii)* and II*(iii)*.
for $k = 0, 1, \ldots$ **do**
$\bar{x}^k = \mathbf{prox}_g^{\Gamma^{-1}}(x^k - \Gamma L^\top u^k - \Gamma \nabla f(x^k))$
$\bar{u}^k = \mathbf{prox}_{h^*}^{\Sigma^{-1}}(u^k + \Sigma L((1-\theta)x^k + \theta \bar{x}^k) - \Sigma \nabla l^*(u^k))$
$\tilde{x}^k = \bar{x}^k - x^k$, $\tilde{u}^k = \bar{u}^k - u^k$
$x^{k+1} = x^k + \lambda(\bar{x}^k - \mu(2-\theta)\Gamma L^\top \tilde{u}^k)$
$u^{k+1} = u^k + \lambda(\bar{u}^k + (1-\mu)(2-\theta)\Sigma L \tilde{x}^k)$

A variant of Algorithm 1 was introduced in [24, Alg. 3] that includes a dynamic stepsize, α_n, in the correction steps. In that work the connection between existing primal-dual algorithms was investigated by enforcing the dynamic stepsize $\alpha_n \equiv 1$. This approach results in cumbersome algebraic steps. By removing the dynamic stepsize we simplify the analysis substantially and provide a simple condition for convergence in Assumption II. Furthermore, in many distributed applications a dynamic stepsize is disadvantageous since it would entail global coordination. Moreover, in comparison to [24, Alg. 3] we generalize the algorithm by employing the matrices Σ and Γ as stepsizes in place of scalars and consider the Lipschitz continuity of ∇f and ∇l^* with respect to $\|\cdot\|_Q$ and $\|\cdot\|_R$.

In Section 4 we show that the sequence $(x^k, u^k)_{k\in\mathbf{N}}$ generated by Algorithm 1 converges to a primal-dual solution if Assumption II holds. Moreover, linear convergence rates are established if either one of four mild regularity assumptions holds for functions f, g, l and h (*cf.* Corollary 4.6).

Assumption II (Convergence condition)**.**

(i) (Algorithm parameters) $\theta \in [0,\infty)$, $\mu \in [0,1]$.

(ii) (Stepsizes) $\Gamma \in \mathcal{S}^n_{++}$, $\Sigma \in \mathcal{S}^r_{++}$ *and (relaxation parameter)* $\lambda \in (0,2)$

(iii) The following condition holds

$$\begin{pmatrix} (\frac{2}{\lambda}-1)\Gamma^{-1} - (1-\mu)(1-\theta)(2-\theta)L^\top\Sigma L - \frac{\beta_f}{2\lambda}Q & (\mu-(1-\mu)(1-\theta)-\frac{\theta}{\lambda})L^\top \\ (\mu-(1-\mu)(1-\theta)-\frac{\theta}{\lambda})L & (\frac{2}{\lambda}-1)\Sigma^{-1} - \mu(2-\theta)L\Gamma L^\top - \frac{\beta_l}{2\lambda}R \end{pmatrix} \succ 0. \tag{9}$$

In Algorithm 1 the linear mappings L and $L^\top$ must be evaluated twice at every iteration. In the special cases when $\mu = 0, 1$ they may be evaluated only once per iteration by keeping track of the value computed in the previous iteration. As is evident from (9) the cases where both the algorithm and the convergence condition simplify are combinations of $\mu = 0, 1$, $\theta = 0, 1, 2$. Here we briefly discuss some of these special cases to demonstrate how (9) leads to simple conditions that are often less conservative than the conditions found in the literature. We have dubbed the algorithms based on whether the two proximal updates can be evaluated in parallel and if a primal or a dual correction step is performed.

The first algorithm is the result of setting $\theta = 2$ (regardless of μ) and leads to the algorithm of Condat and Vũ [15, 32] which itself is a generalization of the Chambolle-Pock algorithm [10]. With this choice of θ, the two proximal updates are performed sequentially while no correction step is required. In the box below a general condition is given for its convergence.

SNCA (Sequential No Corrector Algorithm): $\theta = 2$

Substituting $\theta = 2$ in (9) and dividing by $\frac{2}{\lambda} - 1$ yields

$$\begin{pmatrix} \Gamma^{-1} - \frac{\beta_f}{2(2-\lambda)}Q & -L^\top \\ -L & \Sigma^{-1} - \frac{\beta_l}{2(2-\lambda)}R \end{pmatrix} \succ 0.$$

If $l = \iota_{\{0\}}$, the infimal convolution $h \,\square\, l = h$, $l^* \equiv 0$ and $\beta_l = 0$. Using Schur complement yields the equivalent condition $\Gamma^{-1} - \frac{\beta_f}{2(2-\lambda)}Q - L^\top\Sigma L \succ 0$. If in addition $Q = \mathrm{Id}$ and $\Gamma = \gamma\mathrm{Id}$, $\Sigma = \sigma\mathrm{Id}$ for some scalars γ, σ, then the following sufficient condition may be used

$$\sigma\gamma\|L\|^2 < 1 - \frac{\gamma\beta_f}{2(2-\lambda)}, \quad \lambda \in (0,2).$$

Notice that this condition is less conservative than the condition of [15, Thm. 3.1] (see [24, Rem. 5.6]).

In the next algorithm proximal updates are evaluated sequentially, followed by a correction step for the primal variable, hence the name SPCA. Most notable property of this algorithm is that the generated sequence is Fejér monotone with respect to $\|\cdot\|_S$ where S is block diagonal. The algorithm introduced in [23] can be seen as an application of SPCA to the dual problem when the smooth term is zero.

SPCA (Sequential Primal Corrector Algorithm): $\theta = 1$, $\mu = 1$, $\lambda = 1$

In this case the left-hand side in (9) is block diagonal. Therefore the convergence condition simplifies to

$$\Gamma^{-1} - \tfrac{\beta_f}{2} Q \succ 0, \quad \Sigma^{-1} - L\Gamma L^\top - \tfrac{\beta_l}{2} R \succ 0.$$

If $Q = \mathrm{Id}$, $R = \mathrm{Id}$, $\Gamma = \gamma \mathrm{Id}$, $\Sigma = \sigma \mathrm{Id}$ for some scalars γ, σ, then it is sufficient to have

$$\gamma\beta_f < 2, \quad \sigma\gamma\|L\|^2 < 1 - \tfrac{\sigma\beta_l}{2}. \tag{10}$$

This special case generalizes the recent algorithm proposed in [18]. In particular we allow a third nonsmooth function g as well as the strongly convex function l. In addition to this improvement our convergence condition with $l^* \equiv 0$ (set $\beta_l = 0$ in (10)) is less restrictive and doubles the range of acceptable stepsize γ. The convergence condition in that work is given in our notation as $\gamma\beta_f < 1$ and $\sigma\gamma\|L\|^2 < 1$ [18, Cor. 3.2].

Next algorithm features sequential proximal updates that are followed by a correction step for the dual variable, for all values of $\theta \in (0,\infty)$. The parallel variant of this algorithm, referred to as PDCA, is discussed in a separate box below. We have observed that selecting θ so as to maximize the stepsizes, *i.e.*, $\theta = 1.5$, leads to faster convergence [25]. Moreover, if we set $\Sigma = \Gamma^{-1}$, this choice of μ leads to a three-block ADMM or equivalently a generalization of DRS to include a third cocoercive operator (see [24, Sec. 5.4]).

SDCA (Sequential Dual Corrector Algorithm): $\theta \in (0,\infty)$, $\mu = 0$, $\lambda = 1$

In this case the convergence condition simplifies to

$$\begin{pmatrix} \Gamma^{-1} - (1-\theta)(2-\theta)L^\top \Sigma L - \tfrac{\beta_f}{2} Q & -L^\top \\ -L & \Sigma^{-1} - \tfrac{\beta_l}{2} R \end{pmatrix} \succ 0.$$

If $l = \iota_{\{0\}}$ then $\beta_l = 0$ and using Schur complement we derive the following condition

$$\Gamma^{-1} - \big(\theta^2 - 3\theta + 3\big) L^\top \Sigma L - \tfrac{\beta_f}{2} Q \succ 0.$$

If in addition $Q = \mathrm{Id}$, and $\Gamma = \gamma \mathrm{Id}$, $\Sigma = \sigma \mathrm{Id}$ for some scalars γ, σ, then we have the following sufficient condition

$$(\theta^2 - 3\theta + 3)\sigma\gamma\|L\|^2 < 1 - \tfrac{\gamma\beta_f}{2}.$$

The next algorithm appears to be new and involves parallel proximal updates followed by a primal correction step.

PPCA (Parallel Primal Corrector Algorithm): $\theta = 0$, $\mu = 1$

The convergence condition is given by

$$\begin{pmatrix} (\frac{2}{\lambda} - 1)\Gamma^{-1} - \frac{\beta_f}{2\lambda} Q & L^\top \\ L & (\frac{2}{\lambda} - 1)\Sigma^{-1} - 2L\Gamma L^\top - \frac{\beta_l}{2\lambda} R \end{pmatrix} \succ 0.$$

If $f \equiv 0$ (set $\beta_f = 0$) and $\lambda = 1$, using Schur complement yields the following condition

$$\Sigma^{-1} - 3L\Gamma L^\top - \frac{\beta_l}{2} R \succ 0.$$

If in addition $R = \mathrm{Id}$ and $\Gamma = \gamma \mathrm{Id}$, $\Sigma = \sigma \mathrm{Id}$ for some scalars γ, σ, then the following sufficient condition may be used

$$3\sigma\gamma\|L\|^2 < 1 - \frac{\sigma\beta_l}{2}.$$

The parallel variant of SDCA is considered below. Interestingly, by switching the order of the proximal updates (since $\theta = 0$), PDCA may be seen as PPCA applied to the dual problem (4).

PDCA (Parallel Dual Corrector Algorithm): $\theta = 0$, $\mu = 0$

In this case the convergence condition simplifies to

$$\begin{pmatrix} (\frac{2}{\lambda} - 1)\Gamma^{-1} - 2L^\top \Sigma L - \frac{\beta_f}{2\lambda} Q & -L^\top \\ -L & (\frac{2}{\lambda} - 1)\Sigma^{-1} - \frac{\beta_l}{2\lambda} R \end{pmatrix} \succ 0.$$

If $l = \iota_{\{0\}}$ (set $\beta_l = 0$) and $\lambda = 1$, using Schur complement yields the following condition

$$\Gamma^{-1} - 3L^\top \Sigma L - \frac{\beta_f}{2} Q \succ 0.$$

If in addition $Q = \mathrm{Id}$, and $\Gamma = \gamma \mathrm{Id}$, $\Sigma = \sigma \mathrm{Id}$ for some scalars γ, σ, then the following sufficient condition may be used

$$3\sigma\gamma\|L\|^2 < 1 - \frac{\gamma\beta_f}{2}.$$

The last special case considered here involves sequential proximal updates followed by correction steps for the primal and dual variables. As noted before for this choice of μ, the linear mappings L and $L^\top$ must be evaluated twice at every iteration.

PPDCA (Parallel Primal and Dual Corrector Algorithm): $\theta = 0$, $\mu = 0.5$

In this case condition (9) reduces to:

$$\Gamma^{-1} - \tfrac{\lambda}{(2-\lambda)} L^\top \Sigma L - \tfrac{\beta_f}{2(2-\lambda)} Q \succ 0, \quad \Sigma^{-1} - \tfrac{\lambda}{(2-\lambda)} L \Gamma L^\top - \tfrac{\beta_l}{2(2-\lambda)} R \succ 0.$$

If $Q = \mathrm{Id}$, $R = \mathrm{Id}$, $\lambda = 1$ and $\Gamma = \gamma \mathrm{Id}$, $\Sigma = \sigma \mathrm{Id}$ for some scalars γ, σ, the following sufficient condition may be used

$$\sigma\gamma\|L\|^2 < \mathbf{min}\{1 - \tfrac{\gamma\beta_f}{2}, 1 - \tfrac{\sigma\beta_l}{2}\}.$$

This special case generalizes [8, Algorithm (4.8)] with the addition of the smooth function f and the strongly convex function l.

3 Simplified Asymmetric Forward-Backward-Adjoint Splitting

A new three-term splitting technique was introduced in [24] for the problem of finding $z \in \mathbb{R}^p$ such that

$$0 \in Tz := Az + Cz + Mz, \tag{11}$$

where A is maximally monotone, C is cocoercive and M is a monotone linear mapping. AFBA in its original form includes a dynamic stepsize, see [24, Alg. 1]. In this work we simplify the algorithm by considering a constant stepsize, see Algorithm 2. This variant of AFBA is particularly advantageous in distributed applications where global coordination may be infeasible. Furthermore, unlike [24, Alg. 1], cocoercivity of the operator C is considered with respect to some norm independent of the parameters of the algorithm, and the convergence condition is derived in terms of a matrix inequality. These changes simplify the analysis for the primal-dual algorithms discussed in Section 2. We remind the reader that Algorithm 1 is the result of solving the primal-dual optimality conditions using AFBA. We defer the derivation and convergence analysis of Algorithm 1 until Section 4.

Let us first recall the notion of cocoercivity.

Definition 3.1 (Cocoercivity)**.** *The operator $C : \mathbb{R}^p \to \mathbb{R}^p$ is said to be cocoercive with respect to $\|\cdot\|_U$ if for all $z, z' \in \mathbb{R}^p$*

$$\langle Cz - Cz', z - z' \rangle \geq \|Cz - Cz'\|^2_{U^{-1}}. \tag{12}$$

A basic key inequality that we use is the following.

Lemma 3.2 (Three-point inequality)**.** *Suppose that $C : \mathbb{R}^p \to \mathbb{R}^p$ is cocoercive with respect to $\|\cdot\|_U$, and let $V : \mathbb{R}^p \to \mathbb{R}^p$ be a linear mapping such that $V \circ C = C$ (identity is the trivial choice). Then for any three points $z, z', z'' \in \mathbb{R}^p$ we have*

$$\langle Cz - Cz', z' - z'' \rangle \leq \tfrac{1}{4} \|V^\top (z - z'')\|^2_U. \tag{13}$$

Proof. Use the inequality, valid for any $a, b \in \mathbb{R}^p$,

$$\langle a,b\rangle = 2\langle \tfrac{1}{2}U^{\frac{1}{2}}a, U^{-\frac{1}{2}}b\rangle \leq \tfrac{1}{4}\|a\|_U^2 + \|b\|_{U^{-1}}^2, \tag{14}$$

together with (12) and $V \circ C = C$ to derive

$$\begin{aligned}\langle Cz - Cz', z' - z''\rangle &= \langle V(Cz-Cz'), z - z''\rangle + \langle Cz - Cz', z' - z\rangle \\ &= \langle Cz - Cz', V^\top(z-z'')\rangle + \langle Cz - Cz', z' - z\rangle \\ &\overset{(14)}{\leq} \tfrac{1}{4}\|V^\top(z-z'')\|_U^2 + \|Cz-Cz'\|_{U^{-1}}^2 + \langle Cz - Cz', z' - z\rangle \\ &\overset{(12)}{\leq} \tfrac{1}{4}\|V^\top(z-z'')\|_U^2. \end{aligned}$$ □

Our main motivation for considering V is to avoid conservative bounds in (13). For example, assume that the space is partitioned into two, $z = (z_1, z_2)$, and $Cz = (C_1 z_1, 0)$ where C_1 is cocoercive. Using inequality (14) without taking into account the structure of C, *i.e.*, that $V \circ C = C$, would result in the whole vector appearing in the upper bound in (13).

Algorithm 2 involves two matrices H and S that are instrumental to its flexibility. In Section 4 we discuss a choice for H and S and demonstrate how Algorithm 1 is derived. Below we summarize the assumptions for the monotone inclusion (11) and the convergence conditions for Algorithm 2 (*cf.* Theorem 3.4).

Assumption III.

(i) Assumptions for the monotone inclusion (11):

1) *The operator $A : \mathbf{R}^p \rightrightarrows \mathbf{R}^p$ is maximally monotone.*
2) *The linear mapping $M : \mathbf{R}^p \to \mathbf{R}^p$ is monotone.*
3) *The operator $C : \mathbf{R}^p \to \mathbf{R}^p$ is cocoercive with respect to $\|\cdot\|_U$. In addition, $V : \mathbf{R}^p \to \mathbf{R}^p$ is a linear mapping such that $V \circ C = C$ (identity is the trivial choice).*

(ii) Convergence conditions for Algorithm 2:

1) *The matrix $H := P + K$, where $P \in \mathcal{S}_{++}^p$ and K is a skew-symmetric matrix.*
2) *The matrix $S \in \mathcal{S}_{++}^p$ and*

$$2P - \tfrac{1}{2}VUV^\top - D \succ 0, \tag{15}$$

where

$$D := (H + M^\top)^\top S^{-1}(H + M^\top). \tag{16}$$

Algorithm 2 AFBA with constant stepsize

Require: $z^0 \in \mathbf{R}^p$
Initialize: set S and H according to Assumption III*(ii)*.
for $k = 0, 1, \ldots$ **do**
$\bar{z}^k = (H + A)^{-1}(H - M - C)z^k$
$z^{k+1} = z^k + S^{-1}(H + M^\top)(\bar{z}^k - z^k)$

Lemma 3.3. *Let Assumption III hold. Consider the update for $\bar{z}$ in Algorithm 2*

$$\bar{z} = (H+A)^{-1}(H-M-C)z. \tag{17}$$

For all $z^\star \in \mathbf{zer}\, T$ the following holds

$$\langle z - z^\star, (H+M^\top)(\bar{z}-z)\rangle \leq \tfrac{1}{4}\|V^\top(z-\bar{z})\|_U^2 - \|z-\bar{z}\|_P^2. \tag{18}$$

Proof. Use (17) and the fact that $z^\star \in \mathbf{zer}\, T$, together with monotonicity of A at $z^\star$ and $\bar{z}$ to derive

$$0 \leq \langle -Mz^\star - Cz^\star + Mz + Cz + H(\bar{z}-z), z^\star - \bar{z}\rangle. \tag{19}$$

In Lemma 3.2 set $z' = z^\star$ and $z'' = \bar{z}$

$$\langle Cz - Cz^\star, z^\star - \bar{z}\rangle \leq \tfrac{1}{4}\|V^\top(z-\bar{z})\|_U^2. \tag{20}$$

For the remaining terms in (19) use skew-symmetry of K (twice) and monotonicity of M:

$$\begin{aligned}
&\langle -Mz^\star + Mz + H(\bar{z}-z), z^\star - \bar{z}\rangle \\
=&\langle -Mz^\star + Mz + P(\bar{z}-z) + K(\bar{z}-z) + K(z^\star - \bar{z}), z^\star - \bar{z}\rangle \\
=&\langle (M-K)(z-z^\star) + P(\bar{z}-z), z^\star - \bar{z}\rangle \\
=&\langle (M-K)(z-z^\star) + P(\bar{z}-z), z^\star - z\rangle \\
&+ \langle (M-K)(z-z^\star) + P(\bar{z}-z), z - \bar{z}\rangle \\
\leq&\langle P(\bar{z}-z), z^\star - z\rangle + \langle (M-K)(z-z^\star), z-\bar{z}\rangle - \|\bar{z}-z\|_P^2 \\
\leq&\langle z - z^\star, (M^\top + H)(z-\bar{z})\rangle - \|\bar{z}-z\|_P^2.
\end{aligned}$$

Combining this with (19) and (20) completes the proof. □

Theorem 3.4 (Convergence)**.** *Let Assumption III hold. Consider the sequence $(z^k)_{k\in\mathbf{N}}$ generated by Algorithm 2. Then the following inequality holds for all $k \in \mathbf{N}$*

$$\|z^{k+1} - z^\star\|_S^2 \leq \|z^k - z^\star\|_S^2 - \|z^k - \bar{z}^k\|^2_{2P-\frac{1}{2}VUV^\top - D}, \tag{21}$$

and $(z^k)_{k\in\mathbf{N}}$ converges to a point $z^\star \in \mathbf{zer}\, T$.

Proof. We show that the generated sequence is Fejér monotone with respect to $\mathbf{zer}\, T$ in the space equipped with the inner product $\langle\cdot,\cdot\rangle_S$. For any $z^\star \in \mathbf{zer}\, T$ using the z^{k+1} update in Algorithm 2 we have

$$\begin{aligned}\|z^{k+1}-z^\star\|_S^2 &= \|z^k-z^\star\|_S^2 + \|S^{-1}(H+M^\top)(\bar{z}^k-z^k)\|_S^2 \\ &\quad + 2\langle z^k-z^\star, (H+M^\top)(\bar{z}^k-z^k)\rangle \\ &\overset{(18)}{\leq} \|z^k-z^\star\|_S^2 + \|S^{-1}(H+M^\top)(\bar{z}^k-z^k)\|_S^2 \\ &\quad + \tfrac{1}{2}\|V^\top(z^k-\bar{z}^k)\|_U^2 - 2\|z^k-\bar{z}^k\|_P^2 \\ &\overset{(16)}{\leq} \|z^k-z^\star\|_S^2 - \|z^k-\bar{z}^k\|^2_{2P-\frac{1}{2}VUV^\top-D}.\end{aligned}$$

Therefore the sequence $(z^k-\bar{z}^k)_{k\in\mathbf{N}}$ converges to zero. Convergence of $(z^k)_{k\in\mathbf{N}}$ to a point in $\mathbf{zer}\,T$ follows by standard arguments; see the last part of the proof of [24, Thm. 3.1]. □

Our next goal is to establish linear convergence for Algorithm 2. Before continuing let us recall the notion of metric subregularity.

Definition 3.5 (Metric subregularity)**.** *A mapping F is* metrically subregular *at $\bar{x}$ for $\bar{u}$ if $(\bar{x},\bar{u})\in \operatorname{gra} F$ and there exists a positive constant κ, and neighborhoods $\mathcal{U}$ of $\bar{x}$ and $\mathcal{Y}$ of $\bar{u}$ such that*

$$d(x,F^{-1}\bar{u}) \leq \kappa d(\bar{u},Fx\cap\mathcal{Y}),\ \ \forall x\in\mathcal{U}.$$

If in addition $\bar{x}$ is an isolated point of $F^{-1}\bar{u}$, i.e., $F^{-1}\bar{u}\cap\mathcal{U}=\{\bar{x}\}$, then F is said to be strongly subregular *at $\bar{x}$ for $\bar{u}$. Otherwise stated:*

$$\|x-\bar{x}\| \leq \kappa d(\bar{u},Fx\cap\mathcal{Y}),\ \ \forall x\in\mathcal{U}.$$

The neighborhood $\mathcal{Y}$ in the above definitions can be omitted [17, Ex. 3H.4].

Metric subregularity is a "one-point" version of metric regularity. We refer the interested reader to [17, Chap. 3] and [29, Chap. 9] for an extensive discussion.

The linear convergence for Algorithm 2 is established in Theorem 3.6 under two different assumptions: (i) when the operator T in (11) is metrically subregular at all $z^\star\in\mathbf{zer}\,T$ for 0 or (ii) when the operator $\mathrm{Id}-T_{\mathrm{AFBA}}$ has this property, where T_{AFBA} denotes the operator that maps z^k to z^{k+1} in Algorithm 2. In Section 4 we exploit the first result, Theorem 3.6*(i)*. The advantage of this result is that it is easier to characterize the metric subregularity of T in terms of its component. It is worth mentioning that for the preconditioned proximal point algorithm (a special case of Algorithm 2 with $C=0$, $M=0$, $K=0$, $S=P$) the first assumption implies the second one [30, Prop. IV.2].

Theorem 3.6 (Linear convergence)**.** *Let Assumption III hold. Consider the sequence $(z^k)_{k\in\mathbf{N}}$ generated by Algorithm 2. Suppose that one of the following assumptions holds:*

(i) $T=A+M+C$ is metrically subregular at all $z^\star\in\mathbf{zer}\,T$ for 0.

(ii) Let T_{AFBA} denote the operator that maps z^k to z^{k+1} in Algorithm 2, i.e., $z^{k+1} = T_{\mathrm{AFBA}}(z^k)$. The operator $\mathrm{Id} - T_{\mathrm{AFBA}}$ is metrically subregular at all $z^\star \in \mathbf{zer}\, T$ for 0.

Then $(z^k)_{k\in\mathbf{N}}$ converges R-linearly to some $z^\star \in \mathbf{zer}\, T$ and $(d_S(z^k, \mathbf{zer}\, T))_{k\in\mathbf{N}}$ converges Q-linearly to zero.

Proof. The proof for Theorem 3.6*(i)* can be found in [24, Thm. 3.3]. The proof of the second part follows by noting that

$$\begin{aligned} d_S^2(z^{k+1}, \mathbf{zer}\, T) &\leq \|z^{k+1} - \Pi^S_{\mathbf{zer}\, T}(z^k)\|_S^2 \\ &\overset{(21)}{\leq} \|z^k - \Pi^S_{\mathbf{zer}\, T}(z^k)\|_S^2 - \|z^k - \bar{z}^k\|^2_{2P-\frac{1}{2}VUV^\top - D} \qquad (22) \\ &= d_S^2(z^k, \mathbf{zer}\, T) - \|z^k - z^{k+1}\|_W^2, \qquad (23) \end{aligned}$$

where W is some symmetric positive definite matrix given by replacing $\bar{z}^k - z^k = (H + M^\top)^{-1} S(z^{k+1} - z^k)$ in (22). It remains to bound $d_S(z^k, \mathbf{zer}\, T)$ by $\|z^k - z^{k+1}\|_W$. By Theorem 3.4, z^k converges to some $z^\star \in \mathbf{zer}\, T$. Since $I - T_{\mathrm{AFBA}}$ is metrically subregular at $z^\star$ for 0, there exists $\bar{k} \in \mathbf{N}$, a positive constant κ, and a neighborhood $\mathcal{U}$ of $z^\star$ such that

$$d(z^k, \mathbf{zer}\, T) \leq \kappa \|z^k - z^{k+1}\|, \quad \forall k \geq \bar{k}, \qquad (24)$$

where we used the fact that the set of fixed points of T_{AFBA} is equal to $\mathbf{zer}\, T$. Therefore for $k \geq \bar{k}$ we have

$$\begin{aligned} d_S^2(z^k, \mathbf{zer}\, T) &\leq \|z^k - \Pi_{\mathbf{zer}\, T}(z^k)\|_S^2 \leq \|S\| \|z^k - \Pi_{\mathbf{zer}\, T}(z^k)\|^2 \\ &= \|S\| d^2(z^k, \mathbf{zer}\, T) \overset{(24)}{\leq} \kappa^2 \|S\| \|z^k - z^{k+1}\|^2 \\ &\leq \kappa^2 \|S\| \|W^{-1}\| \|z^k - z^{k+1}\|_W^2 \qquad (25) \end{aligned}$$

Combining (25) with (23) proves the Q-linear convergence rate for the distance from the set of solutions. Using this Q-linear convergence rate for the distance, it follows from (23) that $(\|z^k - z^{k+1}\|_W)_{k\in\mathbf{N}}$ converges R-linearly to zero and hence also $(z^k)_{k\in\mathbf{N}}$ converges R-linearly. □

4 A Unified Convergence Analysis for Primal-Dual Algorithms

Our goal in this section is to describe how Algorithm 1 is derived and to establish its convergence. The idea is to write the primal-dual optimality conditions as a monotone inclusion involving the sum of a maximally monotone, a linear monotone and a cocoercive operator, *cf.* (7). This monotone inclusion is then solved using *asymmetric forward-backward-adjoint* splitting (AFBA) described in Section 3. In order to recover Algorithm 1 simply apply Algorithm 2 to this monotone inclusion with the following parameters: Let $\theta \in [0, \infty)$ and set $H = P + K$ with

$$P = \begin{pmatrix} \Gamma^{-1} & -\frac{\theta}{2}L^\top \\ -\frac{\theta}{2}L & \Sigma^{-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & \frac{\theta}{2}L^\top \\ -\frac{\theta}{2}L & 0 \end{pmatrix}, \tag{26}$$

and $S = \left(\lambda\mu S_1^{-1} + \lambda(1-\mu)S_2^{-1}\right)^{-1}$ where $\mu \in [0,1]$, $\lambda \in (0,2)$ with

$$S_1 = \begin{pmatrix} \Gamma^{-1} & (1-\theta)L^\top \\ (1-\theta)L & \Sigma^{-1} + (1-\theta)(2-\theta)L\Gamma L^\top \end{pmatrix}, \; S_2 = \begin{pmatrix} \Gamma^{-1} + (2-\theta)L^\top \Sigma L & -L^\top \\ -L & \Sigma^{-1} \end{pmatrix}.$$

Notice that with P and K set as in (26), H has a lower (block) triangular structure. Therefore the backward step $(H+A)^{-1}$ in Algorithm 2 can be carried out sequentially [24, Lem. 3.1]. Algorithm 1 is derived by noting this and substituting S and H defined above. We refer the reader to [24, Sec. 5] for a more detailed procedure.

Our next goal is to verify that Assumptions I and II are sufficient for Assumption III to hold. As noted before, the operator A is maximally monotone [4, Thm. 21.2, Prop. 20.33], and the linear mapping M is skew-adjoint and as such monotone. The operator C is cocoercive with respect to the metric induced by $U = \operatorname{blkdiag}(\beta_f Q, \beta_l R)$. In Assumption III*(ii)* we use the linear mapping V in order to avoid conservative requirements. The special cases when $f \equiv 0$ (or $l^* \equiv 0$) are captured by setting $V = \operatorname{blkdiag}(0_n, I_r)$ (or $V = \operatorname{blkdiag}(I_n, 0_r)$) where I_s and 0_s denote the identity and zero matrices in $\mathbb{R}^{s\times s}$. It remains to verify Assumption III*(ii)*. Evaluating D according to (16) yields the following $D = \lambda\mu D_1 + \lambda(1-\mu)D_2$ where

$$D_1 = \begin{pmatrix} \Gamma^{-1} & -L^\top \\ -L & \Sigma^{-1} + (2-\theta)L\Gamma L^\top \end{pmatrix},$$

$$D_2 = \begin{pmatrix} \Gamma^{-1} + (1-\theta)(2-\theta)L^\top \Sigma L & (1-\theta)L^\top \\ (1-\theta)L & \Sigma^{-1} \end{pmatrix}.$$

Noting that $\Gamma, \Sigma \succ 0$, and using Schur complement for D_1 and P defined in (26) we have

$$D_1 \succ 0 \;\Leftrightarrow\; \Sigma^{-1} + (1-\theta)L\Gamma L^\top \succ 0, \quad \Sigma^{-1} - \tfrac{\theta^2}{4}L\Gamma L^\top \succ 0 \;\Leftrightarrow\; P \succ 0.$$

Thus, since $1 - \theta \geq -\frac{\theta^2}{4}$ for all θ, we have that $D_1 \succ 0$ if $P \succ 0$. It can be shown that the same argument applies for S_1, S_2 and D_2. The sum of two positive definite matrices is also positive definite, therefore $S, D \succ 0$ if $P \succ 0$. The matrix P is symmetric positive definite if (15) holds. The convergence conditions in Assumption II are simply the result of replacing D, P, U and V in (15).

We showed that Assumption III holds for the described choice of H and S. Algorithm 1 is a simple application of Algorithm 2 for solving the monotone inclusion (7). Therefore, the convergence of Algorithm 1 follows directly from that of AFBA (*cf.* Theorem 3.4). This result is summarized in the following theorem.

Theorem 4.1 (Convergence)**.** *Let Assumptions I and II hold. The sequence* $(z^k)_{k\in\mathbf{N}} = (x^k, u^k)_{k\in\mathbf{N}}$ *generated by Algorithm 1 converges to a point* $z^\star \in \mathbf{zer}\, T$.

Linear Convergence

In this section we explore sufficient conditions for f, g, h and l under which Algorithm 1 achieves linear convergence. We saw that Algorithm 1 is an instance of AFBA, therefore its linear convergence may be established based on the results developed in Theorem 3.6*(i)*. In Lemma 4.3 and Lemma 4.5 we provide four regularity assumptions under which T defining the primal-dual optimality conditions is metrically subregular at $z^\star \in \mathbf{zer}\, T$ for 0. The linear convergence for Algorithm 1 is then deduced in Corollary 4.6.

Let us first recall the notion of quadratic growth: a proper closed convex function g is said to have *quadratic growth* at $\bar{x}$ for 0 with $0 \in \partial g(\bar{x})$ if there exists a neighborhood $\mathcal{U}$ of $\bar{x}$ such that

$$g(x) \geq \mathbf{inf}\, g + cd^2(x, \partial g^{-1}(0)), \quad \forall x \in \mathcal{U}. \tag{27}$$

Metric subregularity of the subdifferential operator and the quadratic growth condition are known to be equivalent [2, 19]. In particular, ∂g is metrically subregular at $\bar{x}$ for $\bar{u}$ with $\bar{u} \in \partial g(\bar{x})$ if and only if the quadratic growth condition (27) holds for $g(\cdot) - \langle \bar{u}, \cdot \rangle$, *i.e.*, there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that [2, Thm. 3.3]

$$g(x) \geq g(\bar{x}) + \langle \bar{u}, x - \bar{x} \rangle + cd^2(x, \partial g^{-1}(\bar{u})), \quad \forall x \in \mathcal{U}. \tag{28}$$

Strong subregularity has a similar characterization [2, Thm. 3.5]

$$g(x) \geq g(\bar{x}) + \langle \bar{u}, x - \bar{x} \rangle + c\|x - \bar{x}\|^2, \quad \forall x \in \mathcal{U}. \tag{29}$$

Next, let us define the following general growth condition.

Definition 4.2 (Quadratic growth relative to a set)**.** *Consider a proper closed convex function g and a pair $(\bar{x}, \bar{u}) \in \operatorname{gra} \partial g$. We say that g has quadratic growth at $\bar{x}$ for $\bar{u}$ relative to a nonempty closed convex set X containing $\bar{x}$, if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that*

$$g(x) \geq g(\bar{x}) + \langle \bar{u}, x - \bar{x} \rangle + cd^2(x, X), \quad \forall x \in \mathcal{U}. \tag{30}$$

From the above definition it is evident that metric subregularity and strong subregularity characterized in (28) and (29) are recovered when $X = \partial g^{-1}(\bar{u})$ and $X = \{\bar{x}\}$, respectively.

Another regularity assumption used in Lemma 4.3 is the notion of local strong convexity: a proper closed convex function g is said to be locally strongly convex in a neighborhood of $\bar{x}$, denoted by $\mathcal{U}$, if there exists a positive constant c such that

$$g(x') \geq g(x) + \langle v, x' - x \rangle + \tfrac{c}{2}\|x' - x\|^2, \quad \forall x, x' \in \mathcal{U},\ \forall v \in \partial g(x).$$

Notice that local strong convexity in a neighborhood of $\bar{x}$ implies (29), but (29) is *much weaker* than local strong convexity since it holds only at $\bar{x}$ and only for $\bar{u} \in \partial g(\bar{x})$.

In the next lemma we provide three different regularity assumptions that are sufficient for metric subregularity of the operator defining the primal-dual optimality conditions. In Lemma 4.3*(i)* (or Lemma 4.3*(ii)*) we use local strong convexity, as well as the quadratic growth condition (30) relative to the set of primal solutions (or dual solutions). Interestingly, this regularity assumption does not entail a unique primal-dual solution.

Lemma 4.3. *Let Assumption I hold. The operator T defining the primal-dual optimality conditions, cf. (8), is metrically subregular at $z^\star = (x^\star, u^\star)$ for 0 with $0 \in Tz^\star$ if one of the following assumptions holds:*

(i) *$f+g$ has quadratic growth at $x^\star$ for $-L^\top u^\star$ relative to the set of primal solutions $X^\star$, and $h^* + l^*$ is locally strongly convex in a neighborhood of $u^\star$. In this case the set of dual solutions is a singleton, $U^\star = \{u^\star\}$.*

(ii) *$f+g$ is locally strongly convex in a neighborhood of $x^\star$, and $h^* + l^*$ has quadratic growth at $u^\star$ for $Lx^\star$ relative to the set of dual solutions $U^\star$. In this case the set of primal solutions is a singleton, $X^\star = \{x^\star\}$.*

(iii) *$\nabla f + \partial g$ is strongly subregular at $x^\star$ for $-L^\top u^\star$ and $\partial h^* + \nabla l^*$ is strongly subregular at $u^\star$ for $Lx^\star$. In this case the set of primal-dual solutions is a singleton, $\mathbf{zer}\, T = \{(x^\star, u^\star)\}$.*

Proof. 4.3*(i)*- Consider the point $z^\star = (x^\star, u^\star)$. By definition of quadratic growth there exists a neighborhood $\mathcal{U}_{x^\star}$ and a positive constant c_1 such that

$$(f+g)(x) \geq (f+g)(x^\star) + \langle -L^\top u^\star, x - x^\star \rangle + c_1 d^2(x, X^\star), \quad \forall x \in \mathcal{U}_{x^\star}. \tag{31}$$

Let $\mathcal{U}_{u^\star}$ denote the neighborhood of $u^\star$ in which the local strong convexity of $h^* + l^*$ holds. Fix a point $z = (x,u) \in \mathcal{Z}_{z^\star} := \mathcal{U}_{x^\star} \times \mathcal{U}_{u^\star}$. Now take $v = (v_1, v_2) \in Tz = Az + Mz + Cz$, *i.e.*,

$$\begin{cases} v_1 \in \partial g(x) + \nabla f(x) + L^\top u, \\ v_2 \in \partial h^*(u) + \nabla l^*(u) - Lx. \end{cases} \tag{32}$$

Let $z_0 = (x_0, u_0)$ denote the projection of z onto the set of solutions, $\mathbf{zer}\, T$. The subgradient inequality for $f+g$ at x using (32) gives

$$\langle v_1, x - x_0 \rangle \geq (f+g)(x) - (f+g)(x_0) + \langle L^\top u, x - x_0 \rangle. \tag{33}$$

Noting that $0 \in Tz_0$, by the subgradient inequality for $h^* + l^*$ at u_0 we have

$$(h^* + l^*)(u) \geq (h^* + l^*)(u_0) + \langle Lx_0, u - u_0 \rangle. \tag{34}$$

Summing (33) and (34) yields

$$\langle v_1, x - x_0 \rangle \geq \mathcal{L}(x,u) - \mathcal{L}(x_0, u_0) = \mathcal{L}(x,u) - \mathcal{L}(x^\star, u^\star), \tag{35}$$

where $\mathcal{L}$ is the Lagrangian defined in (5). By local strong convexity of h^*+l^* at $u\in\mathcal{U}_{u^\star}$ (for some strong convexity parameter c_2):

$$(h^*+l^*)(u^\star)\geq(h^*+l^*)(u)+\langle v_2+Lx,u^\star-u\rangle+\tfrac{c_2}{2}\|u^\star-u\|^2.$$

Sum this inequality with (31) to derive

$$\mathcal{L}(x,u)-\mathcal{L}(x^\star,u^\star)\geq c_1d^2(x,X^\star)+\langle v_2,u^\star-u\rangle+\tfrac{c_2}{2}\|u^\star-u\|^2,\quad z\in\mathcal{Z}_{z^\star}. \tag{36}$$

It follows from (35) and (36) that

$$\begin{aligned}\langle v_2,u-u^\star\rangle+\langle v_1,x-x_0\rangle&\geq\tfrac{c_2}{2}\|u^\star-u\|^2+c_1d^2(x,X^\star)\\&=\tfrac{c_2}{2}\|u^\star-u\|^2+c_1\|x-x_0\|^2\\&\geq c\big(\|u^\star-u\|^2+\|x-x_0\|^2\big),\end{aligned} \tag{37}$$

where $c=\mathbf{min}\{c_1,\frac{c_2}{2}\}$. By the Cauchy-Schwarz inequality

$$\langle v_1,x-x_0\rangle+\langle v_2,u-u^\star\rangle\leq\|v\|\big(\|u-u^\star\|^2+\|x-x_0\|^2\big)^{\frac{1}{2}}. \tag{38}$$

Combining (37) and (38) yields

$$\|v\|\geq c\big(\|u-u^\star\|^2+\|x-x_0\|^2\big)^{\frac{1}{2}}\geq c\|z-z_0\|=cd(z,T^{-1}0).$$

Since $v\in Tz$ was selected arbitrarily we have that

$$d(z,T^{-1}0)\leq\tfrac{1}{c}d(Tz,0),\quad\forall z\in\mathcal{Z}_{z^\star}.$$

This completes the first claim. Next, consider $\bar{z}^\star=(\bar{x}^\star,\bar{u}^\star)\in\mathbf{zer}\,T$ such that $\bar{z}^\star\in\mathcal{Z}_{z^\star}$. Setting $z=\bar{z}^\star$ in (36) yields

$$0=\mathcal{L}(\bar{x}^\star,\bar{u}^\star)-\mathcal{L}(x^\star,u^\star)\geq\tfrac{c_2}{2}\|u^\star-\bar{u}^\star\|^2.$$

Therefore, $\bar{u}^\star=u^\star$ and since $\mathbf{zer}\,T$ is convex we conclude that $U^\star=\{u^\star\}$.

4.3*(ii)*- The proof of the second part is similar to part 4.3*(i)*. Therefore, we outline the proof. Let $\mathcal{U}_{x^\star},\mathcal{U}_{u^\star}$ be the neighborhoods in the definition of local strong convexity of $f+g$ and quadratic growth of h^*+l^*, respectively. Fix $z\in\mathcal{Z}_{z^\star}=\mathcal{U}_{x^\star}\times\mathcal{U}_{u^\star}$. Take $v=(v_1,v_2)$ as in (32), and let z_0 denote the projection of z onto $\mathbf{zer}\,T$. In contrast to the previous part, sum the subgradient inequalities for $f+g$ at x_0 and for h^*+l^* at u to derive

$$\langle v_2,u-u_0\rangle\geq\mathcal{L}(x_0,u_0)-\mathcal{L}(x,u)=\mathcal{L}(x^\star,u^\star)-\mathcal{L}(x,u). \tag{39}$$

Use local strong convexity of $f+g$ at x, and the quadratic growth of h^*+l^* at $u^\star$ for $Lx^\star$ relative to $U^\star$ to derive

$$\mathcal{L}(x^\star,u^\star)-\mathcal{L}(x,u)\geq c_1d(u,U^\star)+\langle v_1,x^\star-x\rangle+\tfrac{c_2}{2}\|x-x^\star\|^2.$$

Combining this with (39) and arguing as in the previous part completes the proof.

4.3*(iii)*- The proof of this part is slightly different. Let $\mathcal{U}_{x^\star}, \mathcal{U}_{u^\star}$ be the neighborhoods in the definitions of the two strong subregularity assumptions. Fix $z \in \mathcal{Z}_{z^\star} = \mathcal{U}_{x^\star} \times \mathcal{U}_{u^\star}$. Sum the subgradient inequality for $f+g$ at x and for $h^* + l^*$ at u to derive

$$\langle v, z - z^\star \rangle = \langle v_2, u - u^\star \rangle + \langle v_1, x - x^\star \rangle \geq \mathcal{L}(x, u^\star) - \mathcal{L}(x^\star, u) \tag{40}$$

On the other hand by [2, Thm. 3.5], $f+g$ has quadratic growth at $x^\star$ for $-L^\top u^\star$ relative to $\{x^\star\}$ and $h^* + l^*$ has quadratic growth at $u^\star$ for $Lx^\star$ relative to $\{u^\star\}$. Summing the two yields

$$\mathcal{L}(x, u^\star) - \mathcal{L}(x^\star, u) \geq c_2 \|u - u^\star\|^2 + c_1 \|x - x^\star\|^2 \tag{41}$$

Combining this inequality with (40) and using the Cauchy-Schwartz inequality as in previous parts completes the proof. Uniqueness of the solution follows from (41) by setting $z = \bar{z}^\star \in \mathbf{zer}\, T$ such that $\bar{z}^\star \in \mathcal{Z}_{z^\star}$ and using the convexity of $\mathbf{zer}\, T$. □

The assumptions of Lemma 4.3 are much weaker than strong convexity and do not always imply a unique primal-dual solution. Here we present two simple examples for demonstration. Notice that in the next example the assumption of Lemma 4.3*(i)* that $f+g$ has quadratic growth with respect to the set of primal solutions is equivalent to the metric subregularity assumption.

Example 1. Consider the problem

$$\underset{x \in \mathbf{R}}{\textbf{minimize}}\ g(x) + h(x) = \textbf{max}\{1 - x, 0\} + \tfrac{x}{2}\,\textbf{min}\{x, 0\}$$

The solution to this problem is not unique and any $x^\star \in [1, \infty)$ solves this problem. The dual problem is given by

$$\underset{u \in \mathbf{R}}{\textbf{minimize}}\ g^*(-u) + h^*(u), \tag{42}$$

where $g^*(u) = u + \iota_{[-1,0]}(u)$ and $h^*(u) = \frac{1}{2}u^2 + \iota_{\leq 0}(u)$. It is evident that the dual problem has the unique solution $u^\star = 0$. It is easy to verify that g has quadratic growth at all the points $x^\star \in [1, \infty)$ for 0 with respect to $X^\star = [1, \infty)$. Moreover, we have $\partial g^{-1}(0) = [1, \infty)$, *i.e.*, $X^\star = \partial g^{-1}(0)$. In other words in this case the assumption of Lemma 4.3*(i)* for g is equivalent to the metric subregularity of ∂g at $x^\star$ for 0. Notice that ∂g is not strongly subregular at any point in $[1, \infty)$ for 0. Furthermore, h^* is globally strongly convex given that ∇h is Lipschitz. Therefore, according to Lemma 4.3*(i)* one would expect a unique dual solution but not necessarily a unique primal solution, which is indeed the case.

Example 2. Let $c \in [-1, 1]$ and consider

$$\underset{x \in \mathbf{R}}{\textbf{minimize}}\ g(x) + h(x) = |x| + cx.$$

When $c \in (-1,1)$ the problem attains a unique minimum at $x^\star = 0$. When $c = 1$ (or $c = -1$) all $x^\star \in (-\infty, 0]$ (or $x^\star \in [0,\infty)$) solves the problem. The dual problem is given by (42) with $g^*(u) = \iota_{[-1,1]}(u)$ and $h^*(u) = \iota_{\{c\}}(u)$. The unique dual solution is $u^\star = c$. Furthermore, ∂h^* is strongly subregular at $u^\star = c$ for all $x^\star$ given that $x^\star \in \partial h^*(u^\star)$. It is easy to verify that ∂g is metrically subregular at $x^\star = 0$ for $u^\star \in [-1,1]$ but is only strongly subregular at $x^\star = 0$ for $u^\star \in (-1,1)$. Notice that $u^\star = c$, therefore by Lemma 4.3*(iii)* one would expect a unique primal-dual solution when $u^\star = c \in (-1,1)$ which is indeed the case.

Another class of functions prevalent in optimization is the class of *piecewise linear-quadratic* (PLQ) functions, which is closed under scalar multiplication, addition, conjugation and Moreau envelope [29]. Recall the notion of piecewise linear-quadratic functions.

Definition 4.4 (Piecewise Linear-Quadratic function)**.** *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called piecewise linear-quadratic (PLQ) if its domain can be represented as the union of finitely many polyhedral sets, and in each such set $f(x)$ is given by an expression of the form $\frac{1}{2}x^\top Qx + d^\top x + c$, for some $c \in \mathbb{R}$, $d \in \mathbb{R}^n$, and $Q \in \mathcal{S}^n$.*

A wide range of functions used in optimization applications belong to this class, for example: affine functions, quadratic forms, indicators of polyhedral sets, polyhedral norms such as ℓ_1, and regularizers such as elastic net, Huber loss, hinge loss, and many more [29, 3]. For a proper closed convex PLQ function g, ∂g is piecewise polyhedral, and therefore metrically subregular at any z for any z' provided that $z' \in \partial g(z)$ [29, 17].

Lemma 4.5. *Let Assumption I hold. In addition, assume that f, g, l, and h are piecewise linear-quadratic. Then the operator T defining the primal-dual optimality conditions, cf. (8), is metrically subregular at any z for any z' provided that $z' \in Tz$.*

Proof. The subdifferential ∂g, ∇f, ∂h^* and ∇l^* are piecewise polyhedral [29, Prop. 12.30, Thm. 11.14]. Therefore, A and C are piecewise polyhedral. Furthermore, the graph of M is polyhedral since M is linear. Therefore, the graph of $T = A + M + C$ is also piecewise polyhedral. The inverse of a piecewise polyhedral mapping is also piecewise polyhedral. Therefore by [17, Prop. 3H.1, 3H.3] the mapping T is metrically subregular at z for z' whenever $(z, z') \in \operatorname{gra} T$. □

The next corollary summarizes the linear convergence results based on the two previous lemmas and Theorem 3.6*(i)*.

Corollary 4.6 (Linear convergence for Algorithm 1)**.** *Let Assumptions I and II hold. In addition, suppose that one of the following assumptions holds:*

(i) f, g, l, and h are piecewise linear-quadratic.

(ii) f, g, l, and h satisfy at least one of the the conditions of Lemma 4.3 at every $z^\star \in \mathbf{zer}\, T$ (not necessarily the same condition at all the points).

Then the sequence $(z^k)_{k\in\mathbf{N}}$ generated by Algorithm 1 converges R-linearly to some $z^\star \in \mathbf{zer}\, T$, and $(d_S(z^k, \mathbf{zer}\, T))_{k\in\mathbf{N}}$ converges Q-linearly to zero.

References

1. A. Alotaibi, P. L. Combettes, and N. Shahzad. Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn-Tucker set. *SIAM Journal on Optimization*, 24(4):2076–2095, 2014.
2. F. J. Aragón Artacho and M. H. Geoffroy. Characterization of metric regularity of subdifferentials. *Journal of Convex Analysis*, 15(2):365–380, 2008.
3. A. Y. Aravkin, J. V. Burke, and G. Pillonetto. Sparse/robust estimation and Kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *Journal of Machine Learning Research*, 14:2689–2728, 2013.
4. H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.
5. D. P. Bertsekas. Extended monotropic programming and duality. *Journal of Optimization Theory and Applications*, 139(2):209–225, Nov 2008.
6. R. I. Boţ and C. Hendrich. 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(4):2541–2565, 2013.
7. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
8. L. M. Briceño-Arias and P. L. Combettes. A monotone + skew splitting model for composite monotone inclusions in duality. *SIAM Journal on Optimization*, 21(4):1230–1250, 2011.
9. L. M. Briceño-Arias and D. Davis. Forward-backward-half forward algorithm with non self-adjoint linear operators for solving monotone inclusions. *arXiv preprint arXiv:1703.03436*, 2017.
10. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.
11. A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, Sep 2016.
12. P. L. Combettes and J. Eckstein. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Mathematical Programming*, Jul 2016.
13. P. L. Combettes and I.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.
14. P. L. Combettes and J.-C. 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(2):307–330, 2012.
15. L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.
16. D. Davis. Convergence rate analysis of primal-dual splitting schemes. *SIAM Journal on Optimization*, 25(3):1912–1943, 2015.
17. A. L. Dontchev and R. T. Rockafellar. Implicit functions and solution mappings. *Springer Monographs in Mathematics. Springer*, 208, 2009.
18. Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.
19. D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv preprint arXiv:1602.06661*, 2016.
20. E. Esser, X. Zhang, and T. 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(4):1015–1046, 2010.
21. B. He and X. Yuan. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences*, 5(1):119–149, 2012.

22. N. Komodakis and J. C. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, Nov 2015.
23. P. Latafat, N. M. Freris, and P. Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *arXiv preprint arXiv:1706.02882*, 2017.
24. P. Latafat and P. Patrinos. Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators. *Computational Optimization and Applications*, pages 1–37, 2017.
25. P. Latafat, L. Stella, and P. Patrinos. New primal-dual proximal algorithm for distributed optimization. In *55th IEEE Conference on Decision and Control (CDC)*, pages 1959–1964, 2016.
26. J. Liang, J. Fadili, and G. Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1):403–434, Sep 2016.
27. D. R. Luke and R. Shefi. A globally linearly convergent method for pointwise quadratically supportable convex–concave saddle point problems. *Journal of Mathematical Analysis and Applications*, 457(2):1568 – 1590, 2018. Special Issue on Convex Analysis and Optimization: New Trends in Theory and Applications.
28. R. T. Rockafellar. *Convex analysis*. Princeton University Press, 2015.
29. R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
30. P. Sopasakis, A. Themelis, J. Suykens, and P. Patrinos. A primal-dual line search method and applications in image processing. In *25th European Signal Processing Conference (EUSIPCO)*, pages 1065–1069, Aug 2017.
31. S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.
32. B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.