

# Distributed Primal-dual Optimization for Heterogeneous Multi-agent Systems

Yichuan Li<sup>1</sup>, Petros Voulgaris<sup>2</sup>, and Nikolaos M. Freris<sup>3</sup>

**Abstract**—Heterogeneous networks comprise agents with varying capabilities in terms of computation, storage, and communication. In such settings, it is crucial to factor in the operating characteristics in allowing agents to choose appropriate updating schemes, so as to better distribute computational tasks and utilize the network more efficiently. We consider the multi-agent optimization problem of cooperatively minimizing the sum of local strongly convex objectives. We propose an asynchronous distributed primal-dual protocol, which allows for the primal update steps to be agent-dependent (an agent can opt between first-order or Newton updates). Our analysis introduces a unifying framework for such hybrid optimization scheme and establishes global linear convergence in expectation, under strongly convex objectives and general agent activation schemes. Numerical experiments on real life datasets attest to the merits of the proposed algorithm.

## I. INTRODUCTION

By harnessing the computational power of multiple machines, distributed optimization aims to integrate a network of agents to solve a minimization problem. The framework has found success across a wide range of fields, including control, smart-grid networks, and machine learning [1]–[4]. The growing number of mobile devices with on-device computation and communication capabilities further instigates the possibility of ever-larger scale networks with decentralized computing, such as Cyber-Physical Systems (CPS) [5]. In this setting, heterogeneity is ubiquitous and relates to several aspects of the system such as: hardware conditions, network connectivity, data volumes, as well as communication bandwidths. These challenges require revisiting algorithmic designs to account for the operating characteristics to effectively distribute computational tasks.

We consider the following collaborative network problem:

$$\underset{\tilde{x} \in \mathbb{R}^d}{\text{minimize}} \left\{ \sum_{i=1}^m f_i(\tilde{x}) + g(\tilde{x}) \right\}, \quad (1)$$

where  $f_i(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  are local strongly convex objectives, and  $g(\cdot)$  is a convex but possibly nonsmooth regularization function. Each agent has access only to its individual objective function, and performs updates using local data. A network-wide solution is achieved asymptotically via an exchange of information among neighboring agents, which serves to effectively integrate global information.

<sup>1</sup>Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL 61820, USA. yli129@illinois.edu.

<sup>2</sup>Department of Mechanical Engineering, University of Nevada, Reno, NV 89557, USA. pvoulgaris@unr.edu.

<sup>3</sup>School of Computer Science, University of Science and Technology of China, Hefei, Anhui, 230027, China. nfr@ustc.edu.cn.

\*This work was supported by the Ministry of Science and Technology of China under grant 2019YFB2102200. Correspondence to N. Freris.

Distributed first order methods [6]–[9] constitute a popular class of algorithms for solving (1) due to their simplicity and low computational requirements. Agents perform updates using (sub) gradients pertaining to local objectives along with averaged values from neighbors. The method in [6] can be viewed as solving a penalized version of (1) and therefore converges to a biased solution. To converge to the exact solution, diminishing step sizes have to be used and consequently converges slowly. This is remedied by the method in [8], where agents utilize iterates and gradient histories. By using tracking techniques, [8] establish  $\mathcal{O}(1/t)$  convergence rate under general convex objectives and linear convergence rate under strongly convex objectives. Despite its economical costs, first order methods are well-known to suffer from slow convergence speed, especially for ill-conditioned large scale problems.

Second order methods attain acceleration by invoking information pertaining to the (exact or approximate) Hessian matrix of the objective. Recent work on (quasi) Newton methods [10]–[13] incorporate objective curvature information to compute more informed updates. Similar to its first order counterpart, Network Newton [10] and its asynchronous extension [11] solve a penalized version of (1), where Newton steps are approximated via a truncated Taylor expansion. Similar techniques were employed in [12] for the primal-dual formulation of the problem. The ability to reach a high-accuracy solution in significantly fewer iterations makes second order algorithms a desirable candidate when abundant computational resources are available. However, each update step of Newton’s method requires to solve a linear system at the cost of  $\mathcal{O}(d^3)$  computational costs for general objectives.

Methods that allow agents to choose between first and second order solver are rather scarce in the literature. This consideration becomes significant for large scale heterogeneous networks where machines have varying capacities in terms of computation, communication, storage, and battery levels. In this aspect, a *hybrid* scheme serves to effectively avoid placing heavy computational burden on less capable machines thus better utilizing the network-wide resources. An emerging application where such challenges become more pronounced is Federated Learning [14], where massive number of agents collaborate and aim to obtain a model with improved performances by utilizing data distributed across the network. We note a related work [15] where authors proposed a hybrid protocol for the client-server setting where clients may choose to perform either gradient or Newton updates depending on their local environments. Nevertheless, [15] is limited to the client-server setting, and further

requires synchronous computation and communication: this is unrealistic in large-scale networks. In contradistinction, the proposed algorithm can be deployed in general multi-agent settings, (of which client-server is a special case) and achieves a totally asynchronous implementation.

### Contributions:

- We propose **Hybrid Primal-dual Proximal Optimization (HIPPO)**, a framework that encases heterogeneous multi-agent systems by allowing agents to select between gradient and or Newton updates.
- HIPPO admits a *totally asynchronous* implementation that allows a general participation scheme of agents at its iteration. This removes the need for a central clock and network-wide synchronization.
- We establish linear convergence rate in expectation under strongly convex local objectives with Lipschitz continuous gradients, for arbitrarily chosen hybrid updating schemes. Numerical experiments on real-life datasets attest the efficacy of the proposed algorithm.

**Notation:** Vectors  $x \in \mathbb{R}^d$  are denoted using lower case letters and matrices  $A \in \mathbb{R}^{n \times m}$  are denoted using upper case letters, while transpose is denoted as  $A^\top$ . Row stacking and column stacking is expressed as  $[A; B]$  and  $[A, B]$ , respectively (provided dimensions match). We use superscript to denote iterates of the algorithm, while subscripts are reserved for vector components. For example,  $x_i^t$  denotes the vector held by the  $i$ -th agent at time step  $t$ . When a norm specification is not provided,  $\|x\|$  and  $\|A\|$  denote the Euclidean norm of the vector and the induced norm of the matrix, respectively. For a positive definite matrix  $P \succ 0$ , we represent the associated quadratic form as  $\|x\|_P := \sqrt{x^\top P x}$ . For matrices of arbitrary dimensions, we denote their Kronecker product as  $A \otimes B$ . Last, we use the shorthand notation  $[m] := \{1, \dots, m\}$ .

## II. PRELIMINARIES

### A. Proximal operator

The proximal operator [16]  $\text{prox}_{g/\mu}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$  associated with a function  $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  and a constant  $\mu > 0$  is defined as:

$$\text{prox}_{g/\mu}(x) = \underset{v \in \mathbb{R}^d}{\text{argmin}} \left\{ g(v) + \frac{\mu}{2} \|v - x\|^2 \right\}. \quad (2)$$

When  $g(\cdot)$  is convex, the above mapping is single valued. Proximal algorithms [17] were developed for convex composite problems, where the objective is the sum of two convex functions: one smooth and one possibly non-smooth but proximal (i.e., with efficiently computable proximal mapping). Examples where (2) is easy to evaluate are ubiquitous in applications, e.g., the  $\ell_1$ -norm or the indicator functions for hyperplane, halfspace, simplex, orthant, etc. The vast majority of these methods are first order, i.e., they use the gradient pertaining to the smooth term along with the proximal operator of the non-smooth part to iteratively solve the problem. Using second order information, in this framework, typically poses challenges in ensuring convergence guarantees while maintaining affordable computational

costs. This is due to the fact that the norm function in (2) has to be scaled according to the Hessian matrix [18]. Such scaling may render the evaluation of the associated proximal mappings expensive, even when its first-order counterpart admits a closed form solution.

### B. Problem formulation

The network topology is captured by an undirected connected graph:  $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V} := [m]$  is the index set and  $\mathcal{E} \subseteq (\mathcal{V}, \mathcal{V})$  is the edge set. Edge  $(i, j) \in \mathcal{E}$  if and only if a communication channel exists between the  $i$ -th agent and  $j$ -th agent. We do not consider self-loops, i.e.,  $(i, i) \notin \mathcal{E}$  for all  $i \in [m]$ . The number of edges is represented as  $|\mathcal{E}| = n$ , and the set of neighbors of the  $i$ -th agent is defined by  $\mathcal{N}_i := \{j \in [m], (i, j) \in \mathcal{E}\}$ . Using the above definitions, problem (1) can be recast as a consensus optimization problem by introducing local decision vectors  $x_i$  to agents, and consensus vectors  $z_{ij}$  to edges of the graph. In specific:

$$\begin{aligned} & \underset{x_i, \theta, z_{ij} \in \mathbb{R}^d}{\text{minimize}} \left\{ \sum_{i=1}^m f_i(x_i) + g(\theta) \right\} \\ & \text{subject to } x_i = z_{ij} = x_j, \text{ for } i \in [m], j \in \mathcal{N}_i, \\ & \quad x_l = \theta \text{ for some } l \in [m]. \end{aligned} \quad (3)$$

Note that we have also separated the argument of the smooth part and the nonsmooth part of the objective by letting  $x_l = \theta$  for an arbitrary agent. The introduction of intermediate variables  $\{z_{ij}\}$  deviates from the classical setting of consensus optimization and is crucial in obtaining block-diagonal Hessian structures that are needed for our hybrid scheme (See Section III-A). The above formulation can be compactly expressed by defining the source and the destination matrices  $\widehat{A}_s, \widehat{A}_d$  as follows. We first assign an arbitrary order to the edge set and denote  $\mathcal{E}_k = (i, j), k \in [n]$ , if the  $k$ -th edge connects the  $i$ -th and the  $j$ -th agent. Matrices  $\widehat{A}_s, \widehat{A}_d \in \mathbb{R}^{n \times m}$  are constructed by letting  $[\widehat{A}_s]_{ki} = [\widehat{A}_d]_{kj} = 1$  (with all other entries in the  $k$ -th row being zero). The network topology can be captured by the signed and unsigned graph incidence matrices defined respectively as:  $\widehat{E}_s = \widehat{A}_s - \widehat{A}_d$  and  $\widehat{E}_u = \widehat{A}_s + \widehat{A}_d$ . An immediate consequence is that problem (3) is equivalent to:

$$\begin{aligned} & \underset{x \in \mathbb{R}^{md}, z \in \mathbb{R}^{nd}, \theta \in \mathbb{R}^d}{\text{minimize}} \left\{ F(x) + g(\theta) \right\} \\ & \text{subject to } \begin{bmatrix} A_s \\ A_d \end{bmatrix} x = Ax = Bz = \begin{bmatrix} I_{nd} \\ I_{nd} \end{bmatrix} z, \\ & \quad S^\top x = \theta. \end{aligned} \quad (4)$$

where we stack  $x_i$  into a single column vector  $x \in \mathbb{R}^{md} := [x_1^\top, \dots, x_m^\top]^\top$  and similarly for  $z \in \mathbb{R}^{nd}$ . The matrix  $S := (s_l \otimes I_d) \in \mathbb{R}^{md \times d}$ , where  $s_l \in \mathbb{R}^d$  is a zero vector except for the  $l$ -th entry being one, serves to select the  $l$ -th agent. We denote  $A_s := \widehat{A}_s \otimes I_d, A_d := \widehat{A}_d \otimes I_d$  as the Kronecker product between  $\widehat{A}_s, \widehat{A}_d$  and the identity matrix of dimension  $d$ , respectively. Matrices  $A, B$  in (4) are formed by stacking the corresponding matrices.

### C. Primal-dual algorithms

Primal-dual optimization algorithms date back to the work [19]. By introducing the (augmented) Lagrangian, primal-dual algorithms seek to solve the primal and the dual problem simultaneously, through the use of dual variables. The proliferation of primal-dual methods [20] is attributed to their wide applicability to distributed optimization. In the following of this subsection, we review one instance of primal-dual algorithms, termed Alternating Direction Method of Multipliers (ADMM) [21], which serves as the foundation of our subsequent development. To solve (4) using ADMM, we define the associated augmented Lagrangian as follows:

$$\mathcal{L}(x, z, \theta; y, \lambda) = F(x) + g(\theta) + y^\top (Ax - Bz) + \lambda^\top (S^\top x - \theta) + \frac{\mu_z}{2} \|Ax - Bz\|^2 + \frac{\mu_\theta}{2} \|S^\top x - \theta\|^2, \quad (5)$$

where  $y \in \mathbb{R}^{2nd}$ ,  $\lambda \in \mathbb{R}^d$  are the dual vectors and  $\mu_z, \mu_\theta > 0$  are scalars that controls the quadratic penalty. ADMM solves (4) by sequentially minimizing the augmented Lagrangian with respect to the primal variables  $(x, z, \theta)$ , and then performs gradient ascent on the dual variable.

$$x^{t+1} = \underset{x}{\operatorname{argmin}} \mathcal{L}(x, z^t, \theta^t; y^t, \lambda^t), \quad (6a)$$

$$z^{t+1} = \underset{z}{\operatorname{argmin}} \mathcal{L}(x^{t+1}, z, \theta^t; y^t, \lambda^t), \quad (6b)$$

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(x^{t+1}, z^{t+1}, \theta; y^t, \lambda^t) \quad (6c)$$

$$y^{t+1} = y^t + \mu_z (Ax^{t+1} - Bz^{t+1}). \quad (6d)$$

$$\lambda^{t+1} = \lambda^t + \mu_\theta (S^\top x^{t+1} - \theta^{t+1}) \quad (6e)$$

As a result of introducing the consensus variables  $\{z_{ij}\}$ , we obtain an instance of 3-block ADMM in (6), which impose more strict assumptions than 2-block ADMM for convergence[22]. The introduction of the consensus variables  $\{z_{ij}\}$  decouples  $\{x_i\}$  and renders the step (6a) distributedly computable [23], i.e., agents can perform the step (6a) with only local information and information from their immediate neighbors. However, multiple inner loops are required to solve (6a) with general objective functions. To alleviate the computational burden, several approximation schemes [24]-[26] exist wherein instead of aiming for an exact solution, agents perform one or more updating steps using gradient or (quasi) Newton information. Building on these results, we propose a general framework that accommodates system heterogeneity by allowing agents to select their updates and further compute updates in an asynchronous fashion.

### III. ALGORITHMIC DEVELOPMENT

Abbreviating  $\mathcal{L}^t \equiv \mathcal{L}(x^t, z^t, \theta^t; y^t, \lambda^t)$ , we define the approximated augmented Lagrangian as follows,

$$\hat{\mathcal{L}}(x, z^t, \theta^t; y^t, \lambda^t) = \mathcal{L}^t + (x - x^t)^\top \nabla_x \mathcal{L}^t + \frac{1}{2} \|x - x^t\|_{H^t}^2, \quad (7)$$

where  $H^t$  is a matrix (to be defined subsequently) that encapsulates our hybrid updating schemes. By replacing the augmented Lagrangian (5) with its approximation in (7), we obtain the following closed form solution for step (6a):

$$x^{t+1} = x^t - (H^t)^{-1} \nabla_x \mathcal{L}(x^t, z^t; y^t). \quad (8)$$

Moreover, step (6b) corresponds to:

$$B^\top y^t + \mu B^\top (Ax^{t+1} - Bz^{t+1}) = 0. \quad (9)$$

By completion of squares, the update of  $\theta^{t+1}$  can be cast as the following proximal mapping:

$$\theta^{t+1} = \mathbf{prox}_{g/\mu_\theta} (S^\top x^{t+1} + \frac{1}{\mu_\theta} \lambda^t). \quad (10)$$

Dual variables  $(y, \lambda)$  are updated in verbatim as (6d)-(6e). The updates (8)-(10), and (6d)-(6e) serve as the basis for developing our proposed algorithm. An equivalent while more efficient implementation can be achieved, which we formalize in the following lemma.

*Lemma 1:* Consider the iterates generated by (8)-(10), and (6d)-(6e). Let the dual variable  $y^t \in \mathbb{R}^{2nd}$  be expressed as  $y^t = [\alpha^t; \beta^t]$ ,  $\alpha^t, \beta^t \in \mathbb{R}^{nd}$ . If  $y^0$  and  $z^0$  are initialized so that  $\alpha^0 = -\beta^0$  and  $z^0 = \frac{1}{2} E_u x^0$ , then  $\alpha^t = -\beta^t$  and  $z^t = \frac{1}{2} E_u x^t$  for all  $t \geq 0$ . Moreover, by defining  $\phi^t = E_s^\top \alpha^t \in \mathbb{R}^{md}$ , we obtain the following equivalent updates:

$$x^{t+1} = x^t - (H^t)^{-1} \left( \nabla F(x^t) + \phi^t + \frac{\mu_z}{2} E_s^\top E_s x^t + S \left[ \lambda^t + \mu_\theta (S^\top x^t - \theta^t) \right] \right), \quad (11a)$$

$$\theta^{t+1} = \mathbf{prox}_{g/\mu} (S^\top x^{t+1} + \frac{1}{\mu_\theta} \lambda^t), \quad (11b)$$

$$\phi^{t+1} = \phi^t + \frac{\mu_z}{2} E_s^\top E_s x^{t+1}, \quad (11c)$$

$$\lambda^{t+1} = \lambda^t + \mu_\theta (S^\top x^{t+1} - \theta^{t+1}). \quad (11d)$$

*Proof:* See Appendix.

*Remark 1:* The above result is similarly derived in [23] when the subproblem is solved exactly. Note updates (11) are equivalent to updates (8)-(10), and (6d)-(6e) under appropriate initialization, irrespective of choices of  $H^t$ . However, (11) falls into the category of 2-block ADMM through elimination of the  $z$ -update ( $z^t = \frac{1}{2} E_u x^t$ ). This gives more freedom for choosing hyperparameters  $\mu_z, \mu_\theta$ . On the other hand, the use of the consensus variables  $\{z_{ij}\}$  in the formulation (4) is necessary, as we demonstrate next.

#### A. Hybrid updating

We construct  $H^t$  as follows:

$$H^t = J^t + \mu_z D + \mu_\theta S S^\top + \Delta, \quad (12)$$

where  $J^t \in \mathbb{R}^{md \times md}$  is a block diagonal matrix, whose sub-blocks dictate the updating scheme of the corresponding agent,  $D$  is a diagonal matrix with entries equal to the degree of the corresponding agent, and  $\Delta$  is a constant matrix, whose sub-blocks serve to specify step sizes and provide numerical robustness. In specific, if the  $i$ -th agent opts to use Newton updates, then  $J_{ii}^t = \nabla^2 f_i(x_i^t) \in \mathbb{R}^{d \times d}$ ; if the  $i$ -th agent opts to use first-order information, then  $J_{ii}^t = 0$  and the step size of the gradient descent is determined by specifying  $\Delta_{ii}$ . Note that when  $J^t = \nabla^2 F(x^t)$ , i.e., all agents perform Newton updates, we obtain the Hessian of the augmented Lagrangian (5) plus the diagonal step size matrix  $\Delta$ . Irrespective of the choice of  $J_{ii}^t$ ,  $H^t$  in (12) is always block diagonal. This important property translates to the fact that computing the update step  $u^t = (H^t)^{-1} \nabla_x \mathcal{L}^t$  can be carried out *locally* by agents, without additional communication rounds once  $\nabla_x \mathcal{L}^t$  is available. In specific,

---

**Algorithm 1** HIPPO

Zero initialization for all variables. Hyperparameters  $\mu_z, \mu_\theta$  and  $\Delta_{ii}, i \in [m]$ .

```

1: for  $t = 0, 1, 2, \dots$  do
2:   for active agents do
3:     Retrieve  $x_j^t, j \in \mathcal{N}_i$ , from the local buffer.
4:     if agent  $i \in \mathcal{U}_{\text{Gradient}}^t$  then
5:        $J_{ii}^t = 0$ 
6:     else if  $i \in \mathcal{U}_{\text{Newton}}^t$  then
7:        $J_{ii}^t = \nabla^2 f_i(x_i^t)$ 
8:     end if
9:     Compute  $u_i^t$  as in (13).
10:     $x_i^{t+1} = x_i^t - u_i^t$ .
11:    Broadcast  $x_i^{t+1}$ .
12:     $\phi_i^{t+1} = \phi_i^t + \frac{\mu}{2} \sum_{j \in \mathcal{N}_i} (x_i^{t+1} - x_j^{t+1})$ .
13:    if  $i = l$  then
14:       $\theta^{t+1} = \text{prox}_{g/\mu_\theta}(x_l^{t+1} + \frac{1}{\mu_\theta} \lambda^t)$ 
15:       $\lambda^{t+1} = \lambda^t + \mu_\theta (x_l^{t+1} - \theta^{t+1})$ 
16:    end if
17:  end for
18: end for

```

---

the  $i$ -th agent computes the update direction  $u_i^t$  by solving the following linear system:

$$(J_{ii}^t + \mu_z |\mathcal{N}_i| + \delta_{il} \mu_\theta + \Delta_{ii}) u_i^t = \nabla f_i(x_i^t) + \phi_i^t + \frac{\mu_z}{2} \sum_{j \in \mathcal{N}_i} (x_i^t - x_j^t) + \delta_{il} (\lambda^t + \mu_\theta (x_l^t - \theta^t)), \quad (13)$$

where  $\delta_{il} = 1$  if and only if  $i = l$ . The block-diagonal structure of  $H^t$  in (12) is a consequence of the introduction of  $\{z_{ij}\}$ . If instead a direct consensus constraint is imposed as  $x_i = x_j$  in (3), it is not hard to verify that  $H^t$  in (12) would be of the following form:

$$J^t + \mu_z L_s + \mu_\theta S S^\top + \Delta,$$

where  $L_s = E_s^\top E_s$  is the signed graph Laplacian matrix, whose  $ij$ -th block is nonzero if  $(i, j) \in \mathcal{E}$ . The presence of  $L_s$  couples agents with their neighbors. In such cases, solving for the update direction  $u_i^t$  requires either collecting all information from neighbors beforehand, or invoking a truncated Taylor expansion of the Hessian inverse as in [12]. Such schemes not only induce additional communication rounds within agents, but are actually infeasible in the hybrid scheme: if neighboring agents opt to use different updates, the required information to compute approximated Newton steps is not available. The proposed algorithm is described in **Algorithm 1**.

We equip each agent with a local buffer so that agents can receive and store information from neighbors even if the agent is not active. We assume each agent is active according to local Poisson clocks [11]. For a round that agent  $i$  participates, i.e., performs an update, it retrieves the most recent copy of  $x_j^t, j \in \mathcal{N}_i$ , from its local buffer as in line 3 of **Algorithm 1**. Active agents are divided into two groups:  $\mathcal{U}_{\text{Gradient}}^t$  and  $\mathcal{U}_{\text{Newton}}^t$ , where  $\mathcal{U}_{\text{Gradient}}^t \cup \mathcal{U}_{\text{Newton}}^t \subseteq \mathcal{V}$ . Depending on whether an agent opts to use gradient or Newton update, the update direction is computed using  $J_{ii}^t = \nabla^2 f(x_i^t)$  or  $J_{ii}^t = 0$  in (13) accordingly (lines 3-8).

Note that each agent can choose different computing schemes at each time step. Once the update direction  $u_i^t$  is obtained, the primal update is carried out, followed by broadcasting  $x_i^{t+1}$  to neighbors in lines 10-11. For Poisson processes, it is natural to assume only one clock ticks at a time and therefore  $x_j^{t+1} = x_j^t$ . For other cases, agents can check their local buffers before proceeding to the dual update in the line 12. The  $l$ -th agent additionally update the  $(\theta, \lambda)$  pair pertaining to the regularization function.

### B. Operator description of the algorithm

To succinctly describe the proposed asynchronous algorithm for the sake of analysis, we introduce the following vector concatenation  $v := [x; \phi; \theta; \lambda] \in \mathbb{R}^{(2m+2)d}$ . We define an operator  $T : \mathbb{R}^{(2m+2)d} \rightarrow \mathbb{R}^{(2m+2)d}$  as follows:

$$v^{t+1} = T v^t, \quad (14)$$

which maps  $[x^t; \phi^t; \theta^t; \lambda^t] \in \mathbb{R}^{(2m+2)d}$  to  $[x^{t+1}; \phi^{t+1}; \theta^{t+1}; \lambda^{t+1}]$  according to (11). We further define the following activation matrix:

$$\Omega^t = \begin{bmatrix} X^t & 0 & 0 & 0 \\ 0 & X^t & 0 & 0 \\ 0 & 0 & X_{ll}^t & 0 \\ 0 & 0 & 0 & X_{ll}^t \end{bmatrix}, \quad (15)$$

where  $X^t \in \mathbb{R}^{md \times md}$  is a random diagonal matrix with sub-blocks  $X_{ii}^t \in \mathbb{R}^{d \times d}, i \in [m]$ , taking values  $I_d$  or the zero matrix. Using (14)–(15), we can express the asynchronous updates as:

$$v^{t+1} = v^t + \Omega^{t+1} (T v^t - v^t). \quad (16)$$

This update rule simply prescribes that the  $i$ -th agent participates at the  $t$ -th iteration if and only if  $X_{ii}^{t+1} = I_d$ . Moreover, the primal-dual pair  $(x_i^{t+1}, \phi_i^{t+1})$  is updated simultaneously when the corresponding agent is active, by construction. Therefore, by varying  $\mathbb{E}^t [X_{ii}^{t+1}] = \mathbb{P}(X_{ii}^{t+1} = I_d | \mathcal{F}^t)$ , where  $\mathcal{F}^t$  is the filtration generated by  $X^1, X^2, \dots, X^t$  and  $\mathbb{E}^t[\cdot] := \mathbb{E}[\cdot | \mathcal{F}^t]$ , one obtain a wide range of activation schemes.

- *Synchronous updates*:  $X_{ii}^t = I_d$  for  $i \in [m]$  and  $t \geq 1$ . In this case, updates (16) boils down to (14).
- *Stochastic coordinate updates*: By letting the random matrix  $X^t$  take values in:  $\{\text{Blkdiag}(I_d, 0, \dots, 0), \text{Blkdiag}(0, I_d, 0, \dots, 0), \dots, \text{Blkdiag}(0, 0, \dots, I_d)\}$ , we obtain a randomized algorithm where, at each iteration, a single agent performs gradient or Newton update, depending on  $i \in \mathcal{U}_{\text{Gradient}}^t$  or  $i \in \mathcal{U}_{\text{Newton}}^t$ .
- *Multi-agent updates*: By setting  $\mathbb{E}^t [X_{ii}^{t+1}] = p_i^t$ , we obtain a general activation scheme where the  $i$ -th agent performs updates at  $t$ -th iteration with probability  $p_i^t$ .

## IV. ANALYSIS

In this section, we present a unified analysis that establishes linear convergence rate for the proposed algorithm under strongly convex local objectives, with arbitrary updating schemes chosen by agents. We assume that the initialization requirement in Lemma 1 is satisfied (zero initialization for all variables suffices). Recall the definition  $\phi = E_s^\top \alpha$  in

Lemma 1 and the concatenation  $v = [x; \phi; \theta; \lambda] \in \mathbb{R}^{(2m+2)d}$  in Section III-B. Using the tuple  $(x, \phi, \theta, \lambda)$  facilitates the implementation of the algorithm while  $(x, z, \alpha, \theta, \lambda)$  is most suitable for our analysis. We note that their equivalence is established in Lemma 1. In the following, we denote  $v_\alpha = [x; z; \alpha; \theta; \lambda] \in \mathbb{R}^{(m+2n+2)d}$ . The following assumptions are made throughout.

*Assumption 1:* The local cost functions  $f_i(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  are twice continuously differentiable,  $m_f$ -strongly convex, and have  $M_f$ -Lipschitz continuous gradients, i.e.,  $\forall i \in [m], x \in \mathbb{R}^d$ , the following holds:

$$m_f I \preceq \nabla^2 f_i(x) \preceq M_f I. \quad (17)$$

*Assumption 2:* The function  $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  is proper, convex, and closed, i.e.,  $\forall x, y \in \mathbb{R}^d$ , the following holds:

$$(\partial g(x) - \partial g(y))^\top (x - y) \geq 0, \quad (18)$$

where  $\partial g(\cdot)$  denotes the subdifferential set.

*Assumption 3:* The Hessian of the local cost function  $f_i(\cdot)$  is Lipschitz continuous with constant  $L_f$ , i.e.,  $\forall i \in [m], x, y \in \mathbb{R}^d$ , the following holds:

$$\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq L_f \|x - y\|. \quad (19)$$

Since  $F(x) = \sum_{i=1}^m f_i(x_i)$ , the bound in (17) holds for  $\nabla^2 F(x)$  as well, i.e.,  $m_f I \preceq \nabla^2 F(x) \preceq M_f I$ . The Lipschitz continuity assumption is standard in analyzing both first and second order methods while Assumption 3 is only needed for analyzing second order methods [27]. We proceed by stating a lemma that will be handy in quantifying the error introduced when we replace the exact optimization step (6a) with the one-step update (11a).

*Lemma 2:* Consider the iterates generated by (11) and equivalently (14). Recall the definition  $\phi = E_s^\top \alpha^t$  and the identity  $z^t = \frac{1}{2} E_u x^t$ . The following holds:

$$e^t + \nabla F(x^{t+1}) - \nabla F(x^*) + \Delta(x^{t+1} - x^t) + E_s^\top (\alpha^{t+1} - \alpha^*) + \mu_z E_u^\top (z^{t+1} - z^t) + S(\lambda^{t+1} - \lambda^* + \mu_\theta (\theta^{t+1} - \theta^t)) = 0,$$

where the error term  $e^t$  is defined as:

$$e^t := \nabla F(x^t) - \nabla F(x^{t+1}) + J^t(x^{t+1} - x^t). \quad (20)$$

*Proof:* See Appendix.

The following lemma establishes an upper bound for  $\|e^t\|$  that is crucial to establish the convergence of the algorithm.

*Lemma 3:* The error  $e^t$  in (20) is upper bounded as:  $\|e^t\|^2 \leq \|x^{t+1} - x^t\|_{(\Pi^t)^2}^2$ , where  $\Pi^t \in \mathbb{R}^{md \times md}$  is a diagonal matrix with entries equal to:

$$\begin{aligned} \Pi_{ii}^t &= M_f \cdot I, \text{ for } i \in \mathcal{U}_{\text{Gradient}}, \\ \Pi_{ii}^t &= \min \left\{ 2M_f, \frac{L_f}{2} \|x_i^{t+1} - x_i^t\| \right\} \cdot I, \text{ for } i \in \mathcal{U}_{\text{Newton}}. \end{aligned} \quad (21)$$

Lemma 3 reveals the difference between using Newton and gradient updates, i.e., the progression of the error in terms of  $\|x^{t+1} - x^t\|$ . Before we establish the convergence rate of the algorithm, we introduce the following activation matrix

for the tuple  $(x, z, \alpha, \theta, \lambda)$ :

$$\Omega_\alpha^t := \begin{bmatrix} X^t & 0 & 0 & 0 & 0 \\ 0 & Y^t & 0 & 0 & 0 \\ 0 & 0 & Y^t & 0 & 0 \\ 0 & 0 & 0 & X_{ll}^t & 0 \\ 0 & 0 & 0 & 0 & X_{ll}^t \end{bmatrix}. \quad (22)$$

where we have additionally introduced a diagonal random matrix  $Y^t \in \mathbb{R}^{nd \times nd}$ . The activation matrix  $\Omega_\alpha^t$  in (22) differs from  $\Omega^t$  in (15) as we allow  $(z, \alpha)$  to be updated independently, as captured by the activation matrix  $Y^t$ . Similar to (16), we compactly write the iterates as:

$$v_\alpha^{t+1} = v_\alpha^t + \Omega_\alpha^{t+1} (T v_\alpha^t - v_\alpha^t). \quad (23)$$

To compare (16) (**Algorithm 1**) and (23), we assign  $(z, \alpha)$  to edges and  $(x, \theta, \lambda)$  to agents (note that  $(\theta, \lambda)$  is only held by the  $l$ -th agent). We emphasize that this is done *only* for analysis, and **Algorithm 1** does not require storing/updating edge variables. In fact, the activation scheme in (16) is a special case of (23), and has a simple realization: whenever an agent is active, all incident edges become active and update their variables. We first characterize the condition that ensures the convergence of (23) in Theorem 1 and then show that (16) satisfies this condition.

*Theorem 1:* Consider the iterates generated by (23), and let assumptions 1-3 hold. We denote the maximum and minimum eigenvalue of  $L_u = E_u^\top E_u$  as  $\sigma_{\max}^{L_u}$  and  $\sigma_{\min}^{L_u}$  respectively. Under any activation schemes such that  $\mathbb{E}^t[\Omega_\alpha^{t+1}] = \Omega_\alpha \succ 0$ , set  $\mu_z = 2\mu_\theta$ ,  $\Delta_{ii} = \epsilon$  for  $i \in [m]$ . Then the following holds:

$$\mathbb{E}^t \left[ \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 \right] \leq \left( 1 - \frac{\rho_{\min} \eta}{1 + \eta} \right) \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2,$$

where for  $i \in [m], k \in [n]$ ,  $\mathbb{E}^t[X_{ii}^{t+1}] = p_i^X$ ,  $\mathbb{E}^t[Y_{kk}^{t+1}] = p_k^Y$ ,  $p^{\min} := \min_{i \in [m], k \in [n]} \{p_i^X, p_k^Y\}$ ,  $\eta$  is given by:

$$\eta = \min \left\{ \frac{2m_f M_f}{m_f + M_f} \frac{1}{\epsilon + \mu_\theta (\sigma_{\max}^{L_u} + 2)}, \frac{1}{2}, \frac{2}{5} \frac{\mu_\theta \sigma_{\min}^+}{m_f + M_f}, \frac{\mu_\theta \sigma_{\min}^+}{5\epsilon}, \frac{\sigma_{\min}^+}{5 \max\{1, \sigma_{\max}^{L_u}\}} \right\}, \quad (24)$$

and the scaling matrix  $\mathcal{H} = \text{diag}[\epsilon, 2\mu_z, \frac{2}{\mu_z}, \mu_\theta, \frac{1}{\mu_\theta}]$ .

*Proof:* See Appendix.

Note that the only requirement on the activation scheme is  $\mathbb{E}^t[\Omega_\alpha^{t+1}] \succ 0$ , i.e., each agent and each edge has a nonzero probability of being active at each round. We proceed to show that by specifying  $X^{t+1}$  alone (sampling agents and activating incident edges) in (16), the corresponding tuple  $(x, z, \alpha, \theta, \lambda)$  converges linearly in expectation.

*Corollary 1:* Consider the iterates generated by (16). Under the same setting as in Theorem 1, for arbitrary activation scheme such that  $\mathbb{E}^t[X^{t+1}] \succ 0$ , the corresponding tuple  $(x, z, \theta; \alpha, \lambda)$  converges linearly in expectation.

*Proof:* See Appendix.

*Remark 2:* Note that Theorem 1 and Corollary 1 does not impose any conditions on the proportion of agents who perform gradient or Newton updates. This is due to the fact that we can upper bound the matrix  $\Pi^t$  in Lemma 3 as  $\Pi^t \preceq 2M_f I$ . To quantify the impact on the proportion of agents performing Newton updates, we consider a special

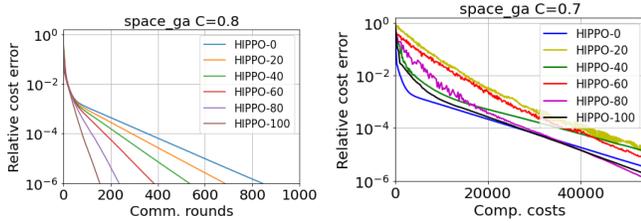


Fig. 1: Performance comparison with different percentage of agents performing Newton updates using comm. rounds (left) and computation costs (right) as the metric.

case of activation scheme in (23): we activate one variable uniformly at random with probability  $p = \frac{1}{m+n}$ , so that  $\mathbb{E}^t[\Omega_\alpha^{t+1}] = pI$ . Moreover, we denote the percentage of agents who perform Newton updates as  $q \in [0, 1]$ . Then the following holds:

$$\mathbb{E}^t \left[ \|e_v^{t+1}\|^2 \right] \leq (1-q) \left( 1 - \frac{p\eta_g}{1+\eta_g} \right) \|e_v^t\|^2 + q \left( 1 - \frac{p\eta_n}{1+\eta_n} \right) \|e_v^t\|^2,$$

where  $\|e_v^t\|^2 = \|v_\alpha^t - v_\alpha^*\|^2$  and  $\eta_g, \eta_n$  denotes the convergence coefficient of gradient and Newton updates respectively. Recall Lemma 3. Note that as  $t \rightarrow \infty$ ,  $\Pi_{ii}^t \rightarrow 0$  for  $i \in \mathcal{U}_{\text{Newton}}$ . This allow us to choose  $\Delta_{ii} = 0$  for  $i \in \mathcal{U}_{\text{Newton}}$ . By inspecting (46), we conclude that  $\eta_g \leq \eta_n$ , which translates to the fact that agents performing Newton updates converges at a faster rate. Therefore, as we increase  $q$  (more agents perform Newton updates), we obtain faster rate in expectation. This is corroborated in the next section.

## V. EXPERIMENTS

In this section, we present numerical experiments for HIPPO. We consider the following distributed  $\gamma$ -weighted LASSO problem:

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ l(x) = \left\{ \sum_{i=1}^m \frac{1}{2} \|A_i x - b_i\|^2 + \gamma \|x\|_1 \right\}.$$

We generate random graphs with  $m = 50$  agents by repetitively drawing edges according to a Bernoulli( $p$ ) distribution (with redrawing if necessary to ensure connectivity). We evenly distribute 3,000 data points from the space\_ga<sup>1</sup> dataset to agents, where  $d = 6$ . The averaged relative loss  $\frac{\frac{1}{m} \sum_{i=1}^m l(x_i^t) - l(x^*)}{\frac{1}{m} \sum_{i=1}^m l(x_i^0) - l(x^*)}$  is plotted against two metrics, namely communication rounds and computation costs. Note that a beneficial attribute of the proposed method is that the communication costs per agent in one iteration of the algorithm, regardless of the computational choice of the agents. The fraction of active agents is denoted by  $C \in (0, 1]$ , where agents are sampled uniformly at random. We compare the performance of the algorithm with different proportion of agents using Newton updates. For example, HIPPO-20 stands for 20% of agents perform Newton updates and the rest perform gradient updates. We observe that in Fig.1, as the number of agents performing Newton updates increases,

the performance strictly increases. However, when computation costs is used as the metric, the candidate with the optimal performance is not obvious.

## REFERENCES

- [1] A. Nedić and J. Liu, "Distributed optimization for control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 77–103, 2018.
- [2] T. Huang, N. M. Freris, P. R. Kumar, and L. Xie, "A synchrophasor data-driven method for forced oscillation localization under resonance conditions," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3927–3939, 2020.
- [3] O. Angatkina, A. G. Alleyne, and A. Wissa, "Robust Design and Evaluation of a Novel Modular Origami-Enabled Mobile Robot (OSCAR)," *Journal of Mechanisms and Robotics*, vol. 15, no. 2, 2022.
- [4] M. I. Qureshi and U. A. Khan, "Stochastic first-order methods over distributed data," in *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2022, pp. 405–409.
- [5] K.-D. Kim and P. R. Kumar, "Cyber-physical systems: A perspective at the centennial," *Proceedings of the IEEE*, vol. 100, pp. 1287–1308, 2012.
- [6] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, pp. 48–61, 2009.
- [7] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, pp. 944–966, 2015.
- [8] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2018.
- [9] R. Xin, S. Pu, A. Nedić, and U. A. Khan, "A general framework for decentralized optimization with first-order methods," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 1869–1889, 2020.
- [10] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network Newton distributed optimization methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 1, pp. 146–161, 2017.
- [11] F. Mansoori and E. Wei, "A fast distributed asynchronous Newton-based optimization algorithm," *IEEE Transactions on Automatic Control*, vol. 65, no. 7, pp. 2769–2784, 2020.
- [12] Y. Li, N. M. Freris, P. Voulgaris, and D. Stipanović, "DN-ADMM: Distributed Newton ADMM for multi-agent optimization," in *CDC*, 2021, pp. 3343–3348.
- [13] Y. Li, Y. Gong, N. M. Freris, P. Voulgaris, and D. Stipanović, "BFGS-ADMM for large-scale distributed optimization," in *CDC*, 2021, pp. 1689–1694.
- [14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, vol. 54, 2017, pp. 1273–1282.
- [15] X. Niu and E. Wei, "Fedhybrid: A hybrid primal-dual algorithm framework for federated optimization," in *arxiv:2106.01279*, 2021.
- [16] R. T. Rockafellar, *Convex analysis*. Princeton University Press, 1970.
- [17] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, p. 127–239, 2014.
- [18] J. D. Lee, Y. Sun, and M. A. Saunders, "Proximal Newton-type methods for minimizing composite functions," *SIAM Journal on Optimization*, vol. 24, no. 3, pp. 1420–1443, 2014.
- [19] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics With Applications*, vol. 2, pp. 17–40, 1976.
- [20] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.
- [21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, p. 1–122, 2011.
- [22] T. Lin, S. Ma, and S. Zhang, "Global convergence of unmodified 3-block ADMM for a class of convex minimization problems," *Journal of Scientific Computing*, vol. 76, no. 1, pp. 69–88, 2018.
- [23] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.
- [24] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 4051–4064, 2015.

<sup>1</sup><https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

- [25] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "DQM: Decentralized quadratically approximated alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5158–5173, 2016.
- [26] Y. Li, P. G. Voulgaris, and N. M. Freris, "A communication efficient quasi-newton method for large-scale distributed multi-agent optimization," in *ICASSP, 2022*, pp. 4268–4272.
- [27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

## APPENDIX

*Proof of Lemma 1:* We denote  $y^t = [\alpha^t; \beta^t]$  for some  $\alpha^t, \beta^t \in \mathbb{R}^{nd}$ . By premultiplying (6d) with  $B^\top$  and using (9), it holds that  $B^\top y^{t+1} = 0$ . By the definition of  $B$  matrix in (4) and the initialization  $\alpha^0 = -\beta^0$ , we conclude that  $\alpha^t = \beta^t$  for  $t \geq 0$ . Therefore, we can rewrite the dual update (6d) as:

$$\alpha^{t+1} = \alpha^t + \mu_z(A_s x^{t+1} - z^{t+1}), \quad (25)$$

$$-\alpha^{t+1} = -\alpha^t + \mu_z(A_d x^{t+1} - z^{t+1}). \quad (26)$$

By taking the sum and the difference of (25) and (26), we obtain for  $t \geq 0$ ,  $z^{t+1} = \frac{\mu_z}{2} E_u x^{t+1}$ ,  $\alpha^{t+1} = \alpha^t + \frac{\mu_z}{2} E_s x^{t+1}$ . Recalling the definition  $\phi^t = E_s^\top \alpha^t$  and by premultiplying the update of  $\alpha^{t+1}$  with  $E_s^\top$ , we obtain the update (11c). It remains to show that the primal update (8) is equivalent to (11a). The primal update (8) can be expressed as:

$$x^{t+1} = x^t - (H^t)^{-1} [\nabla F(x^t) + A^\top y^t + S \lambda^t + \mu_z A^\top (A x^t - B z^t) + \mu_\theta S (S^\top x^t - \theta^t)]. \quad (27)$$

Since  $y^t = [\alpha^t; -\alpha^t]$  and  $z^t = \frac{1}{2} E_u x^t$ , we obtain  $A^\top y^t = E_s^\top \alpha^t = \phi^t$  and  $\mu_z A^\top (A x^t - B z^t) = \frac{\mu_z}{2} L_u x^t$ . After substituting these equalities into (27), we obtain (11a). ■

*Proof of Lemma 2:* From the primal update (11a) and the identities  $\phi = E_s^\top \alpha$ ,  $L_s = E_s^\top E_s$ , we obtain:

$$\nabla F(x^t) + E_s^\top \alpha^t + S \lambda^t + \frac{\mu_z}{2} L_s x^t + \mu_\theta S (S^\top x^t - \theta^t) + H^t (x^{t+1} - x^t) = 0. \quad (28)$$

Using the dual update (11c)-(11d), we further obtain:

$$E_s^\top \alpha^t + \frac{\mu_z}{2} L_s x^t = E_s^\top \alpha^{t+1} - \frac{\mu_z}{2} L_s (x^{t+1} - x^t). \quad (29)$$

$$S \lambda^t + \mu_\theta S (S^\top x^t - \theta^t) = S \lambda^{t+1} - \mu_\theta S [S^\top (x^{t+1} - x^t) - (\theta^{t+1} - \theta^t)]. \quad (30)$$

By substituting (29) and (30) into (28), we obtain:

$$\nabla F(x^t) + E_s^\top \alpha^{t+1} - \frac{\mu_z}{2} L_s (x^{t+1} - x^t) + S [\lambda^{t+1} - \mu_\theta S^\top (x^{t+1} - x^t) + \mu_\theta (\theta^{t+1} - \theta^t)] + H^t (x^{t+1} - x^t) = 0. \quad (31)$$

Recall  $2D = L_s + L_u$ . Adding and subtracting  $(\mu_z D + \mu_\theta S S^\top + \Delta)(x^{t+1} - x^t)$  from (31), we obtain:

$$\begin{aligned} & \nabla F(x^t) + E_s^\top \alpha^{t+1} + \frac{\mu_z}{2} L_u (x^{t+1} - x^t) + S \lambda^{t+1} \\ & + \mu_\theta S (\theta^{t+1} - \theta^t) + \Delta (x^{t+1} - x^t) \\ & + (H^t - \mu_z D - \mu_\theta S S^\top - \Delta) (x^{t+1} - x^t) = 0. \end{aligned} \quad (32)$$

From Lemma 1 and KKT conditions, it holds that:

$$\frac{\mu_z}{2} L_u (x^{t+1} - x^t) = \mu_z E_u^\top (z^{t+1} - z^t), \quad (33)$$

$$\nabla F(x^*) + E_s^\top \alpha^* + S \lambda^* = 0. \quad (34)$$

By substituting (33) into (32) and subtracting (34), we obtain the desired. ■

*Proof of Lemma 3:* By the definition of  $e^t$  in (20):  $e^t =$

$\nabla F(x^t) - \nabla F(x^{t+1}) + J^t(x^{t+1} - x^t)$ , we obtain

$$\|e_i^t\| \leq \|\nabla f_i(x_i^t) - \nabla f_i(x_i^{t+1})\| + \|J_{ii}^t\| \|x_i^{t+1} - x_i^t\|.$$

For  $i \in \mathcal{U}_{\text{Gradient}}^t$ ,  $J_{ii}^t = 0$ . By assumption 1, it holds that:  $\|e_i^t\| \leq M_f \|x_i^{t+1} - x_i^t\|$ . For  $i \in \mathcal{U}_{\text{Newton}}^t$ ,  $J_{ii}^t = \nabla^2 f_i(x_i^t)$ . Using assumption 1 again, we obtain:

$$\|e_i^t\| \leq 2M_f \|x_i^{t+1} - x_i^t\|. \quad (35)$$

Moreover, using the fundamental theorem of calculus, we rewrite  $\nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t)$  as:

$$\begin{aligned} & \nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t) = \\ & \int_0^1 \nabla^2 f_i(s x_i^{t+1} + (1-s)x_i^t) (x_i^{t+1} - x_i^t) ds. \end{aligned}$$

After adding and subtracting  $\int_0^1 \nabla^2 f_i(x_i^t) (x_i^{t+1} - x_i^t) ds$ , we obtain:

$$\begin{aligned} & \nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t) = \int_0^1 \nabla^2 f_i(x_i^t) (x_i^{t+1} - x_i^t) ds \\ & + \int_0^1 (\nabla^2 f_i(s x_i^{t+1} + (1-s)x_i^t) - \nabla^2 f_i(x_i^t)) (x_i^{t+1} - x_i^t) ds. \end{aligned}$$

Therefore, we obtain the following chain of inequality:

$$\begin{aligned} & \|\nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t) - \nabla^2 f_i(x_i^t) (x_i^{t+1} - x_i^t)\| \\ & = \left\| \int_0^1 (\nabla^2 f_i(s x_i^{t+1} + (1-s)x_i^t) - \nabla^2 f_i(x_i^t)) (x_i^{t+1} - x_i^t) ds \right\| \\ & \leq \int_0^1 \|\nabla^2 f_i(s x_i^{t+1} + (1-s)x_i^t) - \nabla^2 f_i(x_i^t)\| \cdot \|x_i^{t+1} - x_i^t\| ds \\ & \leq \int_0^1 s L_f \|x_i^{t+1} - x_i^t\|^2 ds = \frac{L_f}{2} \|x_i^{t+1} - x_i^t\|^2. \end{aligned}$$

Note that for  $i \in \mathcal{U}_{\text{Newton}}^t$ , it holds that:

$$\begin{aligned} \|e_i^t\| &= \|\nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t) - \nabla^2 f_i(x_i^t) (x_i^{t+1} - x_i^t)\| \\ &\leq \frac{L_f}{2} \|x_i^{t+1} - x_i^t\|. \end{aligned}$$

Combining (35) and the above, we obtain: for  $i \in \mathcal{U}_{\text{Newton}}^t$ ,  $\|e_i^t\| \leq \min\{2M_f, \frac{L_f}{2} \|x_i^{t+1} - x_i^t\|\} \|x_i^{t+1} - x_i^t\|$ . Using the notation in (21), we obtain the desired. ■

*Proof of Theorem 1:* Part (i): We proceed by establishing that:

$$\|T v_\alpha^t - v_\alpha^*\|_{\mathcal{H}}^2 \leq \frac{1}{\eta} \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}}^2. \quad (36)$$

We first prove the following useful inequalities:

$$(\lambda^{t+1} - \lambda^t)^\top (\theta^{t+1} - \theta^t) \geq 0, \quad (37)$$

$$(\lambda^{t+1} - \lambda^*)^\top (\theta^{t+1} - \theta^*) \geq 0. \quad (38)$$

From the update (11b), its optimality condition, and dual update 11d, we obtain

$$0 \in \partial g(\theta^{t+1}) - \mu_\theta (\frac{1}{\mu_\theta} \lambda^t + S^\top x^{t+1} - \theta^{t+1}) = \partial g(\theta^{t+1}) - \lambda^{t+1}.$$

Therefore, we obtain (37) by using assumption 3 and (38) by using KKT conditions. We proceed to establish a useful upper bound for  $\|\alpha^{t+1} - \alpha^*\|^2 + \|\lambda^{t+1} - \lambda^*\|^2$  by first showing that  $[\alpha^t; \lambda^t]$  that lies in the column space of  $[E_s; S^\top]$ . Recalling the dual update (11c)-(11d) and the identity  $\phi^t = E_s^\top \alpha^t$ , it suffices to show that the column space of  $[0; \mu_\theta I]$  is a subset of the column space of  $[\frac{\mu_z}{2} E_s; \mu_\theta S^\top]$ . Consider fixed  $r^x \in \mathbb{R}^d$  and construct a vector  $r^y \in \mathbb{R}^{md}$  as  $r^y = [r^x; \dots; r^x]$ . Then it holds that

$$[\frac{\mu_z}{2} E_s; \mu_\theta S^\top] r^y = [0; \mu_\theta I] r^x,$$

which shows the desired inclusion relation. Moreover, it is easy to verify that there exists a unique pair  $[\alpha^*; \lambda^*]$  that lies in the column space of  $[E_s; S^\top]$ . By selecting  $\mu_z = 2\mu_\theta$ , we obtain that  $[\alpha^t - \alpha^*; \lambda^t - \lambda^*]$  lies in the column space of  $[E_s; S^\top]$ . Denoting the smallest positive eigenvalue of  $[E_s; S^\top][E_s^\top; S]$  as  $\sigma_{\min}^+$ , we obtain:

$$\begin{aligned} & \|\alpha^{t+1} - \alpha^*\|^2 + \|\lambda^{t+1} - \lambda^*\|^2 \\ & \leq \frac{1}{\sigma_{\min}^+} \|E_s^\top(\alpha^{t+1} - \alpha^*) + S(\lambda^{t+1} - \lambda^*)\|^2. \end{aligned} \quad (39)$$

Since  $F(x)$  is  $m_f$ -strongly convex and  $M_f$ -Lipschitz continuous, the following holds:

$$\begin{aligned} & \frac{m_f M_f}{m_f + M_f} \|x^{t+1} - x^*\|^2 + \frac{1}{m_f + M_f} \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 \\ & \leq (x^{t+1} - x^*)^\top (\nabla F(x^{t+1}) - \nabla F(x^*)) \\ & \leq -(x^{t+1} - x^*)^\top e^t - \epsilon (x^{t+1} - x^*)^\top (x^{t+1} - x^t) \\ & \quad - (x^{t+1} - x^*)^\top E_s^\top (\alpha^{t+1} - \alpha^*) \\ & \quad - (x^{t+1} - x^*)^\top S (\lambda^{t+1} - \lambda^* + \mu_\theta (\theta^{t+1} - \theta^t)) \\ & \quad - \mu_z (x^{t+1} - x^*)^\top E_u^\top (z^{t+1} - z^t), \end{aligned} \quad (40)$$

where the last inequality follows from Lemma 2. From the dual update and KKT conditions, we obtain

$$\begin{aligned} (x^{t+1} - x^*)^\top E_s^\top &= \frac{2}{\mu_z} (\alpha^{t+1} - \alpha^t)^\top, \\ (x^{t+1} - x^*)^\top S &= (\theta^{t+1} - \theta^*)^\top + \frac{1}{\mu_\theta} (\lambda^{t+1} - \lambda^t)^\top, \\ (x^{t+1} - x^*)^\top E_u^\top &= 2(z^{t+1} - z^*)^\top. \end{aligned}$$

Using these expressions for  $(x^{t+1} - x^*)^\top E_s^\top$ ,  $(x^{t+1} - x^*)^\top S$ , and  $(x^{t+1} - x^*)^\top E_u^\top$ , we rewrite (40) as:

$$\begin{aligned} & \frac{2m_f M_f}{m_f + M_f} \|x^{t+1} - x^*\|^2 + \frac{2}{m_f + M_f} \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 \\ & \leq -2(x^{t+1} - x^*)^\top e^t - 2\epsilon (x^{t+1} - x^*)^\top (x^{t+1} - x^t) \\ & \quad - \frac{4}{\mu_z} (\alpha^{t+1} - \alpha^t)^\top (\alpha^{t+1} - \alpha^*) - \frac{2}{\mu_\theta} (\lambda^{t+1} - \lambda^t)^\top (\lambda^{t+1} - \lambda^*) \\ & \quad - 2\mu_\theta (\theta^{t+1} - \theta^*)^\top (\theta^{t+1} - \theta^t) - 4\mu_z (z^{t+1} - z^*)^\top (z^{t+1} - z^t) \\ & = \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}}^2 - \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2 - \|v_\alpha^{t+1} - v_\alpha^t\|_{\mathcal{H}}^2 \\ & \quad - 2(x^{t+1} - x^*)^\top e^t, \end{aligned} \quad (41)$$

where have used (38), (37), and the identity:  $-2(a-b)^\top (a-c) = \|b-c\|^2 - \|a-b\|^2 - \|a-c\|^2$ . To establish (36), we need to show for some  $\eta > 0$ ,

$$\eta \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2 \leq \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}}^2 - \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2. \quad (42)$$

Using (41), it suffices to show:

$$\begin{aligned} & \eta \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2 \leq \frac{2m_f M_f}{m_f + M_f} \|x^{t+1} - x^*\|^2 \\ & \quad + \frac{2}{m_f + M_f} \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 + 2(x^{t+1} - x^*)^\top e^t. \end{aligned} \quad (43)$$

We expand  $\eta \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2$  as follows:

$$\begin{aligned} & \eta \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2 = \eta \left( \epsilon \|x^{t+1} - x^*\|^2 + 2\mu_z \|z^{t+1} - z^*\|^2 \right. \\ & \quad \left. + \frac{2}{\mu_z} \|\alpha^{t+1} - \alpha^*\|^2 + \mu_\theta \|\theta^{t+1} - \theta^*\|^2 + \frac{1}{\mu_\theta} \|\lambda^{t+1} - \lambda^*\|^2 \right). \end{aligned}$$

We proceed to establish upper bounds for each component. From Lemma 2, the following holds:

$$\begin{aligned} & E_s^\top (\alpha^{t+1} - \alpha^*) + S (\lambda^{t+1} - \lambda^*) = -\{\nabla F(x^{t+1}) - \nabla F(x^*) \\ & \quad + \epsilon (x^{t+1} - x^*) + \mu_z E_u^\top (z^{t+1} - z^t) + \mu_\theta S (\theta^{t+1} - \theta^t) + e^t\}. \end{aligned}$$

Using (39) and selecting  $\mu_z = 2\mu_\theta$ , we further obtain:

$$\begin{aligned} & \frac{2}{\mu_z} \|\alpha^{t+1} - \alpha^*\|^2 + \frac{1}{\mu_\theta} \|\lambda^{t+1} - \lambda^*\|^2 \leq \frac{5}{\mu_\theta \sigma_{\min}^+} (\|e^t\|^2 \\ & \quad + \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 + \epsilon^2 \|x^{t+1} - x^t\|^2 \\ & \quad + \mu_\theta^2 \|\theta^{t+1} - \theta^t\|^2 + \sigma_{\max}^{L_u} \mu_z^2 \|z^{t+1} - z^t\|^2), \end{aligned} \quad (44)$$

where  $\sigma_{\max}^{L_u}$  is the maximum eigenvalue of  $L_u = E_u^\top E_u$ . Moreover, since  $z^{t+1} - z^* = \frac{1}{2} E_u (x^{t+1} - x^*)$ , it holds that  $2\mu_z \|z^{t+1} - z^*\|^2 \leq \frac{\mu_z \sigma_{\max}^{L_u}}{2} \|x^{t+1} - x^*\|^2$ . Similarly, from the dual update (11b), the following holds:  $\mu_\theta \|\theta^{t+1} - \theta^*\|^2 \leq 2\mu_\theta \|x^{t+1} - x^*\|^2 + \frac{2}{\mu_\theta} \|\lambda^{t+1} - \lambda^t\|^2$ . Using these bounds and considering (43), we need to show the following holds for  $\eta > 0$ ,

$$\begin{aligned} & \eta \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}}^2 \leq \eta \left\{ \frac{5}{\mu_\theta \sigma_{\min}^+} \left[ \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 \right. \right. \\ & \quad \left. \left. + \epsilon^2 \|x^{t+1} - x^t\|^2 + \mu_\theta^2 \|\theta^{t+1} - \theta^t\|^2 + \|e^t\|^2 \right] \right. \\ & \quad \left. + \sigma_{\max}^{L_u} \mu_z^2 \|z^{t+1} - z^t\|^2 + \frac{2}{\mu_\theta} \|\lambda^{t+1} - \lambda^t\|^2 \right. \\ & \quad \left. + (\epsilon + 2\mu_\theta + \frac{\mu_z \sigma_{\max}^{L_u}}{2}) \|x^{t+1} - x^*\|^2 \right\} \leq \\ & \frac{2m_f M_f}{m_f + M_f} \|x^{t+1} - x^*\|^2 + \frac{2}{m_f + M_f} \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 \\ & \quad + \|v_\alpha^{t+1} - v_\alpha^t\|_{\mathcal{H}}^2 + 2(x^{t+1} - x^*)^\top e^t. \end{aligned} \quad (45)$$

Note that  $-\zeta \|e^t\|^2 - \frac{1}{\zeta} \|x^{t+1} - x^*\|^2 \leq 2(x^{t+1} - x^*)^\top e^t$  holds for any  $\zeta > 0$ , it is sufficient to show the following to establish (42):

$$\begin{aligned} & \zeta \|x^{t+1} - x^t\|_{(\Pi^t)^2}^2 + \eta \left\{ \frac{5}{\mu_\theta \sigma_{\min}^+} \left[ \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2 \right. \right. \\ & \quad \left. \left. + \|x^{t+1} - x^t\|_{\epsilon^2 I + (\Pi^t)^2}^2 + \mu_\theta^2 \|\theta^{t+1} - \theta^t\|^2 \right] \right. \\ & \quad \left. + \sigma_{\max}^{L_u} \mu_z^2 \|z^{t+1} - z^t\|^2 + \frac{2}{\mu_\theta} \|\lambda^{t+1} - \lambda^t\|^2 \right. \\ & \quad \left. + (\epsilon + 2\mu_\theta + \frac{\mu_z \sigma_{\max}^{L_u}}{2}) \|x^{t+1} - x^*\|^2 \right\} \leq \\ & \left( \frac{2m_f M_f}{m_f + M_f} - \frac{1}{\zeta} \right) \|x^{t+1} - x^*\|^2 + \epsilon \|x^{t+1} - x^t\|^2 \\ & \quad + 2\mu_z \|z^{t+1} - z^t\|^2 + \frac{2}{\mu_z} \|\alpha^{t+1} - \alpha^t\|^2 + \mu_\theta \|\theta^{t+1} - \theta^t\|^2 \\ & \quad + \frac{1}{\mu_\theta} \|\lambda^{t+1} - \lambda^t\|^2 + \frac{2}{m_f + M_f} \|\nabla F(x^{t+1}) - \nabla F(x^*)\|^2. \end{aligned}$$

By selecting  $\eta$  as in (24), we obtain the desired. Moreover observe that:

$$\eta \leq \frac{\mu_\theta \sigma_{\min}^+}{5} \cdot \frac{\|x^{t+1} - x^t\|_{\epsilon I - \zeta (\Pi^t)^2}^2}{\|x^{t+1} - x^t\|_{\epsilon^2 I + (\Pi^t)^2}^2}. \quad (46)$$

We have now established (36). We proceed to show the asynchronous updates converges linearly in expectation. Note that:

$$\begin{aligned} & \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 = \|v_\alpha^t + \Omega_\alpha^{t+1} (T v_\alpha^t - v_\alpha^t) - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 \\ & = \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 + 2(v_\alpha^t - v_\alpha^*)^\top \mathcal{H} \Omega_\alpha^{-1} \Omega_\alpha^{t+1} (T v_\alpha^t - v_\alpha^t) \\ & \quad + (T v_\alpha^t - v_\alpha^t)^\top \Omega_\alpha^{t+1} \mathcal{H} \Omega_\alpha^{-1} \Omega_\alpha^{t+1} (T v_\alpha^t - v_\alpha^t). \end{aligned}$$

Since  $\Omega_\alpha^{t+1}$ ,  $\Omega_\alpha^{-1}$ , and  $\mathcal{H}$  are all diagonal matrices, they commute with each other. Moreover, since each sub-block of

$\Omega_\alpha^{t+1}$  is  $I_d$  or 0, it holds that  $\Omega_\alpha^{t+1}\Omega_\alpha^{t+1} = \Omega_\alpha^{t+1}$ . Therefore,

$$\begin{aligned} \mathbb{E}^t \left[ \|v_\alpha^{t+1} - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 \right] &= \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 + \|Tv_\alpha^t - v_\alpha^t\|_{\mathcal{H}}^2 \\ &\quad + 2(v_\alpha^t - v_\alpha^*)^\top \mathcal{H}(Tv_\alpha^t - v_\alpha^t) \\ &\leq \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2 - \frac{\eta}{1+\eta} \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}}^2 \\ &\leq \left(1 - \frac{\eta}{1+\eta}\right) \|v_\alpha^t - v_\alpha^*\|_{\mathcal{H}\Omega_\alpha^{-1}}^2, \end{aligned}$$

which is the desired.  $\blacksquare$

*Proof of Corollary 1:* We first distribute each  $\alpha_k, k \in [n]$ , to each edge and label agents and edges with an arbitrary order. For each edge  $\mathcal{E}_k$ , we write  $\mathcal{E}_k = (i, j)$  with the convention  $i < j$ . Then for each agent  $i$ , we can divide the incident edges to two groups:  $\mathcal{P}_i = \{k : \mathcal{E}_k = (i, j), j \in \mathcal{N}_i\}$  and  $\mathcal{S}_i = \{k : \mathcal{E}_k = (j, i), j \in \mathcal{N}_i\}$ . Consider the activation scheme using  $\Omega_\alpha^{t+1}$ . Recall  $\alpha_k^{t+1} = \alpha_k^t + \frac{\mu_z}{2}(x_i^{t+1} - x_j^{t+1})$ . Then the dual updates are described by:

$$\begin{aligned} \phi_i^{t+1} &= \phi_i^t + \frac{\mu_z}{2} X_{ii}^{t+1} \sum_{j \in \mathcal{N}_i} (x_i^{t+1} - x_j^{t+1}) \\ &= \phi_i^t + X_{ii}^{t+1} \left\{ \sum_{k \in \mathcal{P}_i} (\alpha_k^{t+1} - \alpha_k^t) + \sum_{k \in \mathcal{S}_i} (\alpha_k^t - \alpha_k^{t+1}) \right\}. \end{aligned}$$

Therefore, if  $X_{ii}^{t+1} = I_d$ , then  $Y_{kk}^{t+1} = I_d$  for  $k \in \mathcal{P}_i \cup \mathcal{S}_i$  for the corresponding  $\Omega_\alpha^{t+1}$ , i.e., all incident edges are active. It can be verified that we can map  $X^{t+1}$  to  $Y^{t+1}$  as:

$$Y^{t+1} = \mathbf{Blkdiag} \left( \left[ \frac{E_u X^{t+1} (\mathbf{1} \otimes I_d)}{2} \right] \right),$$

where  $\lceil \cdot \rceil$  is the entry-wise ceiling operation and  $\mathbf{1} \in \mathbb{R}^m$  is the all one vector. To show  $\mathbb{E}^t[\Omega_\alpha^{t+1}] \succ 0$ , we only need to show  $\mathbb{E}^t[Y^{t+1}] \succ 0$ , which amounts to showing that  $\mathbb{E}^t \left[ \left[ \frac{E_u X^{t+1} (\mathbf{1} \otimes I_d)}{2} \right]_k \right] \in \mathbb{R}^{d \times d}, k \in [n]$ , is positive definite.

We can explicitly express each block as:

$$\left[ \frac{E_u X^{t+1} (\mathbf{1} \otimes I_d)}{2} \right]_k = \left[ \frac{X_{ii}^{t+1} + X_{jj}^{t+1}}{2} \right],$$

where  $(i, j) \in \mathcal{E}_k$ . Therefore,

$$\mathbb{E}^t \left[ \left[ \frac{E_u X^{t+1} (\mathbf{1} \otimes I_d)}{2} \right]_k \right] = \mathbb{E}^t \left[ \left[ \frac{X_{ii}^{t+1} + X_{jj}^{t+1}}{2} \right] \right] \succ 0,$$

which shows that  $\mathbb{E}^t[Y^{t+1}] \succ 0$ .  $\blacksquare$