A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization

A New Randomized Block-Coordinate Primal-Dual Proximal Algorithm for Distributed Optimization

Puya Latafat, Nikolaos M. Freris, Panagiotis Patrinos

***Abstract*—This paper proposes TriPD, a new primal-dual algorithm for minimizing the sum of a Lipschitz-differentiable convex function and two possibly nonsmooth convex functions, one of which is composed with a linear mapping. We devise a randomized block-coordinate version of the algorithm which converges under the same stepsize conditions as the full algorithm. It is shown that both the original as well as the block-coordinate scheme feature linear convergence rate when the functions involved are either piecewise linear-quadratic, or when they satisfy a certain quadratic growth condition (which is weaker than strong convexity). Moreover, we apply the developed algorithms to the problem of multi-agent optimization on a graph, thus obtaining novel synchronous and asynchronous distributed methods. The proposed algorithms are fully distributed in the sense that the updates and the stepsizes of each agent only depend on local information. In fact, no prior global coordination is required. Finally, we showcase an application of our algorithm in distributed formation control.**

***Index Terms*—Primal-dual algorithms, block-coordinate minimization, distributed optimization, randomized algorithms, asynchronous algorithms.**

I. Introduction

In this paper we consider the optimization problem

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

where L is a linear mapping, h and g are proper, closed, convex functions (possibly nonsmooth), and f is convex, continuously differentiable with Lipschitz-continuous gradient. We further assume that the *proximal mappings* associated with h and g are efficiently computable [1]. This setup is quite general and captures a wide range of applications in signal processing, machine learning and control.

In problem (1), it is typically assumed that the gradient of the smooth term f is β_f-Lipschitz for some nonnegative constant β_f. We consider Lipschitz continuity of ∇f with respect to $\|\cdot\|_Q$ with $Q\succ 0$ in place of the canonical norm (*cf.* (3)). This is because in many applications of practical interest, a scalar Lipschitz constant fails to accurately capture the Lipschitz continuity of ∇f. A prominent example lies in distributed optimization, where f is separable, *i.e.*, $f(x)=\sum_{i=1}^m f_i(x_i)$. In this case, the metric Q is taken block-diagonal with blocks containing the Lipschitz constants of the ∇f_i's. Notice that in such settings considering a scalar Lipschitz constant results in using the largest of the Lipschitz constants, which leads to conservative stepsize selection and consequently slower convergence rates.

The main contributions of the paper are elaborated upon in four separate sections below.

A. A New Primal-Dual Algorithm

In this work a new primal-dual algorithm, TriPD (Alg. 1), is introduced for solving (1). The algorithm consists of two proximal evaluations (corresponding to the two nonsmooth terms g and h), one gradient evaluation (for the smooth term f), and one correction step (*cf.* Alg. 1). We adopt the general Lipschitz continuity assumption (3) in our convergence analysis, which is essential for avoiding conservative stepsize conditions that depend on the *global* scalar Lipschitz constant.

In Section II, it is shown that the sequence generated by TriPD (Alg. 1) is S-Fejér monotone (with respect to the set of primal-dual solutions),[1] where S is a block diagonal positive definite matrix. This key property is exploited in Section III to develop a block-coordinate version of the algorithm with a general randomized activation scheme.

The connections of our method to other related primal-dual algorithms in the literature are discussed in Section II-A. Most notably, we recap the Vũ-Condat scheme [2], [3], a popular algorithm used for solving the structured optimization problem (1) (convergence of this method was established independently by Vũ [2] and Condat [3], by casting it in the form of the forward-backward splitting). In the analysis of [2], [3], a scalar constant is used to capture the Lipschitz continuity of the gradient of f, thus resulting in potentially smaller stepsizes (and slower convergence in practice). In [4], the authors assume the more general Lipschitz continuity property (3) by using a preconditioned variable metric forward-backward iteration. Nevertheless, the stepsize matrix is restricted to be proportional to Q^{-1}. In Section II-A, we show how the analysis technique for the new primal-dual algorithm can be used to recover the Vũ-Condat algorithm with general stepsize matrices, and highlight that this line of analysis leads to *less restrictive* sufficient conditions on the selected stepsizes compared to [2]–[4]. More importantly, it is shown that unlike

Puya Latafat[1,2] puya.latafat@kuleuven.be
Nikolaos M. Freris [3] nfreris2@gmail.com
Panagiotis Patrinos[1] panos.patrinos@esat.kuleuven.be

The work of the first and third authors was supported by: FWO PhD grant 1196818N; FWO research projects G086518N and G086318N; KU Leuven internal funding StG/15/043; Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160.

The work of the second author, while with New York University Abu Dhabi and New York University Tandon School of Engineering, was supported by the US National Science Foundation under grant CCF-1717207.

[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.

[3] University of Science and Technology of China, School of Computer Science and Technology, Hefei, 230000, China.

[1]Given a symmetric positive definite matrix S, we say that a sequence is S-Fejér monotone with respect to a set C if it is Fejér monotone with respect to C in the space equipped with $\langle\cdot,\cdot\rangle_S$.

TriPD (Alg. 1), the Vũ-Condat generated sequence is S-Fejér monotone, where S is *not diagonal*. As we discuss in the next subsection, this constitutes the main difficulty in devising a randomized version of the Vũ-Condat algorithm.

B. Randomized Block-Coordinate Algorithm

Block-coordinate (BC) minimization is a simple approach for tackling large-scale optimization problems. At each iteration, a subset of the coordinates is updated while others are held fixed. *Randomized* BC algorithms are of particular interest, and can be divided into two main categories:

Type a) comprises algorithms in which *only one* coordinate is randomly activated and updated at each iteration. The BC versions of gradient [5] and proximal gradient methods [6] belong in this category. A distinctive attribute of the aforementioned algorithms is the fact that the stepsizes are selected to be inversely proportional to the *coordinate-wise* Lipschitz constant of the smooth term rather than the global one. This results in applying larger stepsizes in directions with smaller Lipschitz constant, and therefore leads to faster convergence.

Type b) contains methods where *more than one* coordinate may be randomly activated and simultaneously updated [7], [8]. Note that this class may also capture the single active coordinate (type a) as a special case. The convergence condition for this class of BC algorithms is typically the same as in the full algorithm. In [7], [8] random BC is applied to α-averaged operators by establishing stochastic Fejér monotonicity, while [8] also considers quasi-nonexpansive operators. In [7], [9] the authors obtain randomized BC algorithms based on the primal-dual scheme of Vũ and Condat; the main drawback is that, just as in the full version of these algorithms, the use of conservative stepsize conditions leads to slower convergence in practice.

The BC version of TriPD (Alg. 1) falls into the second class, *i.e.*, it allows for a general randomized activation scheme (*cf.* Alg. 2). The proposed scheme converges under the same stepsize conditions as the full algorithm. As a consequence, in view of the characterization of Lipschitz continuity of ∇f in (3), when f is separable, *i.e.*, $f(x) = \sum_{i=1}^m f_i(x_i)$, our approach leads to algorithms that depend on the *local* Lipschitz constants (of ∇f_i's) rather than the global constant, thus assimilating the benefits of both categories. Notice that when f is separable, the coordinate-wise Lipschitz continuity assumption of [5], [6], [10] is equivalent to (3) with $\beta_f = 1$ and $Q = \mathbf{blkdiag}(\beta_1 I_{n_1}, \ldots, \beta_m I_{n_m})$, where m denotes the number of coordinate blocks, n_i denotes the dimension of the i-th coordinate block, and β_i denotes the Lipschitz constant of f_i. In the general setting, [5, Lem. 2] can be invoked to establish the connection between the metric Q and the coordinate-wise Lipschitz assumption. However, in many cases (most notably the separable case) this lemma is conservative.

As mentioned in the prequel, in Section II-A the Vũ-Condat algorithm is recovered using the same analysis that leads to our proposed primal-dual algorithm. It is therefore natural to consider adapting the approach of Section III so as to devise a block-coordinate variant of the the Vũ-Condat algorithm. However, this is not possible given that the Vũ-Condat generated sequence is S-Fejér monotone, where S is *not diagonal* (*cf.* (20)), while the proof of Theorem III.1 relies heavily on the diagonal structure of S. This presents a distinctive merit of our proposed algorithm over the current state-of-the-art for solving problem (1).

In [10], the authors propose a randomized BC version of the Vũ-Condat scheme. Their analysis does not require the cost functions to be separable and utilizes a different Lyapunov function for establishing convergence. Notice that the block-coordinate scheme of [10] updates a single coordinate at every iteration (*i.e.*, it is a type a) algorithm) as opposed to the more general random sweeping of the coordinates. Additionally, in the case of f being separable, our proposed method (*cf.* Alg. 2) assigns a block stepsize that is inversely proportional to $\frac{\beta_i}{2}$ (where β_i denotes the Lipschitz constant for f_i), in place of β_i required by [10, Assum. 2.1(e)]: larger stepsizes are typically associated with faster convergence in primal-dual proximal algorithms.

C. Linear Convergence

A third contribution of the paper is establishing linear convergence for the full algorithm under an additional *metric subregularity* condition for the monotone operator pertaining to the primal-dual optimality conditions (*cf.* Thm. IV.5). For the BC version, the linear rate is established under a slightly stronger condition (*cf.* Thm. IV.6). We further explicate the required condition in terms of the objective functions, with two special cases of prevalent interest: a) when f, g and h satisfy a *quadratic growth* condition (*cf.* Lem. IV.2) (which is *much weaker than strong convexity*) or b) when f, g and h are *piecewise linear-quadratic* (*cf.* Lem. IV.4), a common scenario in many applications such as LPs, QPs, SVM and fitting problems for a wide range of regularization functions; *e.g.* ℓ_1 norm, elastic nets, Huber loss and many more.

Last but not least, it is shown that the monotone operator defining the primal-dual optimality conditions is metrically subregular if and only if the residual mapping (the operator that maps z^k to $z^k - z^{k+1}$) is metrically subregular (*cf.* Lem. IV.7). This connection enables the use of Lemmas IV.2 and IV.4 to establish linear convergence for a large class of algorithms based on conditions for the cost functions.

D. Distributed Optimization

As an important application, we consider a distributed structured optimization problem over a network of agents. In this context, each agent has its own private cost function of the form (1), while the communication among agents is captured by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\begin{aligned} \underset{x_1,\ldots,x_m}{\mathbf{minimize}} \quad & \sum_{i=1}^m f_i(x_i) + g_i(x_i) + h_i(L_i x_i) \\ \mathbf{subject\ to} \quad & A_{ij}x_i + A_{ji}x_j = b_{(i,j)} \qquad (i,j) \in \mathcal{E}. \end{aligned}$$

We use (i, j) to denote the unordered pair of agents i, j, and ij to denote the ordered pair. The goal is to solve the global optimization problem through local exchange of information. Notice that the linear constraints on the edges of the graph prescribe relations between neighboring agents' variables. This type of edge constraints was also considered in [11]. It is

worthwhile noting that for the special case of two agents $i = 1, 2$, with $f_i, h_i \equiv 0$, one recovers the setup for the celebrated *alternating direction method of multipliers* (ADMM) algorithm. Another special case of particular interest is *consensus optimization*, when $A_{ij} = I$, $A_{ji} = -I$ and $b_{(i,j)} = 0$. A primal-dual algorithm for consensus optimization was introduced in [12] for the case of $f_i \equiv 0$, where a transformation was used to replace the edge variables with node variables.

This multi-agent optimization problem arises in many contexts such as sensor networks, power systems, transportation networks, robotics, water networks, distributed data-sharing, etc. [13]–[15]. In most of these applications, there are computation, communication and/or physical limitations on the system that render centralized management infeasible. This motivates the *fully* distributed synchronous and asynchronous algorithms developed in Section V. Both versions are fully distributed in the sense that not only the iterations are performed locally, but also the stepsizes of each agent are selected based on local information without any prior global coordination (*cf.* Assumption 6). The asynchronous variant of the algorithm is based on an instance of the randomized block-coordinate algorithm in Section III. The protocol is as follows: at each iteration, a) agents are activated at random, and independently from one another, b) active agents perform local updates, c) they communicate the required updated values to their neighbors and d) return to an idle state.

Notation and Preliminaries

In this section, we introduce notation and definitions used throughout the paper; the interested reader is referred to [16], [17] for more details.

For an extended-real-valued function f, we use $\mathbf{dom}\, f$ to denote its domain. For a set C, we denote its relative interior by $\mathbf{ri}\, C$. The identity matrix is denoted by $I_n \in \mathbb{R}^{n\times n}$. For a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$, 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 $\mathbb{R}^d$. We denote the domain of A by $\mathbf{dom}\, A = \{x \in \mathbb{R}^n \mid Ax \neq \emptyset\}$, its graph by $\mathbf{gra}\, A = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^d \mid y \in Ax\}$, the set of its zeros by $\mathbf{zer}\, A = \{x \in \mathbb{R}^n \mid 0 \in Ax\}$, and the set of its fixed points by $\mathbf{fix}\, A = \{x \mid x \in Ax\}$. The mapping A is called monotone if $\langle x - x', y - y'\rangle \geq 0$ for all $(x, y), (x', y') \in \mathbf{gra}\, A$, and is said to be maximally monotone if its graph is not strictly contained by the graph of another monotone operator. The inverse of A is defined through its graph: $\mathbf{gra}\, A^{-1} := \{(y, x) \mid (x, y) \in \mathbf{gra}\, A\}$. The *resolvent* of A is defined by $J_A := (\mathrm{Id} + A)^{-1}$, where Id denotes the identity operator.

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ be a proper closed, convex function. Its subdifferential is the operator $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$

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

It is well-known that the subdifferential of a convex function is maximally monotone. The resolvent of ∂f is called the *proximal operator* (or proximal mapping), and is single-valued. Let V denote a symmetric positive definite matrix. The proximal mapping of f relative to $\|\cdot\|_V$ is uniquely determined by the resolvent of $V^{-1}\partial f$:

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

The *Fenchel conjugate* of f, denoted by f^*, is defined by $f^*(v) := \sup_{x\in\mathbb{R}^n}\{\langle v, x\rangle - f(x)\}$. The *Fenchel-Young* inequality states that $\langle x, u\rangle \leq f(x) + f^*(u)$ holds for all $x, u \in \mathbb{R}^n$; in the special case when $f = \frac{1}{2}\|\cdot\|_V^2$ for some symmetric positive definite matrix V, this gives:

$$\langle x, u\rangle \leq \tfrac{1}{2}\|x\|_V^2 + \tfrac{1}{2}\|u\|_{V^{-1}}^2. \tag{2}$$

Let X be a nonempty closed convex set. The indicator of X is defined by $\delta_X(x) = 0$ if $x \in X$, and $\delta_X(x) = \infty$ if $x \notin X$. The distance from X and the projection onto X with respect to $\|\cdot\|_V$ are denoted by $d_V(\cdot, X)$ and $\mathcal{P}_X^V(\cdot)$, respectively.

We use $(\Omega, \mathcal{F}, \mathbb{P})$ for defining a probability space, where Ω, $\mathcal{F}$ and $\mathbb{P}$ denote the sample space, σ-algebra, and the probability measure. Moreover, *almost surely* is abbreviated as a.s.

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

II. A New Primal-Dual Algorithm

In this section we present a primal-dual algorithm for problem (1). We adhere to the following assumptions throughout sections II to IV:

Assumption 1.

(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.,*

$$\|\nabla f(x) - \nabla f(y)\|_{Q^{-1}} \leq \beta_f\|x - y\|_Q \quad \forall x, y \in \mathbb{R}^n. \tag{3}$$

(iii) The set of solutions to (1) *is nonempty. Moreover, there exists* $x \in \mathbf{ri}\,\mathbf{dom}\, g$ *such that* $Lx \in \mathbf{ri}\,\mathbf{dom}\, h$.

In Assumption 1*(ii)*, the constant $\beta_f \geq 0$ is not absorbed into the metric Q in order to also incorporate the case when ∇f is a constant (by setting $\beta_f = 0$).

The dual problem is to

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

With a slight abuse of terminology, we say that $(u^\star, x^\star)$ is a *primal-dual solution* (in place of dual-primal) if $u^\star$ solves the dual problem (4) and $x^\star$ solves the primal problem (1). We denote the set of primal-dual solutions by $\mathcal{S}$. Assumption 1*(iii)* guarantees that the set of solutions to the dual problem is nonempty and the duality gap is zero [18, Corollary 31.2.1]. Furthermore, the pair $(u^\star, x^\star)$ is a primal-dual solution if and only if it satisfies:

$$\begin{cases} 0 \in \partial h^*(u) - Lx, \\ 0 \in \partial g(x) + \nabla f(x) + L^\top u. \end{cases} \tag{5}$$

We proceed to present the new primal-dual scheme TriPD (Alg. 1). The motivation behind the name becomes apparent in the sequel after equation (13). The algorithm involves two proximal evaluations (respective to the non-smooth terms g, h), and one gradient evaluation (for the Lipschitz-differentiable term f). The stepsizes in TriPD (Alg. 1) are chosen so as to satisfy the following assumption:

Assumption 2 (Stepsize selection)**.** *Both the dual stepsize matrix $\Sigma \in \mathbb{R}^{r\times r}$, and the primal stepsize matrix $\Gamma \in \mathbb{R}^{n\times n}$ are symmetric positive definite. In addition, they satisfy:*

$$\Gamma^{-1} - \tfrac{\beta_f}{2} Q - L^\top \Sigma L \succ 0. \tag{6}$$

Selecting scalar primal and dual stepsizes, along with the standard definition of Lipschitz continuity, as is prevalent in the literature [2], [3], can plainly be treated by setting $\Sigma = \sigma I_r$, $\Gamma = \gamma I_n$, and $Q = I_n$, whence from (6) we require that

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

Algorithm 1 **Tri**angularly Preconditioned **P**rimal-**D**ual algorithm (TriPD)

Inputs: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$
for $k = 0, 1, \ldots$ **do**
$\bar{u}^k = \mathbf{prox}_{h^*}^{\Sigma^{-1}}(u^k + \Sigma L x^k)$
$x^{k+1} = \mathbf{prox}_{g}^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k) - \Gamma L^\top \bar{u}^k)$
$u^{k+1} = \bar{u}^k + \Sigma L (x^{k+1} - x^k)$

Remark II.1. Each iteration of TriPD (Alg. 1) requires one application of L and one of $L^\top$ (even though it appears to require two applications of L). The reason is that, at iteration k, only $L^\top \bar{u}^k$, Lx^{k+1} need to be evaluated since $L(x^{k+1} - x^k) = Lx^{k+1} - Lx^k$ and Lx^k was computed during the previous iteration.

TriPD (Alg. 1) can be compactly written as:

$$z^{k+1} = Tz^k,$$

where $z^k := (u^k, x^k)$, and the operator T is given by:

$$\bar{u} = \mathbf{prox}_{h^*}^{\Sigma^{-1}}(u + \Sigma L x) \tag{7a}$$
$$\bar{x} = \mathbf{prox}_{g}^{\Gamma^{-1}}(x - \Gamma \nabla f(x) - \Gamma L^\top \bar{u}) \tag{7b}$$
$$Tz = (\bar{u} + \Sigma L(\bar{x} - x), \bar{x}). \tag{7c}$$

Remark II.2 (Relaxed iterations)**.** It is also possible to devise a *relaxed* version of TriPD (Alg. 1) as follows:

$$z^{k+1} = z^k + \Lambda(Tz^k - z^k),$$

where Λ is a positive definite matrix and $\Lambda \prec 2I_{n+r}$. For ease of exposition, we present the convergence analysis for the original version (*i.e.*, for $\Lambda = I_{n+r}$). Note that the analysis carries through with minor modifications for relaxed iterations.

For compactness of exposition, we define the following operators:

$$A : (u, x) \mapsto (\partial h^*(u), \partial g(x)), \tag{8a}$$
$$M : (u, x) \mapsto (-Lx, L^\top u), \tag{8b}$$
$$C : (u, x) \mapsto (0, \nabla f(x)). \tag{8c}$$

The optimality condition (5) can then be written in the equivalent form of the *monotone inclusion*:

$$0 \in Az + Mz + Cz =: Fz, \tag{9}$$

where $z = (u, x)$. Observe that the linear operator M is monotone since it is skew-symmetric, *i.e.*, $M^\top = -M$. It is also easy to verify that the operator A is maximally monotone [17, Thm. 21.2 and Prop. 20.23], while operator C is cocoercive, being the gradient of $\tilde{f}(u, x) = f(x)$, and in light of Assumption 1*(ii)* and [17, Thm. 18.16].

We further define

$$P = \begin{pmatrix} \Sigma^{-1} & \frac{1}{2}L \\ \frac{1}{2}L^\top & \Gamma^{-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & -\frac{1}{2}L \\ \frac{1}{2}L^\top & 0 \end{pmatrix}, \tag{10}$$

and set $H = P + K$. It is plain to check that condition (6) implies that the symmetric matrix P is positive definite (by a standard Schur complement argument). In addition, we set

$$S = \mathbf{blkdiag}(\Sigma^{-1}, \Gamma^{-1}). \tag{11}$$

Using these definitions, the operator T defined in (7) can be written as:

$$Tz := z + S^{-1}(H + M^\top)(\bar{z} - z), \tag{12}$$

where

$$\bar{z} = (H + A)^{-1}(H - M - C)z. \tag{13}$$

This compact representation simplifies the convergence analysis. A key consideration for choosing P and K as in (10) is to ensure that $H = P + K$ is lower block-triangular. Notice that when $M \equiv 0$, (12) can be viewed as a *triangularly preconditioned* forward-backward update, followed by a correction step. This motivates the name *TriPD*: **Tri**angularly Preconditioned **P**rimal-**D**ual algorithm. Due to the triangular structure of H, the backward step $(H+A)^{-1}$ in (13) can be carried out sequentially: an updated dual vector $\bar{u}$ is computed (through proximal mapping) using (u, x) and, subsequently, the primal vector $\bar{x}$ is computed using $\bar{u}$ and x, *cf.* (7). Furthermore, it follows from (12) that this choice makes $H + M^\top$ upper block-triangular which, alongside the diagonal structure of S, yields the efficiently computable update (7c) in view of:

$$S^{-1}(H + M^\top) = \begin{pmatrix} I & \Sigma L \\ 0 & I \end{pmatrix}. \tag{14}$$

Remark II.3. The operator in (12) is inspired from [19, Alg. 1], where operators of this form were introduced for devising a splitting method for solving general monotone inclusions of the form in (9). We note, in passing, that the aforementioned algorithm entails an additional dynamic stepsize parameter (α_n, therein). Although we may also adopt this here, for potentially improving the rate of convergence in practice, we opt not to: the reason is that in the context of multi-agent optimization (that we especially target in this paper) such design choice would require global coordination, that is contradictory to our objective of devising distributed algorithms. As a positive side-effect, the convergence analysis is greatly simplified compared to [19, Sec. 5]. Besides, we use stepsize matrices (in place of scalar stepsizes) in TriPD (Alg. 1) along with the general Lipschitz continuity property (*cf.* Assumption 1*(ii)*) as an essential means for avoiding conservative stepsizes, which is especially important for large-scale distributed optimization.

We proceed by showing that the set of primal-dual solutions coincides with the set of fixed points of T, $\mathbf{fix}\, T$:

$$\mathcal{S} = \{z \mid 0 \in Az + Mz + Cz\} = \mathbf{fix}\, T. \tag{15}$$

To see this note that from (12) and (13) we have:

$$\begin{aligned} z \in \mathbf{fix}\, T &\iff z = Tz \iff \bar{z} = z \\ &\iff (H+A)^{-1}(H - M - C)z = z \\ &\iff Hz - Mz - Cz \in Hz + Az \iff z \in \mathcal{S}, \end{aligned}$$

where in the second equivalence we used the fact that S is positive definite and $\langle (H + M^\top)z, z\rangle \geq \|z\|_P^2$ for all $z \in \mathbb{R}^{n+r}$ (since K is skew-adjoint and M is monotone).

Next, let us define

$$\tilde{P} \coloneqq \begin{pmatrix} \Sigma^{-1} & -\frac{1}{2}L \\ -\frac{1}{2}L^\top & \Gamma^{-1} - \frac{\beta_f}{4}Q \end{pmatrix}. \tag{16}$$

Observe that (from Schur complement) Assumption 2 is necessary and sufficient for $2\tilde{P} - S$ to be symmetric positive definite (*cf.* to the convergence result in Thm. II.5). In particular, $\tilde{P}$ is positive definite since S is positive definite.

The next lemma establishes the key property of the operator T that is instrumental in our convergence analysis:

Lemma II.4. *Let Assumptions 1 and 2 hold. Consider the operator T in (7) (equivalently (12)). Then for any $z^\star \in \mathcal{S}$ and any $z \in \mathbb{R}^{n+r}$ we have*

$$\|Tz - z\|_{\tilde{P}}^2 \leq \langle z - z^\star, z - Tz\rangle_S. \tag{17}$$

Proof. See Appendix A. □

The next theorem establishes the main convergence result for TriPD (Alg. 1). In specific, it is shown that the generated sequence is S-Fejér monotone. We emphasize that the diagonal structure of S is the key property used in developing the block-coordinate version of the algorithm in Section III.

Theorem II.5. *Let Assumptions 1 and 2 hold. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by TriPD (Alg. 1). The following Fejér-type inequality holds for all $z^\star \in \mathcal{S}$:*

$$\|z^{k+1} - z^\star\|_S^2 \leq \|z^k - z^\star\|_S^2 - \|z^{k+1} - z^k\|_{2\tilde{P}-S}^2. \tag{18}$$

Consequently, $(z^k)_{k\in\mathbb{N}}$ converges to some $z^\star \in \mathcal{S}$.

Proof. See Appendix A. □

A. Related Primal-Dual Algorithms

Recently, the design of primal-dual algorithms for solving problem (1) (possibly with $f \equiv 0$ or $g \equiv 0$) has received a lot of attention in the literature. Most of the existing approaches can be interpreted as applications of one of the three main splittings techniques: forward-backward (FB), Douglas-Rachford (DR), and forward-backward-forward (FBF) splittings [2], [3], [20], [21], while others employ different tools to establish convergence [22], [23].

A unifying analysis for primal-dual algorithms is proposed in [19, Sec. 5], where in place of FBS, DRS, or FBFS, a new three-term splitting, namely *asymmetric forward-backward adjoint* (AFBA) is used to design primal-dual algorithms. In particular, the algorithms of [2], [3], [20]–[23] are recovered (under less restrictive stepsize conditions) and other new primal-dual algorithms are proposed. As discussed in Remark II.3 the AFBA splitting [19, Alg. 1] is the motivation behind the operator T defined in (12). We refer the reader to [19, Sec. 5] and [24] for a detailed discussion on the relation between primal-dual algorithms.

Next we briefly discuss how the celebrated algorithm of Vũ and Condat [2], [3] can be seen as fixed-point iterations of the operator T in (12) for an appropriate selection of S, P, K.

In [3] Condat considers problem (1), while Vũ [2] considers the following variant:

$$\underset{x\in\mathbb{R}^n}{\mathbf{minimize}} \quad f(x) + g(x) + (h \,\square\, l)(Lx), \tag{19}$$

where l is a strongly convex function and $\square$ represents the infimal convolution [17]. For this problem, an additional assumption is that the conjugate of l is continuously differentiable, and ∇l^* is β_l-Lipschitz continuous with respect to a metric $G \succ 0$, for some $\beta_l \geq 0$, *cf.* (3). Note that it is possible to derive and analyze a variant of TriPD (Alg. 1) for (19), however, we do not pursue this in this paper and focus on problem (1) for clarity of exposition and length considerations.

One can verify that the operator defining the fixed-point iterations in the Vũ-Condat algorithm is given by (12) with $H = P + K$ and S defined as follows:

$$S = \begin{pmatrix} \Sigma^{-1} & L \\ L^\top & \Gamma^{-1} \end{pmatrix}, \tag{20}$$

$$P = \begin{pmatrix} \Sigma^{-1} & L \\ L^\top & \Gamma^{-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & -L \\ L^\top & 0 \end{pmatrix}.$$

For such selection of S, P, K, it holds that $S^{-1}(H + M^\top) = I$, whence in proximal form, the operator defined in (12) becomes:

$$\begin{aligned} \bar{u} &= \mathbf{prox}_{h^*}^{\Sigma^{-1}}(u - \Sigma\nabla l^*(u) + \Sigma L x) \\ \bar{x} &= \mathbf{prox}_{g}^{\Gamma^{-1}}(x - \Gamma\nabla f(x) - \Gamma L^\top(2\bar{u} - u)) \\ Tz &= (\bar{u}, \bar{x}). \end{aligned}$$

Observe the non-diagonal structure of S for the Vũ-Condat algorithm in (20), in contrast with the one for TriPD (Alg. 1) in (11). For the sake of comparison with [2], [3] we consider the relaxed iteration $z^{k+1} = z^k + \lambda(Tz^k - z^k)$ for some $\lambda \in (0, 2)$, in this subsection (which we opted to exclude from TriPD (Alg. 1) solely for the purpose of simplicity).

The analysis in Theorem II.5 can be further used to establish convergence of the Vũ-Condat scheme for problem (19) under the following sufficient conditions (in place of Assumption 2):

$$\Sigma^{-1} - \tfrac{\beta_l}{2(2-\lambda)}G \succ 0, \tag{21a}$$

$$\Gamma^{-1} - \tfrac{\beta_f}{2(2-\lambda)}Q - L^\top\left(\Sigma^{-1} - \tfrac{\beta_l}{2(2-\lambda)}G\right)^{-1}L \succ 0. \tag{21b}$$

Notice that when $l = \delta_{\{0\}}$ (*i.e.*, for problem (1)), $l^* \equiv 0$ whence $\beta_l = 0$, and the condition simplifies to:

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

Given the stepsize condition (21) the following Fejér-type inequality holds.

$$\|z^{k+1} - z^\star\|_S^2 \leq \|z^k - z^\star\|_S^2 - \lambda\|z^{k+1} - z^k\|_{2\hat{P}-\lambda S}^2, \tag{22}$$

with S defined in (20) and $\hat{P}$ given by:

$$\hat{P} \coloneqq \begin{pmatrix} \Sigma^{-1} - \frac{\beta_l}{4}G & L \\ L^\top & \Gamma^{-1} - \frac{\beta_f}{4}Q \end{pmatrix}.$$

This generalizes the result in [3, Thm. 3.1], [2, Cor. 4.2] and [19, Prop. 5.1] where $Q = I$ and the stepsizes are assumed to be scalar.

Our main goal here was to demonstrate the non-diagonal structure of S for the Vũ-Condat algorithm. In the sequel, we highlight that our analysis additionally leads to less conservative conditions as compared to [2]–[4]. Notice that the proofs in the aforementioned papers are based on casting the algorithm in the form of forward-backward iterations. Consequently, the stepsize condition obtained ensures that the underlying operator is *averaged*. In contradistinction, the sufficient condition in (21) only ensures that the Fejér-type inequality (22) holds, which is sufficient for convergence. Therefore, even in the case of scalar stepsizes (as in [2], [3]) condition (21) allows for larger stepsizes compared to [2], [3].

In [4], [9] the authors propose a variable metric version of the algorithm with a preconditioning that accounts for the general Lipschitz metric. This is accomplished by fixing the stepsize matrix to be a constant times the inverse of the Lipschitz metric, and obtaining a condition on the constant. Our approach does not assume this restrictive form for the stepsize matrix; even when such a restriction is imposed it allows for *larger* stepsizes, thus achieving generally faster convergence. As an illustrative example, let us set $\Gamma = \mu Q^{-1}$ and $\Sigma = \nu G^{-1}$ for some $\mu, \nu > 0$. For simplicity and without loss of generality, let $\beta_l = 1$, $\beta_f = 1$. Then (21) simplifies to:

$$(\mu^{-1} - \tfrac{1}{2(2-\lambda)})(\nu^{-1} - \tfrac{1}{2(2-\lambda)})Q - L^\top G^{-1} L \succ 0, \quad (23)$$

whereas the condition required in [4], [9] is $\lambda \in (0, 1]$ and

$$\frac{\delta}{1+\delta} > \frac{\mathbf{max}\{\mu,\nu\}}{2} \text{ with } \delta = \tfrac{1}{\sqrt{\nu\mu}}\|G^{-1/2} L Q^{-1/2}\|^{-1} - 1. \quad (24)$$

It is not difficult to check that condition, (23), is always less restrictive than (24). For instance, let $G^{-1/2} L Q^{-1/2} = I$ and set $\mu = 1.5$, then (23) requires that $\nu < \frac{1}{6.5}$ whereas (24) necessitates that $\nu < \frac{1}{24}$.

III. A Randomized Block-Coordinate Algorithm

In this section, we describe a randomized block-coordinate variant of TriPD (Alg. 1) and discuss important special cases pertaining to the randomized coordinate activation mechanism. The convergence analysis is based on establishing stochastic Fejér monotonicity [8] of the generated sequence. In addition, we establish linear convergence of the method under further assumptions in Section IV.

First, let us define a partitioning of the vector of primal-dual variables into m blocks of coordinates. Notice that each block might include a subset of primal or dual variables, or a combination of both. Respectively, let $U_i \in \mathbb{R}^{(n+r)\times(n+r)}$, for $i = 1, \ldots, m$, be a diagonal matrix with 0-1 diagonal entries that is used to select a subset of the coordinates (selected coordinates correspond to diagonal entries equal to 1). We call such matrix an *activation matrix*, as it is used to activate/select a subset of coordinates to update.

Let $\Phi = \{0, 1\}^m$ denote the set of binary strings of length m (with the elements considered as column vectors of dimension m). At the k-th iteration, the algorithm draws a Φ-valued random activation vector ϵ^{k+1} which determines which *blocks of coordinates* will be updated. The i-th element of the vector ϵ^{k+1} is denoted as ϵ_i^{k+1}: the i-th block is updated at iteration k if $\epsilon_i^{k+1} = 1$. Notice that in general multiple blocks of coordinates may be concurrently updated. The conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_k]$ is abbreviated by $\mathbb{E}_k[\cdot]$, where $\mathcal{F}_k$ is the filtration generated by $(\epsilon^1, \ldots, \epsilon^k)$. The following assumption summarizes the setup of the randomized coordinate selection.

Assumption 3.

(i) $\{U_i\}_{i=1}^m$ *are* 0-1 *diagonal matrices and* $\sum_{i=1}^m U_i = I$.

(ii) $(\epsilon^k)_{k\in\mathbb{N}}$ *is a sequence of i.i.d.* Φ*-valued random vectors with*

$$p_i := \mathbb{P}(\epsilon_i^1 = 1) > 0 \quad i = 1, \ldots, m. \quad (25)$$

(iii) The stepsize matrices Σ, Γ *are diagonal.*

The first condition implies that the activation matrices define a *partition* of the coordinates, while the second that each partition is activated with a positive probability.

We further define the (diagonal) coordinate activation probability matrix Π as follows:

$$\Pi := \sum_{i=1}^m p_i U_i. \quad (26)$$

For $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ we define the operator $\hat{T}^{(\epsilon)}$ by:

$$\hat{T}^{(\epsilon)} z := z + \sum_{i=1}^m \epsilon_i U_i (Tz - z),$$

where T was defined in (7) (equivalently (12)). Observe that this is a compact notation for the update of only the selected blocks. The randomized scheme is then written as an iterative application of $\hat{T}^{(\epsilon^{k+1})}$ for $k = 0, 1, \ldots$ (this operator updates the active blocks of coordinates and leaves the others unchanged, *i.e.*, equal to their previous iterate values). The randomized block-coordinate scheme is summarized below.

Algorithm 2 Block-coordinate TriPD algorithm

Inputs: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$
for $k = 0, 1, \ldots$ **do**
  Select Φ-valued r.v. ϵ^{k+1}
  $z^{k+1} = \hat{T}^{(\epsilon^{k+1})} z^k$

We emphasize that the randomized model that we adopt here is capable of capturing many stationary randomized activation mechanisms. To illustrate this, consider the following activation mechanisms (of specific interest in the realm of distributed multi-agent optimization, *cf.* Section V):

- *Multiple coordinate activation*: at each iteration, the j-th coordinate block is randomly activated with probability $p_j > 0$ independent of other coordinates blocks. This corresponds to the case that the sample space is equal to $\Phi = \{0, 1\}^m$. The general distributed algorithm of Section V assumes this mechanism.
- *Single coordinate activation*: at each iteration, one coordinate block is selected, *i.e.*, the sample space is

$$\{(1,0,\ldots,0),(0,1,0,\ldots,0)\ldots,(0,\ldots,0,1)\}. \quad (27)$$

We assign probability p_i to the event $\epsilon_i = 1$ (and $\epsilon_j = 0$ for $j \neq i$), whence the probabilities must satisfy $\sum_{i=1}^m p_i = 1$.

The next lemma establishes stochastic Fejér monotonicity for the generated sequence, by directly exploiting the diagonal structure of S. The proof technique is adapted from [7, Thm. 3] (see also [25, Thm. 2], [8, Thm. 2.5]), and is based on the Robbins-Siegmund lemma [26].

Theorem III.1. *Let Assumptions 1 to 3 hold. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by TriPD-BC (Alg. 2). The following Fejér-type inequality holds for all $z^\star \in \mathcal{S}$:*

$$\mathbb{E}_k\big[\|z^{k+1}-z^\star\|^2_{\Pi^{-1}S}\big] \le \|z^k-z^\star\|^2_{\Pi^{-1}S} - \|Tz^k-z^k\|^2_{2\tilde{P}-S}. \tag{28}$$

Consequently, $(z^k)_{k\in\mathbb{N}}$ converges a.s. to some $z^\star \in \mathcal{S}$.

Proof. See Appendix A. □

It is important to emphasize that a naive implementation of TriPD-BC (Alg. 2) (with regards to the partitioning of primal-dual variables) may involve wasteful computations. As an example, consider a BC algorithm in which, at every iteration, either all primal or all dual variables are updated. In such a case, if at iteration k the dual vector is to be updated, both x^{k+1}, u^{k+1} are computed (*cf.* Alg. 1), whereas only u^{k+1} is updated. This phenomenon is common to all primal-dual algorithms, and is due to the fact that the primal and dual updates need to be performed sequentially in the full version of the algorithm. As a consequence, the blocks of coordinates must be partitioned in such a way that computations are not discarded, so that the iteration cost of a BC algorithm is (substantially) smaller than computing the full operator T. This choice relies entirely on the structure of the optimization problem under consideration. A canonical example of prominent practical interest is the setting of multi-agent optimization in a network (*cf.* §V), where L is not diagonal, f and g are separable, and additional coupling between (primal) coordinates is present through h, see (32). In this example, the primal and dual coordinates are partitioned in such a way that no computation is discarded (*cf.* §V for more details).

We proceed with another example where the coordinates may be grouped such that the BC algorithm does not incur any wasteful computations: consider problem (1) with $Lx = \mathbf{blkdiag}(L_1x_1,\dots,L_mx_m)$, and g, h separable functions *i.e.*,

$$\underset{x\in\mathbb{R}^n}{\mathbf{minimize}} \quad f(x)+\sum_{i=1}^m \big(g_i(x_i)+h_i(L_ix_i)\big).$$

In this problem, the coupling between the (primal) coordinates is carried via function f. For each $i=1,\dots,m$, we can choose U_i such that it selects the i-th primal-dual coordinate block (u_i,x_i). Under such partitioning of coordinates, one may use TriPD-BC (Alg. 2) with any random activation pattern satisfying Assumption 3. For example, for the case of multiple independently activated coordinates, as discussed above, at iteration k the following is performed

$$\begin{cases} \bullet \text{ each block } (u_i,x_i) \text{ is activated with probability } p_i>0 \\ \bullet \text{ for active block(s) } i \text{ compute:} \\ \bar{u}_i^k = \mathbf{prox}_{\sigma h_i^*}(u_i^k+\sigma L_i x_i^k) \\ x_i^{k+1} = \mathbf{prox}_{\gamma g_i}(x_i^k-\gamma\nabla_i f(x^k)-\gamma L_i^\top \bar{u}_i^k) \\ u_i^{k+1} = \bar{u}_i^k+\sigma L_i(x_i^{k+1}-x_i^k). \end{cases}$$

More generally, when g and h are separable in problem (1), and L is such that either each (block) row only has one nonzero element or each (block) column has one nonzero element, then the coordinates can be grouped together in such a way that no wasteful computations occur: in the first case the primal vector x_i and all dual vectors u_j that are required for its computation are selected by U_i (with the role of primal and dual reversed in the second case).

Remark III.2. Note that in TriPD-BC (Alg. 2) the probabilities p_i are taken fixed, *i.e.*, the matrix Π is constant throughout the iterations. This is a non-restrictive assumption and can be relaxed by considering iteration-varying probabilities p_i^k in (25) and modifying TriPD-BC (Alg. 2) by setting:

$$z^{k+1} = z^k + \sum_{i=1}^m \frac{\epsilon_i^{k+1}}{mp_i^{k+1}} U_i(Tz^k-z^k).$$

Let Π^k denote the probability matrix defined as in (26) using p_i^k. Then, by arguing as in Theorem III.1, it can be shown that the following stochastic Fejér monotonicity holds for the modified sequence:

$$\mathbb{E}_k\big[\|z^{k+1}-z^\star\|^2_S\big] \le \|z^k-z^\star\|^2_S - \|Tz^k-z^k\|^2_{\frac{2}{m}\tilde{P}-\frac{1}{m^2}S(\Pi^{k+1})^{-1}}.$$

IV. Linear Convergence

In this section, we establish linear convergence of Algorithms 1 and 2 under additional conditions on the cost functions f, g and h. To this end, we show that linear convergence is attained if the monotone operator $F=A+M+C$ defining the primal-dual optimality conditions (*cf.* (9)) is *metrically subregular* (globally metrically subregular in the case of TriPD-BC (Alg. 2)). A notable consequence of our analysis is the fact that linear convergence is attained when the cost functions either a) belong in the class of *piecewise linear-quadratic* (PLQ) convex functions or b) when they satisfy a certain *quadratic growth* condition (which is much weaker than strong convexity). Moreover, notice that in the case of PLQ the solution need not be unique (*cf.* Thm.s IV.5 and IV.6).

We first recall the notion of *metric subregularity* [27].

Definition IV.1 (Metric subregularity)**.** *A set-valued mapping $F:\mathbb{R}^n \rightrightarrows \mathbb{R}^d$ is* metrically subregular *at $\bar{x}$ for $\bar{y}$ if $(\bar{x},\bar{y}) \in \mathbf{gra}\, F$ and there exists a positive constant η together with a* neighborhood of subregularity *$\mathcal{U}$ of $\bar{x}$ such that*

$$d(x,F^{-1}\bar{y}) \le \eta d(\bar{y},Fx) \;\; \forall x\in\mathcal{U}.$$

If the following stronger condition holds

$$\|x-\bar{x}\| \le \eta d(\bar{y},Fx) \;\; \forall x\in\mathcal{U},$$

then F is said to be strongly subregular *at $\bar{x}$ for $\bar{y}$.*

Moreover, we say that F is globally (strongly) subregular at $\bar{x}$ for $\bar{y}$ if (strong) subregularity holds with $\mathcal{U}=\mathbb{R}^n$.

We refer the reader to [16, Chap. 9], [27, Chap. 3] and [28, Chap. 2] for further discussion on metric subregularity.

Metric subregularity of the subdifferential operator has been studied thoroughly and is equivalent to the *quadratic growth condition* [29], [30] defined next. In particular, for a proper

closed convex function f, the subdifferential ∂f is metrically subregular at $\bar{x}$ for $\bar{y}$ with $(\bar{x}, \bar{y}) \in \mathbf{gra}\, \partial f$ if and only if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that the following growth condition holds [29, Thm. 3.3]:

$$f(x) \geq f(\bar{x}) + \langle \bar{y}, x - \bar{x} \rangle + c d^2(x, (\partial f)^{-1}(\bar{y})) \quad \forall x \in \mathcal{U}$$

Furthermore, ∂f is strongly subregular at $\bar{x}$ for $\bar{y}$ with $(\bar{x}, \bar{y}) \in \mathbf{gra}\, \partial f$, if and only if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that [29, Thm. 3.5]:

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

Note that strongly convex functions satisfy (29), but (29) is *much weaker* than strong convexity, as it is a *local condition*: it only holds in a neighborhood of $\bar{x}$, and also only for $\bar{y}$.

The lemma below provides a sufficient condition for metric subregularity of the monotone operator $A + M + C$, in terms of strong subregularity of $\nabla f + \partial g$ and ∂h^* (equivalently the quadratic growth of $f + g$ and h^*, *cf.* (29)) as stated in the following assumption:

Assumption 4 (Strong subregularity of $\nabla f + \partial g$ and ∂h^*)**.** *There exists $z^\star = (u^\star, x^\star) \in \mathcal{S}$ satisfying:*

(i) $\nabla f + \partial g$ is strongly subregular at $x^\star$ for $-L^\top u^\star$,

(ii) ∂h^* is strongly subregular at $u^\star$ for $Lx^\star$.

We say that f, g and h satisfy this assumption globally *if the strong subregularity assumption of $\nabla f + \partial g$ and ∂h^* both hold globally (cf. Definition IV.1).*

In particular, Assumption 4 holds globally if either f or g (or both) are strongly convex and h is continuously differentiable with Lipschitz continuous gradient, *i.e.*, h^* is strongly convex.

Lemma IV.2. *Let Assumptions 1 and 4 hold. Then $F = A + M + C$ (cf. (8)) is strongly subregular at $z^\star$ for 0. Moreover, if f, g and h satisfy Assumption 4* globally*, then F is globally strongly subregular at $z^\star$ for 0. In both cases the set of primal-dual solutions is a singleton, $\mathcal{S} = \{z^\star\}$.*

Proof. See Appendix A. □

Our next objective is to show that $A + M + C$ is globally metrically subregular when the functions f, g and h are *piecewise linear-quadratic* (PLQ). Note that this assumption does not imply that the set of solutions $\mathcal{S}$ is a singleton, nevertheless, linear convergence can still be established. Let us recall the definition of PLQ functions [16]:

Definition IV.3 (Piecewise linear-quadratic)**.** *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 $\frac{1}{2}\langle x, Qx \rangle + \langle d, x \rangle + c$, for some $c \in \mathbb{R}$, $d \in \mathbb{R}^n$, and symmetric matrix $Q \in \mathbb{R}^{n \times n}$.*

The class of PLQ functions is closed under scalar multiplication, addition, conjugation and Moreau envelope [16]. 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 (*e.g.*, the ℓ_1-norm), and regularizing functions such as elastic net, Huber loss, hinge loss, to name a few.

Lemma IV.4. *Let Assumption 1 hold. In addition, assume that f, g and h are piecewise linear-quadratic. Then $F = A + M + C$ (cf. (8)) is metrically subregular with the same constant η at any z for any v with $(z, v) \in \mathbf{gra}\, F$.*

Proof. See Appendix A. □

Our main convergence rate results are provided in Theorems IV.5 and IV.6. In this context, Lemmas IV.2 and IV.4 are used to establish sufficient conditions in terms of the cost functions. We omit the proof of Theorem IV.5 for length considerations. The proof is similar to that of Theorem IV.6, the main difference being that in Theorem IV.5 *local* (as opposed to global) metric subregularity is used: due to the Fejér-type inequality (18), $\bar{z}^k$ will eventually be contained in a neighborhood of metric subregularity, where inequality (53) applies.

Theorem IV.5 (Linear convergence of Alg. 1)**.** *Consider TriPD (Alg. 1) under the assumptions of Theorem II.5. Suppose that $F = A + M + C$ is metrically subregular at all $z^\star \in \mathcal{S}$ for 0. Then $(d_S(z^k, \mathcal{S}))_{k \in \mathbb{N}}$ converges Q-linearly to zero, and $(z^k)_{k \in \mathbb{N}}$ converges R-linearly to some $z^\star \in \mathcal{S}$.*

In particular, the metric subregularity assumption holds and the result follows if either one of the following holds:

(i) either f, g and h are PLQ,

(ii) or f, g and h satisfy Assumption 4, in which case the solution is unique.

Theorem IV.6 (Linear convergence of Alg. 2)**.** *Consider TriPD-BC (Alg. 2) under the assumptions of Theorem III.1. Suppose that $F = A + M + C$ is globally metrically subregular for 0 (cf. Def. IV.1), i.e., there exists $\eta > 0$ such that*

$$d(z, F^{-1}0) \leq \eta d(0, Fz) \quad \forall z \in \mathbb{R}^{n+r}.$$

Then $(\mathbb{E}\left[d^2_{\Pi^{-1}S}(z^k, \mathcal{S})\right])_{k \in \mathbb{N}}$ converges Q-linearly to zero. The same holds if

(i) either f, g, h are PLQ and there exists a compact set $\mathcal{C}$ such that $(z^k)_{k \in \mathbb{N}} \subseteq \mathcal{C}$ (as is the case if $\mathbf{dom}\, g$ and $\mathbf{dom}\, h^*$ are compact),

(ii) or f, g and h satisfy Assumption 4 globally, in which case the solution is unique.

Proof. See Appendix A. □

In the recent work [31] the authors establish linear convergence in the framework of non-expansive operators under the assumption that the residual mapping defined as $R = \mathrm{Id} - T$ is metrically subregular. However, such a condition is not easily verifiable in terms of conditions on the cost functions. In the next lemma, we show that R is metrically subregular if and only if the monotone operator F is metrically subregular. This result connects the two assumptions and is interesting in its own right. More importantly, it enables the use of Lemmas IV.2 and IV.4 for establishing linear convergence for a wide array of problems.

Lemma IV.7. *Let Assumptions 1 and 2 hold. Consider the operator T defined in (12) and a point $z^\star \in \mathcal{S}$. Then $F = A + M + C$ (cf. (8)) is metrically subregular at $z^\star$ for 0 if and only if the residual mapping $R \coloneqq \mathrm{Id} - T$ is metrically subregular at $z^\star$ for 0.*

Proof. See Appendix A. □

V. Distributed Optimization

In this section, we consider a general formulation for multi-agent optimization over a network, and leverage Algorithms 1 and 2 to devise both synchronous and randomized asynchronous distributed primal-dual algorithms. The setting is as follows. We consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over a vertex set $\mathcal{V} = \{1, \dots, m\}$ with edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Each vertex is associated with a corresponding *agent*, which is assumed to have a local memory and computational unit, and can only communicate with its neighbors. We define the *neighborhood* of agent i by $\mathcal{N}_i := \{j | (i,j) \in \mathcal{E}\}$. We use the terms vertex, agent, and node interchangeably. The goal is to solve the following global optimization problem in a distributed fashion:

$$\underset{x_1,\dots,x_m}{\textbf{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) + g_i(x_i) + h_i(L_i x_i) \tag{30a}$$

$$\textbf{subject to} \quad A_{ij}x_i + A_{ji}x_j = b_{(i,j)} \qquad (i,j) \in \mathcal{E}, \tag{30b}$$

where $x_i \in \mathbb{R}^{n_i}$. The cost functions f_i, g_i, $h_i \circ L_i$ are taken *private* to agent/node $i \in \mathcal{V}$, *i.e.*, our distributed methods operate solely by exchanging local variables among neighboring nodes that are unaware of each other's objectives. The coupling in the problem is represented through the edge constraints (30b).

Throughout this section the following assumptions hold:

Assumption 5. *For each $i = 1, \dots, m$:*

(i) For $j \in \mathcal{N}_i$, $b_{(i,j)} \in \mathbb{R}^{l_{(i,j)}}$ and $A_{ij} \in \mathbb{R}^{n_i} \to \mathbb{R}^{l_{(i,j)}}$ is a linear mapping.

(ii) $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$, $h_i : \mathbb{R}^{r_i} \to \overline{\mathbb{R}}$ are proper closed convex functions, and $L_i : \mathbb{R}^{n_i} \to \mathbb{R}^{r_i}$ is a linear mapping.

(iii) $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is convex, continuously differentiable, and for some $\beta_i \in [0, \infty)$, ∇f_i is β_i-Lipschitz continuous with respect to the metric $Q_i \succ 0$, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\|_{Q_i^{-1}} \leq \beta_i \|x - y\|_{Q_i} \quad x, y \in \mathbb{R}^{n_i}.$$

(iv) The graph $\mathcal{G}$ is connected.

(v) The set of solutions of (30) is nonempty. Moreover, there exists $x_i \in \mathbf{ri\,dom}\, g_i$ such that $L_i x_i \in \mathbf{ri\,dom}\, h_i$, for $i = 1, \dots, m$, and $A_{ij}x_i + A_{ji}x_j = b_{(i,j)}$ for $(i,j) \in \mathcal{E}$.

Each agent $i \in \mathcal{V}$ maintains its own local primal variable $x_i \in \mathbb{R}^{n_i}$ and dual variables $y_i \in \mathbb{R}^{r_i}$, and $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$ (for each $j \in \mathcal{N}_i$), where the former is related to the linear mapping L_i, and the latter is the local dual variable of agent i corresponding to the edge-constraint (30b). It is important to note that the updates in TriPD-Dist (Alg. 3) are performed locally through communication with neighbors: the only information that agent i shares with its neighbor $j \in \mathcal{N}_i$ is the quantity $A_{ij}x_i$, along with edge variable $w_{(i,j),i}$, while all other variables are kept *private*.

The proposed distributed protocol features both a synchronous as well as an asynchronous implementation. In the synchronous version, at every iteration, all the agents update their variables. In the randomized asynchronous implementation, only a subset of randomly activated agents perform updates, at each iteration, and they do so using their local variables as well as information previously communicated to them by their neighbors. After an update is performed, in both cases, updated values are communicated to neighboring agents. Notice that the asynchronous scheme corresponds to the case of multiple coordinate blocks activation in TriPD-BC (Alg. 2). Other activation schemes can also be considered, and our convergence analysis plainly carries over; notably, the single agent activation which corresponds to the asynchronous model of [32]–[34] in which agents are assumed to 'wake-up' based on independent exponentially distributed tick-down timers.

Furthermore, in TriPD-Dist (Alg. 3) each agent i keeps positive local stepsizes σ_i, τ_i and $\left(\kappa_{(i,j)}\right)_{j \in \mathcal{N}_i}$. The edge weights/stepsizes $\kappa_{(i,j)}$ may alternatively be interpreted as inherent parameters of the communication graph. For example, they may be used to capture edge's 'fidelity,' *e.g.*, the channel quality in a communication link. The stepsizes are assumed to satisfy the following *local* assumption that is sufficient for the convergence of the algorithm (*cf.* Thm.s V.1 and V.2).

Assumption 6 (Stepsizes of TriPD-Dist (Alg. 3))**.**

(i) (node stepsizes) Each agent i keeps two positive stepsizes σ_i, τ_i.

(ii) (edge stepsizes) A positive stepsize $\kappa_{(i,j)}$ is associated with edge $(i,j) \in \mathcal{E}$, and is shared between agents i, j.

(iii) (convergence condition) The stepsizes satisfy the following local *condition*

$$\tau_i < \frac{1}{\frac{\beta_i \|Q_i\|}{2} + \|\sigma_i L_i^\top L_i + \sum_{j \in \mathcal{N}_i} \kappa_{(i,j)} A_{ij}^\top A_{ij}\|}.$$

According to Assumption 6*(iii)* the stepsizes τ_i, σ_i for each agent only depend on the local parameters β_i, $\|Q_i\|$, the edge weights, $\kappa_{(i,j)}$ and the linear mappings L_i, and A_{ij}, which are all known to agent i; therefore the stepsizes can be selected locally, in a decentralized fashion.

We proceed by casting the multi-agent optimization problem (30) in the form of the structured optimization problem (1). In doing so, we describe how TriPD-Dist (Alg. 3) is derived as an instance of Algorithms 1 and 2.

Define the linear operator

$$N_{(i,j)} : \mathsf{x} \mapsto (A_{ij}x_i, A_{ji}x_j),$$

and $\mathsf{N} \in \mathbb{R}^{2\sum_{(i,j)\in\mathcal{E}} l_{(i,j)} \times \sum_{i=1}^m n_i}$ by stacking $N_{(i,j)}$:

$$\mathsf{N} : \mathsf{x} \mapsto (N_{(i,j)}\mathsf{x})_{(i,j)\in\mathcal{E}}.$$

Its transpose is given by:

$$\mathsf{N}^\top : (w_{(i,j)})_{(i,j)\in\mathcal{E}} \mapsto \tilde{\mathsf{x}} = \sum_{(i,j)\in\mathcal{E}} N_{(i,j)}^\top w_{(i,j)},$$

with $\tilde{x}_i = \sum_{j \in \mathcal{N}_i} A_{ij}^\top w_{(i,j),i}$. We have set $w_{(i,j)} = (w_{(i,j),i}, w_{(i,j),j})$, *i.e.*, we consider two dual variables (of dimension $l_{(i,j)}$) for each edge constraint, where $w_{(i,j),i}$ is maintained by agent i and $w_{(i,j),j}$ by agent j.

Consider the set

$$C_{(i,j)} = \{(z_1, z_2) \in \mathbb{R}^{l_{(i,j)}} \times \mathbb{R}^{l_{(i,j)}} \mid z_1 + z_2 = b_{(i,j)}\}.$$

Then problem (30) can then be re-written as:

$$\textbf{minimize} \quad \sum_{i=1}^{m} f_i(x_i) + g_i(x_i) + h_i(L_i x_i) + \sum_{(i,j)\in\mathcal{E}} \delta_{C_{(i,j)}}(N_{(i,j)}\mathsf{x}) \tag{31}$$

Algorithm 3 Synchronous & asynchronous versions of TriPD-Dist algorithm

Inputs: $x_i^0 \in \mathbb{R}^{n_i}$, $y_i^0 \in \mathbb{R}^{r_i}$, for $i = 1, \dots, m$, and $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$ for $j \in \mathcal{N}_i$.

for $k = 0, 1, \dots$ **do**

I: Synchronous version

for all agents $i = 1, \dots, m$ **do**

II: Asynchronous version

Each agent $i = 1. \dots, m$ is activated independently with probability $p_i > 0$

for all active agents **do**

Local updates:

$$\begin{aligned}
\bar{w}^k_{(i,j),i} &= \tfrac{1}{2}\big(w^k_{(i,j),i} + w^k_{(i,j),j}\big) + \tfrac{\kappa_{(i,j)}}{2}\big(A_{ij}x_i^k + A_{ji}x_j^k - b_{(i,j)}\big), \quad \forall j \in \mathcal{N}_i \\
\bar{y}_i^k &= \mathbf{prox}_{\sigma_i h_i^\star}\big(y_i^k + \sigma_i L_i x_i^k\big) \\
x_i^{k+1} &= \mathbf{prox}_{\tau_i g_i}\big(x_i^k - \tau_i L_i^\top \bar{y}_i^k - \tau_i \textstyle\sum_{j\in\mathcal{N}_i} A_{ij}^\top \bar{w}^k_{(i,j),i} - \tau_i \nabla f_i(x_i^k)\big) \\
y_i^{k+1} &= \bar{y}_i^k + \sigma_i L_i (x_i^{k+1} - x_i^k) \\
w^{k+1}_{(i,j),i} &= \bar{w}^k_{(i,j),i} + \kappa_{(i,j)} A_{ij}(x_i^{k+1} - x_i^k), \quad \forall j \in \mathcal{N}_i
\end{aligned}$$

Transmission of information:

Send $A_{ij}x_i^{k+1}$, $w^{k+1}_{(i,j),i}$ to agent j, $\forall j \in \mathcal{N}_i$

Let $C = \bigtimes_{(i,j)\in\mathcal{E}} C_{(i,j)}$, $L = \mathbf{blkdiag}(L_1, \dots, L_m)$, and $\mathsf{Lx} = (L\mathsf{x}, \mathsf{Nx}) =: (\tilde{\mathsf{y}}, \tilde{\mathsf{w}}) \in \mathbb{R}^{n_d}$ with $n_d = 2\sum_{(i,j)\in\mathcal{E}} l_{(i,j)} + \sum_{i=1}^m r_i$, and rewrite (31) in the following compact form:

$$\mathbf{minimize} \quad \mathsf{f}(\mathsf{x}) + \mathsf{g}(\mathsf{x}) + \tilde{\mathsf{h}}(\mathsf{Lx}), \tag{32}$$

where $\mathsf{f}(\mathsf{x}) = \sum_{i=1}^m f_i(x_i)$, $\mathsf{g}(\mathsf{x}) = \sum_{i=1}^m g_i(x_i)$, $\tilde{\mathsf{h}}(\tilde{\mathsf{y}}, \tilde{\mathsf{w}}) = \mathsf{h}(\tilde{\mathsf{y}}) + \delta_C(\tilde{\mathsf{w}})$, $\mathsf{h}(\tilde{\mathsf{y}}) = \sum_{i=1}^m h_i(\tilde{y}_i)$.

In what follows, $\mathcal{S}$ refers to the set of primal-dual solutions of (32). As in Section II, the primal-dual optimality conditions can be written in the form of monotone inclusion (9) with

$$\begin{aligned}
A &: (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (\partial \mathsf{h}^*(\mathsf{y}), \partial \delta_C^*(\mathsf{w}), \partial \mathsf{g}(\mathsf{x})), \\
M &: (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (-L\mathsf{x}, -\mathsf{Nx}, L^\top \mathsf{y} + \mathsf{N}^\top \mathsf{w}), \\
C &: (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (0, 0, \nabla \mathsf{f}(\mathsf{x})),
\end{aligned}$$

where $\mathsf{u} = (\mathsf{y}, \mathsf{w})$ represents the dual vector.

We define the edge weight matrix as follows

$$W = \mathbf{blkdiag}\left((\kappa_{(i,j)} I_{2l_{(i,j)}})_{(i,j)\in\mathcal{E}}\right),$$

where the weights $\kappa_{(i,j)}$ are repeated twice (for each of the two neighboring agents). Furthermore, we set

$$\begin{aligned}
\Sigma &= \mathbf{blkdiag}(\sigma_1 I_{r_1}, \dots, \sigma_m I_{r_m}, W), \\
\Gamma &= \mathbf{blkdiag}(\tau_1 I_{n_1}, \dots, \tau_m I_{n_m}), \\
Q &= \mathbf{blkdiag}(\beta_1 Q_1, \dots, \beta_m Q_m).
\end{aligned}$$

Since $\mathbf{prox}_{\tilde{\mathsf{h}}^\star}(\mathsf{y}, \mathsf{w}) = (\mathbf{prox}_{\mathsf{h}^\star}(\mathsf{y}), \mathsf{w} - \mathcal{P}_C(\mathsf{w}))$ (using $\mathbf{prox}_{\delta_C}(\cdot) = \mathcal{P}_C(\cdot)$ along with Moreau decomposition [17, Thm. 14.3]) the proximal updates of TriPD (Alg. 1), *cf.* (7), become:

$$\begin{aligned}
\bar{y}_i &= \mathbf{prox}_{\sigma_i h_i^\star}(y_i + \sigma_i L_i x_i), \\
\bar{w}_{(i,j)} &= w_{(i,j)} + \kappa_{(i,j)}(N_{(i,j)}\mathsf{x} - \mathcal{P}_{C_{(i,j)}}(\kappa_{(i,j)}^{-1} w_{(i,j)} + N_{(i,j)}\mathsf{x})), \\
\bar{x}_i &= \mathbf{prox}_{\tau_i g_i}(x_i - \tau_i L_i^\top \bar{y}_i - \tau_i (\mathsf{N}^\top \bar{\mathsf{w}})_i - \tau_i \nabla f(x_i)).
\end{aligned}$$

Note that for $w_1, w_2 \in \mathbb{R}^{l_{(i,j)}}$ the projection onto $C_{(i,j)}$ is

$$\mathcal{P}_{C_{(i,j)}}(w_1, w_2) = \frac{1}{2}\big(w_1 - w_2 + b_{(i,j)}, -w_1 + w_2 + b_{(i,j)}\big).$$

By assigning to agent i the primal coordinate x_i and dual coordinate y_i and $w_{(i,j),i}$ for all $j \in \mathcal{N}_i$, TriPD-Dist (Alg. 3) is obtained. Note that this assignment entails non-overlapping sets of coordinates, *i.e.*, Assumption 3*(i)* is satisfied.

The convergence results of TriPD-Dist (Alg. 3) are provided separately for the synchronous and asynchronous schemes in the next two theorems, along with a sufficient condition for linear convergence. The proofs follow directly from Theorems IV.5 and IV.6.

Theorem V.1 (Convergence of Algorithm 3-I)**.** *Let Assumptions 5 and 6 hold. The sequence* $(\mathsf{z}^k)_{k\in\mathbb{N}} = (\mathsf{y}^k, \mathsf{w}^k, \mathsf{x}^k)_{k\in\mathbb{N}}$ *generated by Algorithm 3-I converges to some* $z^\star \in \mathcal{S}$*. Furthermore, if* f_i*,* g_i *and* h_i*,* $i = 1, \dots, m$ *are PLQ, then* $(d_S(\mathsf{z}^k, \mathcal{S}))_{k\in\mathbb{N}}$ *converges Q-linearly to zero, and* $(\mathsf{z}^k)_{k\in\mathbb{N}}$ *converges R-linearly to* $z^\star \in \mathcal{S}$*.*

Theorem V.2 (Convergence of Algorithm 3-II)**.** *Let Assumptions 5 and 6 hold. The sequence* $(\mathsf{z}^k)_{k\in\mathbb{N}} = (\mathsf{y}^k, \mathsf{w}^k, \mathsf{x}^k)_{k\in\mathbb{N}}$ *generated by Algorithm 3-II converges almost surely to some* $z^\star \in \mathcal{S}$*. Furthermore, if* f_i*,* g_i *and* h_i*,* $i = 1, \dots, m$ *are PLQ and* $(\mathsf{z}^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ *where* $\mathcal{C}$ *is a compact set, then* $\left(\mathbb{E}\left[d^2_{\Pi^{-1}S}(\mathsf{z}^k, \mathcal{S})\right]\right)_{k\in\mathbb{N}}$ *converges Q-linearly to zero.*

VI. Application: Formation Control

In this section we consider the problem of formation control of a group of robots [15], [35], where each robot/agent has its own local dynamics and cost function and the goal is to achieve a specific formation by communicating only with neighboring agents.

For simplicity of visualization we consider a 2D problem. Each subsystem (corresponding to a robot) has four states $x_i = (p_{x_i}, p_{y_i}, v_{x_i}, v_{y_i})$, where (p_{x_i}, p_{y_i}) and (v_{x_i}, v_{y_i}) denote the position and the velocity vectors, respectively. The input for each system is given by $u_i = (v^u_{x_i}, v^u_{y_i})$. The discrete-time LTI model of each system is given by

$$x_i(k+1) = \Phi_i x_i(k) + \Delta_i u_i(k), \quad k = 0, 1, \dots.$$

The state and input transition matrices are as follows

$$\Phi_i = \begin{pmatrix} I & 0 & X_1 & 0 \\ 0 & I & 0 & X_1 \\ 0 & 0 & X_2 & 0 \\ 0 & 0 & 0 & X_2 \end{pmatrix}, \quad \Delta_i = \begin{pmatrix} X_3 & 0 \\ 0 & X_3 \\ X_1 & 0 \\ 0 & X_1 \end{pmatrix},$$

where the parameters are $X_1 = -t_d(e^{-\frac{1}{t_d}} - 1)$, $X_2 = e^{-\frac{1}{t_d}}$ and $X_3 = t_d^2(e^{-\frac{1}{t_d}} - 1 + \frac{1}{t_d})$ with time constant $t_d = 5$ (s). This discrete-time model was derived from the continuous-time model of [35] using exact discretization with step length $\Delta T = 1$.

Let N denote the horizon length. Consider the stacked state and input vectors $\boldsymbol{x}_i \in \mathbb{R}^{4N}, \boldsymbol{u}_i \in \mathbb{R}^{2N}$:

$$\boldsymbol{x}_i := (x_i(1), \dots, x_i(N)),\ \boldsymbol{u}_i := (u_i(0), \dots, u_i(N-1)).$$

Then the dynamics of each agent can be represented as $\mathcal{A}_i \boldsymbol{x}_i + \mathcal{B}_i \boldsymbol{u}_i = b_i$ where $\mathcal{A}_i$, $\mathcal{B}_i$ are appropriate matrices and b_i depends on the initial state. The state and input constraints of each agent are represented by the sets $\mathcal{X}_i$, $\mathcal{U}_i$ and are assumed to be easy to project onto, *e.g.*, boxes, halfspaces, norm balls, etc. Moreover, we assume that each agent has its own private objective captured by input and state cost matrices $\mathcal{Q}_i$ and $\mathcal{R}_i$, and vectors q_i, t_i. The specific formation between agents is enforced using another quadratic term that penalizes deviation of two neighbors from the desired relative position. The optimization problem is described as follows:

$$\begin{aligned} \underset{\boldsymbol{x}_i, \boldsymbol{u}_i}{\textbf{minimize}}\ & \sum_{i=1}^{m} \tfrac{1}{2}\|\mathcal{Q}_i \boldsymbol{x}_i - q_i\|^2 + \tfrac{1}{2}\|\mathcal{R}_i \boldsymbol{u}_i - t_i\|^2 \\ & + \sum_{i=1}^{m} \sum_{j \in \mathcal{N}_i} \tfrac{\lambda_i}{2}\|\mathcal{C}(\boldsymbol{x}_i - \boldsymbol{x}_j) - d_{ij}\|^2 \qquad (33) \\ \textbf{subject to}\ & \mathcal{A}_i \boldsymbol{x}_i + \mathcal{B}_i \boldsymbol{u}_i = b_i,\ \boldsymbol{x}_i \in \mathcal{X}_i,\ \boldsymbol{u}_i \in \mathcal{U}_i \\ & i = 1, \dots, m \end{aligned}$$

The relative desired distance of agent i from its neighbor j is given by d_{ij}, $\mathcal{C}$ is an appropriate linear mapping that selects the position variables, and λ_i is an scalar weight to penalize deviation.

For each system that communicates with i, *i.e.*, $j \in \mathcal{N}_i$, we introduce a local variable $\boldsymbol{x}_{ij}$, that can be seen as the estimate of $\boldsymbol{x}_j$ kept locally by agent i. In order to be consistent hereafter the self variables $\boldsymbol{x}_i, \boldsymbol{u}_i$ are denoted by $\boldsymbol{x}_{ii}, \boldsymbol{u}_{ii}$.

For each agent $i = 1, \dots, m$ define the stacked vector

$$z_{\mathcal{N}_i} = \big((\boldsymbol{x}_{ij})_{j \in \mathcal{N}_i \cup \{i\}}, \boldsymbol{u}_{ii}\big) \in \mathbb{R}^{n_i},$$

where $n_i = 4N(|\mathcal{N}_i| + 1) + 2N$.

Let E_i be a linear mapping such that $E_i z_{\mathcal{N}_i} = \mathcal{A}_i \boldsymbol{x}_{ii} + \mathcal{B}_i \boldsymbol{u}_{ii}$. Hence, the set of points satisfying the dynamics are given by $\mathcal{D}_i = \{z \in \mathbb{R}^{n_i} | E_i z = b_i\}$. Consider the linear mapping L_i such that $L_i z_{\mathcal{N}_i} = (\boldsymbol{x}_{ii}, \boldsymbol{u}_{ii})$ and denote $\mathcal{Z}_i := \mathcal{X}_i \times \mathcal{U}_i$. Moreover, let $h_i := \delta_{\mathcal{Z}_i}$, $g_i := \delta_{\mathcal{D}_i}$ and

$$\begin{aligned} f_i(z_{\mathcal{N}_i}) := & \tfrac{1}{2}\|\mathcal{Q}_i \boldsymbol{x}_{ii} - q_i\|^2 + \tfrac{1}{2}\|\mathcal{R}_i \boldsymbol{u}_{ii} - t_i\|^2 \\ & + \tfrac{\lambda_i}{2} \textstyle\sum_{j \in \mathcal{N}_i} \|\mathcal{C}(\boldsymbol{x}_{ii} - \boldsymbol{x}_{ij}) - d_{ij}\|^2. \end{aligned}$$

With these definitions problem (33) is cast in the form of problem (30) (minimizing over $z_{\mathcal{N}_i}$, $i = 1, \dots, m$) where the linear mapping A_{ij}, for $j \in \mathcal{N}_i$, is such that $A_{ij} z_{\mathcal{N}_i} = (\boldsymbol{x}_{ii}, -\boldsymbol{x}_{ij})$ if $i < j$ and $A_{ij} z_{\mathcal{N}_i} = (-\boldsymbol{x}_{ij}, \boldsymbol{x}_{ii})$ otherwise. Therefore, we can readily apply TriPD-Dist (Alg. 3) to solve the problem in a fully distributed fashion yielding both synchronous and randomized asynchronous implementations.

In our simulations we used horizon length $N = 3$. For the input and state constraints of all agents we used box constraints: the positions p_{x_i} and p_{y_i} are assumed to be between 0 and 20 (m). The velocities v_{x_i} and v_{y_i} and inputs $v^u_{x_i}$ and $v^u_{y_i}$ are assumed to be between between 0 and 15 (m/s) (for all agents). The local state cost matrices are set $\mathcal{Q}_i = 0.1I$ for all i. The local input cost matrices are set $\mathcal{R}_i = I$ for half of the agents and $\mathcal{R}_i = 2I$ for the rest. Moreover, the vectors q_i, t_i are set equal to zero, and the penalty parameter $\lambda_i = 10$ is used for all the agents.

The stepsizes of TriPD-Dist (Alg. 3) were selected as follows: i) (edge stepsizes) $\kappa_{(i,j)} = 1$ for all $(i,j) \in \mathcal{E}$, ii) (node stepsizes) $\sigma_i = \beta_i/4$ and $\tau_i = 0.99/(\frac{\beta_i}{2} + \sigma_i + \sum_{j \in \mathcal{N}_i} \kappa_{(i,j)})$ for all i, where we used

$$\beta_i = \mathbf{max}\{\|\mathcal{Q}_i^\top \mathcal{Q}_i\| + \lambda_i(|\mathcal{N}_i| + 1), \|\mathcal{R}_i^\top \mathcal{R}_i\|\},$$

which is an upper bound for the Lipschitz constant of ∇f_i. It is plain to see that the above choice of stepsizes for the agents satisfy Assumption 6*(iii)*. Note that the stepsize selection only requires local parameters $\mathcal{R}_i$, $\mathcal{Q}_i$, λ_i and the number of neighbors $|\mathcal{N}_i|$, *i.e.*, the algorithm can be implemented without *any* global coordination.

In our simulations, we considered m robots initially in a polygon configuration and enforced an arrow formation by appropriate selection of d_{ij} in (33). This scenario is depicted for $m = 5$ in Figure 2. The neighborhood relation in this case is taken to be the same arrow configuration, *i.e.*, all agents have two neighbors apart from two agents with only one neighbor.

For comparison we considered the dual decomposition approach of [15] (based on the subgradient method). Notice that dual decomposition with gradient or accelerated gradient methods can not be applied to this problem since f_i's are convex but not strongly convex. Recently, TriPD-Dist (Alg. 3) was compared against the dual accelerated proximal gradient method, in the context of distributed model predictive control (with strongly convex quadratic cost) [36].

In the simulations for Figure 1, we used the stepsize $10/k$ (as tuned for achieving better performance) for the dual decomposition method where k is the number of iterations. Notice that the dual decomposition approach for this problem can not achieve a full splitting of the operators involved: at every iteration agents need to solve an inner minimization (we used MATLAB's `quadprog` to perform this step), the result of which must be communicated to the neighbors for their computation, and is followed by another communication round. This extra need for synchronization would further slow down the algorithm in practical implementations [37].

Figure 1 demonstrates the superior performance of both the synchronous and asynchronous versions of TriPD-Dist (Alg. 3) compared to the dual decomposition approach. The y-axis is the distance of $v^k := (\boldsymbol{x}^k_{11}, \boldsymbol{u}^k_{11}, \dots, \boldsymbol{x}^k_{mm}, \boldsymbol{u}^k_{mm})$ from the solution ($v^\star$ was computed by solving (33) in a centralized fashion). The x-axis denotes the total number of local transmissions between agents. In the asynchronous implementation we used independent activation probabilities $p_i = 0.5$ for all agents. It is observed that the total number of local iterations is similar to that of the synchronous implementation. Finally, as evident in Figure 1 both versions of TriPD-Dist (Alg. 3) achieve linear convergence rate as predicted by Theorems V.1 and V.2 (the functions f_i, g_i and h_i are PLQ).

VII. Conclusions

The primal-dual algorithm introduced in this paper enjoys several structural properties that distinguish it from other related methods in the literature. A key property, that has been instrumental in developing a block-coordinate version of the algorithm, is the fact that the generated sequence is S-Fejér

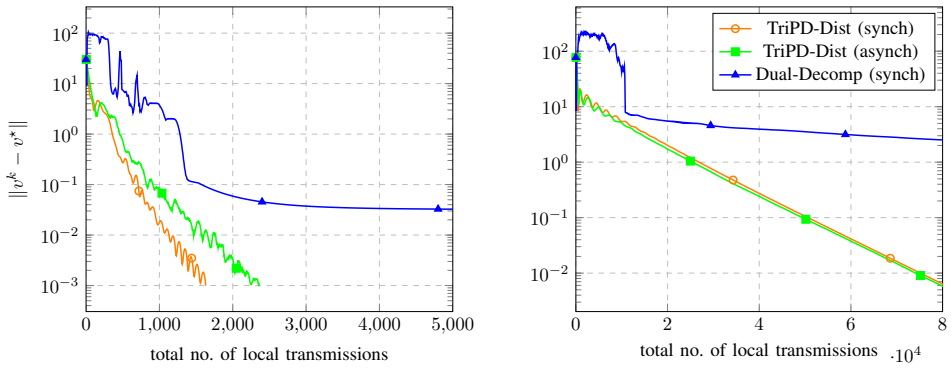


Figure 1. Comparison for the convergence of the algorithms for $m = 5$ (left), and $m = 50$ (right).

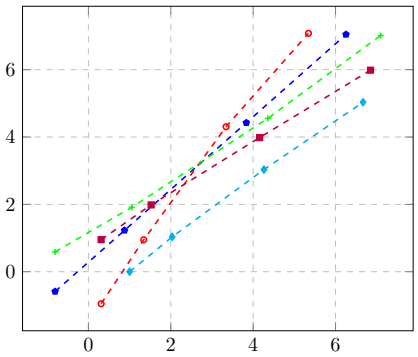


Figure 2. Five agents reorganizing from a polygon to an arrow configuration .

monotone, where S is a block diagonal positive definite matrix. It is shown that the algorithm attains linear convergence under a metric subregularity assumption that holds for a wide range of cost functions that are not necessarily strongly convex. The block-coordinate version of the developed algorithm is exploited to devise a novel fully distributed asynchronous method for multi-agent optimization over graphs. Our future work includes designing a block-coordinate version of the *SuperMann* scheme of [38] that applies to quasi-nonexpansive operators. In light of the fact that this method enjoys superlinear convergence rates, such extension is especially attractive for multi-agent optimization yielding schemes with faster convergence and fewer communication rounds. Other research directions enlist investigating extensions to account for directed and time-varying topologies, communication delays, and designing efficient strategies for selecting activation probabilities and stepsizes.

Appendix A

Proof of Lemma II.4. Consider the operator T as in (12). By monotonicity of A at $z^\star$ and $\bar{z}$ along with (13) we have

$$0 \leq \langle -Mz^\star - Cz^\star + Mz + Cz - Hz + H\bar{z}, z^\star - \bar{z}\rangle. \quad (34)$$

For $\beta_f > 0$, Assumption 1*(ii)* is equivalent to ∇f being *cocoercive* [17, Thm. 18.16], *i.e.*, for all $x, y \in \mathbb{R}^n$:

$$\tfrac{1}{\beta_f}\|\nabla f(x) - \nabla f(y)\|^2_{Q^{-1}} \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle. \quad (35)$$

On the other hand, for $\beta_f > 0$ we have

$$\begin{aligned}\langle Cz - Cz^\star, z^\star - \bar{z}\rangle &= \langle \nabla f(x) - \nabla f(x^\star), x^\star - \bar{x}\rangle\\ &= \langle \nabla f(x) - \nabla f(x^\star), x - \bar{x}\rangle\\ &\quad + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x\rangle\\ &\leq \tfrac{1}{\beta_f}\|\nabla f(x) - \nabla f(x^\star)\|^2_{Q^{-1}} + \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q\\ &\quad + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x\rangle\\ &\leq \langle \nabla f(x) - \nabla f(x^\star), x - x^\star\rangle + \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q\\ &\quad + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x\rangle,\\ &= \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q, \end{aligned} \quad (36)$$

where we have used (2) (with $V = \frac{2}{\beta_f}Q^{-1}$) in the first inequality, and (35) in the second inequality, respectively. Notice that if $\beta_f = 0$ then inequality (36) holds trivially with equality.

Using (36) in (34), along with skew-symmetry of K and M, we have

$$\begin{aligned}0 \leq& \langle -Mz^\star - Cz^\star + Mz + Cz - Hz + H\bar{z}, z^\star - \bar{z}\rangle\\ \leq& \langle (M - K)(z - z^\star) + P(\bar{z} - z), z^\star - \bar{z}\rangle + \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q\\ =& \langle (M - K)(z - z^\star) + P(\bar{z} - z), z^\star - z\rangle + \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q\\ &+ \langle (M - K)(z - z^\star) + P(\bar{z} - z), z - \bar{z}\rangle\\ =& \langle P(\bar{z} - z), z^\star - z\rangle + \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q - \|\bar{z} - z\|^2_P\\ &+ \langle (M - K)(z - z^\star), z - \bar{z}\rangle\\ =& \langle z - z^\star, (H + M^\top)(z - \bar{z})\rangle\\ &+ \tfrac{\beta_f}{4}\|x - \bar{x}\|^2_Q - \|\bar{z} - z\|^2_P. \end{aligned} \quad (37)$$

By definition, $S^{-1}(H + M^\top)(\bar{z} - z) = Tz - z$. Thus

$$\langle z - z^\star, (H + M^\top)(z - \bar{z})\rangle = \langle z - z^\star, z - Tz\rangle_S. \quad (38)$$

On the other hand, we have $\bar{z} - z = (H + M^\top)^{-1}S(Tz - z)$. Using (10), (14) and (7c) we conclude

$$\|\bar{z} - z\|^2_P - \tfrac{\beta_f}{4}\|\bar{x} - x\|^2_Q = \|Tz - z\|^2_{\tilde{P}}, \quad (39)$$

where $\tilde{P}$ is defined in (16). Combining (37), (38) and (39) completes the proof. □

Proof of Theorem II.5. We establish convergence by showing that the sequence $(z^k)_{k\in\mathbb{N}}$ is Fejér monotone with respect

to $\mathcal{S} = \mathbf{fix}\, T$. We have

$$\begin{aligned}\|z^{k+1} - z^\star\|_S^2 &= \|Tz^k - z^k + z^k - z^\star\|_S^2 \\ &= \|z^k - z^\star\|_S^2 + \|Tz^k - z^k\|_S^2 \\ &\quad + 2\langle z^k - z^\star, Tz^k - z^k\rangle_S \\ &\leq \|z^k - z^\star\|_S^2 - \|Tz^k - z^k\|_{2\tilde{P}-S}^2, \end{aligned} \tag{40}$$

where the inequality follows from Lemma II.4. Note that $2\tilde{P} - S$ is symmetric positive-definite if and only if Assumption 2 holds. Therefore, by (40) the sequence $(z^k)_{k\in\mathbb{N}}$ is Fejér monotone in the space equipped with inner product $\langle\cdot,\cdot\rangle_S$; in particular, $(z^k)_{k\in\mathbb{N}}$ is bounded. Furthermore, it follows from (40) and the fact that $2\tilde{P} - S$ is positive-definite that

$$\|Tz^k - z^k\| \to 0. \tag{41}$$

The operator T is continuous (since it involves proximal and linear mappings that are continuous, and since ∇f is assumed continuous). Let z^c be a cluster point of $(z^k)_{k\in\mathbb{N}}$. It follows from the continuity of T and (41) that $Tz^c - z^c = 0$, *i.e.*, $z^c \in \mathbf{fix}\, T$. The result follows from Fejér monotonicity of $(z^k)_{k\in\mathbb{N}}$ with respect to $\mathcal{S} = \mathbf{fix}\, T$ and [17, Thm. 5.5]. □

Proof of Theorem III.1. Let us define the operator $E^k := \sum_{i=1}^m \epsilon_i^k U_i$ that maps the elements of $(\mathbb{R}^{n+r}, \mathcal{F}_{k-1})$ to $(\mathbb{R}^{n+r}, \mathcal{F}_k)$. The iterations of TriPD-BC (Alg. 2) can be written as $z^{k+1} = z^k + E^{k+1}(Tz^k - z^k)$. We have

$$\begin{aligned}\mathbb{E}_k \circ E^{k+1} &= \sum_{\varepsilon\in\Psi} \mathbb{P}(\epsilon^{k+1} = \varepsilon) \sum_{j=1}^m \varepsilon_j U_j \\ &= \sum_{j=1}^m \sum_{\varepsilon\in\Psi} \mathbb{P}(\epsilon^{k+1} = \varepsilon)\varepsilon_j U_j \\ &= \sum_{j=1}^m \sum_{\varepsilon\in\Psi, \varepsilon_j = 1} \mathbb{P}(\epsilon^{k+1} = \varepsilon) U_j = \sum_{j=1}^m p_j U_j = \Pi, \end{aligned} \tag{42}$$

where we used Assumptions 3*(i)* and 3*(ii)*. Therefore, we have

$$\begin{aligned}&\mathbb{E}_k\left[\|z^{k+1} - z^\star\|_{\Pi^{-1}S}^2\right] \\ &= \mathbb{E}_k\left[\|z^k + E^{k+1}(Tz^k - z^k) - z^\star\|_{\Pi^{-1}S}^2\right] \\ &= \|z^k - z^\star\|_{\Pi^{-1}S}^2 + 2\langle z^k - z^\star, \mathbb{E}_k\left[E^{k+1}(Tz^k - z^k)\right]\rangle_{\Pi^{-1}S} \\ &\quad + \mathbb{E}_k\left[\langle E^{k+1}(Tz^k - z^k), E^{k+1}(Tz^k - z^k)\rangle_{\Pi^{-1}S}\right] \\ &= \|z^k - z^\star\|_{\Pi^{-1}S}^2 + \|Tz^k - z^k\|_S^2 \\ &\quad + 2\langle z^k - z^\star, Tz^k - z^k\rangle_S \end{aligned}$$

where we used (42) and the fact E^k is self-adjoint and idempotent (since U_i are 0-1 matrices) in the last equality. Inequality (28) follows by using (17). The convergence of the sequence follows from (28) using the Robbins-Siegmund lemma [26] and arguing as in [7, Thm. 3] and [8, Prop. 2.3]. □

Proof of Lemma IV.2. From the equivalent characterization of strong subregularity in (29) we have that there exists a neighborhood $\mathcal{U}_{x^\star}$ of $x^\star$ such that for all $x \in \mathcal{U}_{x^\star}$

$$\begin{aligned}(f+g)(x) \geq &(f+g)(x^\star) + \langle -L^\top u^\star, x - x^\star\rangle \\ &+ c_1\|x - x^\star\|^2, \end{aligned} \tag{43}$$

and a neighborhood $\mathcal{U}_{u^\star}$ of $u^\star$ such that for all $u \in \mathcal{U}_{u^\star}$

$$h^*(u) \geq h^*(u^\star) + \langle Lx^\star, u - u^\star\rangle + c_2\|u - u^\star\|^2. \tag{44}$$

Fix $z = (u, x)$ with $u \in \mathcal{U}_{u^\star}$ and $x \in \mathcal{U}_{x^\star}$. Consider $v = (v_1, v_2) \in Fz := Az + Mz + Cz$. By definition (cf. (8)) we have

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

Using this together with the definition of subdifferential yields:

$$\langle v_1 + Lx, u - u^\star\rangle \geq h^*(u) - h^*(u^\star), \tag{45}$$

$$\langle v_2 - L^\top u, x - x^\star\rangle \geq (f+g)(x) - (f+g)(x^\star). \tag{46}$$

Combining (45), (46) with (43), (44) and noting that

$$\langle L^\top(u^\star - u), x - x^\star\rangle + \langle L(x - x^\star), u - u^\star\rangle = 0,$$

yields:

$$\begin{aligned}\langle v, z - z^\star\rangle &= \langle v_1, u - u^\star\rangle + \langle v_2, x - x^\star\rangle \\ &\geq c_2\|u - u^\star\|^2 + c_1\|x - x^\star\|^2 \geq c\|z - z^\star\|^2, \end{aligned}$$

where $c = \mathbf{min}\{c_1, c_2\}$. Therefore, by the Cauchy-Schwarz inequality $\|v\| \geq c\|z - z^\star\|$. Since $\|z - z^\star\| \geq d(z, F^{-1}0)$, and $v \in Fz$ was selected arbitrarily, we have

$$d(z, F^{-1}0) \leq \tfrac{1}{c} d(0, Fz) \quad \forall z \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star}. \tag{47}$$

Thus F is metrically subregular at $z^\star$ for 0.

To establish uniqueness of the primal-dual solution consider:

$$\mathcal{L}(u, x) := (f+g)(x) + \langle Lx, u\rangle - h^*(u).$$

Adding (43) and (44) yields

$$\mathcal{L}(u^\star, x) - \mathcal{L}(u, x^\star) \geq c\|z - z^\star\|^2 \quad \forall z \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star} \tag{48}$$

Let $\bar{z}^\star = (\bar{u}^\star, \bar{x}^\star) \in \mathcal{S}$ such that $\bar{z}^\star \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star}$. Since $\bar{z}^\star$ is also a primal-dual solution we have $\mathcal{L}(\bar{u}^\star, x^\star) - \mathcal{L}(u^\star, \bar{x}^\star) \geq 0$. Therefore, using (48) at $\bar{z}^\star$ yields $\bar{z}^\star = z^\star$. Since $\mathcal{S}$ is convex, we conclude that it is a singleton, *i.e.*, $\mathcal{S} = \{z^\star\}$. Consequently it follows from (47) that F is strongly subregular at $z^\star$ for 0.

The second part is a direct consequence of the first part and the fact that if Assumption 4 holds globally then also the quadratic growth conditions (43) and (44) hold globally, *i.e.*, $\mathcal{U}_{x^\star} = \mathbb{R}^n$, $\mathcal{U}_{u^\star} \in \mathbb{R}^r$. This can be shown by adapting the proof of [29, Thm. 3.3]. □

Proof of Lemma IV.4. Since f, g and h are proper closed convex PLQ, the subdifferentials ∂g, ∇f and ∂h^* are piecewise polyhedral mappings [16, Prop. 12.30(b), Thm. 11.14(b)]. The graph of M is polyhedral, since M is linear. Therefore, the sum $F = A + M + C$ is also piecewise polyhedral. Since the inverse of a piecewise polyhedral mapping is piecewise polyhedral, the result follows from [27, 3H.1 and 3H.3]. □

Proof of Theorem IV.6. (Linear convergence of Alg. 2) For notational convenience let $\bar{S} = \Pi^{-1}S$ and note that $\mathcal{S} = \mathbf{zer}\, F$ (*cf.* (15)). By definition we have $\|z^k - \mathcal{P}_{\mathcal{S}}^{\bar{S}}(z^k)\|_{\bar{S}} = d_{\bar{S}}(z^k, \mathcal{S})$ (where the minimum is attained since $\mathcal{S}$ is a closed convex set). Consequently, it follows from (28) that

$$\begin{aligned}\mathbb{E}_k\left[d_{\bar{S}}^2(z^{k+1}, \mathcal{S})\right] &\leq \mathbb{E}_k\left[\|z^{k+1} - \mathcal{P}_{\mathcal{S}}^{\bar{S}}(z^k)\|_{\bar{S}}^2\right] \\ &\leq \|z^k - \mathcal{P}_{\mathcal{S}}^{\bar{S}}(z^k)\|_{\bar{S}}^2 - \|Tz^k - z^k\|_{2\tilde{P}-S}^2 \\ &= d_{\bar{S}}^2(z^k, \mathcal{S}) - \|Tz^k - z^k\|_{2\tilde{P}-S}^2. \end{aligned} \tag{49}$$

By definition (12), we have

$$\begin{aligned}&\|\bar{z}^k - z^k\|^2 = \|(H + M^\top)^{-1}S(Tz^k - z^k)\|^2 \\ &\leq \|(H + M^\top)^{-1}S\|^2 \|(2\tilde{P} - S)^{-1}\| \|Tz^k - z^k\|_{2\tilde{P}-S}^2, \end{aligned} \tag{50}$$

where $\bar{z}^k$ is defined by (13) applied at $z = z^k$. Consider the projection of $\bar{z}^k$ onto $\mathcal{S}$, $\mathcal{P}_{\mathcal{S}}(\bar{z}^k)$. By definition $\|\bar{z}^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\| = d(\bar{z}^k, \mathcal{S})$, and we have

$$\begin{aligned} d^2_{\bar{S}}(z^k, \mathcal{S}) &\le \|z^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\|^2_{\bar{S}} \le \|\bar{S}\|\|z^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\|^2 \\ &\le \|\bar{S}\|\big(\|\bar{z}^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\| + \|\bar{z}^k - z^k\|\big)^2 \\ &= \|\bar{S}\|\big(d(\bar{z}^k, \mathcal{S}) + \|\bar{z}^k - z^k\|\big)^2. \end{aligned} \tag{51}$$

In what follows we bound $d(\bar{z}^k, \mathcal{S})$ by $\|\bar{z}^k - z^k\|$. Define

$$v^k := -(H - M)(\bar{z}^k - z^k) + C\bar{z}^k - Cz^k. \tag{52}$$

It follows from (13) that $(H-M-C)z^k \in (H+D)\bar{z}^k$, which in turn implies

$$v^k \in F\bar{z}^k = (A + M + C)\bar{z}^k.$$

Consequently, using (global) metric subregularity of F yields

$$d(\bar{z}^k, \mathcal{S}) \le \eta\|v^k\|. \tag{53}$$

By the triangle inequality and Lipschitz continuity of C,

$$\begin{aligned} &\|v^k\| = \|(H-M)(\bar{z}^k - z^k) - C\bar{z}^k + Cz^k\| \\ &\le \|(H-M)(\bar{z}^k - z^k)\| + \|C\bar{z}^k - Cz^k\| \le \xi\|\bar{z}^k - z^k\|, \end{aligned} \tag{54}$$

where $\xi = \|H-M\| + \beta_f\|Q\|$. By (53) and (54) we have

$$d(\bar{z}^k, \mathcal{S}) \le \xi\eta\|\bar{z}^k - z^k\|.$$

Combine this with (50) and (51) to derive

$$d^2_{\bar{S}}(z^k, \mathcal{S}) \le \phi\|Tz^k - z^k\|^2_{2\tilde{P}-S}, \tag{55}$$

where $\phi = (\xi\eta + 1)^2\|(H + M^\top)^{-1}S\|^2\|(2\tilde{P} - S)^{-1}\|\|\bar{S}\|$. Therefore, by (49) and (55) we have

$$\mathbb{E}_k\big[d^2_{\bar{S}}(z^{k+1}, \mathcal{S})\big] \le d^2_{\bar{S}}(z^k, \mathcal{S}) - \tfrac{1}{\phi}d^2_{\bar{S}}(z^k, \mathcal{S}).$$

Taking expectation in both sides concludes the proof. For the case of PLQ functions, let $\mathcal{U}_{z_\star}$ denote an open subregularity neighborhood around $z_\star \in \mathcal{S}$, and set $\mathcal{U}_\star := \cup_{z^\star \in \mathcal{S}}\mathcal{U}_{z_\star}$. By Lemma IV.4 there exists a positive η such that $d(z, F^{-1}0) \le \eta d(0, Fz)$ for $z \in \mathcal{U}_\star$. Moreover, since $(z^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ up to possibly enlarging $\mathcal{C}$ we have $(\bar{z}^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$. Note that since $(z^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ and $\mathcal{C}$ is closed, $\mathcal{C} \cap \mathcal{S} \neq \emptyset$ and $\mathcal{C} \cap \mathcal{U}_\star \neq \emptyset$. It is sufficient to show that $d(z, F^{-1}0) \le \eta' d(0, Fz)$ for $z \in \mathcal{C}$. Let us define $D(z) := d(0, Fz)$. Since $\mathbf{gra}\, F$ is closed, $D(z)$ is lower semicontinuous [16, Thm. 5.7, Prop. 5.11(a)]. By [16, Cor. 1.10] $D(z)$ attains a minimum over the compact set $\mathcal{C} \setminus \mathcal{U}_\star$: $c_d := \mathbf{min}_{z\in\mathcal{C}\setminus\mathcal{U}_\star} D(z) > 0$ where the strict inequality is due to the fact that the minimizer belongs to $\mathcal{C} \setminus \mathcal{U}_\star$. Moreover, $c_{\mathcal{C}} := \mathbf{sup}_{z\in\mathcal{C}} d(z, F^{-1}0) < \infty$ due to the fact that $\mathcal{C}$ is bounded. Hence $d(z, F^{-1}0) \le c_{\mathcal{C}} \le \frac{c_{\mathcal{C}}}{c_d} d(0, Fz)$ for $z \in \mathcal{C} \setminus \mathcal{U}_\star$. Therefore, by combining the two cases we obtain $d(z, F^{-1}0) \le \mathbf{max}\{\frac{c_{\mathcal{C}}}{c_d}, \eta\} d(0, Fz)$ for $z \in \mathcal{C}$ as claimed. The second sufficient condition follows from Lemma IV.2. □

Proof of Lemma IV.7. First we show the if statement: assume that $R = \mathrm{Id} - T$ is metrically subregular at $z^\star$ for 0. Then there exists $\eta > 0$ and a neighborhood $\mathcal{U}$ of $z^\star$ such that

$$d(z, R^{-1}0) \le \eta d(0, Rz) \quad \forall z \in \mathcal{U}. \tag{56}$$

The two sets $R^{-1}0$ and $F^{-1}0$ are equal, *cf.* (15). In what follows, we upper bound $d(0, Rz)$ by $d(0, Fz)$. Let $w \in Fz = Az + Mz + Cz$. By (13) we have that

$$Hz - Mz - Cz - H\bar{z} \in A\bar{z}.$$

Using this together with the monotonicity of A at z and $\bar{z}$, we obtain:

$$\begin{aligned} 0 \le &\langle z - \bar{z}, (w - Mz - Cz) - (Hz - Mz - Cz - H\bar{z})\rangle \\ = &\langle z - \bar{z}, w - Hz + H\bar{z}\rangle = \langle z - \bar{z}, w\rangle - \|\bar{z} - z\|^2_P, \end{aligned}$$

where in the last equality we have used the fact that $H = P + K$ and K is skew-symmetric.

By the Cauchy–Schwarz inequality

$$\|\bar{z} - z\|^2_P \le \langle z - \bar{z}, w\rangle \le \|\bar{z} - z\|_P\|w\|_{P^{-1}},$$

therefore

$$\|\bar{z} - z\|_P \le \|w\|_{P^{-1}}. \tag{57}$$

On the other hand by (12):

$$\|Rz\| \le \|S^{-1}(H + M^\top)P^{-1/2}\|\|\bar{z} - z\|_P.$$

Combine this with (56) and (57) to obtain

$$\begin{aligned} d(z, F^{-1}0) = &d(z, R^{-1}0) \le \eta\|Rz\| \\ \le &\eta\|S^{-1}(H + M^\top)P^{-1/2}\|\|P^{-1}\|^{1/2}\|w\|. \end{aligned}$$

Since $w \in Fz$ was arbitrary, we conclude that F is metrically subregular at $z^\star$ for 0 (with a different subregularity modulus).

Next we prove the *only if* statement: assume that F is metrically subregular at $z^\star$ for 0, *i.e.*, there exists $\eta > 0$ and neighborhood $\mathcal{U}$ of $z^\star$ such that

$$d(z, F^{-1}0) \le \eta d(0, Fz) \quad \forall z \in \mathcal{U}. \tag{58}$$

By (37) and the Cauchy–Schwarz inequality we infer that

$$\|\bar{z} - z\| \le c\|z - z^\star\|,$$

for some positive constant c. Hence, there exists a neighborhood $\bar{\mathcal{U}} \subset \mathcal{U}$ of $z^\star$ such that if $z \in \bar{\mathcal{U}}$ then $\bar{z} \in \mathcal{U}$. Fix a point $z \in \bar{\mathcal{U}}$ so that $\bar{z} \in \mathcal{U}$. By (58) it holds that:

$$d(\bar{z}, F^{-1}0) \le \eta d(0, F\bar{z}). \tag{59}$$

Define v as in (52) (dropping the iteration index k). Noting that $v \in F\bar{z}$, it follows from (59) that

$$d(\bar{z}, F^{-1}0) \le \eta\|v\| \le \eta\xi\|\bar{z} - z\|, \tag{60}$$

where we used (54) in the second inequality. Invoking triangle inequality we have

$$\begin{aligned} d(z, R^{-1}0) = &d(z, F^{-1}0) \le d(\bar{z}, F^{-1}0) + \|\bar{z} - z\| \\ \le &(1 + \eta\xi)\|\bar{z} - z\|. \end{aligned} \tag{61}$$

On the other hand by (12) it holds that

$$\|\bar{z} - z\| \le \|(H + M^\top)^{-1}S\|\|Rz\|.$$

Combining this with (61) yields

$$d(z, R^{-1}0) \le (1 + \eta\xi)\|(H + M^\top)^{-1}S\|\|Rz\| \quad \forall z \in \bar{\mathcal{U}},$$

i.e., that R is metrically subregular at $z^\star$ for 0. □

References

[1] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer New York, 2011, pp. 185–212.

[2] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Advances in Computational Mathematics*, vol. 38, no. 3, pp. 667–681, 2013.

[3] L. Condat, "A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, 2013.

[4] P. L. Combettes, L. Condat, J.-C. Pesquet, and B. C. Vũ, "A forward-backward view of some primal-dual optimization methods in image recovery," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4141–4145.

[5] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[6] P. Richtárik and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, no. 1-2, pp. 1–38, 2014.

[7] P. Bianchi, W. Hachem, and F. Iutzeler, "A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 10, pp. 2947–2957, Oct 2016.

[8] P. L. Combettes and J.-C. Pesquet, "Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 1221–1248, 2015.

[9] J.-C. Pesquet and A. Repetti, "A class of randomized primal-dual algorithms for distributed optimization," *Journal of Nonlinear and Convex Analysis*, vol. 16, no. 12, pp. 2453–2490, 2015.

[10] O. Fercoq and P. Bianchi, "A coordinate descent primal-dual algorithm with large step size and possibly non separable functions," *arXiv preprint arXiv:1508.04625*, 2015.

[11] G. Zhang and R. Heusdens, "Bi-alternating direction method of multipliers over graphs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 3571–3575.

[12] P. Latafat, L. Stella, and P. Patrinos, "New primal-dual proximal algorithm for distributed optimization," in *55th IEEE Conference on Decision and Control (CDC)*, 2016, pp. 1959–1964.

[13] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[14] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[15] R. L. Raffard, C. J. Tomlin, and S. P. Boyd, "Distributed optimization for cooperative agents: application to formation flight," in *43rd IEEE Conference on Decision and Control (CDC)*, vol. 3, 2004, pp. 2453–2459.

[16] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[17] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

[18] R. T. Rockafellar, *Convex analysis*. Princeton University Press, 2015.

[19] P. Latafat and P. Patrinos, "Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators," *Computational Optimization and Applications*, vol. 68, no. 1, pp. 57–93, Sep 2017.

[20] L. M. Briceño-Arias and P. L. Combettes, "A monotone + skew splitting model for composite monotone inclusions in duality," *SIAM Journal on Optimization*, vol. 21, no. 4, pp. 1230–1250, 2011.

[21] 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*, vol. 20, no. 2, pp. 307–330, 2012.

[22] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[23] Y. Drori, S. Sabach, and M. Teboulle, "A simple algorithm for a class of nonsmooth convex-concave saddle-point problems," *Operations Research Letters*, vol. 43, no. 2, pp. 209–214, 2015.

[24] P. Latafat and P. Patrinos, "Primal-dual proximal algorithms for structured convex optimization: A unifying framework," in *Large-Scale and Distributed Optimization*, P. Giselsson and A. Rantzer, Eds. Springer International Publishing, 2018, pp. 97–120.

[25] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers," in *52nd IEEE Conference on Decision and Control (CDC)*, 2013, pp. 3671–3676.

[26] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 111–135.

[27] A. L. Dontchev and R. T. Rockafellar, "Implicit functions and solution mappings," *Springer Monographs in Mathematics. Springer*, vol. 208, 2009.

[28] A. Ioffe, *Variational Analysis of Regular Mappings: Theory and Applications*, ser. Springer Monographs in Mathematics. Springer International Publishing, 2017.

[29] F. J. Aragón Artacho and M. H. Geoffroy, "Characterization of metric regularity of subdifferentials," *Journal of Convex Analysis*, vol. 15, no. 2, pp. 365–380, 2008.

[30] D. Drusvyatskiy and A. S. Lewis, "Error bounds, quadratic growth, and linear convergence of proximal methods," *Mathematics of Operations Research*, vol. 43, no. 3, pp. 919–948, 2018.

[31] J. Liang, J. Fadili, and G. Peyré, "Convergence rates with inexact non-expansive operators," *Mathematical Programming*, vol. 159, no. 1, pp. 403–434, Sep 2016.

[32] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[33] N. Freris and A. Zouzias, "Fast distributed smoothing of relative measurements," in *51st IEEE Conference on Decision and Control (CDC)*, 2012, pp. 1411–1416.

[34] A. Zouzias and N. Freris, "Randomized gossip algorithms for solving Laplacian systems," in *European Control Conference (ECC)*, 2015, pp. 1920–1925.

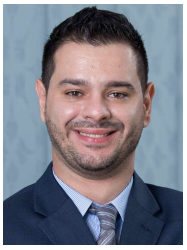
[35] T. Schouwenaars, J. How, and E. Feron, "Decentralized cooperative trajectory planning of multiple aircraft with hard safety guarantees," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2004, pp. 1–14.

[36] P. Latafat, A. Bemporad, and P. Patrinos, "Plug and play distributed model predictive control with dynamic coupling: A randomized primal-dual proximal algorithm," in *European Control Conference (ECC)*, June 2018, pp. 1160–1165.

[37] N. Freris, S. Graham, and P. Kumar, "Fundamental limits on synchronizing clocks over networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1352–1364, 2011.

[38] A. Themelis and P. Patrinos, "Supermann: a superlinearly convergent algorithm for finding fixed points of nonexpansive operators," *arXiv preprint arXiv:1609.06955*, 2016.

Puya Latafat is currently working towards a joint PhD at the Department of Electrical Engineering (ESAT) of KU Leuven (Belgium) and IMT School for Advanced Studies Lucca (Italy). He received his M.Sc. in Mathematical Engineering jointly from University of L'Aquila (Italy) and University of Hamburg (Germany), and his B.Sc. in Electrical Engineering from University of Tabriz (Iran). His research interests revolve around large-scale and distributed optimization with applications to model predictive control and machine learning.



Nikolaos M. Freris is Professor with the School of Computer Science and Technology at the University of Science and Technology of China (USTC). He received a Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece in 2005, an M.S. degree in Electrical and Computer Engineering, an M.S. degree in Mathematics, and a Ph.D. degree in Electrical and Computer Engineering all from the University of Illinois at Urbana-Champaign in 2007, 2008, and 2010, respectively. Dr. Freris's research interests lie in the area of cyberphysical systems: distributed estimation, optimization, and control, machine learning, wireless networks, signal processing, and applications in transportation, sensor networks, robotics, and power systems. His research was recognized with the 1000-talents award, the IBM High Value Patent award, two IBM invention achievement awards, and the Gerondelis foundation award. Previously, Dr. Freris was Assistant Professor of Electrical and Computer Engineering at New York University Abu Dhabi, and Global Network Assistant Professor of Computer Science at NYU Tandon School of Engineering. Dr. Freris is a senior member of IEEE, and a member of ACM and SIAM.

Panagiotis (Panos) Patrinos is assistant professor at the Department of Electrical Engineering (ESAT) of KU Leuven, Belgium since 2015. During fall/winter 2014 he held a visiting assistant professor position at Stanford University. He received his PhD in Control and Optimization, M.S. in Applied Mathematics and M.Eng. from National Technical University of Athens, in 2010, 2005 and 2003, respectively. After his PhD he held postdoc positions at the University of Trento and IMT Lucca, Italy, where he became an assistant professor in 2012. His current research interests are in the theory and algorithms of structured convex and nonconvex optimization and predictive control with a focus on large-scale, distributed, stochastic and embedded optimization and a wide range of application areas including smart grids, water networks, automotive, aerospace, machine learning and signal processing.