Distributed Regularized Primal-Dual Method

Masoud Badiei, Na Li*
Harvard University, Cambridge, MA, USA

June 14, 2016

Abstract

We study deterministic and stochastic primal-dual subgradient methods for distributed optimization of a sepearable objective function with global constraints. In both algorithms, the norm of dual variables is controlled by augmenting the corresponding Lagrangian function with a regularizer on the dual variables. Specifically, for each underlying algorithm we show that as long as its step size satisfies a certain restriction, the norm of dual variables is inversely proportional to the regularizer's curvature. In the deterministic optimization case, we leverage the bound on dual variables to analyze the consensus terms and subsequently establish the convergence rate of the distributed primal-dual algorithm. In the stochastic optimization case, the bound on dual variables is further used to derive a high probablity bound on the convergence rate via the method of bounded martingale difference. For both algorithms, we exhibit a tension between the convergence rate of underlying algorithm and the decay rate associated with the constraint violation.

1 Introduction

Network-structured optimization is a framework to distribute the computational complexity of solving an optimization problem among many nodes in a network. In such a framework, each node i in the network is assigned with a local objective function $f_i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$. Further, each node coordinates its actions with other nodes through exchanging local information with adjacent nodes in the network. In this paper, we study a distributed primal-dual algorithm to optimize a separable convex objective function subject to a set of global inequality constraints

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{1a}$$

$$\text{subject to: } g(\mathbf{x}) \preceq \mathbf{0}, \tag{1b}$$

where $g(\mathbf{x}) \equiv (g_1(\mathbf{x}), \cdots, g_m(\mathbf{x}))$ and $g_k : \mathbb{R}^d \to \mathbb{R}$ are convex constraints, $\mathcal{X} \subseteq (\cap_{k=1}^{m} \text{dom } g_k) \cap (\cap_{i=1}^{n} \text{dom } f_i)$ is non-empty, convex, and closed subspace of $\mathbb{R}^d$, and $\preceq$ denotes the element-wise inequality. In particular, we examine the effect of inequality constraints in Eq. (1b) on the convergence rate of the underlying distributed algorithm.

*Email: mbadieikhuzani@g.harvard.edu,nali@seas.harvard.edu.

Interest in obtaining efficient distributed algorithms for the framework in Eqs.(1a)-(1b) has been re-stimulated by large-scale problems arising in statistics, machine learning, and related areas [CV95]. In particular, a prevalent problem in statistical machine learning is to learn from and make predictions about observed data in a high dimensional data set by minimizing a loss function. However, processing data in a centralized fashion on large data sets is at best inefficient, and often infeasible. An alternative approach is thus to spread the processing task among many computing nodes, where each node has only access to a subset of data set.

In addition to machine learning problems, distributed optimization has been used in variety of domains including the following applications:

Source Localization. Localization problem is concerned with pinpointing the unknown location of a target $\mathbf{x}$ that emits signal isotropically. There are two common techniques for measuring distances between wireless devices, namely Received Signal Strength Indicator (RSSI) and Time Difference of Arrival (TDoA). RSSI measures the ratio of the power present in a received radio signal (P_r) and a reference transmitted power (P_s). The ratio P_r/P_s is inversely proportional to the square of the distance between the receiver and the transmitter. Hence, RSSI can be used to estimate the distance to the target. In particular, a number of detectors are deployed where the received signal energy measurment at the detector i can be described as [WL09],

$$\frac{P_r^i}{P_s} = \frac{C}{\|\mathbf{x} - \mathbf{r}_i\|^\alpha} + w_i.$$

Here, $\mathbf{r}_i$ is the location of the ith detector relative to a fixed reference, w_i is the additive measurement noise, and C and α are two constants. In this case, the optimization problem in Eqs. (1a)-(1b) can be recast into

$$\min_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{n} \left(\frac{P_r^i}{P_s} - \frac{C}{\|\mathbf{x} - \mathbf{r}_i\|^\alpha} \right)^2,$$

where the constraint $\mathcal{X} \subset \mathbb{R}^d$, $d \in \{2, 3\}$ specifies the search region. A similar optimization problem can be characterized for TDoA method.

Estimation in Sensor networks. Estimation problems in sensor networks can be framed as a special case of Eqs. (1a)-(1b). In the estimation problem, the goal is to estimate a parameter $\theta \in \Theta$ over a sensing field. Let $\mathbf{y}_i$ denotes the ith sensor's local measurment of the parameter. The estimation problem can be formulated as follows

$$\min_{\theta\in\Theta} \sum_{i=1}^{n} f_i(\theta; \mathbf{y}_i).$$

A possible choice of the loss function is the Huber function for robust estimation

$$f_i(\theta; \mathbf{y}_i) = \begin{cases} \frac{1}{2}(\theta - \mathbf{y}_i)^2 & \text{if } \|\theta - \mathbf{y}_i\| < z \\ z(\theta - \mathbf{y}_i) - \frac{z^2}{2} & \text{if } \|\theta - \mathbf{y}_i\| \geq z, \end{cases}$$

where z is a constant.

Congestion Control. Consider a network with $\mathcal{L} = \{1, 2, \cdots, L\}$ links that are shared among S sources, where $\mathcal{S} = \{1, 2, \cdots, S\}$. Each source $s \in \mathcal{S}$ is characterized by the tuple $(U_s, \mathcal{L}(s), m_s, M_s)$, where $U_s : \mathbb{R}_+ \to \mathbb{R}$ is the utility function, $\mathcal{L}(s) \subset \mathcal{L}$ is the set of links used by the source, and m_s and M_s are respectively the minimum and maximum transmission rates. Based on this description, the congestion control problem then takes the following form

$$\begin{aligned} \max_{x_1,\cdots,x_s} & \sum_{s=1}^{S} U_s(x_s) \\ & \sum_{s\in\mathcal{S}(l)} x_s \le c_l, \quad l = 1, 2, \cdots, L, \\ & m_s \le x_s \le M_s, \quad s = 1, 2, \cdots, S, \end{aligned}$$

where $\mathcal{S}(l) = \{s \in \mathcal{S} | l \in \mathcal{L}(s)\}$ is the set of sources that share the link l, and c_l is the capacity of the link $l \in \mathcal{L}$. Clearly, the congestion control problem can be written in the form of Eqs.(1a)-(1b). In particular, given the transmission rate vector $\mathbf{x} \equiv (x_1, \cdots, x_s)^T$ the problem in Eqs.(1a)-(1b) can be specialized as $f_s(\mathbf{x}) = U_s(x_s)$, $\mathcal{X} = \prod_{s=1}^{S}[m_s, M_s]$ and $g(\mathbf{x}) = A\mathbf{x} - c$, where $c = (c_1, \cdots, c_L)^T$ and $A = [A_{ls}] \in \mathbb{R}^{L\times S}$ is an incident matrix such that $A_{ls} = 1$ if $x_s \in \mathcal{S}(l)$ and $A_{ls} = 0$ otherwise.

1.1 Related Works

Early works on the distributed optimization were focused on the extremization of a smooth convex function $f(\mathbf{x})$ through distribution of the decision variable vector $\mathbf{x} \in \mathbb{R}^n$ among n different nodes, cf. Tsitsiklis *et al.* [Tsi84], Bertsekas and Tsitsiklis [BT89]. In contrast, later studies developed a framework for optimizing separable objective functions of the form $f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$, where the ith node in the network has information about $f_i(\cdot)$ only.

In the latter line of research, many distributed algorithms have been developed and analyzed in the *unconstrained* settings [NO09a, Ols14, JXM14]. Nedić and Ozdaglar [NO09a] studied a distributed subgradient method where the underlying algorithm utilizes a constant step size and hence, there is a constant error gap present in the convergence bound. In the case of a smooth objective function, the convergence rate can be greatly improved through a distributed Nesterov algorithm [JXM14] which provides an appealing convergence rate $\mathcal{O}(\log(T)/T)$ along with a scaling factor $\mathcal{O}(\sqrt{n})$.

In the context of *constrained* distributed optimization, Duchi *et al.* [DAW12] have studied a dual averaging algorithm where there is a global constraint set $\mathcal{X}$ on agents' actions. However, the step size used in the proposed method depends on the spectral gap of the network. That is, each node must have knowledge of the underlying network structure which poses a challenge in time-varying networks. In the case of optimization with coupled linear equality constraints, i.e., when decision variables of nodes must jointly satisfy a set of linear equality constraints, penalty and barrier function methods are established [LM14]. Moreover, using a game theoretic argument, the convergence of the proposed algorithms are proved.

For distributed optimization with a set of global non-linear inequality constraints, a distributed primal-dual method similar to this paper is studied in [YXZ11]. However, the

proposed method requires projection of the dual variables onto a simplex at each algorithm iteration whereas in our framework this projection is onto the non-negative orthant of the Euclidean space. As a result, the projection is greatly simplified in our proposed scheme. More importantly, the error bound of the PD algorithm in [YXZ11] depends on the quality of the Slater vector for the inequality constraints, i.e., it depends to the inverse of the value $\max_{k=1,2,\cdots,m} g_k(\hat{\mathbf{x}})$, where $\hat{\mathbf{x}}$ is a Slater vector that satisfies $g(\hat{\mathbf{x}}) \prec \mathbf{0}$. However, reliance on the Slater vector is unappealing as it ties the algorithm performance to the structure of the feasible set. We resolve this issue by regularizing the Lagrangian with a smooth and strongly convex function of the dual variables.

1.2 Our Contributions

We study a distributed primal-dual (PD) subgradient method for optimization over a network of fixed topology subject to a set of inequality constraints. Our study is inspired by the work of Mahdvai, *et al.* [MJY12] where it has been shown that a quadratic regularization of the dual variables in an online PD algorithm can achieve a sublinear 'regret' and simultaneously guarantee a vanishing long-term constraint violation. However, the approach in [MJY12] is not easily applicable to the distributed setting as it does not provide a bound on the subgradients of the Lagrangian function. It turns out that this bound is essential in analyzing the 'consensus terms' in the distributed primal-dual method. Moreover, in the distributed stochastic primal-dual method, the bound on subgradients further plays a cruicial role in deriving a high probability bound for the convergence rate by using the concenteration inequalities.

Therefore, herein we take a different approach from the work of Mahdvai, *et al.* [MJY12]. In particular, we establish an upper bound on the norm of dual variables that is modulated by the inverse of regularizer's curvature. In turn, this upper bound allows us to upper bound the subgradients of the Lagrangian function. Moreover, we characterize our result for a general form of regularizer which subsumes the quadratic regularizer in [MJY12] as a special case. Our approach also reveals the tension between the convergence rate of the PD algorithm and the corresponding constraint violation performance. In particular, we show that achieving a fast convergence rate results in a slow decay rate of the constraint violation and vice versa.

We summarize our contributions as follows:

- In both the deterministic and stochastic optimization cases, we establish an upper bound on the norm of dual variables that is inversely proportional to the regularizer's curvature.
- We determine the convergence rate of the distributed regularized primal-dual method. We also derive two asymptotic bounds on the constraint violation performance.
- We characterize the trade-off between the convergence rate of the distributed PD algorithm and the corresponding constraint violation performance. In particular, we show that increasing the speed of convergence degrades the decay rate associated with the constraint violation and vice versa.

- We decribe and analyze a distributed *stochastic* primal-dual method to reduce the computational complexity of its deterministic variant. Specifically, we show that by randomizing the distributed PD method, each node in the network only needs to compute the subgradient of its local objective function and one constraint at each algorithm iteration.

- We use the method of bounded margingale difference to derive a high probability bound on the convergence rate of the stochastic PD method.

1.3 Organization

The rest of this paper is organized as follows. In Section 2 we define the problem setting and describe a distributed, regularized PD algorithm for constrained optimization. In Section 3, we state our main results, including a theorem that establishes the convergence rate of the distributed regularized PD algorithm. In Section 4, we compare the results for the convergence rate of regularized PD algorithm with another distributed algorithm for constrained optimization problem, namely the dual averaging algorithm. In Section 5 we verify our theoretical studies with the numerical simulations. In Section 6 we discuss our results and conclude the paper.

Notation: For ease of notation, we denote the ℓ_2-norm by $\|\cdot\|$. However, we use the standard notation $\|\cdot\|_1$ for the ℓ_1-norm. Furthermore, we denote the *dual norm* by $\|\cdot\|_*$ which is defined as $\|\mathbf{x}\|_* \equiv \sup_{\|\mathbf{y}\|=1}\langle \mathbf{x}, \mathbf{y}\rangle$. We also use standard asymptotic notation. If a_n and b_n are positive sequences, then $a_n = \mathcal{O}(b_n)$ means that $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$. Similarly, $a_n = \Omega(b_n)$ denotes $\liminf \frac{a_n}{b_n} > 0$. When, $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$, we write $a_n = \Theta(b_n)$. For non-neagative sequences a_n and b_n, $a_n \lesssim b_n$ indicates the inequality $a_n \leq c \cdot b_n$ for all $n \in \mathbb{N}$ and for some constant $c < \infty$. We denote the vectors as $a \equiv (a_1, a_2, \cdots, a_n)$. For two vectors a and b, the vector inequality $a \preceq b$ means the element-wise inequality, i.e., $a_i \leq b_i$ for all $i = 1, 2, \cdots, n$. Lastly, we denote the projection of the vector $\mathbf{x}$ onto the closed set $\mathcal{X}$ by $\Pi_{\mathcal{X}}(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_2$.

2 Preliminaries

We consider a multi-agent optimization problem, consisting of n nodes that exchange information on the edge of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a *fixed* topology, where $\mathcal{V} = \{1, 2, \cdots, n\}$ denotes the set of vertices, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges defined on the vertices. At each iteration $t \in [T] \equiv \{1, 2, \cdots, T\}$ of the distributed algorithm, agent $i \in \mathcal{V}$ takes an action $\mathbf{x}_i^t \in \mathcal{X} \subset \mathbb{R}^d$ based on knowledge of a local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$. We also consider a set of global inequality constraints $g_k(\mathbf{x}_i) \leq 0, k \in [m] \equiv \{1, 2, \cdots, m\}$ on the actions of each agent. The objective of agents is to cooperatively minimize the global loss function $f(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ while satisfying the inequality constraints.

More concretely, we study a distributed primal-dual algorithm for the optimization problem in Eqs. (1a)-(1b), where we assume the subspace $\mathcal{X}$ is known at each node of the network and has a finite diameter $R \equiv \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$. To simplify the analysis, we further assume that $\mathbf{0} \in \mathcal{X}$. This last requirement is always attainable by a simple translation

$\varphi : \mathcal{X} \to \mathcal{X} + \Delta, \mathbf{x} \mapsto \mathbf{x} + \Delta$ for some $\Delta \in \mathbb{R}^d$ and optimizing the composite functions $\tilde{f}_i \equiv f_i \circ \varphi^{-1}$ and $\tilde{g}_k \equiv g_k \circ \varphi^{-1}$.

We assume that $f_i : \mathbb{R}^d \to \mathbb{R}, i \in \mathcal{V}$ and $g_k : \mathbb{R}^d \to \mathbb{R}, k \in [m]$ are convex functions. Furthermore, we assume that f_i and g_k are Liptschitz continuous, i.e.,

$$|f_i(\mathbf{x}) - f_i(\mathbf{y})| \leq L_{f_i}\|\mathbf{x} - \mathbf{y}\|, \quad i = 1, 2, \cdots, n,$$
$$|g_k(\mathbf{x}) - g_k(\mathbf{y})| \leq L_{g_k}\|\mathbf{x} - \mathbf{y}\|, \quad k = 1, 2, \cdots, m.$$

Let $L \equiv \max\{L_{f_1}, \cdots, L_{f_n}, L_{g_1}, \cdots, L_{g_m}\}$. We notice that since $g_k(\cdot)$ are Lipschitz continuous on $\mathcal{X}$, they are bounded. In particular, the Liptschitz continuity assumption of $g_k(\cdot)$ implies

$$|g_k(\mathbf{x})| \leq L_{g_k} R \leq LR, \quad k \in [m]. \tag{2}$$

We denote the optimal solution and optimal Lagrangian multiplier associated with the problem in Eqs. (1a)-(1b) with $\mathbf{x}^*$ and λ^*, respecitively. The constrained optimization problem in Eqs. (1a)-(1b) can be reformulated as a saddle point problem

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\lambda \in \mathbb{R}^m_+} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) + \langle \lambda, g(\mathbf{x}) \rangle. \tag{3}$$

where $\lambda \equiv (\lambda_1, \lambda_2, \cdots, \lambda_m)$.

Based on this reformulation, we design a distributed primal-dual algorithm for Eq. (3) such that the inequality constraint $g(\mathbf{x}) \preceq 0$ are satisfied *asymptotically* as $T \to \infty$. The distributed method we present is based on the regularization of the dual variables λ that is formalized in the following definition.

Definition 1. An admissible regularizer $\psi(\cdot) : \mathbb{R}^m \to \mathbb{R}, t \in [T]$ is characterized by the following three conditions:

(i) $\psi(\lambda) \geq 0$, $\psi(\mathbf{0}) = 0$ and $\langle \nabla\psi(\mathbf{0}), \lambda \rangle \geq 0$ for all $\lambda \in \mathbb{R}^m_+$.

(ii) $\psi(\lambda)$ is η-strongly convex with respect to the induced norm $\|\cdot\|$,

$$\psi(\lambda) - \psi(\hat{\lambda}) - \langle \nabla\psi(\hat{\lambda}), \lambda - \hat{\lambda} \rangle \geq \frac{\eta}{2}\|\lambda - \hat{\lambda}\|^2, \quad \forall \lambda, \hat{\lambda} \in \mathbb{R}^m_+, \tag{4}$$

(iii) $\psi(\lambda)$ is γ-smooth function with respect to the induced norm $\|\cdot\|$,

$$\psi(\lambda) - \psi(\hat{\lambda}) - \langle \nabla\psi(\hat{\lambda}), \lambda - \hat{\lambda} \rangle \leq \frac{\gamma}{2}\|\lambda - \hat{\lambda}\|^2, \quad \forall \lambda, \hat{\lambda} \in \mathbb{R}^m_+.$$

Definition 2. The condition number associated with the regularization ψ is defined as the ratio of the smoothness constant γ and the regularizer's curvature η, i.e., $Q_\psi \equiv \gamma/\eta$.

Conditions *(i)* and *(ii)* in Definition 1 are standard requirements of a regularizer (e.g. see [DAW12]). Condition *(iii)*, however, is an additional restriction that we put in order to provide an upper bound on the norm of the Lagrangian dual variables $\|\lambda\|$ (cf. Thm. 1). It

is easy to verify that the squared ℓ_2-norm regularizer $\psi(\lambda) = \theta\|\lambda\|_2^2/2$ satisfies the specified conditions with $\eta = \gamma = \theta$ and is thus admissible. We also note that in the case that the regularizer ψ is twice continuously differentiable, the condition number Q_ψ in Definition 2 corresponds to the ratio of the largest and smallest eigenvalues of the Hessian matrix of ψ. For example, for the quadratic function $\psi(\lambda) = \theta\|\lambda\|_2^2/2$ the condition number is $Q_\psi = 1$.

To simplify the analysis, in the following we assume the regularizer satisfies $\nabla\psi(\mathbf{0}) = 0$. However, the more general case $\langle\nabla\psi(\mathbf{0}), \lambda\rangle \geq 0$ for $\lambda \in \mathbb{R}_+^m$ can be treated similarly.

Based on the definition of the admissible regularizer ψ, we define the augmented Lagrangian as follows

$$\mathfrak{L}_i(\mathbf{x}_i, \lambda_i) \equiv f_i(\mathbf{x}_i) + \langle\lambda_i, g(\mathbf{x}_i)\rangle - \psi(\lambda_i), \tag{5}$$

where $\lambda_i \equiv (\lambda_{i,1}, \lambda_{i,2}, \cdots, \lambda_{i,m})$. Furthermore,

$$\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i, \lambda_i) \equiv \nabla f_i(\mathbf{x}_i) + \langle\lambda_i, \nabla g(\mathbf{x}_i)\rangle \tag{6a}$$
$$\nabla_{\lambda}\mathfrak{L}_i(\mathbf{x}_i, \lambda_i) \equiv g(\mathbf{x}_i) - \nabla\psi(\lambda_i). \tag{6b}$$

Note that in the case that functions f_i and g_k are not differentiable, we use their corresponding subgradients. However, for ease of notation, we use $\nabla f_i(\mathbf{x}_i)$ and $\nabla g_k(\mathbf{x}_i)$ to denote both gradient and subgradient when f_i and g_k are differentiable and non-differentiable, respectively. In the latter case, we define the set of subgradients of f_i and g_k as follows

$$\partial f_i \equiv \{\nu \in \mathbb{R}^d : f_i(\mathbf{x}) - f_i(\mathbf{x}_0) \geq \langle\nu, \mathbf{x} - \mathbf{x}_0\rangle, \forall \mathbf{x}, \mathbf{x}_0 \in \text{dom } f_i\}, \quad i \in \mathcal{V}$$
$$\partial g_k \equiv \{\hat{\nu} \in \mathbb{R}^d : g_k(\mathbf{x}) - g_k(\mathbf{x}_0) \geq \langle\hat{\nu}, \mathbf{x} - \mathbf{x}_0\rangle, \forall \mathbf{x}, \mathbf{x}_0 \in \text{dom } g_k\}, \quad k \in [m].$$

Based on the definition of $\mathfrak{L}_i(\cdot,\cdot)$, we solve the regularized min-max problem characterized below

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\lambda\in\mathbb{R}_+^m} \frac{1}{n}\sum_{i=1}^{n} \mathfrak{L}_i(\mathbf{x}_i, \lambda_i). \tag{7}$$

To describe the distributed primal-dual algorithm, we consider a weight matrix $W \equiv [W]_{ij}$ that fulfills the following conditions:

- (*Doubly stochastic*) The weight matrix is doubly stochastic,

  $$W \times \mathbb{1}_n = \mathbb{1}_n, \quad \mathbb{1}_n^T \times W = \mathbb{1}_n^T,$$

  where $\mathbb{1}_n \in \mathbb{R}^n$ is the column vector with all elements equal to one.

- (*Connectivity*) The weight matrix respects the graph topology

  $$W_{ij} > 0 \quad \text{if} \quad (i,j) \in \mathcal{E}$$
  $$W_{ij} = 0 \quad \text{if} \quad (i,j) \notin \mathcal{E}.$$

REMARK 1. For $n \times n$ doubly stochastic matrices, the singualr values can be sorted in a non-increasing fashion $\sigma_1(W) \geq \sigma_2(W) \geq \cdots \geq \sigma_n(W) \geq 0$, where $\sigma_1(W) = 1$ (cf. [HJ12]). This is due to the fact that for a doubly stochastic matrix $\mathbb{1}_n$ is both the left and right eigenvector, i.e., $W\mathbb{1}_n = \mathbb{1}_n$ and $\mathbb{1}_n^T W = \mathbb{1}_n^T$. Throughout the paper, we refer to $1 - \sigma_2(W)$ as the spectral gap of the matrix W.

Algorithm 1 DISTRIBUTED DETERMINISTIC PRIMAL-DUAL METHOD

1: **Initialize**: $\mathbf{x}_i^0 = \mathbf{0}$, $\lambda_i^0 = \mathbf{0}, \forall i \in \mathcal{V}$ and a constant step size $\alpha \in \mathbb{R}_+$.
2: **for** $t = 0, 1, 2, \cdots, T$ at the i-th node **do**
3: Update the primal and dual variables

$$\widehat{\mathbf{x}}_i^t = \mathbf{x}_i^t - \alpha \nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)$$
$$\widehat{\lambda}_i^t = \lambda_i^t + \alpha \nabla_{\lambda} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t).$$

4: Run the consensus steps

$$\mathbf{x}_i^{t+1} = \Pi_{\mathcal{X}} \left(\sum_{j=1}^{n} [W]_{ij} \widehat{\mathbf{x}}_j^t \right),$$
$$\lambda_i^{t+1} = \Pi_{\mathbb{R}_+^m} \left(\sum_{j=1}^{n} [W]_{ij} \widehat{\lambda}_j^t \right).$$

5: **end for**
6: **Output**: $\tilde{\mathbf{x}}_i^T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_i^t$ for all $i \in \mathcal{V}$.

We are now in position to describe the distributed algorithm for solving the regularized min-max formulation in Eq. (7); see Algorithm 1.

Discussion: The intuition underlying using augmented Lagrangian in Eq. (5) is as follows. Since the vector of Lagrangian dual variables λ_i take values from $\mathbb{R}_+^m$, the subgradients defined in Eqs. (6a) and (6b) can be unbounded which imposes a challenge for the convergence analysis of centeralized and distributed primal-dual method. To circumvent this issue, Nedić and Ozdaglar [NO09b] rely on the *Slater constraint qualification* to bound the dual variables. More specifically, suppose there exists a vector $\hat{\mathbf{x}}$ such that

$$g(\hat{\mathbf{x}}) \prec \mathbf{0}. \tag{8}$$

Furthermore, define

$$\mathfrak{F}(\lambda) \equiv \inf_{\mathbf{x} \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) + \langle \lambda, g(\mathbf{x}) \rangle.$$

Then it is shown that [NO09b],

$$\|\lambda^*\|_1 \leq \mu^{-1} \times (f(\hat{\mathbf{x}}) - \inf_{\lambda \in \mathbb{R}_+^m} \mathfrak{F}(\lambda)), \tag{9}$$

where $\mu \equiv \min_{k=1,2,\cdots,m} \{-g_k(\hat{\mathbf{x}})\}$, and λ^* is the optimal dual variable. The same strategy has also been adapted in [YXZ11] in order to upper bound the dual variables λ_i^t in the distributed primal-dual method. Specifically, in Ref. [YXZ11] a similar min-max problem as in Eq. (3) is considered albeit without the regularization ψ term. Nevertheless, the approach adopted in [YXZ11] has two main drawbacks in the distributed systems:

- In the proposed primal-dual algorithm in [YXZ11], each agent must project the local computation of dual variables λ_i^t onto the following simplex

$$\Lambda \equiv \{\lambda \in \mathbb{R}^m : \|\lambda\|_1 \leq \mu^{-1} \cdot (f(\hat{\mathbf{x}}) - \mathfrak{F}(\hat{\lambda}))\}, \tag{10}$$

  where $\hat{\lambda} \in \mathbb{R}_+^m$ is an arbitrary vector. In contrast, the projection in Algorithm 1 is onto the non-negative orthant $\mathbb{R}_+^m$ which simply corresponds to replacing each negative component of the dual vector λ_i^t with zero. Since the projection step must be executed at each algorithm iteration, our proposed strategy is significantly more efficient.

- The size of the projection set Λ is inversely proportional to μ and the value of μ can be small when $g_k(\hat{\mathbf{x}})$ is small for at least one coordinate $k \in \{1, 2, \cdots, m\}$. This, in turn, results in loose upper bounds on the norm of the gradients in Eqs. (6a) and (6b) and hence a loose convergence bound for the underlying primal-dual algorithm. The upper bound on the constraint violation in [YXZ11] also depends on $\mu^{-1} \cdot (f(\hat{\mathbf{x}}) - \inf_{\lambda \in \mathbb{R}_+^m} \mathfrak{F}(\lambda))$ which can be loose due to the same reason.

3 Main Results

3.1 Distributed Deterministic Primal-Dual Algorithm

In this section, we state our main results. As our first result, we prove a theorem that supplies us with an upper bound on the norm of the Lagrangian dual variables:

Proposition 1. *Under the restriction* $0 < \alpha \leq \frac{1}{2Q_\psi^2 \eta}$ *on the step size of Algorithm 1, the norm of the Lagrangian dual* λ_i^t *is bounded by*

$$\|\lambda_i^t\| \leq \frac{2LR\sqrt{nm}}{\eta}, \tag{11}$$

for all $t \in [T]$. *Specifically, for the choice of* $\eta = \frac{2LR\sqrt{nm}}{\beta}, \beta > 0$ *we have*

$$\|\lambda_i^t\| \leq \beta. \tag{12}$$

The proof of Proposition 1 can be found in Appendix A.2. Proposition 1 highlights the role that the regularizer ψ plays in the augmented Lagrangian $\mathfrak{L}_i(\cdot, \cdot)$. Specifically, the curvature η of regularizer ψ provides a degree of freedom to control the norm of the dual variable λ_i^t in the primal dual method. The upper bound in Eq. (11) is also intuitive. As η becomes larger, the cost associated with choosing a large lagrangian dual variable λ_i^t increases which results in a smaller norm $\|\lambda_i^t\|$.

We now use the result of Theorem 1 to compute upper bounds on the norm of subgradients of $\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)$.

Corollary 2. *For all for all* $t \in [T]$,

$$\|\nabla_\lambda \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\| \leq 3LRQ_\psi\sqrt{mn}$$
$$\|\nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\| \leq L\tilde{\beta},$$

where $\tilde{\beta} \equiv 1 + \beta\sqrt{m}$.

Next, we leverage the result of Corollary 2 to bound the 'consensus' terms $\|\mathbf{x}_i^t - \mathbf{x}_j^t\|$ which is a measure of deviation between agents' decision variables. Specifically, the next theorem provide an upper bound on the consensus terms in Algorithm 1.

Proposition 3. *For all $i, j \in \mathcal{V}$, the deviation in the primal variables of nodes is bounded by*

$$\|\mathbf{x}_i^t - \mathbf{x}_j^t\| \leq 10\alpha\tilde{\beta}L\frac{\log(T\sqrt{n})}{1-\sigma_2(W)},$$

for all $n \geq 2$.

We note that the separation in the primal variables of a pair of nodes is governed by the inverse of the spectral gap $1 - \sigma_2(W)$ which itself is dictated by the choice of the weight matrix W as well as the structure of underlying graph. For example, an admissible choice of the weight matrix W is given by the *lazy Metropolis* matrix [Ols14] that is characterized as $W = \frac{1}{2}I + \frac{1}{2}M$, where given the degrees $d(i)$ and $d(j)$ of nodes i and j, respectively, the matrix $[M]$ has the following elements

$$[M]_{ij} = \begin{cases} \dfrac{1}{\max(d(i), d(j))} & \text{if } (i,j) \in \mathcal{E} \\ 0, & \text{if } (i,j) \notin \mathcal{E}. \end{cases}$$

It is easy to verify that the lazy Metropolis matrix W is stochastic, symmetric, and diagonally dominant. Further, due to symmetry, the singular values are simply the absolute value of the eigenvalues. More importantly, the inverse of the spectral gap has an upper bounded proportional to n^2 [Ols14]. Specifically,

$$\frac{1}{1-\sigma_2(W)} \leq 71n^2. \tag{13}$$

By putting together Propositions 1 and 3, we arrive at the following result:

Theorem 4. *For all $j \in \mathcal{V}$, the following holds*

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{\mathbf{x}}_j^T) - f_i(\mathbf{x}^*) \leq \frac{R^2}{2T\alpha} + \alpha m L^2 R^2 + 13\alpha L^2\tilde{\beta}^2\frac{\log(T\sqrt{n})}{1-\sigma_2(W)}. \tag{14}$$

Specifically, suppose $\alpha = \frac{1}{4L\sqrt{mT}}$ and $\eta = \frac{2LR\sqrt{nm}}{\varrho T^{r/4}}$, where $r \in [0,1)$ and ϱ is such that $\alpha\eta < \frac{1}{2Q_\psi^2}$ (cf. Remark 3). Then,

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{\mathbf{x}}_j^T) - f_i(\mathbf{x}^*) \leq \frac{3R^2L\sqrt{m}}{\sqrt{T}} + \frac{13\varrho^2 L\sqrt{m}}{1-\sigma_2(W)} \cdot \frac{\log(T\sqrt{n})}{T^{\frac{1-r}{2}}}, \tag{15}$$

for all $j \in \mathcal{V}$ and $n \geq 2$.

In the next theorem, we characterize two asymptotic bounds for the constraint violation of Algorithm 1.

Theorem 5. *Consider the step size α and the regularizer's curvature η as defined as in Thm. 4 and $r \in [0, 1)$. The norm of the constraint violation has the following asymptotic for all $i \in \mathcal{V}$,*

$$\left\| \Pi_{\mathbb{R}^m_+} \left(g(\tilde{\mathbf{x}}_i^T) \right) \right\|^2 = \mathcal{O}\left(\frac{LRQ_\psi n\sqrt{nm}}{\varrho T^{\frac{r}{4}}} \right). \tag{16}$$

Furthermore, if the optimal solution $\mathbf{x}^*$ is strictly feasible $g(\mathbf{x}^*) \prec \mathbf{0}$, we have

$$\left\| \Pi_{\mathbb{R}^m_+} \left(g(\tilde{\mathbf{x}}_i^T) \right) \right\|^2 = \mathcal{O}\left(\frac{LRQ_\psi n\sqrt{nm}}{\varrho T^{\frac{1}{2}-\frac{r}{4}}} \right). \tag{17}$$

Remark 2. The case of $r = 1$ in Theorem 4 and Theorem 5 is excluded since it creates an error term in the upper bound in Eq. (15) that grows unboundedly as $T \to \infty$. The case of $r = 0$ is, however, more subtle as it can cause a non-vanishing term in the constraint violation bound under the condition that $g_k(\mathbf{x}^*) = 0$ for at least one coordinate $k \in \{1, 2, \cdots, m\}$, see Eq. (16). In particular, when the optimal solution is strictly feasible $g(\mathbf{x}^*) \prec \mathbf{0}$, the value of $r = 0$ provides the optimal rate in both Eqs. (15) and (17).

Remark 3. The constant parameter ϱ incorporated in η in Theorem 4 and Theorem 5 satisfies

$$\varrho \geq \frac{RQ_\psi^2\sqrt{n}}{T^{\frac{1}{2}+\frac{r}{4}}}. \tag{18}$$

It is easy to verify that with such a choice of ϱ, the condition $\alpha\eta < \frac{1}{2Q_\psi^2}$ is satisfied. However, in most cases of interest, T is a large number in which case we can comfortably put $\varrho = 1$.

From Theorem 4 and Thereom 5, we observe that when one of the constraints is binding, i.e., $g_k(\mathbf{x}^*) = 0$ for at least one $k \in [m]$, there is a tension between the convergence rate of the dustributed primal-dual algorithm and the decay rate of the constraint violation bound. More specifically, adopting a small value for $r \in (0, 1]$ improves the convergence rate in Eq. (15) while detoriates the constraint violation bound in Eq. (16).

This tension can be explained by inspecting the role that the regularizer ψ plays in Algorithm 1. We observe that by selecting a regularizer with a large curvature η, the norm of dual variables $\|\lambda_i^t\|$ can be reduced arbitrarly. We already noted this point in the discussion after Proposition 1. In turn, a small norm $\|\lambda_i^t\|$ results in small subgradients of $\mathfrak{L}_i(\cdot, \cdot)$ which render a fast consensus between agents in the network and thus a fast convergence rate in Theorem 4. Nevertheless, a small norm $\|\lambda_i^t\|$ also reduces the penalty of constraint violation and hence worsens the first asymptotic bound in Theorem 5.

3.2 Distributed Stochastic Primal-Dual Method

Here, we devise a *stochastic* regularized primal-dual (SPD) algorithm for solving the min-max problem in Eq. (7). The motivation for studying a randomized PD algorithm is due to the observation that during each iteration $t \in [T]$ of Algorithm 1, the full subgradient vector $\{\nabla g_k(\mathbf{x}_i^t)\}_{k=1}^m$ must be computed in Eq. (6a) at each node $i \in \mathcal{V}$. But for high dimensional data sets (large d), the computation of subgradient vector is expensive. In particular,

Algorithm 2 Distributed Stochastic Primal-Dual Method

1: **Initialize**: $\mathbf{x}_i{}^0 = \mathbf{0}$, $\lambda_i^0 = \mathbf{0}$ for all $i \in \mathcal{V}$ and a constant step size $\alpha \in \mathbb{R}_+$. Select $p_i^0 = \texttt{Uniform}\{1, 2, \cdots, m\}$.

2: **for** $t = 0, 1, 2, \cdots, T$ at the i-th node $i \in \mathcal{V}$ **do**

3: Draw a random index $K_i^t \in \{1, 2, \cdots, m\}$ according to the distribution $K_i^t \sim p_i^t$.

4: Update the primal and dual variables

$$\widehat{\mathbf{x}}_i^t = \mathbf{x}_i^t - \alpha \nabla_{\mathbf{x}} \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t; K_i^t)$$
$$\widehat{\lambda}_i^t = \lambda_i^t + \alpha \nabla_{\lambda} \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t).$$

5: Run the consensus step

$$\mathbf{x}_i^{t+1} = \Pi_{\mathcal{X}} \left(\sum_{j=1}^{n} [W]_{ij} \widehat{\mathbf{x}}_i^t \right)$$
$$\lambda_i^{t+1} = \Pi_{\mathbb{R}_+^m} \left(\sum_{j=1}^{n} [W]_{ij} \widehat{\lambda}_i^t \right).$$

6: Update $p_i^t(k) = \dfrac{\lambda_{i,k}^t}{\|\lambda_i^t\|_1}$ for $k = 1, 2, \cdots, m$. Set $p_i^t = \texttt{Uniform}\{1, \cdots, m\}$ if $\lambda_i^t = \mathbf{0}$.

7: **end for**

8: **Output**: $\tilde{\mathbf{x}}_i^T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_i^t$ for all $i \in \mathcal{V}$.

the complexity of computing gradient vector for functions defined by an explicit sequence of standard operations is proportional (with a constant proportionality coefficient) to the computational complexity of the value of corresponding function [Nes12]. Even for medium size data sets, evaluation of each iteration of the standard PD algorithm is prohibitive when the number of constraints m is large.

To reduce the complexity associated with computing subgradients of constraint functions $\{g_k\}_{k=1}^m$, we randomize the PD algorithm based on a distribution that is updated according to the dual variables computed at each algorithm iteration. Consequently, at each step of SPD, one constraint $k \in \{1, 2, \cdots, m\}$ is selected randomly and its associated subgradient $\nabla g_k(\cdot)$ is computed at each node.

To make this statement more rigorous, let $K_i^t \in \{1, 2, \cdots, m\}$ denotes a random variable distributed as $p_i^t(k) \equiv \mathbb{P}[K_i^t = k] = \lambda_{i,k}^t / \|\lambda_i^t\|_1$. To have a well-defined formulation for p_i^t, we assume that when the dual parameters are all zero $\lambda_i^t = \mathbf{0} \in \mathbb{R}_+^m$, the distribution p_i^t is uniform, i.e., $p_i^t(k) = 1/m$ for all $k \in \{1, 2, \cdots, m\}$.

With a slight abuse of notation, we define

$$\nabla_{\mathbf{x}} \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t; K^t) \equiv \nabla f_i(\mathbf{x}_i^t) + \|\lambda_i^t\|_1 \cdot \nabla g_{K^t}(\mathbf{x}_i^t) \tag{19a}$$
$$\nabla_{\lambda} \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t) \equiv g(\mathbf{x}_i^t) - \nabla \psi(\lambda_i^t). \tag{19b}$$

The randomization step in Eq. (19a) resembles to the *incremental* gradient methods which has been notably used in the training of neural networks where they are known as 'backprop-

agation' methods; see Refs. [NB01] and [BT89]. In contrast to the incremental methods, however, our proposed strategy uses an adaptive distribution that is updated based on the observed dual variables at each algorithm iteration.

Equipped with these definitions, we outline the pseudocode for the distributed stochastic primal-dual method in Algorithm 2.

Our first observation about Algorithm 2 is that the boundedness of dual variables in Proposition 1 extends to the stochastic optimization setting. This is due to the fact that for any realization of random varibales $\mathbf{x}_i^t \in \mathcal{X}, t \in [T]$ the function $g(\mathbf{x}_i^t)$ is bounded by Eq. (2). Therefore, the proof of Proposition 1 can be carried over without modification to the stochastic case. Consequently, under the restriction $0 < \alpha \leq \frac{1}{2Q_\psi^2 \eta}$ on the step size of Algorithm 2 and with $\eta = \frac{2LR\sqrt{nm}}{\beta}$, any realization of random variable λ_i^t satisfies

$$\|\lambda_i^t\| \leq \beta, \tag{20}$$

for all $t \in [T]$.

Moreover, analogous to Corollarly 2, the inequality in Eq. (20) provides upper bounds on the norm of the subgradients in Eqs. (19a)-(19b). Specifically,

$$\begin{aligned} \|\nabla_{\mathbf{x}} \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t; K_i^t)\| &\leq 4LRQ_\psi\sqrt{nm} \\ \|\nabla_\lambda \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t)\| &\leq L\tilde{\beta}, \end{aligned}$$

almost surely for all $t \in [T]$, where we recall $\tilde{\beta} \equiv 1 + \beta\sqrt{m}$.

Due to the boundedness of subgradients, we can invoke the method of bounded martingale difference to derive a high probability bound. Therefore, in the stochastic optimization case, the regularization ψ is required for two reasons:

i) It provides us with an almost sure bound on the consensus terms $\|\mathbf{x}_i - \mathbf{x}_j\|$ similar to Proposition 8.

ii) It allows us to bound the tail of the difference between the deterministic and stochastic Lagrangian functions $\nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t) - \nabla_{\mathbf{x}} \widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t; K_i^t)$, using Azuma's martingale inequality [CL06].

The first result determines the convergence rate of Algorithm 2. To prove the following theorem, let $\mathcal{F}_t$ denotes the σ-field such that the processes $(\mathbf{x}^\tau)_{\tau=1}^{t+1}$ and $(\lambda^\tau)_{\tau=1}^{t+1}$ defined in Algorithm 2 are $\mathcal{F}_t$-measurable.

Theorem 6. *The following inequality holds for Algorithm 2.*

(a) With probability of at least $1 - \varepsilon$,

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{\mathbf{x}}_j^T) - f_i(\mathbf{x}^*) \leq \frac{R^2}{2T\alpha} + \alpha m L^2 R^2 + 13\alpha L^2 \tilde{\beta}^2 \frac{\log(T\sqrt{n})}{1 - \sigma_2(W)} + 3LR\beta\sqrt{\frac{m \log \frac{1}{\varepsilon}}{T}},$$

for all $j \in \mathcal{V}$.

In particualr, let $\alpha = \frac{1}{4L\sqrt{mT}}$ *and* $\eta = \frac{2LR\sqrt{nm}}{\varrho T^{r/4}}$*, where* $r \in [0,1)$ *and* ϱ *is specified in Eq. (18). With probability of at least* $1 - \frac{1}{T}$*,*

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\tilde{\mathbf{x}}_j^T) - f_i(\mathbf{x}^*) \leq \frac{3R^2L\sqrt{m}}{\sqrt{T}} + \frac{13\varrho^2 L\sqrt{m}}{1-\sigma_2(W)} \cdot \frac{\log(T\sqrt{n})}{T^{\frac{1-r}{2}}} + 3R\varrho L\sqrt{m}\frac{\log^{\frac{1}{2}} T}{T^{\frac{1}{2}-\frac{r}{4}}}. \quad (21)$$

(b) The expected value of the convergence rate is given by

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f_i(\tilde{\mathbf{x}}_j^T)] - f_i(\mathbf{x}^*) \leq \frac{R^2}{2T\alpha} + \alpha m L^2 R^2 + 13\alpha L^2\tilde{\beta}^2 \frac{\log(T\sqrt{n})}{1-\sigma_2(W)}.$$

In particualr, let $\alpha = \frac{1}{4L\sqrt{mT}}$*,* $\beta = \varrho T^{r/4}$*, where* $r \in [0,1)$*,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f_i(\tilde{\mathbf{x}}_j^T)] - f_i(\mathbf{x}^*) \leq \frac{3R^2L\sqrt{m}\varrho}{\sqrt{T}} + \frac{13\varrho^2 L\sqrt{m}}{1-\sigma_2(W)} \cdot \frac{\log(T\sqrt{n})}{T^{\frac{1-r}{2}}}.$$

It is immediate from Eq. (21) that $f(\tilde{\mathbf{x}}_j^T) \to f(\mathbf{x}^*)$ almost surely as $T \to \infty$. Moreover, by comparing Eq. (21) with the deterministic bound in Eq. (15) we observe that both Algorithm 1 and Algorithm 2 supply the same asypmtotic convergence rate $\mathcal{O}(\log(T)/T^{\frac{1-r}{2}})$. This is due to the fact that in both algorithms, the consensus step (Steps 4 of Alg. 1 and Step 5 of Alg. 2) is the bottleneck in the convergence speed.

In the next theorem, we address the constraint violation performance of Algorithm 2. The proof is omited since it borrows similar ideas from the proofs of Theorems 5 and 6.

Theorem 7. *Consider* α *and* β *as specified in Theorem 6. Then, with probability of at least* $1 - \frac{1}{T}$

$$\left\| \Pi_{\mathbb{R}^m_+}\left(g(\tilde{\mathbf{x}}_i^T)\right) \right\|^2 = \mathcal{O}\left(\frac{LRQ_\psi n\sqrt{nm}}{\varrho T^{\frac{r}{4}}}\right). \quad (22)$$

Furthermore, if the optimal solution $\mathbf{x}^*$ *satisfies the inequality constraints strictly* $g(\mathbf{x}^*) \prec \mathbf{0}$*, we have*

$$\left\| \Pi_{\mathbb{R}^m_+}\left(g(\tilde{\mathbf{x}}_i^T)\right) \right\|^2 = \mathcal{O}\left(\frac{LRQ_\psi n\sqrt{nm}}{\varrho T^{\frac{1}{2}-\frac{r}{4}}}\right), \quad (23)$$

with probability of at least $1 - \frac{1}{T}$*.*

4 Comparison with Dual Averaging

To put our work into the context of other distributed optimization methods, here we contrast our results from Subsections 3.1 and 3.2 with those of dual averaging algorithm; cf. [DAW12].

The distributed dual averaging algorithm aims at solving the following constrained optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{24}$$

Notice that the inequality constraints $g(\mathbf{x}) \preceq \mathbf{0}$ in Eq. (1b) are absent in this formulation.

The distributed dual averaging algorithm with the variable step size α consists of two steps:

1. (*Averaging Step*): $\widehat{\mathbf{x}}_i^{t+1} = \sum_{j=1}^{n}[W]_{ij}\mathbf{x}_j^t + \nabla f_i(\mathbf{x}_i^t)$.
2. (*Projection Step*): $\mathbf{x}_i^{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \widehat{\mathbf{x}}_i^{t+1}, \mathbf{x}\rangle + \frac{1}{\alpha(t)}\Psi(\mathbf{x}) \right\}$,

where the initial value is $\mathbf{x}_i^0 \in \mathcal{X}$, and $\Psi : \mathbb{R}^d \to \mathbb{R}_+$ is a *proximal function* that stablizes each step. Based on this approach, the following convergence rate result is established [DAW12],

$$f(\tilde{\mathbf{x}}_j^T) - f(\mathbf{x}^*) \leq 8\frac{LR}{\sqrt{T}} \cdot \frac{\log(T\sqrt{n})}{\sqrt{1-\sigma_2(W)}}. \tag{25}$$

Here, the value of R is such that $\Psi(\mathbf{x}^*) \leq R^2$. However, for a quadratic proximal function $\Psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, the value of R in Eq. (25) corresponds to our definition in Theorem 5.

To compare the upper bounds in Eqs. (15) and (25) we consider two different regimes:

- *Binding constraints*: In this case $g_k(\mathbf{x}^*) = 0$ for at least one $k \in [m]$. We observe that in this regime including the inequality constraints $g(\mathbf{x}) \preceq \mathbf{0}$ in the optimization problem (1a)-(1b) modifies the convergence rate from $\mathcal{O}(\log(T)/T^{\frac{1}{2}})$ to $\mathcal{O}(\log(T)/T^{\frac{1-r}{2}})$ where $r \in [0, 1)$ dictates the rate of decay in the constraint violation bound in Eq. (16).
- *Non-binding constraints*: In this regime $g(\mathbf{x}^*) \prec \mathbf{0}$ and we can plug $r = 0$ in Eq. (15) to obtain the same asymptotic convergence rate $\mathcal{O}(\log(T)/T^{\frac{1}{2}})$ as the distributed dual averaging algorithm. Moreover, from Eq. (17) with $r = 0$ we derive that the constraint violation decays at the rate of $\mathcal{O}(1/T^{\frac{1}{2}})$.

Although the distributed dual averaging provides a better asymptotic rate than the primal-dual method in the first regime, we must note that performing the projection in the latter can be significantly more efficient. To demonstrate this with an example, consider a convex optimization problem involving a set of box constraints as well as a norm constraint,

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \\ \text{subject to} \quad & l_k \leq x_k \leq u_k, \quad k = 1, 2, \cdots, d, \\ & \|\mathbf{x}\| \leq r. \end{aligned}$$

By applying the distributed primal-dual method to this problem with $g_k(\mathbf{x}) = l_k - x_k, k = 1, 2, \cdots, d$ and $g_{k+d}(\mathbf{x}) = x_k - u_k, k = 1, 2, \cdots, d$, we observe that the projection $\Pi_{\mathcal{X}}(\cdot)$ in

Step 4 of Algorithm 1 is onto the ℓ_2-ball of radius r, i.e., $\mathcal{X} = \mathbb{B}(r) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq r\}$. It turns out that in this special case, the projection in Step 4 of Algorithm 1 has a closed form solution corresponding to a rescaled vector (cf. Appendix C.1),

$$\Pi_{\mathbb{B}(r)}(\mathbf{x}) = \frac{r \cdot \mathbf{x}}{\max\{r, \|\mathbf{x}\|\}}. \tag{26}$$

On the contrary, in the distributed dual averaging algorithm, the projection step involves a non-trivial region $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d | \mathbf{x} \in \prod_{i=1}^n [l_i, u_i]\} \cap \{\mathbf{x} \in \mathbb{R}^d | \|\mathbf{x}\| \leq r\}$. Thus, to carry out the projection at each algorithm iteration, a minimization problem must be solved.

By investigating the corresponding step size used in the derivations of Eqs. (15) and (25), a more subtle difference can be distinguished. The step size used in conjunction with the upper bound in Eq. (25) is time variable and network dependent whereas the step size in Theorem 5 is constant and independent of the network topology. In particuar, the step size utilized for computing Eq. (25) is given by $\alpha(t) = \frac{R\sqrt{1-\sigma_2(W)}}{4L\sqrt{t}}$ which depends on the network topology due to incorporating the spectral gap $1 - \sigma_2(W)$. Therefore, in the distributed dual averaging algorithm, each node must have knowledge of the network structure while this information is excessive in Algorithm 1.

5 Numerical Experiments

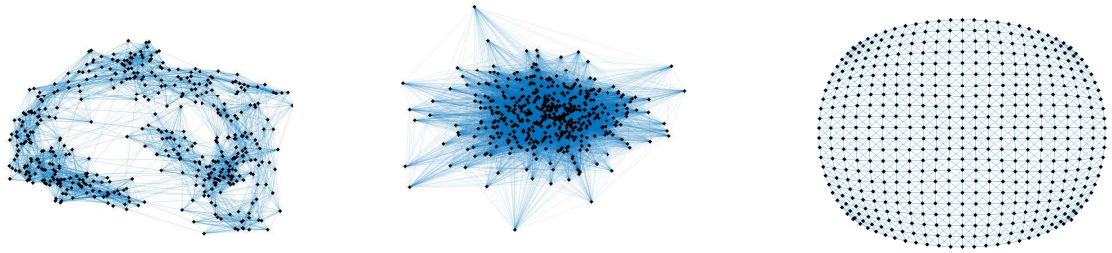
In this section, we report the numerical simulations studying the convergence of the regularized primal-dual method for the distributed regression on synthetic data. To demonstrate the performance of Algorithm 1, we use a logistic loss function with a norm constraint as well as a set of box constraints

$$\min_{\mathbf{x}} f(\mathbf{x}) \equiv \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(\mathbf{b}_i \langle \mathbf{a}_i, \mathbf{x} \rangle)) \tag{27a}$$

$$\begin{aligned} \text{subject to} \quad & g_k(\mathbf{x}) = -l - x_k \leq 0, \\ & g_{k+d}(\mathbf{x}) = x_k - u \leq 0, \quad k = 1, \cdots, d \\ & \mathbf{x} \in \mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\right\}, \end{aligned} \tag{27b}$$

where $(\mathbf{a}_i, \mathbf{b}_i) \in \mathbb{R}^d \times \{-1, +1\}$. Furthermore, we consider vectors of the dimension $d = 5$ (thus $m = 10$) and study three different network sizes, $n \in \{50, 100, 150\}$ and two different upper/lower limits $l = u \in \{0.1, 0.001\}$. To perform the projection onto $\mathcal{X}$ in our simulations, we employ the closed form expression in Eq. (26). In our simulations, we use the ridge penalty function $\psi(\lambda) = \frac{\eta}{2}\|\lambda\|^2$ in Eq. (5).

The optimization problems of the type specified in Eqs. (27a)-(27b) are common in the context of logistic classifiers in supervised learning, where $\{(\mathbf{a}_1, \mathbf{b}_1), \cdots, (\mathbf{a}_n, \mathbf{b}_n)\}$ is the set of n training data such that $\mathbf{a}_i$ is the feature vector (a.k.a. explanatory variables in regression) and $\mathbf{b}_i$ is its associated label. To make a prediction given a new vector $\mathbf{a}$, the classifier outputs $\mathbf{b} = \pm 1$ with probability of $\mathbb{P}(\mathbf{b} = \pm 1 | \mathbf{a}, \mathbf{x}) = \frac{1}{1 + \exp(\pm \langle \mathbf{x}, \mathbf{a} \rangle)}$. In our simulations, we generate $\mathbf{a}_i$ from a uniform distribution on the unit sphere. We then choose a random vector from Gaussian distribution $\mathbf{w} \sim \mathsf{N}(0, I_{d \times d})$ and generate the labels $\mathbf{b}_i \sim \mathsf{Bernoulli}(p)$,



(a) (b) (c)

Figure 1: Illustration of three graph models used in simulations. Panel (a): Watts-Strogatz graph with $K = 20$ and $\vartheta = 0.02$, Panel (b): Erdös-Réyni random graph with $p = 0.06$, Panel (c): unwrapped 8-connected neighbors lattice.

where $p = \dfrac{1}{1+\exp(\langle \mathbf{w}, \mathbf{a}_i \rangle)}$. It is straightforward to verify that $L = \max_{i=1,2,\cdots,n} \|\mathbf{a}_i\| = 1$ and $R = 1$. Note that the solution of the optimization problem in Eq. (27a) approximates $\mathbf{w}$ under the restrictions specified in Eqs. (27b).

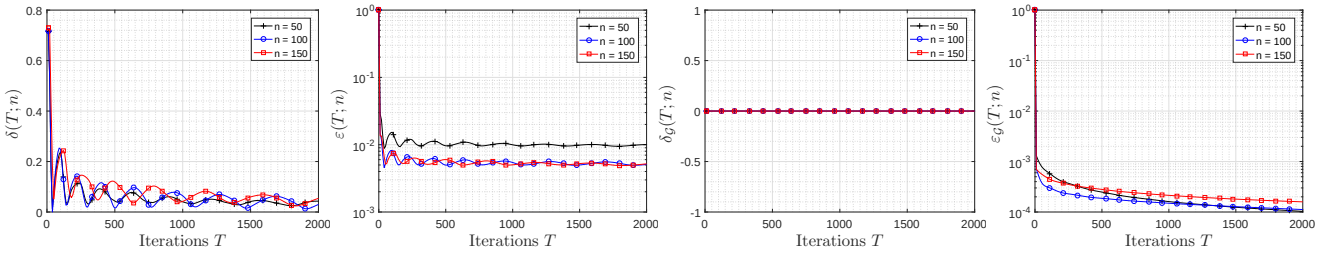
To investigate the performance of Algorithm 1 on different networks, we consider three different classes of graphs in our simulations (a): Watts-Strogatz small-world graph model [WS98], (b) Erdös-Réyni random graph [Bol98], (c) unwrapped 8-connected neighbors lattice. See Fig. 1.

A graph is characterized as a small-world if it is highly clustered locally (like regular lattices) and with a small separation globally. Social networks is an example where each person is only five or six people away from anyone else. Watts-Strogatz model is a framework to generate random graphs with small-world properties based on two structural features, namely the clustering and average path length. These features are captured by two parameters; the mean degree K and a parameter ϑ that interpolates between a lattice ($\vartheta = 0$) and a random graph ($\vartheta = 1$).

In the Erdös-Réyni random graph, the edge between each pair of nodes is included to the graph with the probability p independent from every other edge. Note that the Watts-Strogatz small-world graph model reduces to the Erdös-Réyni random graph model when $\vartheta = 1$, where $p = \frac{K}{N-1}$.

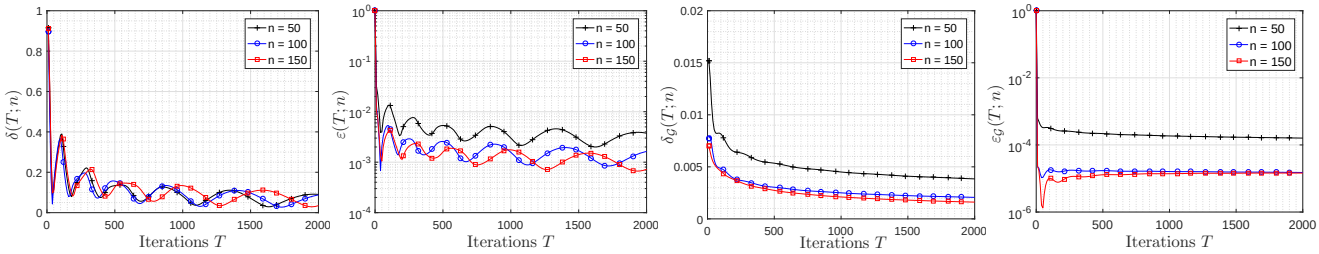
Give a connectivity graph $\mathcal{G}$ with n nodes, let $\varepsilon_{\mathcal{G}}(T, n)$ denotes the maximum relative error of the network, i.e., $\varepsilon_{\mathcal{G}}(T; n) \equiv \max_{j=1,2,\cdots,n} \left| \dfrac{f(\widetilde{\mathbf{x}}_j^T) - f(\mathbf{x}^*)}{f(\mathbf{x}^*)} \right|$ for every node in the graph $i \in \mathcal{V}$. Further, we define $\delta_{\mathcal{G}}(T; n) \equiv \max_{i=1,2,\cdots,n} \|g(\widetilde{\mathbf{x}}_i^T)\|$ as the maximum constraint violation among all the nodes in the network. In the case of the centeralized PD method, we similarly use $\varepsilon(T, n)$ and $\delta(T; n)$ to denote the relative error gap and the constraint violation, respectively. In our simulations, we use `CVX solver` [GB] to compute $f(\mathbf{x}^*)$.

Figure 2 shows the constraint violation as well as convergence rate in the centeralized



(a) (b) (c) (d)

Figure 2: Distributed regression on synthetic data using Watts-Strogratz graph with $K = 20$, $\vartheta = 0.02$, $\eta \propto T^{-1/5}$, $\alpha \propto T^{-1/2}$ and $l = u = 0.1$, Panel (a): Constraint violation $\delta(T;n)$ of the centeralized PD algorithm, Panel (b): Convergence rate $\varepsilon(T;n)$ of the centeralized PD algorithm, Panel (c) Constraint violation $\delta_{\mathcal{G}}(T;n)$ of the decenteralized PD algorithm, Panel (d): Convergence rate $\varepsilon_{\mathcal{G}}(T;n)$ of the decenteralized PD algorithm.



(a) (b) (c) (d)

Figure 3: Distributed regression on synthetic data using Watts-Strogratz graph with $K = 20$, $\vartheta = 0.02$, $\eta \propto T^{-1/5}$, $\alpha \propto T^{-1/2}$ and $l = u = 0.001$, Panel (a): Constraint violation $\delta(T;n)$ of the centeralized PD algorithm, Panel (b): Convergence rate $\varepsilon(T;n)$ of the centeralized PD algorithm, Panel (c) Constraint violation $\delta_{\mathcal{G}}(T;n)$ of the decenteralized PD algorithm, Panel (d): Convergence rate $\varepsilon_{\mathcal{G}}(T;n)$ of the decenteralized PD algorithm.

and decenteralizeed PD algorithms for various iterations T and the value of $l = u = 0.1$ for upper/lower limits [1]. In this particular example, we observe that in the decenteralized PD algorithm, solution $\tilde{\mathbf{x}}_i^T$ is feasible for all $T \in \mathbb{N}$ and $i \in \mathcal{V}$, whereas in the centeralized PD algorithm, the outputs are infeasible. This difference is due to the averaging step (Step 4 of Alg. 1) in the decenteralized PD method, whereby fluctuations in the solution $\tilde{\mathbf{x}}_i^T$ becomes smaller than the standard PD method and thus $\tilde{\mathbf{x}}_i^T$ remains within the boundary of the box constraints.

By tightening the upper and lower bounds in the box constraints (27b), we observe a constraint violation in both the decenteralized and centeralized methods; see Figure 4. However, even in this case the amplitude of $\delta_{\mathcal{G}}(T;n)$ in the decenteralized method is smaller than $\delta(T;n)$ in the centeralized case. Note that in the decenteralized method, clusters of larger size n require more iterations T to achieve a prescribed precision $\varepsilon_{\mathcal{G}}(T;n)$ which conforms to the prediction of the error bound (15).

Lastly, Figure 4 shows the maximum error gap $\max_{i\in\mathcal{V}} f(\widetilde{\mathbf{x}}_i^T) - f(\mathbf{x}^*)$ over three different

[1] The code for these experiments, as well as for the subsequent experiments, is available online at `http://nali.seas.harvard.edu/`.

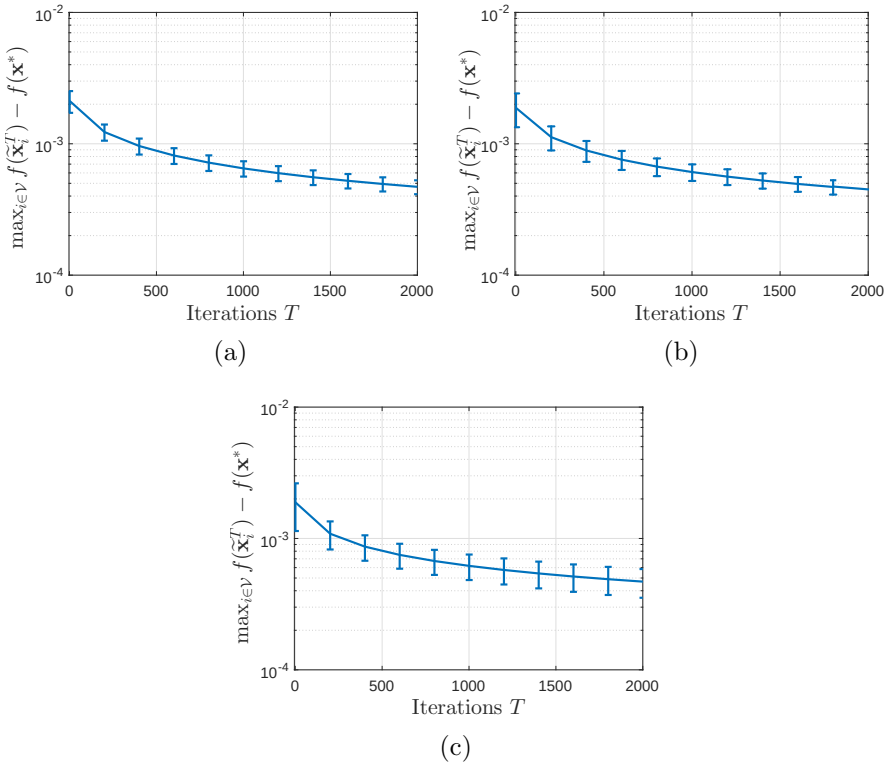


Figure 4: Maximum error gap $\max_{i\in\mathcal{V}} f(\widetilde{x}_i^T) - f(\mathbf{x}^*)$ for three different classes of graphs with $n = 100$ nodes over 20 trials, Panel (a): Watts-Strogratz graph with $K = 20$ and $\vartheta = 0.02$, Panel (b): Erdös-Réyni random graph with $p = 0.06$, Panel (c): unwrapped 8-connected neighbors lattice

classes of graphs demonstrated in Figure 1 and for 20 trials. Each blue curve in Figure 4 shows the average values over the trials and each bar shows the standard deviation. The prediction given by the upper bound in Theorem 1 is also depicted in each panel.

6 Conclusion and Discussion

In this paper, we studied a distributed primal-dual method for solving convex optimization problems with inequality constraints over a network. In the proposed distributed framework, dual variables are regularized with a smooth and strongly convex function. As a result, the norm of dual variables, and hence the subgradients of the Lagrangian function, are bounded. Based on this regularization, we obtained an upper bound on the consensus terms and subsequently an upper bound on the convergence rate of the underlying algorithm. Furthermore, we presented asymptotic results for the diminishing rate of the constraint violation. Our results demonstrates a transition in the behavior of the distributed regularized PD algorithm in the sense that when one of the inequality constraints is binding, there is a tension between the convergence rate speed and the diminishing rate associated with the constraint violation. Nevertheless, this tension vanishes when the constraints are satisfied strictly. We also studied the convergence rate of the distributed stochastic primal-dual method. We showed that in the distributed case, the stochastic algorithm enjoys the same

asymptotic convergence rate as its deterministic counterpart.

There are several interesting questions that can be addressed to supplement the result of this paper. For example, it would be interesting to have a comprehensive analysis of penalty/barrier function methods in the distributed setting and compare the result with the algorithm we developed in this paper. Another interesting question to explore is to determine the effect of local (private) constraints on the convergence rate. Local constraints can reflect the internal dynamics of each node which can be utilized in various applications such as sensor networks.

Acknowledgment

The research of MB was supported by IBM PhD Fellowship and NL was supported by NSF CAREER grant 1553407 and Harvard Center for Green Buildings and Cities.

References

[BB03] Anastasios G Bakirtzis and Pandelis N Biskas. A decentralized solution to the DC-OPF of interconnected power systems. *Power Systems, IEEE Transactions on*, **18**(3):1007–1013, 2003.

[Bol98] Béla Bollobás. *Random graphs*. Springer, 1998.

[BT89] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume **23**. Prentice hall Englewood Cliffs, NJ, 1989.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CL06] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, **3**(1):79–127, 2006.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, **20**(3):273–297, 1995.

[DAW12] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *Automatic control, IEEE Transactions on*, **57**(3):592–606, 2012.

[FM04] J Alexander Fax and Richard M Murray. Information flow and cooperative control of vehicle formations. *Automatic Control, IEEE Transactions on*, **49**(9):1465–1476, 2004.

[GB] Michael Grant and Stephen Boyd. CVX: MATLAB software for disciplined convex programming.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HUL96] JB Hiriart-Urruty and C Lemaréchal. Convex analysis and minimization algorithms, part i: Fundamentals, vol. 305 of grundlehren der mathematischen wissenschaften, 1996.

[JLM03] Ali Jadbabaie, Jie Lin, and A Stephen Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *Automatic Control, IEEE Transactions on*, **48**(6):988–1001, 2003.

[JXM14] Dusan Jakovetic, Joao Xavier, and Jose MF Moura. Fast distributed gradient methods. *Automatic Control, IEEE Transactions on*, **59**(5):1131–1146, 2014.

[LM14] Na Li and Jason R Marden. Decoupling coupled constraints through utility design. *Automatic Control, IEEE Transactions on*, **59**(8):2289–2294, 2014.

[MJY12] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, **13**(1):2503–2528, 2012.

[NB01] Angelia Nedic and Dimitri P Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[NO09a] Angelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *Automatic Control, IEEE Transactions on*, **54**(1):48–61, 2009.

[NO09b] Angelia Nedić and Asuman Ozdaglar. subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, **142**(1):205–228, 2009.

[ÖFL04] Petter Ögren, Edward Fiorelli, and Naomi Ehrich Leonard. Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment. *Automatic Control, IEEE Transactions on*, **49**(8):1292–1302, 2004.

[Ols14] Alex Olshevsky. Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control. *arXiv:1411.4186*, 2014.

[OSFM07] Reza Olfati-Saber, Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, **95**(1):215–233, 2007.

[SITT04] DušAn M Stipanović, GöKhan Inalhan, Rodney Teo, and Claire J Tomlin. Decentralized overlapping control of a formation of unmanned aerial vehicles. *Automatica*, **40**(8):1285–1296, 2004.

[Ste77] Thomas E Stern. A class of decentralized routing algorithms using relaxation. *Communications, IEEE Transactions on*, **25**(10):1092–1102, 1977.

[Tsi84] John Nikolas Tsitsiklis. Problems in decentralized decision making and computation. Technical report, DTIC Document, 1984.

[WL09] Pu Wan and Michael D Lemmon. Event-triggered distributed optimization in sensor networks. In *Information Processing in Sensor Networks, 2009. IPSN 2009. International Conference on*, pages 49–60. IEEE, 2009.

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[XBK07] Lin Xiao, Stephen Boyd, and Seung-Jean Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, **67**(1):33–46, 2007.

[YXZ11] Deming Yuan, Shengyuan Xu, and Huanyu Zhao. Distributed primal–dual subgradient method for multiagent optimization via consensus algorithms. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **41**(6):1715–1724, 2011.

A Proofs of Main Results

A.1 Proof of Proposition 1

From the update rule of λ_i^t in Algorithm 1 we have

$$\begin{aligned}\|\lambda_i^{t+1}\| &= \left\| \Pi_{\mathbb{R}_+^m} \left(\sum_{j=1}^n [W]_{ij} (\lambda_j^t + \alpha \nabla_\lambda \mathfrak{L}_j(\mathbf{x}_j^t, \lambda_j^t)) \right) \right\| \\ &\leq \left\| \sum_{j=1}^n [W]_{ij} (\lambda_j^t + \alpha \nabla_\lambda \mathfrak{L}_j(\mathbf{x}_j^t, \lambda_j^t)) \right\| \\ &\leq \sum_{j=1}^n [W]_{ij} \left\| \lambda_j^t + \alpha \nabla_\lambda \mathfrak{L}_j(\mathbf{x}_j^t, \lambda_j^t) \right\|,\end{aligned}$$

where the last step follows by the triangular inequality of the norm. We now square both sides of the inequality to obtain

$$\begin{aligned}
\|\lambda_i^{t+1}\|^2 &\leq \left(\sum_{j=1}^{n}[W]_{ij}\left\|\lambda_i^t + \alpha\nabla_\lambda \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\right\|\right)^2 \\
&\overset{(a)}{\leq} \sum_{i=1}^{n}[W]_{ij}\left\|\lambda_j^t + \alpha\nabla_\lambda \mathfrak{L}_j(\mathbf{x}_j^t, \lambda_j^t)\right\|^2 \\
&= \sum_{i=1}^{n}[W]_{ij}\left\|\lambda_j^t + \alpha g(\mathbf{x}_j^t) - \alpha\nabla\psi(\lambda_j^t)\right\|^2 \\
&\overset{(b)}{\leq} \sum_{i=1}^{n}[W]_{ij}\left((1+\delta^{-1})\|\alpha g(\mathbf{x}_j^t)\|^2 + (1+\delta)\|\lambda_j^t - \alpha\nabla\psi(\lambda_j^t)\|^2\right) \qquad (28) \\
&= (1+\delta^{-1})\alpha^2\sum_{j=1}^{n}[W]_{ij}\|g(\mathbf{x}_j^t)\|^2 \\
&\quad + (1+\delta)\sum_{j=1}^{n}[W]_{ij}\left(\|\lambda_j^t\|^2 - 2\alpha\langle\nabla\psi(\lambda_j^t), \lambda_j^t\rangle + \alpha^2\|\nabla\psi(\lambda_j^t)\|^2\right),
\end{aligned}$$

where (a) follows from Jensen's inequality, and (b) follows from Young's inequality which is valid for any $\delta > 0$ (cf. Appendix B.1 for proof). In particular, the addition of the parameter δ in (b) allows us to have a tight control over the terms in the bound.

Taking the summation with respect to $i = 1, 2, \cdots, n$ results in

$$\begin{aligned}
\sum_{i=1}^{n}\|\lambda_i^{t+1}\|^2 \leq& (1+\delta^{-1})nm\alpha^2L^2R^2 \\
&+ (1+\delta)\sum_{i=1}^{n}\left(\|\lambda_i^t\|^2 - 2\alpha\langle\nabla\psi(\lambda_i^t), \lambda_i^t\rangle + \alpha^2\|\nabla\psi(\lambda_i^t)\|^2\right), \qquad (29)
\end{aligned}$$

where we used the fact that $\sum_{i=1}^{n}[W]_{ij} = 1$ and $\|g(\mathbf{x}_i^t)\|^2 \leq mL^2R^2$. We now recall that $\psi(\lambda_i^t)$ is η-strongly convex (cf. Def. 1). Therefore,

$$\langle\nabla\psi(\lambda_i^t) - \nabla\psi(\hat{\lambda}_i^t), \lambda_i^t - \hat{\lambda}_i^t\rangle \geq \eta\|\lambda_i^t - \hat{\lambda}_i^t\|^2.$$

Setting $\hat{\lambda}_i^t = \mathbf{0}$ gives us

$$\langle\nabla\psi(\lambda_i^t) - \nabla\psi(\mathbf{0}), \lambda_i^t\rangle \geq \eta\|\lambda_i^t\|^2. \qquad (30)$$

Since $\nabla\psi(\mathbf{0}) = 0$, after a sign change and multiplication by 2α we obtain

$$-2\alpha\langle\nabla\psi(\lambda_i^t), \lambda_i^t\rangle \leq -2\alpha\eta\|\lambda_i^t\|^2. \qquad (31)$$

Moreover, from γ-smoothness assumption, we have

$$\|\nabla\psi(\lambda_i^t)\| \leq \gamma\|\lambda_i^t\|. \qquad (32)$$

Combining Eqs. (31), (32) and (29) results in

$$\sum_{i=1}^{n} \|\lambda_i^{t+1}\|^2 \leq (1+\delta^{-1}) nm\alpha^2 L^2 R^2 + (1+\delta)\left(1 - 2\alpha\eta + \alpha^2\gamma^2\right) \sum_{i=1}^{n} \|\lambda_i^t\|^2. \tag{33}$$

Now, we note that $0 < 1 - 2\alpha\eta + \alpha^2\gamma^2 < 1$ provided $\alpha < 2\eta/\gamma^2$. This is due to the fact that $1 - 2\alpha\eta + \alpha^2\gamma^2 \geq (1 - \gamma\alpha)^2 > 0$ as $\eta \leq \gamma$. Moreover, the constraint $\alpha < 2\eta/\gamma^2$ ensures that $1 - 2\alpha\eta + \alpha^2\gamma^2 < 1$.

Therefore, we can choose $\delta = \dfrac{\varepsilon}{1 - 2\alpha\eta + \alpha^2\gamma^2} - 1$, $1 - 2\alpha\eta + \alpha^2\gamma^2 < \varepsilon < 1$. This choice of δ results in $(1+\delta)\left(1 - 2\alpha\eta + \alpha^2\gamma^2\right) = \varepsilon$ and we can proceed from Eq. (33) as below

$$\begin{aligned}
\sum_{i=1}^{n} \|\lambda_i^{t+1}\|^2 &\leq \varepsilon \sum_{i=1}^{n} \|\lambda_i^t\|^2 + \frac{\varepsilon\alpha^2}{\varepsilon - (1 - 2\alpha\eta + \alpha^2\gamma^2)} nmL^2R^2 \\
&= (\varepsilon^t + \varepsilon^{t-1} + \cdots + 1) \frac{\varepsilon\alpha^2}{\varepsilon - (1 - 2\alpha\eta + \alpha^2\gamma^2)} nmL^2R^2 \\
&= \frac{1 - \varepsilon^{t+1}}{1 - \varepsilon} \frac{\varepsilon\alpha^2}{\varepsilon - (1 - 2\alpha\eta + \alpha^2\gamma^2)} nmL^2R^2 \\
&\leq \frac{1}{1 - \varepsilon} \frac{\varepsilon\alpha^2}{\varepsilon - (1 - 2\alpha\eta + \alpha^2\gamma^2)} nmL^2R^2.
\end{aligned} \tag{34}$$

It is easy to verify that the the upper bound in Eq. (34) is minimized by $\varepsilon^* = \sqrt{1 - 2\alpha\eta + \alpha^2\gamma^2}$. Substituting ε^* in Eq. (34) gives

$$\begin{aligned}
\sum_{i=1}^{n} \|\lambda_i^{t+1}\|^2 &\leq \frac{\alpha^2}{(1 - \sqrt{1 - 2\alpha\eta + \alpha^2\gamma^2})^2} nmL^2R^2 \\
&\leq \frac{nmL^2R^2}{\eta^2(1 - \frac{\alpha\gamma^2}{2\eta})^2}.
\end{aligned} \tag{35}$$

The last step is due to the Bernoulli inequality $(1+x)^r \leq 1 + rx$ for $x \geq -1$ and $r \in [0, 1]$ which results in

$$1 - \sqrt{1 - 2\alpha\eta + \alpha^2\gamma^2} \geq \alpha\eta - \frac{1}{2}\alpha^2\gamma^2.$$

By furhter restricting the step size $\alpha < \eta/(2\gamma^2) = 1/(2Q_\psi^2\eta)$, we can simplify Eq. (35) as

$$\|\lambda_i^{t+1}\|^2 \leq \frac{4nmL^2R^2}{\eta^2}.$$

where we used $\|\lambda_i^t\|^2 \leq \sum_{i=1}^{n} \|\lambda_i^{t+1}\|^2$.

A.2 Proof of Corollarly 2

Based on the upper bound in Theorem 1, we compute

$$\begin{aligned}\|\nabla_\lambda \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\| &\le \|g(\mathbf{x}_i^t) - \nabla\psi(\lambda_i^t)\| \\ &\le \|g(\mathbf{x}_i^t)\| + \|\nabla\psi(\lambda_i^t)\| \\ &\overset{(a)}{\le} \|g(\mathbf{x}_i^t)\| + \gamma\|\lambda_i^t\| \\ &\le LR\sqrt{m}\left(1 + 2\sqrt{n}Q_\psi\right) \\ &\le 3LR\sqrt{mn}Q_\psi, \end{aligned} \tag{36}$$

where in (a) we used the γ-smoothness assumption of $\psi(\cdot)$. Similarly,

$$\begin{aligned}\|\nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\| &\le \|\nabla f_i(\mathbf{x}_i^t)\| + \sum_{k=1}^m \lambda_{i,k}^t \cdot \|\nabla g_k(\mathbf{x}_i^t)\| \\ &\le L\left(1 + \|\lambda_i^t\|_1\right) \\ &\le L\left(1 + \sqrt{m}\|\lambda_i^t\|\right) \qquad (37) \\ &\overset{(b)}{=} L\left(1 + \beta\sqrt{m}\right) \\ &\overset{(c)}{=} L\tilde{\beta}, \qquad (38)\end{aligned}$$

where (b) follows from the bound in Eq. (12), and (c) from the defintion $\tilde{\beta} \equiv 1 + \beta\sqrt{m}$.

A.3 Proof of Proposition 3

We begin by obtaining a recursive formula for $\|\mathbf{x}_i^t - \mathbf{x}_j^t\|$. From the update rule in Algorithm 1 and the non-expansive property of the projection we have (cf. Appendix C.2)

$$\begin{aligned}\|\mathbf{x}_i^t - \mathbf{x}_j^t\| &= \left\|\Pi_{\mathcal{X}}\left(\sum_{k=1}^n [W]_{ik}\widehat{\mathbf{x}}_k^{t-1}\right) - \Pi_{\mathcal{X}}\left(\sum_{k=1}^n [W]_{jk}\widehat{\mathbf{x}}_k^{t-1}\right)\right\| \\ &\le \left\|\sum_{k=1}^n ([W]_{ik} - [W]_{jk})\widehat{\mathbf{x}}_k^{t-1}\right\| \\ &\overset{(a)}{\le} \sum_{k=1}^n |[W]_{ik} - [W]_{jk}| \cdot \left\|\widehat{\mathbf{x}}_k^{t-1}\right\| \\ &= \sum_{k=1}^n |[W]_{ik} - [W]_{jk}| \cdot \|\mathbf{x}_k^{t-1} + \alpha\nabla_{\mathbf{x}}\mathfrak{L}_k(\mathbf{x}_k^{t-1}, \lambda_k^{t-1})\| \\ &\overset{(b)}{\le} \sum_{k=1}^n |[W]_{ik} - [W]_{jk}| \cdot \|\mathbf{x}_k^{t-1}\| + \alpha\sum_{k=1}^n |[W]_{ik} - [W]_{jk}| \cdot \|\nabla_{\mathbf{x}}\mathfrak{L}_k(\mathbf{x}_k^{t-1}, \lambda_k^{t-1})\|, \end{aligned} \tag{39}$$

where in steps (a) and (b) we used the triangle inequality of the norm. Further, since $\mathcal{X}$ contains the origin $\mathbf{x} = \mathbf{0}$, once again from the non-expansive property of the projection we

have

$$\begin{aligned}\|\mathbf{x}_k^{t-1}\| &= \left\|\Pi_{\mathcal{X}}\left(\sum_{\ell=1}^{n}[W]_{k\ell}\widehat{\mathbf{x}}_\ell^{t-2}\right)\right\| \\ &\leq \left\|\sum_{\ell=1}^{n}[W]_{k\ell}\widehat{\mathbf{x}}_\ell^{t-2}\right\| \\ &\leq \sum_{\ell=1}^{n}[W]_{k\ell}\cdot\|\widehat{\mathbf{x}}_\ell^{t-2}\| \\ &\leq \sum_{\ell=1}^{n}[W]_{k\ell}\cdot\|\mathbf{x}_\ell^{t-2}\| + \alpha(t-2)\sum_{\ell=1}^{n}[W]_{k\ell}\cdot\|\nabla_{\mathbf{x}}\mathfrak{L}_k(\mathbf{x}_\ell^{t-2},\lambda_\ell^{t-2})\|. \end{aligned} \tag{40}$$

Plugging (40) in Eq. (39) yields

$$\begin{aligned}\|\mathbf{x}_i^t - \mathbf{x}_j^t\| \leq &\sum_{\ell=1}^{n}|[W^2]_{i\ell} - [W^2]_{j\ell}|\cdot\|\mathbf{x}_\ell^{t-2}\| + \alpha\sum_{\ell=1}^{n}|[W^2]_{i\ell} - [W^2]_{j\ell}|\cdot\|\nabla_{\mathbf{x}}\mathfrak{L}_k(\mathbf{x}_\ell^{t-2},\lambda_\ell^{t-2})\| \\ &+ \alpha\sum_{\ell=1}^{n}|[W]_{i\ell} - [W]_{j\ell}|\cdot\|\nabla_{\mathbf{x}}\mathfrak{L}_k(\mathbf{x}_\ell^{t-1},\lambda_\ell^{t-1})\|. \end{aligned} \tag{41}$$

Pursuing this recursive argument in conjunction with the state transition matrix $\Phi(t,r) \equiv W^{t-r}$ gives us a more compact form of inequality,

$$\|\mathbf{x}_i^t - \mathbf{x}_j^t\| \leq \alpha\sum_{r=0}^{t-1}\sum_{k=1}^{n}|[\Phi(t,r)]_{ik} - [\Phi(t,r)]_{jk}|\cdot\|\nabla_{\mathbf{x}}\mathfrak{L}_k(\mathbf{x}_k^r,\lambda_k^r)\|, \tag{42}$$

where we used the fact that $\mathbf{x}_i^0 = \mathbf{0}$ for all $i \in \mathcal{V}$. We combine the upper bound in Eq. (38) with the expression in Eq. (42) to compute

$$\|\mathbf{x}_i^t - \mathbf{x}_j^t\| \leq \alpha\tilde{\beta}L\sum_{r=0}^{t-1}\|[\Phi(t,r)]_i - [\Phi(t,r)]_j\|_1\,. \tag{43}$$

From the triangle inequality for ℓ_1-norm we have

$$\|[\Phi(t,r)]_i - [\Phi(t,r)]_j\|_1 \leq \|[\Phi(t,r)]_i - \mathbb{1}/n\|_1 + \|[\Phi(t,r)]_j - \mathbb{1}/n\|_1. \tag{44}$$

for all $\forall i,j \in \{1,2,\cdots,n\}, i \neq j$. We now closely follow the argument of Duchi, *et al.* [DAW12] to bound the two terms on the r.h.s. of Eq. (44). We notice that $\Phi(t,r)$ is a doubly stochastic matrix. Borrowing the analysis from [DAW12] we therefore have

$$\|[\Phi(t,r)]_i - \mathbb{1}/n\|_1 \leq \sqrt{n}\sigma_2(W)^{t-r}, \quad \forall i \in \{1,2,\cdots,n\}. \tag{45}$$

Consequently, to achieve the accuracy of $\|[\Phi(t,r)]_i - \mathbb{1}/n\|_1 \leq \frac{1}{T}$, we need

$$t - r \geq \frac{\log(T\sqrt{n})}{\log\sigma_2(W)^{-1}}, \tag{46}$$

Otherwise, the deviation can be bounded as $\|[\Phi(t,r)]_i - \mathbb{1}/n\|_1 \leq 2$. We now define the cutoff time $\tau \equiv \frac{\log(T\sqrt{n})}{\log \sigma_2(W)^{-1}}$. Then, we break the sum in Eq. (43) as below

$$\begin{aligned}\|\mathbf{x}_i^t - \mathbf{x}_j^t\| &\leq \alpha\tilde{\beta}L\Big(\sum_{r=t-\tau+1}^{t-1}\|[\Phi(t,r)]_i - [\Phi(t,r)]_j\|_1 + \sum_{r=0}^{t-\tau}\|[\Phi(t,r)]_i - [\Phi(t,r)]_j\|_1\Big)\\ &\leq 4\alpha\tilde{\beta}L\frac{\log(T\sqrt{n})}{\log\sigma_2(W)^{-1}} + 2\alpha\tilde{\beta}L\\ &\leq 10\alpha\tilde{\beta}L\frac{\log(T\sqrt{n})}{1-\sigma_2(W)},\end{aligned}$$

where the last step follows from the inequality $\log(x^{-1}) \geq 1 - x, x > 0$, and the fact that $\frac{3\log(T\sqrt{n})}{1-\sigma_2(W)} \geq 1$ for $n \geq 2$ and all $T \in \mathbb{N}$ since $3\log(\sqrt{2}) \approx 1.04$.

A.4 Proof of Theorem 4

First, we state a proposition.

Proposition 8. *Let $f(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$. For the optimal solution $\mathbf{x}^* \in \mathcal{X}$ the following inequality holds*

$$\begin{aligned}\sum_{t=1}^T\big(f(\mathbf{x}_j^t) - f(\mathbf{x}^*)\big) \leq \frac{\|\mathbf{x}^*\|^2}{2\alpha} + \frac{\alpha}{2n}\sum_{t=1}^T\sum_{i=1}^n\big(\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2 + \|\nabla_\lambda\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2\big)\\ + \frac{1}{n}\sum_{t=1}^T\sum_{i=1}^n\|\nabla f_i(\mathbf{x}_j^t)\|_*\|\mathbf{x}_j^t - \mathbf{x}_i^t\| - \frac{\eta}{2n}\sum_{t=1}^T\sum_{i=1}^n\|\lambda_i^t\|^2.\end{aligned} \tag{47}$$

Proof. See Appendix B.2. □

Now, due to the second upper bound in Corollarly 2 we obtain

$$\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2 \leq \tilde{\beta}^2L^2. \tag{48}$$

Also, from Eq. (36) we have

$$\begin{aligned}\|\nabla_\lambda\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2 &\leq (LR\sqrt{m} + \gamma\|\lambda_i^t\|)^2 &(49)\\ &\leq 2mL^2R^2 + 2\gamma^2\|\lambda_i^t\|^2. &(50)\end{aligned}$$

Moreover, by employing the upper bound on the consensus term in Proposition 3, we establish the following inequality for the last term of Eq. (47),

$$\begin{aligned}\|\nabla f_i(\mathbf{x}_j^t)\|_*\|\mathbf{x}_j^t - \mathbf{x}_i^t\| &\leq L \times 10\alpha\tilde{\beta}L\frac{\log(T\sqrt{n})}{1-\sigma_2(W)}\\ &\leq 10\alpha\tilde{\beta}^2L^2\frac{\log(T\sqrt{n})}{1-\sigma_2(W)},\end{aligned} \tag{51}$$

where the last step is true since by definition $\tilde{\beta} \geq 1$.

Subsuiting Eqs. (48)-(51) into Eq. (47) and dividing both sides by T we derive

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_j^t)-f(\mathbf{x}^*)\right)\leq\frac{\|\mathbf{x}^*\|^2}{2T\alpha}+\alpha mL^2R^2+\alpha L^2\tilde{\beta}^2+10\alpha L^2\tilde{\beta}^2\frac{\log(T\sqrt{n})}{1-\sigma_2(W)}+\sum_{t=1}^{T}\left(\alpha\gamma^2-\frac{1}{2}\eta\right)\|\lambda_i^t\|^2. \tag{52}$$

Due to the constraint on the step size $\alpha\eta<\frac{1}{2Q_\psi^2}$ in Proposition 1, we have $\alpha\gamma^2=\alpha\eta^2Q_\psi^2\leq\frac{1}{2}\eta$. Therefore, we can eliminate the last term,

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_j^t)-f(\mathbf{x}^*)\right)\leq\frac{\|\mathbf{x}^*\|^2}{2T\alpha}+\alpha mL^2R^2+\alpha L^2\tilde{\beta}^2+10\alpha L^2\tilde{\beta}^2\frac{\log(T\sqrt{n})}{1-\sigma_2(W)}.$$

Since $\|\mathbf{x}^*\|\leq R$ and $\frac{3\log(T\sqrt{n})}{1-\sigma_2(W)}\geq 1$ for $n\geq 2$ we further have

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_j^t)-f(\mathbf{x}^*)\right)\leq\frac{R^2}{2T\alpha}+\alpha mL^2R^2+13\alpha L^2\tilde{\beta}^2\frac{\log(T\sqrt{n})}{1-\sigma_2(W)}.$$

Now, recall the running local average $\tilde{\mathbf{x}}_j^T\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_j^t$. Due to convexity of the function $f(\cdot)$ we have

$$f(\tilde{\mathbf{x}}_j^T)\leq\frac{1}{T}\sum_{t=1}^{T}f(\mathbf{x}_j^t).$$

We thus establish

$$f(\tilde{\mathbf{x}}_j^T)-f(\mathbf{x}^*)\leq\frac{R^2}{2T\alpha}+\alpha mL^2R^2+13\alpha L^2\tilde{\beta}^2\frac{\log(T\sqrt{n})}{1-\sigma_2(W)}. \tag{53}$$

Let $\alpha=\dfrac{1}{4L\sqrt{mT}}$ and $\beta=\varrho T^{\frac{r}{4}}$, where $r\in[0,1)$ and ϱ specified in Eq. (18). Note that since $\varrho\geq 1$, we have $\tilde{\beta}=1+\beta\sqrt{m}\leq 2\sqrt{m}\varrho T^{\frac{r}{4}}$. From Eq. (53) we thus obtain

$$f(\tilde{\mathbf{x}}_j^T)-f(\mathbf{x}^*)\leq\frac{3R^2L\sqrt{m}}{\sqrt{T}}+\frac{13L\sqrt{m}\varrho^2}{1-\sigma_2(W)}\cdot\frac{\log(T\sqrt{n})}{T^{\frac{1-r}{2}}}.$$

A.5 Proof of Theorem 5

We state a proposition first (cf. Appendix B.2):

Proposition 9. *For all $\mathbf{x}\in\mathcal{X}$ and $\lambda\in\mathbb{R}_+^m$ the following inequality holds*

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\mathfrak{L}_i(\mathbf{x}_i^t,\lambda)-\mathfrak{L}_i(\mathbf{x},\lambda_i^t)\right)\leq\frac{1}{2\alpha}\left(\|\mathbf{x}\|^2+\|\lambda\|^2\right)+\frac{\alpha}{2n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2+\|\nabla_{\lambda}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2\right). \tag{54}$$

Let $\mathbf{x} = \mathbf{x}^*$ and use the inequalities of Corollary 2 to obtain

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\mathfrak{L}_i(\mathbf{x}_i^t,\lambda)-\mathfrak{L}_i(\mathbf{x}^*,\lambda_i^t)\right) \leq \frac{\|\mathbf{x}^*\|^2+\|\lambda\|^2}{2\alpha}+\frac{1}{2}T\alpha\left(L^2\tilde{\beta}^2+9L^2R^2Q_\psi^2mn\right). \tag{55}$$

By expanding the l.h.s. of Eq. (55) and dividing by T we derive

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(f_i(\mathbf{x}_i^t)-f_i(\mathbf{x}^*))+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\lambda,g(\mathbf{x}_i^t)\rangle-\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\lambda_i^t,g(\mathbf{x}^*)\rangle-\psi(\lambda)+\frac{1}{n}\sum_{i=1}^{n}\psi(\lambda_i^t)$$
$$\leq \frac{R^2+\|\lambda\|^2}{2T\alpha}+\frac{1}{2}\alpha\left(L^2\tilde{\beta}^2+9L^2R^2Q_\psi^2mn\right),$$

where we used the fact that $\|\mathbf{x}^*\| \leq R$. Due to the positivity of the dual variabels $\lambda_i^t \succeq \mathbf{0}$ as well as $g(\mathbf{x}^*) \preceq \mathbf{0}$ we have $\langle\lambda_i^t, g(\mathbf{x}^*)\rangle \preceq 0$. Furthermore, $\psi(\lambda_i^t) \geq 0$ for an admissible regularizer (cf. Def. 1). Hence, we can eliminate these terms which leaves us

$$\frac{1}{n}\sum_{i=1}^{n}(f_i(\tilde{\mathbf{x}}_i^t)-f_i(\mathbf{x}^*))+\frac{1}{n}\sum_{i=1}^{n}\langle\lambda,g(\tilde{\mathbf{x}}_i^t)\rangle-\psi(\lambda) \tag{56}$$
$$\leq \frac{R^2+\|\lambda\|^2}{2T\alpha}+\frac{1}{2}\alpha\left(L^2\tilde{\beta}^2+9L^2R^2Q_\psi^2mn\right),$$

where we used the definition $\tilde{\mathbf{x}}_i^T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_i^t$ as well as the convexity of the functions f_i and g_k.

Lemma 10. *For all $\lambda \in \mathbb{R}^m$*

$$\psi(\lambda) \leq \frac{\eta Q_\psi}{2}\|\lambda\|^2.$$

Proof. Since ψ is γ-smooth, we have

$$\psi(\lambda)-\psi(\hat{\lambda})-\langle\nabla\psi(\hat{\lambda}),\lambda-\hat{\lambda}\rangle \leq \frac{\gamma}{2}\|\lambda-\hat{\lambda}\|^2.$$

Let $\hat{\lambda} = \mathbf{0}$. From the conditions $\nabla\psi(\mathbf{0}) = \mathbf{0}$ and $\psi(\mathbf{0}) = 0$ and the fact that $\gamma = Q_\psi\eta$ we obtain the desired inequality. □

Recall the definition $\lambda \equiv (\lambda_1, \lambda_2, \cdots, \lambda_m)$. Using the result of Lemma 10 and the inequality (56), we get

$$\frac{1}{n}\sum_{i=1}^{n}(f_i(\tilde{\mathbf{x}}_i^T)-f_i(\mathbf{x}^*))+\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m}\left(\lambda_k g_k(\tilde{\mathbf{x}}_i^T)-\left(\frac{Q_\psi\eta}{2}+\frac{1}{2\alpha T}\right)\lambda_k^2\right)$$
$$\leq \frac{R^2}{2T\alpha}+\frac{1}{2}\alpha\left(L^2\tilde{\beta}^2+9L^2R^2Q_\psi^2mn\right), \tag{57}$$

Let $\mathsf{F} \in \mathbb{R}^+$ be a constant such that $-\mathsf{F} \leq \frac{1}{n}\sum_{i=1}^{n}(f_i(\tilde{\mathbf{x}}_i^T)-f_i(\mathbf{x}^*))$. By maximizng the l.h.s. of Eq. (57) with respect to $\lambda_k, k = 1, 2, \cdots, m$ we derive

$$\left(\frac{Q_\psi\eta}{2}+\frac{1}{2\alpha T}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m}\left[\Pi_{\mathbb{R}_+}\left(g_k(\tilde{\mathbf{x}}_i^T)\right)\right]^2 \leq \mathsf{F}+\frac{R^2}{2T\alpha}+\frac{1}{2}\alpha\left(L^2\tilde{\beta}^2+9L^2R^2Q_\psi^2mn\right).$$

Subsuiting $\alpha = \frac{1}{4L\sqrt{mT}}$ and $\beta = \varrho T^{\frac{r}{4}}$ from Theorem 5 results

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\Pi_{\mathbb{R}^m_+}\left(g(\tilde{\mathbf{x}}_i^T)\right)\right\|^2 \leq \left(\frac{LRQ_\psi\sqrt{nm}}{\varrho T^{\frac{r}{4}}} + \frac{2m}{L\sqrt{T}}\right)\cdot\left(\mathsf{F} + \mathcal{O}\left(\frac{1}{T^{\frac{1-r}{2}}}\right)\right), \tag{58}$$

whence,

$$\left\|\Pi_{\mathbb{R}^m_+}\left(g(\tilde{\mathbf{x}}_i^T)\right)\right\|^2 = \mathcal{O}\left(\frac{LRQ_\psi n\sqrt{nm}}{\varrho T^{\frac{r}{4}}}\right).$$

To prove the remaining part of Theorem 5, we have the following lemma:

Lemma 11. *Suppose the optimal solution* $\mathbf{x}^*$ *satisfies the inequality constraints strictly* $g_k(\mathbf{x}^*) < 0, k = 1, 2, \cdots, m$. *Then,*

$$f_i(\tilde{\mathbf{x}}_i^T) - f_i(\mathbf{x}^*) \geq 0,$$

for all $i \in \mathcal{V}$ *and* $t \in [T]$, *i.e.,* $\mathsf{F} = 0$.

Proof. Due to convexity of $f_i, i = 1, 2, \cdots, n$ we have

$$f_i(\tilde{\mathbf{x}}_i^T) - f_i(\mathbf{x}^*) \geq \langle \nabla f_i(\mathbf{x}^*), \tilde{\mathbf{x}}_i^T - \mathbf{x}^*\rangle. \tag{59}$$

We now write the Karush-Kuhn-Tucker (KKT) conditions [BV04] for the optimal solution $\mathbf{x}^*$ and the optimal Lagrangian multiplier $\lambda^* \equiv (\lambda_1^*, \lambda_2^*, \cdots, \lambda_m^*)$,

[C1] $\mathbf{0} \in \partial f(\mathbf{x}^*) + \sum_{k=1}^m \lambda_k^* \cdot \partial g_k(\mathbf{x}^*)$,

[C2] $\lambda_k^* \cdot g_k(\mathbf{x}^*) = 0$,

[C3] $g(\mathbf{x}^*) \preceq \mathbf{0}$ and $\lambda^* \succeq \mathbf{0}$.

From [C2] we note that $\lambda_k^* = \mathbf{0}$ since $g_k(\mathbf{x}^*) < 0$ for $k \in [m]$. Consequently, from [C1] we conclude $\mathbf{0} \in \partial f(\mathbf{x}^*)$. Choosing $\nabla f_i(\mathbf{x}^*) = \mathbf{0}$ in Eq. (59) completes the proof. □

Equation (17) in Theorem 5 now follows from Eq. (58) by equating $\mathsf{F} = 0$.

A.6 Proof of Theorem 6

Analogous to the derivation in Eq. (68), for any realization of random variables $(x_i^t, \lambda_i^t, K_i^t)_{t=1}^T$ it can be shown that for all $\mathbf{x} \in \mathcal{X}$,

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle \nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t; K_i^t), \mathbf{x}_i^t - \mathbf{x}\rangle \leq n\|\mathbf{x}\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\|\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t, \lambda_i^t; K_i^t)\|_*^2. \tag{60}$$

Recall the definition

$$\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t) \equiv f_i(\mathbf{x}_i^t) + \langle \lambda_i^t, g(\mathbf{x}_i^t)\rangle - \psi(\lambda_i^t). \tag{61}$$

By putting together the ienquality (60) and the definition (61) we obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t),\mathbf{x}_i^t-\mathbf{x}\rangle \leq \frac{n\|\mathbf{x}\|^2}{2\alpha}+\frac{\alpha}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t)\|_*^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}\rangle. \tag{62}$$

Further, similar to Eq. (69) and due to the fact that $\nabla_\lambda\mathfrak{L}(\mathbf{x}^t,\lambda^t)=\nabla_\lambda\widehat{\mathfrak{L}}(\mathbf{x}^t,\lambda^t)$, we derive

$$\sum_{t=1}^{T}\langle\nabla_\lambda\mathfrak{L}(\mathbf{x}_i^t,\lambda_i^t),\lambda-\lambda_i^t\rangle \leq \frac{1}{2\alpha}\|\lambda\|^2+\frac{\alpha}{2}\sum_{t=1}^{T}\|\nabla_\lambda\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2. \tag{63}$$

Thus, following the footsteps of the proof of Proposition 8 (cf. Appendix B.2) we derive for $\mathbf{x}=\mathbf{x}^*\in\mathcal{X}$ that

$$\begin{aligned} f(\tilde{\mathbf{x}}_j^T)-f(\mathbf{x}^*) \leq{}& \frac{\|\mathbf{x}^*\|^2}{2T\alpha}+\frac{\alpha}{2nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\|\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t)\|_*^2+\|\nabla_\lambda\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2\right)\\ &+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}}f_i(\mathbf{x}_j^t)\|_*\|\mathbf{x}_j^t-\mathbf{x}_i^t\|-\frac{\eta}{2nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\|\lambda_i^t\|^2\\ &+\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}\rangle. \end{aligned} \tag{64}$$

We now obtain a high probability bound for the last term of the expression in Eq. (64). First notice that

$$\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\equiv\mathbb{E}_{p_i^t}[\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t)|\mathcal{F}_{t-1}].$$

As the result, we can compute the expectation of the last term with $\mathbf{x}=\mathbf{x}^*$. Specifically, since $\mathbf{x}^t\in\mathcal{F}_{t-1}$ we can write

$$\begin{aligned} &\mathbb{E}\big[\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}^*\rangle\big]\\ &\qquad=\mathbb{E}\big[\mathbb{E}\big[\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}^*\rangle\big]\big|\mathcal{F}_{t-1}\big]\\ &\qquad=\mathbb{E}\big[\langle\mathbb{E}\big[\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t)\big|\mathcal{F}_{t-1}\big],\mathbf{x}_i^t-\mathbf{x}^*\rangle\big]=0. \end{aligned} \tag{65}$$

Moreover, from the upper bound on the dual variables in Eq. (20), we can obtain that

$$\begin{aligned} \langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}^*\rangle &\leq \|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}(\mathbf{x}_i^t,\lambda_i^t;K_i^t)\|_*\|\mathbf{x}_i^t-\mathbf{x}^*\|\\ &\leq 2LR\|\lambda_i^t\|_1\\ &=2LR\beta\sqrt{m}, \end{aligned} \tag{66}$$

where the last step follows from the fact that $\|\lambda_i^t\|_1\leq\sqrt{m}\|\lambda_i^t\|_2$. Applying the Azuma-Hoeffding inequality [CL06] yields the tail bound,

$$\mathbb{P}\left[\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}^*\rangle\geq\delta\right]\leq\exp\left(-\frac{\delta^2}{8mTL^2R^2\beta^2}\right).$$

Hence, with probability of at least $1-\varepsilon$ we have

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)-\nabla_{\mathbf{x}}\widehat{\mathfrak{L}}_i(\mathbf{x}_i^t,\lambda_i^t;K_i^t),\mathbf{x}_i^t-\mathbf{x}\rangle\leq 3LR\beta\sqrt{\frac{m\log\frac{1}{\varepsilon}}{T}}. \tag{67}$$

The rest of the proof can be carried out similar to the proof in Appendix A.4.

To prove the second part of Theorem 6, we compute the expectation of both sides of inequality (64) and use the fact that the expectation of the last term in Eq. (64) is zero due to Eq. (65).

B Auxiliary Results

B.1 Proof of Young's Inequality (28)

For any $\delta>0$ and $a,b\in\mathbb{R}^m$ we have

$$\begin{aligned}\|a+b\|^2&\leq(\|a\|+\|b\|)^2\\&\leq\|a\|^2+\|b\|^2+2\|a\|\|b\|\\&=\|a\|^2+\|b\|^2+(\sqrt{2/\delta}\|a\|)(\sqrt{2\delta}\|b\|)\\&\leq\|a\|^2+\|b\|^2+\frac{1}{2}(2\delta^{-1}\|a\|^2+2\delta\|b\|^2)\\&=(1+\delta^{-1})\|a\|^2+(1+\delta)\|b\|^2.\end{aligned}$$

B.2 Proofs of Proposition 8 and Proposition 9

From the non-expansive property of the projection, we obtain that for all $\mathbf{x}\in\mathcal{X}$,

$$\begin{aligned}\|\mathbf{x}_i^{t+1}-\mathbf{x}\|^2&=\left\|\Pi_{\mathcal{X}}\left(\sum_{j=1}^{n}[W]_{ij}\hat{\mathbf{x}}_i^t\right)-\mathbf{x}\right\|^2\\&\leq\left\|\sum_{j=1}^{n}[W]_{ij}\hat{\mathbf{x}}_j^t-\mathbf{x}\right\|^2\\&\overset{\text{(a)}}{\leq}\sum_{j=1}^{n}[W]_{ij}\|\mathbf{x}_j^t-\mathbf{x}-\alpha\nabla_{\mathbf{x}}\mathfrak{L}_j(\mathbf{x}_j^t,\lambda_j^t)\|^2\\&=\sum_{j=1}^{n}[W]_{ij}\left(\|\mathbf{x}_j^t-\mathbf{x}\|^2-2\alpha\langle\nabla_{\mathbf{x}}\mathfrak{L}_j(\mathbf{x}_j^t,\lambda_j^t),\mathbf{x}_j^t-\mathbf{x}\rangle+\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_j^t(\mathbf{x}_j^t,\lambda_j^t)\|_*^2\right),\end{aligned}$$

where (a) follows from the convexity of the norm square, $\|\cdot\|_*$ is the dual norm. Computing the summation over $i=1,2,\cdots,n$ results

$$\begin{aligned}\sum_{i=1}^{n}\|\mathbf{x}_i^{t+1}-\mathbf{x}\|^2&\leq\sum_{i=1}^{n}\sum_{j=1}^{n}[W]_{ij}\left(\|\mathbf{x}_j^t-\mathbf{x}\|^2-2\alpha\langle\nabla_{\mathbf{x}}\mathfrak{L}_j(\mathbf{x}_j^t,\lambda_j^t),\mathbf{x}_j^t-\mathbf{x}\rangle+\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_j^t(\mathbf{x}_j^t,\lambda_j^t)\|^2\right)\\&=\sum_{i=1}^{n}\left(\|\mathbf{x}_i^t-\mathbf{x}\|^2-2\alpha\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t),\mathbf{x}_i^t-\mathbf{x}\rangle+\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t,\lambda_i^t)\|_*^2\right),\end{aligned}$$

where we used the fact that $\sum_{i=1}^{n}[W]_{ij} = 1$ as W is a doubly stochastic matrix. From this recursion and by noting that $\mathbf{x}_i^0 = \mathbf{0}$, we derive

$$\sum_{i=1}^{n}\|\mathbf{x}_i^{T+1} - \mathbf{x}\|^2 \leq n\|\mathbf{x}\|^2 - \sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t), \mathbf{x}_i^t - \mathbf{x}\rangle + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2.$$

Since the left hand side is non-negative, we obtain the following inequality

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t), \mathbf{x}_i^t - \mathbf{x}\rangle \leq n\|\mathbf{x}\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2. \tag{68}$$

We can similarly show that

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle\nabla_{\lambda}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t), \lambda - \lambda_i^t\rangle \leq n\|\lambda\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\|\nabla_{\lambda}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2. \tag{69}$$

Due to convexity of $\mathfrak{L}_i(\cdot, \lambda_i)$ and concavity of $\mathfrak{L}_i(\mathbf{x}_i, \cdot)$, the following pair of inequalities hold respectively,

$$\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t) - \mathfrak{L}_i(\mathbf{x}, \lambda_i^t) \leq \langle\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t), \mathbf{x}_i^t - \mathbf{x}\rangle \tag{70}$$
$$\mathfrak{L}_i(\mathbf{x}_i^t, \lambda) - \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t) \leq \langle\nabla_{\lambda}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t), \lambda - \lambda_i^t\rangle. \tag{71}$$

Combining Eqs. (68) and (69) coupled with the inequalities (70)-(71) results in

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\left(\mathfrak{L}_i(\mathbf{x}_i^t, \lambda) - \mathfrak{L}_i(\mathbf{x}, \lambda_i^t)\right) \leq & n(\|\mathbf{x}\|^2 + \|\lambda\|^2) \\ & + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\left(\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2 + \|\nabla_{\lambda}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2\right).\end{aligned} \tag{72}$$

This completes the proof of Proposition 9.

To prove Proposition 8, we first combine Eq. (68) with Eq. (70) to obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\left(\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t) - \mathfrak{L}_i(\mathbf{x}, \lambda_i^t)\right) \leq n\|\mathbf{x}\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2. \tag{73}$$

Let $\mathbf{x} = \mathbf{x}^*$. Expanding $\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)$ and $\mathfrak{L}_i(\mathbf{x}^*, \lambda_i^t)$ on the l.h.s. gives us

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\left(f_i(\mathbf{x}_i^t) - f_i(\mathbf{x}^*)\right) + \sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\left(\langle\lambda_i^t, g(\mathbf{x}_i^t)\rangle - \langle\lambda_i^t, g(\mathbf{x}^*)\rangle\right) \\ \leq n\|\mathbf{x}\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n}\alpha^2\|\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2.\end{aligned} \tag{74}$$

Since $\lambda_i^t \succeq \mathbf{0}$ and $g(\mathbf{x}^*) \preceq \mathbf{0}$, we have $-\langle \lambda_i^t, g(\mathbf{x}^*)\rangle \succeq \mathbf{0}$. Therefore, we can eliminate it from the l.h.s. of Eq. (74),

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha \left(f_i(\mathbf{x}_i^t) - f_i(\mathbf{x}^*)\right) + \sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle \lambda_i^t, g(\mathbf{x}_i^t)\rangle \\ \leq n\|\mathbf{x}\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n} \alpha^2 \|\nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2. \tag{75}$$

To obtain an upper bound in terms of $\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha(f_i(\mathbf{x}_j^t) - \sum_{i=1}^{n} f_i(\mathbf{x}))$, we use the convexity of $f_i(\cdot)$,

$$f_i(\mathbf{x}_j^t) + \langle \nabla_{\mathbf{x}} f_i(\mathbf{x}_j^t), \mathbf{x}_i^t - \mathbf{x}_j^t\rangle \leq f_i(\mathbf{x}_i^t). \tag{76}$$

Substituting the inequality (76) in Eq. (77) combined with the Cauchy-Schwarz inequality and the definition $f(\cdot) \equiv \frac{1}{n}\sum_{i=1}^{n} f_i(\cdot)$ gives us

$$\sum_{t=1}^{T} 2\alpha \left(f(\mathbf{x}_j^t) - f(\mathbf{x}^*)\right) + \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle \lambda_i^t, g(\mathbf{x}_i^t)\rangle \\ \leq \|\mathbf{x}\|^2 + \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} \alpha^2 \|\nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2 + \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha \|\nabla_{\mathbf{x}} f_i(\mathbf{x}_j^t)\|_* \cdot \|\mathbf{x}_j - \mathbf{x}_i\|. \tag{77}$$

In the remainder of the proof, we establish a lower bound on $\frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle \lambda_i^t, g(\mathbf{x}_i^t)\rangle$. To do so, we combine Eq. (71) with Eq. (69) to derive

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha \left(\mathfrak{L}_i(\mathbf{x}_i^t, \lambda) - \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\right) \leq n\|\lambda\|^2 + \sum_{t=1}^{T}\sum_{i=1}^{n} \alpha^2 \|\nabla_{\lambda} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2. \tag{78}$$

We let $\lambda = \mathbf{0}$ and then expand the l.h.s. of Eq. (78),

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha \left(\psi(\lambda_i^t) - \psi(\mathbf{0}) - \langle \lambda_i^t, g_i(\mathbf{x}_i^t)\rangle\right) \leq \sum_{t=1}^{T}\sum_{i=1}^{n} \alpha^2 \|\nabla_{\lambda} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2. \tag{79}$$

From Condition *(i)* in Def. 1 we have $\psi(\mathbf{0}) = 0$ and $\psi(\lambda_i^t) \geq 0$. Hence,

$$\sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\psi(\lambda_i^t) - \sum_{t=1}^{T}\sum_{i=1}^{n} \alpha^2 \|\nabla_{\lambda} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2 \leq \sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle \lambda_i^t, g_i(\mathbf{x}_i^t)\rangle. \tag{80}$$

Now, we have the following Lemma:

Lemma 12. *For all $\lambda \in \mathbb{R}^m$ we have*

$$\psi(\lambda) \geq \frac{\eta}{2}\|\lambda\|^2.$$

Proof. Due to the strong convexity of the regularizer ψ, we have

$$\psi(\lambda) - \psi(\hat{\lambda}) - \langle \nabla\psi(\hat{\lambda}), \lambda - \hat{\lambda}\rangle \geq \frac{\eta}{2}\|\lambda - \hat{\lambda}\|^2, \quad \forall \lambda, \hat{\lambda} \in \mathbb{R}^m_+, \tag{81}$$

Let $\hat{\lambda} = \mathbf{0}$. Then, due the first condition in Def. 1 we have $\nabla\psi(\mathbf{0}) = \mathbf{0}$ and $\psi(\mathbf{0}) = 0$ which gives the desired result. □

Based on Lemma 12 we obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{n} \alpha\eta\|\lambda_i^t\|^2 - \sum_{t=1}^{T}\sum_{i=1}^{n} \alpha^2\|\nabla_\lambda \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\|_*^2 \leq \sum_{t=1}^{T}\sum_{i=1}^{n} 2\alpha\langle \lambda_i^t, g_i(\mathbf{x}_i^t)\rangle. \tag{82}$$

Subsuituing this lower bound in Eq. (77) gives us Eq. (47).

C Some Results on Projection

C.1 Projection onto the Ball

In this appendix, we proof a closed form solution for the projection onto the balls.

Lemma 13. *For any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\Pi_{\mathbb{B}(r)}(\mathbf{x}) = \frac{r \cdot \mathbf{x}}{\max\{r, \|\mathbf{x}\|\}},$$

where $\mathbb{B}(r)$ is the ℓ_2-ball of radius r.

Proof. If $\mathbf{x} = \mathbf{0}$, the identity is trivial. So, assume that $\mathbf{x} \neq 0$ and for any vector $\mathbf{y} \in \mathbb{B}(r)$ write the decomposition $\mathbf{y} = \omega\mathbf{x} + \mathbf{w}$, where $\mathbf{w} \perp \mathbf{x}$. Due to the fact that $\|\mathbf{y}\|^2 = \omega^2\|\mathbf{x}\|^2 + \|\mathbf{w}\|^2$ and since $\mathbf{y} \in \mathbb{B}(r)$ we conclude $\omega\mathbf{x} \in \mathbb{B}(r)$. This in turn implies $\omega \leq r/\|\mathbf{x}\|$. Moreover,

$$\|\mathbf{x} - \mathbf{y}\|^2 = (1-\omega)^2\|\mathbf{x}\|^2 + \|\mathbf{w}\|^2.$$

i.e., $(1-\omega)^2\|\mathbf{x}\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$. Therefore, $\Pi_{\mathbb{B}(r)}(\mathbf{x}) = \min_{\omega\mathbf{x}\in\mathbb{B}(r)}(1-\omega)^2\|\mathbf{x}\|^2$. Now, it is easy to see that $\omega = \min\left\{1, \frac{r}{\|\mathbf{x}\|}\right\}$ and thus $\Pi_{\mathbb{B}(r)}(\mathbf{x}) = \min\left\{1, \frac{r}{\|\mathbf{x}\|}\right\}\mathbf{x} = \frac{r \cdot \mathbf{x}}{\max\{r, \|\mathbf{x}\|\}}$. □

C.2 The Non-expansive Property of the Projection

In this appendix we provide a proof for the non-expansive property of the projection. The proof can also be found in [HUL96, Chapter III.3].

Lemma 14. *For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following inequality holds*

$$\|\Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|_*,$$

where $\mathcal{X} \subset \mathbb{R}^d$.

Proof. Recall that $\Pi_{\mathcal{X}}(\mathbf{x}) = \arg\min_{\mathbf{u}\in\mathcal{X}} \|\mathbf{x}-\mathbf{u}\|$. This implies that

$$\langle \mathbf{y} - \Pi_{\mathcal{X}}(\mathbf{y}), \mathbf{z} - \Pi_{\mathcal{X}}(\mathbf{y})\rangle \leq 0, \quad \forall \mathbf{z} \in \mathcal{X}.$$

By setting $\mathbf{z} = \Pi_{\mathcal{X}}(\mathbf{x})$ we thus obtain

$$\langle \mathbf{y} - \Pi_{\mathcal{X}}(\mathbf{y}), \Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y})\rangle \leq 0. \tag{83}$$

Similarly,

$$\langle \mathbf{x} - \Pi_{\mathcal{X}}(\mathbf{x}), \Pi_{\mathcal{X}}(\mathbf{y}) - \Pi_{\mathcal{X}}(\mathbf{x})\rangle \leq 0. \tag{84}$$

Combining Eqs. (83) and (84) gives us

$$\langle \Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y}), \Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y})\rangle \leq \langle \mathbf{x}-\mathbf{y}, \Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y})\rangle.$$

From the Cauchy-Schwarz inequality we derive

$$\|\Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y})\|^2 \leq \|\Pi_{\mathcal{X}}(\mathbf{x}) - \Pi_{\mathcal{X}}(\mathbf{y})\| \cdot \|\mathbf{x}-\mathbf{y}\|_*,$$

which implies the desired inequality.

□