

## AN ACCELERATED PRIMAL DUAL ALGORITHM WITH BACKTRACKING FOR DECENTRALIZED CONSTRAINED OPTIMIZATION

QIUSHUI XU<sup>1</sup>, NECDET S. AYBAT<sup>1,\*</sup>, MERT GÜRBÜZBALABAN<sup>2</sup>

<sup>1</sup>*Department of Industrial Engineering, Pennsylvania State University, University Park, PA 16802, USA*

<sup>2</sup>*Department of Management Science & Information Systems, Rutgers University, Piscataway, NJ 08854, USA*

**Abstract.** We propose a distributed accelerated primal-dual method with backtracking (D-APDB) for cooperative multi-agent constrained consensus optimization problems over an undirected network of agents, where only those agents connected by an edge can directly communicate to exchange large-volume data vectors using a high-speed, short-range communication protocol, e.g., WiFi, and we also assume that the network allows for one-hop simple information exchange beyond immediate neighbors as in LoRaWAN protocol. The objective is to minimize the sum of agent-specific composite convex functions over agent-specific private constraint sets. Unlike existing decentralized primal-dual methods that require knowledge of the Lipschitz constants, D-APDB automatically adapts to local smoothness by employing a distributed backtracking step-size search. Each agent relies only on first-order oracles associated with its own objective and constraint functions and on local communications with the neighboring agents, without any prior knowledge of Lipschitz constants. We establish  $\mathcal{O}(1/K)$  convergence guarantees for sub-optimality, infeasibility and consensus violation, under standard assumptions on smoothness and on the connectivity of the communication graph. To our knowledge, when nodes have private constraints, especially when they are nonlinear convex constraints onto which projections are not cheap to compute, D-APDB is the first distributed method with backtracking that achieves the optimal convergence rate for the class of constrained composite convex optimization problems. We provide numerical results for D-APDB on a distributed QCQP problem and distributed primal SVM training, illustrating the potential performance gains that can be achieved by D-APDB.

**Keywords.** Adaptive methods; Backtracking; Decentralized optimization; Parameter-free optimization methods.

**2020 Mathematics Subject Classification.** 90C30, 90C25, 68W15, 90C46, 90C35, 65K05.

### 1. INTRODUCTION

Modern datasets are large, and data are often acquired by computational agents connected over a communication network. In this context, an agent denotes any computational entity (e.g., device, node, sensor, or processor) capable of local computation and communication. In such systems, information is inherently distributed across agents that exchange messages only with neighbors, rendering centralized processing impractical or infeasible due to bandwidth,

---

\*Corresponding author.

E-mail addresses: [qjx5019@psu.edu](mailto:qjx5019@psu.edu) (Q. Xu), [nsa10@psu.edu](mailto:nsa10@psu.edu) (N.S. Aybat), [mg1366@rutgers.edu](mailto:mg1366@rutgers.edu) (M. Gürbüzbalaban).

Received 9 December 2025; Accepted 3 February 2026; Published online 1 April 2026

©2026 Journal of Nonlinear and Variational Analysis

latency, energy, and privacy constraints. These factors motivate *decentralized optimization* algorithms, wherein data and computation reside across multiple agents and the global objective is optimized via local updates and limited message passing, without sharing raw data. Such methods arise in a wide range of applications, including machine learning with decentralized data, control and coordination in multi-robot systems, smart grids, signal processing and estimation over sensor networks [2, 7, 15, 33, 42]. We refer the reader to [40, 46] for additional examples in a broader context. In these settings, the goal is to collaboratively solve an optimization problem defined by agent-specific objectives and/or constraints through utilizing the computing capability of agents across the network—without a central coordinator—using only local computations and peer-to-peer communication among the neighboring nodes.

Let  $\mathcal{N} \triangleq \{1, \dots, N\}$  denote the set of agents in the network. In many applications, the global feasible set is given by the intersection of agent-specific nonlinear constraint sets that encode local information or individual operational requirements. These constraints are often naturally expressed as convex functional constraints of the form  $-g_i(x) \in \mathcal{K}_i$  in the decision variable  $x$ , where each  $g_i$  is a vector-valued mapping and  $\mathcal{K}_i$  is a closed convex cone, for each  $i \in \mathcal{N}$ . Objective functions are frequently composite as well, combining smooth losses with nonsmooth regularizers or indicator functions that model various constraints. In this paper, we consider the following conic constrained optimization problems:

$$\varphi^* \triangleq \min_{x \in \mathbb{R}^n} \sum_{i \in \mathcal{N}} \varphi_i(x) \triangleq \phi_i(x) + f_i(x) \quad \text{s.t.} \quad -g_i(x) \in \mathcal{K}_i \quad \forall i \in \mathcal{N} \triangleq \{1, \dots, N\}, \quad (1.1)$$

where  $\phi_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is a proper, closed convex (possibly nonsmooth) function,  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  is a smooth convex function,  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}^{m_i}$  is a smooth  $\mathcal{K}_i$ -convex function<sup>1</sup>, and  $\mathcal{K}_i \subseteq \mathbb{R}^{m_i}$  is a closed convex cone for  $i \in \mathcal{N}$ . The data of agent  $i$  defining  $\varphi_i$ ,  $g_i$  and  $\mathcal{K}_i$  are private and not shared globally, and agents cooperate to solve (1.1) using only local computation and limited communication. This is a general class of problems that include many applications and important special cases [9, 30, 33, 39, 40, 46].

Over the past few decades, a rich literature has emerged on decentralized constrained optimization problems that are special cases of the class of problems in (1.1). In particular, distributed (sub)gradient schemes, gradient tracking methods, and their accelerated and proximal variants have been extensively studied under a variety of assumptions on smoothness, convexity, and network connectivity, e.g., [18, 22, 32, 38, 49, 50, 54]. Other approaches include primal–dual type methods based on Augmented Lagrangian and ADMM formulations, e.g., [2, 4, 5, 7, 9, 18, 19, 27]. The latter class of algorithms typically rely on Lagrangian formulations and exploit separability across agents to achieve fully distributed implementations. Except very few [5, 19], none of the aforementioned methods can handle (1.1) in its full generality without employing projections onto  $\{x : -g_i(x) \in \mathcal{K}_i\}$ . More importantly, a persistent challenge in nearly all of these methods is to appropriately choose stepsizes with theoretical convergence guarantees. Such guarantees often require stepsizes that depend on global problem parameters—such as global Lipschitz constants for  $\nabla f_i$  and Jacobians of  $g_i$  for  $i \in \mathcal{N}$ , that are difficult to estimate, rarely known a-priori, and may vary widely across agents. In practice, this leads

<sup>1</sup>Given a closed convex cone  $\mathcal{K} \subset \mathbb{R}^m$ , and  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is  $\mathcal{K}$ -convex if  $f(\lambda x' + (1 - \lambda)x'') \preceq_{\mathcal{K}} \lambda f(x') + (1 - \lambda)f(x'')$  holds for all  $x', x''$  and  $\lambda \in [0, 1]$ , where  $\preceq_{\mathcal{K}}$  denotes the partial order induced by  $\mathcal{K}$ , i.e.,  $y' \preceq_{\mathcal{K}} y''$  for  $y', y'' \in \mathbb{R}^m$  when  $y'' - y' \in \mathcal{K}$ .

practitioners to grid-search or adopt conservative, globally synchronized stepsizes that can significantly degrade the convergence speed. The difficulty is further exacerbated in the presence of nonlinear functional constraints, where the relevant Lipschitz constants for the Lagrangian gradients depend on the dual iterates, which typically evolve within a cone; therefore, the Lipschitz constants may be only locally finite but not globally.

The literature on parameter-free decentralized algorithms—i.e., methods that do not require prior knowledge of Lipschitz constants—remains limited. A small body of work studies decentralized adaptive gradient methods that exploit past gradient information to choose stepsizes adaptively across iterations. For example, [37] develops a distributed adaptive moment estimation method for online smooth convex and nonconvex minimization and establishes regret bounds, while [12] proposes a general framework for converting centralized adaptive gradient schemes such as RMSProp, Adam, and AdaGrad into decentralized counterparts. We should emphasize that for the methods proposed in [12] to have convergence guarantees, it is required that the step size is sufficiently small inversely proportional to  $1/\max_{i=1,\dots,N}\{L_{f_i}\}$ , where  $L_{f_i}$  is the Lipschitz constant of  $\nabla f_i$ , and this would require some information exchange beyond the immediate neighbors on the network. Furthermore, both [37] and [12] consider unconstrained smooth optimization problems and do not handle constraints.

The work [31] proposes a parameter-free decentralized algorithm for nonconvex stochastic optimization and the proposed method does not require knowledge of  $\{L_{f_i}\}_{i \in \mathcal{N}}$  or topological information about the communication network. Another line of work [1] introduces a port-Hamiltonian systems framework for the design and analysis of distributed optimization algorithms, and develops methods for smooth, strongly convex, and unconstrained decentralized problems. Convergence of the proposed decentralized methods is guaranteed for certain special graph classes or when the stepsize is chosen below a threshold that depends on global network information [1, Corollaries 2 and 3]. However, we should emphasize that the proposed approaches in [1, 31] both focus on smooth minimization problems and they do not address either the node-specific constraints or the non-smooth terms in the objective.

More recently, [29] considered distributed smooth strongly convex unconstrained minimization problems and proposed a parameter-free decentralized optimization algorithm that employs a local backtracking line search to select stepsizes adaptively—without requiring global information or heavy communication, and the authors analyze two different implementations of the proposed method. The first implementation uses a global min-consensus to synchronize agents' stepsizes, which can be computed in practice via flooding protocols over low-power wide-area networks, e.g., LoRa [3, 23, 25] (low power, long range), as it requires simple information exchange beyond immediate neighbors. On the other hand, the second one replaces the global min-consensus update with a local min-consensus update, restricting communication to immediate neighbors; however, as mentioned in the follow-up work [28], this increased locality comes at the cost of weaker theoretical guarantees due to stronger assumptions requiring that the iterates stay bounded, and potentially non-monotone convergence trajectories. In [28], the authors introduce a fully decentralized algorithm in which each agent adaptively selects its own stepsize using only neighbor-to-neighbor communication and no global information is required for its implementation—agents need not even know whether the problem is strongly convex. The algorithm retains strong guarantees: it converges at a *linear* rate under strong convexity and at a  $\mathcal{O}(1/T)$ -*sublinear* rate under mere convexity, matching the best-known rates for

parameter-dependent, nonadaptive distributed methods. In the merely convex setting, the convergence guarantees for the method in [28] also require that the primal-dual iterate sequence stays bounded (see [28, Theorem 10]) as in their earlier work [29]—since neither work [28, 29] can handle indicator functions to directly impose boundedness, the authors argue that this requirement can be satisfied by some particular choice of algorithmic parameter sequences.

All of the previously mentioned works are designed to address *unconstrained* decentralized problems of the form  $\min_{x \in \mathbb{R}^n} \sum_{i \in \mathcal{N}} f_i(x)$ , where each  $f_i$  is smooth with a Lipschitz-continuous gradient. Consequently, they do not apply to the more general constrained problems we consider in (1.1), which feature agent-specific local constraints and nonsmooth terms  $\phi_i(x)$  in the objective. A more recent work [14] is the first to accommodate nonsmoothness, focusing on problems of the form  $\min \sum_{i \in \mathcal{N}} f_i(x) + \phi(x)$ , where  $\phi$  is a convex, nonsmooth function known to *all* agents in the network. The work in [14] introduces a novel three-operator splitting technique and proposes a method that requires neither global network information nor extensive inter-agent communication. The resulting adaptive decentralized method enjoys robust convergence guarantees and outperforms existing nonadaptive approaches.

These aforementioned results on adaptive distributed methods are exciting developments for decentralized optimization; nevertheless, none of these existing methods that we discussed above can handle a problem of the form (1.1) in its full generality. Indeed, to our knowledge, there is no decentralized optimization method that does not rely on the a priori knowledge of Lipschitz constants and that can handle agent-specific local constraints defined by nonlinear convex functions.

We should mention that for *centralized unconstrained* optimization, where a central node can aggregate and process all data, a wide range of parameter-free methods is available. These include centralized adaptive gradient schemes such as AdaGrad [16], RMSProp [51], Adam [26], and their variants [34, 44], as well as methods based on Barzilai–Borwein stepsizes [8, 10, 57] and adaptive techniques that estimate local curvature [35, 36]. For centralized constrained convex optimization, backtracking schemes can estimate local Lipschitz constants on the fly, yielding convergence-rate guarantees without prior knowledge of problem-specific smoothness parameters. In particular, Lagrangian reformulations of (1.1) lead to convex–concave saddle-point (SP) problems, for which primal–dual methods with backtracking have been developed [20, 24]—on a different note, there are also backtracking-based algorithms for centralized non-convex min–max formulations [53, 55], which can be used for (1.1) as well. However, these methods fundamentally rely on centralized aggregation; therefore, they do not extend directly to the decentralized computation setting we consider in this paper, where no single node has access to global information. There are also adaptive distributed methods in federated learning or central-server settings, e.g., [13, 43, 52]; but these algorithms still require a central node to aggregate information from the network and are thus not applicable to the fully decentralized setting over an arbitrary connected undirected communication network considered in this paper.

To address this gap, we propose D-APDB, a novel decentralized primal–dual algorithm that incorporates a backtracking mechanism for *local*, agent-specific stepsize selection to solve (1.1). Unlike existing decentralized primal-dual methods that require knowledge of the Lipschitz constants, D-APDB automatically adapts to local smoothness by employing a distributed backtracking step-size search. Each agent relies only on first-order oracles associated with its own objective and constraint functions, without any prior knowledge of Lipschitz constants. At each

iteration of D-APDB, each agent locally communicates a  $n$ -dimensional vector one time with the neighboring agents, and the agents across the network collectively implement a max-consensus one time. As explained in [29], the max-consensus protocol is well-suited to existing wireless mesh network technologies; more precisely, LoRa [3, 23, 25] enables wide-area coverage at low data rates, which is advantageous for network-wide flooding where each transmission reaches all nodes in one hop but conveys limited information. We establish  $\mathcal{O}(1/K)$  convergence guarantees for sub-optimality, infeasibility and consensus violation, under standard assumptions on smoothness and on the connectivity of the communication graph. To our knowledge, D-APDB is the first distributed method with backtracking that achieves the optimal convergence rate for the class of composite convex optimization problems subject to functional agent-specific convex constraints. Furthermore, we propose a variant of our method, which we call D-APDB0, tailored to the setting with  $g_i(\cdot) = 0$  for all  $i \in \mathcal{N}$  while agents can still have closed convex functions  $\phi_i$  in their local objectives. D-APDB0 can also achieve  $\mathcal{O}(1/K)$  convergence guarantees for sub-optimality and consensus violation –to the best of our knowledge, this is the first time a rate result is provided for a decentralized method that can handle node-specific closed convex functions without requiring a priori knowledge on Lipschitz constants. Finally, we present numerical results for D-APDB0 and D-APDB on a distributed  $\ell_1$ -norm regularized QP and QCQP problems, illustrating the potential performance gains achievable with our proposed backtracking framework.

**Notation.** Throughout the text  $\|\cdot\|$  denotes the Euclidean norm and the spectral norm when the argument is a vector and a matrix, respectively. Given two sets  $A$  and  $B$ ,  $A \times B$  denotes the Cartesian product, and  $\prod_{i=1}^M A_i$  denotes the Cartesian product of a collection of sets  $\{A_i\}_{i=1}^M$ . In few places, we abuse the notation  $\prod_{i=1}^M a_i$  to denote the multiplication of a collection of real numbers  $\{a_i\}_{i=1}^M \subset \mathbb{R}$ . Given a set  $S$ , let  $\mathbb{1}_S(\cdot)$  denote its indicator function, i.e.,  $\mathbb{1}_S(x) = 0$  if  $x \in S$  and is equal to  $+\infty$  otherwise. For  $S \subset \mathbb{R}^n$ ,  $\mathcal{P}_S(\cdot)$  denotes the Euclidean projection onto  $S$ ; moreover,  $d_S : \mathbb{R}^n \rightarrow \mathbb{R}_+$  denotes the distance function, i.e.,  $d_S(x) = \|x - \mathcal{P}_S(x)\|$ . Given a convex cone  $\mathcal{K} \subseteq \mathbb{R}^m$ , let  $\mathcal{K}^*$  denote its dual cone, i.e.,  $\mathcal{K}^* \triangleq \{\theta \in \mathbb{R}^m : \langle \theta, w \rangle \geq 0, \forall w \in \mathcal{K}\}$ . Given a discrete set  $\mathcal{S}$ ,  $|\mathcal{S}|$  denotes the cardinality of  $\mathcal{S}$ . The set  $\mathbb{S}_{++}^{n_x}$  denotes the set of  $n_x \times n_x$  positive definite matrices. Throughout the text, we use **diag**( $\cdot$ ) to construct block-diagonal matrices, i.e., given  $G_i \in \mathbb{R}^{m_i \times n}$  for  $i \in \mathcal{N}$ ,  $G = \mathbf{diag}([G_i]_{i \in \mathcal{N}})$  denotes a block diagonal matrix with diagonal blocks being  $G_i$  for  $i \in \mathcal{N}$ ; hence,  $G \in \mathbb{R}^{m \times n|\mathcal{N}|}$  where  $m = \sum_{i \in \mathcal{N}} m_i$ . We use  $I_m$  to denote the  $m \times m$  identity matrix, and  $\mathbf{1}_m$  to denote the  $m$ -dimensional vector of ones.

## 2. ASSUMPTIONS AND THE MAIN RESULTS

We investigate distributed methods for efficiently solving the problem in (1.1) over a network of  $\mathcal{N}$  agents with computing and message passing capability.

**2.1. Assumptions.** This network is modeled as an undirected, static, connected graph  $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ , where  $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$  and  $(i, j) \in \mathcal{E}$  if there is communication link (edge) between  $i \in \mathcal{N}$  and  $j \in \mathcal{N}$ . Our assumptions on  $\mathcal{G}$  are formally stated below.

**Assumption 1.** Let  $\mathcal{G} = (\mathcal{N}, \mathcal{E})$  denote a connected undirected graph of  $N$  computing nodes, where  $\mathcal{N} \triangleq \{1, \dots, N\}$  and  $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$  denotes the set of edges – without loss of generality assume that  $(i, j) \in \mathcal{E}$  implies  $i < j$ . Suppose that nodes  $i \in \mathcal{N}$  and  $j \in \mathcal{N}$  can exchange

$n$ -dimensional data vectors only if  $(i, j) \in \mathcal{E}$ . Moreover, given arbitrary  $\{\eta_i\}_{i \in \mathcal{N}} \subset \mathbb{R}$  such that each  $\eta_i$  is only known to node  $i \in \mathcal{N}$ , suppose that the network is capable of computing  $\max_{i \in \mathcal{N}} \{\eta_i\}$  in such a way that this quantity would be available to all the nodes. Finally, each agent  $i \in \mathcal{N}$  has only access to the agent-specific functions  $\phi_i(\cdot)$ ,  $g_i(\cdot)$ , and the cone  $\mathcal{K}_i$ .

**Definition 2.1.** For  $i \in \mathcal{N}$ , let  $\mathcal{N}_i \triangleq \{j \in \mathcal{N} : (i, j) \in \mathcal{E} \text{ or } (j, i) \in \mathcal{E}\}$  denote the set of neighboring nodes and  $d_i \triangleq |\mathcal{N}_i|$  is its degree, and also let  $d_{\max} \triangleq \max_{i \in \mathcal{N}} \{d_i\}$ .

According to Assumption 1, each agent  $i \in \mathcal{N}$  can communicate (send/receive vector data) only with  $j \in \mathcal{N}_i$ , which denotes the set of immediate neighbors of agent  $i$ . We assume that the agents are *collaborative* and the objective is to solve (1.1) in a distributed manner. For this purpose each agent  $i \in \mathcal{N}$  stores/updates its own copy of the decision vector, i.e.,  $x_i \in \mathbb{R}^n$  corresponds to agent- $i$ ; hence, (1.1) can be equivalently formulated as

$$\min_{\mathbf{x}} \varphi(\mathbf{x}) \triangleq \phi(\mathbf{x}) + f(\mathbf{x}) \quad \text{s.t.} \quad -G(\mathbf{x}) \in \mathcal{K}, \quad \mathbf{A}\mathbf{x} = 0, \quad (2.1)$$

where  $\mathbf{x} = [x_i]_{i \in \mathcal{N}} \in \mathcal{X} \triangleq \mathbb{R}^{n|\mathcal{N}|}$  denotes the long vector obtained by vertically concatenating the local decision variables  $x_i \in \mathbb{R}^n$  for all agents  $i \in \mathcal{N}$ ,  $\phi(\mathbf{x}) = \sum_{i \in \mathcal{N}} \phi_i(x_i)$ ,  $f(\mathbf{x}) = \sum_{i \in \mathcal{N}} f_i(x_i)$ ,  $G(\mathbf{x}) = [g_i(x_i)]_{i \in \mathcal{N}} \in \mathbb{R}^m$  with  $m = \sum_{i \in \mathcal{N}} m_i$ ,  $\mathcal{K} = \prod_{i \in \mathcal{N}} \mathcal{K}_i$  denotes the Cartesian product, and  $A \in \mathbb{R}^{|\mathcal{E}| \times n|\mathcal{N}|}$  is a block matrix such that  $A = H \otimes \mathbf{I}_n$  where  $H \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{N}|}$  is the oriented edge-node incidence matrix, i.e., the entry  $H_{(i,j),l}$ , corresponding to edge  $(i, j) \in \mathcal{E}$  and  $l \in \mathcal{N}$ , is equal to 1 if  $l = i$ , -1 if  $l = j$ , and 0 otherwise. Note that  $A^\top A = H^\top H \otimes \mathbf{I}_n = \Omega \otimes \mathbf{I}_n$ , where  $\Omega \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$  denotes the graph Laplacian of  $\mathcal{G}$ , i.e.,  $\Omega_{ii} = d_i$ ,  $\Omega_{ij} = -1$  if  $(i, j) \in \mathcal{E}$  or  $(j, i) \in \mathcal{E}$ , and equal to 0 otherwise.

We make the following assumptions on  $\{\phi_i\}_{i \in \mathcal{N}}$ ,  $\{f_i\}_{i \in \mathcal{N}}$  and  $\{g_i\}_{i \in \mathcal{N}}$  throughout the paper.

**Assumption 2.** For all  $i \in \mathcal{N}$ ,  $\phi_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is a proper closed convex function with a compact domain, i.e., there exists  $D_i > 0$  such that  $\|x_i\| \leq D_i$  for all  $x_i \in \mathbf{dom} \phi_i$ , and  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex function that is differentiable on an open set containing  $\mathbf{dom} \phi_i$ . Suppose there exists a constant  $L_{f_i} > 0$  such that the gradient  $\nabla f_i$  satisfies

$$\|\nabla f_i(x) - \nabla f_i(\bar{x})\| \leq L_{f_i} \|x - \bar{x}\|, \quad \forall x, \bar{x} \in \mathbf{dom} \phi_i.$$

**Assumption 3.** For all  $i \in \mathcal{N}$ ,  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}^{m_i}$  is a  $\mathcal{K}_i$ -convex function that is differentiable on an open set containing  $\mathbf{dom} \phi_i$ . Suppose there exists a constant  $L_{g_i} \geq 0$  such that the Jacobian  $\mathbf{J}g_i : \mathbb{R}^n \rightarrow \mathbb{R}^{m_i \times n}$  satisfies

$$\|\mathbf{J}g_i(x) - \mathbf{J}g_i(\bar{x})\| \leq L_{g_i} \|x - \bar{x}\|, \quad \forall x, \bar{x} \in \mathbf{dom} \phi_i.$$

Furthermore, we assume that  $g_i$  is Lipschitz on  $\mathbf{dom} \phi_i$ ; hence, there exists  $C_{g_i} > 0$  such that  $\|\mathbf{J}g_i(x)\| \leq C_{g_i}$  for all  $x \in \mathbf{dom} \phi_i$ . Let  $L_G \triangleq \max_{i \in \mathcal{N}} \{L_{g_i}\}$  and  $C_G \triangleq \max_{i \in \mathcal{N}} \{C_{g_i}\}$ .

**Assumption 4.** A primal-dual optimal pair  $(x^*, \theta^*) \in \mathbb{R}^n \times \mathcal{K}^*$  for (1.1) exists, where  $\theta_i^* \in \mathcal{K}_i^*$  denotes an optimal dual variable corresponding to  $-g_i(x) \in \mathcal{K}_i$  constraint for  $i \in \mathcal{N}$ . For  $i \in \mathcal{N}$  such that  $g_i(\cdot)$  is not affine, we assume that agent- $i$  knows a bound  $B_i$  such that  $2\|\theta_i^*\| \leq B_i$ .

Assumption 2 is standard in the analysis of first-order algorithms for distributed optimization, and has been adopted in many works (e.g., [21, 41, 47]). Assumption 3 has also commonly appeared in the literature related to constrained optimization problems with functional constraints [6, 19, 20]. If the  $\mathbf{dom} \phi$  is bounded and the Jacobian is continuous on the closure of the domain, it will be satisfied. Assumption 4 requires existence of a primal-dual solution to (1.1),

which is guaranteed to hold under some mild regularity conditions, e.g., whenever (1.1) admits a Slater point. Moreover, when  $g_i(\cdot)$  is affine, i.e., when  $g_i(x) = A_i x + b_i$  for some  $A_i \in \mathbb{R}^{m_i \times n}$  and  $b_i \in \mathbb{R}^{m_i}$ , the knowledge of a dual bound  $B_i$  for  $-g_i(x) \in \mathcal{K}_i$  is not required. It is essential to emphasize that due to conic structure of the constraints, one can still consider a rich class of nonlinear constraints even when  $g_i$  is affine, one can model convex quadratic inequality constraints, SOCP and SDP constraints. Furthermore, in the following remark, we discuss how the agents can compute dual bounds  $B_i$  (as defined in Assumption 4) for  $\theta_i^*$  corresponding to the constraints with nonlinear  $g_i$  in the distributed computation setting we assume in this paper.

**Remark 2.2.** Let  $\mathcal{X} = \mathbb{R}_+^N$  and  $\phi_i(x) = \mathbb{1}_X(x)$  for  $i \in \mathcal{N}$ , where  $X = \{x \in \mathbb{R}^n : \|x\| \leq r\}$ . Consider (P) :  $\varphi^* = \min_x \{\sum_{i \in \mathcal{N}} f_i(x) : g_i(x) \leq 0, i \in \mathcal{N}, x \in X\}$ . In this remark, we discuss how one can compute a Slater point for (P) in a distributed manner.

Given some small  $\varepsilon \in (0, r)$ , let  $X^\circ \triangleq \{x \in \mathbb{R}^n : \|x\| \leq r - \varepsilon\}$ . Consider the Phase I problem:

$$\begin{aligned} \text{(Phase I) : } & (x^\circ, t^\circ) \in \underset{x, t}{\mathbf{argmin}} \mathbb{1}_{X^\circ}(x) + t \\ & \text{s.t. } g_i(x) \leq t : \theta_i, \quad \forall i \in \mathcal{N}. \end{aligned}$$

Note that  $\bar{x} = 0_n$  and  $\bar{t} = \bar{g} + 1$  is a Slater point for (Phase I) where  $\bar{g} \triangleq \max_{i \in \mathcal{N}} g_i(\bar{x})$ . Moreover, let  $\underline{g}_i = \inf\{g_i(x) : x \in X\}$  for  $i \in \mathcal{N}$ , and  $\underline{g} \triangleq \max_{i \in \mathcal{N}} \underline{g}_i$ . Let  $\mathcal{L}(x, t, \theta)$  denote the Lagrangian for (Phase I) and  $\theta^\circ = [\theta_i^\circ]_{i \in \mathcal{N}} \in \mathbb{R}_+^N$  be an optimal dual solution for (Phase I). Clearly, we have  $\underline{g} \leq t^\circ = \inf_{x, t} \mathcal{L}(x, t, \theta^\circ) \leq \bar{t} + \sum_{i \in \mathcal{N}} \theta_i^\circ (g_i(\bar{x}) - \bar{t})$ , which implies that

$$\bar{g} - \underline{g} + 1 \geq \sum_{i \in \mathcal{N}} \theta_i^\circ (\bar{t} - g_i(\bar{x})) \geq \sum_{i \in \mathcal{N}} \theta_i^\circ;$$

therefore,  $\theta_i^\circ \geq 0$  such that  $\theta_i^\circ \leq \bar{g} - \underline{g} + 1$  for all  $i \in \mathcal{N}$ . Moreover, by construction  $x^\circ$  is a Slater point for the original problem (P). Note that (Phase I) satisfies the assumptions for D-APDB, and we can employ D-APDB to compute a Slater point for (P), a special case of (1.1), in a distributed manner. Next, using the computed Slater point  $x^\circ$ , the network of agents can compute a bound on  $\theta^*$ , i.e., an optimal dual solution of (P), by employing a global max operation twice. Indeed, let  $\underline{\varphi}$  be a lower bound on  $\varphi^*$  and it is known, e.g., when  $f_i$  is a loss function for  $i \in \mathcal{N}$ , then one can set  $\underline{\varphi} = 0$ . Then, using a similar Lagrangian argument with above, one can show that  $\sum_{i \in \mathcal{N}} \theta_i^* \leq (\sum_{i \in \mathcal{N}} f_i(x^\circ) - \underline{\varphi}) / (-\max_{i \in \mathcal{N}} g_i(x^\circ)) \leq (\underline{\varphi} - N \max_{i \in \mathcal{N}} f_i(x^\circ)) / \max_{i \in \mathcal{N}} g_i(x^\circ)$ .

We next provide two definitions before presenting our algorithm and our main results.

**Definition 2.3.** Under Assumption 4, for each  $i \in \mathcal{N}$ , if  $g_i(\cdot)$  is not affine, let  $\mathcal{B}_i \triangleq \{\theta_i \in \mathbb{R}^{m_i} : \|\theta_i\| \leq B_i\}$ ; otherwise, if  $g_i(\cdot)$  is an affine function, let  $\mathcal{B}_i = \mathbb{R}^{m_i}$ . Define  $\mathcal{B} \triangleq \prod_{i \in \mathcal{N}} \mathcal{B}_i$ .

**Definition 2.4.** For any  $\mathbf{x} \in \mathbf{dom} \phi$ , let  $\mathbf{JG}(\mathbf{x}) \triangleq \mathbf{diag}([\mathbf{J}g_i(x_i)]_{i \in \mathcal{N}}) \in \mathbb{R}^{m \times n|\mathcal{N}|}$ .

**2.2. Main Results.** In this paper, we propose D-APDB, displayed in Algorithm 1, for solving the constrained, composite convex consensus optimization problem in (1.1), and propose D-APDB0, displayed in Algorithm 2, for solving the *unconstrained*<sup>2</sup> version of the composite convex consensus optimization problem. Given the parameters  $\alpha_i^k, \beta_i^k$  and  $\tilde{\alpha}_i^{k+1}, \tilde{\beta}_i^{k+1}, \tilde{\zeta}_i^{k+1}$  as

<sup>2</sup>Here “unconstrained” means  $g_i(\cdot) = 0$  for all  $i \in \mathcal{N}$ . Any node  $i \in \mathcal{N}$  can still have a simple set constraint incorporated in (1.1) through choosing  $\phi_i$  as an indicator function.

set in D-APDB, the backtracking test function  $E_i^k(\cdot, \cdot)$  in Line 18 of D-APDB is defined as follows:

$$\begin{aligned} E_i^k(x, \theta) \triangleq & - \left( \frac{1}{\bar{\tau}_i^k} - \eta_i^k(\alpha_i^k + \beta_i^k) - \zeta_i^{k+1} \right) \|x - x_i^k\|^2 - \frac{1}{\bar{\alpha}_i^k} \|\theta - \theta_i^k\|^2 \\ & + \frac{2}{\bar{\alpha}_i^{k+1}} \left\| \mathbf{J}g_i(x)^\top (\theta - \theta_i^k) \right\|^2 + \frac{1}{\bar{\beta}_i^{k+1}} \left\| \left( \mathbf{J}g_i(x) - \mathbf{J}g_i(x_i^k) \right)^\top \theta_i^k \right\|^2 + 2\Lambda_i(x), \end{aligned} \quad (2.2)$$

where  $\Lambda_i(x) \triangleq f_i(x) - f_i(x_i^k) - \langle \nabla f_i(x_i^k), x - x_i^k \rangle$ , for all  $i \in \mathcal{N}$  – in case  $L_{g_i} = 0$ , setting  $\beta_i^k = 0$  and  $\bar{\beta}_i^{k+1} = 0$ , we adopt  $0^2/0 = 0$ ; hence, we set  $\left\| \left( \mathbf{J}g_i(x) - \mathbf{J}g_i(x_i^k) \right)^\top \theta_i^k \right\|^2 / \bar{\beta}_i^{k+1} = 0$ .

**Remark 2.5.** From convexity of  $f_i(\cdot)$ , one has  $f_i(x) - f_i(x_i^k) \leq \langle \nabla f_i(x), x - x_i^k \rangle$ ; therefore, one can bound the last term in the test function as follows:  $\Lambda_i(x) \leq \langle \nabla f_i(x) - \nabla f_i(x_i^k), x - x_i^k \rangle$ . In the definition of  $E_i^k(\cdot, \cdot)$  replacing  $\Lambda_i(x)$  with this inner product leads to a stronger condition; that said, in practice, we have found this condition to be numerically more stable.

Now we formally state our main results for D-APDB and D-APDB0.

**Theorem 2.6.** *Suppose that Assumptions 1 to 4 hold, and  $\delta, c_\alpha, c_\beta, c_\zeta > 0$  are given such that  $\delta + c < 1$ , where  $c \triangleq c_\alpha + c_\beta + c_\zeta$ . Let  $(x^*, \theta^*) \in \mathbb{R}^n \times \mathcal{H}^*$  denote an arbitrary primal-dual optimal pair satisfying Assumption 4 and  $\lambda^* \in \mathbb{R}^{n|\mathcal{E}|}$  be an optimal dual variable corresponding to the consensus constraint  $\mathbf{A}x = 0$  in (2.1). For all  $i \in \mathcal{N}$ , it holds for all step size parameter values  $\bar{\tau}_i, \zeta_i > 0$  that when initialized from arbitrary  $x_i^0 \in \mathbf{dom} \phi_i$  and  $\theta_i^0 = 0_{m_i}$ , the ergodic iterate sequences  $\{(\bar{x}_i^k, \bar{\theta}_i^k)\}_{k \geq 0}$  generated by D-APDB, displayed in Algorithm 1, satisfy*

$$\begin{aligned} \text{(i) Suboptimality:} \quad & \left| \sum_{i \in \mathcal{N}} \varphi_i(\bar{x}_i^K) - \varphi^* \right| = \mathcal{O}\left(\frac{1}{K}\right), \\ \text{(ii) Infeasibility:} \quad & \sum_{i \in \mathcal{N}} \|\theta_i^*\| d_{-\mathcal{X}_i}\left(g_i(\bar{x}_i^K)\right) + \|\lambda^*\| \|\mathbf{A}\bar{\mathbf{x}}^K\| = \mathcal{O}\left(\frac{1}{K}\right), \end{aligned}$$

where  $(\bar{x}_i^K, \bar{\theta}_i^K) = \sum_{k=0}^{K-1} t_k(x_i^k, \theta_i^k) / \sum_{k=0}^{K-1} t_k$  for  $i \in \mathcal{N}$ , and  $\{t_k\}_{k \geq 0} \in \mathbb{R}_{++}$  is defined recursively such that  $t_{k+1} = t_k / \eta^{k+1}$  for  $k \geq 0$  and  $t_0 = 1$ . Moreover, there exists  $(x^*, \theta^*)$  a primal-dual optimal solution to (1.1) such that the actual primal-dual iterate sequence  $\{(x_i^k, \theta_i^k)\}_{k \geq 0}$  converges to  $(x^*, \theta_i^*)$  for all  $i \in \mathcal{N}$ , i.e.,  $\lim_{k \rightarrow \infty} x_i^k = x^*$  and  $\lim_{k \rightarrow \infty} \theta_i^k = \theta_i^*$  for  $i \in \mathcal{N}$ .

*Proof.* The proof of this result is given in Section 6 in three parts. The first part is on establishing the rate result which is stated in more detail with explicit  $\mathcal{O}(1)$  constants – see Corollary 6.7. In the second part, we provide an upper bound on the total number of gradient and projection evaluations required for each  $i \in \mathcal{N}$  to collectively compute an  $\varepsilon$ -optimal solution – see Corollary 6.8. Finally, in Theorem 6.9, we show that the iterate sequence converges to a primal-dual optimal solution of to (1.1).  $\square$

**Theorem 2.7.** *Consider the consensus optimization problem  $\varphi^* \triangleq \min_{x \in \mathbb{R}^n} \sum_{i \in \mathcal{N}} \varphi_i(x)$ , where  $\varphi_i(x) \triangleq \phi_i(x) + f_i(x)$  for  $x \in \mathbf{dom} \phi_i$ . Suppose Assumptions 1 to 3 hold, and  $\delta, c_\alpha, c_\zeta > 0$  are given such that  $\delta + c < 1$ , where  $c = c_\alpha + c_\beta + c_\zeta$  with  $c_\beta = 0$ . For all  $i \in \mathcal{N}$ , it holds for all step size parameter value  $\bar{\tau}_i > 0$  that when initialized from arbitrary  $x_i^0 \in \mathbf{dom} \phi_i$ , the ergodic*

iterate sequence  $\{\bar{x}_i^k\}_{k \geq 0}$  generated by D-APDBO, displayed in Algorithm 2, satisfy

$$\left| \sum_{i \in \mathcal{N}} \varphi_i(\bar{x}_i^K) - \varphi^* \right| = \mathcal{O}\left(\frac{1}{K}\right), \quad \|\mathbf{A}\bar{\mathbf{x}}^K\| = \mathcal{O}\left(\frac{1}{K}\right),$$

where  $\bar{x}_i^K = \sum_{k=0}^{K-1} t_k x_i^k / \sum_{k=0}^{K-1} t_k$  for  $i \in \mathcal{N}$ , and  $\{t_k\}_{k \geq 0} \in \mathbb{R}_{++}$  is defined recursively such that  $t_{k+1} = t_k / \eta^{k+1}$  for  $k \geq 0$  and  $t_0 = 1$ . Moreover, there exists an optimal solution  $x^*$  such that the actual iterate sequence  $\{x_i^k\}_{k \geq 0}$  converges to  $x^*$  for all  $i \in \mathcal{N}$ , i.e.,  $\lim_{k \rightarrow \infty} x_i^k = x^*$  for  $i \in \mathcal{N}$ .

*Proof.* This result follows from the proof of Theorem 2.6 for the case  $g_i(\cdot) = 0$  for  $i \in \mathcal{N}$ .  $\square$

### 3. MOTIVATION AND DESIGN OF THE PROPOSED ALGORITHMS: D-APDB AND D-APDBO

Under Assumption 4, the constrained convex optimization problem in (2.1) can be equivalently written as a minimax problem through the use of Lagrangian duality:

$$\min_{\mathbf{x}} \max_{\theta, \lambda} \mathcal{L}(\mathbf{x}, \theta, \lambda) \triangleq \varphi(\mathbf{x}) + \langle \theta, G(\mathbf{x}) \rangle + \langle \lambda, \mathbf{A}\mathbf{x} \rangle - h(\theta), \quad (3.1)$$

where  $h: \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$  denotes the indicator function of  $\mathcal{K}^* \cap \mathcal{B}$ , i.e.,  $h(\theta) = 0$  if  $\theta \in \mathcal{K}^* \cap \mathcal{B}$ , and  $+\infty$  otherwise, i.e.,  $h(\theta) = \sum_{i \in \mathcal{N}} \mathbb{1}_{\mathcal{K}_i^* \cap \mathcal{B}_i}(\theta_i)$ .

**3.1. Preliminaries.** We briefly present a related previous work [20], where we proposed an accelerated primal-dual (APD) algorithm for solving convex-concave saddle-point (SP) problems. As it is shown in [20], APD can be viewed as an extension of the primal-dual algorithm proposed in [11] from bilinear SP problems to a more general setting with a non-bilinear coupling term. Let  $\mathcal{X} \subset \mathbb{R}^{n_x}$  and  $\mathcal{Y} \subset \mathbb{R}^{n_y}$  be finite-dimensional vector spaces. Here we present a slightly modified version of APD proposed in [19] to solve the following problem:

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}(\mathbf{x}, \mathbf{y}) \triangleq (\psi_x + q_x)(\mathbf{x}) + \mathcal{H}(\mathbf{x}, \mathbf{y}) - (\psi_y + q_y)(\mathbf{y}), \quad (3.2)$$

where  $\psi_x, \psi_y$  are possibly non-smooth convex functions, and  $q_x, q_y$  are convex and differentiable on open sets containing  $\text{dom } \psi_x$  and  $\text{dom } \psi_y$  satisfying a descent property governed by  $L_{q_x} \in \mathbb{R}^{n_x \times n_x}$  and  $L_{q_y} \in \mathbb{R}^{n_y \times n_y}$ , i.e.,  $q_x(\mathbf{x}') \leq q_x(\mathbf{x}) + \langle \nabla q_x(\mathbf{x}), \mathbf{x}' - \mathbf{x} \rangle + \frac{1}{2}(\mathbf{x}' - \mathbf{x})^\top L_{q_x}(\mathbf{x}' - \mathbf{x})$  for all  $\mathbf{x}', \mathbf{x} \in \text{dom } \psi_x$ , and similar inequality also holds for  $q_y$ . The coupling  $\mathcal{H}: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  is a continuously differentiable function that is convex in  $\mathbf{x}$  for any  $\mathbf{y} \in \mathcal{Y}$ , and concave in  $\mathbf{y}$  for any  $\mathbf{x} \in \mathcal{X}$ . Moreover, for any  $\mathbf{x} \in \mathcal{X}$ ,  $\nabla_{\mathbf{y}} \mathcal{H}(\mathbf{x}, \cdot)$  is  $L_{yy}$ -Lipschitz for some  $L_{yy} \in \mathbb{R}_+$  and  $\nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}, \cdot)$  is  $\mathbb{R}_+^{p_{xy} \times n_y} \ni L_{xy}$ -Lipschitz for some  $p_{xy} \geq 1$ , i.e.,  $\|\nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}, \mathbf{y}') - \nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}, \mathbf{y})\| \leq \|L_{xy}(\mathbf{y}' - \mathbf{y})\|$  for all  $\mathbf{y}, \mathbf{y}' \in \mathcal{Y}$ ; for any  $\mathbf{y} \in \mathcal{Y}$ ,  $\nabla_{\mathbf{x}} \mathcal{H}(\cdot, \mathbf{y})$  is  $L_{xx}$ -Lipschitz for some  $L_{xx} \in \mathbb{R}_+$  and  $\nabla_{\mathbf{y}} \mathcal{H}(\cdot, \mathbf{y})$  is  $\mathbb{R}_+^{p_{yx} \times n_x} \ni L_{yx}$ -Lipschitz for some  $p_{yx} \geq 1$ . Given  $Q_x, Q_y \in \mathbb{S}_{++}^{n_x} \times \mathbb{S}_{++}^{n_y}$  and initial iterates  $\mathbf{x}^0 \in \mathcal{X}$  and  $\mathbf{y}^0 \in \mathcal{Y}$ , the slightly modified version of APD iterations consist of the following updates:

$$\mathbf{p}^k \leftarrow 2 \nabla_{\mathbf{y}} \mathcal{H}(\mathbf{x}^k, \mathbf{y}^k) - \nabla_{\mathbf{y}} \mathcal{H}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1}), \quad (3.3a)$$

$$\mathbf{y}^{k+1} \leftarrow \arg \min_{\mathbf{y} \in \mathcal{Y}} \psi_y(\mathbf{y}) + \langle \nabla q_y(\mathbf{y}^k) - \mathbf{p}^k, \mathbf{y} \rangle + \frac{1}{2} \|\mathbf{y} - \mathbf{y}^k\|_{Q_y}^2, \quad (3.3b)$$

$$\mathbf{x}^{k+1} \leftarrow \arg \min_{\mathbf{x} \in \mathcal{X}} \psi_x(\mathbf{x}) + \langle \nabla q_x(\mathbf{x}^k) + \nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}^k, \mathbf{y}^{k+1}), \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}^k\|_{Q_x}^2, \quad (3.3c)$$

**Algorithm 1** Distributed APD with Backtracking (D-APDB)

---

1: **Inputs:** constants  $\delta, c_\alpha, c_\beta, c_\zeta, c_\gamma > 0$ :  $c_\alpha + c_\beta + c_\zeta < 1 - \delta$ ,  $c_\gamma \leq 1/(2|\mathcal{E}|)$ ,  $\rho \in (0, 1)$

2: **Inputs for each**  $i \in \mathcal{N}$ :  
     initial states  $x_i^0 \in \text{dom } \phi_i, \theta_i^0 \in \mathcal{X}_i^*$ ; step size parameters  $\bar{\tau}_i, \zeta_i > 0$ ; dual bound  $B_i$  if  $L_{g_i} > 0$

3:  $(x_i^{-1}, \theta_i^{-1}) \leftarrow (x_i^0, \theta_i^0)$ ,  $s_i^0 \leftarrow 0 \quad \forall i \in \mathcal{N}$

4:  $\tau_i^{-1} \leftarrow \bar{\tau}_i$ ,  $\sigma_i^{-1} \leftarrow \zeta_i \bar{\tau}_i$ ,  $\tau_i^0 \leftarrow \bar{\tau}_i$ ,  $\sigma_i^0 \leftarrow \zeta_i \bar{\tau}_i \quad \forall i \in \mathcal{N}$

5:  $\alpha_i^0 \leftarrow c_\alpha / \bar{\tau}_i$ ,  $\beta_i^0 \leftarrow c_\beta / \bar{\tau}_i$ ,  $\zeta_i^0 \leftarrow c_\zeta / \bar{\tau}_i \quad \forall i \in \mathcal{N}$   $\triangleright \beta_i^0 = 0$  if  $L_{g_i} = 0$

6:  $r_i^0 \leftarrow \mathbf{J}g_i(x_i^0)^\top \theta_i^0 + \sum_{j \in \mathcal{N}_i}(s_i^0 - s_j^0)$ ,  $r_i^{-1} \leftarrow r_i^0$ ,  $\forall i \in \mathcal{N}$

7:  $\bar{\tau} \leftarrow \max_{i \in \mathcal{N}} \{\bar{\tau}_i\}$

8: **for**  $k = 0, 1, 2, \dots$  **do**

9:      $\eta^k \leftarrow 1$

10:    **for all**  $i \in \mathcal{N}$  **do**

11:        $\tilde{\tau}_i^k \leftarrow \tau_i^{k-1}$

12:       **loop**  $\triangleright$  Backtracking Loop for  $i \in \mathcal{N}$

13:           $\tilde{\sigma}_i^k \leftarrow \zeta_i \tilde{\tau}_i^k$ ,  $\eta_i^k \leftarrow \tau_i^{k-1} / \tilde{\tau}_i^k$

14:           $\tilde{\alpha}_i^{k+1} \leftarrow c_\alpha / \tilde{\tau}_i^k$ ,  $\tilde{\beta}_i^{k+1} \leftarrow c_\beta / \tilde{\tau}_i^k$ ,  $\tilde{\zeta}_i^{k+1} \leftarrow c_\zeta / \tilde{\tau}_i^k$   $\triangleright \tilde{\beta}_i^{k+1} = 0$  if  $L_{g_i} = 0$

15:           $\tilde{p}_i^k \leftarrow r_i^k + \eta_i^k (r_i^k - r_i^{k-1})$

16:           $\tilde{x}_i^{k+1} \leftarrow \text{prox}_{\tilde{\tau}_i^k \phi_i} (x_i^k - \tilde{\tau}_i^k (\nabla f_i(x_i^k) + \tilde{p}_i^k))$

17:           $\tilde{\theta}_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^* \cap \mathcal{B}_i} (\theta_i^k + \tilde{\sigma}_i^k g_i(\tilde{x}_i^{k+1}))$

18:          **if**  $E_i^k(\tilde{x}_i^{k+1}, \tilde{\theta}_i^{k+1}) \leq -\frac{\delta}{\tilde{\tau}_i^k} \|\tilde{x}_i^{k+1} - x_i^k\|^2 - \frac{\delta}{\tilde{\sigma}_i^k} \|\tilde{\theta}_i^{k+1} - \theta_i^k\|^2$  **then**  $\triangleright$  See (2.2) for  $E_i^k(\cdot, \cdot)$

19:             **break**

20:          **else**

21:              $\tilde{\tau}_i^k \leftarrow \rho \tilde{\tau}_i^k$   $\triangleright$  Backtracking for primal step size

22:          **end if**

23:       **end loop**

24:    **end for**

25:      $\eta^k \leftarrow \max_{i \in \mathcal{N}} \eta_i^k$   $\triangleright$  max-consensus step

26:      $\gamma^k \leftarrow \frac{c_\gamma}{\bar{\tau}} (\frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta})^{-1}$

27:    **for all**  $i \in \mathcal{N}$  **do**

28:        $\tau_i^k \leftarrow \tau_i^{k-1} / \eta^k$ ,  $\sigma_i^k \leftarrow \zeta_i \tau_i^k$

29:        $s_i^{k+1} \leftarrow s_i^k + \gamma^k ((1 + \eta^k)x_i^k - \eta^k x_i^{k-1})$

30:        $p_i^k \leftarrow r_i^k + \eta^k (r_i^k - r_i^{k-1})$

31:       **if**  $\eta^k > 1$  **then**  $\triangleright \eta^k > 1$ : At least one node did backtracking

32:           $x_i^{k+1} \leftarrow \text{prox}_{\tau_i^k \phi_i} (x_i^k - \tau_i^k (\nabla f_i(x_i^k) + p_i^k))$

33:           $\theta_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^* \cap \mathcal{B}_i} (\theta_i^k + \sigma_i^k g_i(x_i^{k+1}))$

34:       **else**  $\triangleright \eta^k = 1$ : no node did backtracking

35:           $x_i^{k+1} \leftarrow \tilde{x}_i^{k+1}$

36:           $\theta_i^{k+1} \leftarrow \tilde{\theta}_i^{k+1}$

37:       **end if**

38:        $r_i^{k+1} \leftarrow \mathbf{J}g_i(x_i^{k+1})^\top \theta_i^{k+1} + \sum_{j \in \mathcal{N}_i}(s_i^{k+1} - s_j^{k+1})$

39:        $\alpha_i^{k+1} \leftarrow c_\alpha / \tau_i^k$ ,  $\beta_i^{k+1} \leftarrow c_\beta / \tau_i^k$ ,  $\zeta_i^{k+1} \leftarrow c_\zeta / \tau_i^k$   $\triangleright \beta_i^{k+1} = 0$  if  $L_{g_i} = 0$

40:    **end for**

41:      $k \leftarrow k + 1$

42: **end for**

---

---

**Algorithm 2** Distributed APD with Backtracking (D-APDB0)
 

---

1: **Inputs:** constants  $\delta, c_\alpha, c_\zeta, c_\gamma > 0$ :  $c_\alpha + c_\zeta < 1 - \delta$ ,  $c_\gamma \leq 1/(2|\mathcal{E}|)$ ,  $\rho \in (0, 1)$   
 2: **Inputs for each**  $i \in \mathcal{N}$ :  
     initial state  $x_i^0 \in \text{dom } \phi_i$ ; step size parameter  $\bar{\tau}_i > 0$   
 3:  $x_i^{-1} \leftarrow x_i^0$ ,  $s_i^0 \leftarrow 0$ ,  $\tau_i^{-1} \leftarrow \bar{\tau}_i$ ,  $\tau_i^0 \leftarrow \bar{\tau}_i \quad \forall i \in \mathcal{N}$   
 4:  $r_i^0 \leftarrow \sum_{j \in \mathcal{N}_i} (s_i^0 - s_j^0)$ ,  $r_i^{-1} \leftarrow r_i^0$ ,  $\forall i \in \mathcal{N}$   
 5:  $\bar{\tau} \leftarrow \max_{i \in \mathcal{N}} \{\bar{\tau}_i\}$   
 6: **for**  $k = 0, 1, 2, \dots$  **do**  
 7:      $\eta^k \leftarrow 1$   
 8:     **for all**  $i \in \mathcal{N}$  **do**  
 9:          $\tilde{\tau}_i^k \leftarrow \tau_i^{k-1}$   
 10:         **loop** ▷ Backtracking Loop for  $i \in \mathcal{N}$   
 11:              $\eta_i^k \leftarrow \tau_i^{k-1} / \tilde{\tau}_i^k$   
 12:              $\tilde{p}_i^k \leftarrow r_i^k + \eta_i^k (r_i^k - r_i^{k-1})$   
 13:              $\tilde{x}_i^{k+1} \leftarrow \text{prox}_{\tilde{\tau}_i^k \phi_i} (x_i^k - \tilde{\tau}_i^k (\nabla f_i(x_i^k) + \tilde{p}_i^k))$   
 14:             **if**  $f_i(\tilde{x}_i^{k+1}) - f_i(x_i^k) - \langle \nabla f_i(x_i^k), \tilde{x}_i^{k+1} - x_i^k \rangle \leq \frac{1}{2\tilde{\tau}_i^k} (1 - \delta - c_\alpha - c_\zeta) \|\tilde{x}_i^{k+1} - x_i^k\|^2$  **then**  
 15:                 **break**  
 16:             **else**  
 17:                  $\tilde{\tau}_i^k \leftarrow \rho \tilde{\tau}_i^k$  ▷ Backtracking for primal step size  
 18:             **end if**  
 19:         **end loop**  
 20:         **end for**  
 21:          $\eta^k \leftarrow \max_{i \in \mathcal{N}} \eta_i^k$  ▷ max-consensus step  
 22:          $\gamma^k \leftarrow \frac{c_\gamma}{\bar{\tau}} (\frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta})^{-1}$   
 23:         **for all**  $i \in \mathcal{N}$  **do**  
 24:              $\tau_i^k \leftarrow \tau_i^{k-1} / \eta^k$   
 25:              $s_i^{k+1} \leftarrow s_i^k + \gamma^k ((1 + \eta^k)x_i^k - \eta^k x_i^{k-1})$   
 26:              $p_i^k \leftarrow r_i^k + \eta^k (r_i^k - r_i^{k-1})$   
 27:             **if**  $\eta^k > 1$  **then** ▷  $\eta^k > 1$ : At least one node did backtracking  
 28:                  $x_i^{k+1} \leftarrow \text{prox}_{\tau_i^k \phi_i} (x_i^k - \tau_i^k (\nabla f_i(x_i^k) + p_i^k))$  ▷  $\eta^k = 1$ : no node did backtracking  
 29:             **else**  
 30:                  $x_i^{k+1} \leftarrow \tilde{x}_i^{k+1}$   
 31:             **end if**  
 32:              $r_i^{k+1} \leftarrow \sum_{j \in \mathcal{N}_i} (s_i^{k+1} - s_j^{k+1})$   
 33:         **end for**  
 34:          $k \leftarrow k + 1$   
 35: **end for**

---

where  $\|\mathbf{x}\|_{Q_x} = \sqrt{\mathbf{x}^\top Q_x \mathbf{x}}$  for  $\mathbf{x} \in \mathcal{X}$  and  $\|\mathbf{y}\|_{Q_y} = \sqrt{\mathbf{y}^\top Q_y \mathbf{y}}$  for  $\mathbf{y} \in \mathcal{Y}$ . Based on the discussion in [20], if  $Q_x$  and  $Q_y$  are chosen such that there exist some  $c \geq 1$ , and  $\alpha, \beta \geq 0$  satisfying

$$Q_x - L_{q_x} \succeq L_{xx} I_{n_x} + \frac{1}{\alpha} L_{yx}^\top L_{yx}, \quad Q_y - L_{q_y} \succeq c(\alpha + \beta) I_{n_y} + \frac{1}{\beta} L_{yy}^2 I_{n_y}, \quad (3.4)$$

– see [20, Assumption 3] with  $\delta = 0$ , then according to [20, Eq. (3.3)] and [20, Lemma 3.4], it holds for all  $K \geq 0$  that

$$0 \leq \mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K) \leq \frac{1}{2K} \left( \|\mathbf{x} - \mathbf{x}^0\|_{Q_x}^2 + \|\mathbf{y} - \mathbf{y}^0\|_{Q_y}^2 \right), \quad \forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}, \quad (3.5)$$

where  $(\bar{\mathbf{x}}^K, \bar{\mathbf{y}}^K) = \frac{1}{K} \sum_{k=1}^K (\mathbf{x}^k, \mathbf{y}^k)$ .

Our objective is to design a momentum-based primal-dual method built on APD framework in (3.3) for the saddle-point formulation in (3.1), which can be seen as a special case of (3.2) with a very particular structure:  $\mathbf{y} = [\boldsymbol{\theta}^\top \boldsymbol{\lambda}^\top]^\top$ ,  $\psi_x(\mathbf{x}) = \phi(\mathbf{x})$ ,  $q_x(\mathbf{x}) = f(\mathbf{x})$ ,  $\mathcal{H}(\mathbf{x}, \mathbf{y}) = \langle \boldsymbol{\theta}, G(\mathbf{x}) \rangle + \langle \boldsymbol{\lambda}, A\mathbf{x} \rangle$ ,  $\psi_y(\mathbf{y}) = h(\boldsymbol{\theta})$  and  $q_y(\mathbf{y}) = 0$ . Therefore, we will focus on a variant of the primal-dual iterations stated in (3.3) with two goals in mind: **(i)** we prefer using momentum acceleration for the primal updates rather than having momentum term in dual updates as in (3.3); **(ii)** one should be able to adaptively select node-specific step sizes via adopting a backtracking scheme, without relying on some prior knowledge on global Lipschitz constants.

The first goal is mainly motivated by the primal-dual dynamics for the particular saddle point formulation we focus on. Indeed,  $\mathcal{H}(\mathbf{x}, \mathbf{y})$  has a curvature in  $\mathbf{x}$ , while it is affine in  $\mathbf{y}$ , and that is why using momentum in  $\mathbf{x}$ -updates helps dampening the oscillatory behavior that naturally arises due to primal-descent-dual-ascent-type updates, e.g., see [48, 56] for the oscillatory behavior of gradient-descent-ascent updates; on the other hand, the use of momentum in  $\mathbf{y}$ -update as in (3.3) may even exacerbate the oscillations as  $\nabla_{\mathbf{y}} \mathcal{H}(\mathbf{x}, \mathbf{y})$  only depends on  $\mathbf{x}$ , i.e., if  $\{\mathbf{x}^k\}$  oscillates, not only there is no damping mechanism for  $\mathbf{x}$ -updates in (3.3), but oscillations in  $\{\mathbf{x}^k\}$  also causes  $\{\mathbf{y}^k\}$  to oscillate even more.

The first goal requires a change in the update order of (3.3), which can be justified assuming (3.2) has a saddle point. Indeed, applying the extended version of APD given in (3.3) on an equivalent problem  $\min_{\mathbf{y} \in \mathcal{Y}} \max_{\mathbf{x} \in \mathcal{X}} -\mathcal{L}(\mathbf{x}, \mathbf{y})$  with  $q_y(\cdot) = 0$  would lead to the following iterations with the desired momentum acceleration on the primal updates:

$$\mathbf{p}^k \leftarrow (1 + \eta^k) \nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}^k, \mathbf{y}^k) - \eta^k \nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1}), \quad (3.6a)$$

$$\mathbf{x}^{k+1} \leftarrow \arg \min_{\mathbf{x} \in \mathcal{X}} \psi_x(\mathbf{x}) + \langle \nabla q_x(\mathbf{x}^k) + \mathbf{p}^k, \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{x}^k\|_{Q_x^k}^2, \quad (3.6b)$$

$$\mathbf{y}^{k+1} \leftarrow \arg \min_{\mathbf{y} \in \mathcal{Y}} \psi_y(\mathbf{y}) - \langle \nabla_{\mathbf{y}} \mathcal{H}(\mathbf{x}^{k+1}, \mathbf{y}^k), \mathbf{y} \rangle + \frac{1}{2} \|\mathbf{y} - \mathbf{y}^k\|_{Q_y^k}^2, \quad (3.6c)$$

with  $Q_x^k = Q_x$ ,  $Q_y^k = Q_y$ , and  $\eta^k = 1$  for all  $k \geq 0$ . A convergence rate result similar to (3.5) continues to hold for this variant with  $Q_x$  and  $Q_y$  chosen slightly different than (3.4) due to primal-dual switch:

$$Q_y \succeq L_{yy} I_{n_y} + \frac{1}{\alpha} L_{xy}^\top L_{xy}, \quad Q_x - L_{q_x} \succeq c(\alpha + \beta) I_{n_x} + \frac{1}{\beta} L_{xx}^2 I_{n_x},$$

for some  $c \geq 1$  and  $\alpha, \beta \geq 0$ . Our motivation for giving (3.6) with iteration dependent step-size matrices  $Q_x^k, Q_y^k$  and momentum parameter  $\eta^k$  is to set the grounds for a setting that will allow us to analyze a variant of APD method employing node-specific local step-size search. In the rest, we use the notation  $\mathbf{x}^k = [x_i^k]_{i \in \mathcal{N}}$  for  $k \geq 0$ .

Towards the second goal stated above, i.e., to ensure that the nodes can adaptively select their primal and dual step sizes at each iteration through backtracking, we still need to slightly modify the updates given in (3.6) since the  $\boldsymbol{\lambda}$ -update in (3.6), related to the dual variable vector

$\lambda$  corresponding to the consensus constraint  $\mathbf{A}\mathbf{x} = 0$ , is *not* suitable for a distributed local step size search mechanism. Indeed, given some  $\tau_i^k, \sigma_i^k > 0$  for  $i \in \mathcal{N}$  and  $\gamma^k > 0$  for  $k \geq 0$ , let

$$\mathcal{Q}_x^k = \mathbf{D}_{\tau^k}, \quad \mathcal{Q}_y^k = \begin{pmatrix} \mathbf{D}_{\sigma^k} & 0 \\ 0 & \frac{1}{\gamma^k} \mathbf{I}_{n|\mathcal{E}|} \end{pmatrix},$$

where  $\mathbf{D}_{\tau^k} = \mathbf{diag} \left( \left[ \frac{1}{\tau_i^k} \mathbf{I}_n \right]_{i \in \mathcal{N}} \right)$ ,  $\mathbf{D}_{\sigma^k} = \mathbf{diag} \left( \left[ \frac{1}{\sigma_i^k} \mathbf{I}_{m_i} \right]_{i \in \mathcal{N}} \right)$ . For any iteration  $k \geq 0$ , after a candidate point  $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$  for the next iteration is computed according to (3.6) using some given step sizes  $\{\tau_i^k, \sigma_i^k\}_{i \in \mathcal{N}}$  and  $\gamma^k$ , one needs to check a backtracking condition to ensure that  $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$  provides a sufficient decrease in a suitably defined potential function, where  $\mathbf{y}^{k+1} = (\theta^{k+1}, \lambda^{k+1})$  –if the condition holds, we accept  $(\mathbf{x}^{k+1}, \mathbf{y}^{k+1})$ ; otherwise, we decrease the given step sizes and compute a new candidate point. In this procedure, one needs to compute a term involving the candidate dual point  $\lambda^{k+1}$  every time the condition is checked, e.g.,  $\|\lambda^{k+1} - \lambda^k\|^2$ . However, the update rule in (3.6) implies that  $\lambda^{k+1} = \lambda^k + \gamma^k \mathbf{A}\mathbf{x}^{k+1}$ , which is a function of  $\mathbf{x}^{k+1}$  for  $k \geq 0$ ; hence, checking a test function involving  $\lambda^{k+1}$  is not suitable for distributed computation. Indeed, as an example, we consider the quantity  $\|\lambda^{k+1} - \lambda^k\|^2$  which can be computed as below:

$$\begin{aligned} \|\lambda^{k+1} - \lambda^k\|^2 &= (\gamma^k)^2 \langle \mathbf{A}\mathbf{x}^{k+1}, \mathbf{A}\mathbf{x}^{k+1} \rangle = (\gamma^k)^2 \langle \mathbf{x}^{k+1}, (\mathbf{\Omega} \otimes \mathbf{I}_n) \mathbf{x}^{k+1} \rangle \\ &= (\gamma^k)^2 \sum_{i \in \mathcal{N}} \langle x_i^{k+1}, \sum_{j \in \mathcal{N}_i} (x_i^{k+1} - x_j^{k+1}) \rangle. \end{aligned} \quad (3.7)$$

According to (3.7), at iteration  $k \geq 0$ , for each agent  $i \in \mathcal{N}$ , checking its *local* backtracking condition involves computing the term  $\langle x_i^{k+1}, \sum_{j \in \mathcal{N}_i} (x_i^{k+1} - x_j^{k+1}) \rangle$ , which requires the decision vectors  $x_j^{k+1}$  of its neighbors  $j \in \mathcal{N}_i$  that satisfy their own backtracking conditions. Therefore, if one adopts the update scheme in (3.6), it is not trivial or practical to compute  $\{x_i^{k+1}\}_{j \in \mathcal{N}}$  in a distributed manner that satisfy all local backtracking conditions simultaneously. To avoid this issue, we slightly change the update order as stated in (3.8), i.e., at iteration  $k \geq 0$ , consensus related dual variable  $\lambda$  is updated first using the information related to  $\mathbf{x}^k$  and  $\mathbf{x}^{k-1}$ , then the  $\mathbf{x}$ -update is executed, and finally the dual variable  $\theta_i$  related to the local constraint  $-g_i(x_i) \in \mathcal{K}_i$  is updated for all  $i \in \mathcal{N}$ . Hence, for the variant of APD in (3.8), the term  $\|\lambda^{k+1} - \lambda^k\|^2$  only relies on the points  $\{x_i^k\}_{i \in \mathcal{N}}$  that satisfy the local backtracking conditions at iteration  $k-1$ . It should be emphasized that as an artifact of the proof technique we adopted, rather than setting  $\lambda^{k+1} = \lambda^k + \gamma^k \mathbf{A}\mathbf{x}^k$ , we set  $\lambda^{k+1} = \lambda^k + \gamma^k \mathbf{A}((1 + \eta^k)\mathbf{x}^k - \eta^k \mathbf{x}^{k-1})$  for some properly chosen  $\eta^k \geq 1$ , i.e., we also use momentum term in  $\lambda$ -update as well.<sup>3</sup>

Given some arbitrary initial points  $\mathbf{x}^0 \in \mathbf{dom} \phi$ ,  $\theta^0 \in \mathcal{X}^*$ ,  $\lambda^0 \in \mathbb{R}^{n|\mathcal{E}|}$ , step sizes  $\tau_i^k, \sigma_i^k > 0$  for  $i \in \mathcal{N}$ ,  $\gamma^k > 0$  and the momentum parameter  $\eta^k > 0$  for all  $k \geq 0$ , we set  $\mathbf{x}^{-1} = \mathbf{x}^0$ ,  $\theta^{-1} = \theta^0$ ,

<sup>3</sup>Later in Lemma 4.2, we establish how duality gap changes after one iteration of D-APDB. This choice of  $\lambda$ -update leads to a telescoping sum for the duality gap bound in (4.4) by ensuring that  $t_{k+1} \mathcal{Q}^{k+1}(\mathbf{z}) - t_k \mathcal{R}^{k+1}(\mathbf{z}) \leq 0$  holds for all  $k \geq 0$  and  $\mathbf{z} \in \mathcal{X} \times \mathcal{Y}$ .

$\lambda^{-1} = \lambda^0$ , and we consider the following iterations for  $k \geq 0$ :

$$\lambda^{k+1} \leftarrow \underset{\lambda}{\mathbf{argmin}} - \left\langle A \left( (1 + \eta^k) \mathbf{x}^k - \eta^k \mathbf{x}^{k-1} \right), \lambda \right\rangle + \frac{1}{2\gamma^k} \|\lambda - \lambda^k\|^2, \quad (3.8a)$$

$$\mathbf{p}^k \leftarrow A^\top \left( (1 + \eta^k) \lambda^k - \eta^k \lambda^{k-1} \right) + (1 + \eta^k) \mathbf{J}G(\mathbf{x}^k)^\top \boldsymbol{\theta}^k - \eta^k \mathbf{J}G(\mathbf{x}^{k-1})^\top \boldsymbol{\theta}^{k-1}, \quad (3.8b)$$

$$\mathbf{x}^{k+1} \leftarrow \underset{\mathbf{x}}{\mathbf{argmin}} \sum_{i \in \mathcal{N}} \left[ \phi_i(x_i) + \langle \nabla f(x_i^k) + p_i^k, x_i \rangle + \frac{1}{2\tau_i^k} \|x_i - x_i^k\|^2 \right], \quad (3.8c)$$

$$\boldsymbol{\theta}_i^{k+1} \leftarrow \underset{\boldsymbol{\theta}_i \in \mathcal{X}_i^* \cap \mathcal{B}_i}{\mathbf{argmin}} - \langle \boldsymbol{\theta}_i, g_i(x_i^{k+1}) \rangle + \frac{1}{2\sigma_i^k} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^k\|^2, \quad \forall i \in \mathcal{N}. \quad (3.8d)$$

Here, update (3.8a) enforces the consensus constraint  $A\mathbf{x} = 0$  via an ascent step; (3.8b) forms a momentum/correction term using past dual/primal states; (3.8c) performs a proximal gradient-type step on each block  $x_i$  whereas (3.8d) is a projected ascent ensuring  $\boldsymbol{\theta}_i^{k+1} \in \mathcal{X}_i^*$ . At this point, we have an abstract algorithm as described in (3.8) with momentum acceleration in  $\mathbf{x}$ -updates to exploit the curvature of  $G(\mathbf{x})$ , and it is in a suitable form to design a backtracking condition for it. Next, we discuss how this abstract update rule can be implemented in a distributed manner using only local communications for  $n$ -dimensional data vectors.

**3.2. Distributed implementation of the abstract method.** Since we set  $\mathbf{x}^0 = \mathbf{x}^{-1}$  and (3.8a) implies that  $\lambda^{k+1} = \lambda^k + \gamma^k A \left( (1 + \eta^k) \mathbf{x}^k - \eta^k \mathbf{x}^{k-1} \right)$  for all  $k \geq 0$ , we get

$$\lambda^{k+1} = \sum_{\ell=0}^k \gamma^\ell A \left( (1 + \eta^\ell) \mathbf{x}^\ell - \eta^\ell \mathbf{x}^{\ell-1} \right), \quad \forall k \geq 0,$$

and we choose  $\lambda^0 \triangleq 0$ . To implement these steps efficiently using only local communications among the neighboring nodes, we eliminate  $\{\lambda^k\}$  via an auxiliary consensus state sequence  $\{\mathbf{s}^k\}_{k \geq 0}$  such that  $\mathbf{s}^0 \triangleq 0$  and  $\mathbf{s}^{k+1} = \mathbf{s}^k + \gamma^k (1 + \eta^k) \mathbf{x}^k - \gamma^k \eta^k \mathbf{x}^{k-1}$  for all  $k \geq 0$ ; hence,  $\lambda^k = A\mathbf{s}^k$  for all  $k \geq 0$ . Using  $A^\top A = \Omega \otimes \mathbf{I}_n$ , we obtain that

$$\langle A^\top \lambda^k, \mathbf{x} \rangle = \langle \mathbf{x}, (\Omega \otimes \mathbf{I}_n) \mathbf{s}^k \rangle = \sum_{i \in \mathcal{N}} \langle x_i, \sum_{j \in \mathcal{N}_i} (s_i^k - s_j^k) \rangle, \quad \forall \mathbf{x}.$$

Therefore, (3.8) can be implemented in a distributed manner, i.e., each node  $i \in \mathcal{N}$ , starting from arbitrary  $x_i^0 \in \mathbf{dom} \phi_i$  and  $\boldsymbol{\theta}_i^0 \in \mathcal{X}_i^*$ , initializes  $s_i^0 = s_i^{-1} = 0$ ,  $x_i^{-1} = x_i^0$  and  $\boldsymbol{\theta}_i^{-1} = \boldsymbol{\theta}_i^0$ , sets  $r_i^{-1} = \mathbf{J}g_i(x_i^0)^\top \boldsymbol{\theta}_i^0$ , and computes the following updates<sup>4</sup>:

$$s_i^{k+1} \leftarrow s_i^k + \gamma^k \left( (1 + \eta^k) x_i^k - \eta^k x_i^{k-1} \right) \quad (3.9a)$$

$$r_i^k \leftarrow \mathbf{J}g_i(x_i^k)^\top \boldsymbol{\theta}_i^k + \sum_{j \in \mathcal{N}_i} (s_i^k - s_j^k) \quad (3.9b)$$

$$p_i^k \leftarrow r_i^k + \eta^k (r_i^k - r_i^{k-1}) \quad (3.9c)$$

$$x_i^{k+1} \leftarrow \text{prox}_{\tau_i^k \phi_i} \left( x_i^k - \tau_i^k (\nabla f_i(x_i^k) + p_i^k) \right) \quad (3.9d)$$

$$\boldsymbol{\theta}_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{X}_i^* \cap \mathcal{B}_i} \left( \boldsymbol{\theta}_i^k + \sigma_i^k g_i(x_i^{k+1}) \right), \quad (3.9e)$$

<sup>4</sup>Our initialization implies that  $p_i^0 = r_i^0 = \mathbf{J}g_i(x_i^0)^\top \boldsymbol{\theta}_i^0$  for all  $i \in \mathcal{N}$ .

for all  $k \geq 0$ , where the proximal map,  $\text{prox}_{\tau\phi_i}(\cdot)$ , is defined for any  $\tau > 0$  and  $x \in \mathbb{R}^n$  as follows:

$$\text{prox}_{\tau\phi_i}(x) \triangleq \arg \min_{w \in \mathbb{R}^n} \left\{ \tau\phi_i(w) + \frac{1}{2}\|w - x\|^2 \right\}.$$

Thus each node  $i \in \mathcal{N}$  updates its local variables  $(x_i, \theta_i, s_i)$  using only local gradients/Jacobians and communicating only with the neighboring nodes in  $\mathcal{N}_i$ .

The accelerated primal-dual algorithm with backtracking (ADPB) introduced in [20] is a version of ADP that supports backtracking; however it was designed for centralized minimax problems. Our goal in this paper is to design a *distributed* stepsize selection mechanism for the abstract method given in (3.9) so that we can extend ADPB [20] from centralized to the decentralized setting. The *distributed* stepsize selection procedure we propose is based on local backtracking and it does not require a priori knowledge of global Lipschitz constants.

**3.3. From the abstract method towards D-APDB in Algorithm 1.** Here we propose a step-size choice mechanism for the abstract distributed update scheme in (3.9), which would lead to D-APDB displayed in Algorithm 1. Our goal in each iteration  $k \geq 0$  is to employ an Armijo-type local backtracking condition for determining local step-sizes  $\tau_i^k, \sigma_i^k > 0$  for each  $i \in \mathcal{N}$ , and to use a global-max consensus across the network to coordinate the momentum parameter  $\eta^k > 0$ .

At each iteration  $k \geq 0$ , every node  $i \in \mathcal{N}$  calls a backtracking subroutine to compute candidate step sizes and a momentum parameter:  $(\tilde{\tau}_i^k, \tilde{\sigma}_i^k, \eta_i^k)$ . Within the backtracking subroutine, each node  $i \in \mathcal{N}$ , employs Armijo-type step-size search to determine  $\tilde{\tau}_i^k$  using only local information –for the sake of notational simplicity, in the coming argument we do not explicitly show the iteration index  $k \geq 0$ . For each node  $i \in \mathcal{N}$ , given the current primal-dual iterate  $(x_i, \theta_i)$ , and the gradient terms related to the agent-specific constraint function  $g_i$  and consensus violation for the current and the previous iterations, i.e.,  $r_i$  and  $r_i^-$ , the trial primal stepsize  $\tilde{\tau}_i$  is initialized using the step size of the previous iteration  $\tau_i^-$ , i.e.,  $\tilde{\tau}_i \leftarrow \tau_i^-$ , and node- $i$  computes the candidate primal-dual iterate  $(\tilde{x}_i, \tilde{\theta}_i)$  as follows:

$$\begin{aligned} \tilde{p}_i &\leftarrow r_i + \frac{\tau_i^-}{\tilde{\tau}_i}(r_i - r_i^-), \\ \tilde{x}_i &\leftarrow \text{prox}_{\tilde{\tau}_i\phi_i}\left(x_i - \tilde{\tau}_i(\nabla f_i(x_i) + \tilde{p}_i)\right), \\ \tilde{\theta}_i &\leftarrow \mathcal{P}_{K_i^* \cap \mathcal{B}_i}(\theta_i + \tilde{\sigma}_i g_i(\tilde{x}_i)), \end{aligned}$$

where  $\rho \in (0, 1)$  is a fixed contraction factor,  $\delta \in (0, 1)$  is a backtracking parameter, and  $\tilde{p}_i$  denotes a local momentum term mimicking the update in (3.9c). We adopted an Armijo-type backtracking condition based on a local merit function  $E_i$  as defined in (2.2), where for given trial primal stepsize  $\tilde{\tau}_i$ , the parameters within  $E_i$  depend only on  $\tilde{\tau}_i$  and  $\tau_i^-$ , i.e.,

$$\tilde{\alpha}_i^{k+1} = \frac{c_\alpha}{\tilde{\tau}_i^k}, \quad \tilde{\beta}_i^{k+1} = \frac{c_\beta}{\tilde{\tau}_i^k}, \quad \tilde{\zeta}_i^{k+1} = \frac{c_\zeta}{\tilde{\tau}_i^k}, \quad \tilde{\sigma}_i^k = \zeta_i \tilde{\tau}_i^k, \quad \eta_i^k = \frac{\tau_i^{k-1}}{\tilde{\tau}_i^k},$$

are all derived from  $\tilde{\tau}_i^k$  (with fixed  $c_\alpha, c_\beta, c_\zeta > 0$  and  $\zeta_i > 0$ ), while we set  $\alpha_i^k = c_\alpha/\tau_i^{k-1}$  and  $\beta_i = c_\beta/\tau_i^{k-1}$ . The merit function  $E_i$  is a quadratic surrogate tailored to the primal-dual structure, and the trial pair  $(\tilde{\tau}_i, \tilde{\sigma}_i)$  is *accepted* if the condition,

$$E_i(\tilde{x}_i, \tilde{\theta}_i) \leq -\frac{\delta}{2\tilde{\tau}_i}\|\tilde{x}_i - x_i\|^2 - \frac{\delta}{2\tilde{\sigma}_i}\|\tilde{\theta}_i - \theta_i\|^2, \tag{3.10}$$

holds; otherwise, both step sizes are shrunk, i.e.,  $\tilde{\tau}_i \leftarrow \rho \tilde{\tau}_i$ ,  $\tilde{\sigma}_i \leftarrow \rho \tilde{\sigma}_i$ , and the loop repeats until (3.10) holds eventually. Upon the acceptance of the trial pair  $(\tilde{\tau}_i, \tilde{\sigma}_i)$ , the routine returns the current step  $\tilde{\tau}_i$  together with  $\eta_i = \frac{\tau_i^-}{\tilde{\tau}_i}$ . Condition (3.10) enforces a decrease proportional to the “step energy”  $\|\tilde{x}_i - x_i\|^2$  and  $\|\tilde{\theta}_i - \theta_i\|^2$  in a model that captures the curvature of  $f_i$ , and the local constraint sensitivity via  $\mathbf{J}g_i$ . As  $\tilde{\sigma}_i$  and  $\tilde{\tau}_i$  are coupled through some  $\zeta_i > 0$ , the primal and dual step sizes always remain properly scaled. More precisely, once  $\tilde{\tau}_i^k$  is accepted, node- $i$  sets  $\tilde{\sigma}_i^k = \zeta_i \tilde{\tau}_i^k$  and  $\eta_i^k = \frac{\tau_i^{k-1}}{\tilde{\tau}_i^k}$ ; and next, D-APDB sets  $\eta^k$  via

$$\eta^k = \max_{i \in \mathcal{N}} \eta_i^k.$$

This step requires a *global max consensus*, i.e., the cooperation of the nodes to compute a maximum over the node variables  $\eta_i^k$ . It can be typically implemented with protocols such as LoRaWAN (low power, long range) which supports long-distance low-power wireless communication between the nodes [14, 25]. Then, D-APDB derives  $\tau_i^k = \tau_i^{k-1}/\eta^k$ ,  $\alpha_i^{k+1} = c_\alpha/\tau_i^k$ ,  $\beta_i^{k+1} = c_\beta/\tau_i^k$ ,  $\zeta_i^{k+1} = c_\zeta/\tau_i^k$ ,  $\sigma_i^{k+1} = \zeta_i \tau_i^k$ , and chooses  $\gamma_k$  via

$$\gamma_k = \frac{c_\gamma}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta} \right)^{-1},$$

with  $\bar{\tau} \triangleq \max_{i \in \mathcal{N}} \{\tilde{\tau}_i\}$  and for some  $c_\gamma > 0$  such that  $c_\gamma \leq 1/(2|\mathcal{E}|)$ , to update the  $s_i^k$  variable. With this information at hand, the subsequent updates consist of updating the node-based  $p_i^k$ ,  $r_i^k$  and  $x_i^k$  variables as outlined in Algorithm 1. Basically, these updates correspond to a fully decentralized implementation of (3.9) –more precisely, Algorithm 1 describes how primal-dual stepsize sequence  $\{\tau_i^k, \sigma_i^k\}_{k \geq 0}$  for  $i \in \mathcal{N}$  and the momentum parameter sequence  $\{\eta^k\}_{k \geq 0}$  defining the update rule in (3.9) can be chosen through employing node-specific backtracking and running the max-consensus protocol across the network once per iteration  $k \geq 0$ .

#### 4. PRELIMINARY TECHNICAL RESULTS

Now, we are ready to give our main technical results. We will derive some key inequalities below for D-APDB iterates  $\{\mathbf{x}_k, \lambda_k, \theta_k\}_{k \geq 0}$  generated by Algorithm 1. Let  $\mathbf{y} = [\theta^\top, \lambda^\top]^\top$  denote the concatenation of the dual variables and  $\Phi : \mathbf{dom} \phi \times \mathcal{X}^* \times \mathbb{R}^{n|\mathcal{E}|} \rightarrow \mathbb{R}$  denote the coupling function:  $\Phi(\mathbf{x}, \theta, \lambda) = \langle G(\mathbf{x}), \theta \rangle + \langle A\mathbf{x}, \lambda \rangle$ . Thus, we have  $\mathcal{L}(\mathbf{x}, \theta, \lambda) = \varphi(\mathbf{x}) + \Phi(\mathbf{x}, \theta, \lambda) - h(\theta)$ , where  $\mathcal{L}$  and  $h(\cdot)$  as defined in (3.1), and  $\varphi(\cdot)$  as defined in (2.1). For  $k \geq 0$ , define

$$\begin{aligned} p^k &\triangleq \nabla_{\mathbf{x}} \Phi(\mathbf{x}^k, \theta^k, \lambda^k) + \eta^k q^k, & q^k &\triangleq \nabla_{\mathbf{x}} \Phi(\mathbf{x}^k, \theta^k, \lambda^k) - \nabla_{\mathbf{x}} \Phi(\mathbf{x}^{k-1}, \theta^{k-1}, \lambda^{k-1}), \\ q^{k,\mathbf{y}} &\triangleq \nabla_{\mathbf{x}} \Phi(\mathbf{x}^k, \theta^k, \lambda^k) - \nabla_{\mathbf{x}} \Phi(\mathbf{x}^k, \theta^{k-1}, \lambda^{k-1}), \\ q^{k,\mathbf{x}} &\triangleq \nabla_{\mathbf{x}} \Phi(\mathbf{x}^k, \theta^{k-1}, \lambda^{k-1}) - \nabla_{\mathbf{x}} \Phi(\mathbf{x}^{k-1}, \theta^{k-1}, \lambda^{k-1}); \end{aligned} \quad (4.1)$$

hence,  $q^k = q^{k,\mathbf{y}} + q^{k,\mathbf{x}}$ . Moreover, for all  $i \in \mathcal{N}$  and  $k \geq 0$ , we similarly define

$$\begin{aligned} q_i^{k,\mathbf{y}} &\triangleq \nabla_{x_i} \Phi(\mathbf{x}^k, \theta^k, \lambda^k) - \nabla_{x_i} \Phi(\mathbf{x}^k, \theta^{k-1}, \lambda^{k-1}), \\ q_i^{k,\mathbf{x}} &\triangleq \nabla_{x_i} \Phi(\mathbf{x}^k, \theta^{k-1}, \lambda^{k-1}) - \nabla_{x_i} \Phi(\mathbf{x}^{k-1}, \theta^{k-1}, \lambda^{k-1}) \end{aligned} \quad (4.2)$$

based on the differences of partial derivatives with respect to  $x_i$ . Given the initial primal-dual point  $\mathbf{x}^0 \in \mathbf{dom} \phi$  and  $\mathbf{y}^0$  such that  $\theta^0 \in \mathcal{X}^* \cap \mathcal{B}$  and  $\lambda^0 = 0$ , we set  $\mathbf{x}^{-1} = \mathbf{x}^0$  and  $\mathbf{y}^{-1} = \mathbf{y}^0$ ;

hence,  $q^0 = \mathbf{0}$ . Moreover, for  $k \geq 0$ , the Cauchy-Schwarz inequality implies that

$$\langle q^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle \leq \sum_{i \in \mathcal{N}} \left( \frac{1}{2\alpha_i^k} \|q_i^{k,y}\|^2 + \frac{1}{2\beta_i^k} \|q_i^{k,x}\|^2 + \frac{\alpha_i^k + \beta_i^k}{2} \|x_i^{k+1} - x_i^k\|^2 \right) \quad (4.3)$$

for any  $\alpha_i^k, \beta_i^k > 0$ . We will use (4.3) later in the proof a few times.

**Remark 4.1.** For  $i \in \mathcal{N}$  such that  $L_{g_i} = 0$ , for all  $k \geq 0$ , one trivially has  $q_i^{k,x} = 0$ ; hence, setting  $\beta_i^k = 0$ , we adopt  $0^2/0 = 0$  and set  $\|q_i^{k,x}\|^2/\beta_i^k = 0$ .

Recall that Lines 29-33 in Algorithm 1 correspond to (3.9), which can be written in a compact form as in (3.8). Therefore, in the next lemma we provide a crucial result that analyzes the effect of one step update in the form of (3.8). The result holds for any positive  $\{\alpha_i^k, \alpha_i^{k+1}, \beta_i^k, \beta_i^{k+1}, \varsigma_i^k, \varsigma_i^{k+1}\}_{i \in \mathcal{N}}$  which are indeed free design parameters of D-APDB. Later in Section 5 we discuss how these parameters should be selected.

**Lemma 4.2 (One Step Result).** *Suppose that Assumptions 1 to 3 hold. Fix  $k \geq 0$  and let  $\mathbf{x}^k, \mathbf{x}^{k-1} \in \mathbf{dom} \phi$ ,  $\theta^k, \theta^{k-1} \in \mathbf{dom} h$  and  $\lambda^k, \lambda^{k-1} \in \mathbb{R}^{n|\mathcal{E}|}$  be arbitrary. Let  $\{\tau_i^k\}_{i \in \mathcal{N}}$ ,  $\{\sigma_i^k\}_{i \in \mathcal{N}}$ ,  $\eta^k$  and  $\gamma^k$  be arbitrary positive scalars, and suppose  $(\mathbf{x}^{k+1}, \theta^{k+1}, \lambda^{k+1})$  is generated from  $(\mathbf{x}^k, \theta^k, \lambda^k)$  according to (3.8). For any  $\mathbf{x} \in \mathbf{dom} \phi$ ,  $\theta \in \mathbf{dom} h$  and  $\lambda \in \mathbb{R}^{n|\mathcal{E}|}$ , let  $\mathbf{z} \triangleq (\mathbf{x}, \mathbf{y})$  with  $\mathbf{y} \triangleq (\theta, \lambda)$ , and similarly  $\mathbf{y}^{k+1} \triangleq (\theta^{k+1}, \lambda^{k+1})$ . Then, for any collection of positive parameters  $\{\alpha_i^k, \alpha_i^{k+1}, \beta_i^k, \beta_i^{k+1}, \varsigma_i^k, \varsigma_i^{k+1}\}_{i \in \mathcal{N}}$ , the following inequality holds:*

$$\mathcal{L}(\mathbf{x}^{k+1}, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \mathbf{y}^{k+1}) \leq Q^k(\mathbf{z}) - R^{k+1}(\mathbf{z}) + P^k, \quad (4.4)$$

where

$$Q^k(\mathbf{z}) \triangleq \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i - x_i^k\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i - \theta_i^k\|^2 + \frac{\eta^k}{2\alpha_i^k} \|q_i^{k,y}\|^2 + \frac{\eta^k}{2\beta_i^k} \|q_i^{k,x}\|^2 \right) \quad (4.5a)$$

$$+ \frac{1}{2\gamma^k} \|\lambda - \lambda^k\|^2 + \eta^k \langle q^k, \mathbf{x} - \mathbf{x}^k \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^k - \lambda \rangle + \sum_{i \in \mathcal{N}} \frac{\eta^k \varsigma_i^k}{2} \|x_i^k - x_i^{k-1}\|^2$$

$$R^{k+1}(\mathbf{z}) \triangleq \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i - x_i^{k+1}\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i - \theta_i^{k+1}\|^2 + \frac{1}{2\alpha_i^{k+1}} \|q_i^{k+1,y}\|^2 + \frac{1}{2\beta_i^{k+1}} \|q_i^{k+1,x}\|^2 \right) \quad (4.5b)$$

$$+ \frac{1}{2\gamma^k} \|\lambda - \lambda^{k+1}\|^2 + \langle q^{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + \langle A\mathbf{x}^{k+1} - A\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle + \sum_{i \in \mathcal{N}} \frac{\varsigma_i^{k+1}}{2} \|x_i^{k+1} - x_i^k\|^2$$

$$P^k \triangleq \sum_{i \in \mathcal{N}} P_i^k - \frac{1}{2} \left( \frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} \frac{\eta^k}{\varsigma_i^k} d_i \right) \|\lambda^{k+1} - \lambda^k\|^2 \quad (4.5c)$$

$$P_i^k \triangleq -\frac{1}{2} \left( \frac{1}{\tau_i^k} - \eta^k (\alpha_i^k + \beta_i^k) - \varsigma_i^{k+1} \right) \|x_i^{k+1} - x_i^k\|^2 - \frac{1}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \quad (4.5d)$$

$$+ \frac{1}{2\alpha_i^{k+1}} \|q_i^{k+1,y}\|^2 + \frac{1}{2\beta_i^{k+1}} \|q_i^{k+1,x}\|^2 + f_i(x_i^{k+1}) - f_i(x_i^k) - \langle \nabla f_i(x_i^k), x_i^{k+1} - x_i^k \rangle, \forall i \in \mathcal{N},$$

for  $q_i^{k,x}$ ,  $q_i^{k,y}$ ,  $q_i^{k+1,x}$ , and  $q_i^{k+1,y}$  defined as in (4.2) for all  $i \in \mathcal{N}$ , and  $q^k$  defined as in (4.1).

*Proof.* Fix arbitrary  $\mathbf{x} \in \mathbf{dom} \phi$ ,  $\boldsymbol{\theta} \in \mathbf{dom} h$  and  $\lambda$ . Using [20, Lemma 7.1] for the  $\mathbf{x}$ -,  $\boldsymbol{\theta}$ - and  $\lambda$ - subproblems in (3.8), we get

$$\begin{aligned} & \phi(\mathbf{x}^{k+1}) - \phi(\mathbf{x}) + \langle \nabla f(\mathbf{x}^k) + \mathbf{p}^k, \mathbf{x}^{k+1} - \mathbf{x} \rangle \\ & \leq \sum_{i \in \mathcal{N}} \frac{1}{2\tau_i^k} (\|x_i - x_i^k\|^2 - \|x_i - x_i^{k+1}\|^2 - \|x_i^{k+1} - x_i^k\|^2), \end{aligned} \quad (4.6a)$$

$$\begin{aligned} & h(\boldsymbol{\theta}^{k+1}) - h(\boldsymbol{\theta}) - \langle G(\mathbf{x}^{k+1}), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta} \rangle \\ & \leq \sum_{i \in \mathcal{N}} \frac{1}{2\sigma_i^k} (\|\theta_i - \theta_i^k\|^2 - \|\theta_i - \theta_i^{k+1}\|^2 - \|\theta_i^{k+1} - \theta_i^k\|^2), \end{aligned} \quad (4.6b)$$

$$\begin{aligned} & \frac{1}{\gamma^k} \langle \lambda - \lambda^{k+1}, \lambda^{k+1} - \lambda^k \rangle = \langle \lambda - \lambda^{k+1}, A((1 + \eta^k)\mathbf{x}^k - \eta^k \mathbf{x}^{k-1}) \rangle \\ & \leq \frac{1}{2\gamma^k} (\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2 - \|\lambda^{k+1} - \lambda^k\|^2). \end{aligned} \quad (4.6c)$$

The convexity of  $f(\cdot)$  implies that

$$\begin{aligned} \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x} \rangle &= \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x} \rangle \\ &\geq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle - f(\mathbf{x}) \end{aligned} \quad (4.7)$$

holds for all  $\mathbf{x} \in \mathbf{dom} \phi$  and  $k \geq 0$ . Note that  $G(\mathbf{x}^{k+1}) = \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1})$ ; hence, the inner product in (4.6b) can be written equivalently using the linearity of  $\Phi(\mathbf{x}^{k+1}, \cdot)$ :

$$\begin{aligned} \langle G(\mathbf{x}^{k+1}), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta} \rangle &= \langle \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta} \rangle \\ &= \langle \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}), \boldsymbol{\theta}^k - \boldsymbol{\theta} \rangle + \langle \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k \rangle \\ &= \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}) - \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}, \lambda^{k+1}) + \langle \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k \rangle. \end{aligned} \quad (4.8)$$

By combining (4.8) and (4.6b), we have

$$\begin{aligned} h(\boldsymbol{\theta}^{k+1}) - h(\boldsymbol{\theta}) &\leq \sum_{i \in \mathcal{N}} \frac{1}{2\sigma_i^k} (\|\theta_i - \theta_i^k\|^2 - \|\theta_i - \theta_i^{k+1}\|^2 - \|\theta_i^{k+1} - \theta_i^k\|^2) \\ &+ \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}) - \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}, \lambda^{k+1}) + \langle \nabla_{\boldsymbol{\theta}} \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^k, \lambda^{k+1}), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k \rangle. \end{aligned} \quad (4.9)$$

On the other hand, by combining Equations (4.6a) and (4.7), and adding both  $f(\mathbf{x}^{k+1})$  and  $-\Phi(\mathbf{x}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1})$  to both sides, for  $k \geq 0$ , we get

$$\begin{aligned} & \phi(\mathbf{x}^{k+1}) + f(\mathbf{x}^{k+1}) - \phi(\mathbf{x}) - f(\mathbf{x}) - \Phi(\mathbf{x}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1}) \\ & \leq -\Phi(\mathbf{x}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1}) + \langle \mathbf{p}^k, \mathbf{x} - \mathbf{x}^{k+1} \rangle + \Lambda^k, \\ & + \sum_{i \in \mathcal{N}} \frac{1}{2\tau_i^k} \left[ \|x_i - x_i^k\|^2 - \|x_i - x_i^{k+1}\|^2 - \|x_i^{k+1} - x_i^k\|^2 \right], \end{aligned} \quad (4.10)$$

where  $\Lambda^k \triangleq f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) - \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle$ . Then, summing (4.9) and (4.10) leads to

$$\begin{aligned} & \mathcal{L}(\mathbf{x}^{k+1}, \boldsymbol{\theta}, \lambda) - \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1}) \\ & = \phi(\mathbf{x}^{k+1}) + f(\mathbf{x}^{k+1}) + \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}, \lambda) - h(\boldsymbol{\theta}) - \phi(\mathbf{x}) - f(\mathbf{x}) - \Phi(\mathbf{x}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1}) + h(\boldsymbol{\theta}^{k+1}) \\ & \leq \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1}) - \Phi(\mathbf{x}, \boldsymbol{\theta}^{k+1}, \lambda^{k+1}) + \langle \mathbf{p}^k, \mathbf{x} - \mathbf{x}^{k+1} \rangle \\ & + \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}, \lambda) - \Phi(\mathbf{x}^{k+1}, \boldsymbol{\theta}, \lambda^{k+1}) + \Lambda^k \end{aligned}$$

$$\begin{aligned}
 & + \sum_{i \in \mathcal{N}} \frac{1}{2\sigma_i^k} \left[ \|\theta_i - \theta_i^k\|^2 - \|\theta_i - \theta_i^{k+1}\|^2 - \|\theta_i^{k+1} - \theta_i^k\|^2 \right] \\
 & + \sum_{i \in \mathcal{N}} \frac{1}{2\tau_i^k} \left[ \|x_i - x_i^k\|^2 - \|x_i - x_i^{k+1}\|^2 - \|x_i^{k+1} - x_i^k\|^2 \right], \quad \forall k \geq 0,
 \end{aligned} \tag{4.11}$$

where we used  $\Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^{k+1}) - \Phi(\mathbf{x}^{k+1}, \theta^{k+1}, \lambda^{k+1}) + \langle \nabla_{\theta} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^{k+1}), \theta^{k+1} - \theta^k \rangle = 0$  due to linearity of  $\Phi(\mathbf{x}^{k+1}, \cdot, \lambda^{k+1})$ . Moreover, using the convexity of  $\Phi(\cdot, \mathbf{y}^{k+1})$ , we get

$$\begin{aligned}
 & \Phi(\mathbf{x}^{k+1}, \theta^{k+1}, \lambda^{k+1}) - \Phi(\mathbf{x}, \theta^{k+1}, \lambda^{k+1}) + \langle \mathbf{p}^k, \mathbf{x} - \mathbf{x}^{k+1} \rangle \\
 & \leq \langle \nabla_{\mathbf{x}} \Phi(\mathbf{x}^{k+1}, \theta^{k+1}, \lambda^{k+1}), \mathbf{x}^{k+1} - \mathbf{x} \rangle + \langle \nabla_{\mathbf{x}} \Phi(\mathbf{x}^k, \theta^k, \lambda^k) + \eta^k q^k, \mathbf{x} - \mathbf{x}^{k+1} \rangle \\
 & = -\langle q^{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + \eta^k \langle q^k, \mathbf{x} - \mathbf{x}^k \rangle + \eta^k \langle q^k, \mathbf{x}^k - \mathbf{x}^{k+1} \rangle.
 \end{aligned}$$

Next, we bound the term  $\Phi(\mathbf{x}^{k+1}, \theta, \lambda) - \Phi(\mathbf{x}^{k+1}, \theta, \lambda^{k+1})$  as follows:

$$\begin{aligned}
 & \Phi(\mathbf{x}^{k+1}, \theta, \lambda) - \Phi(\mathbf{x}^{k+1}, \theta, \lambda^{k+1}) \\
 & = \langle A\mathbf{x}^{k+1}, \lambda - \lambda^{k+1} \rangle \\
 & = \langle A\mathbf{x}^{k+1}, \lambda - \lambda^{k+1} \rangle - \langle A\mathbf{x}^k, \lambda - \lambda^{k+1} \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^{k+1} - \lambda \rangle \\
 & \quad + \langle A((1 + \eta^k)\mathbf{x}^k - \eta^k \mathbf{x}^{k-1}), \lambda - \lambda^{k+1} \rangle \\
 & \leq -\langle A\mathbf{x}^{k+1} - A\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^k - \lambda \rangle \\
 & \quad + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^{k+1} - \lambda^k \rangle + \frac{1}{2\gamma^k} (\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2 - \|\lambda^{k+1} - \lambda^k\|^2),
 \end{aligned}$$

where the inequality follows from (4.6c). Thus, the above inequalities with (4.11) implies

$$\begin{aligned}
 & \mathcal{L}(\mathbf{x}^{k+1}, \theta, \lambda) - \mathcal{L}(\mathbf{x}, \theta^{k+1}, \lambda^{k+1}) \\
 & \leq -\langle q^{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + \eta^k \langle q^k, \mathbf{x} - \mathbf{x}^k \rangle + \eta^k \langle q^k, \mathbf{x}^k - \mathbf{x}^{k+1} \rangle + \Lambda^k \\
 & \quad - \langle A\mathbf{x}^{k+1} - A\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^k - \lambda \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^{k+1} - \lambda^k \rangle \\
 & \quad + \sum_{i \in \mathcal{N}} \frac{1}{2\sigma_i^k} \left[ \|\theta_i - \theta_i^k\|^2 - \|\theta_i - \theta_i^{k+1}\|^2 - \|\theta_i^{k+1} - \theta_i^k\|^2 \right] \\
 & \quad + \sum_{i \in \mathcal{N}} \frac{1}{2\tau_i^k} \left[ \|x_i - x_i^k\|^2 - \|x_i - x_i^{k+1}\|^2 - \|x_i^{k+1} - x_i^k\|^2 \right] \\
 & \quad + \frac{1}{2\gamma^k} (\|\lambda - \lambda^k\|^2 - \|\lambda - \lambda^{k+1}\|^2 - \|\lambda^{k+1} - \lambda^k\|^2).
 \end{aligned}$$

Thus, given some arbitrary  $\alpha_i^k, \beta_i^k > 0$  for  $i \in \mathcal{N}$  and  $k \geq 0$ , (4.3) implies that

$$\begin{aligned}
 & \mathcal{L}(\mathbf{x}^{k+1}, \theta, \lambda) - \mathcal{L}(\mathbf{x}, \theta^{k+1}, \lambda^{k+1}) \\
 & \leq \left[ \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i - x_i^k\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i - \theta_i^k\|^2 \right) + \frac{1}{2\gamma^k} \|\lambda - \lambda^k\|^2 \right. \\
 & \quad \left. + \eta^k \langle q^k, \mathbf{x} - \mathbf{x}^k \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^k - \lambda \rangle \right] \\
 & \quad - \left[ \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i - x_i^{k+1}\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i - \theta_i^{k+1}\|^2 \right) + \frac{1}{2\gamma^k} \|\lambda - \lambda^{k+1}\|^2 \right]
 \end{aligned}$$

$$\begin{aligned}
& + \langle q^{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + \langle A\mathbf{x}^{k+1} - A\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle \Big] \\
& - \left[ \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right) + \frac{1}{2\gamma^k} \|\lambda^{k+1} - \lambda^k\|^2 \right] \\
& + \eta^k \langle q^k, \mathbf{x}^k - \mathbf{x}^{k+1} \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^{k+1} - \lambda^k \rangle + \Lambda^k \\
\leq & \left[ \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i - x_i^k\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i - \theta_i^k\|^2 + \frac{\eta^k}{2\alpha_i^k} \|q_i^{k,y}\|^2 + \frac{\eta^k}{2\beta_i^k} \|q_i^{k,x}\|^2 \right) + \frac{1}{2\gamma^k} \|\lambda - \lambda^k\|^2 \right] \\
& + \eta^k \langle q^k, \mathbf{x} - \mathbf{x}^k \rangle + \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^k - \lambda \rangle + \sum_{i \in \mathcal{N}} \frac{\eta^k \zeta_i^k}{2} \|x_i^k - x_i^{k-1}\|^2 \Big] \\
& - \left[ \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^k} \|x_i - x_i^{k+1}\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i - \theta_i^{k+1}\|^2 + \frac{1}{2\alpha_i^{k+1}} \|q_i^{k+1,y}\|^2 + \frac{1}{2\beta_i^{k+1}} \|q_i^{k+1,x}\|^2 \right) \right. \\
& + \frac{1}{2\gamma^k} \|\lambda - \lambda^{k+1}\|^2 + \langle q^{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + \langle A\mathbf{x}^{k+1} - A\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle \\
& \left. + \sum_{i \in \mathcal{N}} \frac{\zeta_i^{k+1}}{2} \|x_i^{k+1} - x_i^k\|^2 \right] \\
& - \left[ \sum_{i \in \mathcal{N}} \left( \frac{1}{2} \left( \frac{1}{\tau_i^k} - \eta^k (\alpha_i^k + \beta_i^k) \right) \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right) + \frac{1}{2\gamma^k} \|\lambda^{k+1} - \lambda^k\|^2 \right] \\
& + \Lambda^k + \sum_{i \in \mathcal{N}} \left( \frac{1}{2\alpha_i^{k+1}} \|q_i^{k+1,y}\|^2 + \frac{1}{2\beta_i^{k+1}} \|q_i^{k+1,x}\|^2 + \frac{\zeta_i^{k+1}}{2} \|x_i^{k+1} - x_i^k\|^2 \right. \\
& \left. + \frac{\eta^k}{2\zeta_i^k} d_i \|\lambda^{k+1} - \lambda^k\|^2 \right),
\end{aligned}$$

where in the last inequality we invoke Young's inequality the third time, i.e.,

$$\begin{aligned}
& \eta^k \langle A\mathbf{x}^k - A\mathbf{x}^{k-1}, \lambda^{k+1} - \lambda^k \rangle \\
& = \eta^k \langle \mathbf{x}^k - \mathbf{x}^{k-1}, A^\top (\lambda^{k+1} - \lambda^k) \rangle \\
& \leq \sum_{i \in \mathcal{N}} \left( \frac{\eta^k \zeta_i^k}{2} \|x_i^k - x_i^{k-1}\|^2 + \frac{\eta^k}{2\zeta_i^k} \|A_i A_i^\top\| \|\lambda^{k+1} - \lambda^k\|^2 \right) \tag{4.12} \\
& \leq \sum_{i \in \mathcal{N}} \left( \frac{\eta^k \zeta_i^k}{2} \|x_i^k - x_i^{k-1}\|^2 + \frac{\eta^k}{2\zeta_i^k} d_i \|\lambda^{k+1} - \lambda^k\|^2 \right)
\end{aligned}$$

holds for any  $\zeta_i^k > 0$  –in the first inequality  $A_i = \Omega_i \otimes \mathbf{I}_n$  with  $\Omega_i \in \mathbb{R}^{|\mathcal{N}|}$  denoting the  $i$ -th column of the Laplacian  $\Omega$  for  $i \in \mathcal{N}$ ; hence, the final inequality follows from the fact that the spectral norm  $\|A_i A_i^\top\| = \|\Omega_i\|^2 = d_i$  for  $i \in \mathcal{N}$ . Thus, using the definitions of  $Q^k(\mathbf{z})$ ,  $R^{k+1}(\mathbf{z})$  and  $P^k$  given in (4.5), we obtain the desired inequality in (4.4).  $\square$

**Lemma 4.3.** *Given some arbitrary initial points  $\mathbf{x}^0 \in \text{dom } \phi$ ,  $\theta^0 \in \mathcal{X}^*$ ,  $\lambda^0 \in \mathbb{R}^{n|\mathcal{E}|}$ , step sizes  $\gamma^k > 0$ ,  $\tau_i^k, \sigma_i^k > 0$  for  $i \in \mathcal{N}$ , and the momentum parameter  $\eta^k > 0$  for all  $k \geq 0$ , set  $\mathbf{x}^{-1} = \mathbf{x}^0$ ,  $\theta^{-1} = \theta^0$ ,  $\lambda^{-1} = \lambda^0$ , and consider the iterations as in (3.8). Suppose that there exists  $\{t_k\}_{k \geq 0} \subset$*

$\mathbb{R}_{++}$  such that the algorithm parameters satisfy the following inequalities:

$$\max_{i \in \mathcal{N}} \left\{ \frac{\tau_i^k}{\tau_i^{k+1}} \right\} \leq \frac{t_k}{t_{k+1}}, \quad \max_{i \in \mathcal{N}} \left\{ \frac{\sigma_i^k}{\sigma_i^{k+1}} \right\} \leq \frac{t_k}{t_{k+1}}, \quad \frac{\gamma^k}{\gamma^{k+1}} \leq \frac{t_k}{t_{k+1}}, \quad \eta^{k+1} = \frac{t_k}{t_{k+1}} \quad \forall k \geq 0. \quad (4.13)$$

Then, for any  $\mathbf{z} = (\mathbf{x}, \mathbf{y})$  such that  $\mathbf{x} \in \mathbf{dom} \phi$ ,  $\boldsymbol{\theta} \in \mathcal{X}^* \cap \mathcal{B}$ , it holds that

$$T_K (\mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K)) \leq t_0 Q^0(\mathbf{z}) - t_{K-1} R^K(\mathbf{z}) + \sum_{k=0}^{K-1} t_k P^k, \quad \forall K \geq 1, \quad (4.14)$$

where  $T_K = \sum_{k=0}^{K-1} t_k$ , and  $(\bar{\mathbf{x}}^K, \bar{\mathbf{y}}^K) \triangleq \frac{1}{T_K} \sum_{k=0}^{K-1} t_k (\mathbf{x}^k, \mathbf{y}^k)$ .

*Proof.* Fix an arbitrary  $\mathbf{z}$  as given in the statement of the lemma. Note that  $\mathcal{L}(\cdot, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \cdot)$  is a convex function; thus, if we multiply  $t_k$  for both sides of (4.4) and sum the resulting inequality from  $k = 0$  to  $K - 1$ , then using Jensen's lemma, we get

$$T_K (\mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K)) \leq \sum_{k=0}^{K-1} t_k (Q^k(\mathbf{z}) - R^{k+1}(\mathbf{z}) + P^k). \quad (4.15)$$

Next, we argue that under the parameter rule in (4.13), for  $k = 1$  to  $K - 2$ , we have  $t_{k+1} Q^{k+1}(\mathbf{z}) - t_k R^{k+1}(\mathbf{z}) \leq 0$ ; indeed, it is easy to verify this inequality after writing it equivalently as follows, using the definitions of  $Q^k(\mathbf{z})$  and  $R^{k+1}(\mathbf{z})$  given in (4.5a) and (4.5b), respectively:

$$\begin{aligned} & \sum_{i \in \mathcal{N}} \left( \frac{t_{k+1}}{2\tau_i^{k+1}} \|x_i - x_i^{k+1}\|^2 + \frac{t_{k+1}}{2\sigma_i^{k+1}} \|\theta_i - \theta_i^{k+1}\|^2 + \frac{t_k}{2\alpha_i^{k+1}} \|q_i^{k+1, \mathbf{y}}\|^2 + \frac{t_k}{2\beta_i^{k+1}} \|q_i^{k+1, \mathbf{x}}\|^2 \right) \\ & + \frac{t_{k+1}}{2\gamma^{k+1}} \|\lambda - \lambda^{k+1}\|^2 + t_k \langle q_{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + t_k \langle \mathbf{A}\mathbf{x}^{k+1} - \mathbf{A}\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle \\ & + \sum_{i \in \mathcal{N}} \frac{t_k \varsigma_i^{k+1}}{2} \|x_i^{k+1} - x_i^k\|^2 \\ & \leq \sum_{i \in \mathcal{N}} \left( \frac{t_k}{2\tau_i^k} \|x_i - x_i^{k+1}\|^2 + \frac{t_k}{2\sigma_i^k} \|\theta_i - \theta_i^{k+1}\|^2 + \frac{t_k}{2\alpha_i^{k+1}} \|q_i^{k+1, \mathbf{y}}\|^2 + \frac{t_k}{2\beta_i^{k+1}} \|q_i^{k+1, \mathbf{x}}\|^2 \right) \\ & + \frac{t_k}{2\gamma^k} \|\lambda - \lambda^{k+1}\|^2 + t_k \langle q^{k+1}, \mathbf{x} - \mathbf{x}^{k+1} \rangle + t_k \langle \mathbf{A}\mathbf{x}^{k+1} - \mathbf{A}\mathbf{x}^k, \lambda^{k+1} - \lambda \rangle \\ & + \sum_{i \in \mathcal{N}} \frac{t_k \varsigma_i^{k+1}}{2} \|x_i^{k+1} - x_i^k\|^2, \end{aligned}$$

where we used the condition  $t_k = t_{k+1} \eta^{k+1}$  for  $k \geq 0$ . Therefore, the desired inequality follows from (4.15).  $\square$

Lemma 4.3 indicates that bounding  $\sum_{k=0}^{K-1} t_k P^k$  is essential to derive rate results for D-APDB. Next, we show that  $P^k$  can be bounded above by some quantity that can be written as a node-specific consecutive iterate differences whenever  $\gamma^k$  is sufficiently small. This observation will play an important role in arguing for each  $i \in \mathcal{N}$  that (i) there is a non-zero lower bound  $\hat{\tau}_i > 0$  such that  $\tau_i^k \geq \hat{\tau}_i$  for  $k \geq 0$ , and that (ii) backtracking condition for agent- $i$  should be satisfied after a finite number of contractions for all iterations  $k \geq 0$ .

**Lemma 4.4.** For arbitrary  $k \geq 0$ ,  $P^k$  defined in (4.5c) can be bounded from above as follows:

$$\begin{aligned} P^k &\leq \sum_{i \in \mathcal{N}} \bar{P}_i^k - \frac{1}{2} \left[ \frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} d_i \left( \frac{\eta^k}{\zeta_i^k} + \frac{2}{\alpha_i^{k+1}} \right) \right] \|\lambda^{k+1} - \lambda^k\|^2, \\ \bar{P}_i^k &\triangleq \Lambda_i^k + \frac{1}{\alpha_i^{k+1}} \|\mathbf{J}g_i(x_i^{k+1})^\top (\theta_i^{k+1} - \theta_i^k)\|^2 + \frac{1}{2\beta_i^{k+1}} \|(\mathbf{J}g_i(x_i^{k+1}) - \mathbf{J}g_i(x_i^k))^\top \theta_i^k\|^2 \\ &\quad - \frac{1}{2} \left( \frac{1}{\tau_i^k} - \eta^k (\alpha_i^k + \beta_i^k) - \zeta_i^{k+1} \right) \|x_i^{k+1} - x_i^k\|^2 - \frac{1}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2, \end{aligned} \quad (4.16)$$

and  $\Lambda_i^k \triangleq f_i(x_i^{k+1}) - f_i(x_i^k) - \langle \nabla f_i(x_i^k), x_i^{k+1} - x_i^k \rangle$  for  $i \in \mathcal{N}$ . Thus, it holds that  $P^k \leq \sum_{i \in \mathcal{N}} \bar{P}_i^k$  if  $\gamma^k \leq \left( \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^{k+1}} + \frac{\eta^k}{\zeta_i^k} \right) \right)^{-1}$ .

*Proof.* Give arbitrary  $\mathbf{x}', \mathbf{x}'' \in \mathbf{dom} \phi$  and  $\theta', \theta'' \in \mathcal{X}^* \cap \mathcal{B}$ . Recall that  $\Phi(\mathbf{x}, \theta, \lambda) = \langle G(\mathbf{x}), \theta \rangle + \langle A\mathbf{x}, \lambda \rangle$  for any  $\mathbf{x} \in \mathbf{dom} \phi$ ,  $\theta \in \mathcal{X}^* \cap \mathcal{B}$  and  $\lambda \in \mathbb{R}^{n|\mathcal{E}|}$ . Thus, for any  $i \in \mathcal{N}$  and  $\lambda$ , we have

$$\|\nabla_{x_i} \Phi(\mathbf{x}'', \theta', \lambda) - \nabla_{x_i} \Phi(\mathbf{x}', \theta', \lambda)\| = \|(\mathbf{J}g_i(x_i'') - \mathbf{J}g_i(x_i'))^\top \theta_i'\|, \quad (4.17a)$$

$$\|\nabla_{x_i} \Phi(\mathbf{x}'', \theta'', \lambda) - \nabla_{x_i} \Phi(\mathbf{x}'', \theta', \lambda)\| = \|\mathbf{J}g_i(x_i'')^\top (\theta_i'' - \theta_i')\|. \quad (4.17b)$$

We consider the terms  $\frac{1}{\alpha_i^{k+1}} \|q_i^{k+1, \mathbf{y}}\|^2$  and  $\frac{1}{\beta_i^{k+1}} \|q_i^{k+1, \mathbf{x}}\|^2$  that appear in the definition of  $P^k$  given in (4.5c). For any  $i \in \mathcal{N}$ , using (4.17b), we get

$$\begin{aligned} \frac{1}{2\alpha_i^{k+1}} \|q_i^{k+1, \mathbf{y}}\|^2 &\triangleq \frac{1}{2\alpha_i^{k+1}} \|\nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^{k+1}, \lambda^{k+1}) - \nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^k)\|^2 \\ &\leq \frac{1}{\alpha_i^{k+1}} \|\nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^{k+1}, \lambda^{k+1}) - \nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^{k+1})\|^2 \\ &\quad + \frac{1}{\alpha_i^{k+1}} \|\nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^{k+1}) - \nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^k)\|^2 \\ &\leq \frac{1}{\alpha_i^{k+1}} \|\mathbf{J}g_i(x_i^{k+1})^\top (\theta_i^{k+1} - \theta_i^k)\|^2 + \frac{d_i}{\alpha_i^{k+1}} \|\lambda^{k+1} - \lambda^k\|^2, \end{aligned} \quad (4.18)$$

where in the first inequality we use  $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$  after adding and subtracting the term  $\nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^{k+1})$ , and the last inequality follows from (4.17b),  $A^\top (\lambda^{k+1} - \lambda^k) = \nabla_{\mathbf{x}} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^{k+1}) - \nabla_{\mathbf{x}} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^k)$  and the fact that the spectral norm  $\|A_i A_i^\top\| = \|\Omega_i\|^2 = d_i$  for  $i \in \mathcal{N}$ . Similarly, using (4.17a), we can obtain that

$$\begin{aligned} \frac{1}{2\beta_i^{k+1}} \|q_i^{k+1, \mathbf{x}}\|^2 &\triangleq \frac{1}{2\beta_i^{k+1}} \|\nabla_{x_i} \Phi(\mathbf{x}^{k+1}, \theta^k, \lambda^k) - \nabla_{x_i} \Phi(\mathbf{x}^k, \theta^k, \lambda^k)\|^2 \\ &= \frac{1}{2\beta_i^{k+1}} \|(\mathbf{J}g_i(x_i^{k+1}) - \mathbf{J}g_i(x_i^k))^\top \theta_i^k\|^2. \end{aligned} \quad (4.19)$$

Therefore, using the bounds in (4.18) and (4.19) within the definition of  $P^k$  in (4.5c), we get

$$\begin{aligned}
 P^k &= -\frac{1}{2} \sum_{i \in \mathcal{N}} \left[ \left( \frac{1}{\tau_i^k} - \eta^k (\alpha_i^k + \beta_i^k) - \varsigma_i^{k+1} \right) \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right] \\
 &\quad - \frac{1}{2} \left( \frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} \frac{\eta^k}{\varsigma_i^k} d_i \right) \|\lambda^{k+1} - \lambda^k\|^2 + \sum_{i \in \mathcal{N}} \left( \Lambda_i^k + \frac{1}{2\alpha_i^{k+1}} \|q_i^{k+1, \mathbf{y}}\|^2 + \frac{1}{2\beta_i^{k+1}} \|q_i^{k+1, \mathbf{x}}\|^2 \right) \\
 &\leq -\frac{1}{2} \sum_{i \in \mathcal{N}} \left[ \left( \frac{1}{\tau_i^k} - \eta^k (\alpha_i^k + \beta_i^k) - \varsigma_i^{k+1} \right) \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right] \\
 &\quad + \sum_{i \in \mathcal{N}} \left[ \Lambda_i^k + \frac{1}{\alpha_i^{k+1}} \|\mathbf{J}g_i(x_i^{k+1})^\top (\theta_i^{k+1} - \theta_i^k)\|^2 + \frac{1}{2\beta_i^{k+1}} \|(\mathbf{J}g_i(x_i^{k+1}) - \mathbf{J}g_i(x_i^k))^\top \theta_i^k\|^2 \right] \\
 &\quad - \frac{1}{2} \left[ \frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} d_i \left( \frac{\eta^k}{\varsigma_i^k} + \frac{2}{\alpha_i^{k+1}} \right) \right] \|\lambda^{k+1} - \lambda^k\|^2;
 \end{aligned}$$

hence,  $P^k \leq \sum_{i \in \mathcal{N}} \bar{P}_i^k$  holds whenever  $\gamma^k \leq \left( \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^{k+1}} + \frac{\eta^k}{\varsigma_i^k} \right) \right)^{-1}$ .  $\square$

## 5. PARAMETER CHOICES

In the previous section, we considered a meta algorithm, i.e., given some arbitrary initial points  $\mathbf{x}^0 \in \mathbf{dom} \phi$ ,  $\theta^0 \in \mathcal{X}^*$ ,  $\lambda^0 \in \mathbb{R}^{n|\mathcal{E}|}$ , step sizes  $\gamma^k > 0$ ,  $\tau_i^k, \sigma_i^k > 0$  for  $i \in \mathcal{N}$ , and the momentum parameter  $\eta^k > 0$  for all  $k \geq 0$ , we set  $\mathbf{x}^{-1} = \mathbf{x}^0$ ,  $\theta^{-1} = \theta^0$ ,  $\lambda^{-1} = \lambda^0$ , and consider the iterations as in (3.8). For this generic framework, we established a duality gap result given in (4.14) assuming that there exists  $\{t_k\}_{k \geq 0} \subset \mathbb{R}_{++}$  such that the algorithm parameters satisfy (4.13); furthermore, we were able to bound the error term  $P^k$  assuming that  $\gamma^k \leq \left( \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^{k+1}} + \frac{\eta^k}{\varsigma_i^k} \right) \right)^{-1}$ . In this section, we will show that D-APDB displayed in Algorithm 1 satisfies all these conditions stated above. Towards this goal, we first argue that D-APDB is well defined by showing that the agent-specific local backtracking conditions can be satisfied in finite backtracking iterations for all  $i \in \mathcal{N}$ .

For all  $i \in \mathcal{N}$ , let  $\tilde{\tau}_i^k, \tilde{\sigma}_i^k, \eta_i^k$  denote the candidate stepsizes and momentum term agent- $i$  will be testing by checking the stepsize-search condition  $E_i^k(\tilde{x}_i^{k+1}, \tilde{\theta}_i^{k+1}) \leq -\frac{\delta}{2\tilde{\tau}_i^k} \|\tilde{x}_i^{k+1} - x_i^k\|^2 - \frac{\delta}{2\tilde{\sigma}_i^k} \|\tilde{\theta}_i^{k+1} - \theta_i^k\|^2$  in Line 18 of D-APDB at the current iteration  $k \geq 0$  by setting  $\alpha_i^k, \beta_i^k, \tilde{\alpha}_i^{k+1}, \tilde{\beta}_i^{k+1}$ , and  $\tilde{\zeta}_i^{k+1}$  in a particular manner as stated in D-APDB, where  $E_i^k(\cdot, \cdot)$  is defined in (2.2). Note that at the time test condition is checked in Line 18, according to Line 13 and Line 14 of D-APDB, given some *candidate* primal step size  $\tilde{\tau}_i^k \in (0, \tau_i^{k-1}]$ , we set

$$\eta_i^k = \tau_i^{k-1} / \tilde{\tau}_i^k, \quad \tilde{\sigma}_i^k = \zeta_i \tilde{\tau}_i^k, \quad \tilde{\alpha}_i^{k+1} = c_\alpha / \tilde{\tau}_i^k, \quad \tilde{\beta}_i^{k+1} = c_\beta / \tilde{\tau}_i^k, \quad \tilde{\zeta}_i^{k+1} = c_\varsigma / \tilde{\tau}_i^k; \quad (5.1)$$

and the test point  $(\tilde{x}_i^{k+1}, \tilde{\theta}_i^{k+1})$  is computed according to Line 16 and Line 17 of D-APDB. Moreover, according to Line 39, we also have

$$\alpha_i^k = c_\alpha / \tau_i^{k-1}, \quad \beta_i^k = c_\beta / \tau_i^{k-1}, \quad \varsigma_i^k = c_\varsigma / \tau_i^{k-1}. \quad (5.2)$$

Therefore, evaluating  $E_i^k(\cdot, \cdot)$ , defined in (2.2), at the test point  $(\tilde{x}_i^{k+1}, \tilde{\theta}_i^{k+1})$  using the parameters as stated in (5.1) and (5.2), the test condition for agent- $i$  can be equivalently written as follows:

$$\begin{aligned} E_i^k(\tilde{x}_i^{k+1}, \tilde{\theta}_i^{k+1}) &= 2\tilde{\Lambda}_i^k - \frac{1}{\tilde{\tau}_i^k} \left(1 - (c_\alpha + c_\beta + c_\zeta)\right) \|\tilde{x}_i^{k+1} - x_i^k\|^2 - \frac{1}{\tilde{\sigma}_i^k} \|\tilde{\theta}_i^{k+1} - \theta_i^k\|^2 \\ &\quad + 2\frac{\tilde{\tau}_i^k}{c_\alpha} \|\mathbf{J}_{g_i}(\tilde{x}_i^{k+1})^\top (\tilde{\theta}_i^{k+1} - \theta_i^k)\|^2 + \frac{\tilde{\tau}_i^k}{c_\beta} \left\| \left(\mathbf{J}_{g_i}(\tilde{x}_i^{k+1}) - \mathbf{J}_{g_i}(x_i^k)\right)^\top \theta_i^k \right\|^2 \\ &\leq -\frac{\delta}{\tilde{\tau}_i^k} \|\tilde{x}_i^{k+1} - x_i^k\|^2 - \frac{\delta}{\tilde{\sigma}_i^k} \|\tilde{\theta}_i^{k+1} - \theta_i^k\|^2, \end{aligned} \quad (5.3)$$

where  $\tilde{\Lambda}_i^k \triangleq f_i(\tilde{x}_i^{k+1}) - f_i(x_i^k) - \langle \nabla f_i(x_i^k), \tilde{x}_i^{k+1} - x_i^k \rangle$ . Note that this condition can be easily checked by agent- $i$  locally without requiring any local communication.

Next, we provide a sufficient condition on the stepsizes for (5.3) to hold. First, we provide some useful bounds on partial gradient differences. Consider some arbitrary  $\mathbf{x}', \mathbf{x}'' \in \mathbf{dom} \phi$  and  $\theta', \theta'' \in \mathcal{X}^* \cap \mathcal{B}$ . Let  $\bar{L}_{x_i x_i} \triangleq L_{g_i} B_i$  for  $i \in \mathcal{N}$  such that  $L_{g_i} > 0$ ; thus, for any  $\lambda \in \mathbb{R}^{n|\mathcal{E}|}$ , Assumption 3 implies that

$$\|(\mathbf{J}_{g_i}(x_i'') - \mathbf{J}_{g_i}(x_i'))^\top \theta_i'\| \leq L_{g_i} \|\theta_i'\| \|x_i'' - x_i'\| \leq \bar{L}_{x_i x_i} \|x_i'' - x_i'\|, \quad (5.4)$$

which follows from  $\theta_i' \in \mathcal{B}_i$ , i.e.,  $\|\theta_i'\| \leq B_i$ . Similarly, Assumption 3 also implies

$$\|\mathbf{J}_{g_i}(x_i'')^\top (\theta_i'' - \theta_i')\| \leq C_{g_i} \|\theta_i'' - \theta_i'\|. \quad (5.5)$$

In the next result, we provide sufficient conditions for the backtracking line search to terminate.

**Theorem 5.1.** *Consider node  $i \in \mathcal{N}$ . Given D-APDB parameters  $\delta, c_\alpha, c_\beta, c_\zeta > 0$ , suppose  $\delta + c < 1$  where  $c \triangleq c_\alpha + c_\beta + c_\zeta$ . Moreover, given a candidate primal step size  $\tilde{\tau}_i^k \in (0, \tau_i^{k-1}]$ , suppose that candidate dual step size  $\tilde{\sigma}_i^k$  and momentum parameter  $\eta_i^k$  together with test function parameters  $\tilde{\alpha}_i^{k+1}, \tilde{\beta}_i^{k+1}, \tilde{\zeta}_i^{k+1}$  and  $\alpha_i^k, \beta_i^k, \zeta_i^k$  are set as in (5.1) and (5.2).*

*Then, for all  $k \geq 0$ , the test condition in Line 18 of D-APDB holds whenever at the time of checking it,  $\tilde{\tau}_i^k$  satisfies the following two conditions, with the convention  $L_{g_i} B_i = 0$  for  $L_{g_i} = 0$ ,*

$$\frac{1 - \delta}{\tilde{\tau}_i^k} \geq \frac{c}{\tilde{\tau}_i^k} + L_{f_i} + \frac{L_{g_i}^2 B_i^2}{c_\beta} \tilde{\tau}_i^k, \quad \frac{1 - \delta}{\tilde{\tau}_i^k} \geq \frac{2\zeta_i C_{g_i}^2}{c_\alpha} \tilde{\tau}_i^k, \quad (5.6)$$

*moreover, (5.6) is guaranteed to hold whenever  $\tilde{\tau}_i^k \in (0, \hat{\tau}_i]$ , where*

$$\hat{\tau}_i \triangleq \min \left\{ \frac{-L_{f_i} + \sqrt{L_{f_i}^2 + 4(1 - (\delta + c))L_{g_i}^2 B_i^2 / c_\beta}}{2L_{g_i}^2 B_i^2 / c_\beta}, \frac{1}{C_{g_i}} \sqrt{\frac{c_\alpha(1 - \delta)}{2\zeta_i}} \right\} \quad (5.7)$$

*for  $i \in \mathcal{N}$  such that  $L_{g_i} > 0$ , and  $\hat{\tau}_i \triangleq \min\{\frac{1 - (\delta + c)}{2L_{f_i}}, \frac{1}{C_{g_i}} \sqrt{\frac{c_\alpha(1 - \delta)}{2\zeta_i}}\}$  for  $i \in \mathcal{N}$  such that  $L_{g_i} = 0$ .*

*Finally, one also has<sup>5</sup>*

$$\hat{\tau}_i \geq \min \left\{ \frac{1 - (\delta + c)}{2L_{f_i}}, \frac{1}{L_{g_i} B_i} \sqrt{\frac{c_\beta(1 - (\delta + c))}{2}}, \frac{1}{C_{g_i}} \sqrt{\frac{c_\alpha(1 - \delta)}{2\zeta_i}} \right\}, \quad \forall i \in \mathcal{N}.$$

<sup>5</sup>With the convention that  $1/0 = \infty$  when  $L_{g_i} B_i = 0$  for  $i \in \mathcal{N}$  such that  $L_{g_i} = 0$ .

*Proof.* First, note that  $\tilde{\Lambda}_i^k \leq \frac{L_{f_i}}{2} \|\tilde{x}_i^{k+1} - x_i^k\|^2$  as  $\nabla f_i$  is Lipschitz with constant  $L_{f_i}$  over  $\mathbf{dom} \phi_i$ . Moreover, since  $g_i$  and  $\mathbf{J}g_i$  are Lipschitz on  $\mathbf{dom} \phi_i$  with constants  $C_{g_i}$  and  $L_{g_i}$ , respectively, and  $\|\theta_i^k\| \leq B_i$ , using (5.4) and (5.5), we get

$$\begin{aligned} & 2 \frac{\tilde{\tau}_i^k}{c_\alpha} \|\mathbf{J}g_i(\tilde{x}_i^{k+1})^\top (\tilde{\theta}_i^{k+1} - \theta_i^k)\|^2 + \frac{\tilde{\tau}_i^k}{c_\beta} \left\| \left( \mathbf{J}g_i(\tilde{x}_i^{k+1}) - \mathbf{J}g_i(x_i^k) \right)^\top \theta_i^k \right\|^2 \\ & \leq \frac{\tilde{\tau}_i^k}{c_\alpha} 2C_{g_i}^2 \|\tilde{\theta}_i^{k+1} - \theta_i^k\|^2 + \frac{\tilde{\tau}_i^k}{c_\beta} L_{g_i}^2 B_i^2 \|\tilde{x}_i^{k+1} - x_i^k\|^2. \end{aligned}$$

Since  $\tilde{\sigma}_i^k = \zeta_i \tilde{\tau}_i^k$ , we can conclude that (5.3) holds if (5.6) holds.

Clearly, the roots of the two quadratic inequalities in (5.6) immediately imply that any  $\tilde{\tau}_i^k \in (0, \hat{\tau}_i]$  satisfies (5.6) for  $\hat{\tau}_i$  given as in (5.7). Finally, the lower bound on  $\hat{\tau}_i$  is obtained by solving for the roots of another quadratic system, which provides us with a stronger condition, i.e.,  $\frac{1-(\delta+c)}{2\tilde{\tau}_i^k} \geq \max \left\{ L_{f_i}, \frac{L_{g_i}^2 B_i^2}{c_\beta} \tilde{\tau}_i^k \right\}$  and  $\frac{1-\delta}{\tilde{\tau}_i^k} \geq \frac{2\zeta_i C_{g_i}^2}{c_\alpha} \tilde{\tau}_i^k$ .  $\square$

**Definition 5.2.** Given a candidate primal step size  $\tilde{\tau}_i^k \in (0, \tau_i^{k-1}]$ , let candidate dual step size  $\tilde{\sigma}_i^k$  and momentum parameter  $\eta_i^k$  together with test function parameters  $\tilde{\alpha}_i^{k+1}, \tilde{\beta}_i^{k+1}, \tilde{\zeta}_i^{k+1}$  and  $\alpha_i^k, \beta_i^k, \zeta_i^k$  are set as in (5.1) and (5.2). The D-APDB parameters  $(\tilde{\tau}_i^k, \tilde{\sigma}_i^k, \eta_i^k)$  are called *admissible* if the test condition in Line 18 of D-APDB holds with these parameters.

In Theorem 5.1 we established that for each  $k \geq 0$ , the test condition in Line 18 of D-APDB holds after a finite number of backtracking iterations. For  $k \geq 0$ , let the *candidate* primal step size is  $\tilde{\tau}_i^k$  when the condition in Line 18 holds for  $i \in \mathcal{N}$ ; hence,  $\eta_i^k = \frac{\tau_i^{k-1}}{\tilde{\tau}_i^k}$ , which implies that the number of times node- $i$  employs backtracking (i.e., shrinks its primal step size) within the  $k$ -th iteration is equal to  $n_i^k \triangleq \log_{1/\rho} \eta_i^k$ . Thus,  $\eta^k = \max_{i \in \mathcal{N}} \{\eta_i^k\} = \max_{i \in \mathcal{N}} \{\tau_i^{k-1} / \tilde{\tau}_i^k\}$  denotes the largest amount of contraction of primal step size among all the nodes.

**Corollary 5.3.** *Under the premise of Theorem 5.1, for all  $k \geq 0$ , the number of times node- $i$  employs backtracking (i.e., shrinks its primal step size) within the  $k$ -th iteration of D-APDB, i.e.,  $n_i^k \triangleq \log_{1/\rho} \eta_i^k$ , is bounded above by  $\bar{n}_i \triangleq \log_{1/\rho} \lceil \tau_i / \hat{\tau}_i \rceil + 1$  for  $i \in \mathcal{N}$ .*

The *accepted* parameters  $\tau_i^k, \sigma_i^k, \eta^k, \alpha_i^{k+1}, \beta_i^{k+1}, \zeta_i^{k+1}$  are set after the max-consensus step  $\eta^k = \max_{i \in \mathcal{N}} \{\eta_i^k\}$ , where  $\tau_i^k = \tau_i^{k-1} / \eta^k$ ,  $\sigma_i^k = \zeta_i \tau_i^k$ , and  $\alpha_i^{k+1} = c_\alpha / \tau_i^k$ ,  $\beta_i^{k+1} = c_\beta / \tau_i^k$ ,  $\zeta_i^{k+1} = c_\zeta / \tau_i^k$ . Then, each node- $i$  computes  $x_i^{k+1}$  and  $\theta_i^{k+1}$  according to Line 32 and Line 33 of D-APDB, respectively, for  $i \in \mathcal{N}$ . It is essential to note that  $\eta^k = 1$  implies no backtracking is employed, i.e.,  $\eta_i^k = 1$  for all  $i \in \mathcal{N}$ ; hence,  $\tau_i^k = \tilde{\tau}_i^k = \tau_i^{k-1}$  for all  $i \in \mathcal{N}$ , which also implies that  $\tilde{x}_i^{k+1} = \text{prox}_{\tau_i^k \phi_i} \left( x_i^k - \tau_i^k (\nabla f_i(x_i^k) + p_i^k) \right)$  and  $\tilde{\theta}_i^{k+1} = \mathcal{P}_{\mathcal{X}_i^* \cap \mathcal{B}_i} \left( \theta_i^k + \sigma_i^k g_i(x_i^{k+1}) \right)$  – that is why we set  $x_i^{k+1} = \tilde{x}_i^{k+1}$  and  $\theta_i^{k+1} = \tilde{\theta}_i^{k+1}$  in Line 35 and Line 36 of D-APDB for the scenario  $\eta^k = 1$ .

Below we collect all the parameter conditions stated so far within the hypothesis of Lemmas 4.3 and 4.4 and Theorem 5.1 together with (5.1) and (5.2). All of these required conditions are formally stated below in **Condition 1**.

**Condition 1.** *For the given parameters  $\delta, c_\alpha, c_\beta, c_\zeta > 0$ , the primal-dual step-size sequence  $\{(\tau_i^k, \sigma_i^k)\}_{k \geq 0} \subset \mathbb{R}_{++} \times \mathbb{R}_{++}$  for  $i \in \mathcal{N}$  and the momentum parameter sequence  $\{(\gamma^k, \eta^k)\}_{k \geq 0} \subset \mathbb{R}_{++} \times \mathbb{R}_{++}$  is acceptable if there exists  $\{t_k\}_{k \geq 0} \subset \mathbb{R}_{++}$  such that (4.13) holds together with*

$\gamma^k \leq \left( \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^{k+1}} + \frac{\eta^k}{\zeta_i^k} \right) \right)^{-1}$  where the auxiliary parameter sequences  $\{(\alpha_i^k, \beta_i^k, \zeta_i^k)\}_{k \geq 0}$  are defined as in (5.2) such that  $\delta + c < 1$  for  $c \triangleq c_\alpha + c_\beta + c_\zeta$ , and  $(\alpha_i^0, \beta_i^0, \zeta_i^0)$  is initialized as  $\alpha_i^0 = c_\alpha / \bar{\tau}_i$ ,  $\beta_i^0 = c_\beta / \bar{\tau}_i$ ,  $\zeta_i^0 = c_\zeta / \bar{\tau}_i$  for some  $\bar{\tau}_i > 0$  such that  $\tau_i^k \leq \bar{\tau}_i$  for  $k \geq 0$  for all  $i \in \mathcal{N}$ .

Next, we show that the particular parameter choice in D-APDB satisfy **Condition 1**.

**Theorem 5.4.** *Suppose Assumptions 1 to 3 hold, and  $\delta, c_\alpha, c_\beta, c_\zeta > 0$  are given that satisfy  $\delta + c < 1$ , where  $c \triangleq c_\alpha + c_\beta + c_\zeta$ . For each  $i \in \mathcal{N}$ , let  $x_i^0 \in \mathbf{dom} \phi_i$ ,  $\theta_i^0 \in \mathcal{X}_i^*$  be the initial primal-dual variables, and let  $\bar{\tau}_i, \zeta_i > 0$  and  $\rho \in (0, 1)$  denote the step size parameters of node  $i$  and the contraction coefficient. Then, the primal-dual step-size sequence  $\{(\tau_i^k, \sigma_i^k)\}_{k \geq 0} \subset \mathbb{R}_{++} \times \mathbb{R}_{++}$  for  $i \in \mathcal{N}$  and the momentum parameter sequence  $\{(\gamma^k, \eta^k)\}_{k \geq 0} \subset \mathbb{R}_{++} \times \mathbb{R}_{++}$  generated according to D-APDB are acceptable with respect to **Condition 1**.*

*Proof.* For all  $k \geq 0$ , according to Theorem 5.1, the test condition in Line 18 of D-APDB holds within finite backtracking iterations, let  $(\tilde{\tau}_i^k, \tilde{\sigma}_i^k, \eta_i^k)$  be an admissible set of local parameters in terms of Definition 5.2. According to D-APDB, we have  $\eta_i^k = \tau_i^{k-1} / \tilde{\tau}_i^k \geq 1$  since  $\tilde{\tau}_i^k \in (\rho_i \hat{\tau}_i, \tau_i^{k-1}]$  for all  $i \in \mathcal{N}$ ; therefore,  $\eta^k = \max_{i \in \mathcal{N}} \eta_i^k \geq 1$  as well. Moreover, according to Line 26 and Line 28 of D-APDB, we have  $\gamma^k = \frac{c_\gamma}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta} \right)^{-1}$  for some  $c_\gamma \in \left( 0, \frac{1}{2|\mathcal{E}|} \right)$  while the primal-dual step sizes are set to  $\tau_i^k = \tau_i^{k-1} / \eta^k$  and  $\sigma_i^k = \zeta_i \tau_i^k$ , where  $\bar{\tau} = \max_{i \in \mathcal{N}} \bar{\tau}_i$ . Hence, for any  $i \in \mathcal{N}$ , it holds that

$$\tau_i^k = \tau_i^{k-1} / \eta^k \leq \tau_i^{k-1} / \eta_i^k \leq \tau_i^{k-1}. \quad (5.8)$$

The auxiliary parameter update in Line 39 of D-APDB shows that  $(\alpha_i^{k+1}, \beta_i^{k+1}, \zeta_i^{k+1})$  are defined based on (5.2) such that  $\delta + c < 1$  for  $c \triangleq c_\alpha + c_\beta + c_\zeta$ , and  $(\alpha_i^0, \beta_i^0, \zeta_i^0)$  is indeed initialized as  $\alpha_i^0 = c_\alpha / \bar{\tau}_i$ ,  $\beta_i^0 = c_\beta / \bar{\tau}_i$ ,  $\zeta_i^0 = c_\zeta / \bar{\tau}_i$  for some  $\bar{\tau}_i > 0$  such that  $\tau_i^k \leq \bar{\tau}_i$  for  $k \geq 0$  for all  $i \in \mathcal{N}$ —the last condition holds as it has been shown that  $\{\tau_i^k\}_{k \geq 0}$  is a nonincreasing sequence such that  $\tau_i^k \leq \bar{\tau}_i$  for all  $k \geq 0$  by construction for all  $i \in \mathcal{N}$ .

Next, we check the condition on  $\{\gamma^k\}$  to be acceptable by verifying that

$$\begin{aligned} \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^{k+1}} + \frac{\eta^k}{\zeta_i^k} \right) &= \sum_{i \in \mathcal{N}} d_i \left( \frac{2\tau_i^k}{c_\alpha} + \frac{\eta^k \tau_i^{k-1}}{c_\zeta} \right) \leq \sum_{i \in \mathcal{N}} d_i \tau_i^{k-1} \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta} \right) \\ &\leq \bar{\tau} \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta} \right) \leq \frac{\bar{\tau}}{c_\gamma} \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta} \right); \end{aligned} \quad (5.9)$$

hence, we can conclude that  $\gamma^k \leq \left( \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^{k+1}} + \frac{\eta^k}{\zeta_i^k} \right) \right)^{-1}$  as in **Condition 1**, where the first inequality follows from (5.8), the second inequality uses  $\tau_i^k \leq \bar{\tau}_i$  for  $k \geq -1$  and  $\bar{\tau} = \max_{i \in \mathcal{N}} \bar{\tau}_i$ , finally in the last inequality we use  $\sum_{i \in \mathcal{N}} d_i = 2|\mathcal{E}|$ .

To verify that **Condition 1** holds for the sequences  $\{(\tau_i^k, \sigma_i^k)\}_{k \geq 0}$  for  $i \in \mathcal{N}$  and  $\{(\gamma^k, \eta^k)\}_{k \geq 0}$  generated according to D-APDB, one needs to show that there exists  $\{t_k\}_{k \geq 0} \subset \mathbb{R}_{++}$  such that (4.13) holds. At this point, we have shown that  $\{\eta^k\}_{k \geq 0}$  exists that satisfy  $\eta^k \geq 1$ . Thus, we recursively define  $\{t_k\}_{k \geq 0}$  as follows:  $t_0 = 1$  and  $t_{k+1} = t_k / \eta^{k+1}$ , which implies that  $t_{k+1} \leq t_k \leq 1$  for  $k \geq 0$ . Moreover, (5.8) and our choice of  $\sigma_i^k = \zeta_i \tau_i^k$  imply that  $\max_{i \in \mathcal{N}} \left\{ \frac{\tau_i^{k-1}}{\tau_i^k} \right\} = \eta^k = \frac{t_{k-1}}{t_k}$  and  $\max_{i \in \mathcal{N}} \left\{ \frac{\sigma_i^{k-1}}{\sigma_i^k} \right\} = \eta^k = \frac{t_{k-1}}{t_k}$  both hold by construction. For  $k \geq 0$ , the only thing that

remains to be shown at this point is  $\frac{\gamma^k}{\gamma^{k+1}} \leq \frac{t_k}{t_{k+1}} = \eta^{k+1}$ ; indeed,

$$\frac{\gamma^k}{\gamma^{k+1}} = \frac{\frac{c_\gamma}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\zeta} \right)^{-1}}{\frac{c_\gamma}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{\eta^{k+1}}{c_\zeta} \right)^{-1}} = \frac{2c_\zeta + \eta^{k+1}c_\alpha}{2c_\zeta + \eta^k c_\alpha} \leq \eta^{k+1} \quad (5.10)$$

holds if and only if  $2c_\zeta(\eta^{k+1} - 1) + c_\alpha\eta^{k+1}(\eta^k - 1) \geq 0$ , which can be easily verified as  $\eta^k \geq 1$  and  $\eta^{k+1} \geq 1$ .  $\square$

## 6. CONVERGENCE GUARANTEES FOR D-APDB AND THE PROOF OF MAIN RESULTS

Now we are ready to establish convergence guarantees for D-APDB. We first analyze how we can bound the duality gap for D-APDB primal-dual iterate sequence through carefully controlling the difference between the admissible parameters  $(\bar{\tau}_i^k, \bar{\sigma}_i^k, \eta_i^k)$  according to Definition 5.2 and the updated parameters after computing max-consensus among the nodes  $(\tau_i^k, \sigma_i^k, \eta^k)$ .

We first define an index set that plays a crucial role in the analysis of D-APDB. Recall that for all  $k \geq 0$ ,  $\eta^k = \max_{i \in \mathcal{N}} \frac{\tau_i^{k-1}}{\bar{\tau}_i^k}$ ; thus,  $\eta^k > 1$  implies that in the  $k$ -th D-APDB iteration at least one node has employed backtracking, i.e., its candidate primal step size has contracted.

**Definition 6.1.** Let  $\mathcal{S} \triangleq \{k \in \mathbb{Z}_+ : \eta^k > 1\}$  be the set of iterations with contraction.

**Lemma 6.2.** *Under the premise of Theorem 5.4, let  $\{\eta^k\}_{k \geq 0}$  be the momentum sequence generated by D-APDB displayed in algorithm 1. Then,  $\{t_k\}_{k \geq 0}$  such that  $t_0 = 1$  and  $t_{k+1} = t_k/\eta^{k+1}$  for  $k \geq 0$  satisfies Condition 1 such that  $t_k \leq t_0 = 1$  and  $t_k \geq \hat{t} \triangleq \rho \min_{i \in \mathcal{N}} \hat{\tau}_i/\bar{\tau}_i$ , which implies that  $T_K = \sum_{k=0}^{K-1} t_k \geq \hat{t}K$  for all  $K \geq 1$ , i.e.,  $T_K = \Omega(K)$ . Moreover,  $|\mathcal{S}| \leq \left\lfloor \log_{1/\rho} \left( \max_{i \in \mathcal{N}} \frac{\bar{\tau}_i}{\hat{\tau}_i} \right) \right\rfloor + 1$ .*

*Proof.* Since  $t_0 = 1$ , from the recursive definition of  $\{t_k\}$ , it follows that  $t_k = \left( \prod_{\ell=1}^k \eta^\ell \right)^{-1}$  for  $k \geq 1$ . Moreover, for  $i \in \mathcal{N}$  and  $k \geq 1$ , since  $\tau_i^k = \tau_i^{k-1}/\eta^k$ , we have  $\prod_{\ell=1}^k \eta^\ell = \frac{\tau_i^0}{\tau_i^k}$ ; hence,  $t_k = \frac{\tau_i^k}{\tau_i^0}$ . Due to  $\tau_i^k = \tau_i^{k-1}/\eta^k$  for  $i \in \mathcal{N}$ , whenever at least one node shrinks its step size, all the other nodes must shrink theirs as well. Therefore, for all  $k \geq 0$  such that  $\eta^k > 1$ , we have  $\log_{1/\rho} \eta^k \geq 1$ , i.e., every node must shrink their step size from  $\tau_i^{k-1}$  to  $\tau_i^k$  by using at least one contraction by  $\rho \in (0, 1)$ . On the other hand, since  $\{\tau_i^k\}_{k \geq 0}$  is nonincreasing sequences for all  $i \in \mathcal{N}$ , it follows from Theorem 5.1 that whenever  $\tau_i^{k^*-1} \leq \hat{\tau}_i$  for all  $i \in \mathcal{N}$  for some  $k^* \geq 0$ , the test condition in Line 18 of D-APDB is guaranteed to hold for all  $k \geq k^*$ , i.e.,  $\eta^k = 1$  for all  $k \geq k^*$ . Therefore, we can conclude that  $|\mathcal{S}| \leq \left\lfloor \log_{1/\rho} \left( \max_{i \in \mathcal{N}} \frac{\bar{\tau}_i}{\hat{\tau}_i} \right) \right\rfloor + 1$ ; thus,  $\tau_i^k \geq \bar{\tau}_i \rho^{|\mathcal{S}|} \geq \rho \bar{\tau}_i \min_{j \in \mathcal{N}} \frac{\hat{\tau}_j}{\bar{\tau}_j}$  for all  $i \in \mathcal{N}$ . Using our earlier observation  $t_k = \frac{\tau_i^k}{\tau_i^0}$  for all  $i \in \mathcal{N}$  and  $k \geq 1$ , we obtain the following lower bound:  $t_k \geq \rho \min_{i \in \mathcal{N}} \frac{\hat{\tau}_i}{\bar{\tau}_i}$ , which follows from  $\tau_i^0 \leq \bar{\tau}_i$  for  $i \in \mathcal{N}$ .  $\square$

**Lemma 6.3.** *Under the premise of Theorem 5.4, given some arbitrary initial points  $(x_i^0, \theta_i^0) \in \text{dom } \phi_i \times \mathcal{X}_i^*$  and the step size parameters  $\bar{\tau}_i, \zeta_i > 0$  for  $i \in \mathcal{N}$ , and the contraction coefficient  $\rho \in (0, 1)$ , let  $\{x_i^k, \theta_i^k\}_{k \geq 0}$  denote the D-APDB iterate sequence for  $i \in \mathcal{N}$ . Then, it holds that*

$$\sum_{i \in \mathcal{N}} \bar{P}_i^k \leq -\frac{\delta}{2} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right) + \mathbf{1}_{\mathcal{S}}(k) \sum_{i \in \mathcal{N}} \Xi_i^k, \quad (6.1)$$

where  $\mathcal{I}$  is given in Definition 6.1,  $\bar{P}_i^k$  is defined in (4.16), and  $\mathbf{1}_{\mathcal{I}}(k) = 1$  if  $k \in \mathcal{I}$ , equal to 0 otherwise; moreover, for  $i \in \mathcal{N}$ , for the case  $L_{g_i} > 0$ ,  $\Xi_i^k \triangleq 2\left(L_{f_i} + \frac{\bar{\tau}_i}{c_\beta} B_i^2 L_{g_i}^2\right) D_i^2 + 4\frac{\bar{\tau}_i}{c_\alpha} C_{g_i}^2 B_i^2$ , and for the case  $L_{g_i} = 0$ ,  $\Xi_i^k = 2L_{f_i} D_i^2 + \frac{\bar{\tau}_i}{c_\alpha} C_{g_i}^2 \|\theta_i^{k+1} - \theta_i^k\|^2$  for  $k \geq 0$ .

*Proof.* Let  $\mathcal{I} \subset \mathbb{Z}_+$  be the index set as given in Definition 6.1. First, consider a fixed iteration  $k \in \mathbb{Z}_+$  such that  $k \notin \mathcal{I}$ . Since  $\eta^k \geq 1$  for all  $k \geq 0$ ,  $k \in \mathcal{I}^c$  implies that  $\eta^k = 1$ ; hence, according to Line 35 and Line 36 of D-APDB, we get  $\tilde{x}_i^{k+1} = x_i^{k+1}$  and  $\tilde{\theta}_i^{k+1} = \theta_i^{k+1}$  for all  $i \in \mathcal{N}$ , and we have  $\tilde{\tau}_i^k = \tau_i^k = \tau_i^{k-1}$ , which also implies  $\tilde{\sigma}_i^k = \sigma_i^k = \sigma_i^{k-1}$  together with  $\tilde{\alpha}_i^{k+1} = \alpha_i^{k+1}$ ,  $\tilde{\beta}_i^{k+1} = \beta_i^{k+1}$  and  $\tilde{\zeta}_i^{k+1} = \zeta_i^{k+1}$ . Therefore, it is easy to verify that for all  $i \in \mathcal{N}$ , it holds that

$$\bar{P}_i^k = \frac{1}{2} E_i^k(x_i^{k+1}, \theta_i^{k+1}) \leq -\frac{\delta}{2\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 - \frac{\delta}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2, \quad \forall k \in \mathcal{I}^c, \quad (6.2)$$

which follows from Line 18 of D-APDB and (5.3).

Next, consider a fixed iteration  $k \in \mathbb{Z}_+$  such that  $k \in \mathcal{I}$ . For each  $i \in \mathcal{N}$ , using  $\Lambda_i^k \leq \frac{L_{f_i}}{2} \|x_i^{k+1} - x_i^k\|^2$ ,  $\alpha_i^{k+1} = c_\alpha / \tau_i^k$ ,  $\beta_i^{k+1} = c_\beta / \tau_i^k$ ,  $\zeta_i^{k+1} = c_\zeta / \tau_i^k$  and  $\tau_i^k = \tau_i^{k-1} / \eta^k$ , we can bound  $\bar{P}_i^k$  defined in (4.16) as follows:

$$\begin{aligned} \bar{P}_i^k &\leq \frac{L_{f_i}}{2} \|x_i^{k+1} - x_i^k\|^2 - \frac{1}{2\tau_i^k} \left(1 - (c_\alpha + c_\beta + c_\zeta)\right) \|x_i^{k+1} - x_i^k\|^2 - \frac{1}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \\ &\quad + \frac{\tau_i^k}{c_\alpha} \|\mathbf{J}g_i(x_i^{k+1})^\top (\theta_i^{k+1} - \theta_i^k)\|^2 + \frac{\tau_i^k}{2c_\beta} \left(\mathbf{J}g_i(x_i^{k+1}) - \mathbf{J}g_i(x_i^k)\right)^\top \theta_i^k \|^2 \\ &\leq \frac{1}{2} \left(L_{f_i} + \frac{\tau_i^k}{c_\beta} B_i^2 L_{g_i}^2 - \frac{\delta}{\tau_i^k}\right) \|x_i^{k+1} - x_i^k\|^2 + \left(\frac{\tau_i^k}{c_\alpha} C_{g_i}^2 - \frac{1}{2\sigma_i^k}\right) \|\theta_i^{k+1} - \theta_i^k\|^2, \end{aligned}$$

where we used  $\eta^k(\alpha_i^k + \beta_i^k) = (c_\alpha + c_\beta) / \tau_i^k$  for the first inequality, and  $\delta \leq 1 - (c_\alpha + c_\beta + c_\zeta)$  in the second inequality. For all  $i \in \mathcal{N}$ , recall that for the case  $L_{g_i} > 0$ , we have  $\|\theta_i^k\| \leq B_i$ ; hence,  $\|\theta_i^{k+1} - \theta_i^k\|^2 \leq 4B_i^2$  for all  $k \geq 0$ . Therefore, for all  $i \in \mathcal{N}$ , since  $\tau_i^k \leq \bar{\tau}_i$ , using  $\|x_i^{k+1} - x_i^k\|^2 \leq 4D_i^2$  together with  $\delta \in (0, 1)$ , we get  $\bar{P}_i^k \leq -\frac{\delta}{2\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 - \frac{\delta}{2\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 + \Xi_i^k$ , for all  $k \in \mathcal{I}$ .  $\square$

**Theorem 6.4.** *Under the premise of Theorem 5.4, given some arbitrary initial points  $(x_i^0, \theta_i^0) \in \text{dom } \phi_i \times \mathcal{K}_i^*$  and the step size parameters  $\bar{\tau}_i, \zeta_i > 0$  for  $i \in \mathcal{N}$ , and the contraction coefficient  $\rho \in (0, 1)$ , it holds for  $i \in \mathcal{N}$  that the D-APDB dual iterate sequence  $\{\theta_i^k\}_{k \geq 0}$  is bounded; hence, for  $i \in \mathcal{N}$  such that  $L_{g_i} = 0$ , there exists  $B_i > 0$  such that  $\|\theta_i^k\| \leq B_i$  for all  $k \geq 0$ .*

*Proof.* Let  $(\hat{x}, \hat{\theta}, \hat{\lambda})$  be an arbitrary saddle point of  $\mathcal{L}$ . Using Young's inequality in a similar manner to (4.3), it follows from the definition of  $Q^k(\hat{z})$  in (4.5a) that

$$\begin{aligned} t_k Q^k(\hat{z}) &\geq \sum_{i \in \mathcal{N}} \left[ \frac{t_k}{2} \left( \frac{1}{\tau_i^k} - \eta^k(\alpha_i^k + \beta_i^k) \right) \|x_i^k - \hat{x}_i\|^2 + \frac{t_k}{2\sigma_i^k} \|\theta_i^k - \hat{\theta}_i\|^2 \right] \\ &\quad + \frac{t_k}{2} \left( \frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} d_i \frac{\eta^k}{\zeta_i^k} \right) \|\lambda^k - \hat{\lambda}\|^2 \\ &\geq \frac{\hat{t}}{2} \sum_{i \in \mathcal{N}} \left( \frac{\delta}{\bar{\tau}_i} \|x_i^k - \hat{x}_i\|^2 + \frac{1}{\zeta_i \bar{\tau}_i} \|\theta_i^k - \hat{\theta}_i\|^2 \right) + \frac{2\hat{t}\rho}{c_\alpha} |\mathcal{E}| \min_{i \in \mathcal{N}} \{\hat{\tau}_i\} \|\lambda^k - \hat{\lambda}\|^2 > 0, \end{aligned} \quad (6.3)$$

where in the second inequality we used  $t_k \geq \hat{t}$ ,  $\tau_i^k \leq \bar{\tau}_i$ ,  $\sigma_i^k = \zeta_i \tau_i^k$ , and  $1 - c_\alpha - c_\beta \geq \delta$  together with the fact that

$$\frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} d_i \frac{\eta^k}{\varsigma_i^k} \geq \sum_{i \in \mathcal{N}} d_i \frac{2}{\alpha_i^{k+1}} = \frac{2}{c_\alpha} \sum_{i \in \mathcal{N}} d_i \tau_i^k \geq \frac{4\rho}{c_\alpha} |\mathcal{E}| \min_{i \in \mathcal{N}} \hat{\tau}_i > 0,$$

which holds due to Condition 1,  $|\mathcal{E}| = \sum_{i \in \mathcal{N}} d_i/2$  and  $\tau_i^k \geq \rho \min_{i \in \mathcal{N}} \hat{\tau}_i$  for all  $k \geq 0$ .

Moreover, since Condition 1 holds, (4.13) immediately implies  $t_k R^{k+1}(\hat{\mathbf{z}}) \geq t_{k+1} Q^{k+1}(\hat{\mathbf{z}})$ . Therefore, after multiplying both sides of (4.4) by  $t_k$ , and using the fact that  $\mathcal{L}(\mathbf{x}^{k+1}, \hat{\mathbf{y}}) - \mathcal{L}(\hat{\mathbf{x}}, \mathbf{y}^{k+1}) \geq 0$ , Lemma 4.4 implies that

$$0 \leq t_k Q^k(\hat{\mathbf{z}}) - t_{k+1} Q^{k+1}(\hat{\mathbf{z}}) + t_k \sum_{i \in \mathcal{N}} \bar{P}_i^k - \frac{t_k}{2} \left[ \frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} d_i \left( \frac{\eta^k}{c_\varsigma} \tau_i^{k-1} + \frac{2}{c_\alpha} \tau_i^k \right) \right] \|\lambda^{k+1} - \lambda^k\|^2.$$

Recall that  $\gamma^k = \frac{c_\gamma}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{\eta^k}{c_\varsigma} \right)^{-1}$  for some  $c_\gamma \in \left( 0, \frac{1}{2|\mathcal{E}|} \right)$ ; hence, using  $\eta^k \geq 1$ ,  $\tau_i^k \leq \tau_i^{k-1} \leq \bar{\tau}_i \leq \bar{\tau}$  and  $\sum_{i \in \mathcal{N}} d_i = 2|\mathcal{E}|$ , we get  $\frac{1}{\gamma^k} - \sum_{i \in \mathcal{N}} d_i \left( \frac{\eta^k}{c_\varsigma} \tau_i^{k-1} + \frac{2}{c_\alpha} \tau_i^k \right) \geq \bar{\tau} \left( \frac{1}{c_\gamma} - \sum_{i \in \mathcal{N}} d_i \right) \left( \frac{2}{c_\alpha} - \frac{1}{c_\varsigma} \right) > 0$ . Finally, we define  $\delta' \triangleq \hat{t} \cdot \bar{\tau} \left( \frac{1}{c_\gamma} - \sum_{i \in \mathcal{N}} d_i \right) \left( \frac{2}{c_\alpha} - \frac{1}{c_\varsigma} \right) > 0$ . Thus, using (6.1) and the fact that  $0 < \hat{t} \leq t_k \leq t_0 = 1$ , we obtain a crucial inequality:

$$\begin{aligned} 0 \leq t_k Q^k(\hat{\mathbf{z}}) - t_{k+1} Q^{k+1}(\hat{\mathbf{z}}) + \mathbf{1}_{\mathcal{S}}(k) \sum_{i \in \mathcal{N}} \Xi_i^k - \frac{\delta'}{2} \|\lambda^{k+1} - \lambda^k\|^2 \\ - \frac{\delta}{2} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right), \quad \forall k \geq 0. \end{aligned} \tag{6.4}$$

Next, let  $a_k \triangleq t_k Q^k(\hat{\mathbf{z}})$ ,  $b_k \triangleq \frac{\delta}{2} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right) + \frac{\delta'}{2} \|\lambda^{k+1} - \lambda^k\|^2$ , and  $c_k \triangleq \mathbf{1}_{\mathcal{S}}(k) \sum_{i \in \mathcal{N}} \Xi_i^k$  for all  $k \geq 0$ . Recall that  $\Xi_i^k = 2L_{f_i} D_i^2 + \frac{\bar{\tau}_i}{c_\alpha} C_{g_i}^2 \|\theta_i^{k+1} - \theta_i^k\|^2$  for  $i \in \mathcal{N}$  such that  $L_{g_i} = 0$ ; on the other hand,  $\Xi_i^k = 2L_{f_i} D_i^2 + 4\frac{\bar{\tau}_i}{c_\alpha} C_{g_i}^2 B_i^2$  for  $i \in \mathcal{N}$  such that  $L_{g_i} > 0$ . Clearly,  $a_k, b_k, c_k \geq 0$  such that  $a_{k+1} \leq a_k - b_k + c_k$  for all  $k \geq 0$ . Recall that  $|\mathcal{S}|$  is bounded from above according to Lemma 6.2; therefore, we can conclude that  $\sum_{k=0}^{+\infty} c_k < \infty$ . This implies that  $\lim_{k \rightarrow \infty} a_k \geq 0$  exists and  $\sum_{k=0}^{+\infty} b_k < +\infty$ . Since  $\{a_k\}$  is convergent, it must be bounded; hence, (6.3) implies that  $\{(x_i^k, \theta_i^k)\}_{k \geq 0}$  for all  $i \in \mathcal{N}$  and  $\{\lambda^k\}_{k \geq 0}$  are bounded sequences—note that even if  $\{x_i^k\}_{k \geq 0}$  must be bounded by Assumption 2, for the case  $g_i(\cdot)$  is affine we set  $\mathcal{B}_i = \mathbb{R}^{m_i}$ , and this step of the proof is essential to establish  $\{\theta_i^k\}_{k \geq 0}$  stays bounded even for the setting  $g_i(\cdot)$  is affine and we set  $\mathcal{B}_i = \mathbb{R}^{m_i}$ .  $\square$

**Lemma 6.5.** *Under the premise of Lemma 6.3, it holds that*

$$\sum_{k=0}^{\infty} t_k \left( \sum_{i \in \mathcal{N}} \bar{P}_i^k \right) \leq -\frac{\delta}{2} \sum_{i \in \mathcal{N}} \sum_{k=0}^{\infty} \left( \frac{1}{\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right) + \Xi, \tag{6.5}$$

where  $\Xi \triangleq \left( \left\lceil \log_{1/\rho} \left( \max_{i \in \mathcal{N}} \frac{\bar{\tau}_i}{\tau_i} \right) \right\rceil + 1 \right) \sum_{i \in \mathcal{N}} \Xi_i$  such that for  $i \in \mathcal{N}$  with  $L_{g_i} > 0$ ,  $\Xi_i \triangleq 2 \left( L_{f_i} + \frac{\bar{\tau}_i}{c_\beta} B_i^2 L_{g_i}^2 \right) D_i^2 + 4 \frac{\bar{\tau}_i}{c_\alpha} C_{g_i}^2 B_i^2$  using  $B_i$  given in Assumption 4, and for  $i \in \mathcal{N}$  with  $L_{g_i} = 0$ ,  $\Xi_i = 2L_{f_i} D_i^2 + 4\frac{\bar{\tau}_i}{c_\alpha} C_{g_i}^2 B_i^2$  for  $B_i > 0$  given in Theorem 6.4.

*Proof.* The desired results immediately follows from the bound on the cardinality of  $\mathcal{S}$  given in Lemma 6.2 and the fact that  $t_k \leq t_0 = 1$  for all  $k \geq 0$ .  $\square$

**Theorem 6.6.** *Under the premise of Lemma 6.3. Let  $\{x_i^k, \theta_i^k\}_{k \geq 0}$  be the D-APDB iterate sequence for  $i \in \mathcal{N}$ , and  $\{s_i^k\}_{k \geq 0}$  denote the D-APDB auxiliary sequence with  $s_i^0 = 0$  for  $i \in \mathcal{N}$ , and define  $\{\lambda^k\}_{k \geq 0}$  such that  $\lambda^k = \mathbf{A}s^k$  with  $\mathbf{s}^k = [s_i^k]_{i \in \mathcal{N}}$  for all  $k \geq 0$ . For any  $\mathbf{x} \in \mathbf{dom} \phi$ ,  $\theta \in \mathcal{H}^* \cap \mathcal{B}$  and  $\lambda \in \mathbb{R}^{n|\mathcal{E}|}$ , let  $\mathbf{z} \triangleq (\mathbf{x}, \mathbf{y})$  with  $\mathbf{y} \triangleq (\theta, \lambda)$ . Then, for any  $K \geq 1$ , it holds that*

$$\begin{aligned} & \mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K) + \frac{u_K}{2T_K} \|\lambda - \lambda^K\|^2 \\ & + \frac{t_K}{2T_K} \sum_{i \in \mathcal{N}} \left( \frac{1 - (c_\alpha + c_\beta)}{\tau_i^K} \|x_i - x_i^K\|^2 + \frac{1}{\sigma_i^K} \|\theta_i - \theta_i^K\|^2 \right) \leq \frac{1}{T_K} \bar{\Gamma}(\mathbf{x}, \mathbf{y}), \end{aligned} \quad (6.6)$$

$$\bar{\Gamma}(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^0} \|x_i - x_i^0\|^2 + \frac{1}{2\sigma_i^0} \|\theta_i - \theta_i^0\|^2 \right) + \frac{1}{2\gamma^0} \|\lambda\|^2 + \Xi,$$

where  $u_K \triangleq t_{K-1} \left( \frac{1}{\gamma^{K-1}} - \sum_{i \in \mathcal{N}} \frac{d_i}{\zeta_i^{K-1}} \right) > 0$ ,  $T_K = \sum_{k=0}^{K-1} t_k$ , and  $(\bar{\mathbf{x}}^K, \bar{\mathbf{y}}^K) \triangleq \frac{1}{T_K} \sum_{k=0}^{K-1} t_k (\mathbf{x}^k, \mathbf{y}^k)$ .

*Proof.* Fix an arbitrary  $\mathbf{z} = (\mathbf{x}, \mathbf{y})$  such that  $\mathbf{x} \in \mathbf{dom} \phi$  and  $\theta \in \mathcal{H}^* \cap \mathcal{B}$ . Since the hypothesis of Lemma 4.3 holds for D-APDB parameters and  $t_0 = 1$ , it follows that

$$T_K (\mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K)) \leq Q^0(\mathbf{z}) - t_{K-1} R^K(\mathbf{z}) + \sum_{k=0}^{K-1} t_k P^k,$$

holds for any  $K \geq 1$ , where  $T_K = \sum_{k=0}^{K-1} t_k$  and  $(\bar{\mathbf{x}}^K, \bar{\mathbf{y}}^K) \triangleq \frac{1}{T_K} \sum_{k=0}^{K-1} t_k (\mathbf{x}^k, \mathbf{y}^k)$ . Therefore, Lemma 4.4 and Lemma 6.3 together imply that

$$\mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K) + \frac{t_{K-1}}{T_K} R^K(\mathbf{z}) \leq (Q^0(\mathbf{z}) + \Xi) / T_K, \quad (6.7)$$

where  $\Xi$  is defined in (6.5). Next, using the definitions of  $Q^0(\mathbf{z})$  in (4.5a) and  $R^K(\mathbf{z})$  in (4.5b) together with the fact that  $q^0 = 0$ ,  $\mathbf{x}^{-1} = \mathbf{x}^0$ ,  $\theta^{-1} = \theta^0$ ,  $\lambda^0 = 0$  and  $t_0 = 1$ , (6.7) implies that

$$\begin{aligned} \Gamma^K(\mathbf{z}) & \triangleq \mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K) + \frac{1}{2\gamma^{K-1}} \frac{t_{K-1}}{T_K} \|\lambda - \lambda^K\|^2 + \sum_{i \in \mathcal{N}} \frac{\zeta_i^K}{2} \frac{t_{K-1}}{T_K} \|x_i^K - x_i^{K-1}\|^2 \\ & + \frac{t_{K-1}}{2T_K} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^{K-1}} \|x_i - x_i^K\|^2 + \frac{1}{\sigma_i^{K-1}} \|\theta_i - \theta_i^K\|^2 + \frac{1}{\alpha_i^K} \|q_i^{K,\mathbf{y}}\|^2 + \frac{1}{\beta_i^K} \|q_i^{K,\mathbf{x}}\|^2 \right) \\ & + \frac{t_{K-1}}{T_K} \langle q^K, \mathbf{x} - \mathbf{x}^K \rangle + \frac{t_{K-1}}{T_K} \langle \mathbf{A}\mathbf{x}^K - \mathbf{A}\mathbf{x}^{K-1}, \lambda^K - \lambda \rangle \\ & \leq \frac{1}{2T_K} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^0} \|x_i - x_i^0\|^2 + \frac{1}{\sigma_i^0} \|\theta_i - \theta_i^0\|^2 \right) + \frac{1}{2\gamma^0 T_K} \|\lambda\|^2 + \frac{\Xi}{T_K}. \end{aligned} \quad (6.8)$$

Using the same arguments we employed while deriving (4.3) and (4.12), we can further lower bound the left hand side of (6.8) as below:

$$\begin{aligned}
 \Gamma^K(\mathbf{z}) &\geq \mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K) + \frac{1}{2\gamma^{K-1}} \frac{t_{K-1}}{T_K} \|\lambda - \lambda^K\|^2 + \frac{t_{K-1}}{2T_K} \sum_{i \in \mathcal{N}} \zeta_i^K \|x_i^K - x_i^{K-1}\|^2 \\
 &\quad + \frac{t_{K-1}}{2T_K} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^{K-1}} \|x_i - x_i^K\|^2 + \frac{1}{\sigma_i^{K-1}} \|\theta_i - \theta_i^K\|^2 + \frac{1}{\alpha_i^K} \|q_i^{K,\mathbf{y}}\|^2 + \frac{1}{\beta_i^K} \|q_i^{K,\mathbf{x}}\|^2 \right) \\
 &\quad - \frac{t_{K-1}}{2T_K} \sum_{i \in \mathcal{N}} \left( \frac{1}{\alpha_i^K} \|q_i^{K,\mathbf{y}}\|^2 + \frac{1}{\beta_i^K} \|q_i^{K,\mathbf{x}}\|^2 + (\alpha_i^K + \beta_i^K) \|x_i - x_i^K\|^2 \right) \\
 &\quad - \frac{t_{K-1}}{2T_K} \sum_{i \in \mathcal{N}} \left( \zeta_i^{K-1} \|x_i^K - x_i^{K-1}\|^2 + \frac{d_i}{\zeta_i^{K-1}} \|\lambda - \lambda^K\|^2 \right) \\
 &\geq \mathcal{L}(\bar{\mathbf{x}}^K, \mathbf{y}) - \mathcal{L}(\mathbf{x}, \bar{\mathbf{y}}^K) + \frac{t_{K-1}}{2T_K} \left( \frac{1}{\gamma^{K-1}} - \sum_{i \in \mathcal{N}} \frac{d_i}{\zeta_i^{K-1}} \right) \|\lambda - \lambda^K\|^2 \\
 &\quad + \frac{t_{K-1}}{2T_K} \sum_{i \in \mathcal{N}} \left( \frac{1 - (c_\alpha + c_\beta)}{\tau_i^{K-1}} \|x_i - x_i^K\|^2 + \frac{1}{\sigma_i^{K-1}} \|\theta_i - \theta_i^K\|^2 \right),
 \end{aligned}$$

where in the first inequality we used Young's inequality twice, and the second inequality is due to  $\zeta_i^K \geq \zeta_i^{K-1}$ , which follows from  $\tau_i^K \leq \tau_i^{K-1}$  for all  $i \in \mathcal{N}$ . Furthermore, since D-APDB parameter satisfy **Condition 1**, we have  $\gamma^{K-1} \leq \left( \sum_{i \in \mathcal{N}} d_i \left( \frac{2}{\alpha_i^K} + \frac{\eta^{K-1}}{\zeta_i^{K-1}} \right) \right)^{-1}$ , which together with  $\eta^{K-1} \geq 1$  implies that  $\frac{1}{\gamma^{K-1}} - \sum_{i \in \mathcal{N}} \frac{d_i}{\zeta_i^{K-1}} > 0$ , and  $\frac{1 - (c_\alpha + c_\beta)}{2\tau_i^{K-1}} > 0$  follows from  $1 - c > \delta$ ; hence, we can conclude using (4.13) to argue that  $t_{K-1}/\tau_i^{K-1} \geq t_K/\tau_i^K$  and  $t_{K-1}/\sigma_i^{K-1} \geq t_K/\sigma_i^K$  for  $i \in \mathcal{N}$ .  $\square$

**Corollary 6.7.** Let  $(x^*, \theta^*) \in \mathbb{R}^n \times \mathcal{X}^*$  denote an arbitrary primal-dual optimal pair for (1.1) as defined in Assumption 4 such that  $\|\theta_i^*\| \leq B_i/2$  for  $i \in \mathcal{N}$  with  $L_{g_i} > 0$ , and  $\lambda^* \in \mathbb{R}^{n|\mathcal{E}|}$  be an optimal dual variable corresponding to the consensus constraint  $\mathbf{A}\mathbf{x} = 0$  in (2.1). Under the premise of Lemma 6.3, initialized from arbitrary  $x_i^0 \in \text{dom } \phi_i$  and  $\theta_i^0 = 0$  for all  $i \in \mathcal{N}$ , the ergodic iterate sequences of D-APDB, i.e.,  $\{\bar{x}_i^k, \bar{\theta}_i^k\}_{k \geq 0}$  for all  $i \in \mathcal{N}$ , satisfy

$$(i) \text{ Suboptimality: } \quad \left| \sum_{i \in \mathcal{N}} \varphi_i(\bar{x}_i^K) - \varphi^* \right| \leq \frac{\Gamma_0}{T_K} = \mathcal{O}(1/K), \quad (6.9)$$

$$(ii) \text{ Infeasibility: } \quad \sum_{i \in \mathcal{N}} \|\theta_i^*\| d_{-\mathcal{X}_i}(g_i(\bar{x}_i^K)) + \|\lambda^*\| \|A\bar{\mathbf{x}}^K\| \leq \frac{\Gamma_0}{T_K} = \mathcal{O}(1/K), \quad (6.10)$$

where  $\Gamma_0 \triangleq \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^0} \|x^* - x_i^0\|^2 + \frac{2}{\sigma_i^0} \|\theta_i^*\|^2 \right) + \frac{2}{\gamma^0} \|\lambda^*\|^2 + \Xi$ ,  $\mathbf{x}^* \triangleq \mathbf{1}_N \otimes x^*$  and  $\mathbf{y}^* = (\theta^*, \lambda^*)$ .

*Proof.* Consider  $\mathbf{x}^* = \mathbf{1}_N \otimes x^*$  and note that  $(\mathbf{x}^*, \theta^*, \lambda^*) \in \text{dom } \phi \times (\mathcal{X}^* \cap \mathcal{B}) \times \mathbb{R}^{n|\mathcal{E}|}$  is a saddle point for  $\min_{\mathbf{x} \in \mathcal{X}} \max_{\theta, \lambda} \mathcal{L}(\mathbf{x}, \theta, \lambda)$ . Therefore, it follows from (6.6) that

$$\mathcal{L}(\bar{\mathbf{x}}^K, \theta^*, \lambda^*) - \mathcal{L}(\mathbf{x}^*, \bar{\theta}^K, \bar{\lambda}^K) \leq \bar{\Gamma}(\mathbf{x}^*, \mathbf{y}^*)/T_K.$$

Define  $\tilde{\mathbf{w}} \triangleq [\tilde{w}_i]_{i \in \mathcal{N}}$  such that  $\tilde{w}_i \triangleq g_i(\bar{x}_i^K) \in \mathbb{R}^{m_i}$  for all  $i \in \mathcal{N}$ . Since  $\mathcal{X}_i$  is a closed convex cone, it induces a decomposition on  $\mathbb{R}^{m_i}$ :

$$\tilde{w}_i = \tilde{w}_i^{(1)} + \tilde{w}_i^{(2)} \quad \text{s.t.} \quad \tilde{w}_i^{(1)} \triangleq \mathcal{P}_{-\mathcal{X}_i}(\tilde{w}_i), \quad \tilde{w}_i^{(2)} \triangleq \mathcal{P}_{\mathcal{X}_i^*}(\tilde{w}_i), \quad (6.11)$$

which also satisfy  $\langle \tilde{w}_i^{(1)}, \tilde{w}_i^{(2)} \rangle = 0$  for  $i \in \mathcal{N}$ . Thus, for all  $i \in \mathcal{N}$ ,  $\|\tilde{w}_i^{(2)}\| = \|\mathcal{P}_{-\mathcal{X}_i}(\tilde{w}_i) - \tilde{w}_i\| \triangleq d_{-\mathcal{X}_i}(\tilde{w}_i)$ . Next, we define  $\tilde{\theta} = [\tilde{\theta}_i]_{i \in \mathcal{N}} \in \mathcal{X}^* \cap \mathcal{B}$  such that

$$\tilde{\theta}_i \triangleq 2\|\theta_i^*\| \cdot \frac{\tilde{w}_i^{(2)}}{\|\tilde{w}_i^{(2)}\|} \in \mathcal{X}_i^* \cap \mathcal{B}_i, \quad \forall i \in \mathcal{N}.$$

Therefore, for each  $i \in \mathcal{N}$ , we get

$$\langle g_i(\bar{x}_i^K), \tilde{\theta}_i \rangle = 2\|\theta_i^*\| \cdot \left\langle \tilde{w}_i^{(1)} + \tilde{w}_i^{(2)}, \frac{1}{\|\tilde{w}_i^{(2)}\|} \tilde{w}_i^{(2)} \right\rangle = 2\|\theta_i^*\| d_{-\mathcal{X}_i}(\tilde{w}_i), \quad (6.12)$$

where the second equality follows from the orthogonality  $\tilde{w}_i^{(1)} \perp \tilde{w}_i^{(2)}$ .

Similarly, we define  $\tilde{\lambda} \triangleq 2\|\lambda^*\| \cdot \frac{A\bar{x}^K}{\|A\bar{x}^K\|}$ ; hence, from the definition, we get

$$\langle A\bar{x}^K, \tilde{\lambda} \rangle = 2\|\lambda^*\| \cdot \|A\bar{x}^K\|. \quad (6.13)$$

Note that  $\mathcal{L}(\mathbf{x}^*, \bar{\theta}^K, \bar{\lambda}^K) \leq \mathcal{L}(\mathbf{x}^*, \theta^*, \lambda^*) = \varphi^*$  since  $(\mathbf{x}^*, \theta^*, \lambda^*)$  is a saddle point of  $\mathcal{L}$ , and that  $\tilde{\theta} \in \mathcal{X}^* \cap \mathcal{B}$ . Thus (6.12) and (6.13) imply

$$\begin{aligned} \varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}^*) + 2\|\lambda^*\| \cdot \|A\bar{\mathbf{x}}^K\| + \sum_{i \in \mathcal{N}} 2\|\theta_i^*\| d_{-\mathcal{X}_i}(g_i(\bar{x}_i^K)) \\ = \mathcal{L}(\bar{\mathbf{x}}^k, \tilde{\theta}, \tilde{\lambda}) - \mathcal{L}(\mathbf{x}^*, \theta^*, \lambda^*) \leq \mathcal{L}(\bar{\mathbf{x}}^k, \tilde{\theta}, \tilde{\lambda}) - \mathcal{L}(\mathbf{x}^*, \bar{\theta}^K, \bar{\lambda}^K). \end{aligned} \quad (6.14)$$

Moreover, since  $\mathbf{x}^* \in \mathbf{dom} \phi$  and  $\tilde{\theta} \in \mathcal{X}^* \cap \mathcal{B}$ , (6.6) also implies that

$$\mathcal{L}(\bar{\mathbf{x}}^k, \tilde{\theta}, \tilde{\lambda}) - \mathcal{L}(\mathbf{x}^*, \bar{\theta}^K, \bar{\lambda}^K) \leq \bar{\Gamma}(\mathbf{x}^*, \tilde{\mathbf{y}})/T_K, \quad (6.15)$$

where  $\tilde{\mathbf{y}} = (\tilde{\theta}, \tilde{\lambda})$ , and (6.6) implies that

$$\bar{\Gamma}(\mathbf{x}^*, \tilde{\mathbf{y}}) = \sum_{i \in \mathcal{N}} \left( \frac{1}{2\tau_i^0} \|x^* - x_i^0\|^2 + \frac{1}{2\sigma_i^0} \|\tilde{\theta}_i\|^2 \right) + \frac{1}{2\gamma^0} \|\tilde{\lambda}\|^2 + \Xi = \Gamma_0. \quad (6.16)$$

Thus, combining (6.14), (6.15), and (6.16), we obtain

$$\varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}^*) + 2\|\lambda^*\| \cdot \|A\bar{\mathbf{x}}^K\| + \sum_{i \in \mathcal{N}} 2\|\theta_i^*\| d_{-\mathcal{X}_i}(g_i(\bar{x}_i^K)) \leq \Gamma_0/T_K. \quad (6.17)$$

Moreover, using the saddle point property of  $(\mathbf{x}^*, \theta^*, \lambda^*)$  one more time, we see that

$$\mathcal{L}(\bar{\mathbf{x}}^k, \theta^*, \lambda^*) - \mathcal{L}(\mathbf{x}^*, \theta^*, \lambda^*) \geq 0.$$

It follows that

$$\varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}^*) + \langle \lambda^*, A\bar{\mathbf{x}}^K \rangle + \sum_{i \in \mathcal{N}} \langle \theta_i^*, g_i(\bar{x}_i^K) \rangle \geq 0. \quad (6.18)$$

Recall that  $g_i(\bar{x}_i^K) = \tilde{w}_i$  such that (6.11) holds for  $i \in \mathcal{N}$ . Hence, we immediately obtain

$$\langle \theta_i^*, g_i(\bar{x}_i^K) \rangle = \langle \theta_i^*, \tilde{w}_i - \tilde{w}_i^{(1)} + \tilde{w}_i^{(1)} \rangle \leq \langle \theta_i^*, \tilde{w}_i - \tilde{w}_i^{(1)} \rangle \leq \|\theta_i^*\| d_{-\mathcal{X}_i}(g_i(\bar{x}_i^K)).$$

Thus (6.18) implies

$$\varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}^*) + \|\lambda^*\| \cdot \|A\bar{\mathbf{x}}^K\| + \sum_{i \in \mathcal{N}} \|\theta_i^*\| d_{-\mathcal{X}_i}(g_i(\bar{x}_i^K)) \geq 0. \quad (6.19)$$

Combining (6.17) and (6.19) immediately yields the desired result.  $\square$

**Corollary 6.8.** *Under the premise of Corollary 6.7, let  $\hat{t} \triangleq \rho \min_{i \in \mathcal{N}} \hat{\tau}_i / \bar{\tau}_i$ . Then, for any given  $\varepsilon > 0$ , the  $D$ -APDB ergodic iterate sequence  $\{\bar{x}_i^k\}_{k \geq 0}$  for  $i \in \mathcal{N}$  satisfies*

$$\left| \sum_{i \in \mathcal{N}} \varphi_i(\bar{x}_i^K) - \varphi^* \right| \leq \varepsilon, \quad \sum_{i \in \mathcal{N}} d_{-\mathcal{X}_i}(g_i(\bar{x}_i^K)) \|\theta_i^*\| + \|A\bar{\mathbf{x}}^K\| \|\lambda^*\| \leq \varepsilon, \quad \forall K \geq K_\varepsilon \triangleq \frac{\Gamma_0}{\hat{t}\varepsilon}. \quad (6.20)$$

Moreover, for each agent  $i \in \mathcal{N}$ , the total number of backtracking steps is upper bounded by  $\log_{1/\rho}(1/\hat{t})$ . Then, to compute an  $\varepsilon$ -optimal solution as in (6.20), the total number of gradient and projection evaluations required for each  $i \in \mathcal{N}$  is upper bounded by

$$\mathcal{O} \left( \log_{1/\rho} \left( \frac{1}{\hat{t}} \right) \cdot \frac{\Gamma_0}{\hat{t}} \cdot \frac{1}{\varepsilon} \right). \quad (6.21)$$

*Proof.* Due to Corollary 6.7, in order to achieve the complexity, we have that  $\frac{\Gamma_0}{T_K} \leq \varepsilon$ . Furthermore, according to Lemma 6.2, we have  $t_k \geq \hat{t} \triangleq \rho \min_{i \in \mathcal{N}} \hat{\tau}_i / \bar{\tau}_i$ ; therefore,  $T_K \geq \hat{t}K$ . Therefore, for all  $K \geq K_\varepsilon$ , we observe that  $\frac{\Gamma_0}{T_K} \leq \varepsilon$  is guaranteed to hold. Finally, combining the iteration complexity with Corollary 5.3 gives us the total gradient complexity for each node.  $\square$

**Theorem 6.9.** *Under the premise of Theorem 5.4, given some arbitrary initial points  $(x_i^0, \theta_i^0) \in \text{dom } \phi_i \times \mathcal{X}_i^*$  and the step size parameters  $\bar{\tau}_i, \zeta_i > 0$  for  $i \in \mathcal{N}$ , and the contraction coefficient  $\rho \in (0, 1)$ , let  $\{x_i^k, \theta_i^k\}_{k \geq 0}$  denote the  $D$ -APDB iterate sequence for  $i \in \mathcal{N}$ . Then, there exists  $(x^*, \theta^*)$  a primal-dual optimal solution to (1.1) such that  $\{(x_i^k, \theta_i^k)\}_{k \geq 0}$  converges to  $(x^*, \theta_i^*)$  for all  $i \in \mathcal{N}$ , i.e.,  $\lim_{k \rightarrow \infty} x_i^k = x^*$  and  $\lim_{k \rightarrow \infty} \theta_i^k = \theta_i^*$  for  $i \in \mathcal{N}$ .*

*Proof.* Let  $(\hat{\mathbf{x}}, \hat{\theta}, \hat{\lambda})$  be an arbitrary saddle point of  $\mathcal{L}$ . In the proof of Theorem 6.4 we show that  $\{a_k\}$  converges, where  $a_k \triangleq t_k Q^k(\hat{\mathbf{z}})$  for  $k \geq 0$ ; hence, it follows from (6.3) that  $\{(x_i^k, \theta_i^k)\}_{k \geq 0}$  for all  $i \in \mathcal{N}$  and  $\{\lambda^k\}_{k \geq 0}$  are bounded sequences. Let  $\{k_n\}_{n \geq 0} \subset \mathbb{Z}_+$  such that  $\lim_{n \rightarrow \infty} (x_i^{k_n}, \theta_i^{k_n}) = (x_i^*, \theta_i^*)$  for  $i \in \mathcal{N}$  and  $\lambda^{k_n} \rightarrow \lambda^*$ , where  $\mathbf{x}^* = [x_i^*]_{i \in \mathcal{N}}$  and  $\mathbf{y}^* = (\theta^*, \lambda^*)$  be an arbitrary limit point—here,  $\theta^* = [\theta_i^*]_{i \in \mathcal{N}}$ . Let  $\mathbf{y}^k = (\theta^k, \lambda^k)$  for  $k \geq 0$ . In the proof of Theorem 6.4, we also showed  $\sum_{k \geq 0} b_k < \infty$  where  $b_k \triangleq \frac{\delta}{2} \sum_{i \in \mathcal{N}} \left( \frac{1}{\tau_i^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{\sigma_i^k} \|\theta_i^{k+1} - \theta_i^k\|^2 \right) + \frac{\delta'}{2} \|\lambda^{k+1} - \lambda^k\|^2$  for  $k \geq 0$ . Since  $\tau_i^k \leq \bar{\tau}_i > 0$  and  $\sigma_i^k \leq \zeta_i \bar{\tau}_i > 0$  for all  $k \geq 0$  and  $i \in \mathcal{N}$ ,  $\sum_{k \geq 0} b_k < \infty$  implies that  $\sum_{k \geq 0} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 < \infty$ . Thus, for any given  $\varepsilon > 0$ , there exists  $N_1 \in \mathbb{Z}_+$  such that for  $n \geq N_1$ , one has  $\|\mathbf{x}^{k_n+1} - \mathbf{x}^{k_n}\|^2 + \|\mathbf{y}^{k_n+1} - \mathbf{y}^{k_n}\|^2 \leq \frac{\varepsilon}{4}$  and  $\|\mathbf{x}^{k_n-1} - \mathbf{x}^{k_n}\|^2 + \|\mathbf{y}^{k_n-1} - \mathbf{y}^{k_n}\|^2 \leq \frac{\varepsilon}{4}$ . Moreover, there also exists  $N_2 \in \mathbb{Z}_+$  such that for  $n \geq N_2$ , one has  $\|\mathbf{x}^{k_n} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k_n} - \mathbf{y}^*\|^2 \leq \frac{\varepsilon}{4}$ . Hence, for all  $n \geq \max\{N_1, N_2\}$ , it must hold that  $\|\mathbf{x}^{k_n+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k_n+1} - \mathbf{y}^*\|^2 \leq \varepsilon$  and  $\|\mathbf{x}^{k_n-1} - \mathbf{x}^*\|^2 + \|\mathbf{y}^{k_n-1} - \mathbf{y}^*\|^2 \leq \varepsilon$ , i.e.,

$$\lim_{n \rightarrow \infty} (x_i^{k_n}, y_i^{k_n}) = \lim_{n \rightarrow \infty} (x_i^{k_n+1}, y_i^{k_n+1}) = \lim_{n \rightarrow \infty} (x_i^{k_n-1}, y_i^{k_n-1}) = (x_i^*, y_i^*), \quad \forall i \in \mathcal{N}. \quad (6.22)$$

Next, we argue that  $(\mathbf{x}^*, \theta^*, \lambda^*)$  is indeed a saddle point of  $\mathcal{L}$ . First, recall that the sequence  $\{\lambda^k\}_{k \geq 0}$  defined as  $\lambda^k = A\mathbf{s}^k$  for  $\mathbf{s}^k = [s_i^k]_{i \in \mathcal{N}}$  satisfies

$$(\lambda^{k_n+1} - \lambda^{k_n}) / \gamma^{k_n} = A((1 + \eta^{k_n})\mathbf{x}^{k_n} - \eta^{k_n}\mathbf{x}^{k_n-1}) \quad (6.23)$$

$$\mathbf{p}^{k_n} = A^\top \left( (1 + \eta^{k_n})\lambda^{k_n} - \eta^{k_n}\lambda^{k_n-1} \right) + (1 + \eta^{k_n})\mathbf{JG}(\mathbf{x}^{k_n})^\top \theta^{k_n} - \eta^{k_n}\mathbf{JG}(\mathbf{x}^{k_n-1})^\top \theta^{k_n-1}, \quad (6.24)$$

which follows from  $\mathbf{s}^{k+1} = \mathbf{s}^k + \gamma^k(1 + \eta^k)\mathbf{x}^k - \gamma^k\eta^k\mathbf{x}^{k-1}$  in (3.9a) and (3.9c). Since  $\gamma^k = \frac{c\gamma}{\bar{\tau}} \left( \frac{2}{c\alpha} + \frac{\eta^k}{c\zeta} \right)^{-1}$  for some  $c\gamma \in \left( 0, \frac{1}{2|\mathcal{O}|} \right)$  and  $1 \leq \eta^k = \max_{i \in \mathcal{N}} \eta_i^k = \max_{i \in \mathcal{N}} \frac{\tau_i^{k-1}}{\bar{\tau}_i^k} \leq$

$\max_{i \in \mathcal{N}} \frac{\bar{\tau}_i}{\rho \bar{\tau}_i} \leq 1/\hat{\tau}$ ; hence,  $\frac{c_Y}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{1/\hat{\tau}}{c_\zeta} \right)^{-1} \leq \gamma^k \leq \frac{c_Y}{\bar{\tau}} \left( \frac{2}{c_\alpha} + \frac{1}{c_\zeta} \right)^{-1}$  for all  $k \geq 0$ . Therefore,  $(\lambda^{k_n+1} - \lambda^{k_n})/\gamma^{k_n} \rightarrow 0$  and  $\eta^{k_n}(\mathbf{x}^{k_n} - \mathbf{x}^{k_n-1}) \rightarrow 0$  as  $n \rightarrow \infty$ . Thus, (6.23) implies that

$$A\mathbf{x}^* = 0; \quad (6.25)$$

hence, there exists some  $x^* \in \mathbb{R}^n$  such that  $\mathbf{x}^* = \mathbf{1}_N \otimes x^*$ . Furthermore, (6.24) implies that  $\mathbf{p}^* \triangleq \lim_{n \rightarrow \infty} \mathbf{p}^{k_n} = A^\top \lambda^* + \mathbf{J}G(\mathbf{x}^*)^\top \theta^*$ .

Finally, consider the first-order optimality conditions for (3.9d) and (3.9e) along the subsequence  $\{k_n\}_{n \geq 0} \subset \mathbb{Z}_+$ :

$$\frac{1}{\tau_i^{k_n}} (x_i^{k_n} - x_i^{k_n+1}) \in \nabla f_i(x_i^{k_n}) + p_i^{k_n} + \partial \phi_i(x_i^{k_n+1}) \quad (6.26)$$

$$\left\langle \theta_i^{k_n+1} - \theta_i^{k_n} - \sigma_i^{k_n} g_i(x_i^{k_n+1}), \theta_i - \theta_i^{k_n+1} \right\rangle \geq 0, \quad \forall \theta_i \in \mathcal{K}_i^* \cap \mathcal{B}_i; \quad (6.27)$$

hence, because  $\bar{\tau}_i \geq \tau_i^k \geq \rho \min_{i \in \mathcal{N}} \hat{\tau}_i > 0$  and  $\sigma_i^k = \zeta_i \tau_i^k$  for some  $\zeta_i > 0$  for all  $k \geq 0$ , taking the limit on both sides of (6.26) and (6.27), and using [45, Theorem 24.4], we get

$$0 \in \nabla f_i(x^*) + p_i^* + \partial \phi_i(x^*), \quad \langle -g_i(x^*), \theta_i - \theta_i^* \rangle \geq 0, \quad \forall \theta_i \in \mathcal{K}_i^* \cap \mathcal{B}_i; \quad (6.28)$$

hence, (6.25) implies  $(\theta^*, \lambda^*) \in \mathbf{argmax}\{\mathcal{L}(\mathbf{x}^*, \theta, \lambda) : \theta \in \mathcal{K}^* \cap \mathcal{B}, \lambda\}$  and

$$0 \in \nabla f(\mathbf{x}^*) + \mathbf{J}G(\mathbf{x}^*)^\top \theta^* + A^\top \lambda^* + \partial \phi(\mathbf{x}^*), \quad (6.29)$$

which shows that  $\mathbf{z}^* \triangleq (\mathbf{x}^*, \theta^*, \lambda^*) \in \mathbf{dom} \phi \times \mathbf{dom} h \times \mathbb{R}^{n|\mathcal{E}|}$  is a saddle point of  $\mathcal{L}$  as well. Since (6.4) holds for any saddle point  $\hat{\mathbf{z}}$  of  $\mathcal{L}$ , we can replace  $\hat{\mathbf{z}}$  with  $\mathbf{z}^*$  and the same arguments we used in the proof of Theorem 6.4 continue to hold. Accordingly, we have  $\bar{a}_{k+1} \leq \bar{a}_k - b_k + c_k$  for  $k \geq 0$  where  $\{\bar{a}_k\}_{k \geq 0}$  is such that  $\bar{a}_k = t_k Q^k(\mathbf{z}^*) \geq 0$  for  $k \geq 0$ . Since  $\sum_{k \geq 0} c_k < \infty$ , we have  $\bar{a} \triangleq \lim_{k \rightarrow \infty} \bar{a}_k$  exists. Moreover, the subsequence  $\{k_n\}_{n \geq 0}$  defined above is independent of  $\mathbf{z}^*$  or  $\hat{\mathbf{z}}$ ; hence, it follows from the definition of  $Q^k(\mathbf{z}^*)$  in (4.5a) and (6.22) that  $\lim_{n \rightarrow \infty} a_{k_n} = t_{k_n} Q^{k_n}(\mathbf{z}^*) = 0$ . Since for a convergent sequence, every subsequence converges to the same limit point, we must have  $\bar{a} = \lim_{k \rightarrow \infty} \bar{a}_k = 0$ . Thus, the inequality in (6.3) with  $\hat{\mathbf{z}}$  replaced with  $\mathbf{z}^*$  and taking the limit as  $k \rightarrow \infty$  implies that  $\lim_{k \rightarrow \infty} \mathbf{x}^k = \mathbf{x}^*$ . Since we have already observed that  $\mathbf{x}^* = \mathbf{1}_N \otimes x^*$  for some  $x^* \in \mathbb{R}^n$ , we have  $\lim_{k \rightarrow \infty} x_i^k = x^*$  for all  $i \in \mathcal{N}$ . Furthermore,  $t_k \geq \hat{\tau} > 0$  for all  $k \geq 0$  and  $T_K \rightarrow \infty$  as  $K \rightarrow \infty$  implies that the weighted sequence  $\lim_{K \rightarrow \infty} \bar{x}_i^K = \lim_{K \rightarrow \infty} \bar{x}_i^K = x^*$ ; thus,  $A\mathbf{x}^* = 0$  implies that  $\lim_{K \rightarrow \infty} \|A\bar{\mathbf{x}}^K\| = 0$ . Finally, taking the limit on both sides of (6.9) and (6.10) as  $K \rightarrow \infty$  implies that  $\varphi^* = \sum_{i \in \mathcal{N}} \varphi_i(x^*)$  and  $d_{-\mathcal{K}_i}(g_i(x^*)) = 0$ , i.e.,  $-g_i(x^*) \in \mathcal{K}_i$ , for all  $i \in \mathcal{N}$ , which shows that  $x^*$  is an optimal solution to (1.1). On the other hand, setting  $\theta_i = 0$  within (6.28) implies that  $\langle -g_i(x^*), \theta_i^* \rangle \leq 0$  for  $i \in \mathcal{N}$ ; moreover,  $-g_i(x^*) \in \mathcal{K}_i$  and  $\theta_i^* \in \mathcal{K}^*$  imply that  $\langle -g_i(x^*), \theta_i^* \rangle \geq 0$ ; thus, the complementary slackness conditions  $\langle -g_i(x^*), \theta_i^* \rangle = 0$  hold for all  $i \in \mathcal{N}$ , which together with (6.29),  $\theta_i^* \in \mathcal{K}_i^*$  and  $-g_i(x^*) \in \mathcal{K}_i$  for all  $i \in \mathcal{N}$ , and  $A\mathbf{x}^* = 0$  together imply that  $\theta^* = [\theta_i^*]_{i \in \mathcal{N}}$  is an optimal dual solution.  $\square$

## 7. EXPERIMENTS

In this section, we provide numerical experiments to illustrate the performance of D-APDB and D-APDB0. The former algorithm can handle node-specific functional constraints defined by  $g_i(\cdot)$ ; whereas the latter one is tailored for problems that are either unconstrained or with simple constraint sets onto which projections are cheap to compute. We test D-APDB over  $\ell_1$ -norm-regularized quadratically constrained quadratic programming (QCQP) problems and on

primal support vector machine (SVM) training problems while D-APDB0 is tested over  $\ell_1$ -norm-regularized unconstrained quadratic programming problems. For all experiments, we test our algorithms on a random small-world network  $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ , i.e., we choose  $|\mathcal{N}|$  edges creating a random cycle over nodes, then the remaining  $|\mathcal{E}| - |\mathcal{N}|$  edges are selected uniformly at random. In our experiments, we set  $|\mathcal{E}| = 24$  edges and  $|\mathcal{N}| = 12$  nodes. Our experiments are run on a MacBook Air equipped with an Apple M2 CPU and 8 GB of unified memory. Our code is available at <https://github.com/Qiushui-Xu/D-APDB>.

**7.1. QCQP problems with  $\ell_1$ -norm regularization.** We consider a QCQP problem of the following form:

$$\varphi^* \triangleq \min_{x \in \mathcal{X}} \|x\|_1 + \sum_{i \in \mathcal{N}} \frac{1}{2} x^\top Q_i x \tag{7.1a}$$

$$\text{s.t. } g_i(x) \triangleq \frac{1}{2} (x - \bar{x}_i)^\top A_i (x - \bar{x}_i) \leq 1, \quad i \in \mathcal{N}, \tag{7.1b}$$

i.e., for  $i \in \mathcal{N}$ ,  $f_i(x) = \frac{1}{2} x^\top Q_i x$  for some randomly generated  $Q_i \succeq 0$ ,  $\mathcal{X} = [-10, 10]^n$  and  $\phi_i(x) = \frac{1}{N} \|x\|_1$ ; hence, the local objective  $\varphi_i = \phi_i + f_i$  is convex. For every agent  $i \in \mathcal{N}$ ,  $Q_i \in \mathbb{S}^n$  is generated randomly as follows:  $Q_i = V_i \Gamma_i V_i^\top$ , where  $V_i$  is a random orthonormal matrix satisfying  $V_i V_i^\top = I$ , and  $\Gamma_i = \text{diag}(\gamma_{i,1}, \gamma_{i,2}, \dots, \gamma_{i,n})$  is a diagonal matrix whose first element is  $\gamma_{i,1} = 5i$ , whose last two elements are  $\gamma_{i,n-1} = \gamma_{i,n} = 0$ , whose third-to-last element is  $\gamma_{i,n-2} = 1$ , and whose remaining elements are sampled from the uniform distribution  $\mathcal{U}[1, 5i]$  supported on the interval  $[1, 5i]$ . By construction, the elements of  $\Gamma_i$ , which coincide with the eigenvalues of  $Q_i$ , are sorted in a decreasing order. With this setup,  $Q_i$  has zero eigenvalues and therefore  $f_i(x)$  is convex (but not strongly convex). For generating matrices  $A_i$ , we follow a similar approach and set  $A_i = U_i R_i U_i^\top$  where  $U_i$  is a random orthonormal matrix such that  $U_i U_i^\top = I$  and  $R_i = \text{diag}(r_{i,1}, r_{i,2}, \dots, r_{i,n})$  is a diagonal matrix whose first element  $r_{i,1} = \frac{1}{4}$ , the last element is  $r_{i,n} = \frac{1}{16}$ , and all the other elements are sampled from the uniform distribution  $\mathcal{U}[\frac{1}{16}, \frac{1}{4}]$  and as before all the elements are sorted in a decreasing order. The constraints result in ellipsoid sets with a center at  $\bar{x}_i$ . We set its  $j$ -th coordinate to  $\bar{x}_i^j = 2 + \xi_i^j$ , where  $\xi_i^j$  is uniformly sampled from  $[-\frac{1}{2\sqrt{n}}, \frac{1}{2\sqrt{n}}]$ . In our experiments, we set  $n = 20$  and generate the initial point  $x_0$  such that its entries are i.i.d. with uniform distribution  $\mathcal{U}[-10, 10]$ .

For this class of randomly generated QCQP problems, we compare our D-APDB with D-APD in terms of log relative suboptimality, relative consensus error and infeasibility of the average iterate sequence  $\{\bar{x}^k\}_{k \geq 0} \subset \mathbb{R}^n$ , i.e., in Fig. 1, we plot  $\log(|\varphi(\bar{x}^k) - \varphi^*|/|\varphi^*| + 1)$  in (A),  $\sum_{i \in \mathcal{N}} \|x_i^k - \bar{x}^k\|^2 / (N \|\bar{x}^k\|^2)$  in (B) and  $\max_{i \in \mathcal{N}} \|(g_i(\bar{x}^k))_+\| / \max_{i \in \mathcal{N}} \|(g_i(\bar{x}^0))_+\|$  in (C), where  $\bar{x}^k = \sum_{i \in \mathcal{N}} x_i^k / N \in \mathbb{R}^n$  for  $k \geq 0$ . The benchmark algorithm D-APD is essentially our proposed method D-APDB without backtracking; that is, it uses a constant stepsize  $\bar{\tau}_i$  and sets  $\eta_i^k = 1$  for all  $k \geq 0$ , rather than performing backtracking on these parameters<sup>5</sup>. Comparing against this benchmark allows us to assess the benefits of incorporating backtracking.

For the benchmark algorithm D-APD which is based on constant stepsize, we set the initial stepsizes  $\bar{\tau}_i = \hat{\tau}_i$  where  $\hat{\tau}_i$  is as in (5.7),  $\zeta_i = 1$ ,  $c_\alpha, c_\beta, c_\zeta = 0.1$  and  $\eta^k = 1$  for  $k \geq 0$ . This stepsize requires knowledge of the Lipschitz constants  $L_{f_i}$  and constants  $L_{g_i}$ ; therefore, we compute these

<sup>5</sup>We note that D-APD is closely related to the method proposed in [19] for decentralized constrained strongly convex problems when the strong convexity parameter is set to zero.

constants using the randomly generated matrices  $A_i, Q_i$ <sup>6</sup>. For D-APDB, because it can potentially support larger stepsizes that adapt to the local curvature, we set the initial stepsizes to be larger than  $\hat{\tau}_i$ . More specifically, we set the initial stepsize of D-APDB (in Line 2 of D-APDB) as  $\bar{\tau}_i = \kappa \hat{\tau}_i$  with  $\kappa = 20$ , and we set the shrink factor  $\rho = 0.9$ ,  $\zeta_i = 1$ ,  $c_\alpha, c_\beta, c_\zeta = 0.1$ , and  $\delta = 0.1$ . All the other parameters are set according to our theoretical results. The results in Fig. 1 shows that our method outperforms D-APD on average over 20 simulations, each corresponding to a randomly generated problem instance, where the shaded areas demonstrate the standard deviation of the performance over the runs.

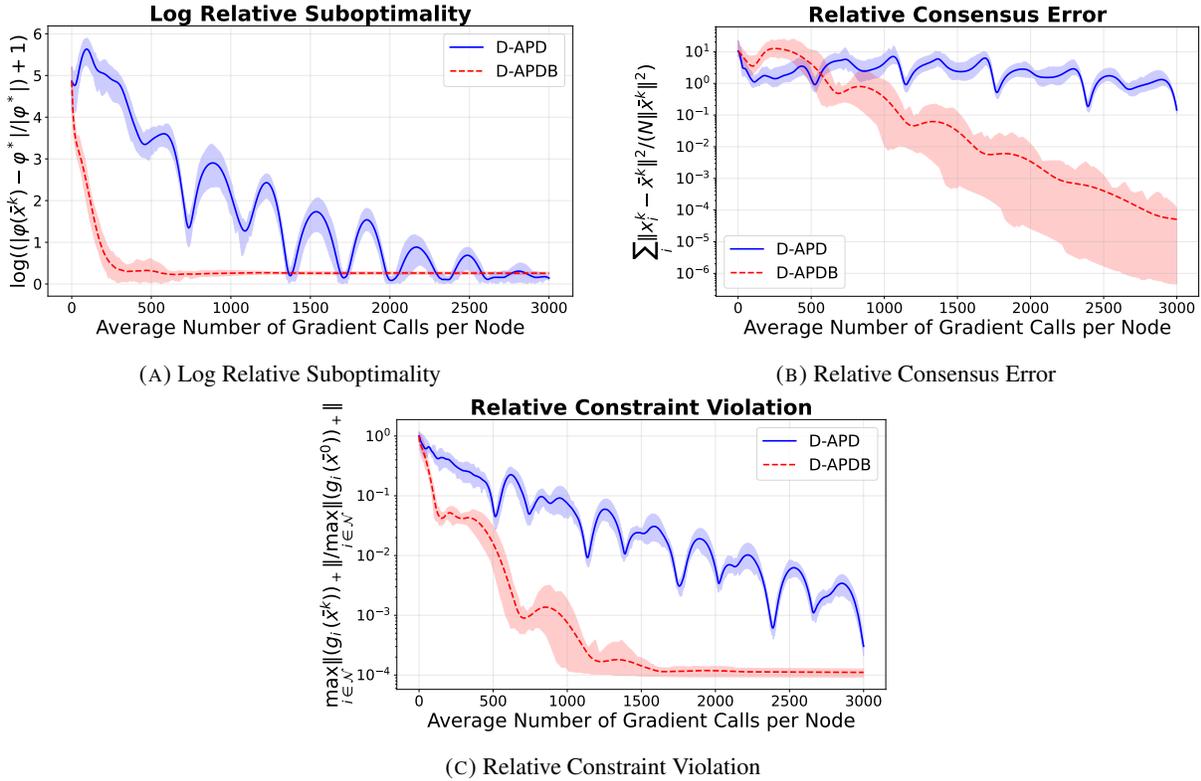


FIGURE 1. Comparison of D-APDB against D-APD for solving (7.1) over 20 runs. “Average number of gradient calls per node” means  $\frac{1}{N} \sum_{i=1}^N n_i^k$  where  $n_i^k$  counts how many calls for gradients for node  $i$  at iteration  $k$ .

**7.2. Unconstrained  $\ell_1$ -norm-regularized quadratic programming.** We consider an unconstrained quadratic programming problem of the following form

$$\varphi^* \triangleq \min_{x \in \mathcal{X}} \|x\|_1 + \sum_{i \in \mathcal{N}} \frac{1}{2} x^\top Q_i x + q_i^\top x + c_i, \quad (7.2)$$

i.e., for  $i \in \mathcal{N}$ ,  $f_i(x) = \frac{1}{2} x^\top Q_i x + q_i^\top x + c_i$  for some randomly generated  $Q_i \succeq 0$  and  $\phi_i(x) = \frac{1}{N} \|x\|_1$ ; hence, the local objective  $\varphi_i = \phi_i + f_i$  is convex.

Let  $\{L_{f_i}\}_{i \in \mathcal{N}}$  be i.i.d. random variables sampled from the normal distribution with mean 1000 and standard deviation 100. The matrix  $Q_i$  is generated randomly where we set  $Q_i =$

<sup>6</sup>Note  $\nabla f_i(x) = Q_i x$ ; hence, we can take  $L_{f_i} = \|Q_i\|_2$  and  $L_{g_i}$  can be computed similarly as  $L_{g_i} = \|A_i\|_2$ .

$V_i \Gamma_i V_i^\top$  such that  $V_i$  is a random orthonormal matrix satisfying  $V_i V_i^\top = I$  and  $\Gamma_i = \text{diag}(\gamma_{i,1}, \gamma_{i,2}, \dots, \gamma_{i,n})$  is a diagonal matrix with the first diagonal element  $\gamma_{i,1} = L_{f_i}$ , the last element  $\gamma_{i,n} = 0$ , and all the other elements are independently sampled from  $\mathcal{U}[0, \min\{100, L_{f_i}\}]$ . Moreover, every entry of  $q_i$  is sampled from standard normal distribution independently and  $c_i$  is sampled from  $\mathcal{U}[0, 1]$ . In this way, we can expect some significant variation among node-specific Lipschitz constants  $\{L_{f_i}\}_{i \in \mathcal{N}}$  as  $L_{f_i} = \|Q_i\|_2$  due to  $\nabla f_i(x) = Q_i x + q_i$  for  $i \in \mathcal{N}$ .

For this class of randomly generated unconstrained  $\ell_1$ -norm-regularized QP problems, we compare our D-APDBO with D-APD and global\_DATOS [14] in terms of log relative suboptimality, relative consensus error of the iterate sequence, i.e., in the two panels of Fig. 2, from left to right, we plot  $\log(|\varphi(\bar{x}^k) - \varphi^*|/|\varphi^*| + 1)$ ,  $\sum_{i=1}^N \|x_i^k - \bar{x}^k\|^2 / (N \|\bar{x}^k\|^2)$ . Benchmarking against these methods allows us to evaluate the benefits of incorporating node-specific backtracking and employing node-specific stepsize choices. global\_DATOS uses two communication rounds among the neighboring nodes per iteration while both D-APDBO and D-APD require one communication round per iteration. In Fig. 2 we plotted sub-optimality and consensus violation against the number of communication rounds in the x-axis.

In the experiments, for the benchmark algorithm D-APD, we set the initial stepsizes to  $\bar{\tau}_i = \frac{1}{2L_{f_i}}$  for each  $i \in \mathcal{N}$ , and choose  $c_\alpha = c_\zeta = 0.4$  and  $\eta_i^k = 1$  for all  $k \geq 0$ . For D-APDBO, because it can support larger stepsizes, we use a more aggressive initialization. Specifically, in Line 2 of D-APDBO, we set the initial stepsizes to  $\bar{\tau}_i = \kappa \hat{\tau}_i$  with  $\kappa = 5$ , and choose the shrink factor  $\rho = 0.9$ ,  $c_\alpha = c_\zeta = 0.4$ , and  $\delta = 0.1$ . For global\_DATOS, node-specific stepsizes are not allowed, but backtracking enables the use of larger global stepsizes. Therefore, we set the initial stepsize for all nodes to  $\bar{\alpha} = \kappa \hat{\alpha}$ , where  $\hat{\alpha} \triangleq \max_{i \in \mathcal{N}} \hat{\tau}_i$  and again choose  $\kappa = 5$  for a fair comparison between D-APDBO and global\_DATOS. We follow the definition of Laplacian-based constant edge weight matrix  $W = I - \frac{\Omega}{d_{\max} + 1}$  in [49] to generate the gossip matrix for global\_DATOS, where  $\Omega$  is the Laplacian matrix of the graph  $\mathcal{G}$  and  $d_{\max}$  is the max degree of the graph  $\mathcal{G}$ . All other parameters follow the theoretical prescriptions. The results in Fig. 2 show that our method consistently outperforms both D-APD and global\_DATOS on average for 20 simulations, where the shaded regions indicate the standard deviation across runs.

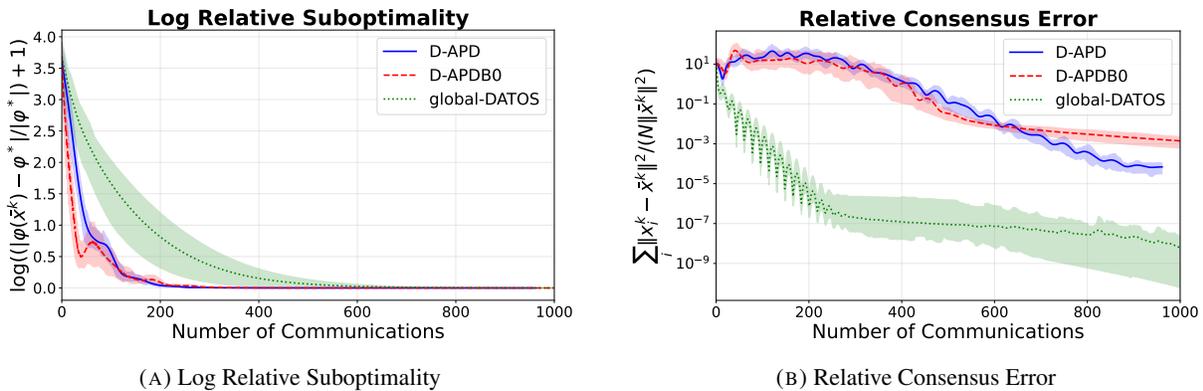


FIGURE 2. Comparison of D-APDBO against D-APD and global\_DATOS for solving (7.2) with 20 simulation.

**7.3. Primal linear SVM problems.** We consider a primal linear SVM problem with data distributed among computing nodes in  $\mathcal{N}$ . Let  $\mathcal{S} \triangleq \{1, 2, \dots, s\}$  denote the set of indices for

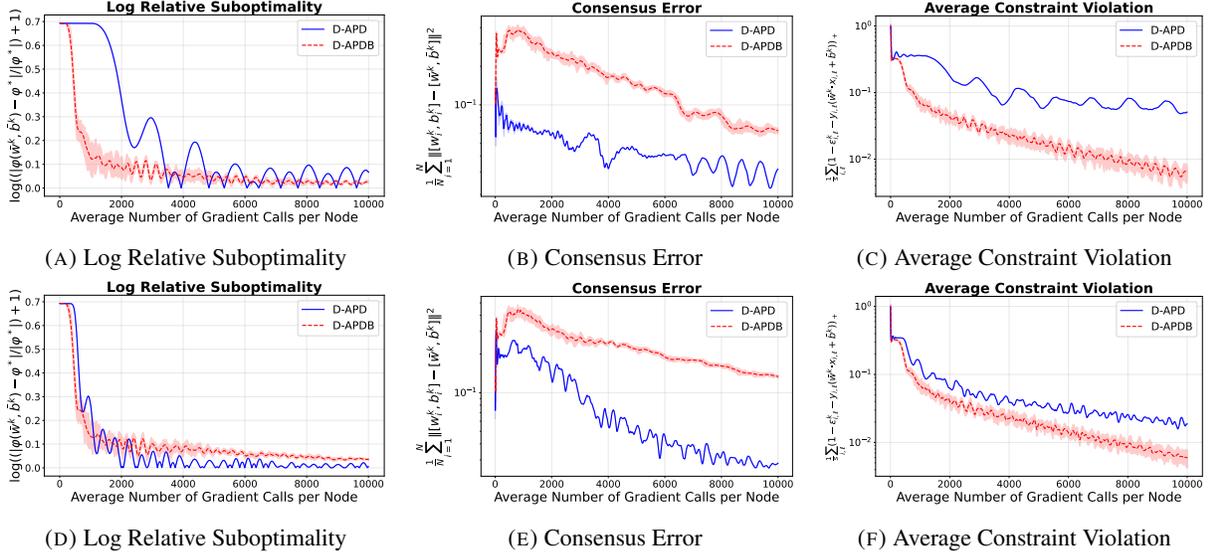


FIGURE 3. Comparison of D-APDB against D-APD for solving the linear SVM problem in (7.3) over 20 random problem instances. “Average number of gradient calls per node” means  $\frac{1}{N} \sum_{i=1}^N n_i^k$  where  $n_i^k$  counts how many calls for gradients for node  $i$  at iteration  $k$ .

training data and  $\mathcal{D} \triangleq \{(x_\ell, y_\ell) \in \mathbb{R}^n \times \{-1, +1\} : \ell \in \mathcal{S}\}$  be a set of feature vector and label pairs. Let  $\{\mathcal{S}_i\}_{i \in \mathcal{N}}$  be a partition of  $\mathcal{S}$  among the nodes  $\mathcal{N}$ .

Define  $\mathbf{w} = [w_i]_{i \in \mathcal{N}}$ ,  $\mathbf{b} = [b_i]_{i \in \mathcal{N}}$ , and  $\boldsymbol{\varepsilon} \in \mathbb{R}^s$  such that  $w_i \in \mathbb{R}^n$  and  $b_i \in \mathbb{R}$  for  $i \in \mathcal{N}$ . Consider the following distributed SVM problem:

$$\begin{aligned}
 \varphi(\mathbf{w}, \mathbf{b}) &\triangleq \min_{\mathbf{w}, \mathbf{b}, \boldsymbol{\varepsilon}} \frac{1}{2} \sum_{i \in \mathcal{N}} \|w_i\|^2 + NC \sum_{i \in \mathcal{N}} \sum_{\ell \in \mathcal{S}_i} \varepsilon_\ell \\
 \text{s.t.} \quad &y_\ell(w_i^\top x_\ell + b_i) \geq 1 - \varepsilon_\ell, \quad \varepsilon_\ell \geq 0, \quad \ell \in \mathcal{S}_i, \quad i \in \mathcal{N}, \\
 &w_i = w_j, \quad b_i = b_j, \quad (i, j) \in \mathcal{E}.
 \end{aligned} \tag{7.3}$$

Similar to [17],  $\{x_\ell\}_{\ell \in \mathcal{S}}$  is generated from two-dimensional multivariate Gaussian distribution with covariance matrix  $\Sigma = [1, 0; 0, 2]$  and with mean vector either  $m_1 = [-1, -1]^\top$  or  $m_2 = [1, 1]^\top$  with equal probability. The experiment was performed for  $s = 600$  training points such that for each  $i \in \mathcal{N}$ , we randomly sample 50 training data points from  $\mathcal{S}$ , i.e.,  $|\mathcal{S}_i| = 50$  for  $i \in \mathcal{N}$ . For this class of randomly generated primal linear SVM problems, we compare our D-APDB with D-APD in terms of log relative suboptimality, consensus error and average constraint violation of the iterate sequence, i.e., in Fig. 3, we plot  $\log(|\varphi(\bar{w}^k, \bar{b}^k) - \varphi^*|/|\varphi^*| + 1)$  in (A),  $\frac{1}{N} \sum_{i \in \mathcal{N}} \|[w_i^k, b_i^k] - [\bar{w}^k, \bar{b}^k]\|^2$  in (B) and  $\frac{1}{s} \sum_{i \in \mathcal{N}} \sum_{\ell \in \mathcal{S}_i} (1 - \varepsilon_\ell^k - y_\ell(\bar{w}^{k\top} x_\ell + \bar{b}^k))_+$  in (C), where  $\bar{w}^k = \sum_{i \in \mathcal{N}} w_i^k / N \in \mathbb{R}^n$  and  $\bar{b}^k = \sum_{i \in \mathcal{N}} b_i^k / N \in \mathbb{R}$  for  $k \geq 0$ .

For the benchmark algorithm D-APD, which uses constant stepsize, we set the step size  $\bar{\tau}_i = 0.005$  (the first row in Fig. 3) to mimic the scenario where one picks a conservative stepsize in case the Lipschitz constants are not readily available. For the case when one knows the Lipschitz constants, we set  $\bar{\tau}_i = \min \left\{ \frac{1 - (\delta + c)}{2L_{f_i}}, \frac{1}{C_{g_i}} \sqrt{\frac{c_\alpha(1 - \delta)}{2\zeta_i}} \right\}$  shown in the second row of Fig. 3. For both cases we set  $\zeta_i = 1$  (for dual step size choice),  $c_\alpha, c_\beta, c_\zeta = 0.1$  and  $\eta^k = 1$  for  $k \geq 0$ . For D-APDB,

because it can potentially support larger stepsizes, we set the initial stepsizes to be larger than  $\bar{\tau}_i$  choices for D-APD stated above. More specifically, we set the initial stepsize of D-APDB (in Line 2 of D-APDB) as 10 and 5 times larger than  $\bar{\tau}_i$  of D-APD for the first and second scenario, respectively. Moreover, we set the contraction factor  $\rho = 0.9$ ,  $\zeta_i = 1$ ,  $c_\alpha, c_\beta, c_\zeta = 0.1$ , and  $\delta = 0.1$  for both settings. All the other parameters are set according to our theoretical results. The results in Fig. 3 shows that our method outperforms D-APD on average over 20 randomly generated problem instances, where the shaded areas demonstrate the standard deviation of the performance over the runs.

## 8. CONCLUSION

In this work, we consider cooperative multi-agent constrained consensus optimization problems over an undirected network of agents. Agents have local objective functions with a composite structure—each is the sum of smooth and nonsmooth terms—and each agent is subject to its own private convex conic constraints. We consider the setting when only agents connected by an edge can directly communicate to exchange large-volume data vectors via a high-speed, short-range protocol (e.g., WiFi), and when the network supports one-hop, low-rate information exchange beyond immediate neighbors, as in the LoRaWAN protocol. In this setting, unlike existing decentralized primal–dual methods that require knowledge of Lipschitz constants, we propose a new algorithm, D-APDB, which adapts to local smoothness via a distributed backtracking step-size search. We obtain  $\mathcal{O}(1/K)$  convergence rate guarantees for suboptimality, infeasibility and consensus violation. When nodes have private constraints for which projections are expensive to compute, D-APDB is, to the best of our knowledge, the first distributed method with backtracking that attains the optimal convergence rate for constrained composite convex optimization problems. We also provide a variant of our algorithm, D-APDB0, for problems that are either unconstrained or with simple constraint sets for which projections are cheap to compute. D-APDB0 can also achieve the same  $\mathcal{O}(1/K)$  rate for suboptimality and consensus violation. Finally, we present numerical experiments demonstrating the performance of D-APDB0 and D-APDB on distributed  $\ell_1$ -norm-regularized QP and QCQP problems, and distributed primal SVM training.

## Acknowledgements

This paper builds on momentum acceleration, and we dedicate it to Prof. Yurii Nesterov (UCLouvain, Belgium), whose seminal contributions to the development and analysis of optimization algorithms have been profoundly influential in this field. This research was supported in part by the Office of Naval Research under award number N00014-24-1-2628 and N00014-24-1-2666.

## REFERENCES

- [1] R. Aldana-López, A. Macchelli, G. Notarstefano, R. Aragüés, C. Sagüés, Towards parameter-free distributed optimization: A port-Hamiltonian approach, arXiv preprint arXiv:2404.13529, 2024.
- [2] Y. Arjevani, J. Bruna, B. Can, M. Gurbuzbalaban, S. Jegelka, H. Lin, IDEAL: Inexact decentralized accelerated augmented Lagrangian method, In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, (ed.), *Advances in Neural Information Processing Systems*, vol. 33, pp. 20648–20659, Curran Associates, Inc., 2020.

- [3] A. Askhedkar, B. Chaudhari, M. Zennaro, Hardware and software platforms for low-power wide-area networks, In: LPWAN technologies for IoT and M2M applications, Elsevier, pp. 397–407, 2020.
- [4] N.S. Aybat, Z. Wang, G. Iyengar, An asynchronous distributed proximal gradient method for composite convex optimization, In: International Conference on Machine Learning, pp. 2454–2462, 2015.
- [5] N.S. Aybat, E.Y. Hamedani, A primal-dual method for conic constrained distributed optimization problems. In *Advances in Neural Information Processing Systems*, pages 5049–5057, 2016.
- [6] N.S. Aybat, E.Y. Hamedani, A distributed ADMM-like method for resource sharing over time-varying networks, *SIAM J. Optim.* 29 (2019) 3036–3068.
- [7] N.S. Aybat, Z. Wang, T. Lin, S. Ma, Distributed linearized alternating direction method of multipliers for composite convex consensus optimization, *IEEE Trans. Auto. Control*, 63 (2017) 5–20.
- [8] J. Barzilai, J.M. Borwein, Two-point step size gradient methods, *IMA J. Numer. Anal.* 8 (1088), 141–148.
- [9] S.P. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (2011) 1–122.
- [10] O. Burdakov, Y. Dai, N. Huang, Stabilized Barzilai-Borwein method, *J. Comput. Math.* 37 (2019), 916–936.
- [11] A. Chambolle, T. Pock, On the ergodic convergence rates of a first-order primal–dual algorithm, *Math. Program.* 159 (2016) 253–287.
- [12] X. Chen, B. Karimi, W. Zhao, P. Li, On the convergence of decentralized adaptive gradient methods, In: E. Khan, M. Gonen, (ed.) *Proceedings of The 14th Asian Conference on Machine Learning (ACML)*, vol. 189, pp. 217–232, PMLR, 2023.
- [13] X. Chen, X. Li, P. Li, Toward communication efficient adaptive gradient method, In: *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pp. 119–128, 2020.
- [14] X. Chen, I. Kuruzov, G. Scutari, A. Gasnikov, A parameter-free decentralized algorithm for composite convex optimization, arXiv preprint arXiv:2508.01466, 2025.
- [15] E. Dall’Anese, H. Zhu, G.B. Giannakis, Distributed optimal power flow for smart microgrids, *IEEE Trans. Smart Grid*, 4 (2013), 1464–1475.
- [16] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.* 12 (2011) 2121–2159.
- [17] P.A. Forero, A. Cano, G.B. Giannakis, Consensus-based distributed linear support vector machines, In: *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 35–46, 2010.
- [18] D. Ghaderyan, N.S. Aybat, A.P. Aguiar, F.L. Pereira, A fast row-stochastic decentralized method for distributed optimization over directed graphs, *IEEE Trans. Auto. Control*, 69 (2023), 275–289.
- [19] E.Y. Hamedani, N.S. Aybat, A decentralized primal-dual method for constrained minimization of a strongly convex function, *IEEE Trans. Auto. Control*, 67 (2021), 5682–5697.
- [20] E.Y. Hamedani, N.S. Aybat, A primal-dual algorithm with line search for general convex-concave saddle point problems, *SIAM J. Optim.* 31 (2021), 1299–1329.
- [21] D. Jakovetić, A unification and generalization of exact distributed first-order methods, *IEEE Trans. Auto. Control*, 63 (2018), 1115–1129.
- [22] D. Jakovetić, J.M.F. Xavier, J.M. F. Moura, Fast distributed gradient methods, *IEEE Trans. Auto. Control*, 59 (2011), 1131–1146.
- [23] T. Janssen, N. BuiLam, M. Aernouts, R. Berkvens, M. Weyn, Lora 2.4 ghz communication link and range, *Sensors*, 20 (2020), 4366.
- [24] R. Jiang, A. Mokhtari, Generalized optimistic methods for convex-concave saddle point problems, *SIAM J. Optim.* 35 (2025), 2066–2097.
- [25] D.H. Kim, J.Y. Lim, J.D. Kim, Low-power, long-range, high-data transmission using wi-fi and lora, In: 2016 6th international conference on IT convergence and security (ICITCS), pp. 1–3, IEEE, 2016.
- [26] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, arXiv:1412.6980, 2015.
- [27] J. Koshal, A. Nedić, U.V. Shanbhag, Multiuser optimization: Distributed algorithms and error analysis, *SIAM J. Optim.* 21 (2011), 1046–1081.
- [28] I. Kuruzov, X. Chen, G. Scutari, A. Gasnikov, Adaptive stepsize selection in decentralized convex optimization, arXiv preprint arXiv:2507.23725, 2025.

- [29] I. Kuruzov, G. Scutari, A. Gasnikov, Achieving linear convergence with parameter-free algorithms in decentralized optimization, *Adv. Neural Info. Process. Sys.* 37 (2024), 96011–96044.
- [30] S. Lee, A. Nedic, Distributed random projection algorithm for convex optimization, *EEE J. Selected Topics Signal Process.* 7 (2013), 221–229.
- [31] J. Li, X. Chen, S. Ma, M. Hong, Problem-parameter-free decentralized nonconvex stochastic optimization, arXiv preprint arXiv:2402.08821, 2024.
- [32] Z. Li, W. Shi, M. Yan, A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates, *IEEE Trans. Signal Process.* 67 (2019), 4494–4506.
- [33] S. Lu, K. Zhang, T. Chen, T. Başar, L. Horesh, Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning, In: *AAAI Conference on Artificial Intelligence*, 2021.
- [34] L. Luo, Y. Xiong, Y. Liu, X. Sun, Adaptive gradient methods with dynamic bound of learning rate, In: *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, 2019.
- [35] Y. Malitsky, K. Mishchenko, Adaptive gradient descent without descent, In: *International Conference on Machine Learning*, pp. 6702–6712, PMLR, 2020.
- [36] Y. Malitsky, K. Mishchenko, Adaptive proximal gradient method for convex optimization, *Adv. Neural Info. Process. Sys.* 37 (2024), 100670–100697.
- [37] P. Nazari, D.A. Tarzanagh, G. Michailidis, Dadam: A consensus-based distributed adaptive gradient method for online optimization, *IEEE Trans. Signal Process.* 70 (2022), 6065–6079.
- [38] A. Nedic, A. Ozdaglar, P.A. Parrilo, Constrained consensus and optimization in multi-agent networks, *IEEE Trans. Auto. Control*, 55 (2010), 922–938.
- [39] A. Nedić, A.E. Ozdaglar, Distributed subgradient methods for multi-agent optimization, *IEEE Trans. Auto. Control*, 54 (2009), 48–61.
- [40] A. Nedich, Convergence rate of distributed averaging dynamics and optimization in networks, *Found. Trends Signal Process.* 2 (2015), 1–100.
- [41] S. Pu, A. Nedić, Distributed stochastic gradient tracking methods, *Math. Program.* 182 (2020), 95–129.
- [42] M. Rabbat, R. Nowak, Distributed optimization in sensor networks, In: *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, pp. 20–27, 2004.
- [43] S.J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, H.B. McMahan, Adaptive federated optimization, In: *International Conference on Learning Representations*, 2021.
- [44] S.J. Reddi, S. Kale, S. Kumar, On the convergence of adam and beyond, In: *International Conference on Learning Representations (ICLR)*, 2018.
- [45] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1997.
- [46] A.H. Sayed, Adaptation, learning, and optimization over networks, *Foundations and Trends® in Machine Learning*, 7 (2014), 311–801.
- [47] K. Scaman, F. Bach, S. Bubeck, Y.T. Lee, L. Massoulié, Optimal algorithms for smooth and strongly convex distributed optimization in networks, In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 3027–3036, PMLR, 2017.
- [48] F. Schäfer, A. Anandkumar, Competitive gradient descent, In: *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [49] W. Shi, Q. Ling, G. Wu, W. Yin, Extra: An exact first-order algorithm for decentralized consensus optimization, *SIAM J. Optim.* 25 (2015), 944–966.
- [50] W. Shi, Q. Ling, G. Wu, W. Yin, A proximal gradient algorithm for decentralized composite optimization, *IEEE Trans. Signal Process.* 63 (2015), 6013–6023.
- [51] T. Tieleman, G. Hinton, Lecture 6.5–RMSProp: Divide the gradient by a running average of its recent magnitude, Coursera: *Neural Networks for Machine Learning*, 2012. Lecture 6.5.
- [52] C. Xie, O. Koyejo, I. Gupta, H. Lin, Local AdaAlter: Communication-efficient stochastic gradient descent with adaptive learning rates, arXiv preprint arXiv:1911.09030, 2020.
- [53] Q. Xu, X. Zhang, N.S. Aybat, M. Gürbüzbalaban, A stochastic GDA method with backtracking for solving nonconvex (strongly) concave minimax problems, arXiv preprint arXiv:2403.07806, 2024.
- [54] H. Ye, Z. Zhou, L. Luo, T. Zhang, Decentralized accelerated proximal gradient descent, In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, (ed.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 18308–18317, Curran Associates, Inc., 2020.

- [55] X. Zhang, Q. Xu, N.S. Aybat, AGDA+: Proximal alternating gradient descent ascent method with a non-monotone adaptive step-size search for nonconvex minimax problems, arXiv:2406.14371, 2024.
- [56] Y. Zheng, N. Loizou, J. You, E. Mallada, Dissipative gradient descent ascent method, arXiv preprint arXiv:2403.09090, 2024.
- [57] D. Zhou, S. Ma, J. Yang, AdaBB: Adaptive Barzilai-Borwein method for convex optimization, arXiv preprint arXiv:2401.08024, 2024.