Multi-agent constrained optimization of a strongly convex function over time-varying directed networks

Erfan Yazdandoost Hamedani∗, and Necdet Serhat Aybat∗

∗Industrial & Manufacturing Engineering Department, The Pennsylvania State University, PA, USA.
Emails: evy5047@psu.edu, nsa10@psu.edu

Abstract

We consider cooperative multi-agent consensus optimization problems over both static and time-varying communication networks, where only local communications are allowed. The objective is to minimize the sum of agent-specific possibly non-smooth composite convex functions over agent-specific private conic constraint sets; hence, the optimal consensus decision should lie in the intersection of these private sets. Assuming the sum function is strongly convex, we provide convergence rates in sub-optimality, infeasibility and consensus violation; examine the effect of underlying network topology on the convergence rates of the proposed decentralized algorithms.

I. INTRODUCTION

Decentralized optimization over communication networks has various applications: i) distributed parameter estimation in wireless sensor networks [1], [2]; ii) multi-agent cooperative control and coordination in multirobot networks [3], [4]; iii) distributed spectrum sensing in cognitive radio networks [5], [6]; iv) processing distributed big-data in (online) machine learning [7], [8], [9], [10], [11]; v) power control problem in cellular networks [12], to name a few application areas. In many of these network applications the communication network may be directed, i.e., communication links can be unidirectional, and/or the network in the wireless setting may be time-varying, e.g., communication links can be on/off over time due to failures or the links may exist among agents depending on their inter-distances. In the context of decentralized optimization, time-varying directed networks can also arise in wired networks as uni-directional asynchronous protocols are desired over bi-directional communication protocols which create deadlocks due to lack of enforcement rule to block a third node when the other two neighbors
are exchanging local variables between themselves [7]. In majority of the applications discussed above, other than the topology being time-invariant (static) or time-varying, or the network having undirected or directed links, one common characteristic shared by today’s big-data networks is that the network size is usually prohibitively large for centralized optimization, which requires a fusion center that collects the physically distributed data and runs a centralized optimization method. This process has expensive communication overhead, requires large enough memory to store and process the data, and also may violate data privacy in case agent are not willing to share their data even though they are collaborative agents [13], [14].

In this paper, from a broader perspective, we aim to study constrained distributed optimization of a strongly convex function over static or time-varying communication networks $G^t = (\mathcal{N}, \mathcal{E}^t)$ for $t \geq 0$; in particular, from an application perspective, we are motivated to design an efficient decentralized solution method for constrained LASSO (C-LASSO) problems [15] with distributed data. C-LASSO, having the generic form $\min_x \{ \lambda \|x\|_1 + \|Cx - d\|_2^2 : Ax \leq b \}$, is an important class of problems in statistics, which includes fused LASSO, constrained regression, and generalized LASSO problems as its special cases [16], [15], [17] to name a few. In the rest, we provide our results for a more general setting of constrained decentralized optimization. We assume that i) each node $i \in \mathcal{N}$ has a local conic convex constraint set $\chi_i$, for which projections are not easy to compute, and a local convex objective function $\varphi_i$ (possibly non-smooth) such that $\sum_{i \in \mathcal{N}} \varphi_i(x)$ is strongly convex, and ii) nodes are willing to collaborate, without sharing their private data defining $\chi_i$ and $\varphi_i$, to compute an optimal consensus decision minimizing the sum of local functions and satisfying all local constraints; moreover, iii) nodes are only allowed to communicate with the neighboring nodes over the links in the network. Although we assume that $\sum_{i \in \mathcal{N}} \varphi_i(x)$ is strongly convex, it is possible that none of the local functions $\{\varphi_i\}_{i \in \mathcal{N}}$ are strongly convex. This kind of structure arises in LASSO problems; in particular, let $\varphi_i : \mathbb{R}^n \to \mathbb{R}$ such that $\varphi_i(x) = \lambda \|x\|_1 + \|C_i x - d_i\|_2^2$ for $C_i \in \mathbb{R}^{m_i \times n}$ and $d_i \in \mathbb{R}^{m_i}$ for $i \in \mathcal{N}$. Note that while $\varphi_i$ is merely convex for all $i \in \mathcal{N}$, $\sum_{i \in \mathcal{N}} \varphi_i(x)$ is strongly convex when $m_i < n$ for $i \in \mathcal{N}$ and $\text{rank}(C) = n \leq \sum_{i \in \mathcal{N}} m_i \equiv m$ where $C = [C_i]_{i \in \mathcal{N}} \in \mathbb{R}^{m \times n}$. Therefore, it is important to note that in the centralized formulation of this problem $\min_x \sum_{i \in \mathcal{N}} \varphi_i(x)$, the objective is strongly convex; however, in the decentralized formulation, this is not the case where we minimize $\sum_{i \in \mathcal{N}} \varphi_i(x_i)$ while imposing consensus among local variables $\{x_i\}_{i \in \mathcal{N}}$. In the numerical section, we considered a distributed C-LASSO problem under a similar strong convexity setting.

June 27, 2017
Many of the real-life application problems discussed above are special cases of the generic conic constrained decentralized optimization framework discussed in this paper. With the motivation of designing an efficient decentralized solution method for the distributed conic constrained problem over static or time-varying communication networks, as we briefly described above, we propose distributed primal-dual algorithms: DPDA for static and DPDA-TV for time-varying communication networks. DPDA and DPDA-TV are both based on the primal-dual algorithm (PDA), recently proposed in [18] for convex-concave saddle-point problems which for sake of completeness will be discussed in detail in Section I-A.

**Problem Description.** Let \( \{G^t\}_{t \in \mathbb{R}_+} \) denote a time-varying graph of \( N \) computing nodes. More precisely, for all \( t \geq 0 \), the graph has the form \( G^t = (\mathcal{N}, \mathcal{E}^t) \), where \( \mathcal{N} \triangleq \{1, \ldots, N\} \) is the set of nodes and \( \mathcal{E}^t \subseteq \mathcal{N} \times \mathcal{N} \) is the set of (possibly directed) edges at time \( t \). Suppose that each node \( i \in \mathcal{N} \) has a private (local) cost function \( \varphi_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) such that
\[
\varphi_i(x) \triangleq \rho_i(x) + f_i(x),
\]
where \( \rho_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is a possibly non-smooth convex function, and \( f_i : \mathbb{R}^n \to \mathbb{R} \) is a smooth convex function. We assume \( f_i \) is differentiable on an open set containing \( \text{dom} \rho_i \) with a Lipschitz continuous gradient \( \nabla f_i \), of which Lipschitz constant is \( L_i \); and the prox map of \( \rho_i \),
\[
\text{prox}_{\rho_i}(x) \triangleq \arg\min_{y \in \mathbb{R}^n} \{ \rho_i(y) + \frac{1}{2} \|y - x\|^2 \},
\]
is efficiently computable for \( i \in \mathcal{N} \), where \( \|\cdot\| \) denotes the Euclidean norm. Consider the following minimization problem:
\[
x^* \in \arg\min_{x \in \mathbb{R}^n} \bar{\varphi}(x) \triangleq \sum_{i \in \mathcal{N}} \varphi_i(x) \quad \text{s.t.} \quad A_i x - b_i \in K_i, \quad \forall i \in \mathcal{N},
\]
where \( A_i \in \mathbb{R}^{m_i \times n} \), \( b_i \in \mathbb{R}^{m_i} \) and \( K_i \subseteq R^{m_i} \) is a closed, convex cone. Suppose that projections onto \( K_i \) can be computed efficiently, while the projection onto the preimage \( \chi_i \triangleq A_i^{-1}(K_i + b_i) \) is assumed to be impractical, e.g., when \( K_i \) is the positive semidefinite cone, projection to preimage requires solving an SDP.

**Assumption I.1.** The duality gap for (3) is zero, and a primal-dual solution to (3) exists.

A sufficient condition is the existence of a Slater point, i.e., there exists \( \bar{x} \in \text{relint(dom } \bar{\varphi} \text{) such that } A_i \bar{x} - b_i \in \text{int}(K_i) \) for \( i \in \mathcal{N} \), where \( \text{dom } \bar{\varphi} = \bigcap_{i \in \mathcal{N}} \text{dom } \varphi_i. \)
Definition 1. A differentiable function \( f : \mathbb{R}^n \to \mathbb{R} \) is strongly convex with modulus \( \mu > 0 \) if the following inequality holds

\[
f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\mu}{2} \|x - \bar{x}\|^2 \quad \forall x, \bar{x} \in \mathbb{R}^n.
\]

Assumption I.2. Suppose \( \tilde{f}(x) \triangleq \sum_{i \in \mathcal{N}} f_i(x) \) is strongly convex with modulus \( \bar{\mu} > 0 \); and each \( f_i \) is strongly convex with modulus \( \mu_i \geq 0 \) for \( i \in \mathcal{N} \), and define \( \bar{\mu} \triangleq \min_{i \in \mathcal{N}} \{ \mu_i \} \geq 0 \).

Remark I.1. Clearly \( \bar{\mu} \geq \sum_{i \in \mathcal{N}} \mu_i \) is always true, and it is possible that \( \mu_i = 0 \) for all \( i \in \mathcal{N} \) but still \( \bar{\mu} > 0 \); moreover, \( \bar{\mu} > 0 \) implies that \( x^* \) is the unique optimal solution to (3).

Previous Work. Consider \( \min_{x \in \mathbb{R}^n} \{ \varphi(x) : x \in \cap_{i \in \mathcal{N}} \chi_i \} \) over a communication network of computing agents \( \mathcal{N} \), where \( \varphi(x) = \sum_{i \in \mathcal{N}} \varphi_i(x) \). Although the unconstrained consensus optimization, i.e., \( \chi_i = \mathbb{R}^n \), is well studied for static or time-varying networks – see \cite{19}, \cite{20}, \cite{21}, \cite{22}, \cite{23}, \cite{24}, \cite{25}, \cite{26}, \cite{27}, \cite{28} and the references therein, the constrained case is still an area of active research, e.g., \cite{19}, \cite{20}, \cite{21}, \cite{22}, \cite{23}, \cite{24}, \cite{25}, \cite{26}, \cite{27}, \cite{28}. Our focus in this paper is on the case where \( \varphi \) is strongly convex such that each \( \varphi_i = \rho_i + f_i \) is composite convex, and \( \chi_i \) has the form \( \mathcal{A}_i^{-1}(K_i + b_i) \) for \( i \in \mathcal{N} \). In this section, we briefly review the existing work related to our setup.

Unconstrained minimization of a strongly convex objective function \( \tilde{f}(x) \triangleq \sum_{i \in \mathcal{N}} f_i(x) \) in the multi-agent setting has been investigated in many papers, e.g., \cite{29}, \cite{30}, \cite{31}, \cite{32}, \cite{33} considered static communication networks \( \mathcal{G} = (\mathcal{N}, \mathcal{E}) \) while \cite{34}, \cite{35} studied the time-varying networks. In the rest, suppose that \( \mu_i \geq 0 \) denotes the convexity modulus of \( f_i \) for \( i \in \mathcal{N} \). In \cite{29}, Makhdoumi and Ozdaglar proposed a distributed ADMM to solve \( \min_x \tilde{f}(x) \) over a time-invariant (static), undirected network; they show that when \( f_i \) has Lipschitz continuous gradient with constant \( L_i \) and when \( \mu_i > 0 \) for each \( i \in \mathcal{N} \), the local iterates at all nodes are within an \( \epsilon \)-ball of the optimal solution after at most \( O(\sqrt{\kappa} \log(1/\epsilon)) \) iterations, where \( \kappa = L_{\max}/\mu, \ L_{\max} \triangleq \max_{i \in \mathcal{N}} L_i \) and \( \mu \triangleq \min_{i \in \mathcal{N}} \{ \mu_i \} \); on the other hand, since each iteration requires exact minimization of an augmented function involving \( f_i \) at each \( i \in \mathcal{N} \), iterations can be very costly depending on \( f_i \). In \cite{36}, Chang et al. considered the composite convex minimization problem, \( \min_x \sum_{i \in \mathcal{N}} \rho_i(x) + f_i(C_i x) \), over a static undirected network \( \mathcal{G} \), where \( \rho_i \) is merely convex and \( f_i \) is strongly convex with a Lipschitz continuous gradient for \( i \in \mathcal{N} \). A method based on ADMM taking proximal-gradient steps, IC-ADMM, is proposed to reduce the computational work of ADMM due to exact minimizations required in each iteration. Under the
assumption that the smallest eigenvalue of the un-oriented Laplacian of $\mathcal{G}$ is known at all agents, it is shown that IC-ADMM sequence converges when each $f_i$ is strongly convex – no rate result is provided for this case; on the other hand, linear convergence is established in the absence of the merely convex (possibly non-smooth) term $\rho_i$ and assuming each $C_i$ has full column-rank in addition to the previous assumptions required for establishing the convergence result. In a similar spirit, to overcome the costly exact minimizations required in ADMM, an exact first-order algorithm (EXTRA) is proposed in \cite{30} for minimizing $\bar{f}$ over an undirected static network $\mathcal{G}$. When $\bar{f}$ is smooth and strongly convex with modulus $\bar{\mu} > 0$, it is shown that the algorithm has linear convergence without assuming each $f_i$ to be strongly convex provided that the step-size $\alpha > 0$, constant among all the nodes, is sufficiently small, i.e., $\alpha = \mathcal{O}(\bar{\mu}/L_{\text{max}}^2)$. In a follow up work, Extra-Push \cite{31} has been proposed that extends EXTRA to handle strongly connected, directed static networks using push-sum protocol. Convergence of Extra-Push, without providing any rate, has been shown under boundedness assumption on the iterate sequence; moreover, under the assumption that the stationary distribution, $\phi \in \mathbb{R}^{|\mathcal{N}|}$, of the column-stochastic mixing matrix that represents the static directed network is known, i.e., each node $i \in \mathcal{N}$ knows $\phi_i > 0$, they relax the boundedness assumption on the iterate sequence, and show that a variant of Extra-Push converges at a linear rate if each $f_i$ is smooth and strongly convex with $\mu_i > 0$ for $i \in \mathcal{N}$ – note that assuming each node $i \in \mathcal{N}$ knows $\phi_i$ exactly is a fairly strong assumption in a decentralized optimization setting. In \cite{32}, Xi et al. also combined EXTRA with the push-sum protocol to obtain DEXTRA to minimize strongly convex $\bar{f}$ over a static directed network. In addition to assumptions on $\{f_i\}_{i \in \mathcal{N}}$ in \cite{31}, by further assuming that $\nabla f_i$ bounded over $\mathbb{R}^n$ for $i \in \mathcal{N}$, which implies boundedness of the iterate sequence, it is shown that the iterate sequence converges linearly when the constant step-size $\alpha$, fixed for all $i \in \mathcal{N}$, is chosen carefully belonging to a non-trivial interval $[\alpha_{\text{min}}, \alpha_{\text{max}}]$ such that $\alpha_{\text{min}} > 0$ – note that the boundedness on each $\nabla f_i$ is a strong requirement and clearly it is not satisfied by commonly used quadratic loss function. In a follow up paper \cite{33}, Xi and Khan proposed Accelerated Distributed Directed Optimization (ADD-OPT) where they improved on the nontrivial step-size condition of DEXTRA and showed that the iterates converge linearly when the constant step-size $\alpha$ is chosen sufficiently small – assuming that the directed network topology is static and each $f_i$ is strongly convex with Lipschitz continuous gradients (without assuming boundedness as in \cite{32}). In a more general setting, Nedić and Olshevsky \cite{34} proposed a stochastic (sub)gradient-push for minimizing strongly convex $\bar{f}$ on time-varying directed graphs without assuming differentiability when the stochastic
error in subgradient samples has zero mean and bounded standard deviation. When \( \mu_i > 0 \) for all \( i \in \mathcal{N} \), choosing a diminishing step-size sequence, they were able to show \( O(\log(k)/k) \) rate result provided that the iterate sequence stays bounded – the boundedness assumption on the iterate sequence can be removed by assuming that functions are smooth, having Lipschitz continuous gradients. In [35], Nedić et al. proposed distributed inexact gradient methods referred to as DIGing and Push-DIGing for time-varying undirected and directed networks, respectively. Assuming \( f_i \) is strongly convex with Lipschitz continuous gradient for each \( i \in \mathcal{N} \), it is shown that the iterate sequence converges linearly provided that the constant step-size \( \alpha \), fixed for all \( i \in \mathcal{N} \), is chosen sufficiently small.

For constrained consensus optimization, other than few exceptions, e.g., [23], [24], [25], [26], [27], [28], the existing methods require that each node compute a projection on the local set \( \chi_i \) in addition to consensus and (sub)gradient steps, e.g., [21], [22]. Moreover, among those few exceptions, only [25], [26], [27], [28] can handle agent-specific constraints without assuming global knowledge of the constraints by all agents. However, no rate results in terms of suboptimality, local infeasibility, and consensus violation exist for the primal-dual distributed methods in [25], [26], [27] when implemented for the agent-specific conic constraint sets \( \chi_i = \{x : A_i x - b_i \in K_i \} \) studied in this paper. In [25], a consensus-based distributed primal-dual perturbation (PDP) algorithm using a diminishing step-size sequence is proposed. The objective is to minimize a composition of a global network function (smooth) with the sum of local objective functions (smooth), i.e., \( F(\sum_{i \in \mathcal{N}} f_i(x)) \), subject to local compact sets and inequality constraints on the summation of agent specific constrained functions, i.e., \( \sum_{i \in \mathcal{N}} g_i(x) \leq 0 \), over a time-varying directed network. They showed that the local primal-dual iterate sequence converges to a global optimal primal-dual solution; however, no rate result was provided. The proposed PDP method can also handle non-smooth constraints with similar convergence guarantees. In a recent work [26], the authors proposed a distributed algorithm on time-varying directed networks for solving saddle-point problems subject to consensus constraints. The algorithm can also solve consensus optimization problems with inequality constraints that can be written as summation of local convex functions of local and global variables. It is shown that using a carefully selected decreasing step-size sequence, the ergodic average of primal-dual sequence converges with \( O(1/\sqrt{k}) \) rate in terms of saddle-point evaluation error; however, when applied to constrained optimization problems, no rate in terms of either suboptimality or infeasibility is provided. In [27], a closely related paper to ours, a proximal dual consensus ADMM method, PDC-ADMM,
is proposed by Chang to minimize $\bar{\varphi}$ subject to a coupling equality and agent-specific constraints over both static and time-varying undirected networks – for the time-varying topology, they assumed that agents are on/off and communication links fail randomly with certain probabilities. Each agent-specific set is assumed to be an intersection of a polyhedron and a “simple” compact set. More precisely, the goal is to solve $\min_x \{ \sum_i \varphi_i(x_i) : \sum_{i \in \mathcal{N}} C_i x_i = d, \ x_i \in \chi_i, i \in \mathcal{N} \}$ where $\varphi_i = \rho_i + f_i$ is composite convex, $\chi_i = \{ x_i : A_i x_i \geq b_i, x_i \in \mathcal{S}_i \}$ and $\mathcal{S}_i$ is a convex compact set. Clearly, by properly choosing the primal constraint $\sum_{i \in \mathcal{N}} C_i x_i = d$ one can impose consensus on $\{x_i\}_{i \in \mathcal{N}}$. The polyhedral constraints defining $\chi_i$ are handled using a penalty formulation without requiring projection onto them. It is shown that both for static and time-varying cases, PCD-ADMM have $O(1/k)$ ergodic convergence rate in the mean for suboptimality and infeasibility when each $f_i$ is strongly convex and differentiable with a Lipschitz continuous gradient for $i \in \mathcal{N}$. More recently, in [28], Aybat and Yazdandoost Hamedani proposed a distributed primal-dual method to solve (3) when $\varphi_i = \rho_i + f_i$ is composite convex. Assuming $f_i$ is smooth, $O(1/k)$ ergodic rate is shown for suboptimality and infeasibility. In this paper, we aim to improve on this rate by further assuming $\sum_i \varphi_i$ is strongly convex to achieve $O(1/k^2)$ ergodic rate.

**Contribution.** To the best of our knowledge, only a handful of methods, e.g., [25], [26], [27], [28] can handle consensus problems, similar to (3), with agent-specific local constraint sets $\{\chi_i\}_{i \in \mathcal{N}}$ without requiring each agent $i \in \mathcal{N}$ to project onto $\chi_i$. However, no rate results in terms of suboptimality, local infeasibility, and consensus violation exist for the distributed methods in [25], [26], [27] when implemented for conic sets $\{\chi_i\}_{i \in \mathcal{N}}$ studied in this paper; moreover, none of these four methods exploits the strong convexity of the sum function $\bar{\varphi} = \sum_{i \in \mathcal{N}} \varphi_i$. We believe DPDA and DPDA-TV proposed in this paper is one of the first decentralized algorithms to solve (3) with $O(1/k^2)$ ergodic rate guarantee on both sub-optimality and infeasibility. More precisely, we show that when $\bar{\varphi}$ is strongly convex and each $\varphi_i$ is composite convex with smooth $f_i$ for $i \in \mathcal{N}$, our proposed method reduces the suboptimality and infeasibility with $O(1/k^2)$ rate as $k$, the number primal-dual iterations, increases, and it requires $O(k)$ and $O(k \log(k))$ local communications for all $k$ iterations in total when the network topologies are static and time-varying, respectively. To the best of our knowledge, this is the best rate result for our setting. Moreover, the proposed methods do not require the agents to know any global parameter depending on the entire network topology, e.g., the second smallest eigenvalue of the Laplacian.

**Notation.** Throughout $||.||$ denotes either the Euclidean norm or the spectral norm, and $\langle \theta, w \rangle \triangleq \theta^T w$ for $\theta, w \in \mathbb{R}^n$. Given a convex set $\mathcal{S}$, let $\sigma_\mathcal{S}(.)$ denote its support function, i.e., $\sigma_\mathcal{S}(\theta) \triangleq$
sup_{w \in S} \langle \theta, w \rangle$, let \( I_S(\cdot) \) denote the indicator function of \( S \), i.e., \( I_S(w) = 0 \) for \( w \in S \) and equal to \(+\infty\) otherwise, and let \( P_S(w) \triangleq \arg\min\{\|v - w\| : \, v \in S\} \) denote the projection onto \( S \).

For a closed convex set \( S \), we define the distance function as \( d_S(w) \triangleq \|P_S(w) - w\| \). Given a convex cone \( K \subseteq \mathbb{R}^m \), let \( K^\circ \) denote its dual cone, i.e., \( K^\circ \triangleq \{ \theta \in \mathbb{R}^m : \langle \theta, w \rangle \geq 0 \ \forall w \in K \} \), and \( K^\circ = -K^* \) denote the polar cone of \( K \). Note that for a given cone \( K \subseteq \mathbb{R}^m \), \( \sigma_K(\theta) = 0 \) for \( \theta \in K^\circ \) and equal to \(+\infty\) if \( \theta \not\in K^\circ \), i.e., \( \sigma_K(\theta) = \|\theta\|_{K^\circ} \) for all \( \theta \in \mathbb{R}^m \). Given a convex function \( g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\} \), its convex conjugate is defined as \( g^*(w) \triangleq \sup_{\theta \in \mathbb{R}^n} \langle w, \theta \rangle - g(\theta) \).

\( \otimes \) denotes the Kronecker product, \( 1_n \in \mathbb{R}^n \) be the vector all ones, \( I_n \) is the \( n \times n \) identity matrix. \( S_+^n \) (\( S^n_+ \)) denotes the cone of symmetric positive (semi)definite matrices. For \( Q \succ 0 \), i.e., \( Q \in S_+^n \), \( Q \)-norm is defined as \( \|z\|_Q \triangleq \sqrt{z^T Q z} \). Given \( Q \in S_+^n \), \( \lambda_{\min}(W) \) denotes the smallest positive eigenvalue of \( Q \). \( \Pi \) denotes the Cartesian product. Finally, for \( \theta \in \mathbb{R}^n \), we adopt \( (\theta)_+ \in \mathbb{R}^n_+ \) to denote \( \max\{\theta, 0\} \) where max is computed componentwise.

A. Preliminary

Let \( \mathcal{X} \) and \( \mathcal{Y} \) be finite-dimensional vector spaces. In a recent paper, Chambolle and Pock [18] proposed a primal-dual algorithm (PDA) for the following convex-concave saddle-point problem:

\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y) \triangleq \Phi(x) + \langle Tx, y \rangle - h(y), \quad \text{where} \quad \Phi(x) \triangleq \rho(x) + g(x) \tag{4}
\]

is a strongly convex function with modulus \( \mu \) such that \( \rho \) and \( h \) are possibly non-smooth convex functions, \( g \) is convex and has a Lipschitz continuous gradient defined on \( \text{dom} \rho \) with Lipschitz constant \( L \), and \( T \) is a linear map. Given some positive step-size sequences \( \{\tau^k, \kappa^k, \eta^k\}_{k \geq 0} \) and the initial iterates \( x^0, y^0 \), PDA consists of two proximal-gradient steps:

\[
y^{k+1} \leftarrow \arg\min_y h(y) - \langle T(x^k + \eta^k(x^k - x^{k-1})), y \rangle + D_k(y, y^k), \tag{5a}
\]

\[
x^{k+1} \leftarrow \arg\min_x \rho(x) + g(x^k) + \langle \nabla g(x), x - x^k \rangle + \langle Tx, y^{k+1} \rangle + \frac{1}{2\tau^k} \|x - x^k\|^2, \tag{5b}
\]

where \( D_k \) is a Bregman distance function such that \( D_k(y, \bar{y}) \geq \frac{1}{2\tau_k} \|y - \bar{y}\|^2 \) for any \( y \) and \( \bar{y} \) and \( k \geq 0 \). In [18], a simple proof for the ergodic convergence is provided for (5); indeed, it is shown that when the convexity modulus for \( \rho \) and \( g \) are \( \mu \) and 0, resp., and if \( \tau^k, \kappa^k, \eta^k > 0 \) are chosen such that \( \frac{1}{\tau^k} + \mu \geq \frac{1}{\tau^k + \eta^k + \kappa^k}, \left( \frac{1}{\tau^k} - L \right) \geq \|T\|^2 \kappa^k \), and \( \kappa^k = \kappa^{k+1} \eta^{k+1} \) for all \( k \geq 0 \), then

\[
N_K \left( \mathcal{L}(\bar{x}^K, y) - \mathcal{L}(x, y^K) \right) + \frac{\kappa^K}{\tau^k} \frac{1}{2\tau^k} \|x - \bar{x}^K\|^2 \leq \frac{1}{2\tau^K} \|x - x^0\|^2 + D_0(y, y^0), \tag{6}
\]

for all \( x, y \in \mathcal{X} \times \mathcal{Y} \), where \( N_K \triangleq \sum_{k=1}^K \kappa^{k-1} \kappa^0, \bar{x}^K \triangleq N_K^{-1} \sum_{k=1}^K \kappa^{k-1} x^k \) and \( y^K \triangleq N_K^{-1} \sum_{k=1}^K \kappa^{k-1} y^k \) for all \( K \geq 1 \). In [18], it is shown that \( \{\tau^k, \kappa^k, \eta^k\}_{k \geq 0} \) can be chosen such that \( N_k = \mathcal{O}(k^2) \), \( \tau^k = \mathcal{O}(1/k) \) and \( \kappa^k = \mathcal{O}(k) \) for \( k \geq 0 \).
that \( \theta \) denotes the graph Laplacian of \( G \). One can reformulate (7) as a saddle point problem. Indeed, let

\[
0 - \text{consensus constraints}
\]

\( x \) is a block matrix such that \( M \), denote the set of neighboring nodes of \( i \), e.g., the entry \( H_{ij} \) of \( G \) and \( \theta \) are connected, and \( \theta \) can exchange information only if \( (i, j) \in E \). Let \( N_i \triangleq \{ j \in N : (i, j) \in E \} \text{ or } (j, i) \in E \} \) denote the set of neighboring nodes of \( i \in N \), and \( d_i \triangleq |N_i| \) is the degree of node \( i \in N \).

Let \( x_i \in \mathbb{R}^n \) denote the local decision vector of node \( i \in N \). By taking advantage of the fact that \( G \) is connected, we can reformulate (3) as a consensus optimization problem:

\[
\min_{x_i \in \mathbb{R}^n, i \in N} \left\{ \sum_{i \in N} \varphi_i(x_i) \mid x_i = x_j : \lambda_{ij}, \forall (i, j) \in E, \quad A_ix_i - b_i \in K_i : \theta_i, \forall i \in N \right\}, \tag{7}
\]

where \( \lambda_{ij} \in \mathbb{R}^n \) and \( \theta_i \in \mathbb{R}^{m_i} \) are the corresponding dual variables. Let \( x = [x_i]_{i \in N} \in \mathbb{R}^{n|N|} \). The consensus constraints \( x_i = x_j \) for \( (i, j) \in E \) can be formulated as \( Mx = 0 \), where \( M \in \mathbb{R}^{n|E| \times n|N|} \) is a block matrix such that \( M = H \otimes I_n \) where \( H \) is the oriented edge-node incidence matrix, i.e., the entry \( H_{(i,j),l} \) corresponding to edge \( (i, j) \in E \) and node \( l \in N \), is equal to 1 if \( l = i \), \(-1 \) if \( l = j \), and 0 otherwise. Note that \( M^TM = H^TH \otimes I_n = \Omega \otimes I_n \), where \( \Omega \in \mathbb{R}^{N \times N} \) denotes the graph Laplacian of \( G \), i.e., \( \Omega_{ii} = d_i \), \( \Omega_{ij} = -1 \) if \( (i, j) \in E \) or \( (j, i) \in E \), and equal to 0 otherwise.

Since \( x^* \) is the unique solution to (3) and (7), and since \( x^* \triangleq 1 \otimes x^* \) satisfies \( (\Omega \otimes I_n)x^* = 0 \), one can reformulate (7) as a saddle point problem. Indeed, let \( x = [x_i]_{i \in N} \), \( y = [\theta \otimes \lambda] \) such that \( \theta = [\theta_i]_{i \in N} \) and \( \lambda = [\lambda_{ij}]_{(i,j) \in E} \), then for any \( \alpha \geq 0 \), one can compute a primal-dual optimal solution to (3) through solving

\[
\min_x \max_y \mathcal{L}(x, y) \triangleq \frac{\alpha}{2} \|x\|_{\Omega \otimes I_n}^2 + \sum_{i \in N} \left( \varphi_i(x_i) + (\theta_i, A_ix_i - b_i) - \sigma_{K_i}(\theta_i) \right) + \langle \lambda, Mx \rangle. \tag{8}
\]
Next, we consider implementation of PDA in (5) to solve (8) for some $\alpha \geq 0$.

**Definition 2.** Let $\mathcal{X} \triangleq \prod_{i \in \mathcal{N}} \mathbb{R}^{n_i}$ and $\mathcal{X} \ni x = [x_i]_{i \in \mathcal{N}}$; $\mathcal{Y} \triangleq \prod_{i \in \mathcal{N}} \mathbb{R}^{m_i} \times \mathbb{R}^{|\mathcal{E}|}$, $\mathcal{Y} \ni y = [\theta^T \lambda^T]^T$ such that $\theta = [\theta_i]_{i \in \mathcal{N}} \in \mathbb{R}^m$ and $\lambda = [\lambda_{ij}]_{(i,j) \in \mathcal{E}} \in \mathbb{R}^{m_0}$, where $m \triangleq \sum_{i \in \mathcal{N}} m_i$, and $m_0 \triangleq n|\mathcal{E}|$. Given parameters $\gamma^k > 0$ and $\kappa^k_i > 0$ for $i \in \mathcal{N}$, let $D_{\gamma^k} \triangleq \frac{1}{\gamma^k} I_{m_0}$, $D_{\kappa^k_i} \triangleq \text{diag}(\frac{1}{\kappa^k_i} I_{m_i})_{i \in \mathcal{N}}$, and $D_{\kappa^k_i \gamma^k} \triangleq \begin{bmatrix} D_{\kappa^k_i} & 0 \\ 0 & D_{\gamma^k} \end{bmatrix}$. 

**Definition 3.** Let $\Phi, \varphi : \mathcal{X} \rightarrow \mathbb{R} \cup \{\infty\}$ such that $\Phi(x) = \rho(x) + g(x)$ and $\varphi(x) = \rho(x) + f(x)$ where $\rho(x) \triangleq \sum_{i \in \mathcal{N}} \rho_i(x_i)$, $f(x) \triangleq \sum_{i \in \mathcal{N}} f_i(x_i)$, and $g(x) \triangleq f(x) + \frac{\alpha}{2} \|x\|^2_{\Omega \otimes I_n}$, and let $h : \mathcal{Y} \rightarrow \mathbb{R} \cup \{\infty\}$ such that $h(y) \triangleq \sum_{i \in \mathcal{N}} \sigma_{K_i}(\theta_i) + \langle b_i, \theta_i \rangle$. Define the block-diagonal matrix $A \triangleq \text{diag}(A_i)_{i \in \mathcal{N}} \in \mathbb{R}^{m \times n|\mathcal{N}|}$ and $T = [A^T \ M^T]^T$.

Given some positive parameters $\gamma^k, \tau^k > 0$, $\kappa^k_i > 0$ for $i \in \mathcal{N}$ - we shortly discuss how to select them, we define the Bregman function $D_{\gamma^k}(y, \bar{y}) = \frac{1}{2} \|y - \bar{y}\|^2_{D^{-1}_{\gamma^k, \kappa^k}}$ for each $k \geq 0$. Hence, given $\Phi, h$ and $T$ as in Definition 3 and the initial iterates $x^0$ and $y^0 = [\theta^0^T \lambda^0^T]^T$, the PDA iterations given in (5) take the following form:

$$
\theta_i^{k+1} \leftarrow \text{argmin}_{\theta_i} \sigma_{K_i}(\theta_i) - \langle A_i(x_i^k + \eta^k(x_i^k - x_i^{k-1})) - b_i, \theta_i \rangle + \frac{1}{2\kappa^k_i} \|\theta_i - \theta_i^k\|^2, \quad i \in \mathcal{N}
$$

$$
\mathbf{x}^{k+1} \leftarrow \text{argmin}_{\lambda} - (M(x^k + \eta^k(x^k - x^{k-1})), \lambda) + \frac{1}{2\gamma^k} \|\lambda - \lambda^k\|^2 = \lambda^k + \gamma^k M(x^k + \eta^k(x^k - x^{k-1}))
$$

$$
\mathbf{x}^{k+1} \leftarrow \text{argmin}_{\mathbf{x}} \langle \lambda^{k+1}, \mathbf{Mx} \rangle + \langle \nabla g(x^k), \mathbf{x} \rangle + \sum_{i \in \mathcal{N}} \left[ \rho_i(x_i) + \langle A_i x_i - b_i, \theta_i^{k+1} \rangle + \frac{1}{2\tau^k} \|x_i - x_i^k\|^2 \right]
$$

$$
\text{argmin}_{\mathbf{x}} \sum_{i \in \mathcal{N}} \left[ \rho_i(x_i) + \langle \nabla f_i(x_i^k), x_i \rangle + \langle A_i x_i - b_i, \theta_i^{k+1} \rangle + \frac{1}{2\tau^k} \|x_i - x_i^k\|^2 \right] + \langle \lambda^{k+1}, \mathbf{Mx} \rangle + \alpha \langle (\Omega \otimes \mathbf{I}_n)\mathbf{x}, \mathbf{x} \rangle.
$$

Since $K_i$ is a cone, $\text{prox}_{\kappa^k_i \sigma_{K_i}}(\cdot) = \mathcal{P}_{K_i}(\cdot)$; hence, $\theta_i^{k+1}$ can be written in closed form as

$$
\theta_i^{k+1} = \mathcal{P}_{K_i^k}(\theta_i^k + \kappa^k_i (A_i(x_i^k + \eta^k(x_i^k - x_i^{k-1})) - b_i)), \quad i \in \mathcal{N}.
$$

Using recursion in (9c), we can write $\lambda^{k+1}$ as a partial summation of primal iterates $\{\mathbf{x}^\ell\}_{\ell=0}^k$, i.e., $\lambda^{k+1} = \lambda^0 + \sum_{\ell=0}^k \gamma^\ell M(\mathbf{x}^\ell + \eta^\ell(\mathbf{x}^\ell - \mathbf{x}^{\ell-1}))$ for $k \geq 0$. Let $\mathbf{x}^0 \leftarrow \mathbf{0}$, and define $\{s^k\}_{k \geq 0}$ such that $s^0 = 0$ and $s^{k+1} = s^k + \gamma^k (\mathbf{x}^k + \eta^k(\mathbf{x}^k - \mathbf{x}^{k-1}))$ for $k \geq 0$; hence, $\lambda^k = \mathbf{M} s^k$ for $k \geq 0$. Using the fact that $\mathbf{M}^T \mathbf{M} = \Omega \otimes \mathbf{I}_n$, we obtain

$$
\langle \mathbf{Mx}, \lambda^{k+1} \rangle = \langle \mathbf{x}, (\Omega \otimes \mathbf{I}_n) s^{k+1} \rangle = \sum_{i \in \mathcal{N}} (x_i, \sum_{j \in \mathcal{N}_i} (s_{i}^{k+1} - s_{j}^{k+1})).
$$
Thus, PDA iterations given in (9) for the static graph \( G \) can be computed in a decentralized way, via the node-specific computations as in time-invariant distributed primal dual algorithm displayed in Fig. [1] below.

![Algorithm DPDA ( \( x^0, \theta^0, \alpha, \delta_1, \delta_2, \mu \) )](image)

**Algorithm DPDA ( \( x^0, \theta^0, \alpha, \delta_1, \delta_2, \mu \) )**  

**Initialization:** \( x^{-1} \leftarrow x^0, s^0 \leftarrow 0, \delta_1, \delta_2 > 0, \mu \in (0, \max\{\mu, \mu_0\}] \)  

**Step k:** (\( k \geq 0 \))  

1. \( \theta_i^{k+1} \leftarrow \mathcal{P}_{\mathcal{K}_1^0}(\theta_i^k + \kappa_i^k(A_i(x_i^k + \eta_i(x_i^k - x_i^{k-1}) - b_i)) \), \( \forall i \in \mathcal{N} \)  
2. \( s_i^{k+1} \leftarrow s_i^k + \kappa_i^k(x_i^k + \eta_i(x_i^k - x_i^{k-1})) \), \( \forall i \in \mathcal{N} \)  
3. \( x_i^{k+1} \leftarrow \text{prox}_{\tau \kappa_i^k}(x_i^k - \tau \nabla f_i(x_i^k) + A_i^T \theta_i^{k+1} + \sum_{j \in \mathcal{N}_i}(s_j^{k+1} - s_j^{k-1}) + \alpha(x_i^k - x_i^j)) \), \( \forall i \in \mathcal{N} \)  
4. \( \eta_i^{k+1} \leftarrow \frac{1}{\sqrt{1+\mu \eta_i^{k}}} \), \( \tau \leftarrow \eta_i^{k+1} \), \( \tau_i^k \leftarrow (\frac{1}{\eta_i^{k+1}} + \mu)^{-1} \)  
5. \( \gamma_i^{k+1} \leftarrow \gamma_i^k / \eta_i^{k+1}, \kappa_i^k \leftarrow \gamma_i^k(\frac{\delta_1}{||A_i||^2} \) )  

**Fig. 1:** Distributed Primal Dual Algorithm for static \( G \) (DPDA)

**Definition 4.** Let \( W \in \mathbb{S}^{|\mathcal{N}|} \) such that \( W_{ij} = W_{ji} < 0 \) for \( (i, j) \in \mathcal{E} \), \( W_{ii} = W_{ji} = 0 \) for \( (i, j) \notin \mathcal{E} \), and \( W_{ii} = -\sum_{j \in \mathcal{N}_i} W_{ij} \) for \( i \in \mathcal{N} \).

**Remark II.1.** According Assumption [I.2] when \( \mu > 0 \), \( f(x) = \sum_{i \in \mathcal{N}} f_i(x_i) \) is strongly convex with modulus \( \mu \). That said, as emphasized in the introduction, although \( \bar{f}(x) = \sum_{i \in \mathcal{N}} f_i(x_i) \) is strongly convex with modulus \( \bar{\mu} > 0 \), it is possible that \( f \) may not when \( \mu = 0 \).

Inspired from Proposition 3.6. in [30], we show that by suitably regularizing \( f \), one can obtain a strongly convex function when \( \mu = 0 \).

**Lemma II.1.** Consider \( f(x) = \sum_{i \in \mathcal{N}} f_i(x_i) \) under Assumption [I.2] suppose \( \mu = 0 \). Given \( \alpha > 0 \) and \( W \) as in Definition [4] let \( f_\alpha(x) \triangleq f(x) + \alpha r(x) \), where \( r(x) \triangleq \frac{1}{2} ||x||^2_{W \otimes I_n} \). Then \( f_\alpha \) is strongly convex with modulus \( \mu_\alpha \triangleq \frac{\bar{\mu}|\mathcal{N}| + \alpha \lambda_2}{2} - \left( \left( \frac{\mu|\mathcal{N}| - \alpha \lambda_2}{2} \right)^2 + 4 \bar{L}^2 \right)^{1/2} > 0 \) for any \( \alpha > \frac{4}{\mu \lambda_{\text{min}}(W)} \sum_{i \in \mathcal{N}} L_i \), where \( \bar{L} = \sqrt{\sum_{i \in \mathcal{N}} L_i^2 / |\mathcal{N}|} \) and \( \lambda_2 = \lambda_{\text{min}}^+(W) \).

**Remark II.2.** When \( \mu > 0 \), i.e., all \( f_i \)'s are strongly convex, the parameter \( \alpha \) can be set to zero; hence, \( g(x) = f(x) \) is strongly convex with modulus \( \mu_g = \mu \). Otherwise, when \( \mu = 0 \), \( \alpha \) should be chosen according to Lemma [II.1], hence, \( g(x) = f_\alpha(x) \) is strongly convex with modulus \( \mu_g = \mu_\alpha \). The condition \( \alpha > \frac{4}{\mu \lambda_{\text{min}}(W)} \sum_{i \in \mathcal{N}} L_i \) is similar to the one in [30], where they also have a parameter \( W \in \mathbb{S}^{|\mathcal{N}|} \) for their algorithm and \( \alpha \) should be greater than \( \frac{|\mathcal{N}|L_{\text{max}}^2}{2\mu \lambda_{\text{min}}^+(W)} \) and \( L_{\text{max}} = \max_{i \in \mathcal{N}} L_i \).
Next, we quantify the suboptimality and infeasibility of the DPDA iterate sequence.

**Theorem II.2.** Suppose Assumption I.1 holds. Let \( \{x^k, \theta^k\}_{k \geq 0} \) be the sequence generated by Algorithm DPDA, displayed in Fig. 1, initialized from an arbitrary \( x^0 \) and \( \theta^0 = 0 \). Then \( \{x^k\}_{k \geq 0} \) converges to \( x^* = 1 \otimes x^* \) such that \( x^* \) is the optimal solution to (3); moreover, the following error bounds hold for all \( K \geq 1 \):

\[
\max \left\{ |\Phi(\bar{x}^K) - \varphi(x^*)|, \|M\bar{x}^K\| + \sum_{i \in \mathcal{N}} \|\theta^*_i\| d_{ki} (A_i \bar{x}^K - b_i) \right\} \leq \Theta_0 / N_K, \quad \|x^K - x^*\|^2 \leq \frac{\hat{\tau}^K}{\gamma^K} 2\gamma^0 \Theta_0,
\]

where \( \Theta_0 \triangleq \frac{1}{2\gamma^0} + \sum_{i \in \mathcal{N}} \left[ \frac{1}{2\gamma^0} \|x^0_i - x^*\|^2 + \frac{2}{\gamma^0} \|\theta_i^*\|^2 \right] \). Moreover, \( \frac{\hat{\tau}^K}{\gamma^K} = O(1/K^2) \).

**Remark II.3.** Note that the result in Theorem II.2 can be extended to weighted graphs by replacing the Laplacian matrix \( \Omega \) in \( g(x) = f(x) + \frac{1}{2} \|x\|^2_{\Omega \otimes I_n} \) with a weighted Laplacian \( W \) as in Definition 4 and also replacing consensus constraint \( Mx = 0 \) in (7) with \( (W \otimes I_n)x = 0 \).

### III. A DISTRIBUTED METHOD FOR A TIME-VARYING COMMUNICATION NETWORK

In this section we develop a distributed primal-dual algorithm for solving (3) when the communication network topology is time-varying. We will adopt the following definition and assumption for the time-varying network model.

**Definition 5.** Given \( t \geq 0 \), for an undirected graph \( \mathcal{G}^t = (\mathcal{N}^t, \mathcal{E}^t) \), let \( \mathcal{N}^t_i \triangleq \{ j \in \mathcal{N} : (i, j) \in \mathcal{E}^t \} \) denote the set of neighboring nodes of \( i \in \mathcal{N} \), and \( d^t_i \triangleq |\mathcal{N}^t_i| \) represent the degree of node \( i \in \mathcal{N} \) at time \( t \); for a directed graph \( \mathcal{G}^t = (\mathcal{N}^t, \mathcal{E}^t) \), let \( \mathcal{N}_i^{t,\text{in}} \triangleq \{ j \in \mathcal{N} : (j, i) \in \mathcal{E}^t \} \cup \{ i \} \) and \( \mathcal{N}_i^{t,\text{out}} \triangleq \{ j \in \mathcal{N} : (i, j) \in \mathcal{E}^t \} \cup \{ i \} \) denote the in-neighbors and out-neighbors of node \( i \) at time \( t \), respectively; and \( d^t_i \triangleq |\mathcal{N}_i^{t,\text{out}}| \) be the out-degree of node \( i \).

**Assumption III.1.** Suppose that \( \{\mathcal{G}^t\}_{t \in \mathbb{R}_+} \) is a collection of either all directed or all undirected graphs. When \( \mathcal{G}^t \) is an undirected graph, node \( i \in \mathcal{N} \) can send and receive data to and from \( j \in \mathcal{N} \) at time \( t \) only if \( j \in \mathcal{N}^t_i \), i.e., \((i, j) \in \mathcal{E}^t \) or \((j, i) \in \mathcal{E}^t \); on the other hand, when \( \mathcal{G}^t \) is a directed graph, node \( i \in \mathcal{N} \) can receive data from \( j \in \mathcal{N} \) only if \( j \in \mathcal{N}_i^{t,\text{in}} \), i.e., \((j, i) \in \mathcal{E}^t \), and can send data to \( j \in \mathcal{N} \) only if \( j \in \mathcal{N}_i^{t,\text{out}} \), i.e., \((i, j) \in \mathcal{E}^t \).

We assume a compact domain, i.e., let \( \Delta_i \triangleq \max_{x_i, x'_i \in \text{dom} \rho_i} \|x - x'\| \) and \( \Delta \triangleq \max_{i \in \mathcal{N}} \Delta_i < \infty \). Let \( \mathcal{B}_0 \triangleq \{ x \in \mathbb{R}^n : \|x\| \leq 2\Delta \} \) and \( \mathcal{B} \triangleq \Pi_{i \in \mathcal{N}} \mathcal{B}_0 \); and let \( \mathcal{C} \) and \( \tilde{\mathcal{C}} \) be the sets of consensus and bounded consensus decisions respectively.
Since $x^*$ is the unique solution to (3) and since $x^* \triangleq 1 \otimes x^*$ satisfies $P_C(x^*) = 0$, one can reformulate (3) as a saddle point problem using $\tilde{C}$. Indeed, let $x = [x_i]_{i \in \mathcal{N}} \in \mathbb{R}^{n|\mathcal{N}|}$, $y = [\theta^T \lambda^T]^T$ such that $\theta = [\theta_i]_{i \in \mathcal{N}}$ and $\lambda \in \mathbb{R}^{n|\mathcal{N}|}$, then for any $\alpha \geq 0$, one can compute a primal-dual optimal solution to (3) through solving

$$\min_x \max_y \mathcal{L}(x, y) = \frac{\alpha}{2} d^2_C(x) + \sum_{i \in \mathcal{N}} \left( \varphi_i(x_i) + \langle \theta_i, A_i x_i - b_i \rangle - \sigma_{K_i}(\theta_i) \right) + \langle \lambda, x \rangle - \sigma_C(\lambda).$$

\[ \text{(11)} \]

Next, we consider a slightly different implementation of PDA in (5) to solve (11).

**Definition 6.** Let $\mathcal{X} \triangleq \Pi_{i \in \mathcal{N}} \mathbb{R}^n$ and $\mathcal{X}' \ni x = [x_i]_{i \in \mathcal{N}}$; $\mathcal{Y} \triangleq \prod_{i \in \mathcal{N}} \mathbb{R}^{m_i} \times \mathbb{R}^{m_0}$, $\mathcal{Y} \ni y = [\theta^T \lambda^T]^T$ and $\theta = [\theta_i]_{i \in \mathcal{N}} \in \mathbb{R}^m$, where $m \triangleq \sum_{i \in \mathcal{N}} m_i$ and $m_0 \triangleq n|\mathcal{N}|$. Given parameters $\gamma^k > 0$, $\kappa^k_i > 0$ for $i \in \mathcal{N}$, let $D_{\gamma^k} \triangleq \frac{1}{\gamma^k} I_{m_0}$, $D_{\kappa^k} \triangleq \text{diag}(\kappa^k i_{i \in \mathcal{N}})$, and $D_{\kappa^k, \gamma^k} \triangleq \begin{bmatrix} D_{\kappa^k} & 0 \\ 0 & D_{\gamma^k} \end{bmatrix}$.

**Definition 7.** Let $\Phi, \varphi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ such that $\Phi(x) = \rho(x) + g(x)$ and $\varphi(x) = \rho(x) + f(x)$ where $\rho(x) \triangleq \sum_{i \in \mathcal{N}} \rho_i(x_i)$, $g(x) \triangleq f(x) + \frac{\alpha}{2} d^2_C(x)$ and $f(x) \triangleq \sum_{i \in \mathcal{N}} f_i(x_i)$, and let $h : \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ such that $h(y) \triangleq \sigma_C(\lambda) + \sum_{i \in \mathcal{N}} \sigma_{K_i}(\theta_i) + \langle b_i, \theta_i \rangle$. Define the block-diagonal matrix $A \triangleq \text{diag}((A_i)_{i \in \mathcal{N}}) \in \mathbb{R}^{m \times n|\mathcal{N}|}$ and $T = [A^T I_{n|\mathcal{N}}]^T$.

Given some positive parameters $\gamma^k, \tau^k > 0$, $\kappa^k_i > 0$ for $i \in \mathcal{N}$ – we shortly discuss how to select them, we define the Bregman function $D_k(y, \tilde{y}) = \frac{1}{2} \| y - \tilde{y} \|^2_{D_{\kappa^k, \gamma^k}}$ for each $k \geq 0$. Hence, given $\Phi, h$ and $T$ as in Definition 7 and the initial iterates $x^0$ and $y^0 = [\theta^0^T \lambda^0^T]^T$, the PDA iterations given in (5) take the following form for $k \geq 0$:

\[
\begin{align*}
\theta_i^{k+1} & \leftarrow \arg\min_{\theta_i} \sigma_{K_i}(\theta_i) - \langle A_i (\xi^k_i + \eta^k (\xi^k_i - \xi^{k-1}_i)) - b_i, \theta_i \rangle + \frac{1}{2\kappa^k_i} \| \theta_i - \theta^k_i \|^2_2, \quad i \in \mathcal{N} \quad \text{(12a)} \\
\lambda^{k+1} & \leftarrow \arg\min_{\nu} \sigma_{\tilde{C}}(\nu) - \langle \xi^k + \eta^k (\xi^k - \xi^{k-1}), \nu \rangle + \frac{1}{2\gamma^k} \| \nu - \nu^k \|^2_2, \quad \text{(12b)} \\
\nu^{k+1} & \leftarrow \lambda^{k+1}, \quad \text{(12c)} \\
x^{k+1} & \leftarrow \arg\min_{\xi} \rho(\xi) + \langle \nabla g(\xi^k), \xi \rangle + \langle A \xi - b, \theta^{k+1} \rangle + \langle \xi, \nu^{k+1} \rangle + \frac{1}{2\gamma^k} \| \xi - \xi^k \|^2, \quad \text{(12d)} \\
\xi^{k+1} & \leftarrow x^{k+1}, \quad \text{(12e)}
\end{align*}
\]

where $\xi^{-1} = \xi^0 = x^0$ and $\nu^0 = \lambda^0$. For $k \geq 0$, using extended Moreau decomposition for proximal operators, $\lambda^{k+1}$ in (12b) can be computed as

\[
\lambda^{k+1} = \text{prox}_{\gamma^k \sigma_{\tilde{C}}}(\nu^{k} + \gamma^k (\xi^k + \eta^k (\xi^k - \xi^{k-1}))) = \gamma^k (\omega^k - P_{\tilde{C}}(\omega^k)), \quad \text{(13)}
\]
where $\omega^k \triangleq \frac{1}{\gamma_k} \nu^k + \xi^k + \eta^k(\xi^k - \xi^{k-1})$ for $k \geq 0$. Moreover, $\nabla g$ for the $x$-step in (12d) can be computed as

$$\nabla g(x^k) = \nabla f(x^k) + \alpha \mathcal{P}_C(x^k) = \nabla f(x^k) + \alpha(x^k - \mathcal{P}_C(x^k)).$$

(14)

For any $x = [x_i]_{i \in \mathcal{N}} \in \mathcal{X}$, $\mathcal{P}_C(x)$ and $\mathcal{P}_C(x)$ can be computed as

$$\mathcal{P}_C(x) = 1 \otimes \arg\min_{\xi \in \mathcal{B}_0} \|\xi - \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} x_i\|^2 = \mathcal{P}_B(\mathcal{P}_C(x)), \text{ and } \mathcal{P}_C(x) = 1 \otimes p(x),$$

(15)

where $p(x) \triangleq \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} x_i$, $\mathcal{P}_B(x) = [\mathcal{P}_{B_0}(x_i)]_{i \in \mathcal{N}}$ and $\mathcal{P}_{B_0}(x_i) = x_i \min\{1, \frac{2\Delta}{\|x_i\|}\}$ for $i \in \mathcal{N}$. Equivalently, $\mathcal{P}_C(x) = (W \otimes I_n) x$ for $W \triangleq \frac{1}{|\mathcal{N}|} 1^T 1 \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$.

Although $\theta$-step of the PDA implementation in (12) can be computed locally at each node, computing $x$-step and $\lambda$-step require communication among the nodes to evaluate $\mathcal{P}_C(\omega^k)$ and $\mathcal{P}_C(\xi^k)$. Indeed, evaluating the average operator $p(.)$ is not a simple operation in a decentralized computational setting which only allows for communication among the neighbors. In order to overcome this issue, we will approximate the average operator $p(.)$ using multi-communication rounds, and analyze the resulting iterations as an inexact primal-dual algorithm. We define a communication round at time $t$ as an operation over $\mathcal{G}^t$ such that every node simultaneously sends and receives data to and from its neighboring nodes according to Assumption III.1 – the details of this operation will be discussed shortly. We assume that communication among neighbors occurs instantaneously, and nodes operate synchronously; and we further assume that for each PDA iteration $k \geq 0$, there exists an approximate averaging operator $\mathcal{R}^k(.)$ which can be computed in a decentralized fashion and approximate $\mathcal{P}_C(.)$ with decreasing approximation error as $k$, the number of PDA iterations, increases. This inexact version of PDA using approximate averaging operator $\mathcal{R}^k(.)$ and running on time-varying communication network $\{\mathcal{G}^t\}$ will be called DPDA-TV, of which details will be explained next.

**Assumption III.2.** Given a time-varying network $\{\mathcal{G}^t\}_{t \in \mathbb{R}_+}$ such that $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ for $t \geq 0$, suppose that there is a global clock known to all $i \in \mathcal{N}$. Assume that the local operations requiring to compute $\Pi_{\mathcal{K}_i}$ as in (12a), and $\text{prox}_{\rho_i}$ and $\nabla f_i$ as in (12e) can be completed between two ticks of the clock for all $i \in \mathcal{N}$ and $k \geq 0$; and every time the clock ticks a communication round with instantaneous messaging between neighboring nodes takes place subject to Assumption III.1. Suppose that for each $k \geq 0$ there exists $\mathcal{R}^k(.) = [\mathcal{R}^k_i(.)]_{i \in \mathcal{N}}$ such that $\mathcal{R}^k_i(.)$ can be computed with local information available to node $i \in \mathcal{N}$, and decentralized.
computation of $R^k$ requires $q_k$ communication rounds. Furthermore, we assume that there exist $\Gamma > 0$ and $\alpha \in (0, 1)$ such that for all $k \geq 0$, $R^k$ satisfies
\[
\|R^k(w) - P_C(w)\| \leq N \Gamma \beta^{q_k} \|w\|, \quad \forall w \in \mathbb{R}^{mo}.
\] (16)

Now we briefly talk about such operators. Let $V^t \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$ be a matrix encoding the topology of $G^t = (\mathcal{N}, \mathcal{E}^t)$ in some way for $t \in \mathbb{Z}_+$. We define $W^{t,s} \triangleq V^t V^{t-1} \cdots V^{s+1}$ for any $t, s \in \mathbb{Z}_+$ such that $t \geq s + 1$. For directed time-varying graph $G^t$, set $V^t \in \mathbb{R}^{|\mathcal{N}| \times |\mathcal{N}|}$ as:
\[
V^t_{ij} = \frac{1}{d_j} \text{ if } j \in \mathcal{N}_i^t, \quad V^t_{ij} = 0 \text{ if } j \notin \mathcal{N}_i^t, \quad i \in \mathcal{N}.
\] (17)

Let $t_k \in \mathbb{Z}_+$ be the total number of communication rounds done before the $k$-th iteration of DPDA-TV, and let $q_k \in \mathbb{Z}_+$ be the number of communication rounds to be performed within the $k$-th iteration while evaluating $R^k$. For $x = [x_i]_{i \in \mathcal{N}} \in \mathcal{X}$ such that $x_i \in \mathbb{R}^n$ for $i \in \mathcal{N}$, define
\[
R^k(x) \triangleq \text{diag}(W^{t_k + q_k, t_k} 1_{|\mathcal{N}|}^{-1}(W^{t_k + q_k, t_k} \otimes I_n)) x
\] (18)
to approximate $P_C(\cdot)$. Note that $R^k(\cdot)$ can be computed in a distributed fashion requiring $q_k$ communication rounds — $R^k$ is nothing but the push-sum protocol [37]. Assuming that the digraph sequence $\{G^t\}_{t \in \mathbb{Z}_+}$ is uniformly strongly connected (M-strongly connected), it follows from [37], [38] that $R^k$ satisfies Assumption III.2. When $\{G^t\}_{t \in \mathbb{Z}_+}$ is undirected time-varying network, then choosing $\{V^t\}$ according to Metropolis weights, one can show that
\[
R^k(x) \triangleq (W^{t_k + q_k, t_k} \otimes I_m)x
\] (19)
satisfies Assumption III.2 under certain conditions, e.g., see [39].

Note that for $\tilde{R}^k(\cdot) \triangleq P_B(R^k(\cdot))$, we have $\tilde{R}^k(w) \in B$, and $\|\tilde{R}^k(w) - P_C(w)\| \leq N \Gamma \beta^{q_k} \|w\|$ for $w \in \mathbb{R}^{mo}$ due to non-expansivity of $P_B$. Consider the $k$-th iteration of PDA as shown in (12). Instead of setting $\nu^{k+1}$ to $\lambda^{k+1}$ and $\xi^{k+1}$ to $x^{k+1}$, which require computing $P_C$, we propose replacing these assignment operations in (12c) and (12e) with an operation that uses the inexact averaging operator $R^k$ to approximate $P_C$. This way, we obtain inexact variant of (12) replacing (12c) and (12e) with
\[
\nu^{k+1} \leftarrow \gamma^k (\omega^k - P_B(R^k(\omega^k))) \quad \text{where} \quad \omega^k = \frac{1}{\gamma^k} \nu^k + \xi^k + \eta^k (\xi^k - \xi^{k-1}), \quad (20a)
\]
\[
\xi^{k+1} \leftarrow \text{prox}_{\tau \rho}(\xi^k - \tau^k \left(\nabla f(\xi^k) + A^T \theta^{k+1} + \nu^{k+1} + \alpha (\xi^k - R^k(\xi^k))\right)). \quad (20b)
\]
Thus, PDA iterations given in (12) can be computed inexactly, but in \textit{decentralized} way for a time-varying connectivity network \( \{G_i^t\}_{i \geq 0} \), via the node-specific computations as in time-varying distributed primal dual algorithm displayed in Fig. 2 below. Indeed, the iterate sequence \( \{\xi^k, \nu^k, \theta^k\}_{k \geq 0} \) generated by DPDA-TV displayed in Fig. 2 is the same sequence generated by the recursion in (12a), (20a), and (20b). The sequences \( \{x^k\}_{k \geq 0} \) and \( \{\lambda^k\}_{k \geq 0} \) will not be explicitly computed, instead we will use it in the analysis of the inexact algorithm.

\begin{center}
\begin{minipage}{0.9\textwidth}
\begin{algorithm}
\textbf{Algorithm DPDA-TV (} \( x^0, \theta^0, \alpha, \delta_1, \delta_2, \mu, \{q_k\} \) \)
\begin{enumerate}
\item \( \theta^{k+1} \leftarrow \mathcal{P}_{\mathcal{C}_i} \left( \theta^k + \kappa^k \left( A_i (\xi^k + \eta^k (\xi^k - \xi^{k-1}) - b_i) \right) \right) \), \( \omega^k \leftarrow \frac{1}{\tau} \nu^k + \xi^k + \eta^k (\xi^k - \xi^{k-1}) \), \( i \in \mathcal{N} \)
\item \( \nu^{k+1} \leftarrow \gamma^k \left( \omega^k - \mathcal{P}_{\mathcal{B}_0} (\mathcal{R}_i^k (\omega^k)) \right) \), \( i \in \mathcal{N} \)
\item \( \xi^{k+1} \leftarrow \text{prox}_{\tau \varphi_i} \left( \xi^k - \tau \left( \nabla f_i (\xi^k) + A_i^\top \theta^k + \nu_i^{k+1} + \alpha (\xi^k - \mathcal{R}_i^k (\xi^k)) \right) \right) \), \( i \in \mathcal{N} \)
\item \( \eta^{k+1} \leftarrow \frac{1}{\tau} \nu^k + \xi^k + \kappa_i (\xi^k - \mathcal{R}_i^k (\xi^k)) \), \( \kappa^{k+1} \leftarrow \beta \kappa^k \), \( i \in \mathcal{N} \)
\item \( \gamma^{k+1} \leftarrow \gamma^k / \eta^{k+1} \), \( \delta^{k+1} \leftarrow \delta^k + \beta \kappa^k \), \( i \in \mathcal{N} \)
\end{enumerate}
\end{algorithm}
\end{minipage}
\end{center}

\textbf{Fig. 2:} time-varying Distributed Primal Dual Algorithm (DPDA-TV)

Recall Remark [II.1] it is possible that \( \mu = 0 \). In the next lemma, similar to Lemma [II.1] we generalize the result in Proposition 3.6. of [30], making it suitable for time-varying topology, and show that by suitably regularizing \( f \), one can obtain a strongly convex function when \( \mu = 0 \).

\textbf{Lemma III.1.} Consider \( f(x) = \sum_{i \in \mathcal{N}} f_i(x_i) \) under Assumption [I.2] and suppose \( \mu = 0 \). Given \( \alpha > 0 \), let \( f_\alpha(x) \triangleq f(x) + \alpha \cdot r(x) \), where \( r(x) \triangleq \frac{1}{2} d_i^2(x) \). Then \( f_\alpha \) is strongly convex with modulus \( \mu_\alpha \triangleq \frac{\bar{\mu} / |N| - \alpha}{2} + 4 \bar{L}^2 > 0 \) for any \( \alpha > \frac{4}{\bar{\mu}} \sum_{i \in \mathcal{N}} L_i^2 \), where \( \bar{L} = \sqrt{\sum_{i \in \mathcal{N}} L_i^2 / |\mathcal{N}|} \).

Next, we quantify the suboptimality and infeasibility of the DPDA-TV iterate sequence. Recall that if \( \mu > 0 \), then we set \( \alpha = 0 \) and set \( g = f \); otherwise, when \( \mu = 0 \), it follows from Lemma [III.1] that for any \( \alpha > \frac{4}{\bar{\mu}} \sum_{i \in \mathcal{N}} L_i^2 \), \( f_\alpha \) is strongly convex with modulus \( \mu_\alpha > 0 \); hence, we set \( g = f_\alpha \) − See also Remark [II.1].

\textbf{Theorem III.2.} Suppose Assumptions [I.1] [I.2] [III.1] and [III.2] hold. Starting from \( \nu^0 = 0, \theta^0 = 0 \), and an arbitrary \( x^0 \), let \( \{\xi^k, \theta^k, \nu^k\}_{k \geq 0} \) be the iterate sequence generated by Algorithm DPDA-TV, displayed in Fig. 2 using \( q_k \geq (5 + c) \log_{1/\beta}(k + 1) \) communication rounds for the \( k \)-th iteration for \( k \geq 0 \). Then \( \{\xi^k\}_{k \geq 0} \) converges to \( x^* = 1 \otimes x^* \) such that \( x^* \) is the optimal solution to (3).
Moreover, the following bounds hold for all $K \geq 1$:

\[
\max \left\{ |\Phi(\xi^K) - \varphi(x^*)|, \ d_c(\xi^K) + \sum_{i \in N} \|\theta_i^*\|_2 d_{K_i}(A_i \xi^K - b_i) \right\} \leq \frac{\Theta(K)}{N_K} = O\left(\frac{1}{K^2}\right),
\]

\[
\left\|\bar{x}^K - x^*\right\|^2 \leq \frac{\bar{\tau}^K}{\gamma^K} 2\gamma^0 \Theta(K) = O\left(\frac{1}{K^2}\right),
\]

and the parameters satisfy $N_K = O(K^2)$ and $\bar{\tau}^K / \gamma^K = O(1/K^2)$, where $N_K = \sum_{k=1}^{K} \gamma_k^{-1}$, $\bar{x}^K = N_K^{-1} \sum_{k=1}^{K} \gamma_k^{-1} x^k$, and $\Theta(K) = O\left(\sum_{k=1}^{K} \beta^{\log_2(k+4)}\right)$; hence, $\sup_{K \in \mathbb{Z}^+} \Theta(K) < \infty$.

**Remark III.1.** Note that, at the $K$-th iteration, the suboptimality, infeasibility and consensus violation are $O\left(\frac{1}{N_K} \Theta(K)\right)$ in the ergodic sense, and the distance of iterates to $x^*$ is $O\left(\frac{\bar{\tau}^K}{\gamma^K} \Theta(K)\right)$ where $\Theta(K)$ denotes the error accumulations due to average approximation. Moreover, $\Theta(K)$ can be bounded above for all $K \geq 1$ as $\Theta_2(K) \leq C_1 \sum_{k=1}^{K} \beta^{\log_2(k+4)}$; therefore, for any $c > 0$, choosing $\{q_k\}_{k \in \mathbb{Z}^+}$ as stated in Theorem III.2 ensures that $\sum_{k=1}^{\infty} \beta^{\log_2(k+4)} < 1 + \frac{1}{c}$.

Moreover, for any $c > 0$, setting $q_k = (5 + c) \log_2 (k+1)$ for $k \geq 0$ implies that the total number of communication rounds right before the $K$-th iteration is equal to $t_K = \sum_{k=0}^{K-1} q_k \leq (5 + c) K \log_2 (K)$.

**IV. Numerical Section**

In this section, we illustrate the performance of DPDA and DPDA-TV for solving synthetic C-LASSO problems. We first test the effect of network topology on the performance of proposed algorithms, and then we compare DPDA and DPDA-TV with other distributed primal-dual algorithms, DPDA-S and DPDA-D, proposed in [28] for solving (3) – it is shown in [28] that both DPDA-S and DPDA-D converge with $O(1/K)$ ergodic rate when $\varphi$ is merely convex. In fact, when $\varphi$ is strongly convex with modulus $\mu > 0$, using the fact that $\varphi(x^*) - \varphi(\bar{x}^K) \geq \frac{\mu}{2} \left\|\bar{x}^K - x^*\right\|^2$, it immediately follows that $\left\|\bar{x}^K - x^*\right\|^2 \leq O(1/K)$.

We consider an isotonic C-LASSO problem over network $\mathcal{G}^t = (\mathcal{N}, \mathcal{E}^t)$ for $t \geq 0$. This problem can be formulated in a centralized form as $x^* \triangleq \arg\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Cx - d\|^2 + \lambda \|x\|_1 : Ax \leq 0 \right\}$, where the matrix $C = [C_i]_{i \in \mathcal{N}} \in \mathbb{R}^{m \times n}$, $d = [d_i]_{i \in \mathcal{N}} \in \mathbb{R}^{m \times |\mathcal{N}|}$, and $A \in \mathbb{R}^{n \times n}$. In fact, the matrix $A$ captures the isotonic feature of vector $x^*$, and can be written explicitly as, $A(\ell, \ell + 1) = -1$, for $1 \leq \ell \leq n - 1$, otherwise it is zero. Each agent $i$ has access to $C_i$, $d_i$, and $A$; hence, by making local copies of $x$, the decentralized formulation can be expressed as

\[
\min_{x = [x_i]_{i \in \mathcal{N}} \in \mathcal{C}} \left\{ \frac{1}{2} \sum_{i \in \mathcal{N}} \|C_i x_i - d_i\|^2 + \frac{\lambda}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \|x_i\|_1 : Ax_i \leq 0, \ i \in \mathcal{N} \right\},
\]

where $\mathcal{C}$ is the consensus set - see [10].
In the rest, we set \( n = 20, m = n + 2, \lambda = 0.05\) and \( K_i = -\mathbb{R}_+^{n-1}\) for \( i \in \mathcal{N}\). Moreover, for each \( i \in \mathcal{N}\), we generate \( C_i \in \mathbb{R}^{m \times n} \) as follows: after \( mn \) entries i.i.d. with Gaussian distribution are sampled, the condition number of \( C_i \) is normalized by sampling the singular values from \([1, 3]\) uniformly at random. We generate the first 5 and the last 5 components of \( x^* \) by sampling from \([-10, 0]\) and \([0, 10]\) uniformly at random in ascending order, respectively, and the other middle 10 components are set to zero; hence, \( [x^*]_j \leq [x^*]_{j+1} \) for \( j = 1, \ldots, n-1 \). Finally, we set \( d_i = C_i(x^* + \epsilon_i) \), where \( \epsilon_i \in \mathbb{R}^n \) is a random vector with i.i.d. components following Gaussian distribution with zero mean and standard deviation of \( 10^{-3} \).

**Generating static undirected network:** \( G = (\mathcal{N}, \mathcal{E}) \) is generated as a random small-world network. Given \(|\mathcal{N}|\) and the desired number of edges \(|\mathcal{E}|\), we choose \(|\mathcal{N}|\) edges creating a random cycle over nodes, and then the remaining \(|\mathcal{E}| - |\mathcal{N}|\) edges are selected uniformly at random.

**Generating time-varying undirected network:** Given \(|\mathcal{N}|\) and the desired number of edges \(|\mathcal{E}_0|\) for the initial graph, we generate a random small-world \( G_0 = (\mathcal{N}, \mathcal{E}_0) \) as described above. Given \( M \in \mathbb{Z}_+\), and \( p \in (0, 1)\), for each \( k \in \mathbb{Z}_+\), we generate \( G^t = (\mathcal{N}, \mathcal{E}^t)\), the communication network at time \( t \in \{(k-1)M, \ldots, kM - 2\} \) by sampling \( \lceil p|\mathcal{E}_0| \rceil \) edges of \( G_0 \) uniformly at random and we set \( \mathcal{E}^{kM-1} = \mathcal{E}_0 \setminus \bigcup_{t=(k-1)M}^{kM-2} \mathcal{E}^t \). In all experiments, we set \( M = 5, p = 0.8\) and the number of communications per iteration is set to \( q_k = 10 \ln(k + 1)\).

**A. Effect of Network Topology**

In this section, we test the performance of DPDA and DPDA-TV on undirected communication networks. To illustrate the effect of network topology, we consider four scenarios in which the number of nodes \(|\mathcal{N}| \in \{10, 40\}\) and the average number of edges per node (\(|\mathcal{E}|/|\mathcal{N}|\)) is either \( \approx 1.5 \) or \( \approx 4.5\). For each scenario, we plot both the relative error, i.e., \( \max_{i \in \mathcal{N}} \|x_i^k - x^*\| / \|x^*\| \) and the infeasibility, i.e., \( \max_{i \in \mathcal{N}} d_{K_i}(A\bar{x}_i^k) = \max_{i \in \mathcal{N}} \| (A\bar{x}_i^k)_+ \| \) versus iteration number \( k \). All the plots show the average statistics over all 25 randomly generated replications.
Testing DPDA on static undirected communication networks: We generated the static small-world networks \( G = (\mathcal{N}, \mathcal{E}) \) as described above for \((|\mathcal{N}|, |\mathcal{E}|) \in \{(10, 15), (10, 45), (40, 60), (40, 180)\}\) and solve the saddle-point formulation \((8)\) corresponding to \((22)\) using DPDA. For DPDA, displayed in Fig. 1, we chose \( \delta_1 = \max_{i \in \mathcal{N}} d_i = d_{\max} \) and \( \delta_2 = 2 \max_{i \in \mathcal{N}} L_i = 2L_{\max}, \) which lead to the initial step-sizes as \( \gamma_0 = \frac{2}{3}L_{\max}, \) \( \tau_0 = \frac{1}{3L_{\max}}, \) and \( \kappa_0 = \frac{2}{3} \|A\|. \)

In Fig. 3 we plot \( \max_{i \in \mathcal{N}} \|x^k_i - x^*\| / \|x^*\| \) and \( \max_{i \in \mathcal{N}} \|(A\bar{x}^k)_i^+\| \) statistics for DPDA versus iteration number \( k. \) Note that compared to average edge density, the network size has more influence on the convergence rate, i.e., the smaller the network faster the convergence is. On the other hand, for fixed size network, as expected, higher the density faster the convergence is.

![Fig. 3: Effect of network topology on the convergence rate of DPDA](image)

Testing DPDA-TV on time-varying undirected communication networks: We first generated an undirected graph \( G_u = (\mathcal{N}, \mathcal{E}_u) \) as in the static case, and let \( G_0 = G_u. \) Next, we generated \( \{G^t\}_{t \geq 1} \) as described above by setting \( M = 5 \) and \( p = 0.8. \) For each consensus round \( t \geq 1, \) \( V^t \) is formed according to Metropolis weights, i.e., for each \( i \in \mathcal{N}, \) \( V^t_{ij} = 1 / (\max\{d_i, d_j\} + 1) \) if \( j \in \mathcal{N}_t \), \( V^t_{ii} = 1 - \sum_{j \in \mathcal{N}_t} V^t_{ij}, \) and \( V^t_{ij} = 0 \) otherwise – see \((19)\) for our choice of \( \mathcal{R}^k. \)

For DPDA-TV, displayed in Fig. 2, we chose \( \delta_1 = \delta_2 = 1, \) which lead to the initial step-sizes as \( \gamma_0 = \frac{1}{2}, \) \( \tau_0 = \frac{1}{2\|A\|}, \) and \( \kappa_0 = \frac{1}{2\|A\|^2}. \) In Fig. 4 we plot \( \max_{i \in \mathcal{N}} \|\xi^k_i - x^*\| / \|x^*\| \) and \( \max_{i \in \mathcal{N}} \|(A\bar{x}^k)^+_i\| \) statistics for DPDA-TV versus iteration number \( k. \) – we used \( \{\xi^k\} \) to compute the error statistics instead of \( \{x^k\} \) as \( x^k \) is never actually computed for DPDA-TV. Note that network size and average edge density have the same impact on the rate as in the static case.
B. Comparison with other methods

We also compared our methods with DPDA-S and DPDA-D, in terms of the relative error and infeasibility of the ergodic iterate sequence, i.e., \( \max_{i \in \mathcal{N}} \| \bar{x}_i^k - x^* \| / \| x^* \| \) and \( \max_{i \in \mathcal{N}} \| (A \bar{x}_i^k)_+ \| \).

We further report the performance of our algorithms in terms of relative error of the actual iterate sequence, i.e., \( \max_{i \in \mathcal{N}} \| x_i^k - x^* \| / \| x^* \| \). For DPDA-D and DPDA-TV, we used \( \{ \xi^k \} \) sequence to compute the error statistics instead of \( \{ x^k \} \) as \( x^k \) is never actually computed. In this section we fix the number of nodes to \( |\mathcal{N}| = 10 \) and the average edge density to \( |\mathcal{E}| / |\mathcal{N}| = 4.5 \). We observed the same convergence behavior for the other network scenarios discussed in the previous section.

**Static undirected network:** We generated \( \mathcal{G} = (\mathcal{N}, \mathcal{E}) \) and chose the algorithm parameters as in the previous section. Moreover, the step-sizes of DPDA-S are set to the initial steps-sizes of DPDA. As it can be seen in Fig. 5, DPDA has faster convergence when compared to DPDA-S.

**Time-varying undirected network:** We generated the network sequence \( \{ \mathcal{G}^t \}_{t \geq 0} \) and chose the parameters as in the previous section. Moreover, the step-sizes of DPDA-D are set to the initial steps-sizes of DPDA-TV. As it can be seen in Fig. 6, DPDA-TV has faster convergence when compared to DPDA-D.
Fig. 5: Comparison of DPDA and DPDA-S over undirected static network

Fig. 6: Comparison of DPDA-TV and DPDA-D over undirected time-varying network

**Time-varying directed network:** In this scenario, we generated time-varying communication networks similar to [35]. Let $G_d = (\mathcal{N}_d, \mathcal{E}_d)$ be the directed graph shown in Fig. 8 where it has $|\mathcal{N}| = 12$ nodes and $|\mathcal{E}_d| = 12$ directed edges. We set $G_0 = G_d$, and we generate $\{G^t\}_{t \geq 0}$ generated as in the undirected case with parameters $M = 5$ and $p = 0.8$; hence, $\{G^t\}_{t \geq 0}$ is $M$-strongly-connected. Moreover, communication weight matrices $V^t$ are formed according to rule (17). We chose the initial step-sizes for DPDA-TV as in the time-varying undirected case, and the constant step-sizes of DPDA-D is set to the initial steps-sizes of DPDA-TV. In Fig. 7 we compare DPDA-TV against DPDA-D. We observe that over time-varying directed networks DPDA-TV again outperforms DPDA-D for both statistics.
Fig. 7: Comparison of DPDA-TV and DPDA-D over directed time-varying network.

Fig. 8: $G_d = (\mathcal{N}, \mathcal{E}_d)$ directed strongly connected graph

REFERENCES

[1] Joel B Predd, SB Kulkarni, and H Vincent Poor. Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):56–69, 2006.

[2] Ioannis D Schizas, Alejandro Ribeiro, and Georgios B Giannakis. Consensus in ad hoc WSNs with noisy links - Part I: Distributed estimation of deterministic signals. *Signal Processing, IEEE Transactions on*, 56(1):350–364, 2008.

[3] Ke Zhou and Stergios I Roumeliotis. Multirobot active target tracking with combinations of relative observations. *IEEE Transactions on Robotics*, 27(4):678–695, 2011.

[4] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics*, 9(1):427–438, 2013.

[5] Juan Andrés Bazerque and Georgios B Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, 2010.

[6] Juan Andrés Bazerque, Gonzalo Mateos, and Georgios B Giannakis. Group-lasso on splines for spectrum cartography. *IEEE Transactions on Signal Processing*, 59(10):4648–4663, 2011.

[7] Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 1543–1550. IEEE, 2012.

[8] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2012.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
V. APPENDIX

A. Proof of Lemma II.1

Let $x^* = 1_{|N|} \otimes x^*$, where $x^*$ is the unique optimal solution to (3), and according to Assumption 1.2, $\bar{f}$ is strongly convex with modulus $\bar{\mu} > 0$. Note that any $W$ as given in Definition 4 is positive semidefinite, and $\text{Null}(W) = \text{Span}\{1\}$. In the rest, we will use these properties of $W$. Fix some arbitrary $\alpha > \frac{4}{\lambda_2 \bar{\mu}} \sum_{i \in N} L_i^2$ and $\bar{x} \in \mathbb{R}^{n|N|}$.

$x \in \mathbb{R}^{n|N|}$ can be decomposed into $u \in \text{Span}\{1\}$ and $v \in \text{Span}\{1\}^\perp$ where $x = u + v$ and $\|x\|^2 = \|u\|^2 + \|v\|^2$. From definition of $f_0$ we have that,

$$\langle \nabla f_0(x) - \nabla f_0(x^*), x - x^* \rangle = \langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle + \alpha \|x - x^*\|^2_{W \otimes I_n}.$$  

(23)

Let $N \triangleq |N|$ and $\bar{L} \triangleq \sqrt{\frac{\sum_{i \in N} L_i^2}{N}}$. The inner product of $\langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle$ can be bounded by using the following inequalities:

$$\langle \nabla f(u) - \nabla f(x^*), u - x^* \rangle \geq \frac{\mu}{N} \|u - x^*\|^2,$$  

(24a)

$$\langle \nabla f(u) - \nabla f(x^*), x - u \rangle \geq - \sum_{i \in N} L_i \|u - x^*\| \|x_i - u\| \geq -\bar{L} \|u - x^*\| \|v\|,$$  

(24b)

$$\langle \nabla f(x) - \nabla f(u), x - u \rangle \geq 0,$$  

(24c)

$$\langle \nabla f(x) - \nabla f(u), u - x^* \rangle \geq - \sum_{i \in N} L_i \|x_i - u\| \|u - x^*\| \geq \bar{L} \|u - x^*\| \|v\|,$$  

(24d)

which follow from convexity, Lipschitz differentiability, and strong convexity of $f$. Summing above inequalities leads to,

$$\langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle \geq \frac{\mu}{N} \|u - x^*\|^2 - 2\bar{L} \|u - x^*\| \|v\|.$$  

(25)

Hence, strong convexity of $f_\alpha$ follows from (23), (25). Indeed, it follows from $W \in S^N_+$ and $\text{Null}(W) = \text{Span}\{1\}$ that we have $\|x - x^*\|^2_{W \otimes I_n} = v^T(W \otimes I_n) v \geq \lambda_2 \|v\|^2$, where $\lambda_2 = \lambda_{\text{min}}^+(W)$ is the second smallest eigenvalue of $W$. Therefore,

$$\langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle \geq \frac{\mu}{N} \|u - x^*\|^2 - 2\bar{L} \|u - x^*\| \|v\| + \alpha \lambda_2 \|v\|^2.$$  

(26)

Next, fix some arbitrary $\omega \geq 0$. Then either (i) $\|v\| \leq \omega \|u - x^*\|$, or (ii) $\|v\| \geq \omega \|u - x^*\|$ holds. If (i) is true, then (26) implies

$$\langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle \geq \left( \frac{\mu}{N} - 2\omega \bar{L} \right) \|u - x^*\|^2 + \alpha \lambda_2 \|v\|^2 \geq \min \left\{ \frac{\mu}{N} - 2\omega \bar{L}, \alpha \lambda_2 \right\} \|x - x^*\|^2;$$  

(27)

on the other hand, if (ii) is true, then (26) implies

$$\langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle \geq \frac{\mu}{N} \|u - x^*\|^2 + \left( \alpha \lambda_2 - \frac{2\bar{L}}{\omega} \right) \|v\|^2 \geq \min \left\{ \frac{\mu}{N}, \alpha \lambda_2 - \frac{2\bar{L}}{\omega} \right\} \|x - x^*\|^2.$$  

(28)

Combining (27) and (28) we conclude that,

$$\langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle \geq \min \left\{ \frac{\mu}{N} - 2\omega \bar{L}, \alpha \lambda_2 - \frac{2\bar{L}}{\omega} \right\} \|x - x^*\|^2,$$  

(29)
Since \( \omega \geq 0 \) is arbitrary, \( f_a \) is strongly convex with modulus \( \mu = \max_{\omega \geq 0} \min \left\{ \frac{\mu}{N} - 2L \omega, \; \alpha \lambda_2 - \frac{2L}{\omega} \right\} \).

Note \( \mu \) is attained for \( \omega_\alpha \geq 0 \) such that \( \frac{\mu}{N} - 2L \omega_\alpha = \alpha \lambda_2 - \frac{2L}{\omega_\alpha} \), which implies that \( \omega_\alpha = \frac{1}{2} \left( \frac{\mu/N - \alpha \lambda_2}{2L} + \sqrt{\left( \frac{\mu/N - \alpha \lambda_2}{2L} \right)^2 + 4} \right) \). Moreover, \( \mu = \frac{\mu}{N} - 2L \omega_\alpha \) is the value given in the statement of the lemma, and we have \( \frac{\mu}{N} > \mu > 0 \) for any \( \alpha > \frac{4}{\lambda_2 \mu} \sum_{i \in N} L_{i}^2 \). It is worth mentioning that \( \mu \) is a concave increasing function of \( \alpha \) over \( \mathbb{R}_+ \), and \( \sup_{\alpha > 0} \mu = \lim_{\alpha \to \infty} \mu_\alpha = \frac{\mu}{N} \).

### B. Key Lemmas for the Proof of Theorem [II.2]

**Definition 8.** Let \( T = [A^\top M^\top]^\top \) for \( A \triangleq \text{diag}(\{A_i\}_i \in \mathcal{N}) \in \mathbb{R}^{n \times |\mathcal{N}|} \). Given \( \alpha, \mu, \delta_1 > 0 \), and arbitrary sequences \( \{\tau^k\}, \{\gamma^k\} \subset \mathbb{R}_{++}, \{\kappa^i_k\}_{k \geq 0} \subset \mathbb{R}_{++} \) for \( i \in \mathcal{N} \). For \( k \geq 0 \), define \( D_{\tau^k} \triangleq \left( \frac{\tau^k}{\gamma^k} - 1 \right) I_n \), \( D_{\tau^k} \triangleq \left( \frac{\tau^k}{\gamma^k} - 1 \right) I_n \), \( D_{\tau^k} \triangleq \text{diag}(\{(\frac{1}{\tau^k} - (L + 2\alpha d_i))I_n\}_i \in \mathcal{N}) \), and \( Q_k \triangleq \left[ \begin{array}{cc} A^k & -\gamma^k A^\top \\ -\gamma^k A & D_{\tau^k} \end{array} \right] \), where \( A^k \triangleq (\gamma^k)^2 \gamma^k \text{diag}(\{(2d_i + \delta_1)I_n\}_i \in \mathcal{N}) \geq 0 \), and \( D_{\kappa^i_k, \gamma^k} \) is defined in Definition 2.

In order to prove Theorem II.2 we first prove Lemma V.3 below which help us to appropriately bound \( L(x_{\kappa^i_k}^\kappa, y) - L(x^\kappa, y^\kappa) \) for any \( y \in \mathcal{Y} \) and \( \|x^\kappa - x^\kappa\| \).

In order to prove Lemma V.3 we first need to show the following two lemmas, Lemma V.1 and Lemma V.2, describing a proper choice for the step size sequences.

**Lemma V.1.** Given \( \delta_1 > 0 \). For any \( k \geq 0 \), \( Q_k \geq 0 \) if \( \gamma^k > 0 \), and positive numbers \( \{\kappa^i_k\}_i \in \mathcal{N} \) and \( \gamma^k \) are chosen such that

\[
\frac{\kappa^i_k}{\gamma^k} \frac{\|A_i\|^2}{\gamma^k} \leq \delta_1, \quad \forall i \in \mathcal{N}.
\]

**Proof.** Let \( D_{\kappa^i_k, \gamma^k} \) be as in Definition 2. Since \( D_{\tau^k} \succ 0 \), Schur complement condition implies that \( Q_k \geq 0 \) if and only if

\[
\begin{bmatrix}
A^k & -\gamma^k A^\top \\
-\gamma^k A & D_{\tau^k}
\end{bmatrix} - \gamma^k (\gamma^k)^2 \begin{bmatrix}
M^\top M & 0 \\
0 & 0
\end{bmatrix} \succeq 0.
\]

Moreover, since \( D_{\tau^k} \succ 0 \), again using Schur complement and the fact that \( M^\top M = \Omega \otimes I_n \), one can conclude that (31) holds if and only if \( \frac{1}{(\gamma^k)^2} A^k - \gamma^k \Omega \otimes I_n - A^\top D_{\tau^k}^{-1} A \succeq 0 \). Moreover, by definition \( \Omega = \text{diag}(\{d_i\}_{i \in \mathcal{N}}) - E \), where \( E_{ii} = 0 \) for all \( i \in \mathcal{N} \) and \( E_{ij} = E_{ji} = 1 \) if \( (i, j) \in \mathcal{E} \) or \( (j, i) \in \mathcal{E} \). Note that \( \text{diag}(\{d_i\}_{i \in \mathcal{N}}) + E \succeq 0 \) since it is diagonally dominant. Therefore, \( \Omega \preceq 2 \text{diag}(\{d_i\}_{i \in \mathcal{N}}) \). Hence, one can conclude that (31) holds if \( \frac{1}{(\gamma^k)^2} A^k - 2\gamma^k \text{diag}(\{d_i\}_{i \in \mathcal{N}}) - A^\top D_{\tau^k}^{-1} A \succeq 0 \). This condition holds if (30) is true.

**Lemma V.2.** Let \( D_{\kappa^i_k, \gamma^k} \) be as given in Definition 2 and \( D_{\tau^k}, D_{\tau^k}, D_{\tau^k} \) and \( Q_k \) be as in Definition 8 for \( \alpha \geq 0 \) chosen according to Lemma II.1 and Remark II.2, and \( \mu \in (0, \max\{\mu, \mu_\alpha\}) \). Suppose \( \{\tau^k\}, \{\gamma^k\} \subset \mathbb{R}_{++}, \{\kappa^i_k\}_{k \geq 0} \subset \mathbb{R}_{++} \) for \( i \in \mathcal{N} \) are chosen as in DPDA displayed...
Moreover, \(\eta^k \in (0, 1), 0 < \frac{1}{\tau^k} < \frac{1}{\tau} = \mathcal{O}(k),\) and \(0 < \gamma^k = \mathcal{O}(k).\)

**Proof.** It is trivial to check that the parameter sequence constructed in Fig. 1 satisfies (32). Indeed, Lemma V.1 shows that (32a) is true since \(k_i^k\) for \(i \in \mathcal{N}\) and \(\gamma^k\) as chosen in Fig. 1 satisfy (30) for all \(k \geq 0\). This specific choice of parameters satisfy (32b), (32c), (32d), and (32e) with equality. Moreover, one can use induction to show (32f) using the relations \(\tau^k > \tau^k\), \(\tilde{\tau}^k > \tilde{\tau}^k\), and \(\gamma^k \tilde{\tau}^k = \gamma^k \tilde{\tau}^k + 1\) for all \(k \geq 0\).

**Lemma V.3.** For any \(y \in \mathcal{Y}\), the iterate sequence \(\{x^k, y^k\}_{k \geq 1}\) generated using Algorithm DPDA as in Fig. 1, where \(y^k = [\theta^T \lambda^T]^T\), satisfies for all \(k \geq 0\),

\[
\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x^*, y^{k+1}) \leq \left[ \frac{1}{2} \|x^* - x^k\|^2_{D_{k, \gamma_k}} + \frac{1}{2} \|y - y^k\|^2_{D_{k, \gamma_k}} - \eta^k \langle T(x^k - x^{k-1}), y - y^k \rangle + \frac{1}{2} \|x^k - x^{k-1}\|^2_{A_k} \right]
- \left[ \frac{1}{2} \|x^* - x^{k+1}\|^2_{D_{k, \gamma_k}} + \frac{1}{2} \|y - y^{k+1}\|^2_{D_{k, \gamma_k}} - \langle T(x^k - x^k), y - y^{k+1} \rangle + \frac{1}{2} \|x^k - x^{k-1}\|^2_{D_{k, \gamma_k}} \right].
\]

**Proof.** Note that \(x\)-subproblem in (5b) is separable in local decisions \(\{x_i\}_{i \in \mathcal{N}}\); and for each \(i \in \mathcal{N}\) the local subproblem over \(x_i\) is strongly convex with constant \(1/\tau^k\). Indeed, let \(p^k = T^Ty^k\) and define \(\{p_i^k\}_{i \in \mathcal{N}}\) such that \(p_i^k\) is the subvector corresponding to the components of \(x_i\), i.e., \(p^k = [p_i^k]_{i \in \mathcal{N}}\). In addition, \(\nabla g_i(x^k) = [\nabla g_i(x^k)]_{i \in \mathcal{N}}\) where \(\nabla g_i(x^k) \triangleq \nabla f_i(x^k) + ([\Omega \otimes I_n]x^k)_i\) where \(([\Omega \otimes I_n]x^k)_i = \sum_{j \in \mathcal{N}} (x_j - x_j^k)\). Thus, for all \(i \in \mathcal{N}\)

\[
x_i^{k+1} = \arg \min_{x_i} \rho_i(x_i) + \langle \nabla g_i(x^k), x_i - x_i^k \rangle + \langle p_i^{k+1}, x_i \rangle + \frac{1}{2\tau^k} \|x_i - x_i^k\|^2.
\]

Therefore, for \(i \in \mathcal{N}\), the strong convexity of the objective in local subproblem (34) implies

\[
\rho_i(x^*) + \langle \nabla g_i(x^k), x^* - x_i^k \rangle + \langle p_i^{k+1}, x_i \rangle + \frac{1}{2\tau^k} \|x^* - x_i^k\|^2 \geq
\rho_i(x_i^{k+1}) + \langle \nabla g_i(x^k), x_i^{k+1} - x_i^k \rangle + \langle p_i^{k+1}, x_i^{k+1} \rangle + \frac{1}{2\tau^k} \|x_i^{k+1} - x_i^k\|^2 + \frac{1}{2\tau^k} \|x^* - x_i^{k+1}\|^2.
\]

Now, we show that \(\nabla g\) is Lipschitz continuous. First, recall that as we discussed in the proof of Lemma V.1, we have \(\Omega \preceq 2 \text{ diag}([d_i]_{i \in \mathcal{N}})\). Second, since \(\|x\|^2_{\Omega \otimes I_n}\) is a quadratic term, for any \(x\) we have

\[
\frac{1}{2} \|x\|^2_{\Omega \otimes I_n} = \frac{1}{2} \|x\|^2_{\Omega \otimes I_n} + \langle (\Omega \otimes I_n)x, x - \bar{x} \rangle + \frac{1}{2} (x - \bar{x})^T (\Omega \otimes I_n)(x - \bar{x})
\leq \frac{1}{2} \|x\|^2_{\Omega \otimes I_n} + \langle (\Omega \otimes I_n)x, x - \bar{x} \rangle + (x - \bar{x})^T \text{ diag}([d_i]_{i \in \mathcal{N}})(x - \bar{x}).
\]
In addition, since each $f_i$ has a Lipschitz continuous gradient, we have for any $x$ and $\bar{x}$ that
\[
f(x) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \sum_{i \in \mathcal{N}} \frac{L_i}{2} \|x_i - \bar{x}_i\|^2.
\] (37)

Let $L_g \triangleq \text{diag}([[(L_i + 2d_i \alpha)I_i]_{i \in \mathcal{N}}]) \in \mathbb{S}^{n_{\mathcal{N}}^2}$. Summing (36) and (37), for any $x$ and $\bar{x}$, we have
\[
g(x) \leq g(\bar{x}) + \langle \nabla g(\bar{x}), x - \bar{x} \rangle + \sum_{i \in \mathcal{N}} \frac{L_i + 2d_i \alpha}{2} \|x_i - \bar{x}_i\|^2
\]
\[
= g(\bar{x}) + \langle \nabla g(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \|x - \bar{x}\|^2_{L_g}.
\] (38)

It follows from strong convexity of $\tilde{f}$ that choosing $\alpha \geq 0$ according to Lemma II.1 and Remark II.2 we conclude that for any $\mu \in (0, \max\{\mu, \mu_\alpha\})$ we have
\[
g(x^*) \geq g(x^k) + \langle \nabla g(x^k), x^* - x^k \rangle + \frac{\mu}{2} \|x^* - x^k\|^2
\]
\[
\geq g(x^{k+1}) + \langle \nabla g(x^k), x^* - x^{k+1} \rangle + \frac{\mu}{2} \|x^* - x^k\|^2 - \frac{1}{2} \|x^{k+1} - x^k\|^2_{L_g}.
\] (39)

Since $\sum_{i \in \mathcal{N}} \theta_i \langle p_i^{k+1}, x^* \rangle = \langle T x^*, y^{k+1} \rangle$, first summing (35) over $i \in \mathcal{N}$, next summing the resulting inequality with (39), and then adding $g(x^k)$ to both hand-sides, we get
\[
\Phi(x^*) + \frac{1}{2} \|x^* - x^k\|^2_{D_{x,k}} \geq \Phi(x^{k+1}) + \langle T(x^{k+1} - x^*), y^{k+1} \rangle + \frac{1}{2} \|x^* - x^{k+1}\|^2_{D_{x,k}} + \frac{1}{2} \|x^{k+1} - x^k\|^2_{D_{x,k}}.
\] (40)

Similarly, let $q^k \triangleq T(x^k + \eta^k(x^k - x^{k-1}))$ and define $q^k_0 \in \mathbb{R}^{m_0}$ and $q^k_i \in \mathbb{R}^{m_i}$ for $i \in \mathcal{N}$ such that $q^k_0$ is the subvector corresponding to the components of $\lambda$, and $q^k_i$ is the subvector corresponding to the components of $\theta_i$ for $i \in \mathcal{N}$, i.e., $q^k = [q^k_1 \ldots q^k_N] q^k_0 \ldots q^k_N]$. Thus, from (9a) and (9b), we have
\[
\lambda^{k+1} = \arg\min_{\lambda} - \langle q^k_0, \lambda \rangle + \frac{1}{2 \gamma^k} \|\lambda - \lambda^k\|^2,
\]
\[
\theta^{k+1}_i = \arg\min_{\theta_i} \sigma_{\mathcal{K}_i}(\theta_i) - \langle q^k_i - b_i, \theta_i \rangle + \frac{1}{2 \kappa^k_i} \|\theta_i - \theta^k_i\|^2, \quad \forall i \in \mathcal{N}.
\]

Using the strong convexity of these subproblems, for any $y = [\theta^T, \lambda^T]^T$, we get
\[
- \langle q^k_0, \lambda \rangle + \frac{1}{2 \gamma^k} \|\lambda - \lambda^k\|^2 \geq - \langle q^k_0, \lambda^{k+1} \rangle + \frac{1}{2 \gamma^k} \|\lambda^{k+1} - \lambda^k\|^2 + \frac{1}{2 \gamma^k} \|\lambda - \lambda^{k+1}\|^2,
\]
\[
\sigma_{\mathcal{K}_i}(\theta_i) - \langle q^k_i - b_i, \theta_i \rangle + \frac{1}{2 \kappa^k_i} \|\theta_i - \theta^k_i\|^2 \geq \sigma_{\mathcal{K}_i}((\theta^{k+1})_i) - \langle q^k_i - b_i, \theta^{k+1}_i \rangle + \frac{1}{2 \kappa^k_i} \|\theta^{k+1}_i - \theta^k_i\|^2 + \frac{1}{2 \kappa^k_i} \|\theta_i - \theta^{k+1}_i\|^2.
\]

Since $\langle q^k_0, \lambda \rangle + \sum_{i \in \mathcal{N}} \langle q^k_i, \theta_i \rangle = \langle T(x^k + \eta^k(x^k - x^{k-1})), y \rangle$ for all $y$, summing the second inequality over $i \in \mathcal{N}$ and then summing the resulting inequality with the first one, we get
\[
h(y) + \frac{1}{2} \|y - y^k\|^2_{D_{x,k},x^k} \geq \frac{1}{2} \|y - y^{k+1}\|^2_{D_{x,k},x^{k+1}} + \frac{1}{2} \|y^{k+1} - y\|^2_{D_{x,k},x^k}.
\] (41)
Next, summing (40), (41), and rearranging the terms, we obtain

$$\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x^*, y^{k+1}) \leq$$

$$\eta^k \langle T(x^k - x^{k-1}), y^{k+1} - y \rangle - \frac{1}{2} \|y^{k+1} - y^k\|^2_{D_{\gamma_k, \gamma_k}} + \left[ \frac{1}{2} \|x^* - x^k\|^2_{D_{\gamma_k, \gamma_k}} + \frac{1}{2} \|y - y^k\|^2_{D_{\gamma_k, \gamma_k}} \right]$$

$$- \left[ \frac{1}{2} \|x^* - x^{k+1}\|^2_{D_{\gamma_k, \gamma_k}} + \frac{1}{2} \|y - y^{k+1}\|^2_{D_{\gamma_k, \gamma_k}} - \langle T(x^{k+1} - x^k), y - y^{k+1} \rangle + \frac{1}{2} \|x^{k+1} - x^k\|^2_{D_{\gamma_k, \gamma_k}} \right]$$

Note that we have

$$\eta^k \langle T(x^k - x^{k-1}), y^{k+1} - y \rangle = -\eta^k \langle T(x^k - x^{k-1}), y - y^k \rangle + \eta^k \langle T(x^k - x^{k-1}), y^{k+1} - y^k \rangle;$$

moreover, using (32a), i.e., \(\bar{Q}_k \geq 0\), the last term can be bounded as follows:

$$\eta^k \langle T(x^k - x^{k-1}), y^{k+1} - y^k \rangle \leq \frac{1}{2} \|y^{k+1} - y^k\|^2_{D_{\gamma_k, \gamma_k}} + \frac{1}{2} \|x^k - x^{k-1}\|^2_{\Delta_k}$$

(44)

Then, combining (42), (43) and (44) gives the desired result. \(\square\)

C. Proof of Theorem II.2

Under Assumption 1.1, a saddle point \((x^*, y^*)\) for \(\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \mathcal{L}(x, y)\) in (8) exists, where \(y^* = [\theta^* \top, \lambda^* \top] \top\); moreover, any saddle point \((x^*, \theta^*, \lambda^*)\) satisfies that \(x^* = 1 \otimes x^*\) such that \((x^*, \theta^*)\) is a primal-dual solution to (3). Thus, \(\theta^*_i \in \mathcal{K}_i^0\) and \(\mathcal{L}(x^*, \theta^*, \lambda^*) = \Phi(x^*)\). Recall Definition 3 since \(\|x^*\|^2_{\|x\|_{\Delta_k}} = 0\), we have \(g(x^*) = f(x^*)\); hence, \(\Phi(x^*) = \varphi(x^*) = \sum_{i\in\mathcal{N}} \varphi_i(x^*)\). Therefore, \(\mathcal{L}(x^*, \theta^*, \lambda^*) = \varphi(x^*)\). Moreover, note that if \((x^*, \theta^*, \lambda^*)\) is a saddle point of \(\mathcal{L}\) such that \(\lambda^* \neq 0\), then it trivially follows that \((x^*, \theta^*, 0)\) is another saddle point of \(\mathcal{L}\).

Multiplying both sides of (33) by \(\frac{\gamma^k}{\gamma^0}\) and using Lemma V.2, we get

$$\frac{\gamma^k}{\gamma^0} \left[ \mathcal{L}(x^{k+1}, y) - \mathcal{L}(x^*, y^{k+1}) \right] \leq$$

$$\frac{\gamma^k}{\gamma^0} \left[ \frac{1}{2} \|x^* - x^k\|^2_{D_{\gamma_k, \gamma_k}} + \frac{1}{2} \|y - y^k\|^2_{D_{\gamma_k, \gamma_k}} - \eta^k \langle T(x^k - x^{k-1}), y - y^k \rangle + \frac{1}{2} \|y^k - x^{k-1}\|^2_{\Delta_k} \right]$$

$$- \frac{\gamma^k}{\gamma^0} \left[ \frac{1}{2} \|x^* - x^{k+1}\|^2_{D_{\gamma_k, \gamma_k}} + \frac{1}{2} \|y - y^{k+1}\|^2_{D_{\gamma_k, \gamma_k}} - \eta^k \langle T(x^{k+1} - x^k), y - y^{k+1} \rangle + \frac{1}{2} \|x^{k+1} - x^k\|^2_{\Delta_k} \right].$$

Next, we sum (45) from \(k = 0\) to \(K - 1\); using Jensen inequality and the following facts: \(\bar{Q}_K \geq 0\) and \(x^{k+1} = x^*\), we get

$$2N_K (\mathcal{L}(x^K, y) - \mathcal{L}(x^*, \tilde{y}^K)) \leq \left[ \|x^* - x^0\|^2_{D_{\gamma, \gamma}} + \|y - y^0\|^2_{D_{\gamma, \gamma}} \right] - \frac{\gamma^K}{\gamma^0} \left[ \|x^* - x^k\|^2_{D_{\gamma_k, \gamma_k}} + \|\tilde{y}^K - y^k\|^2_{\bar{Q}_K} \right].$$

where \(\tilde{y}^K \triangleq (x^K - x^{K-1}) \top (y - y^K) \top, N_K = \sum_{k=1}^K \frac{\gamma^{k-1}}{\gamma^0}, \tilde{x}^K = N_K^{-1} \sum_{k=1}^K \gamma^{-1} \gamma^0 x^k\), and

$$\tilde{y}^K = N_K^{-1} \sum_{k=1}^K \gamma^{-1} y^k.$$

Since \(\|z^K\|^2_{\bar{Q}_K} \geq 0\) and \(\tilde{z}^k > x^k\) for \(k \geq 0\), we get the following bounds for all \(K \geq 1\): \(\mathcal{L}(x^K, \theta, \lambda) - \mathcal{L}(x^*, \bar{\theta}^K, \bar{\lambda}^K) \leq \frac{1}{N_K} \Theta(x^*, \theta, \lambda), \frac{1}{2} \|x^K - x^*\|^2 \leq \frac{\tilde{z}^K}{\gamma^0} \gamma^0 \Theta(x^*, \theta, \lambda), (46)\)

$$\Theta(x^*, \theta, \lambda) \triangleq \frac{1}{2\gamma^0} \|\lambda - \lambda^0\|^2 + \sum_{i\in\mathcal{N}} \left[ \frac{1}{2\gamma^0} \|x_i^0 - x^*\|^2 + \frac{1}{2\kappa_i^0} \|\theta_i - \theta_i^0\|^2 \right].$$

Under Assumption 1.1, one can construct a saddle point \((x^*, \theta^*, \lambda^*)\) for \(\mathcal{L}\) in (8) such that \(\lambda^* = 0\); hence, \(\mathcal{L}(x^*, \theta^*, \lambda^*) = \varphi(x^*)\) and \(\theta^*_i \in \mathcal{K}_i^0\) for \(i \in \mathcal{N}\). Define \(\bar{\theta} = [\bar{\theta}_i]_{i\in\mathcal{N}}\) such that
\[ \tilde{\theta}_i \triangleq 2\|\theta^*\| (\|P_{K_i}(A_i \bar{x}_i^K - b_i)\|)^{-1} P_{K_i}(A_i \bar{x}_i^K - b_i) \in K_i^c, \] which implies
\[ \langle A_i \bar{x}_i^K - b_i, \tilde{\theta}_i \rangle = 2\|\theta^*\| d_{K_i}(A_i \bar{x}_i^K - b_i). \] (47)

Similarly, define \( \bar{\lambda} \triangleq M \bar{x}_K / \|M \bar{x}_K\| \); hence, \( \langle M \bar{x}_K, \bar{\lambda} \rangle = \|M \bar{x}_K\| \). Together with (47), we get
\[ \mathcal{L}(\bar{x}_K, \tilde{\theta}, \bar{\lambda}) = \Phi(\bar{x}_K) + 2 \sum_{i \in \mathcal{N}} \|\theta^*_i\| d_{K_i}(A_i \bar{x}_i^K - b_i) + \|M \bar{x}_K\|. \] (48)

Note that for any \( i \in \mathcal{N} \), \( \tilde{\theta}_i^K \in K_i^c \); hence, \( \sigma_{K_i}(\tilde{\theta}_i^K) = 0 \). In addition, since \( \tilde{\theta}_i^K \in K_i^c \) and \( A_i x^* - b_i \in K_i \), we have
\[ \langle A_i x^* - b_i, \tilde{\theta}_i^K \rangle \leq 0. \] (49)

Therefore, using (49) and the fact that \( M x^* = 0 \) we get that,
\[ \mathcal{L}(x^*, \tilde{\theta}_i^K, \bar{\lambda}) \leq \varphi(x^*). \] (50)

Thus, (48), (50) and (46) together with the definitions of \( \tilde{\theta}, \bar{\lambda} \) and the fact that \( \lambda^0 = 0 \) and \( \theta^0 = 0 \) imply that
\[ \Phi(\bar{x}_K) - \varphi(x^*) + 2 \sum_{i \in \mathcal{N}} d_{K_i}(A_i \bar{x}_i^K - b_i)\|\theta^*_i\| + \|M \bar{x}_K\| \leq \frac{1}{N_K} \Theta(x^*, \tilde{\theta}, \bar{\lambda}) = \frac{\Theta_0}{N_K}. \] (51)

Since \((x^*, \theta^*, \lambda^*)\) is a saddle-point for \( \mathcal{L} \) in (8), we have \( \mathcal{L}(\bar{x}_K, \theta^*, \lambda^*) - \mathcal{L}(x^*, \theta^*, \lambda^*) \geq 0 \); therefore,
\[ \Phi(\bar{x}_K) - \varphi(x^*) + \sum_{i \in \mathcal{N}} \langle \theta^*_i, A_i \bar{x}_i^K - b_i \rangle \geq 0. \] (52)

Using the conic decomposition of \( A_i \bar{x}_i^K - b_i \) and the fact that \( \theta^*_i \in K_i^c \), we immediately get
\[ \langle A_i \bar{x}_i^K - b_i, \theta^*_i \rangle \leq \|\theta^*_i\| d_{K_i}(A_i \bar{x}_i^K - b_i). \]

Together with (52), we conclude that
\[ \Phi(\bar{x}_K) - \varphi(x^*) + \sum_{i \in \mathcal{N}} \|\theta^*_i\| d_{K_i}(A_i \bar{x}_i^K - b_i) \geq 0. \] (53)

Finally, combining inequalities (51) and (53) immediately implies the desired result. Moreover, the bound on \( \|x^* - \bar{x}_K\| \) follows from (46). In fact, possibly a tighter bound can be derived using \( \Theta(x^*, \theta^*, \lambda^*) \) for \( \lambda^* = 0 \).

D. Proof of Lemma III.1

Let \( x^* = 1_{|\mathcal{N}|} \otimes x^* \), where \( x^* \) is the unique optimal solution to (3), and according to Assumption I.2, \( f \) is strongly convex with modulus \( \bar{\mu} > 0 \). Fix some arbitrary \( \alpha > \frac{4}{\bar{\mu}} \sum_{i \in \mathcal{N}} L_i^2 \) and \( x \in \mathbb{R}^{nN^2} \). Since \( \mathcal{C} \) is a closed convex cone, \( x \) can be decomposed into \( u = P_{\mathcal{C}}(x) \) and \( v = P_{\mathcal{C}^c}(x) \), i.e., \( x = u + v \) and \( \|x\|^2 = \|u\|^2 + \|v\|^2 \). From the definition of \( f_\alpha \),
\[ \langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle = \langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle + \alpha (x - x^*, v), \] (54)

which follows from the fact that \( \nabla r(x) = x - P_{\mathcal{C}}(x) \); hence \( \nabla r(x^*) = 0 \). Let \( N \triangleq |\mathcal{N}| \) and \( \bar{L} \triangleq \sqrt{\sum_{i \in \mathcal{N}} L_i^2 / N} \). Since \( x^*, u \in \mathcal{C} \) and \( f \) is convex, Lipschitz differentiable, and strongly convex,
the same inequalities in (24) implies:
\[
\langle \nabla f(x) - f(x^*), x - x^* \rangle \geq \frac{\mu}{N} \Vert u - x^* \Vert^2 - 2\bar{L} \Vert u - x^* \Vert \Vert v \Vert. \tag{55}
\]
Note that \( u - x^* \in \mathcal{C} \); hence, \( \langle u - x^*, v \rangle = 0 \) since \( v \in \mathcal{C}^0 \). Thus, \( \langle x - x^*, v \rangle = \Vert v \Vert^2 \); this together with (54) and (55) implies that
\[
\langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle \geq \frac{\mu}{N} \Vert u - x^* \Vert^2 - 2\bar{L} \Vert u - x^* \Vert \Vert v \Vert + \alpha \Vert v \Vert^2. \tag{56}
\]
Next, fix some arbitrary \( \omega \geq 0 \). Then either (i) \( \Vert v \Vert \leq \omega \Vert u - x^* \Vert \), or (ii) \( \Vert v \Vert \geq \omega \Vert u - x^* \Vert \) holds. Using the same arguments to obtain (27), (28) and (29), we can conclude that
\[
\langle \nabla f_\alpha(x) - \nabla f_\alpha(x^*), x - x^* \rangle \geq \min \left\{ \frac{\mu}{N} - 2\bar{L} \omega, \alpha - \frac{2\bar{L}}{\omega} \right\} \Vert x - x^* \Vert^2. \tag{57}
\]
Since \( \omega \geq 0 \) is arbitrary, \( f_\alpha \) is restricted strongly convex with respect to \( x^* \) with modulus \( \mu_\alpha = \max_{\omega \geq 0} \min \left\{ \frac{\mu}{N} - 2\bar{L} \omega, \alpha - \frac{2\bar{L}}{\omega} \right\} \). Note \( \mu_\alpha \) is attained for \( \omega_\alpha \geq 0 \) such that \( \frac{\mu}{N} - 2\bar{L} \omega_\alpha = \alpha - \frac{2\bar{L}}{\omega_\alpha} \), which implies that \( \omega_\alpha = \frac{\mu}{N} - \frac{2\bar{L} \omega_\alpha}{\alpha} + \sqrt{\left( \frac{\mu}{N} - \frac{2\bar{L} \omega_\alpha}{\alpha} \right)^2 + 1} \). Moreover, \( \mu_\alpha = \frac{\mu}{N} - 2\bar{L} \omega_\alpha \) is the value given in the statement of the lemma, and we have \( \frac{\mu}{N} - \mu_\alpha > 0 \) for any \( \alpha > \frac{4}{\mu} \sum_{i \in \mathcal{N}} I_i^2 \). It is worth mentioning that \( \mu_\alpha \) is a concave increasing function of \( \alpha \) over \( \mathbb{R}_{++} \), and \( \sup_{\alpha > 0} \mu_\alpha = \lim_{\alpha \to \infty} \mu_\alpha = \frac{\mu}{N} \).

E. Key Lemmas for the Proof of Theorem III.2

We first define the proximal error sequences \( \{e_i^k\}_{k \geq 1}, \{e_2^k\}_{k \geq 1}, \text{ and } \{e_3^k\}_{k \geq 1} \) which will be used for analyzing the convergence of Algorithm DPDA-TV displayed in Fig. 2. For \( k \geq 0 \), let
\[
e_{1,k+1} \triangleq P\mathcal{C}(\omega_k) \ominus \vec{R}^k(\omega_k), \quad e_{2,k+1} \triangleq P\mathbb{C}(\xi^k) \ominus \bar{R}^k(\xi^k), \quad e_{3,k+1} \triangleq \xi^{k+1} - x^{k+1}, \tag{58}
\]
where \( \omega_k = \alpha \nu^k + \xi^k - \bar{R}^k(\xi^k) \) and \( \bar{R}^k(x) = P\mathcal{B}(R^k(x)) \), i.e., \( \bar{R}^k(x) = [\bar{R}^k_i(x)]_{i \in \mathcal{N}} \) and \( \bar{R}^k_i(x) = P_{\mathcal{B}_i}(R^k_i(x)) \), for \( x \in \mathcal{X} \). Thus, for \( k > 0 \), \( \nu^{k+1} = \lambda^{k+1} + \gamma^{k+1} \) since (12c) is replaced with (20a), and \( \xi^{k+1} = x^{k+1} + e_{3,k+1} \) since (12e) is replaced with (20b). In the rest, we set \( \nu^0 \) to 0.

The following observation will also be useful to prove error bounds for DPDA-TV iterate sequence. Note that (20a) implies for each \( i \in \mathcal{N}, \)
\[
\|e_{1,i}^{k+1}\| \leq \gamma^{k} \|\omega_{i,k}^{k}\| + \gamma^{k} \|\bar{R}^{k}_{i}(\omega^{k})\| \leq \|e_{1,i}^{k}\| + \gamma^{k} \[(1 + \eta^{k})\|\xi_{i,k}^{k}\| + \eta^{k} \|\xi_{i,k}^{k-1}\| + 2\Delta].
\]
Thus, we trivially get the following bound on \( \|\nu^{k+1}\|: \)
\[
\|\nu^{k+1}\| \leq \sum_{i=0}^{k} \gamma^{i} \left( 2\sqrt{N} \Delta + (1 + \eta^{i}) \|\xi_{i}^{i}\| + \eta^{i} \|\xi_{i}^{i-1}\| \right). \tag{59}
\]
Moreover, we will also need the following relation: for any \( \nu \) and \( \lambda \) we have that
\[
\sigma_{\mathcal{C}}(\nu) = \sup_{x \in \mathcal{C}} \langle \lambda, x \rangle + \langle \nu - \lambda, x \rangle \leq \sigma_{\mathcal{C}}(\lambda) + 2\sqrt{N} \Delta \|\nu - \lambda\|. \tag{60}
\]

Definition 9. Let \( T = [A^T, I_{n|\mathcal{N}}] \) for \( A \triangleq \text{diag}([A_i]_{i \in \mathcal{N}}) \in \mathbb{R}^{m \times n|\mathcal{N}|} \). Given \( \alpha, \mu, \delta_1 > 0 \), and arbitrary sequences \( \{\tau^k\}, \{\gamma^k\} \subset \mathbb{R}_{++}, \{\kappa_i^k\}_{k \geq 0} \subset \mathbb{R}_{++} \) for \( i \in \mathcal{N} \), define \( D_{\tau} \triangleq \frac{1}{\tau} \text{diag}([I_n]_{i \in \mathcal{N}}) \), \( D_{\tau} \triangleq \text{diag}([\frac{1}{\tau^k} - (L_i + \alpha)]I_n)_{i \in \mathcal{N}} \), and
\[ Q_k \triangleq \begin{bmatrix} A_k & -\eta^k T^T \\ -\eta^k T & D_{\kappa^k,\gamma^k} \end{bmatrix} \text{ for } k \geq 0, \text{ where } A_k \triangleq (\eta^k)^2 \gamma^k (1 + \delta_1) I_{n|X|} > 0 \text{ and } D_{\kappa^k,\gamma^k} \text{ is defined in Definition 6.} \]

In order to prove Theorem III.2, we first prove Lemma V.6 below which help us to appropriately bound \( L(x^k, y) - L(x^*, y^k) \) for any \( y \in \mathcal{Y} \) and \( \|\xi^k - x^*\| \). That said to show the result in Lemma V.6 we need to show the following two lemmas, Lemma V.4 and Lemma V.5, describing a proper choice for the primal-dual step size sequences.

**Lemma V.4.** Given \( \delta_1 > 0. \) For any \( k \geq 0, \) \( Q_k \geq 0 \) if \( \eta^k > 0, \) and positive numbers \( \{\kappa_i^k\}_{i \in \mathcal{N}}, \) and \( \gamma^k \) are chosen such that
\[
\frac{\kappa_i^k \|A_i\|^2}{\gamma^k} \leq \delta_1, \quad \forall i \in \mathcal{N}. \tag{61}
\]

**Proof.** Let \( D_{\gamma^k} \) be as in Definition 6. Since \( D_{\gamma^k} > 0, \) Schur complement condition implies that \( Q_k \geq 0 \) if and only if
\[
\begin{bmatrix} A_k & -\eta^k A^T \\ -\eta^k A & D_{\gamma^k} \end{bmatrix} - \gamma^k (\eta^k)^2 \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} \succeq 0. \tag{62}
\]

Moreover, since \( D_{\gamma^k} > 0, \) again using Schur complement one can conclude that (62) holds if and only if \( \frac{1}{(\eta^k)^2} A_k - \gamma^k I_n - A^T D_{\gamma^k}^{-1} A \succeq 0. \) This condition holds if (61) is true. \( \square \)

**Lemma V.5.** Let \( D_{\alpha^k} \) be as given in Definition 6, and \( D_{\alpha^k,\gamma^k} \) be as in Definition 9 for \( \alpha > 0 \) chosen according to Lemma III.1 and Remark II.2 and \( \gamma \in (0, \max\{\mu, \mu_\alpha\}). \) Suppose \( \{\tau^k\}, \{\eta^k\}, \{\gamma^k\} \subset \mathbb{R}_{++}, \{\kappa^k_i\}_{i \geq 0} \subset \mathbb{R}_{++}, \) for \( i \in \mathcal{N} \) are chosen as in DPDA-TV displayed in Fig. 2 then the following relations hold for all \( k \geq 0: \)
\[
\begin{align*}
\bar{Q}_k &= \begin{bmatrix} A_k & -\eta^k T^T \\ -\eta^k T & D_{\kappa^k,\gamma^k} \end{bmatrix} \succeq 0, \tag{63a} \\
\gamma^k D_{\alpha^k} &\geq \gamma^k+1 D_{\alpha^k+1}, \quad \tag{63b} \\
\gamma^k D_{\alpha^k} &\geq \gamma^k+1 D_{\alpha^k+1}, \tag{63c} \\
\gamma^k D_{\alpha^k} &\geq \gamma^k+1 D_{\alpha^k+1}, \tag{63d} \\
\gamma^k &\geq \gamma^k+1, \tag{63e} \\
\gamma^k &\geq \gamma^k+1, \tag{63f}
\end{align*}
\]

Moreover, \( \eta^k \in (0, 1), 0 < \frac{1}{\tau^k} < \frac{1}{\tau_\infty} = \mathcal{O}(k), \) and \( 0 < \gamma^k = \mathcal{O}(k). \)

**Proof.** Using the result of Lemma V.4 it is trivial to check that the parameter sequence constructed in Fig. 2 satisfies (63) – see also the discussion in the proof of Lemma V.2. \( \square \)

In order to prove Theorem III.2 we need Lemma V.6 which help us to appropriately bound \( L(x^k, y) - L(x^*, y^k) \) for all \( y \in \mathcal{Y} \) and \( \|\xi^k - x^*\| \) for all \( K \geq 1. \) In particular, Lemma V.6 is similar to Lemma V.3 for the static case, but it also accounts for the approximation errors for the time-varying case, arising due to use of \( \mathcal{R}^k. \)

**Lemma V.6.** Let \( \{\xi^k, y^k\}_{k \geq 0} \) be the iterate sequence generated using Algorithm DPDA-TV as displayed in Fig. 2 which is initialized from an arbitrary \( x^0 \) and \( y^0, \) where \( y^k = [\theta^k] \nu^k \) for \( k \geq 0; \) and let \( \{e_{1}^k\}_{k \geq 1} \) and \( \{e_{2}^k\}_{k \geq 1} \) be the error sequence defined as in (58). For any \( y \in \mathcal{Y}, \)

June 27, 2017
Moreover, it follows from the strong convexity of the objective in (12a) that the iterate sequence 
\[ \{ \mathbf{x}^k, y^k \}_{k \geq 0} \text{ satisfies for all } k \geq 0, \]
\[ \mathcal{L}(\mathbf{x}^{k+1}, y) - \mathcal{L}(\mathbf{x}^*, y^{k+1}) \leq E_{1}^{k+1}(\nu) + E_{2}^{k+1} \]
\[ + \left[ \frac{1}{2} \| \mathbf{x}^* - \mathbf{x}^k \|_{D_{\nu,k}}^2 + \frac{1}{2} \| y - y^k \|_{D_{\nu,k}}^2 - \eta^k \left( T(\mathbf{x}^k - \mathbf{x}^{k-1}), y - y^k \right) + \frac{1}{2} \| \mathbf{x}^k - \mathbf{x}^{k-1} \|_{A_{\nu,k}}^2 \right] \]
\[- \left[ \frac{1}{2} \| \mathbf{x}^* - \mathbf{x}^{k+1} \|_{D_{\nu,k}}^2 + \frac{1}{2} \| y - y^{k+1} \|_{D_{\nu,k}}^2 - \left( T(y^{k+1} - \mathbf{x}^k), y - y^{k+1} \right) + \frac{1}{2} \| y^{k+1} - \mathbf{x}^k \|_{D_{\nu,k}}^2 \right]. \]

where \( E_{1}^{k+1}(\nu) \triangleq \| e^{k+1} \| \left( 4\gamma \sqrt{N} \Delta + \| \nu - \nu^{k+1} \| \right), \text{ and } E_{2}^{k+1} \triangleq \| e^{k+1}_3 \| \left( \frac{2}{\pi} \sqrt{N} \Delta + \alpha \| e^{k+1}_2 \| \right) \]
for \( k \geq 0. \)

**Proof.** Fix \( y = [\theta^T \nu^T]^T \in \mathcal{Y}. \) For \( k \geq 0, \) let \( q^k \triangleq \mathbf{x}^k + \eta^k (\mathbf{x}^k - \mathbf{x}^{k-1}) \) and define \( q_i^k \in \mathbb{R}^n \) for \( i \in \mathcal{N} \) such that \( q_i^k = [q_i^k, \ldots, q_i^N]^T. \) It follows from (12b) that using strong convexity of \( \sigma_C(\nu) - \langle q^k, \nu \rangle + \frac{1}{2\gamma^k} \| \nu - \nu^k \|_2^2 \) in \( \nu \) and the fact that \( \lambda^{k+1} \) is its minimizer, we conclude that
\[ \sigma_C(\nu) - \langle q^k, \nu \rangle + \frac{1}{2\gamma^k} \| \nu - \nu^k \|_2^2 \geq \sigma_C(\lambda^{k+1}) - \langle q^k, \lambda^{k+1} \rangle + \frac{1}{2\gamma^k} \| \lambda^{k+1} - \nu^k \|_2^2 + \frac{1}{2\gamma^k} \| \nu - \lambda^{k+1} \|_2^2. \]

According to (58), \( \nu^{k+1} = \lambda^{k+1} + \gamma^k e^{k+1}_1 \) for all \( k \geq 1; \) hence, from (60) we have
\[ \sigma_C(\nu) - \langle q^k, \nu \rangle + \frac{1}{2\gamma^k} \| \nu - \nu^k \|_2^2 \geq \sigma_C(\nu^{k+1}) - \langle q^k, \nu^{k+1} \rangle + \frac{1}{2\gamma^k} \| \nu^{k+1} - \nu^k \|_2^2 + \frac{1}{2\gamma^k} \| \nu - \nu^{k+1} \|_2^2 - S_{1}^{k+1}(\nu), \]

where the error term \( S_{1}^{k+1}(\nu) \) is defined as
\[ S_{1}^{k+1}(\nu) \triangleq 2\gamma^k \sqrt{N} \Delta \| e^{k+1}_1 \| - \gamma^k \| e^{k+1}_1 \|_2^2 - \langle e^{k+1}_1, \nu - 2\nu^{k+1} + \nu^k + \gamma^k q^k \rangle. \]

Note that for all \( k \geq 0, \) we have \( \nu^k + \gamma^k q^k = \gamma^k \omega^k, \nu^{k+1} = \lambda^{k+1} + \gamma^k e^{k+1}_1, \) and \( \lambda^{k+1} = \gamma^k (\omega^k - P_{\tilde{C}}(\omega^k)). \) Using these we get \( \nu^k + \gamma^k q^k - \nu^{k+1} = \gamma^k (P_{\tilde{C}}(\omega^k) - e^{k+1}_1); \) therefore, (66) can be written as
\[ S_{1}^{k+1}(\nu) = 2\gamma^k \sqrt{N} \Delta \| e^{k+1}_1 \| - \langle e^{k+1}_1, \nu - \nu^{k+1} + \gamma^k P_{\tilde{C}}(\omega^k) \rangle \leq E_{1}^{k+1}(\nu), \]
where the inequality follows from Cauchy-Schwarz and \( \| P_{\tilde{C}}(\omega^k) \| \leq 2\sqrt{N} \Delta \) since \( P_{\tilde{C}}(\omega^k) \in \tilde{C}. \)

Moreover, it follows from the strong convexity of the objective in (12a) that
\[ \sigma_{\mathcal{K}_i}(\theta_i) - \langle A_i q_i^k - b_i, \theta_i \rangle + \frac{1}{2\kappa_i^k} \| \theta_i - \theta_i^k \|_2^2 \]
\[ \geq \sigma_{\mathcal{K}_i}(\theta_i^{k+1}) - \langle A_i q_i^k - b_i, \theta_i^{k+1} \rangle + \frac{1}{2\kappa_i^k} \| \theta_i^{k+1} - \theta_i^k \|_2^2 + \frac{1}{2\kappa_i^k} \| \theta_i - \theta_i^{k+1} \|_2^2. \]

Summing the above inequality over \( i \in \mathcal{N}, \) then summing the resulting inequality with (65) and using (67), we get
\[ h(y) + \frac{1}{2} \| y - y^k \|_{D_{\nu,k}}^2 + E_{1}^{k+1}(\nu) \geq h(y^{k+1}) - \langle T(\mathbf{x}^k + \eta^k (\mathbf{x}^k - \mathbf{x}^{k-1}), y - y^k), y^{k+1} - y \rangle + \frac{1}{2} \| y - y^{k+1} \|_{D_{\nu,k}}^2 + \frac{1}{2} \| y^{k+1} - y^k \|_{D_{\nu,k}}^2. \]

Let \( p^k = T^+ y^k \) for \( k \geq 1. \) Strong convexity of the objective in (12d) implies that
\[ \rho(x^*) + \langle \nabla g(\mathbf{x}^k), x^* \rangle + \langle p^{k+1}, x^* \rangle + \frac{1}{2\gamma^k} \| x^* - \mathbf{x}^k \|_2^2 \geq \rho(x^{k+1}) + \langle \nabla g(\mathbf{x}^k), x^{k+1} \rangle + \langle p^{k+1}, x^{k+1} \rangle + \frac{1}{2\gamma^k} \| x^{k+1} - \mathbf{x}^k \|_2^2 + \frac{1}{2\gamma^k} \| x^* - x^{k+1} \|_2^2, \]

June 27, 2017
Moreover, since June 27, 2017 DRAFT 1 continuous gradient with constant which together with (58) implies that, there exist \( s^{k+1} \in \partial \rho(\xi^{k+1}) \) such that
\[
s^{k+1} + \nabla f(\xi^k) + \alpha(\xi^k - R(\xi^k)) + p^{k+1} + \frac{1}{\tau_k}(\xi^{k+1} - \xi^k) = 0,
\]
which together with (58) implies that,
\[
s^{k+1} + \nabla g(\xi^k) + p^{k+1} = s^{k+1} + \nabla f(\xi^k) + \alpha(\xi^k - R(\xi^k)) + p^{k+1} + \frac{1}{\tau_k}(\xi^{k+1} - \xi^k) - \alpha \epsilon^{2+1}.
\tag{70}
\]
Moreover, since \( \rho(\cdot) \) is a convex function and \( s^{k+1} \in \partial \rho(\xi^{k+1}) \), using (58) we obtain
\[
\rho(s^{k+1}) \geq \rho(s^{k+1}) + \langle s^{k+1}, \xi^k - \xi^{k+1} \rangle = \rho(\xi^{k+1}) - \langle s^{k+1}, \epsilon^{2+1} \rangle.
\tag{71}
\]
Now, using (58) and (71) within (69), we conclude that
\[
\rho(x^*) + \langle \nabla g(\xi^k), x^* \rangle + \langle p^{k+1}, x^* \rangle + \frac{1}{2\tau_k} \| x^* - \xi^k \|^2 \geq \rho(\xi^{k+1}) + \langle \nabla g(\xi^k), \xi^{k+1} \rangle + \langle p^{k+1}, \xi^{k+1} \rangle + \frac{1}{2\tau_k} \| \xi^{k+1} - \xi^k \|^2 + \frac{1}{2\tau_k} \| x^* - \xi^{k+1} \|^2 - S_{k+1}^2,
\tag{72}
\]
where the error term \( S_{k+1}^2 \) is given as follows
\[
S_{k+1}^2 = - \frac{1}{\tau_k} \| e^{k+1} \|^2 + \langle e^{k+1}, s^{k+1} + \nabla g(\xi^k) + p^{k+1} \rangle + \frac{1}{\tau_k} \langle e^{k+1}, 2\xi^{k+1} - \xi^k - x^* \rangle.
\tag{73}
\]
Note that using (70), the definition of \( S_{k+1}^2 \) can be simplified:
\[
S_{k+1}^2 = \langle e^{k+1}, \frac{1}{\tau_k}(x^{k+1} - x^*) - \alpha \epsilon^{2+1} \rangle \leq E_{k+1}^2,
\tag{74}
\]
where we used the fact that \( \| x^{k+1} \| \leq \sqrt{N} \Delta \). In addition, since each \( f_i \) has a Lipschitz continuous gradient with constant \( L_i \) and \( \frac{1}{2}d(\xi) \) has a Lipschitz continuous gradient with constant 1, we have for any \( x \) and \( \bar{x} \) that
\[
g(x) \leq g(\bar{x}) + \langle \nabla g(\bar{x}), x - \bar{x} \rangle + \sum_{i \in N} \frac{L_i + \alpha}{2} \| x_i - \bar{x}_i \|^2.
\tag{75}
\]
Define \( L_q = \text{diag}([L_i + \alpha]_{i \in N}) \). It follows from strong convexity of \( \hat{f} \) that choosing \( \alpha \geq 0 \) according to Lemma III.1 and Remark II.2 we conclude that for any \( \mu \in (0, \max\{\mu, \mu_\alpha\}) \) we have
\[
g(x^*) \geq g(\xi^k) + \langle \nabla g(\xi^k), x^* - \xi^k \rangle + \frac{\mu}{2} \| x^* - \xi^k \|^2
\]
\[
\geq g(\xi^{k+1}) + \langle \nabla g(\xi^k), x^* - \xi^{k+1} \rangle + \frac{\mu}{2} \| x^* - \xi^k \|^2 - \frac{1}{2} \| \xi^{k+1} - \xi^k \|^2 L_q,
\tag{76}
\]
where the last inequality follows from (75). Next, summing inequalities (72) and (76), and using (74), we get
\[
\Phi(x^*) + \frac{1}{2} \| x^* - \xi^k \|^2_{D_{k+1}} + E_{k+1}^2 \geq \Phi(x^{k+1}) + \langle \tau(\xi^{k+1} - x^*), y^{k+1} \rangle + \frac{1}{2} \| x^* - \xi^{k+1} \|^2_{D_{k+1}} + \frac{1}{2} \| \xi^{k+1} - \xi^k \|^2_{D_{k+1}}.
\tag{77}
\]
Next, summing (68) and (77), and rearranging terms, we obtain

\[
\mathcal{L}(\xi^{k+1}, y) - \mathcal{L}(x^*, y^{k+1}) \leq E_1^{k+1}(\nu) + E_2^{k+1} + \eta^k \left( T(\xi^k - \xi^{k-1}), y^{k+1} - y \right) - \frac{1}{2} \left\| y^{k+1} - y \right\|^2_{D_{\nu,k}} + \left[ \frac{1}{2} \left\| x^* - \xi^k \right\|^2_{D_{x,k}} + \frac{1}{2} \left\| y - y^k \right\|^2_{D_{x,k}} \right] - \frac{1}{2} \left\| x^* - \xi^k \right\|^2_{D_{x,k}} + \frac{1}{2} \left\| y - y^k \right\|^2_{D_{x,k}} - \left\langle T(\xi^k - \xi^{k-1}), y - y^k \right\rangle + \frac{1}{2} \left\| \xi^{k+1} - \xi^k \right\|^2_{D_{x,k}}.
\]

(78)

Note that we have,

\[
\eta^k \left\langle T(\xi^k - \xi^{k-1}), y^{k+1} - y \right\rangle = -\eta^k \left\langle T(\xi^k - \xi^{k-1}), y - y^k \right\rangle + \eta^k \left\langle T(\xi^k - \xi^{k-1}), y^{k+1} - y^k \right\rangle;
\]

moreover, the last term can be bounded using the fact that \(Q_k \succeq 0\) as follows:

\[
\eta^k \left\langle T(\xi^k - \xi^{k-1}), y^{k+1} - y^k \right\rangle \leq \frac{1}{2} \left\| y^{k+1} - y^k \right\|^2_{D_{x,k}} + \frac{1}{2} \left\| \xi^{k+1} - \xi^k \right\|^2_{D_{x,k}}.
\]

(79)

Combining (78) and (79) gives the desired result.

Now we are ready to prove Theorem 3.2.

F. Proof of Theorem 3.2

Under Assumption 1.1, a saddle point \((x^*, y^*)\) for \(\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y)\) in (11) exists, where \(y^* = [\theta^*^T, \lambda^*^T]^T\); moreover, any saddle point \((\mathbf{x}^*, \mathbf{\theta}^*, \mathbf{\lambda}^*)\) satisfies that \(x^* = 1 \otimes x^*\) such that \((x^*, \mathbf{\theta}^*)\) is a primal-dual solution to (3). Thus, \(\theta_1^* \in \mathcal{K}_1^0\) and \(\mathcal{L}(x^*, \mathbf{\theta}^*, \mathbf{\lambda}^*) = \Phi(x^*)\). Recall Definition 7, we have \(g(x^*) = f(x^*)\) since \(d_c(x^*) = 0\); hence, \(\Phi(x^*) = \varphi(x^*) = \sum_{i \in \mathcal{N}} \varphi_i(x^*)\). Therefore, \(\mathcal{L}(x^*, \mathbf{\theta}^*, \mathbf{\lambda}^*) = \varphi(x^*)\). Indeed, this implies \(\langle x^*, \mathbf{\lambda}^* \rangle - \sigma^*_C(\lambda^*) = 0\) which leads to \(\sum_{i \in \mathcal{N}} \lambda_i^* = 0\), i.e., \(\mathbf{\lambda}^* \in \mathcal{C}_0\). Therefore, we have \(0 = \langle x^*, \mathbf{\lambda}^* \rangle = \sigma^*_C(\mathbf{\lambda}^*)\). In the rest of the proof, we provide the error bounds for a saddle point \((x^*, \mathbf{\theta}^*, \mathbf{\lambda}^*)\) of \(\mathcal{L}\) such that \(\mathbf{\lambda}^* = 0\). Note that if \((x^*, \mathbf{\theta}^*, \mathbf{\lambda}^*)\) is a saddle point of \(\mathcal{L}\) such that \(\mathbf{\lambda}^* \neq 0\), then it trivially follows that \((x^*, \mathbf{\theta}^*, 0)\) is another saddle point of \(\mathcal{L}\).

Multiplying both sides of (64) by \(\frac{\gamma_k}{\gamma^0}\) and using Lemma V.5, we get

\[
\frac{\gamma_k}{\gamma^0} \mathcal{L}(\xi^{k+1}, y) - \mathcal{L}(x^*, y^{k+1}) \leq \frac{\gamma_k}{\gamma^0} (E_1^{k+1}(\nu) + E_2^{k+1}) + \frac{\gamma_k}{\gamma^0} \left[ \frac{1}{2} \left\| x^* - \xi^k \right\|^2_{D_{x,k}} + \frac{1}{2} \left\| y - y^k \right\|^2_{D_{x,k}} - \eta^k \left\langle T(\xi^k - \xi^{k-1}), y - y^k \right\rangle + \frac{1}{2} \left\| \xi^k - \xi^{k-1} \right\|^2_{D_{x,k}} - \eta^k \left\langle T(\xi^k - \xi^{k-1}), y - y^k \right\rangle + \frac{1}{2} \left\| \xi^{k+1} - \xi^k \right\|^2_{D_{x,k}} \right].
\]

(80)

Next, we sum (80) over \(k = 0\) to \(K - 1\); using Jensen’s inequality and the following facts: \(Q_K \succeq 0\) and \(\xi^{-1} = \xi^0 = x^0\), we get

\[
N_K(\mathcal{L}(\xi^K, y) - \mathcal{L}(x^*, \mathbf{\tilde{y}^K})) \leq \left[ \frac{1}{2} \left\| x^* - x^0 \right\|^2_{D_{x,\nu}} + \frac{1}{2} \left\| y - y^0 \right\|^2_{D_{x,\nu},\gamma} + \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma^0} (E_1^{k+1}(\nu) + E_2^{k+1}) \right] - \frac{\gamma^K}{\gamma^0} \left[ \frac{1}{2} \left\| x^* - \xi^K \right\|^2_{D_{x,K}} + \frac{1}{2} \left\| z^K \right\|^2_{Q_K} \right],
\]

(81)

where \(z^K = ([\xi^K - \xi^{K-1}]^T (y - y^K)^T)^T, N_K = \sum_{k=1}^{K} \frac{\gamma_k}{\gamma^0}, \tilde{\xi^K} = N_K^{-1} \sum_{k=1}^{K} \frac{\gamma_k}{\gamma^0} \xi^k\) and \(\mathbf{\tilde{y}^K} = N_K^{-1} \sum_{k=1}^{K} \frac{\gamma_k}{\gamma^0} y^k\) for \(y^k = [\theta^k^T \nu^k]^T\) for \(k \geq 0\). Note that \(E_1^{k+1}(\nu)\) and \(E_2^{k+1}\) appearing
in (81) are the error terms due to approximating \( P_C \) with \( R^k \) in the \( k \)-th iteration of the algorithm for \( k \geq 0 \). Furthermore, since \( \| z^K \|_{Q_K} \geq 0 \) and \( \tau^K > \tau^k \) for \( k \geq 0 \), (81) can be written more explicitly as follows: for any \( [\theta^T, \nu^T] \in \mathcal{Y} \) and for all \( K \geq 1 \), we have

\[
\mathcal{L}(\xi^K, \theta, \nu) - \mathcal{L}(x^*, \tilde{\theta}^K, \tilde{\nu}^K) \leq \Theta(x^*, \theta, \nu) / N_K, \quad \| \xi^K - x^* \| \leq \frac{\tau^K}{\gamma} \frac{\tau^k}{\gamma} \Theta(x^*, \theta, \nu) \tag{82}
\]

where

\[
\Theta(x^*, \theta, \nu) \triangleq \frac{1}{2} \gamma \| \nu - \lambda_x^0 \|^2 + \sum_{i \in N} \left[ \frac{1}{2} \kappa_i \| x_i^0 - x_i^0 \|^2 + \frac{1}{2} \kappa_i \| \theta_i - \theta_i^0 \|^2 \right] + \sum_{k=0}^{K-1} \frac{\gamma^k}{\gamma_0} \left( \kappa_i E_{k+1}(\nu) + E_{k+1}^2 \right).
\]

Under Assumption [11], one can construct a saddle point \((x^*, \theta^*, \lambda^*)\) for \( \mathcal{L} \) in (11) such that \( \lambda^* = 0 \); hence, \( \mathcal{L}(x^*, \theta^*, \lambda^*) = \varphi(x^*) \) and \( \theta_i^* \in K_i^\circ \) for \( i \in N \). Define \( \tilde{\theta} = [\tilde{\theta}_i]_{i \in N} \) such that \( \tilde{\theta}_i \triangleq 2\| \theta_i^* \| (\| P_{K_i}(A_i \xi_i^K - b_i) \|) \| \mathcal{P}_{K_i}(A_i \xi_i^K - b_i) \| K_i^\circ, \) which implies

\[
\langle A_i \xi_i^K - b_i, \tilde{\theta}_i \rangle = 2\| \theta_i^* \| d_{K_i}(A_i \xi_i^K - b_i).
\]

Note that \( \mathcal{C} \) is a closed convex cone, and the projection \( P_{\mathcal{C}}(x) = 1 \otimes p(x) \) – see [15]. Similarly, define \( \tilde{\nu} = \frac{P_{\mathcal{C}_1}(\xi^K)}{\| P_{\mathcal{C}_1}(\xi^K) \|} \in \mathcal{C}_1^\circ \), where \( \mathcal{C}_1^\circ \) denotes polar cone of \( \mathcal{C}_1 \). Hence, it can be verified that \( \langle \tilde{\nu}^K, \xi^K \rangle = d_{\mathcal{C}}(\xi^K) \). Note that \( \tilde{\nu} \in \mathcal{C}_1^\circ \) implies that \( \sigma_{\mathcal{C}}(\tilde{\nu}) = 0 \); moreover, we also have \( \tilde{\mathcal{C}} \subseteq \mathcal{C} \); hence, \( \sigma_{\mathcal{C}^\circ}(\tilde{\nu}) \leq \sigma_{\mathcal{C}}(\tilde{\nu}) = 0 \). Therefore, we can conclude that \( \sigma_{\mathcal{C}^\circ}(\tilde{\nu}) = 0 \) since \( 0 \in \tilde{\mathcal{C}} \). Together with (83), we get

\[
\mathcal{L}(\xi^K, \tilde{\theta}, \tilde{\nu}) = \Phi(\xi^K) + 2 \sum_{i \in N} d_{K_i}(A_i \xi_i^K - b_i) \| \theta_i^* \| + d_{\mathcal{C}}(\xi^K).
\]

Since, \( x^* \in \tilde{\mathcal{C}} \) we also have that

\[
\langle x^*, \tilde{\nu}^K \rangle - \sigma_{\mathcal{C}^\circ}(\tilde{\nu}^K) \leq \sup_{\nu} \langle x^*, \nu \rangle - \sigma_{\mathcal{C}}(\nu) = I(\mathcal{C})(x^*) = 0.
\]

Note that for any \( i \in N \), \( \tilde{\theta}_i^K \in K_i^\circ \); hence, \( \sigma_{K_i}(\tilde{\theta}_i^K) = 0 \). In addition, since \( \tilde{\theta}_i^K \in K_i^\circ \), and \( A_i x^* - b_i \in K_i \), we have

\[
\langle A_i x^* - b_i, \tilde{\theta}_i^K \rangle \leq 0.
\]

Therefore, (85) and (86) imply

\[
\mathcal{L}(x^*, \theta^K, \nu^K) \leq \varphi(x^*).
\]

Thus, (84), (87) and (82) together with the definitions of \( \tilde{\theta}, \tilde{\nu} \) and the fact that \( \nu^0 = 0 \) and \( \theta^0 = 0 \) imply that

\[
\Phi(\xi^K) - \varphi(x^*) + 2 \sum_{i \in N} \| \theta_i^* \| d_{K_i}(A_i \xi_i^K - b_i) + d_{\mathcal{C}}(\xi^K)
\]

\[
\leq \Theta(x^*, \tilde{\theta}, \tilde{\nu}) / N_K \leq \frac{1}{N_K} \left( \Theta_1 + \sum_{k=0}^{K-1} \frac{\gamma^k}{\gamma_0} \left( E_1^{k+1}(\tilde{\nu}) + E_2^{k+1} \right) \right).
\]

Since \( (x^*, \theta^*, \lambda^*) \) is a saddle-point for \( \mathcal{L} \) in (11), we have \( \mathcal{L}(\xi^K, \theta^*, \lambda^*) - \mathcal{L}(x^*, \theta^*, \lambda^*) \geq 0 \); therefore,

\[
\Phi(\xi^K) - \varphi(x^*) + \sum_{i \in N} \langle \theta_i^*, A_i \xi_i^K - b_i \rangle \geq 0.
\]
Using conic decomposition of $\langle A_i \tilde{\xi}^k - b_i, \theta^*_i \rangle$ and the fact that $\theta^*_i \in K^c_i$, we have that,

$$\langle A_i \tilde{\xi}^k - b_i, \theta^*_i \rangle \leq \| \theta^*_i \| d_{K_i}(A_i \tilde{\xi}^k - b_i).$$

Thus, together with (89), we conclude that

$$\Phi(\xi^K) - \varphi(x^*) + \sum_{i \in N} \| \theta^*_i \| d_{K_i}(A_i \tilde{\xi}^k - b_i) \geq 0. \tag{90}$$

Provided that we show $\Theta_1 + \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma_0} \left( E_{k+1}^k(\tilde{v}) + E_{k+1}^{k+1} \right) \leq \Theta(K)$ for some $\Theta(K) = \mathcal{O}\left( \sum_{k=0}^{K-1} \beta^{q_k} k^4 \right)$, the desired result in (21) follows from (88) and (90). Moreover, the bound on $\| \xi^K - x^* \|$ follows from (82). In fact, possibly a tighter bound can be derived using $\Theta(\mathbf{x}^*, \theta^*, \lambda^*)$ for $\lambda^* = 0$. In the rest of the proof, we construct the $\Theta(K)$ bound with properties as specified above.

Note that using (16) and the non-expansivity of projection, $\mathcal{P}_B(\cdot)$, we conclude that

$$\| \tilde{R}^k(\mathbf{x}) - \mathcal{P}_C(\mathbf{x}) \| \leq N \Gamma \beta^{q_k} \| \mathbf{x} \| \quad \forall \mathbf{x}. $$

Moreover, since we assumed that each $\rho_i$ has a compact domain with diameter at most $\Delta$, we immediately conclude that $\| \mathbf{x}^k \| \leq \sqrt{N} \Delta$ and $\| \xi^k \| \leq \sqrt{N} \Delta$ for $k \geq 1$. Hence, from (58) and nonexpansivity of prox operator we obtain

$$\| e^{k+1}_i \| \leq \alpha \| e^{k+1}_2 \| \leq \alpha N \Gamma \beta^{q_k} \| \xi^k \| \leq \alpha N^{\frac{1}{2}} \Delta \Gamma^{\beta^{q_k}}. \tag{91}$$

Let $q^k = \xi^k + \eta^k(\xi^k - \xi^{k-1})$ for $k \geq 0$. Note that for $\{\eta^k\}$ as specified in Algorithm DPDA-TV displayed in Fig. 2, we have $\eta^k \leq 1$. Therefore, it follows from (58) and (59) that

$$\| e^{k+1}_i \| = \| \mathcal{P}_C(\frac{1}{\gamma} \nu^k + q^k) - \tilde{R}^k(\frac{1}{\gamma} \nu^k + q^k) \| \leq N \Gamma \beta^{q_k} \left( \frac{1}{\gamma} \nu^k + q^k \right) \leq N \gamma \Gamma \beta^{q_k} \left( \frac{5 \sqrt{N} \Delta}{\gamma} \sum_{t=0}^{k-1} \gamma^t + 3 \sqrt{N} \Delta \right) = N \gamma \Gamma \beta^{q_k} \left( \frac{5 k}{\gamma} \sum_{t=0}^{k-1} \gamma^t + 3 \right). \tag{92}$$

Therefore, using (59) and (92) we obtain

$$\sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma_0} E_{k+1}(\tilde{v}) = \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma_0} \| e^{k+1}_i \| \left( 4 \gamma^k \sqrt{N} \Delta + \| \tilde{v} - \nu^{k+1} \| \right) \leq \sum_{k=0}^{K-1} \frac{\Delta}{\gamma_0} N^{\frac{1}{2}} \Gamma \beta^{q_k} \left( 5 \sum_{t=0}^{k-1} \gamma^t + 3 \gamma^k \right) \left( 4 \gamma^k \sqrt{N} \Delta + 1 + 5 \sqrt{N} \Delta \sum_{t=0}^{k-1} \gamma^t \right) \leq \sum_{k=0}^{K-1} \frac{\Delta^2}{\gamma^2_0} N^2 \Gamma \beta^{q_k} \left( 5 \sum_{t=0}^{k-1} \gamma^t + 3 \gamma^k \right) \left( 5 \sum_{t=0}^{k-1} \gamma^t + 1 + 4 \gamma^k \right) \leq \Theta_2(K). \tag{93}$$

Moreover, from (91) we obtain

$$\sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma_0} E_{k+1}^2 = \sum_{k=0}^{K-1} \frac{\gamma_k}{\gamma_0} \left( \frac{2}{\gamma} \sqrt{N} \Delta + \alpha \| e^{k+1}_2 \| \right) \leq \sum_{k=0}^{K-1} \alpha N^{\frac{1}{2}} \Delta \Gamma \beta^{q_k} \frac{\gamma_k}{\gamma_0} \left( \frac{2}{\gamma} \sqrt{N} \Delta + \alpha N^{\frac{1}{2}} \Delta \Gamma \beta^{q_k} \right) \leq \sum_{k=0}^{K-1} \alpha N^2 \Delta \Gamma \beta^{q_k} \frac{\gamma_k}{\gamma_0} \left( \frac{2}{\gamma} \sqrt{N} \Delta + \alpha N^{\frac{1}{2}} \Delta \Gamma \beta^{q_k} \right) \leq \Theta_3(K). \tag{94}$$

Therefore, by letting $\Theta(K) = \Theta_1 + \Theta_2(K) + \Theta_3(K)$ it is easy to see that $\Theta(K) = \mathcal{O}(\sum_{k=0}^{K-1} \beta^{q_k} k^4)$; thus, $\sup_{K \in \mathbb{Z}_+} \Theta(K) < \infty$ due to our choice of $\{q_k\}$, and this completes the proof.