Open access • Journal Article • DOI:10.1109/TAC.2019.2906924

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

Puya Latafat, Nikolaos M. Freris, Panagiotis Patrinos

Institutions: Katholieke Universiteit Leuven, University of Science and Technology of China

Published on: 25 Mar 2019 - IEEE Transactions on Automatic Control (IEEE)

Topics: Piecewise, Convex function and Convexity

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

Puya Latafat, Nikolaos M. Freris, Panagiotis Patrinos

***Abstract*—We consider the problem of minimizing the sum of a Lipschitz differentiable function and two nonsmooth proximable functions one of which is composed with a linear mapping. We propose a novel primal-dual algorithm that takes into account a metric for the Lipschitz continuity instead of a scalar. Moreover, we devise a randomized block-coordinate version of our scheme that captures many random coordinate activation scenarios in a unified fashion. The block-coordinate version of the algorithm converges under identical stepsize conditions as the full algorithm. We show that both the full and block-coordinate schemes feature linear convergence rates if the functions involved are either piecewise linear-quadratic, or if they satisfy a quadratic growth condition. We apply the proposed algorithms to the problem of distributed multi-agent optimization, thus yielding synchronous and asynchronous distributed algorithms. The proposed algorithms are fully distributed in the sense that the stepsizes of each agent only depend on local information. In fact, no prior global coordination is required. Finally, we showcase our distributed algorithms for Network Utility Maximization.**

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

I. Introduction

In this paper we consider the optimization problem

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

where L is a linear mapping, h and g are proper closed convex (possibly nonsmooth) functions, 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.

A primal-dual algorithm was proposed separately by Vũ and Condat [2], [3] to tackle problem (1). The idea is to show that the algorithm takes the form of forward-backward splitting and thus to use Krasnosel'skiĭ-Mann iterations. In both aforementioned papers, the gradient of f is assumed to have a scalar Lipschitz constant. However, in many cases of practical interest, a scalar Lipschitz constant can not properly capture the Lipschitz continuity of ∇f. For this reason, we consider a Lipschitz metric $Q\succ 0$, *i.e.*, for all $x,y\in\mathbb{R}^n$ we assume that:

$$\|\nabla f(x)-\nabla f(y)\|_{Q^{-1}}\leq\|x-y\|_Q. \tag{2}$$

For example, in distributed multi-agent optimization where f is separable, *i.e.*, $f(x)=\sum_{i=1}^m f_i(x_i)$, the metric Q is block diagonal, with blocks containing the Lipschitz constants of ∇f_i's. Adapting the proof of [2], [3] to the case of general Lipschitz condition (2), leads to conservative conditions for selecting stepsizes that depend on $\|Q\|$ (see [3, proof of Thm. 3.1]): for instance, for separable f, $\|Q\|$ equals the maximum of the individual Lipschitz constants. In [4], the authors consider a preconditioned variable metric forward-backward iteration with the stepsize matrix set to be proportional to Q^{-1}; consequently, the selection of stepsizes is limited to a narrow special case. In contrast, we assume a general stepsize matrix, and our convergence analysis yields less conservative conditions for the stepsizes.

Our main contribution is a new primal-dual algorithm for solving (1). The method is based on applying a special case of *Asymmetric Forward-Backward Adjoint* (AFBA) splitting, proposed recently by the authors [5, Algorithm 1], to the primal-dual optimality conditions for (1). The sequence generated by the proposed algorithm is Fejér monotone with respect to $\|\cdot\|_S$ where S is a block diagonal positive definite matrix. This is the key property that is exploited to develop a block-coordinate version of the algorithm without introducing additional variables. In addition, the dynamic stepsize in [5, Algorithm 1] is replaced with a *constant* matrix Λ: this is especially important for distributed optimization in large-scale systems where global coordination may be infeasible.

Block-Coordinate (BC) minimization is a simple approach to tackling large-scale optimization problems. In BC schemes, at each iteration, a subset of the coordinates is updated while the others are held fixed. In particular, we are interested in randomized BC algorithms. Such schemes can be divided into two categories: a) algorithms in which only one coordinate is randomly activated and updated at each iteration [6]–[8], and b) algorithms where several coordinates are randomly activated and simultaneously updated [9], [10]. Our proposed method captures both cases in a unified way (cf. Section III).

A key property of the randomized BC versions of gradient and proximal gradient methods [6], [7], is that the stepsizes are inversely proportional to the coordinate-wise Lipschitz constant rather than the global one. This leads to larger stepsizes in directions with smaller Lipschitz constant and results in faster convergence. In [9], [10] random BC is applied to α-averaged operators by establishing stochastic Fejér monotonicity. In [9], [11] the authors use this analysis to derive random BC algorithms based on the primal-dual algorithm of Vũ and

Puya Latafat[1,2] ✉ puya.latafat@imtlucca.it
Nikolaos M. Freris [3] ✉ nf47@nyu.edu
Panagiotis Patrinos[2] ✉ panos.patrinos@esat.kuleuven.be

The work of P. Patrinos was supported by KU Leuven Research Council BOF/STG-15-043.

[1]IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100 Lucca, Italy.
[2]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS) and Optimization in Engineering Center (OPTEC), Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium.
[3] New York University Abu Dhabi, Division of Engineering, P.O. Box 129188, Abu Dhabi, UAE.

Condat [2], [3]. The main drawback of these approaches is that (just as in the full version of the algorithms), the metric of Lipschitz continuity is not captured by the algorithms thus yielding conservative ranges for the stepsizes.

In the recent paper [8], the authors seek to overcome this issue by proposing a scheme based on the Vũ-Condat algorithm: their analysis does not require the cost functions to be separable and uses a different Lyapunov function for establishing convergence. However, in its general form, the method requires introducing duplicate dual variables; this is because unlike the new primal-dual algorithm proposed here, the Fejér monotonicity of the Vũ-Condat algorithm holds with respect to $\|\cdot\|_S$, where S is not diagonal (cf. (28)).

In the BC version of our algorithm, we use the same metric Q for the Lipschitz continuity as in the full algorithm. This is in contrast to [6]–[8] where coordinate-wise Lipschitz continuity is used. In the special case when f is separable, the two assumptions are equivalent and we have $Q = \operatorname{blkdiag}(\beta_1 I_{n_1}, \ldots, \beta_m I_{n_m})$, where m is the number of blocks of coordinates, n_i is the dimension of the i-th coordinate block, and β_i denotes the Lipschitz constant of f_i. In this setting, our proposed block-coordinate algorithm leads to less restrictive conditions as compared to [8, Assumption 1(e)], in that the stepsizes of the new algorithm are inversely proportional to $\frac{\beta_i}{2}$ rather than β_i. Notice that in the general case [6, Lemma 2] can be used to establishes the connection between the metric Q and coordinate-wise Lipschitz assumption. However, in many cases such as the separable case this result is conservative.

As an application, we consider the distributed optimization problem over a network of agents. Each agent has its own private cost function of the form (1), and the communication between agents is characterized by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\underset{x_i \in \mathbb{R}^{n_i}}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) + g_i(x_i) + h_i\left(L_i x_i\right) \tag{3a}$$

$$\text{subject to} \quad A_{ij}x_i + A_{ji}x_j = b_{(i,j)} \qquad (i,j) \in \mathcal{E} \tag{3b}$$

Throughout the paper (i, j) is used to denote the unordered pair of i, j; and ij for the ordered pair. The goal is to solve the global optimization problem through local exchange of information. Notice that the linear constraints (3b) prescribe relations between neighboring agents' variables. This type of edge constraints is considered in [12]. Most notably, for two agents $i = 1, 2$, with $f_i, h_i \equiv 0$, such constrains are reminiscent of the Alternating Direction Method of Multipliers (ADMM). A special case of particular interest is *consensus*, when $A_{ij} = I$, $A_{ji} = -I$ and $b_{(i,j)} = 0$.

Another primal-dual algorithm was introduced in [13] for problem (3) when $f_i \equiv 0$ and with consensus constraints. The approach in that work is different and consists of a transformation to replace the edge variables with node variables. The main drawback of the algorithm is that as in the Vũ-Condat algorithm the Fejér monotonicity of the generated sequence holds with respect to $\|\cdot\|_S$ where S is not diagonal.

The multi-agent optimization problem (3) arises in many contexts such as sensor networks, power systems, transportation networks, robotics, water networks, distributed data-sharing, etc. [14]–[16]. In most of these applications, there are computation, communication, real-time constraints and/or physical limitations on the system that render centralized management infeasible. This motivates devising fully distributed asynchronous algorithms for problem (3). Therefore, randomized BC methods are of particular interest: they amount to a random activation of nodes (independent of each other) that perform local calculations and, subsequently, communicate updated values to their neighbors at the end of each iteration.

Contributions and Paper Structure

The main contributions of the paper can be summarized as follows:

- In Section II we propose a new primal-dual algorithm with Gauss-Seidel type updates for solving structured optimization problem (1). The proposed algorithm considers a metric for Lipschitz continuity rather than a scalar. Furthermore, we show how our analysis can be applied to the primal-dual method of Vũ-Condat, and emphasize that we obtain less conservative conditions in selecting stepsizes than [2]–[4].
- We establish linear convergence under an additional *metric subregularity* assumption on the monotone operator defining the primal-dual optimality conditions (cf. Theorem 2). Moreover, we show that the metric subregularity assumption holds if the involved functions f, g and h are either (1): *piecewise linear-quadratic* (cf. Lemma 3), or if (2): they satisfy a quadratic growth condition (cf. Lemma 2).
- In Section III, we propose a randomized *block-coordinate* (BC) version of our algorithm with *identical stepsize conditions* as in the original scheme. Our proof relies heavily on the fact that the sequence generated by our algorithm is stochastic Fejér monotone with respect to $\|\cdot\|_S$, where S is a diagonal matrix. This is a particularly attractive feature of our method that allows us to devise *fully distributed asynchronous* schemes.
- In Section IV we adapt the developed algorithm to the distributed optimization problem (3). The resulting algorithm is *fully* distributed in the sense that the stepsizes of each agent is selected based on local information without any prior global coordination (cf. Theorem 6). In fact, the *edge weight* $\kappa_{(i,j)}$ is the only parameter that must be fixed by the neighboring agents i and j. In addition, an asynchronous version of the algorithm is developed based on an instance of the block-coordinate algorithm in Section III. In this setting, at each iteration agents wake up at random independently of one another, *i.e.*, several agents might update their values at a given iteration. Moreover, under the metric subregularity assumption both synchronous and asynchronous versions of the algorithm achieve linear convergence rates. Finally, in Section V we consider the *Network Utility Maximization* problem as an application.

Notation and Preliminaries

We first introduce definitions and notation used throughout the paper; we refer the reader to [17], [18] for more details.

For an extended-real-valued function f, we use $\operatorname{dom} f$ to denote its domain. For a set C, we denote its relative interior

by $\operatorname{ri} C$. For a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$, we define the induced Euclidean norm $\|\cdot\|_P$: for $x \in \mathbb{R}^n$, $\|x\|_P = \sqrt{x'Px}$.

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 $\operatorname{dom} A = \{x \in \mathbb{R}^n \mid Ax \neq \emptyset\}$, its graph by $\operatorname{gra} A = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^d \mid y \in Ax\}$, the set of its zeros by $\operatorname{zer} A = \{x \in \mathbb{R}^n \mid 0 \in Ax\}$, and the set of its fixed points by $\operatorname{fix} A = \{x \mid x \in Ax\}$. The mapping A is called monotone if $\langle x - x', y - y'\rangle \geq 0$ for all $(x, y), (x', y') \in \operatorname{gra} A$, 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: $\operatorname{gra} A^{-1} := \{(y, x) \mid (x, y) \in \operatorname{gra} A\}$. The *resolvent* of A is defined by $J_A := (\mathrm{Id} + A)^{-1}$, where Id denotes the identity operator.

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

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

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

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

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 $x, u \in \mathbb{R}^n$; we use it throughout in the special case when $f = \frac{1}{2}\|\cdot\|_V^2$ for some symmetric positive definite matrix V. The distance from a proper closed convex set X with respect to $\|\cdot\|_V$ is denoted by $d_V(\cdot, X)$. We denote by $\mathcal{P}_X^V(\cdot)$ the projection onto set X with respect to $\|\cdot\|_V$.

II. A New Primal-Dual Algorithm

The following are assumed throughout Sections II and III:

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 $L \in \mathbb{R}^{r\times n}$ is a matrix.

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

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

or, equivalently, ∇f is *cocoercive*:

$$\|\nabla f(x) - \nabla f(y)\|_{Q^{-1}}^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle. \quad (4)$$

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

The primal-dual optimality condition for problem (1) is

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

With a slight abuse of terminology, we say that $(u^\star, x^\star)$ is a primal-dual solution (in place of dual-primal) if it satisfies (5), where $u^\star$ solves the dual problem and $x^\star$ solves the primal problem (1). We denote the set of primal-dual solutions by $\mathcal{S}$. Assumption 1*(iii)* guarantees that $\mathcal{S}$ is non empty (see [19, Corollary 31.2.1] and [20, Proposition 4.3(iii)]).

Let us define the following operators

$$D : (u, x) \mapsto (\partial h^*(u), \partial g(x)), \quad (6a)$$

$$M : (u, x) \mapsto (-Lx, L^\top u), \quad (6b)$$

$$F : (u, x) \mapsto (0, \nabla f(x)). \quad (6c)$$

The optimality condition (5) can be written in the form of the monotone inclusion

$$0 \in Dz + Mz + Fz, \quad (7)$$

where $z = (u, x)$. The linear operator M is monotone since it is skew-symmetric, *i.e.*, $M^\top = -M$. It is straightforward to see that the operator D is maximally monotone [18, Theorem 21.2 and Proposition 20.23], while operator F, being the gradient of $\tilde{f}(u, x) = f(x)$, is cocoercive.

Consider the operator T defined as follows:

$$Tz := z + S^{-1}(H + M^\top)(\bar{z} - z), \quad (8)$$

where

$$\bar{z} = (H + D)^{-1}(H - M - F)z, \quad (9)$$

and $H = P + K$, is the sum of a symmetric positive definite matrix P, and a skew-symmetric matrix K. This structure of H is the key to deriving Gauss-Seidel type updates. The matrix S in (8) is another design parameter and is symmetric positive definite. Notice that when $M \equiv 0$, (9) can be viewed as an asymmetrically preconditioned forward-backward update. Furthermore, from (8), (9) it follows that $z \in \mathcal{S}$ if and only if $z = Tz$, *i.e.*,

$$\mathcal{S} = \{z \mid 0 \in Dz + Mz + Fz\} = \operatorname{fix} T. \quad (10)$$

Set:

$$P = \begin{bmatrix} \Sigma^{-1} & \frac{1}{2}L \\ \frac{1}{2}L^\top & \Gamma^{-1} \end{bmatrix}, \quad K = \begin{bmatrix} 0 & -\frac{1}{2}L \\ \frac{1}{2}L^\top & 0 \end{bmatrix}. \quad (11)$$

Note that $H = P + K$ is lower block triangular. Therefore, in view of [5, Lemma 3.1] the backward step $(H+D)^{-1}$ in (9) is carried out sequentially, in that the dual vector $\bar{u}$ is computed (through proximal mapping) using (u, x), and the primal $\bar{x}$ is subsequently computed using $\bar{u}$, as well as x, cf. (14). Furthermore, it follows from (8) that this selection results in $H + M^\top$ being upper block triangular which allows to take S block diagonal while maintaining efficiently computable iterations. We let $\Sigma \in \mathbb{R}^{n\times n}$ and $\Gamma \in \mathbb{R}^{r\times r}$ be two symmetric positive definite matrices, and set

$$S = \operatorname{blkdiag}(\Sigma^{-1}, \Gamma^{-1}), \quad (12)$$

whence we have

$$S^{-1}(H + M^\top) = \begin{bmatrix} I & \Sigma L \\ 0 & I \end{bmatrix}. \quad (13)$$

We emphasize that the diagonal structure of S is the key property used in developing the block-coordinate version of the algorithm (cf. Section III).

In proximal form, the operator T defined in (8) for problem (1) is given by:

$$\bar{u} = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(u + \Sigma L x) \tag{14a}$$
$$\bar{x} = \operatorname{prox}_{g}^{\Gamma^{-1}}(x - \Gamma\nabla f(x) - \Gamma L^\top \bar{u}) \tag{14b}$$
$$Tz = (\bar{u} + \Sigma L(\bar{x} - x), \bar{x}), \tag{14c}$$

where $z = (u, x)$. The next lemma establishes a very important property of the operator T and is essential in our convergence analysis.

Lemma 1. *Consider the operator T in (14). Suppose that Σ and Γ are such that*

$$\tilde{P} := \begin{bmatrix} \Sigma^{-1} & -\frac{1}{2}L \\ -\frac{1}{2}L^\top & \Gamma^{-1} - \frac{1}{4}Q \end{bmatrix} \succ 0. \tag{15}$$

Then for any $z^\star \in \mathcal{S}$ and any $z \in \mathbb{R}^{n+r}$ we have

$$\|Tz - z\|_{\tilde{P}}^2 \leq \langle z - z^\star, z - Tz\rangle_S. \tag{16}$$

Proof. Consider the definition of the operator T in (8). From monotonicity of D at $z^\star$ and $\bar{z}$ along with (9) we have

$$0 \leq \langle -Mz^\star - Fz^\star + Mz + Fz - Hz + H\bar{z}, z^\star - \bar{z}\rangle. \tag{17}$$

On the other hand, we have

$$\begin{aligned}
\langle Fz - Fz^\star, z^\star - \bar{z}\rangle &= \langle \nabla f(x) - \nabla f(x^\star), x^\star - \bar{x}\rangle \\
&= \langle \nabla f(x) - \nabla f(x^\star), x - \bar{x}\rangle \\
&\quad + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x\rangle \\
&\leq \|\nabla f(x) - \nabla f(x^\star)\|_{Q^{-1}}^2 + \tfrac{1}{4}\|x - \bar{x}\|_Q^2 \\
&\quad + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x\rangle \\
&\leq \langle \nabla f(x) - \nabla f(x^\star), x - x^\star\rangle + \tfrac{1}{4}\|x - \bar{x}\|_Q^2 \\
&\quad + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x\rangle, \\
&= \tfrac{1}{4}\|x - \bar{x}\|_Q^2,
\end{aligned} \tag{18}$$

where we have used the Fenchel-Young inequality for $\frac{1}{4}\|\cdot\|_Q^2$ in the first and (4) in the second inequality, respectively.

Using (18) in (17), along with the skew-symmetricity of K and M, we have

$$\begin{aligned}
0 \leq& \langle -Mz^\star - Fz^\star + Mz + Fz - Hz + H\bar{z}, z^\star - \bar{z}\rangle \\
\leq& \langle (M-K)(z - z^\star) + P(\bar{z} - z), z^\star - \bar{z}\rangle + \tfrac{1}{4}\|x - \bar{x}\|_Q^2 \\
=& \langle (M-K)(z - z^\star) + P(\bar{z} - z), z^\star - z\rangle + \tfrac{1}{4}\|x - \bar{x}\|_Q^2 \\
&+ \langle (M-K)(z - z^\star) + P(\bar{z} - z), z - \bar{z}\rangle \\
=& \langle P(\bar{z} - z), z^\star - z\rangle + \tfrac{1}{4}\|x - \bar{x}\|_Q^2 - \|\bar{z} - z\|_P^2 \\
&+ \langle (M-K)(z - z^\star), z - \bar{z}\rangle \\
=& \tfrac{1}{4}\|x - \bar{x}\|_Q^2 - \|\bar{z} - z\|_P^2 + \langle z - z^\star, (H + M^\top)(z - \bar{z})\rangle.
\end{aligned} \tag{19}$$

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

$$\langle z - z^\star, (H + M^\top)(z - \bar{z})\rangle = \langle z - z^\star, z - Tz\rangle_S. \tag{20}$$

On the other hand, we have $\bar{z} - z = (H + M^\top)^{-1}S(Tz - z)$. Using (11), (13) and (14c) we conclude

$$\|\bar{z} - z\|_P^2 - \tfrac{1}{4}\|\bar{x} - x\|_Q^2 = \|Tz - z\|_{\tilde{P}}^2, \tag{21}$$

where $\tilde{P}$ is defined in (15). Combining (19), (20) and (21) completes the proof. □

Let us define a relaxation matrix $\Lambda = \operatorname{blkdiag}(\Lambda_d, \Lambda_p)$, where $\Lambda_d \in \mathbb{R}^{r\times r}$ and $\Lambda_p \in \mathbb{R}^{n\times n}$ satisfy the following assumption:

Assumption 2 (Parameters)**.**

(i) The matrices Λ, Σ, and Γ are symmetric positive definite.

(ii) Both $\Sigma\Lambda_d$ and $\Gamma\Lambda_p$ are symmetric.

Note that Assumption 2*(ii)* implies that $\Sigma\Lambda_d$ and $\Gamma\Lambda_p$ are positive definite (see [21, Theorem 7.6.3]).

In the case when Λ, Γ, and Σ are selected to be *diagonal* positive definite matrices, Assumption 2 is automatically satisfied. Furthermore, in this case S, defined in (12), is also diagonal. This choice of parameters is exploited in Section III to derive block-coordinate algorithms, which are further used in Section IV to develop a randomized fully distributed algorithm. The proposed primal-dual algorithm is described next:

Algorithm 1 A new primal-dual algorithm

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

Compactly, Algorithm 1 is written as

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

where the operator T is defined in (14) and $z^k := (u^k, x^k)$.

Theorem 1. *Let Assumptions 1 and 2 hold true. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 1. Let $\Lambda = \operatorname{blkdiag}(\Lambda_d, \Lambda_p) \prec 2I$ and assume that the following condition holds:*

$$\Gamma^{-1}(2I - \Lambda_p) - \tfrac{1}{2}Q - L^\top(2I - \Lambda_d)^{-1}\Sigma L \succ 0. \tag{22}$$

Then the sequence $(z^k)_{k\in\mathbb{N}}$ converges to some $(u^\star, x^\star) \in \mathcal{S}$.

Proof. We establish convergence by showing that the sequence $(z^k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathcal{S} = \operatorname{fix} T$. By Assumption 2*(ii)*, $\Lambda^{-1}S$ is symmetric positive definite. For any $z^\star \in \mathcal{S}$ we have

$$\begin{aligned}
\|z^{k+1} - z^\star\|_{\Lambda^{-1}S}^2 &= \|z^k + \Lambda(Tz^k - z^k) - z^\star\|_{\Lambda^{-1}S}^2 \\
&= \|z^k - z^\star\|_{\Lambda^{-1}S}^2 + \|\Lambda(Tz^k - z^k)\|_{\Lambda^{-1}S}^2 \\
&\quad + 2\langle z^k - z^\star, Tz^k - z^k\rangle_S \\
&\leq \|z^k - z^\star\|_{\Lambda^{-1}S}^2 + \|Tz^k - z^k\|_{2\tilde{P} - S\Lambda}^2 \\
&\quad - 2\|Tz^k - z^k\|_{\tilde{P}}^2,
\end{aligned} \tag{23}$$

where the inequality follows from Lemma 1. Next notice that

$$2\tilde{P} - S\Lambda = \begin{bmatrix} \Sigma^{-1}(2I - \Lambda_d) & -L \\ -L^\top & \Gamma^{-1}(2I - \Lambda_p) - \frac{1}{2}Q \end{bmatrix},$$

and by Schur complement it is symmetric positive-definite if and only if (22) holds. This together with (23) shows that the sequence $(z^k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\mathcal{S}$ in the

space equipped with inner product $\langle \cdot, \cdot \rangle_S$. Therefore, $(z^k)_{k\in\mathbb{N}}$ is bounded. Furthermore, it follows from (23) and the fact that $2\tilde{P} - S\Lambda$ is positive definite, that

$$\|Tz^k - z^k\| \to 0. \tag{24}$$

The operator T is continuous (since it involves proximal and linear mappings, that are continuous, and since ∇f is continuous). Let z^c be a cluster point of $(z^k)_{k\in\mathbb{N}}$. It follows from the continuity of T and (24) that $Tz^c - z^c = 0$, *i.e.*, $z^c \in \operatorname{fix} T$. The result follows from Fejér monotonicity of $(z^k)_{k\in\mathbb{N}}$ with respect to $\mathcal{S} = \operatorname{fix} T$ and [18, Theorem 5.5]. □

We next proceed to establish (local) linear convergence rate, often observed in practice. We first recall the notion of *metric subregularity*.

Definition 1 (Metric subregularity)**.** A mapping R is *metrically subregular* at $\bar{x}$ for $\bar{y}$ if $(\bar{x}, \bar{y}) \in \operatorname{gra} R$ and there exists $\eta \in [0, \infty)$, and neighborhoods $\mathcal{U}$ of $\bar{x}$ and $\mathcal{Y}$ of $\bar{y}$ such that

$$d(x, R^{-1}\bar{y}) \leq \eta d(\bar{y}, Rx \cap \mathcal{Y}), \quad \forall x \in \mathcal{U}. \tag{25}$$

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

Metric subregularity of the subdifferential operator has been studied thoroughly and is equivalent to the *quadratic growth condition* [22], [23]. In particular, for a proper closed convex function f, the subdifferential ∂f is metrically subregular at $\bar{x}$ for $\bar{y}$ with $(\bar{x}, \bar{y}) \in \operatorname{gra} \partial f$ if and only if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that the following quadratic growth condition holds [22, Theorem 3.3]:

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

Furthermore, ∂f is strongly subregular at $\bar{x}$ for $\bar{y}$ with $(\bar{x}, \bar{y}) \in \operatorname{gra} \partial f$, if and only if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that [22, Theorem 3.5]:

$$f(x) \geq f(\bar{x}) + \langle \bar{y}, x - \bar{x} \rangle + c\|x - \bar{x}\|^2, \quad \forall x \in \mathcal{U} \tag{26}$$

Note that strongly convex functions satisfy (26), but (26) is *much weaker* than the strong convexity requirement as it only holds in a neighborhood of $\bar{x}$, and only for $\bar{y}$. We refer the reader to [17, Chapter 9] and [24, Chapter 3] for further discussion on metric subregularity.

In Theorem 2 we establish linear convergence under metric subregularity of the monotone operator $D + M + F$. We omit the proof here for length considerations, and note that it follows the same steps as in Theorem 5. In the sequel, we provide sufficient conditions for metric subregularity of $D + M + F$ expressed in terms of conditions on the functions f, g and h; cf. Lemmas 2 and 3.

Theorem 2 (Linear convergence)**.** *Consider Algorithm 1 under the assumptions of Theorem 1. Suppose that $D + M + F$ is metrically subregular at all $z^\star \in \mathcal{S}$ for 0. Then $(d_{\Lambda^{-1}S}(z^k, \mathcal{S}))_{k\in\mathbb{N}}$ converges Q-linearly to zero, and $(z^k)_{k\in\mathbb{N}}$ converge R-linearly to some $z^\star \in \mathcal{S}$*[1].

[1]The sequence $(x_n)_{n\in\mathbb{N}}$ converges to $x^\star$ Q-linearly, with Q-factor $\sigma \in (0, 1)$, if for n sufficiently large $\|x_{n+1} - x^\star\| \leq \sigma \|x_n - x^\star\|$ holds. The sequence $(x_n)_{n\in\mathbb{N}}$ converges to $x^\star$ R-linearly if there is a sequence of nonnegative scalars $(v_n)_{n\in\mathbb{N}}$ such that $\|x_n - x^\star\| \leq v_n$ and $(v_n)_{n\in\mathbb{N}}$ converges to zero Q-linearly.

Lemma 2 provides a sufficient condition for the metric subregularity of $D + M + F$ in terms of the quadratic growth of $f + g$ and h, cf. (26), under which the linear convergence of Algorithm 1 follows from Theorem 2.

Lemma 2. *Let Assumption 1 hold true, and let $z^\star = (u^\star, x^\star) \in \mathcal{S}$. Suppose that $\nabla f + \partial g$ is strongly subregular at $x^\star$ for $-L^\top u^\star$, and ∂h^* is strongly subregular at $u^\star$ for $Lx^\star$. Then $D + M + F$ (cf. (6)) is metrically subregular at $z^\star$ for 0. Furthermore, if such a point exists then the set of primal-dual solutions is a singleton, $\mathcal{S} = \{z^\star\}$.*

Proof. See Appendix A. □

Our next objective is to show that the metric subregularity of $D + M + F$ holds everywhere 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, however, linear convergence can still be established. Let us recall the definition of PLQ functions [17].

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

The class of PLQ functions is closed under scalar multiplication, addition, conjugation and Moreau envelope [17]. A wide range of functions used in optimization applications belong to this class, for example: affine functions, quadratic forms, indicators of polyhedral sets, polyhedral norms such as ℓ_1, and regularizers such as elastic net, Huber loss, hinge loss, and many more.

Lemma 3. *Let Assumption 1 hold true. In addition, assume that f, g and h are piecewise linear-quadratic. Then $D + M + F$ (cf. (6)) is metrically subregular at any z for any z' provided that $(z, z') \in \operatorname{gra}(D + M + F)$.*

Proof. Since f, g and h are closed, proper, convex PLQ, the subdifferentials ∂g, ∇f and ∂h^* are piecewise polyhedral mappings [17, Proposition 12.30(b), Theorem 11.14(b)]. The graph of M is polyhedral, since M is linear. Therefore, the sum $D+M+F$ is also a piecewise polyhedral mapping. Since the inverse of a piecewise polyhedral mapping is piecewise polyhedral, it follows from [24, Proposition 3H.1 and 3H.3] that $D + M + F$ is metrically subregular at z for any z', provided that $(z, z') \in \operatorname{gra}(D + M + F)$. □

Vũ-Condat Primal-Dual Algorithm

In this section, we show how the primal-dual algorithm of Vũ and Condat [2], [3] can be recovered using the operator defined in (8). This analysis yields less restrictive stepsizes conditions and provide further insights into how Algorithm 1 differs from other closely related primal-dual algorithms. In particular, we highlight the non-diagonal structure of S for the Vũ-Condat algorithm (compare (12) and (28)).

In [3] the author considers problem (1), while in [2], the nonsmooth term h, is replaced with $h \,\square\, l$ where l is strongly convex and $\square$ denotes the infimal convolution [18].

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x)+g(x)+(h\,\square\, l)(Lx) \tag{27}$$

In this setting, in addition to Assumption 1, we assume the function l to be strongly convex with respect to the metric $R^{-1}\succ 0$ or equivalently, ∇l^* to be Lipschitz continuous with respect to the metric R.

In both [2], [3], the metrics Q and R are assumed to be scalar. In [4] the authors propose a variable metric version of the algorithm with a preconditioning that accounts for the general Lipschitz metric. This is done by setting the stepsizes to be proportional to the inverse of the Lipschitz metrics. In comparison, our approach yields a wider range of parameters, where the stepsizes are not limited to be proportional to the inverse of the Lipschitz metric, and can take *larger* values thus achieving generally faster convergence.

The Vũ-Condat scheme can be cast in the setting of (8), (9), for a different selection of $H=P+K$ and S:

$$S=\begin{bmatrix}\Sigma^{-1} & L\\ L^\top & \Gamma^{-1}\end{bmatrix}, \tag{28}$$

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

Noticing that $S^{-1}(H+M^\top)=I$, the operator defined in (8), (9) becomes:

$$\bar{u}=\operatorname{prox}_{h^*}^{\Sigma^{-1}}(u-\Sigma\nabla l^*(u)+\Sigma Lx) \tag{29a}$$

$$\bar{x}=\operatorname{prox}_{g}^{\Gamma^{-1}}(x-\Gamma\nabla f(x)-\Gamma L^\top(2\bar{u}-u)) \tag{29b}$$

$$Tz=(\bar{u},\bar{x}). \tag{29c}$$

The proof of the next lemma and theorem follow the same lines as Lemma 1 and Theorem 1 and therefore are omitted.

Lemma 4. *Consider the operator T in (29). Suppose that Σ and Γ are such that*

$$\tilde{P}\coloneqq\begin{bmatrix}\Sigma^{-1}-\frac{1}{4}R & L\\ L^\top & \Gamma^{-1}-\frac{1}{4}Q\end{bmatrix}\succ 0.$$

Then for any $z^\star\in\mathcal{S}$ and any $z\in\mathbb{R}^{n+r}$ we have

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

The iterates of the algorithm are written as follows:

$$z^{k+1}=z^k+\lambda(Tz^k-z^k), \tag{30}$$

where T is defined in (29). When Σ and Γ are scalar the algorithm of [2] is recovered; if in addition $l=\delta_{\{0\}}$, it reduces to the algorithm of [3, Algorithm 3.2] (where in both cases a fixed relaxation parameter λ is used.)

It is possible to replace λ with a matrix Λ similar to Algorithm 1 under the assumption that $S\Lambda$ is symmetric positive definite. However, this is not useful given that (unlike Section II), S *is not block diagonal.*

Theorem 3. *Let Assumption 1 hold true. In addition, assume that $\nabla l^\star$ is Lipschitz continuous with respect to the metric $R\succ 0$. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated according to (30). Let $\lambda\in(0,2)$, and assume that the following conditions hold:*

(i) $\Sigma^{-1}-\frac{1}{2(2-\lambda)}R\succ 0.$

(ii) $\Gamma^{-1}-\frac{1}{2(2-\lambda)}Q-L^\top\left(\Sigma^{-1}-\frac{1}{2(2-\lambda)}R\right)^{-1}L\succ 0.$

Then the sequence $(z^k)_{k\in\mathbb{N}}$ converges to some $(u^\star,x^\star)\in\mathcal{S}$.

When $l=\delta_{\{0\}}$, problem (27) boils down to problem (1), and the convergence conditions of Theorem 3 become

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

This generalizes the result in [3, Theorem 3.1] and [5, Proposition 5.1(ii)] where the Lipschitz metric and the stepsizes are assumed to be scalar. We emphasize that modifying the proof of [3, Theorem 3.1] to account for the metric of Lipschitz continuity and stepsize matrices leads to *more conservative* conditions, even in the scalar case (see [5, Remark 5.6]).

In order to compare our results to [4], [11], we note that the stepsize matrices in their approach are fixed to be $\Gamma=\mu Q^{-1}$ and $\Sigma=\nu R^{-1}$ for some $\mu,\nu>0$. With such a special choice the conditions of Theorem 3 simplify to $\lambda\in(0,2-\frac{\nu}{2})$ and

$$(\mu^{-1}-\tfrac{1}{2(2-\lambda)})(\nu^{-1}-\tfrac{1}{2(2-\lambda)})Q-L^\top R^{-1}L\succ 0, \tag{31}$$

whereas the condition required in [4], [11] is $\lambda\in(0,1]$ and

$$\frac{\delta}{1+\delta}>\frac{\max\{\mu,\nu\}}{2}\ \text{with}\ \delta=\tfrac{1}{\sqrt{\nu\mu}}\|R^{-1/2}LQ^{-1/2}\|^{-1}-1. \tag{32}$$

It is not hard to check that our condition, (31), is always less restrictive than (32). As an example, let $R^{-1/2}LQ^{-1/2}=I$ and set $\lambda=1$, $\mu=1.5$, then (31) simplifies to $\nu<\frac{1}{6.5}$ whereas (32) becomes $\nu<\frac{1}{24}$.

III. A Randomized Block-Coordinate Algorithm

Throughout this section $(\Omega,\mathcal{F},\mathbb{P})$ is a probability space, where Ω, $\mathcal{F}$ and $\mathbb{P}$ denote the sample space, σ-algebra, and probability measure, respectively. Let $\mathscr{F}=(\mathcal{F}_k)_{k\in\mathbb{N}}$ be a *filtration of $\mathcal{F}$, i.e.*, a sequence of sub-sigma algebras of $\mathcal{F}$ such that $\mathcal{F}_k\subset\mathcal{F}_{k+1}$ for all $k\in\mathbb{N}$. The set of sequences of $[0,+\infty)$-valued random variables is denoted by $\ell_+(\mathscr{F})$ and $\ell_+^1(\mathscr{F})\coloneqq\{(\xi^k)_{k\in\mathbb{N}}\in\ell_+(\mathscr{F})\mid\sum_{k\in\mathbb{N}}\xi^k<\infty\ \text{a.s.}\}$. The conditional expectation $\mathbb{E}[\cdot\mid\mathcal{F}_k]$ is denoted by $\mathbb{E}_k[\cdot]$, while *almost surely* is abbreviated as a.s.

We proceed to devise a randomized *block-coordinate* version of our algorithm where the 'coordinates' are the primal and dual variables in Algorithm 1. Let us fix a partitioning of the primal and dual variables into m blocks of coordinates. Let $\mathcal{I}=\{1,\dots,m\}$ and denote by $2^{\mathcal{I}}$ its power set. Set

$$U_i\coloneqq\operatorname{blkdiag}(U_{i,d},U_{i,p}),\quad i\in\mathcal{I}$$

where $U_{i,p}\in\mathbb{R}^{n\times n}$ is a diagonal matrix with 0-1 entries such that $U_{i,p}x$ selects a subset of primal variables (and sets the rest equal to zero), and similarly, $U_{i,d}\in\mathbb{R}^{r\times r}$ is a diagonal matrix that selects a subset of dual variables.

At the kth iteration, the algorithm draws a random activation vector $\epsilon^k\in\{0,1\}^m$, which determines which *blocks of coordinates* will be updated. We assume that ϵ^k is drawn

from some $\Phi \subseteq 2^{\mathcal{I}}$, following probability measure $\mathbb{P}$. We summarize our assumptions next.

Assumption 3.

(i) The matrices $U_{i,p}$ and $U_{i,d}$ are such that

$$\sum_{i=1}^{m} U_{i,p} = I_n, \quad \sum_{i=1}^{m} U_{i,d} = I_r.$$

Furthermore, U_i satisfy $U_i U_j = 0$ for all $i \neq j$.

(ii) $\epsilon^k = (\epsilon_i^k)_{i=\{1,\dots,m\}}$ is an identically distributed Φ-valued random variable such that for each $i \in \mathcal{I}$

$$\sum_{\varepsilon \in \Phi, \varepsilon_i = 1} \mathbb{P}(\epsilon^k = \varepsilon) = p_i > 0.$$

Furthermore, ϵ^k, ϵ^l are independent for $k \neq l$.

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

Let us also define the coordinate activation probability matrix Π, and relaxation matrix $\Lambda = \mathrm{blkdiag}(\Lambda_d, \Lambda_p)$

$$\Pi = \sum_{i=1}^{m} p_i U_i, \quad \Lambda_p = \sum_{i=1}^{m} \lambda_i U_{i,p}, \quad \Lambda_d = \sum_{i=1}^{m} \lambda_i U_{i,d}, \quad (33)$$

where λ_i is the relaxation parameter used in updating coordinate block $i \in \mathcal{I}$.

The Block-Coordinate (BC) primal-dual algorithm is as follows:

Algorithm 2 Block-coordinate algorithm

Inputs: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$
for $k = 0, \dots$ **do**
Select Φ-valued r.v. $\epsilon^k \in \{0,1\}^m$
Calculate Tz^k according to (14)
$z^{k+1} = z^k + \sum_{i=1}^m \epsilon_i^k \lambda_i U_i (Tz^k - z^k)$

The random model we consider is very general, and can capture many randomized mechanisms of selecting primal and dual variables, at each iteration. To showcase, consider two important cases:

- Several active coordinates: $\Phi = \{0,1\}^m$. This model is equivalent to the case in which at each iteration coordinate i is randomly activated with probability p_i independent of other coordinates (cf. Assumption 3*(ii)*).
- Single active coordinate:

$$\Phi = \{(1,0,\dots,0),(0,1,0,\dots,0)\dots,(0,\dots,0,1)\}. \quad (34)$$

In this case, only one coordinate is updated at each iteration: coordinate i will be updated with probability p_i. In addition, notice that in this case the probabilities must satisfy $\sum_{i=1}^m p_i = 1$.

The next lemma establishes a fundamental equality for our scheme by exploiting the diagonal structure of S, Λ and Π.

Lemma 5. *Let Assumptions 1 and 3 hold true. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2. Let $\lambda_i \in (0,2)$ for $i = 1,\dots,m$ and assume that (22) holds. Consider $\tilde{P}$ defined in (15). Then for any $z^\star \in \mathcal{S}$ we have:*

$$\mathbb{E}_k\left[\|z^{k+1} - z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}\right] \leq \|z^k - z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S} - \|Tz^k - z^k\|^2_{2\tilde{P}-S\Lambda} \quad (35)$$

Proof. See Appendix A. □

Remark 1. The relaxation matrix Λ can be selected as arbitrary diagonal so long as it is positive definite and $\Lambda \prec 2I$; we select the same relaxation parameter λ_i for all variables in block i, as this simplifies our notation. Additionally, note that in Algorithm 2 the probabilities p_i are fixed, *i.e.*, the matrix Π is constant through the iterations. This is a non-restrictive assumption and can be relaxed by considering varying probabilities p_i^k and replacing λ_i in Algorithm 2 with, for instance, $\frac{\lambda_i}{m p_i^k}$. This modification results in the following stochastic Fejér monotonicity:

$$\mathbb{E}_k\left[\|z^{k+1} - z^\star\|^2_{\Lambda^{-1}S}\right] \leq \|z^k - z^\star\|^2_{\Lambda^{-1}S} - \|Tz^k - z^k\|^2_{\frac{1}{m}(2\tilde{P} - \frac{1}{m}S\Lambda\Pi(k)^{-1})},$$

where $\Pi(k)$ denotes the probability matrix at iteration k, defined as in (33) using p_i^k. The subsequent convergence analysis holds with minor modifications and is omitted. ◊

The proof of the next lemma is based on the Robbins-Siegmund lemma [25], and is similar to that of [9, Theorem 3].

Theorem 4. *Let Assumptions 1 and 3 hold true. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2. Let $\lambda_i \in (0,2)$ for $i = 1,\dots,m$ and assume that (22) holds. Then the sequence $(z^k)_{k\in\mathbb{N}}$ converges almost surely to an $\mathcal{S}$-valued random variable, i.e., converges to some $(u^\star, x^\star) \in \mathcal{S}$ a.s.*

Proof. From (35) and the Robbins-Siegmund lemma, [25], we have that $\|z^k - z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}$ converges a.s. to a $[0,\infty)$-valued random variable. By the same argument as in [10, Proposition 2.3(iii)], there exists an event $\tilde{\Omega}$ such that $\mathbb{P}(\tilde{\Omega}) = 1$ and, for every $\omega \in \tilde{\Omega}$ and every $z^\star \in \mathcal{S}$,

$$(\|z^k(\omega) - z^\star\|_{\Lambda^{-1}\Pi^{-1}S})_{k\in\mathbb{N}} \text{ converges},$$

and thus $(z_k(\omega))_{k\in\mathbb{N}}$ is bounded.

A second consequence of the Robbins-Siegmund lemma is that $\sum_{k=0}^{\infty} \|Tz^k - z^k\|^2_{2\tilde{P}-S\Lambda} < +\infty$ a.s., *i.e.*, $\|Tz^k - z^k\|^2_{2\tilde{P}-S\Lambda}$ converges to zero a.s. Thus, there exists $\hat{\Omega}$ with $\mathbb{P}(\hat{\Omega}) = 1$ and, for every $\omega \in \hat{\Omega}$ we have

$$Tz^k(\omega) - z^k(\omega) \to 0. \quad (36)$$

The mapping T is continuous since it is defined by proximity operators. With a slight abuse of the notation, define $\Omega = \hat{\Omega} \cap \tilde{\Omega}$, where $\mathbb{P}(\Omega) = 1$. Let $\omega \in \Omega$ and take z^c to be a cluster point of $z^k(\omega)$. It follows from continuity of T and (36) that $Tz^c - z^c = 0$, *i.e.*, $z^c \in \operatorname{fix} T$, whence $z^c \in \mathcal{S}$. Since $(\|z^k(\omega) - z^\star\|_{\Lambda^{-1}\Pi^{-1}S})_{k\in\mathbb{N}}$ converges for every $z^\star \in \mathcal{S}$ we conclude that $\|z^k(\omega) - z^c\|_{\Lambda^{-1}\Pi^{-1}S}$ converges. Thus $\|z^k(\omega) - z^c\|_{\Lambda^{-1}\Pi^{-1}S}$ converges to zero since it does so on some subsequence. □

Theorem 5 (Linear Convergence)**.** *Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 under the assumptions of Theorem 4. Suppose that $D + M + F$ is metrically subregular at all $z^\star \in \mathcal{S}$ for 0. Then almost surely $(d_{\Lambda^{-1}\Pi^{-1}S}(z^k, \mathcal{S}))_{k\in\mathbb{N}}$ converges Q-linearly in expectation to zero.*

Proof. See Appendix A. □

We remark that the result of Theorem 5 holds if either (1): f, g, h are piecewise linear-quadratic (see Lemma 3), or if (2): the growth conditions in Lemma 2 hold at all $z^\star \in \mathcal{S}$.

IV. Distributed Optimization

In this section, we devise both synchronous and asynchronous algorithms for the multi-agent optimization problem (3). The synchronous version is based on Algorithm 1 while the asynchronous one is based on Algorithm 2 in the case $\Phi = \{0,1\}^m$. For length considerations other possibilities of random activation are not presented here. For example, if Φ is set as in (34), the analysis can be carried out via *exponential clocks* [26], [27]: agents are assumed to 'wake-up' based on independent exponentially distributed tick-down timers.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph over a vertex set $V = \{1, \dots, m\}$ with edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Each node $i \in \mathcal{V}$ is associated with an *agent*, and maintains its own local primal variable $x_i \in \mathbb{R}^{n_i}$ and dual variables $y_i \in \mathbb{R}^{r_i}$ and $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$, where the former corresponds to L_i and the latter is the local dual variable of agent i corresponding to the edge-constraint (3b) for $(i,j) \in \mathcal{E}$. We assume that agent i can both send and receive information from its neighbors $j \in \mathcal{N}_i := \{j \in \mathcal{V} \mid (i,j) \in \mathcal{E}\}$. The only information that agent i shares with its neighbor j is $A_{ij}x_i$, along with edge variable $w_{(i,j),i}$. The cost functions f_i, g_i, h_i, along with the matrix L_i and all other variables are kept *private*.

Let us restate the distributed optimization problem (3):

$$\underset{\mathsf{x} \in \mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^m f_i(x_i) + g_i(x_i) + h_i\left(L_i x_i\right) \tag{37a}$$

$$\text{subject to} \quad A_{ij}x_i + A_{ji}x_j = b_{(i,j)} \qquad (i,j) \in \mathcal{E} \tag{37b}$$

where $\mathsf{x} = (x_1, \dots, x_m)$. For every $i = 1, \dots, m$, $x_i \in \mathbb{R}^{n_i}$, and for every $(i,j) \in \mathcal{E}$, $b_{(i,j)} \in \mathbb{R}^{l_{(i,j)}}$. Let the following assumptions hold:

Assumption 4. For $i = 1, \dots, m$:

(i) $A_{ij} \in \mathbb{R}^{l_{(i,j)} \times n_i}$ for $j \in \mathcal{N}_i$, and $L_i \in \mathbb{R}^{r_i \times n_i}$.

(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.

(iii) $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ are convex, continuously differentiable, and ∇f_i are Lipschitz continuous with respect to the metric $Q_i \succ 0$, *i.e.*, for all $x, y \in \mathbb{R}^{n_i}$

$$\|\nabla f_i(x) - \nabla f_i(y)\|_{Q_i^{-1}} \le \|x - y\|_{Q_i}.$$

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

(v) The set of solutions of (37) is nonempty. Moreover, there exists $x_i \in \operatorname{ri}\operatorname{dom} g_i$ such that $L_i x_i \in \operatorname{ri}\operatorname{dom} h_i$, for $i = 1, \dots, m$, and $A_{ij}x_i + A_{ji}x_j = b_{(i,j)}$ for $(i,j) \in \mathcal{E}$.

For each edge (i,j), let $\kappa_{(i,j)} > 0$ denote the corresponding weight that is inherent to the communication graph. In essence, this can be used to capture edge's 'fidelity,' for example the channel quality in a communication link. Edge weights affect agents' stepsizes, cf. (41).

Define the linear operator

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

We define $\mathsf{N} \in \mathbb{R}^{2\sum_{(i,j)\in\mathcal{E}} l_{(i,j)} \times \sum_{i=1}^m n_i}$ 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 for each constraint, where $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$ is maintained by agent i and $w_{(j,i),j} \in \mathbb{R}^{l_{(i,j)}}$ by agent j.

Consider the set

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

The problem (37) can be expressed as

$$\begin{aligned}\text{minimize} \sum_{i=1}^m \left(f_i(x_i) + g_i(x_i) + h_i\left(L_i x_i\right)\right) \\ + \sum_{(i,j)\in\mathcal{E}} \delta_{C_{(i,j)}}\left(N_{(i,j)}\mathsf{x}\right)\end{aligned} \tag{38}$$

where δ_X denotes the indicator function of a closed nonempty convex set.

Let $C = \bigtimes_{(i,j)\in\mathcal{E}} C_{(i,j)}$, $L = \operatorname{blkdiag}(L_1, \dots, L_m)$, and $\mathsf{L}\mathsf{x} = (L\mathsf{x}, \mathsf{N}\mathsf{x}) =: (\tilde{\mathsf{y}}, \tilde{\mathsf{w}}) \in \mathbb{R}^{n_d}$ with $n_d = \sum_{(i,j)\in\mathcal{E}} 2l_{(i,j)} + \sum_{i=1}^m r_i$. Rewrite problem (38) in the following compact form:

$$\text{minimize} \;\; \mathsf{f}(\mathsf{x}) + \mathsf{g}(\mathsf{x}) + \tilde{\mathsf{h}}(\mathsf{L}\mathsf{x}) \tag{39}$$

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 this section, $\mathcal{S}$ refers to the primal-dual solution of (39). As in Section II the primal-dual optimality conditions are written in the form of monotone inclusion (7) with

$$D : (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (\partial\mathsf{h}^*(\mathsf{y}), \partial\delta_C^*(\mathsf{w}), \partial\mathsf{g}(\mathsf{x})), \tag{40a}$$

$$M : (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (-L\mathsf{x}, -\mathsf{N}\mathsf{x}, L^\top\mathsf{y} + \mathsf{N}^\top\mathsf{w}), \tag{40b}$$

$$F : (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (0, 0, \nabla\mathsf{f}(x)), \tag{40c}$$

where $\mathsf{u} = (\mathsf{y}, \mathsf{w})$ is the dual vector. Let us define the edge weight matrix as follows

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

where the weights $\kappa_{(i,j)}$ are repeated twice, once for each agent. Furthermore, set

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

where $\sigma_i, \tau_i > 0$ are the stepsizes, and Q_i for $i = 1, \dots, m$ are defined in Assumption 4*(iii)*.

The next step is to apply Algorithms 1 and 2 to (39). We simplify the proximal updates in (14) using separability of the involved functions.

Algorithm 3 Synchronous & asynchronous distributed primal-dual algorithm

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

I: Synchronous version

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

II: Asynchronous version

draw r.v. ϵ_i^k according to $\mathbb{P}(\epsilon_i^0 = 1) = p_i$
for all agents i with $\epsilon_i^k = 1$ **do**

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

Since $\operatorname{prox}_{\tilde{\mathsf{h}}^\star}(\mathsf{y},\mathsf{w}) = (\operatorname{prox}_{\mathsf{h}^\star}(\mathsf{y}), \mathsf{w} - \mathcal{P}_C(\mathsf{w}))$ (using $\operatorname{prox}_{\delta_C}(\cdot) = \mathcal{P}_C(\cdot)$ along with Moreau decomposition) we have

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

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

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

By assigning the coordinates x_i, y_i and $w_{(i,j),i}$ for all $j \in \mathcal{N}_i$ to agent i we propose Algorithm 3. In the synchronous version of Algorithm 3, at each iteration all the agents must update their values. In the asynchronous version, at every iteration agents wake up randomly. In both versions agent i only requires $A_{ji}x_j^k$ and $w_{(i,j),j}^k$ from neighbors $j \in \mathcal{N}_i$ to compute $\bar{w}_{(i,j),i}^k$. Notice that when an agent is activated, it updates its values and subsequently broadcasts the relevant information to its neighbors, then goes idle until next activation.

The next theorem establishes convergence of our distributed algorithm based on the theory developed in Sections II and III.

Theorem 6. *Let Assumption 4 hold true. Denote by $\mathcal{S}$ the set of primal-dual solutions of (37). Consider the sequences $(\mathsf{x}^k)_{k\in\mathbb{N}} = (x_1^k, \dots, x_m^k)_{k\in\mathbb{N}}$, $(\mathsf{y}^k)_{k\in\mathbb{N}} = (y_1^k, \dots, y_m^k)_{k\in\mathbb{N}}$ and $(\mathsf{w}^k)_{k\in\mathbb{N}} = ((w_{(i,j)})_{(i,j)\in\mathcal{E}}^k)_{k\in\mathbb{N}}$ generated by Algorithm 3-I (or 3-II), and set $(z^k)_{k\in\mathbb{N}} = (\mathsf{y}^k, \mathsf{w}^k, \mathsf{x}^k)_{k\in\mathbb{N}}$. Suppose that the following stepsize condition holds:*

$$
\tau_i < \frac{2 - \lambda_i}{\frac{\|Q_i\|}{2} + \frac{1}{2-\lambda_i}\|\sigma_i L_i^\top L_i + \sum_{j\in\mathcal{N}_i} \kappa_{(i,j)} A_{ij}^\top A_{ij}\|}. \tag{41}
$$

Then, in the case of Algorithm 3-I $(z^k)_{k\in\mathbb{N}}$ converges to $z^\star$, and in the case of Algorithm 3-II it converges a.s. to a $z^\star$-valued random variable, for some $z^\star \in \mathcal{S}$.

An important feature of Algorithm 3 is that according to (41) the stepsizes τ_i, σ_i, and the relaxation parameter λ_i for each agent only depend on the local parameters such as $\|Q_i\|$, the edge weights, $\kappa_{(i,j)}$, and the linear operators L_i, and A_{ij}, which are all known to agent i and require no global coordination.

The linear convergence of Algorithm 3-I (or 3-II) is established next under the metric subregularity assumption.

Theorem 7 (Linear Convergence)**.** *Consider Algorithm 3-I (or 3-II), under the assumptions of Theorem 6. Set $(z^k)_{k\in\mathbb{N}} = (\mathsf{y}^k, \mathsf{w}^k, \mathsf{x}^k)_{k\in\mathbb{N}}$. Suppose that $D + M + F$, defined in (40), is metrically subregular at all $z^\star \in \mathcal{S}$ for 0. Then*

(i) In the case of Algorithm 3-I, $(d_{\Lambda^{-1}S}(z^k, \mathcal{S}))_{k\in\mathbb{N}}$ converges Q-linearly to zero, and $(z^k)_{k\in\mathbb{N}}$ converges R-linearly to some $z^\star \in \mathcal{S}$.

(ii) In the case of Algorithm 3-II, a.s. $(d_{\Lambda^{-1}\Pi^{-1}S}(z^k, \mathcal{S}))_{k\in\mathbb{N}}$ converges Q-linearly in expectation to zero.

Furthermore, if the functions f_i, g_i and h_i, for $i = 1, \dots, m$, are piecewise linear-quadratic then metric subregularity assumption holds at any z for any z' provided that $(z, z') \in \operatorname{gra}(D + M + F)$.

V. Application: Network Utility Maximization

In this section, we showcase an application of our distributed algorithm to *Network Utility Maximization* (NUM), a popular framework used for resource allocation in multi-commodity networks, with applications ranging from wireline and wireless communication networks [28], to industrial assembly lines and smart transportation systems. For illustration, we assume a wireless communication network with topology captured by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where node i can transmit data to node j if and only if $(i,j) \in \mathcal{E}$. For a node $i \in \mathcal{V}$ we define its neighborhood $\mathcal{N}_i := \{j \mid (i,j) \in \mathcal{E}\}$.

Let $\subseteq \mathcal{V} \times \mathcal{V}$ be a set of source-destination pairs, alternatively called *users*, which correspond to end-to-end traffic flows in the network. For simplicity, we allow for two users to share a common source node or destination node, but not both. We denote by $\mathcal{S} \subseteq \mathcal{V}$ and $\mathcal{D} \subseteq \mathcal{V}$ the set of source and destination nodes respectively. For every $i \in \mathcal{S}$, $\mathcal{D}_i$ denotes the set of nodes $d \in \mathcal{D}$ for which $(i,d) \in \mathcal{F}$, *i.e.*, the set of destination nodes whose source is node i. If $i \notin \mathcal{S}$, *i.e.*, if i is not the source of any user, then we define $\mathcal{D}_i = \varnothing$. For $i \in \mathcal{V}$, we denote by $\mathcal{D}_{-i} := \mathcal{D} \setminus \{\mathcal{D}_i \cup \{i\}\}$ the set of destination nodes different than i with a source other than i. For each $i \in \mathcal{S}$, $d \in \mathcal{D}_i$, we use x_i^d to denote the *flow rate* for user $(i,d) \in \mathcal{F}$.

For each $(i,j) \in \mathcal{E}$, we call the directed pair ij the *link* with node i being the transmitter and node j the receiver. We define

link rates, *separately for each destination*[2] [28] as follows: for $(i,j) \in \mathcal{E}$, let r_{ij}^d be the transmission rate from node i to node j, cumulative over all flows with common destination d; these are decision variables related to scheduling (MAC) as well as dynamic routing (when user routes are not determined beforehand). The total link rate (over all flows) is given by $r_{ij} := \sum_d r_{ij}^d$; without any loss of generality, we adopt a link bandwidth constraint $r_{ij} \leq 1$.

Each user $(i,d) \in$ is assigned a utility function U_i^d of its rate x_i^d, which is assumed to be upper-semicontinuous, proper, concave and non-decreasing. Indicative choices are $U_i^d(x) = wU(x)$ with $w \geq 0$ and U is chosen as [28]: a) $U(x) = \log x$ (proportional fairness), b) $U(x) = (1-\alpha)^{-1}x^{1-\alpha},\ 0 < \alpha < 1$ (generalized fairness), and many more. The goal of NUM is twofold: i) the users $(i,d) \in$ select the flow rates $\{x_i^d\}$ (congestion control), and ii) the nodes $i \in V$ select the link rates $\{r_{ij}^d\}$ (joint routing and MAC), so as to maximize the sum of user utilities subject to *stability* and *interference* constraints.

The NUM problem that we consider is given by:

$$\begin{aligned}
\underset{\{x_i^d\},\{r_{ij}^d\}}{\text{minimize}} \ & \sum_{i\in\mathcal{S}}\sum_{d\in\mathcal{D}_i} -U_i^d(x_i^d) \\
\text{subject to} \ & \sum_{j\in\mathcal{N}_i} r_{ji}^d + x_i^d \leq \sum_{j\in\mathcal{N}_i} r_{ij}^d,\ \ i\in\mathcal{S},\, d\in\mathcal{D}_i && \text{(42a)} \\
& \sum_{j\in\mathcal{N}_i} r_{ji}^d \leq \sum_{j\in\mathcal{N}_i} r_{ij}^d,\ \ i\in\mathcal{V},\, d\in\mathcal{D}_{-i} && \text{(42b)} \\
& \sum_{d\in\mathcal{D}}\sum_{j\in\mathcal{N}_i} (r_{ij}^d + r_{ji}^d) \leq 1,\ \ i\in\mathcal{V},\, d\in\mathcal{D} && \text{(42c)} \\
& r_{ij}^d \geq 0,\ \ (i,j)\in\mathcal{E},\, d\in\mathcal{D}, \\
& r_{ij}^d = 0,\ \ (i,j)\in\mathcal{E},\, i=d\in\mathcal{D}, && \text{(42d)} \\
& x_i^d \geq 0,\ \ i\in\mathcal{S},\, d\in\mathcal{D}_i
\end{aligned}$$

Notice that constraint (42d) captures that a user destination is a sink node with no outgoing traffic for that flow.

Stability refers to the requirement for bounded queues (*i.e.*, bounded delays), and is defined separately for each destination d: for each node $i \in \mathcal{V}$ the cumulative incoming rate of traffic is less than or equal to the outgoing traffic. This is captured by the constraints (42a), (42b).

In case the routes are predetermined, we may use a *link-centric* model: for each user f let H^f be a $M-$dimensional 0-1 vector indicating whether each link is used by user f, *i.e.*, $H_{ij}^f = 1$ if the route for flow f contains link ij, and $H_{ij}^f = 0$, else. Then, each link maintains a single queue and the stability is captured by the (in-flow $\leq$ out-flow) constraint:

$$\sum_{f\in} H_{ij}^f x_f \leq r_{ij}.$$

Interference captures conflict constraints in simultaneous transmissions over the same frequency band. We adopt here a simplistic interference model (see [28] for a more general model) in which a node *cannot* simultaneously a) transmit data to more than one nodes, b) receive data from more than one nodes, and c) transmit and receive at the same time, cf. (42c).

The traditional cross-layer optimization [28] amounts to solving the above minimization via dual decomposition, which gives rise to the celebrated backpressure (or MaxWeight) algorithm [28], [29]. The algorithm has a natural intepretation of Lagrange multipliers as queue lengths. Nevertheless, in general (including the simplified interference model that we consider here), it requires solving a combinatorial problem at each iteration which is not amenable to distributed implementation. In contradistinction, our method is fully distributed and entails simple operations.

We now show how to cast NUM into the setup of Problem (37). For each node i, we define the primal vector z_i to contain the incoming and outgoing link rates, as well as the exogenous flow rates (if some flows originate at node i). In essence, a link rate variable "belongs" to the node from which it originates. We introduce "replicas," *i.e.*, define $r_{ij,i}^d$ and $r_{ij,j}^d$, where the first variable "belongs" to node i and the second is its estimate at node j. Therefore, the local decision vector for node $i \in \mathcal{V}$ is defined as

$$z_i := \left((r_{ij,i}^d, r_{ji,i}^d)_{j\in\mathcal{N}_i, d\in\mathcal{D}}, (x_i^d)_{d\in\mathcal{D}_i}\right),$$

and its dimension is $n_i = |\mathcal{N}_i||\mathcal{D}| + |\mathcal{D}_i|$. Note that, throughout this section the variables are stacked by increasing order.

The edge-based consensus constraint (3b) is given by:

$$r_{ij,i}^d = r_{ij,j}^d, \quad r_{ji,i}^d = r_{ji,j}^d, \qquad d\in\mathcal{D}, (i,j)\in\mathcal{E}.$$

In terms of the notation defining the generic distributed optimization problem (37), for $(i,j) \in \mathcal{E}$ one has $b_{(i,j)} = 0$, and (assuming $i < j$ without loss of generality):

$$\begin{aligned}
A_{ij} &: z_i \mapsto (r_{ij,i}^d, -r_{ji,i}^d)_{d\in\mathcal{D}}, \\
A_{ji} &: z_j \mapsto (-r_{ij,j}^d, r_{ji,j}^d)_{d\in\mathcal{D}}.
\end{aligned}$$

Utility functions are typically smooth, but may not have Lipschitz gradients so we set $f_i \equiv 0$, and proceed to define g_i[3] to incorporate utilities and non-negativity constraints. The objective function for agent $i \in \mathcal{V}$ is given by:

$$g_i(z_i) = \sum_{d\in\mathcal{D}_i} -U_i^d(x_i^d) + \delta_{Z_i}(z_i).$$

where

$$\begin{aligned}
Z_i = \{ \ & \left((r_{ij,i}^d, r_{ji,i}^d)_{j\in\mathcal{N}_i, d\in\mathcal{D}}, (x_i^d)_{d\in\mathcal{D}_i}\right) \in \mathbb{R}_+^{n_i}, \\
& \text{and } r_{ij,i}^d = 0, i = d \in \mathcal{D}\}.
\end{aligned}$$

We define $h_i : \mathbb{R}^{|\mathcal{D}\setminus\{i\}|+1} \to \overline{\mathbb{R}}$ to capture the constraints (42a)-(42c). Let $h_i = \delta_{C_i}$, with

$$C_i = \left\{(v^{(1)}, v^{(2)}) \in \mathbb{R}^{|\mathcal{D}\setminus\{i\}|} \times \mathbb{R} \mid v^{(1)} \geq 0, v^{(2)} \leq 1\right\}.$$

Therefore, proximal mapping of h_i is inexpensive.

[2]Note that two different users may have the same destination (but different source). It is customary to maintain, for each link, a single *queue* for the cumulative traffic towards a given destination, which justifies the convention for per-destination rates $\{r_{ij}^d\}$.

[3]The negative utility is a scalar function and has therefore inexpensive proximal operator, which in several cases (e.g., logarithm) may be computed analytically.

We proceed by defining $L_i := (L_i^{(1)}, L_i^{(2)})$. Let us first define the following notation:

$$c_i^d(x) = \begin{cases} x & d \in \mathcal{D}_i \\ 0 & d \in \mathcal{D}_{-i} \end{cases}$$

If $i \in \mathcal{S}$ then

$$L_i^{(1)} : z_i \mapsto v_i^{(1)} = \Big(\sum_{j \in \mathcal{N}_i} (r_{ij,i}^d - r_{ji,i}^d) - c_i^d(x_i^d) \Big)_{d \in \mathcal{D} \setminus \{i\}}.$$

Otherwise, if $i \in \mathcal{V} \setminus \mathcal{S}$ then

$$L_i^{(1)} : z_i \mapsto v_i^{(1)} = \Big(\sum_{j \in \mathcal{N}_i} (r_{ij,i}^d - r_{ji,i}^d) \Big)_{d \in \mathcal{D} \setminus \{i\}}.$$

As for, $L_i^{(2)}$ we have

$$L_i^{(2)} : z_i \mapsto v_i^{(2)} = \sum_{d \in \mathcal{D}} \sum_{j \in \mathcal{N}_i} (r_{ij,i}^d + r_{ji,i}^d) \in \mathbb{R}.$$

The transpose of L_i is given by:

$$L_i^\top : ((y_i^d)_{d \in \mathcal{D} \setminus \{i\}}, y_i^{(2)}) \mapsto \big((u_{ij,i}^d, u_{ji,i}^d)_{j \in \mathcal{N}_i, d \in \mathcal{D}}, (-y_i^d)_{d \in \mathcal{D}_i} \big)$$

where

$$u_{ij,i}^d = y_i^{(2)} + y_i^d,\ u_{ji,i}^d = y_i^{(2)} - y_i^d, \quad \forall d \in \mathcal{D} \setminus \{i\},$$

and $u_{ij,i}^d = u_{ji,i}^d = y_i^{(2)}$ if $i = d \in \mathcal{D}$.

It remains to write $\sum_{j \in \mathcal{N}_i} A_{ij}^\top \bar{w}_{(i,j),i}$ explicitly:

$$\sum_{j \in \mathcal{N}_i} A_{ij}^\top \bar{w}_{(i,j),i} = \big((\bar{w}_{ij,i}^d, -\bar{w}_{ji,i}^d)_{j \in \mathcal{N}_i, d \in \mathcal{D}}, (0)_{d \in \mathcal{D}_i} \big). \quad (43)$$

Algorithm 4 summarizes the result of applying Algorithm 3 to the NUM problem. For simplicity, all relaxation parameters λ_i are set equal to 1.

Since for the NUM problem the smooth terms $f_i \equiv 0$, in (41), the term $\|Q_i\|$ vanishes. Furthermore, it is an easy exercise to check that $\|L_i^\top L_i\| = \|L_i L_i^\top\| = 2|\mathcal{N}_i||\mathcal{D}|$. Given the definition of A_{ij}, a sufficient condition for convergence is to have

$$\tau_i < \frac{1}{2\sigma_i |\mathcal{N}_i||\mathcal{D}| + \max_{j \in \mathcal{N}_i} \kappa_{(i,j)}}. \quad (44)$$

In our simulations we consider small, medium, and large networks with $m = 5, 10, 50$ nodes/agents, respectively. In all three cases we consider 5 random flows. The communication graphs are generated randomly according to the Erdős-Renyi model with parameter 0.05. We report the performance of Algorithm 4 for three cases: $U(x) = \log(x)$, $U(x) = (1-\alpha)^{-1} x^{1-\alpha}$, with $\alpha = 0.5$, and linear utility $U(x) = x$.

For the stepsizes, we set $\kappa_{(i,j)} = 2|\mathcal{D}|$ for all $(i,j) \in \mathcal{E}$, and $\sigma_i = 2/|\mathcal{N}_i|$, $\tau_i = 0.99/(6|\mathcal{D}|)$ (cf. (44)). These values were selected empirically based on better performance of the algorithm, noting that in general larger stepsizes yield faster convergence.

First, we consider the synchronous version of the algorithm. The performance of the algorithm is reported in Table 1 and Figure 1. The termination criteria is based on achieving a given relative error, $\|R^k\|/\|R^0\|$, where $R^k = \bar{v}^k - v^k$, and v is the stacked primal and dual variables of all agents. It is observed that with all three utilities Algorithm 4 achieves a moderate relative error fairly quickly, even for the large network. In fact, in most of the cases the residue is becoming three orders of magnitude smaller in less than 3000 iterations. Furthermore, a higher accuracy is harder to achieve in the larger network, Figure 1 (right). This is not unexpected, since Algorithm 4 is a first order method.

In Figure 2 we compare the synchronous and asynchronous versions of the algorithm for the logarithmic utility function in the network with $m = 10$. The x-axis denotes the total number of local updates (note that in the synchronous case at each iteration m local updates are performed). For the asynchronous case, the activation probabilities of all the agents, $i = 1, \ldots, m$, are set equal to $p_i = 0.5$. We observe that the algorithm takes roughly the same number of total local updates as the full version, *i.e.*, even with random activation of agents the algorithm maintains its speed. This behavior is consistent and is observed for other utility functions and larger networks.

VI. Conclusions

The primal-dual algorithm introduced in this paper enjoys several structural features that distinguish it from other related primal-dual algorithms. The main property that has allowed us to develop a block-coordinate version of the algorithm is the diagonal structure of the metric under which Fejér monotonicity is established. In addition, linear convergence is achieved under a metric subregularty assumption that does not require uniqueness of the saddle point solution. The developed algorithms are employed to derive a fully distributed asynchronous algorithm for optimization over graphs. Future works include extending the *SuperMann* scheme, [30], to the block-coordinate case and for quasi-nonexpansive operators. This algorithm enjoys superlinear convergence rates, therefore, we can achieve faster convergence and consequently less communication rounds. Investigating the effects of communication delays and efficient strategies for selecting activation probabilities and stepsizes are other open research directions.

Appendix A
Omitted Proofs

Proof of Lemma 2. From the equivalent characterization of strong subregularity in (26) we have that there exists a neighborhood $\mathcal{U}_{x^\star}$ of $x^\star$ such that for all $x \in \mathcal{U}_{x^\star}$

$$\begin{aligned}(f+g)(x) \geq & (f+g)(x^\star) + \langle -L^\top u^\star, x - x^\star \rangle \\ & + c_1 \|x - x^\star\|^2 \end{aligned} \quad (45)$$

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

$$h^*(u) \geq h^*(u^\star) + \langle Lx^\star, u - u^\star \rangle + c_2 \|u - u^\star\|^2. \quad (46)$$

Fix $z = (u, x)$ with $u \in \mathcal{U}_{u^\star}$ and $x \in \mathcal{U}_{x^\star}$. Consider $v = (v_1, v_2) \in \tilde{T} z := Dz + Mz + Fz$. By definition (cf. (6)) 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:

$$\langle v_1 + Lx, u - u^\star \rangle \geq h^*(u) - h^*(u^\star), \quad (47)$$

$$\langle v_2 - L^\top u, x - x^\star \rangle \geq (f+g)(x) - (f+g)(x^\star). \quad (48)$$

Table 1: Number of iterations to reach $\|R^k\|/\|R^0\| < \epsilon_{tol} = 10^{-2}, 10^{-3}, 10^{-4}$.

	Logarithmic			Linear			Generalized fairness		
m	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}	10^{-2}	10^{-3}	10^{-4}
5	40	137	229	28	145	353	118	1360	1748
10	45	317	663	36	168	350	132	2491	7818
50	290	3210	17760	320	2360	15980	310	21390	95510

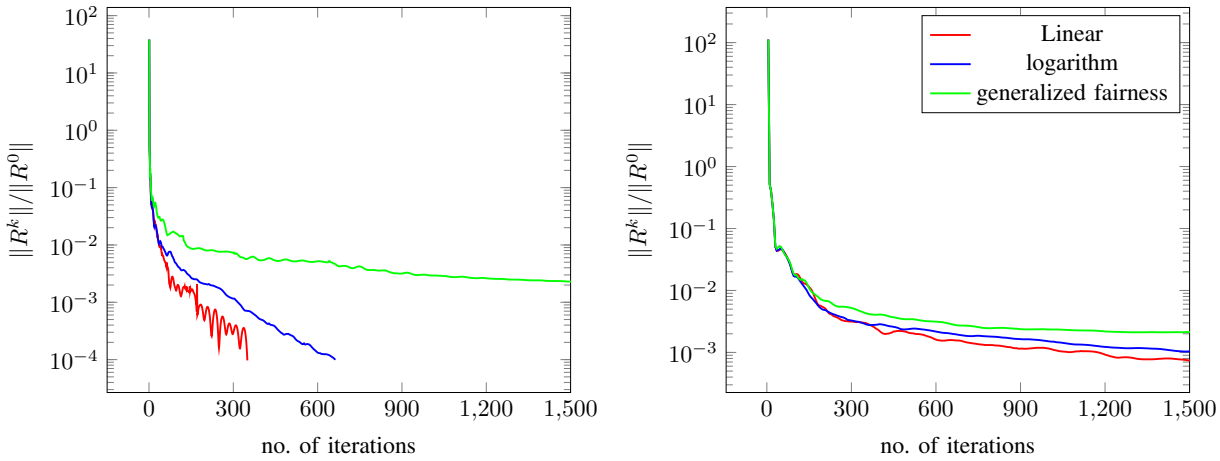

Figure 1: The relative error for $m = 10$ (left), and $m = 50$ (right).

Algorithm 4 Synchronous & asynchronous distributed primal-dual algorithm for NUM

Inputs: $z_i^0 \in \mathbb{R}^{n_i}$, $y_i^0 \in \mathbb{R}^{|\mathcal{D}\setminus\{i\}|+1}$, and $w_{ij,i}^{d,0}, w_{ji,i}^{d,0}, r_{ij,i}^{d,0}, r_{ji,i}^{d,0} \in \mathbb{R}$, for $j \in \mathcal{N}_i$, $d \in \mathcal{D}$, $i = 1, \dots, m$.
for $k = 0, \dots$ **do**

I: Synchronous version

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

II: Asynchronous version
draw r.v. ϵ_i^k according to $\mathbb{P}(\epsilon_i^0 = 1) = p_i$
for all agents i with $\epsilon_i^k = 1$ **do**

Local updates:

$$\bar{w}_{ij,i}^{d,k} = \tfrac{1}{2}\big(w_{ij,i}^{d,k} + w_{ij,j}^{d,k}\big) + \tfrac{\kappa_{(i,j)}}{2}\big(r_{ij,i}^{d,k} - r_{ij,j}^{d,k}\big), \quad \forall j \in \mathcal{N}_i, \forall d \in \mathcal{D}$$
$$\bar{w}_{ji,i}^{d,k} = \tfrac{1}{2}\big(w_{ji,i}^{d,k} + w_{ji,j}^{d,k}\big) + \tfrac{\kappa_{(i,j)}}{2}\big(-r_{ji,i}^{d,k} + r_{ji,j}^{d,k}\big), \quad \forall j \in \mathcal{N}_i, \forall d \in \mathcal{D}$$
$$\bar{y}_i^k = y_i^k + \sigma_i L_i z_i^k - \sigma_i \mathcal{P}_{C_i}\big(\sigma_i^{-1} y_i^k + L_i z_i^k\big)$$
$$z_i^{k+1} = \operatorname{prox}_{\tau_i g_i}\big(z_i^k - \tau_i L_i^\top \bar{y}_i^k - \tau_i \textstyle\sum_{j\in\mathcal{N}_i} A_{ij}^\top \bar{w}_{(i,j),i}^k\big), \text{ where the last term is given by (43).}$$
$$y_i^{k+1} = \bar{y}_i^k + \sigma_i L_i (z_i^{k+1} - z_i^k)$$
$$w_{ij,i}^{d,k+1} = \bar{w}_{ij,i}^{d,k} + \kappa_{(i,j)}(r_{ij,i}^{d,k+1} - r_{ij,i}^{d,k}), \quad \forall j \in \mathcal{N}_i, \forall d \in \mathcal{D}$$
$$w_{ji,i}^{d,k+1} = \bar{w}_{ji,i}^{d,k} - \kappa_{(i,j)}(r_{ji,i}^{d,k+1} - r_{ji,i}^{d,k}), \quad \forall j \in \mathcal{N}_i, \forall d \in \mathcal{D}$$

Broadcast of information:
Send $r_{ij,i}^{d,k+1}$, $r_{ji,i}^{d,k+1}$, $w_{ij,i}^{d,k+1}$, and $w_{ji,i}^{d,k+1}$ to agent j, for all $j \in \mathcal{N}_i$, $d \in \mathcal{D}$.

Combining (47), (48) with (45), (46) 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 $c = \min\{c_1, c_2\}$. Therefore, by the Cauchy-Schwarz inequality $\|v\| \geq c\|z - z^\star\|$. Since $\|z - z^\star\| \geq d(z, \tilde{T}^{-1}0)$, and $v \in \tilde{T}z$ was selected arbitrarily, we have

$$d(z, \tilde{T}^{-1}0) \leq \tfrac{1}{c} d(0, \tilde{T}z), \quad \forall z \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star}$$

which is equivalent to metric subregularity of $\tilde{T}$ at $z^\star$ for 0 [24, Excersize 3H.4].

For the second part, consider the Lagrangian:

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

Sum (45) and (46) to derive

$$\mathcal{L}(u^\star, x) - \mathcal{L}(u, x^\star) \geq c\|z - z^\star\|^2, \quad \forall z \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star} \tag{49}$$

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 (49) at $\bar{z}^\star$ yields $\bar{z}^\star = z^\star$. The convexity of $\mathcal{S}$ concludes the proof. □

Proof of Lemma 5. The update equation in Algorithm 2 can be equivalently written as:

$$z^{k+1} = z^k + \Lambda E^k (Tz^k - z^k), \tag{50}$$

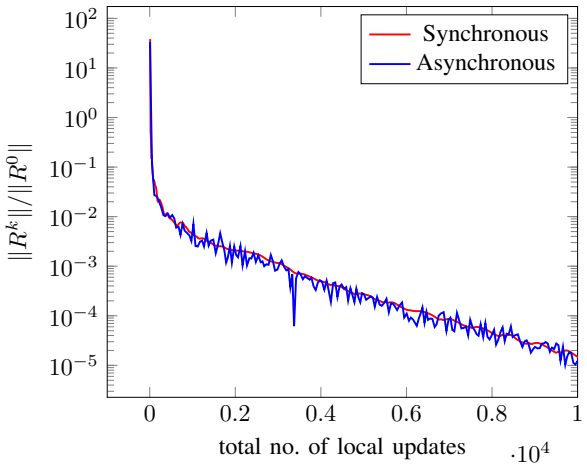

Figure 2: The relative error for the logarithmic utility, synchronous and asynchronous schemes ($m = 10$).

where $E^k = \sum_{i=1}^m \epsilon_i^k U_i$ is a diagonal 0-1 matrix. We have

$$\mathbb{E}_k\left[E^k\right] = \sum_{\varepsilon\in\Psi}\mathbb{P}(\epsilon^k=\varepsilon)\sum_{j=1}^m \varepsilon_j U_j = \sum_{j=1}^m\sum_{\varepsilon\in\Psi}\mathbb{P}(\epsilon^k=\varepsilon)\varepsilon_j U_j$$
$$= \sum_{j=1}^m \sum_{\varepsilon\in\Psi,\varepsilon_j=1}\mathbb{P}(\epsilon^k=\varepsilon)U_j = \sum_{j=1}^m p_j U_j = \Pi,$$

where we used Assumptions 3*(i)* and 3*(ii)*. Furthermore, it is plain to check that E^k is symmetric and idempotent, i.e., $E^k = (E^k)^\top = (E^k)^2$. Therefore

$$\mathbb{E}[E^k] = \mathbb{E}[(E^k)^\top E^k] = \Pi. \tag{51}$$

Given the update (50) we have:

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

where we used (51) and that Λ, Π are diagonal. Lemma 1 completes the proof. We have

$$\begin{aligned}
&\mathbb{E}_k\left[\|z^{k+1}-z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}\right]\\
&=\mathbb{E}_k\left[\|z^k+\Lambda\sum_{j=1}^m\epsilon_j^k U_j(Tz^k-z^k)-z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}\right]\\
&=\|z^k-z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}+\mathbb{E}_k\left[\|\sum_{j=1}^m\epsilon_j^k U_j(Tz^k-z^k)\|^2_{\Pi^{-1}S\Lambda}\right.\\
&\quad\left.+2\langle z^k-z^\star,\sum_{j=1}^m\epsilon_j^k U_j(Tz^k-z^k)\rangle_{\Pi^{-1}S}\right].
\end{aligned}\tag{52}$$

Let $\Psi = \{0,1\}^m$. Then we have

$$\begin{aligned}
&\mathbb{E}_k\left[\|\sum_{j=1}^m\epsilon_j^k U_j(Tz^k-z^k)\|^2_{\Pi^{-1}S\Lambda}\right]\\
&=\sum_{\varepsilon\in\Psi}\mathbb{P}(\epsilon^k=\varepsilon)\|\sum_{j=1}^m\varepsilon_j U_j(Tz^k-z^k)\|^2_{\Pi^{-1}S\Lambda}\\
&=\sum_{j=1}^m\sum_{\varepsilon\in\Psi}\mathbb{P}(\epsilon^k=\varepsilon)\|\varepsilon_j U_j(Tz^k-z^k)\|^2_{\Pi^{-1}S\Lambda}\\
&=\sum_{j=1}^m\left(\sum_{\varepsilon\in\Psi,\varepsilon_j=1}\mathbb{P}(\epsilon^k=\varepsilon)\|U_j(Tz^k-z^k)\|^2_{\Pi^{-1}S\Lambda}\right)\\
&=\sum_{j=1}^m p_j\|U_j(Tz^k-z^k)\|^2_{\Pi^{-1}S\Lambda}\\
&=\|Tz^k-z^k\|^2_{S\Lambda},
\end{aligned}\tag{53}$$

where we used the fact $\Pi^{-1}S\Lambda$ is diagonal and U_js don't overlap in the second equality. Following the same argument for the inner product yields:

$$\begin{aligned}
&\mathbb{E}_k\left[\langle z^k-z^\star,\sum_{j=1}^m\epsilon_j^k U_j(Tz^k-z^k)\rangle_{\Pi^{-1}S}\right]\\
&=\sum_{\varepsilon\in\Psi}\mathbb{P}(\epsilon^k=\varepsilon)\langle z^k-z^\star,\sum_{j=1}^m\varepsilon_j U_j(Tz^k-z^k)\rangle_{\Pi^{-1}S}\\
&=\sum_{j=1}^m\sum_{\varepsilon\in\Psi}\mathbb{P}(\epsilon^k=\varepsilon)\langle z^k-z^\star,\varepsilon_j U_j(Tz^k-z^k)\rangle_{\Pi^{-1}S}\\
&=\sum_{j=1}^m\sum_{\varepsilon\in\Psi,\varepsilon_j=1}\mathbb{P}(\epsilon^k=\varepsilon)\langle z^k-z^\star,U_j(Tz^k-z^k)\rangle_{\Pi^{-1}S}\\
&=\sum_{j=1}^m p_j\langle z^k-z^\star,U_j(Tz^k-z^k)\rangle_{\Pi^{-1}S}\\
&=\langle z^k-z^\star,Tz^k-z^k\rangle_S.
\end{aligned}\tag{54}$$

Combining (52), (53) and (54) yield:

$$\begin{aligned}
\mathbb{E}_k\left[\|z^{k+1}-z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}\right]\leq&\|z^k-z^\star\|^2_{\Lambda^{-1}\Pi^{-1}S}\\
&+\|Tz^k-z^k\|^2_{S\Lambda}\\
&+2\langle z^k-z^\star,Tz^k-z^k\rangle_S.
\end{aligned}$$

Lemma 1 completes the proof. □

Proof of Theorem 5. For notational convenience let $\bar{S} = \Lambda^{-1}\Pi^{-1}S$ and $\tilde{T} = D+M+F$, noting that $\mathcal{S} = \operatorname{zer}\tilde{T}$ (cf.(10)). 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 (35) that

$$\begin{aligned}
\mathbb{E}_k\left[d^2_{\bar{S}}(z^{k+1},\mathcal{S})\right]&\leq\mathbb{E}_k\left[\|z^{k+1}-\mathcal{P}_{\mathcal{S}}^{\bar{S}}(z^k)\|^2_{\bar{S}}\right]\\
&\leq\|z^k-\mathcal{P}_{\mathcal{S}}^{\bar{S}}(z^k)\|^2_{\bar{S}}-\|Tz^k-z^k\|^2_{2\tilde{P}-S\Lambda}\\
&=d^2_{\bar{S}}(z^k,\mathcal{S})-\|Tz^k-z^k\|^2_{2\tilde{P}-S\Lambda}.
\end{aligned}\tag{55}$$

By definition (8) we have

$$\begin{aligned}
&\|\bar{z}^k-z^k\|^2=\|(H+M^\top)^{-1}S(Tz^k-z^k)\|^2\\
&\leq\|(H+M^\top)^{-1}S\|^2\|(2\tilde{P}-S\Lambda)^{-1}\|\|Tz^k-z^k\|^2_{2\tilde{P}-S\Lambda},
\end{aligned}\tag{56}$$

where $\bar{z}^k$ is defined by (9) 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

$$\begin{aligned} d^2_{\bar{S}}(z^k, \mathcal{S}) &\leq \|z^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\|^2_{\bar{S}} \leq \|\bar{S}\|\|z^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\|^2 \\ &\leq \|\bar{S}\| \left(\|\bar{z}^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\| + \|\bar{z}^k - z^k\|\right)^2 \\ &= \|\bar{S}\| \left(d(\bar{z}^k, \mathcal{S}) + \|\bar{z}^k - z^k\|\right)^2. \end{aligned} \tag{57}$$

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

$$v^k := -(H - M)(\bar{z}^k - z^k) + F\bar{z}^k - Fz^k.$$

It follows from (9) that $(H - M - F)z^k \in (H + D)\bar{z}^k$, which in turn implies

$$v^k \in \tilde{T}\bar{z}^k = (D + M + F)\bar{z}^k. \tag{58}$$

Let Ω be defined as in the proof of Theorem 4. Let $\omega \in \Omega$. Combine (8) and (36) together with the fact that $S^{-1}(H + M^\top)$ has full rank to derive

$$\bar{z}^k(\omega) - z^k(\omega) \to 0. \tag{59}$$

On the other hand, metric subregularity of $\tilde{T}$ at all $z^\star \in \mathcal{S} = \operatorname{zer}\tilde{T}$ for 0 implies that there exists a neighborhood $\mathcal{U}$ of $z^\star$ such that (see the equivalent formulation of metric subregularity in [24, Exercise 3H.4]):

$$d(x, \mathcal{S}) \leq \eta d(0, y), \quad \forall x \in \mathcal{U}, \text{ and } y \in \tilde{T}x, \tag{60}$$

for some $\eta \in [0, \infty)$. From (59) and Theorem 4 we have that $\bar{z}^k(\omega) \to z^\star \in \mathcal{S} = \operatorname{zer}\tilde{T}$ which implies that there exists $\bar{k} \in \mathbb{N}$ such that for $k > \bar{k}$ a neighborhood $\mathcal{U}$ of $z^\star$ exists with $\bar{z}^k(\omega) \in \mathcal{U}$. Consequently (58) and (60) yield

$$d(\bar{z}^k(\omega), \mathcal{S}) \leq \eta\|v^k(\omega)\|. \tag{61}$$

From triangle inequality and Lipschitz continuity of F we derive

$$\begin{aligned} \|v^k\| &= \|(H - M)(\bar{z}^k - z^k) - F\bar{z}^k + Fz^k\| \\ &\leq \|(H - M)(\bar{z}^k - z^k)\| + \|F\bar{z}^k - Fz^k\| \leq \xi\|\bar{z}^k - z^k\|, \end{aligned}$$

where ξ is a constant depending on $\|H - M\|$ and the Lipschitz constants β_i for $i = 1, \ldots, m$. Therefore, using (61) we derive

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

Using this in (57) together with (56) yields

$$d^2_{\bar{S}}(z^k(\omega), \mathcal{S}) \leq \phi\|Tz^k(\omega) - z^k(\omega)\|^2_{2\tilde{P} - S\Lambda}, \tag{62}$$

where $\phi = (\xi\eta + 1)^2\|(H + M^\top)^{-1}S\|^2\|(2\tilde{P} - S\Lambda)^{-1}\|\|\bar{S}\|$. Combine (62) with (55) to derive

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

This concludes the proof. □

References

[1] P. L. Combettes and I.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 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] P. Latafat and P. Patrinos, "Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators," *Computational Optimization and Applications*, pp. 1–37, 2017.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

[20] 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.

[21] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 2012.

[22] 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.

[23] D. Drusvyatskiy and A. S. Lewis, "Error bounds, quadratic growth, and linear convergence of proximal methods," *arXiv preprint arXiv:1602.06661*, 2016.

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

[25] 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.

[26] 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.

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

[28] R. Srikant and L. Ying, *Communication networks: an optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.

[29] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

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