Distributed Regularized Primal-Dual Method

Masoud Badiei, Na Li
John Paulson School of Engineering and Applied Sciences
Harvard University
mbadieikhuzani@g.harvard.edu.
nali@seas.harvard.edu.

***Abstract*—We study a deterministic primal-dual subgradient method for distributed optimization of a separable objective function with global inequality constraints. To control the norm of dual variables, we augment the Lagrangian function with a regularizer on dual variables. Specifically, we show that under a certain restriction on the step size of the underlying algorithm, the norm of dual variables is inversely proportional to the regularizer's curvature. We leverage this result to bound the consensus terms and subsequently establish the convergence rate of the distributed primal-dual algorithm. We also establish an asymptotic decay rate on the norm of the constraint violation. We exhibit a tension between the convergence rate of the underlying algorithm and the decay rate associated with the constraint violation. Specifically, we show that when one of the inequality constraints is binding at optimal solutions, improving the convergence rate in the objective value deteriorates the decay rate of the constraint violation bound and vice versa. However, in the case that one optimal solution satisfies the inequality constraints strictly, the tension vanishes.**

I. Introduction

Network optimization is a framework for distributing 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 $f_i : \mathbb{R}^d \to \mathbb{R}$. Further, each node coordinates its actions with other nodes through local communication with adjacent nodes in the network. In this paper, we study a distributed primal-dual (PD) 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, $\mathcal{X} \subset \mathbb{R}^d$ is a non-empty, closed, and convex proper subset, and $\preceq$ denotes the element-wise inequality. Including some constraints in the explicit form as in Eq. (1b) simplifies the constraint set $\mathcal{X}$ which in turn results in a computationally efficient projection step in the algorithm. In this paper, we will introduce dual variables to handle the constraints (1b) and we will examine the effect of inequality constraints in Eq. (1b) on the convergence rate of the distributed algorithm.

The important practical applications of distributed constrained optimization have led to much interest in this problem. In most of the existing frameworks, the constraints take implicit forms [1], [2], [3]. For instance, in the absent of explicit inequality constraints in Eq. (1b), a dual averaging algorithm is proposed [1] where there is a global constraint set $\mathcal{X}$ on agents' actions. When the constraint set has a further structure that can be written as the intersection of finitely many simple convex constraints, a distributed random projection algorithm is proposed [2]. In the proposed distributed scheme, projections are performed locally by each agent based on the random observations of the local constraint components [2].

In the case of optimization with coupled linear equality constraints, i.e., when decision variables of agents must jointly satisfy a set of linear equality constraints, penalty and barrier function methods are established [4]. Moreover, based on a game theoretic argument, the convergence of those methods are established. For distributed optimization with a set of global non-linear inequality constraints, distributed primal-dual methods similar to this paper are studied in [5], [6]. A variation of the primal-dual method is also studied [7] for the case where each agent has local (private) inequality constraints as well as a constraint set. However, those proposed methods require 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 and convergence of the proposed PD algorithms depend on a Slater vector for the inequality constraints, i.e., it is inversely proportional to 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, dependency 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. Our approach also manifests the natural trade-off between the convergence rate of the objective value and the rate at which the constraint violation in the inequality constraints vanishes. In summary, our contributions of this paper includes i) a primal-dual method with regularizer on dual variables, ii) an upper bound on the norm of dual variables that is inversely proportional to the regularizer's curvature, and iii) convergence rates of the distributed regularized primal-dual method for both the objective value and constraint violation. Lastly, we verify our proposed method through numerical simulations for distributed optimization of logistic loss function.

Notation: For ease of notation, we denote the ℓ_2-norm by $\|\cdot\|$. We also use standard asymptotic notation. Specifically, if a_n and b_n are positive sequences, then $a_n = \mathcal{O}(b_n)$ means that $\limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$. We denote the vectors as $\mathbf{a} \equiv (a_1, a_2, \cdots, a_d)$. For two vectors $\mathbf{a}$ and $\mathbf{b}$, the vector inequality $\mathbf{a} \preceq \mathbf{b}$ means the element-wise inequality, i.e., $a_i \leq b_i$ for all $i = 1, 2, \cdots, d$. Lastly, we denote the projection of the vector $\mathbf{x}$ onto the closed set $\mathcal{X}$ by $\Pi_{\mathcal{X}}(\mathbf{x}) \equiv \arg\min_{\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|_2$.

978-1-5090-4545-7/16/$31.00 ©2016 IEEE

II. Problem Statement

We consider a multi-agent optimization problem, consisting of n nodes that exchange information on edges of a given 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. By default, we assume $n \geq 2$. 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 $g_k(\mathbf{x}) \leq 0, k \in [m] \equiv \{1, 2, \cdots, m\}$ on the actions of agents. 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 subset $\mathcal{X}$ is known to each node of the network and has a finite diameter $R \equiv \sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$. Without loss of generality, 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 vector $\Delta \in \mathbb{R}^d$ 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 Lipschitz continuous, i.e.,

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

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. We notice that since $g_k(\cdot)$ are Lipschitz continuous on $\mathcal{X}$, they satisfy the linear growth condition $|g(\mathbf{x})| \leq L'_{g_k}(1 + \|\mathbf{x}\|)$ for some constant $L'_{g_k} > 0$ and thus are bounded. In particular, the Lipschitz continuity assumption of $g_k(\cdot)$ implies

$$|g_k(\mathbf{x})| \leq L'_{g_k}(1 + R), \quad k \in [m]. \tag{2}$$

Let $L \equiv \max\{\{L_{f_i}\}_{i=1}^n, \{L_{g_k}, L'_{g_k}\}_{k=1}^m\}$.

We assume that the optimal solution and optimal Lagrangian multiplier of the probelm (1a)-(1b) exist, and we denote them by $\mathbf{x}^*$ and λ^*, respectively. 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 $\lambda \equiv (\lambda^1, \lambda^2, \cdots, \lambda^m)$. Each λ^k corresponds to one constraint $g_k(\mathbf{x})$.

Based on this formulation, we design a distributed primal-dual algorithm for Eq. (3) such that the inequality constraint $g(\mathbf{x}) \preceq 0$ is satisfied *asymptotically* as $T \to \infty$. The distributed method we present is based on a 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 $\|\cdot\|$,

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

(*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., $Q_\psi \equiv \gamma/\eta$.

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 where $\langle \nabla\psi(\mathbf{0}), \lambda\rangle \geq 0$ for all $\lambda \in \mathbb{R}_+^m$ can be treated similarly.

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

$$\mathfrak{L}_i(\mathbf{x}, \lambda) \equiv f_i(\mathbf{x}) + \langle \lambda, g(\mathbf{x})\rangle - \psi(\lambda). \tag{4}$$

Furthermore,

$$\nabla_{\mathbf{x}}\mathfrak{L}_i(\mathbf{x}, \lambda) \equiv \nabla f_i(\mathbf{x}) + \langle \lambda, \nabla g(\mathbf{x})\rangle \tag{5a}$$
$$\nabla_\lambda\mathfrak{L}_i(\mathbf{x}, \lambda) \equiv g(\mathbf{x}) - \nabla\psi(\lambda). \tag{5b}$$

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 $\nabla f_i(\mathbf{x})$ and $\nabla g_k(\mathbf{x})$ to denote both gradient and subgradient when f_i and g_k are differentiable and non-differentiable, respectively.

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

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\lambda\in\mathbb{R}_+^m} \frac{1}{n}\sum_{i=1}^n \mathfrak{L}_i(\mathbf{x}, \lambda). \tag{6}$$

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 singular 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. [9]). 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 Regularized Primal-Dual Method

```
Initialize: $\mathbf{x}_i^0 = \mathbf{0}$, $\lambda_i^0 = \mathbf{0}, \forall i \in \mathcal{V}$ and a constant step size $\alpha \in \mathbb{R}_+$.
for $t = 0, 1, 2, \cdots, T$ at the $i$-th node do
    Update the primal and dual variables
        $\widehat{\mathbf{x}}_i^t = \mathbf{x}_i^t - \alpha \nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)$
        $\widehat{\lambda}_i^t = \lambda_i^t + \alpha \nabla_{\lambda} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)$.
    Run the consensus steps
        $\mathbf{x}_i^{t+1} = \Pi_{\mathcal{X}}\left(\sum_{j=1}^{n} [W]_{ij} \widehat{\mathbf{x}}_j^t\right)$,
        $\lambda_i^{t+1} = \Pi_{\mathbb{R}_+^m}\left(\sum_{j=1}^{n} [W]_{ij} \widehat{\lambda}_j^t\right)$.
end for
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. (6); see Algorithm 1.

III. Convergence Analysis

In this section, we analyze the convergence of Algorithm 1. Due to the space limitations, we omit the proofs of the following results. The proofs can be found in [10].

A. Main Results

As the first result, we prove 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 variables λ_i^t is bounded by*

$$\|\lambda_i^t\| \leq \frac{2L(1+R)\sqrt{nm}}{\eta}, \tag{7}$$

for all $t \in [T]$. Specifically, for the choice of $\eta = \frac{2L(1+R)\sqrt{nm}}{\beta}, \beta > 0$ we have

$$\|\lambda_i^t\| \leq \beta. \tag{8}$$

Proposition 1 highlights the role of the regularizer ψ 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. (7) is also intuitive. The cost associated with choosing a large lagrangian dual variable λ_i^t increases for larger η which results in a smaller norm $\|\lambda_i^t\|$.

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

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

$$\|\nabla_\lambda \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\| \leq 3L(1+R)Q_\psi\sqrt{mn}$$
$$\|\nabla_{\mathbf{x}} \mathfrak{L}_i(\mathbf{x}_i^t, \lambda_i^t)\| \leq L\tilde{\beta},$$

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 measure deviation between agents' decision variables.

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

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

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.

By putting together Propositions 1 and 3, we arrive at the main result of this paper:

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

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

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

$$f(\tilde{\mathbf{x}}_j^T) - f(\mathbf{x}^*) \leq \frac{3(1+R)^2 L\sqrt{m}}{\sqrt{T}} + \frac{13\varrho^2 L\sqrt{m}}{1-\sigma_2(W)}\frac{\log(T\sqrt{n})}{T^{\frac{1-r}{2}}}, \tag{9}$$

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

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

Theorem 5. *Consider the step size α and the regularizer's curvature η as defined in Theorem 4 and $r \in [0,1)$. The norm of the constraint violation asympotically decays to zero with the following rate*

$$\left\|\Pi_{\mathbb{R}_+^m}\left(g(\tilde{\mathbf{x}}_i^T)\right)\right\|^2 = \mathcal{O}\left(\frac{L(1+R)Q_\psi n\sqrt{nm}}{\varrho T^{\frac{r}{4}}}\right), \tag{10}$$

for all $i \in \mathcal{V}$. Furthermore, if the optimal solution $\mathbf{x}^*$ is strictly feasible $g(\mathbf{x}^*) \prec \mathbf{0}$, we have

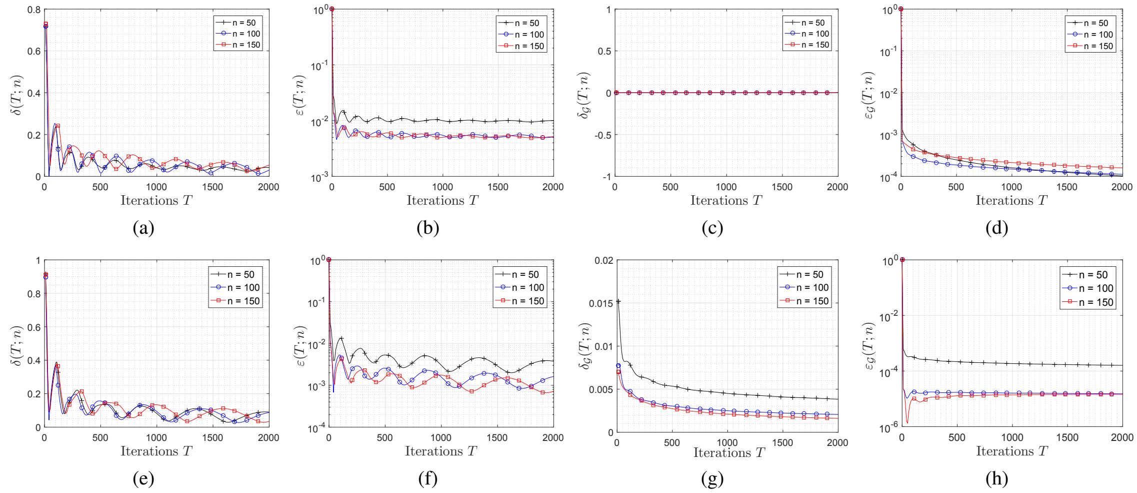
$$\left\|\Pi_{\mathbb{R}_+^m}\left(g(\tilde{\mathbf{x}}_i^T)\right)\right\|^2 = \mathcal{O}\left(\frac{L(1+R)Q_\psi n\sqrt{nm}}{\varrho T^{\frac{1}{2}-\frac{r}{4}}}\right). \tag{11}$$

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. (9) that grows unboundedly as $T \to \infty$. The case of $r = 0$ is however, more subtle. 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. (10). However, 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. (9) and (11).

Remark 3. It is easy to verify that if the constant parameter ϱ incorporated in η in Theorem 4 and Theorem 5 satisfies

$$\varrho \geq \frac{(1+R)Q_\psi^2\sqrt{n}}{T^{\frac{1}{2}+\frac{r}{4}}}, \tag{12}$$

then 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 simply put $\varrho = 1$.



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

(e) (f) (g) (h)

Fig. 1: 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 $l = u = 0.1$ (top) and $l = u = 0.001$ (bottom), Panels (a)-(e): Constraint violation $\delta(T;n)$ of the centeralized PD algorithm without regularization, Panels (b)-(f): Convergence rate $\varepsilon(T;n)$ of the centeralized PD algorithm without regularization, Panels (c)-(g) Constraint violation $\delta_{\mathcal{G}}(T;n)$ of the decenteralized regularized PD algorithm, Panels (d)-(h): Convergence rate $\varepsilon_{\mathcal{G}}(T;n)$ of the decenteralized regularized PD algorithm.

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 objective value and the decay rate of the constraint violation bound for the distributed primal-dual algorithm. More specifically, adopting a small value for $r \in (0,1]$ improves the convergence rate in Eq. (9) while detoriates the constraint violation bound in Eq. (10). 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.

IV. 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 consider a logistic loss function with a norm constraint as well as a set of box constraints

$$\min_{\mathbf{x}} f(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(\mathbf{b}_i\langle \mathbf{a}_i, \mathbf{x}\rangle)) \tag{13a}$$

$$\begin{aligned} \text{subject to} \quad & g_k(\mathbf{x}) = -l - x_k \leq 0, && \text{(13b)} \\ & g_{k+d}(\mathbf{x}) = x_k - u \leq 0, \quad k = 1,\cdots,d \\ & \mathbf{x} \in \mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\right\}, \end{aligned}$$

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 $l = u \in \{0.1, 0.001\}$. In our simulations, we use Watts-Strogratz graph model [11] with parameters $K = 20$, $\vartheta = 0.02$. For this graph $\mathcal{G}$ and for 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. Fig. 1 shows the convergence results for the centeralized primal-dual algorithm without regularization as well as decenteralized regularized primal-dual method. Further, the constraint violation associated with each method is shown. Interestingly, due to presence of a regularizer on dual variables, we observe that the decenteralized scheme has a smaller constraint violation than the centeralized scheme without regularization.

V. Conclusion

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 primal dual algorithm in the sense that when one of the inequality constraints is binding at optimal solutions, there is a tension between the convergence rate of the objective value and the diminishing rate of the constraint violation. Nevertheless, this tension vanishes when the constraints are satisfied strictly.

References

[1] J. C. Duchi, A. Agarwal, and M. J. Wainwright, “Dual averaging for distributed optimization: convergence analysis and network scaling,” *Automatic control, IEEE Transactions on*, vol. 57, no. 3, pp. 592–606, 2012.

[2] S. Lee and A. Nedic, “Distributed random projection algorithm for convex optimization,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 221–229, 2013.

[3] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.

[4] N. Li and J. R. Marden, “Decoupling coupled constraints through utility design,” *Automatic Control, IEEE Transactions on*, vol. 59, no. 8, pp. 2289–2294, 2014.

[5] D. Yuan, S. Xu, and H. Zhao, “Distributed primal–dual subgradient method for multiagent optimization via consensus algorithms,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 6, pp. 1715–1724, 2011.

[6] M. Zhu and S. Martínez, “On distributed convex optimization under inequality and equality constraints,” *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.

[7] T.-H. Chang, A. Nedić, and A. Scaglione, “Distributed constrained optimization by consensus-based primal-dual perturbation method,” *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[8] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient projections onto the l 1-ball for learning in high dimensions,” in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 272–279.

[9] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[10] [Online]. Available: http://nali.seas.harvard.edu/files/nali/files/primaldual.pdf

[11] D. J. Watts and S. H. Strogatz, “Collective dynamics of small-worldnetworks,” *nature*, vol. 393, no. 6684, pp. 440–442, 1998.