Gradient-Free Multi-Agent Nonconvex Nonsmooth Optimization

Davood Hajinezhad and Michael M. Zavlanos

***Abstract*— In this paper, we consider the problem of minimizing the sum of nonconvex and possibly nonsmooth functions over a connected multi-agent network, where the agents have partial knowledge about the global cost function and can only access the zeroth-order information (i.e., the functional values) of their local cost functions. We propose and analyze a distributed primal-dual gradient-free algorithm for this challenging problem. We show that by appropriately choosing the parameters, the proposed algorithm converges to the set of first order stationary solutions with a provable global sublinear convergence rate. Numerical experiments demonstrate the effectiveness of the proposed method for optimizing nonconvex and nonsmooth problems over a network.**

I. Introduction

Consider a network with N distributed agents that collectively solve the following optimization problem

$$\min_{x\in\mathbb{R}^M} \; f(x) := \sum_{i=1}^{N} f_i(x). \tag{1}$$

Here, $f_i : \mathbb{R}^M \to \mathbb{R}$ is possibly a nonconvex nonsmooth function. Such distributed optimization problems arise in many applications; see e.g., [1]. Many distributed optimization methods have been proposed to solve problem (1). One of the first such methods is the distributed subgradient (DSG) algorithm [2] and its related algorithms; see, e.g., [3], [4]. These methods converge to a neighborhood of the solution set unless they use diminishing stepsizes, which often makes them slow. Faster algorithms using constant stepsizes include the incremental aggregated gradient (IAG) method [5], the exact first-order algorithm (EXTRA) [6], and Accelerated Distributed Augmented Lagrangian (ADAL) algorithm [7], [8] for optimization of convex problems. Optimization of nonconvex functions is a much more challenging problem. Only recently, have there been developed a few nonconvex distributed optimization algorithms; see e.g., [9]–[13].

Regardless of convexity and/or smoothness, all the aforementioned methods require that either the first order (gradient/subgradient) information or the explicit form of the objective function is available to the agents. However, in many important practical situations such information can be expensive to obtain, or even impossible. Examples include, simulation-based optimization where the objective function can only be evaluated using repeated simulation, training deep neural networks where the relationship between the variables and the cost function is too complicated to derive an explicit form of the gradient, and bandit optimization where a player optimizes a sequence of cost functions having only knowledge of a single function value each time. In these cases, the zeroth-order information about the objective function values is often readily available. Such zeroth-order information can be obtained through a stochastic zeroth-order oracle ($\mathcal{SZO}$). In particular, suppose that $\hat{x} \in \text{dom}(f_i)$, then the ith $\mathcal{SZO}$ at agent i returns a noisy version of $f_i(\hat{x})$ denoted by $\mathcal{H}_i(\hat{x},\xi)$ that satisfies $\mathbb{E}_\xi[\mathcal{H}_i(\hat{x},\xi)] = f_i(\hat{x})$, where ξ is a random variable representing the noise.

Recently, centralized zeroth-order optimization has received significant attention. In [14], Nesterov proposed a general framework for analyzing zeroth-order algorithms and provided the global convergence rates for both convex and nonconvex problems. In [15], the authors established a stochastic zeroth-order method, which again can deal with both convex and nonconvex (but smooth) optimization problems. Both these zeroth-order algorithms are centralized and cannot be implemented over a multi-agent network. A few recent works [16]–[18] considered zeroth-order distributed convex (possibly nonsmooth) problems, but none of these works can address nonconvex problems.

In this paper, we propose a new algorithm for distributed nonconvex and nonsmooth optimization with zeroth-order information. Specifically, we first show that this problem can be reformulated as a linearly constrained optimization problem over a connected multi-agent network. Then, we propose a nonconvex primal-dual based algorithm, which requires only local communication among the agents, and utilizes local zeroth-order information. Theoretically, We show that the current solution converges approximately to a stationary solution of the problem. We also provide numerical results that corroborate the theoretical findings.

Notation. Given a vector a and a matrix A, we use $\|a\|$ and $\|A\|$ to denote the Euclidean norm of vector a, and spectral norm of matrix A, respectively. $A^\top$ represents the transpose of matrix A. We define $\|a\|_A^2 := a^\top Aa$. The notation $\langle a, b\rangle$ is used to denote the inner product of two vectors a, b. For matrices A and B, $A\otimes B$ is the Kronecker product of A and B. To denote an $M \times M$ identity matrix we use I_M.

II. Problem Definition and Proposed Algorithm

Let us define a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the node set with $|\mathcal{V}| = N$, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set with $|\mathcal{E}| = E$. We assume that $\mathcal{G}$ is undirected, meaning that if $(i, j) \in \mathcal{E}$ then $(j, i) \in \mathcal{E}$. Moreover, every agent i can only communicate with its direct neighbors in the set $\mathcal{N}_i = \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$, and let $d_i = |\mathcal{N}_i|$ denote the degree of node i. We assume that the graph $\mathcal{G}$ is connected, meaning that

Davood Hajinezhad and Michael M. Zavlanos are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA {davood.hajinezhad, michael.zavlanos}@duke.edu

there is a path, i.e., a sequence of nodes where consecutive nodes are neighbors, between any two nodes in $\mathcal{G}$.

In order to decompose problem (1) let us introduce N new variables $x_i \in \mathbb{R}^M$ that are local to every agent i. Then, problem (1) can be reformulated as follows:

$$\min_{\{x_i\}} \quad \sum_{i=1}^{N} f_i(x_i), \quad \text{s.t.} \quad x_i = x_j \ \forall \ (i,j) \in \mathcal{E}. \tag{2}$$

The set of constraints $x_i = x_j$ enforce consensus on the local variables x_i and x_j for all neighbors $j \in \mathcal{N}_i$. We stack all the local variables x_i in a vector $x := \{x_i\} \in \mathbb{R}^{Q\times 1}$, where $Q = NM$. Moreover, we define the Degree matrix $\tilde{D} \in \mathbb{R}^{N\times N}$ to be a diagonal matrix where $\tilde{D}(i,i) = d_i$; let $D = \tilde{D} \otimes I_M \in \mathbb{R}^{Q\times Q}$. For a given graph $\mathcal{G}$, the incidence matrix $\tilde{A} \in \mathbb{R}^{E\times N}$ is a matrix where $\tilde{A}(k,i) = 1$ and $\tilde{A}(k,j) = -1$, where (i,j) is the kth edge of $\mathcal{G}$; the rest of the entries of $\tilde{A}$ are all zero. Let $A = \tilde{A} \otimes I_M \in \mathbb{R}^{EM\times Q}$. Finally, we define the Signed Laplacian matrix as $L^- := A^\top A \in \mathbb{R}^{Q\times Q}$ and Signless Laplacian matrix as $L^+ := 2D - A^\top A \in \mathbb{R}^{Q\times Q}$. Using these notations, problem (2) can be written in the following compact form:

$$\min_{x\in\mathbb{R}^Q} \ f(x), \quad \text{s.t.} \ \ Ax = 0. \tag{3}$$

A. Preliminaries

In this section, we first introduce some standard techniques presented in [14] for approximating and smoothing the gradient of a given function. Suppose that $\phi \in \mathbb{R}^Q$ is a Gaussian random vector and let $\mu > 0$ be some smoothing parameter. The smoothed version of function f is defined as

$$f_\mu(x) = \frac{1}{(2\pi)^{\frac{Q}{2}}} \int f(x+\mu\phi) e^{-\frac{1}{2}\|\phi\|^2} \, d\phi. \tag{4}$$

Then it can be shown that the function f_μ is differentiable and its gradient is given by [14, Eq. (22)]

$$\nabla f_\mu(x) = \frac{1}{(2\pi)^{\frac{Q}{2}}} \int \frac{f(x+\mu\phi) - f(x)}{\mu} \phi e^{-\frac{1}{2}\|\phi\|^2} \, d\phi. \tag{5}$$

Further, assuming that the original function f is Lipschitz continuous, i.e., there exists L_0 such that $|f(x) - f(y)| \leq L_0|x-y|$ for all $x, y \in \text{dom}\,(f)$, it can be shown that (see [14, Lemma 2]) ∇f_μ is also Lipschitz continuous with constant $L_1 = \frac{2L_0\sqrt{Q}}{\mu}$. In other words, for all $x, y \in \text{dom}\,(f_\mu)$ we have

$$\|\nabla f_\mu(x) - \nabla f_\mu(y)\| \leq L_1 \|x - y\|. \tag{6}$$

Let $\mathcal{H}(x,\xi)$ denote the noisy functional value of the function f obtained from an associated $\mathcal{SZO}$. In view of (5), the gradient of $f_\mu(x)$ can be approximated as

$$G_\mu(x,\phi,\xi) = \frac{\mathcal{H}(x+\mu\phi,\xi) - \mathcal{H}(x,\xi)}{\mu}\phi, \tag{7}$$

where the constant $\mu > 0$ is the smoothing parameter. It can be easily checked that $G_\mu(x,\phi,\xi)$ is an unbiased estimator of $\nabla f_\mu(x)$. For simplicity we define $\zeta := (\xi, \phi)$. For a given number J of independent samples of $\{\zeta_j\}_{j=1}^J$, we define the sample average $\bar{G}_\mu(x,\zeta) := \frac{1}{J}\sum_{j=1}^J G_\mu(x,\zeta_j)$, where $\zeta := \{\zeta_j\}_{j=1}^J$. It is easy to see that for any $J \geq 1$, $\bar{G}_\mu(x,\zeta)$ is also an unbiased estimator of $\nabla f_\mu(x)$.

B. The Proposed Algorithm

In this part we propose a primal-dual algorithm for the distributed optimization problem (3). Let $\lambda_{ij} \in \mathbb{R}^M$ be the multiplier associated with the consensus constraint $x_i - x_j = 0$ for each $(i,j) \in \mathcal{E}$. Moreover, stack all λ_{ij}'s in a vector $\lambda = \{\lambda_{ij}\}_{(i,j)\in\mathcal{E}} \in \mathbb{R}^{EM}$. Then, the augmented Lagrangian (AL) function for problem (3) is given by

$$U_\rho(x,\lambda) := f(x) + \langle \lambda, Ax\rangle + \frac{\rho}{2}\|Ax\|_2^2, \tag{8}$$

where $\rho > 0$ is a constant. Moreover, as in (4) define the smoothed version $f_{i,\mu}$ of the local function f_i. At iteration r of the algorithm we obtain an unbiased estimation of the gradient of local function $f_{i,\mu}(x_i^r)$ as follows. For every sample $j \in \{1, 2, \cdots, J\}$ we generate a random vector $\phi_{i,j}^r \in \mathbb{R}^M$ from an i.i.d standard Gaussian distribution and calculate $\bar{G}_{\mu,i}(x_i^r, \zeta_i^r) \in \mathbb{R}^M$ similar to (7) by

$$\begin{aligned}&\bar{G}_{\mu,i}(x_i^r, \zeta_i^r)\\ &= \frac{1}{J}\sum_{j=1}^J \frac{\mathcal{H}_i(x_i^r + \mu\phi_{i,j}^r, \xi_{i,j}^r) - \mathcal{H}_i(x_i^r, \xi_{i,j}^r)}{\mu}\phi_{i,j}^r.\end{aligned} \tag{9}$$

Define $G_\mu^{J,r} := \{\bar{G}_{\mu,i}(z_i^r, \zeta_i^r)\}_{i=1}^N \in \mathbb{R}^Q$. The following theorem bounds the norm of $G_\mu^{J,r}$.

Theorem 1 (Theorem 4 [14]) *If f is a Lipschitz continuous function with constant L_0, then*

$$\mathbb{E}_\zeta[\|G_\mu^{J,r}\|^2] \leq \frac{L_0^2(Q+4)^2}{J^2}. \tag{10}$$

Our proposed algorithm is summarized in Algorithm 1. In

Algorithm 1 The Proposed Algorithm for problem (1)

1: **Input:** $D \in \mathbb{R}^{Q\times Q}$, $T \geq 1$, $J \geq 1$, and $\mu > 0$
2: **Initialize:** $x^0 \in \mathbb{R}^Q$ and $\lambda^0 = 0 \in \mathbb{R}^{EM}$
3: **for** $r = 0$ **to** $T-1$ **do**

$$\begin{aligned}x^{r+1} = \underset{x}{\text{argmin}} \ &\langle G_\mu^{J,r} + A^\top\lambda^r + \rho A^\top A x^r, x - x^r\rangle\\ &+ \rho\|x - x^r\|_D^2, \ \texttt{(Primal Step)}\end{aligned} \tag{11}$$

$$\lambda^{r+1} = \lambda^r + \rho A x^{r+1}. \qquad \texttt{(Dual Step)} \tag{12}$$

4: **end for**
5: Choose uniformly randomly $u \in \{0, 1, \cdots, T-1\}$
6: **Output:** (z^u, λ^u).

the primal step (11), an *approximate gradient descent* step is taken towards minimizing the AL function with respect to x. In particular, in the first-order approximation of the AL function the true gradient of the function $f(x^r)$ is approximated by the noisy zeroth-order estimate $G_\mu^{J,r}$ and then, a matrix-weighted quadratic penalty $\rho\|x - x^r\|_D$ is used. The dual step (12) is then performed, which is a *gradient ascent* step over the dual variable λ.

To see how Algorithm 1 can be implemented in a distributed way, consider the optimality condition for (11) as

$$G_\mu^{J,r} + A^\top(\lambda^r + \rho A x^r) + 2\rho D(x^{r+1} - x^r) = 0. \tag{13}$$

Using the Signed and Signless Laplacian matrices, we have

$$x^{r+1} = \frac{1}{2\rho} D^{-1}\Big[\rho L^+ x^r - G_\mu^{J,r} + A^\top \lambda^r\Big].$$

To implement this primal iteration, each agent i only requires local information as well as information from its neighbors $\mathcal{N}_i$. This is because D is a diagonal matrix and the structure of the matrix L^+ ensures that the ith block vector of L^+x^r is only related to $x_j^r,\ j \in \mathcal{N}_i$. For the dual step w.l.o.g we assign the dual variable λ_{ij} to node i and therefore, from (12) we have

$$\lambda_{ij}^{r+1} = \lambda_{ij}^r + \rho(x_i^{r+1} - x_j^{r+1}), \tag{14}$$

which only requires the local information as well as information from the neighbors in $\mathcal{N}_i$.

III. The Convergence Analysis

In this section we study the convergence of Algorithm 1. We make the following assumptions.

Assumptions A. We assume that

A1. The function f is Lipschitz continuous.

A2. The function f is lower bounded.

To simplify notation let $\mathcal{F}^r := \sigma(\zeta_1, \zeta_2, \cdots \zeta_r)$ be the σ-field generated by the entire history of Algorithm 1 up to iteration r, $\sigma_{\min}$ be the smallest nonzero eigenvalue of $A^\top A$, and $w^r := (x^{r+1} - x^r) - (x^r - x^{r-1})$ be the successive difference of the differences of the primal iterates. In the analysis that follows we will make use of the following relations:

- For any given vectors a and b we have

$$\langle b - a, b\rangle = \frac{1}{2}(\|b\|^2 + \|a-b\|^2 - \|a\|^2), \tag{15}$$

$$\langle a, b\rangle \le \frac{1}{2\epsilon}\|a\|^2 + \frac{\epsilon}{2}\|b\|^2; \quad \forall\ \epsilon > 0. \tag{16}$$

- For n given vectors a_i we have that

$$\Big\|\sum_{i=1}^n a_i\Big\|^2 \le n\sum_{i=1}^n \big\|a_i\big\|^2. \tag{17}$$

First we bound the difference between $\nabla f_\mu(x^r)$ and its unbiased estimation $G_\mu^{J,r}$ as follows:

$$\mathbb{E}_\zeta\|G_\mu^{J,r} - \nabla f_\mu(x^r)\|^2 = \mathbb{E}_\zeta\|G_\mu^{J,r} - \mathbb{E}_\zeta[G_\mu^{J,r}]\|^2 \le \mathbb{E}_\zeta\|G_\mu^{J,r}\|^2 \le \frac{L_0^2(Q+4)^2}{J^2}, \tag{18}$$

where the last inequality follows from (10). The next lemma bounds the change of the dual variables by that of the primal variables. The proof of the results the follows can be found in the appendix.

Lemma 1 *Suppose Assumptions A hold true. Let L_1 denote the gradient Lipschitz constant for function f_μ. Then we have the following inequality:*

$$\frac{1}{\rho}\mathbb{E}\|\lambda^{r+1} - \lambda^r\|^2 \le \frac{9L_0^2(Q+4)^2}{\rho\sigma_{\min}J^2} + \frac{6L_1^2}{\rho\sigma_{\min}}\mathbb{E}\|x^r - x^{r-1}\|^2 + \frac{3\rho\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2. \tag{19}$$

The next step is the key in our analysis. We define the smoothed version of the AL function in a similar way as (4) and denote it by $U_{\rho,\mu}(x,\lambda)$. For notational simplicity let us define $U_{\rho,\mu}^{r+1} := U_{\rho,\mu}(x^{r+1}, \lambda^{r+1})$. From equation (6) we know that the function f_μ is Lipschitz continuous with constant L_1. Now let $c > 0$ be some positive constant and set $k := 2\big(\frac{6{L_1}^2}{\rho\sigma_{\min}} + \frac{3cL_1}{2}\big)$. Moreover, we define $V^{r+1} := \frac{\rho}{2}\big(\|Ax^{r+1}\|^2 + \|x^{r+1} - x^r\|_B^2\big)$, where $B := L^+ + \frac{k}{c\rho}I_Q$. Finally we define the following potential function:

$$P^{r+1} := U_{\rho,\mu}^{r+1} + cV^{r+1}. \tag{20}$$

We study the behavior of the proposed potential function as the algorithm proceeds.

Lemma 2 *Suppose Assumptions A hold true. We have that*

$$\mathbb{E}\big[P^{r+1} - P^r\big] \le -\alpha_1\mathbb{E}\|x^{r+1} - x^r\|^2 - \alpha_2\mathbb{E}\|w^r\|_{L^+}^2 + \alpha_3\frac{L_0^2(Q+4)^2}{J^2}, \tag{21}$$

where

$$\alpha_1 := \rho^2 - (2cL_1 + L_1/2 + L_1^2/2 + 1/2)\rho - \frac{6L_1^2}{\sigma_{\min}},$$
$$\alpha_2 := \frac{3\rho\|L^+\|}{\sigma_{\min}} - \frac{c\rho}{2},\ \alpha_3 = \frac{9}{\rho\sigma_{\min}} + \frac{6c+1}{L_1}. \tag{22}$$

Note that the constants α_1 and α_2 in (21) can be made positive as long as we choose constants c and ρ large enough. In particular, the following conditions are sufficient to have positive α_1, α_2

$$c > \frac{6\|L^+\|}{\sigma_{\min}}, \quad \rho > b + \sqrt{b^2 + 6L_1^2/\sigma_{\min}}, \tag{23}$$

where $b = (cL_1 + L_1/4 + L_1^2/4 + 1/4)$.

In the next lemma we show that P^{r+1} is lower bounded.

Lemma 3 *Suppose Assumptions A hold, and the constant c is selected as $c \ge \frac{2\|L^+\|}{\sigma_{\min}}$. Then there exists a constant $\underline{P}$ that is independent of the total number of iterations T so that*

$$\mathbb{E}[P^{r+1}] \ge \underline{P} > -\infty, \quad \forall\ r \ge 1. \tag{24}$$

To characterize the convergence rate of Algorithm 1, let us define the *stationarity gap* of the smoothed version of problem (3) as

$$\Phi_\mu(x^r, \lambda^{r-1}) := \mathbb{E}\left[\|\nabla_x U_{\rho,\mu}(x^r, \lambda^{r-1})\|^2 + \|Ax^r\|^2\right]. \tag{25}$$

It can be easily checked that $\Phi_\mu(x^*, \lambda^*) = 0$ if and only if (x^*, λ^*) is a KKT point of the smoothed version of problem (3). For simplicity let us denote $\Phi_\mu^r := \Phi_\mu(x^r, \lambda^{r-1})$.

At this point we are ready to combine the previous results to obtain our main theorem.

Theorem 2 *Suppose Assumptions A hold true, the penalty parameter ρ satisfies the condition given in Lemma 2, and the constant c satisfies $c \ge \frac{6\|L^+\|}{\sigma_{\min}}$. Then, there exist constants $\gamma_1, \gamma_2 > 0$ such that*

$$\mathbb{E}_u[\Phi_\mu^u] \le \frac{\gamma_1}{T} + \gamma_2\frac{L_0(Q+4)^2}{J^2}. \tag{26}$$

From Theorem 2 we can observe that there exists always a constant term in the right-hand-side of the stationarity gap. Therefore, no matter how many iterations we run the algorithm, we always converge to a neighborhood of a stationary point. However, if we choose the number of samples $J \in \mathcal{O}(\sqrt{T})$, we have the following bound:

$$\mathbb{E}_u[\Phi_\mu^u] \le \frac{\gamma_1}{T} + \frac{\gamma_2 L_0(Q+4)^2}{T}, \tag{27}$$

which verifies the sublinear convergence rate for the algorithm.

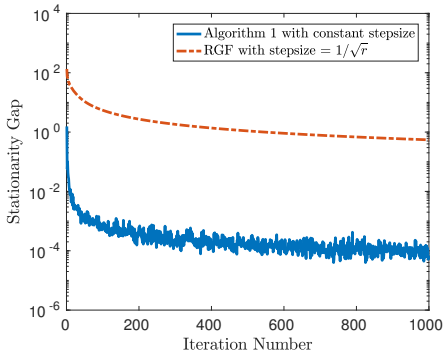

Fig. 1: Comparison of proposed Algorithm 1, and RGF algorithm [16] in terms of the stationarity gap for binary classification problem.

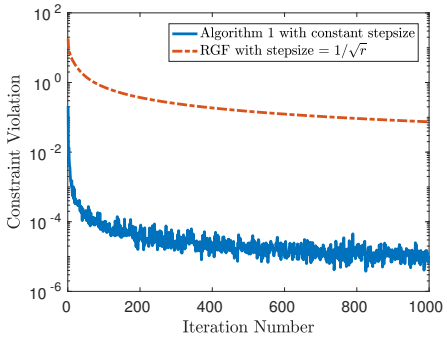

Fig. 2: Comparison of proposed Algorithm 1, and RGF algorithm [16] in terms of the constraint violation (i.e. $\|Ax\|$) for binary classification problem.

IV. Numerical Results

In this section we illustrate the proposed algorithm through numerical simulations. For our experiment we study a mini-batch binary classification problem using nonconvex nonsmooth regularizers, where each node stores b (batch size) data points. For this problem the local function is given by

$$f_i(x_i) = \frac{1}{Nb}\Bigg[\sum_{j=1}^{b}\log(1+\exp(-y_{ij}x_i^\top v_{ij})) + \alpha\log(\epsilon+\|x_i\|_1)\Bigg],$$

where $v_{ij}\in\mathbb{R}^M$ and $y_{ij}\in\{1,-1\}$ are the feature vector and the label for the jth data point of ith agent [19]. The nonconvex nonsmooth regularization term $\log(\epsilon+\|x_i\|_1)$ imposes sparsity to vector x_i, the constant α controls the sparsity level, and $\epsilon>0$ is a small number. The network has $N=15$ nodes and each node contains randomly generated $b=100$ data point. We compare Algorithm 1 using a constant stepsize that satisfies (23) and the Randomize Gradient Free (RGF) algorithm proposed in [16] using diminishing stepsize $\frac{1}{\sqrt{r}}$. Note that in theory RGF only works for the convex problems. However, we include it here for the purpose of comparison only. Algorithm 1 and RGF run for $T=1000$ iterations and Figures 1 and 2 illustrate the stationarity gap and the constraint violation versus the iteration counter. From these plots we can observe that Algorithm 1 is faster than the RGF. Note again that, in theory, RGF is designed for convex problems only. To the best of our knowledge, our algorithm is the first provable distributed zeroth-order method for nonconvex and nonsmooth problems.

V. Conclusion

In this work, we proposed a distributed gradient-free optimization algorithm to solve nonconvex and nonsmooth problems utilizing local zeroth-order information. We rigorously analyzed the convergence rate of the proposed algorithm and demonstrated its performance via simulation. To the best of our knowledge, this is the first distributed framework for the solution of nonconvex and nonsmooth distributed optimization problems that also has a provable sublinear convergence rate.

Appendix

Due to the space limitation we present only the concise proofs here. See the online version for the complete proofs.

A. Proof of Lemma 1

From equation (12) we have

$$\lambda^{r+1}-\lambda^r=\rho Ax^{r+1}. \tag{28}$$

Equation (28) implies that $\lambda^{r+1}-\lambda^r$ lies in the column space of A, therefore we have

$$\sqrt{\sigma_{\min}}\|\lambda^{r+1}-\lambda^r\|\le\|A^\top(\lambda^{r+1}-\lambda^r)\|, \tag{29}$$

where $\sigma_{\min}$ denotes the smallest non-zero eigenvalue of $A^\top A$. Utilizing equation (28) and equation (13), we obtain

$$G_\mu^{J,r}+A^\top\lambda^{r+1}+\rho L^+(x^{r+1}-x^r)=0. \tag{30}$$

Replacing r with $r-1$ in equation (30) and then using the definition of $w^r:=(x^{r+1}-x^r)-(x^r-x^{r-1})$ we obtain

$$\begin{aligned}\frac{1}{\rho}\left\|\lambda^{r+1}-\lambda^r\right\|^2&\le\frac{1}{\rho\sigma_{\min}}\left\|G_\mu^{J,r}-G_\mu^{J,r-1}+\rho L^+w^r\right\|^2\\&\le\frac{3}{\rho\sigma_{\min}}\|G_\mu^{J,r}-\nabla f_\mu(x^r)\|^2+\frac{3\rho^2}{\rho\sigma_{\min}}\|L^+w^r\|^2\\&+\frac{3}{\rho\sigma_{\min}}\|\nabla f_\mu(x^r)-G_\mu^{J,r-1}\|^2,\end{aligned} \tag{31}$$

where the last inequality follows from (17). Adding and subtracting $\nabla f_\mu(x^{r-1})$ to the second term on the r.h.s of (31) and taking the expectation on both sides gives

$$\begin{aligned}&\frac{1}{\rho}\mathbb{E}\|\lambda^{r+1}-\lambda^r\|^2\le\frac{3}{\rho\sigma_{\min}}\mathbb{E}\|G_\mu^{J,r}-\nabla f_\mu(x^r)\|^2\\&+\frac{6}{\rho\sigma_{\min}}\mathbb{E}\|\nabla f_\mu(x^r)-\nabla f_\mu(x^{r-1})\|^2+\frac{3\rho^2}{\rho\sigma_{\min}}\mathbb{E}\|L^+w^r\|^2\\&+\frac{6}{\rho\sigma_{\min}}\mathbb{E}\|\nabla f_\mu(x^{r-1})-G_\mu^{J,r-1}\|^2\\&\le\frac{9L_0^2(Q+4)^2}{\rho\sigma_{\min}J^2}+\frac{6L_1^2}{\rho\sigma_{\min}}\mathbb{E}\|x^r-x^{r-1}\|^2\\&+\frac{3\rho\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|^2_{L^+},\end{aligned}$$

where the last inequity is true because of (18), the fact that $\nabla f_\mu(z)$ is gradient Lipschitz with constant L_1, and the inequality $\|L^+w^r\|^2\le\|L^+\|\|w^r\|^2_{L^+}$. The proof is complete.

B. Proof of Lemma 2

First we prove that the function $g(x):=U_{\rho,\mu}(x,\lambda)+\frac{\rho}{2}\|x-x^r\|^2_{L^+}$ is strongly convex with respect to variable x when $2\rho\ge L_1$. From Assumption A.1 and the fact that $D\succeq I$ we have

$$\begin{aligned}
&\langle \nabla g(x_1) - \nabla g(x_2), x_1 - x_2\rangle \\
&= \langle \nabla f_\mu(x_1) - \nabla f_\mu(x_2), x_1 - x_2\rangle + 2\rho\|x_1 - x_2\|_D^2 \\
&\geq \langle \nabla f_\mu(x_1) - \nabla f_\mu(x_2), x_1 - x_2\rangle + 2\rho\|x_1 - x_2\|^2 \\
&\geq (2\rho - L_1)\|x_1 - x_2\|^2.
\end{aligned}$$

This proves that $U_{\rho,\mu}(x,\lambda)+\frac{\rho}{2}\|x-x^r\|_{L^+}^2$ is strongly convex with modulus $2\rho-L_1$. Using this fact, we can bound $U_{\rho,\mu}^{r+1}-U_{\rho,\mu}^r$ as follows:

$$\begin{aligned}
&U_{\rho,\mu}^{r+1} - U_{\rho,\mu}^r = U_{\rho,\mu}(x^{r+1},\lambda^{r+1}) - U_{\rho,\mu}(x^{r+1},\lambda^r) \\
&+ U_{\rho,\mu}(x^{r+1},\lambda^r) - U_{\rho,\mu}(x^r,\lambda^r) \\
&\leq \frac{1}{\rho}\|\lambda^{r+1} - \lambda^r\|^2 \\
&+ \langle \nabla_x U_{\rho,\mu}(x^{r+1},\lambda^r) + \rho L^+(x^{r+1} - x^r), x^{r+1} - x^r\rangle \\
&- \frac{2\rho - L_1}{2}\|x^{r+1} - x^r\|^2, \qquad (32)
\end{aligned}$$

where the last inequity holds true due to the strong convexity of $U_{\rho,\mu}(x,\lambda)+\frac{\rho}{2}\|x-x^r\|_{L^+}^2$ and (28). Using (30) we have

$$\begin{aligned}
U_{\rho,\mu}^{r+1} - U_{\rho,\mu}^r &\leq \frac{L_1^2 - 2\rho + L_1}{2}\|x^{r+1} - x^r\|^2 \\
&+ \frac{1}{\rho}\|\lambda^{r+1} - \lambda^r\|^2 + \frac{1}{2L_1^2}\|\nabla f_\mu(x^{r+1}) - G_\mu^{J,r}\|^2,
\end{aligned}$$

where the inequality follows from (16) with $\epsilon = L_1^2$. Taking expectation on both sides, and utilizing (19) and (18) we have

$$\begin{aligned}
&\mathbb{E}\big[U_{\rho,\mu}^{r+1} - U_{\rho,\mu}^r\big] \leq \big(\frac{9}{\rho\sigma_{\min}} + \frac{1}{L_1^2}\big)\frac{L_0^2(Q+4)^2}{J^2} \\
&+ \frac{6L_1^2}{\rho\sigma_{\min}}\mathbb{E}\|x^r - x^{r-1}\|^2 + \frac{3\rho\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2 \\
&+ \frac{L_1^2 - 2\rho + L_1 + 1}{2}\mathbb{E}\|x^{r+1} - x^r\|^2. \qquad (33)
\end{aligned}$$

Now we bound $V^{r+1} - V^r$. From the optimality condition for problem (11) and the dual update (12) we have for all $x \in \mathbb{R}^Q$

$$\langle G_\mu^{J,r} + A^\top\lambda^{r+1} + \rho L^+(x^{r+1} - x^r), x^{r+1} - x\rangle \leq 0.$$

Similarly, for the $(r-1)$th iteration, we have

$$\langle G_\mu^{J,r-1} + A^\top\lambda^r + \rho L^+(x^r - x^{r-1}), x^r - x\rangle \leq 0.$$

Setting $x = x^r$ in first equation, $x = x^{r+1}$ in second equation, and adding them, we obtain

$$\begin{aligned}
&\langle A^\top(\lambda^{r+1} - \lambda^r), x^{r+1} - x^r\rangle \leq \\
&- \langle G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r, x^{r+1} - x^r\rangle. \qquad (34)
\end{aligned}$$

The l.h.s can be expressed as follows:

$$\begin{aligned}
&\langle A^\top(\lambda^{r+1} - \lambda^r), x^{r+1} - x^r\rangle = \rho\langle Ax^{r+1}, Ax^{r+1} - Ax^r\rangle \\
&= \frac{\rho}{2}\bigg(\|Ax^{r+1}\|^2 - \|Ax^r\|^2 + \|A(x^{r+1} - x^r)\|^2\bigg), \qquad (35)
\end{aligned}$$

where the first equality follows from (12) and the second equality follows from (15). For the r.h.s of (34) using (16) and (17) we have

$$\begin{aligned}
&- \langle G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r, x^{r+1} - x^r\rangle \\
&\leq \frac{3}{2L_1}\bigg(\|G_\mu^{J,r} - \nabla f_\mu(x^r)\|^2 + \|\nabla f_\mu(x^{r-1}) - G_\mu^{J,r-1}\|^2 \\
&+ \|\nabla g_\mu(x^r) - \nabla g_\mu(x^{r-1})\|^2\bigg) + \frac{L_1}{2}\|x^{r+1} - x^r\|^2 \\
&- \rho\langle L^+ w^r, x^{r+1} - x^r\rangle.
\end{aligned}$$

Taking expectation on both sides and utilizing (18) and (15) we have

$$\begin{aligned}
&- \mathbb{E}[\langle G_\mu^{J,r} - G_\mu^{J,r-1} + \rho L^+ w^r, x^{r+1} - x^r\rangle] \\
&\leq \frac{6L_0^2(Q+4)^2}{L_1 J^2} + \frac{3L_1}{2}\mathbb{E}\|x^r - x^{r-1}\|^2 \\
&+ \frac{\rho}{2}\mathbb{E}\bigg[\|x^r - x^{r-1}\|_{L^+}^2 - \|x^{r+1} - x^r\|_{L^+}^2 - \|w^r\|_{L^+}^2\bigg] \\
&+ \frac{L_1}{2}\mathbb{E}\|x^{r+1} - x^r\|^2. \qquad (36)
\end{aligned}$$

Combining (35) and (36), we obtain

$$\begin{aligned}
&\frac{\rho}{2}\mathbb{E}\bigg(\|Ax^{r+1}\|^2 - \|Ax^r\|^2 + \|A(x^{r+1} - x^r)\|^2\bigg) \\
&\leq \frac{6L_0^2(Q+4)^2}{L_1 J^2} + \frac{3L_1}{2}\mathbb{E}\|x^r - x^{r-1}\|^2 \\
&+ \frac{\rho}{2}\mathbb{E}\bigg(\|x^r - x^{r-1}\|_{L^+}^2 - \|x^{r+1} - x^r\|_{L^+}^2 - \|w^r\|_{L^+}^2\bigg) \\
&+ \frac{L_1}{2}\mathbb{E}\|x^{r+1} - x^r\|^2. \qquad (37)
\end{aligned}$$

The rest of the proof is only rearranging terms in (37), and using the definition of V^r and B.

C. Proof of Lemma 3

From (30) we have

$$\|\lambda^{r+1}\|^2 \leq \frac{2}{\sigma_{\min}}\|G_\mu^{J,r}\|^2 + \frac{2\rho^2}{\sigma_{\min}}\|L^+(x^{r+1} - x^r)\|^2. \qquad (38)$$

From the definition of the potential function we have

$$\begin{aligned}
P^{r+1} &= f_\mu(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} + \frac{1}{\rho}\lambda^{r+1}\|^2 - \frac{1}{2\rho}\|\lambda^{r+1}\|^2 \\
&+ \frac{c\rho}{2}\|Ax^{r+1}\|^2 + \frac{c\rho}{2}\|x^{r+1} - x^r\|_B^2, \qquad (39)
\end{aligned}$$

where $B := L^+ + \frac{k}{c\rho}I$. From Assumption A.2 and utilizing the fact that $\|Ax^{r+1} + \frac{1}{\rho}\lambda^{r+1}\|^2 \geq 0$ we obtain

$$\begin{aligned}
P^{r+1} &\geq \frac{-1}{\rho\sigma_{\min}}\|G_\mu^{J,r}\|^2 - \frac{\rho}{\sigma_{\min}}\|L^+(x^{r+1} - x^r)\|^2 \\
&+ \frac{c\rho}{2}\|x^{r+1} - x^r\|_B^2 + \underline{f} \\
&\geq \frac{-1}{\rho\sigma_{\min}}\|G_\mu^{J,r}\|^2 + \frac{\rho}{\sigma_{\min}}\|x^{r+1} - x^r\|^2_{\frac{c\sigma_{\min}}{2}L^+ - (L^+)^2} + \underline{f}.
\end{aligned}$$

Notice that L^+ is a symmetric PSD matrix. Therefore, picking constant c large enough such that $c \geq \frac{2\|L^+\|}{\sigma_{\min}}$, we have $\frac{c\sigma_{\min}}{2}L^+ - (L^+)^2 \succeq 0$. The rest of the proof is quite straightforward thus omitted.

D. Proof of Theorem 2

First we bound the stationarity gap given in (25). Utilizing equations (12), (30) and (17) we have

$$\begin{aligned}
&\|\nabla_x U_{\rho,\mu}(x^{r+1},\lambda^r)\|^2 \leq 4\|\nabla f_\mu(x^{r+1}) - \nabla f_\mu(x^r)\| \\
&+ 4\|\nabla f_\mu(x^r) - G_\mu^{J,r}\|^2 + 2\rho^2\|L^+(x^{r+1} - x^r)\|^2.
\end{aligned}$$

Taking expectation on both sides gives

$$\mathbb{E}\|\nabla_x U_{\rho,\mu}(x^{r+1},\lambda^r)\|^2 \le 4L_1^2\mathbb{E}\|x^{r+1}-x^r\| + \frac{4L_0^2(Q+4)^2}{J^2} + 2\rho^2\mathbb{E}\|L^+(x^{r+1}-x^r)\|^2, \quad (40)$$

where the inequality follows from (18). Next, we bound the expected value of the constraint violation. Utilizing the equation (12) we have $\|Ax^{r+1}\|^2 = \frac{1}{\rho^2}\|\lambda^{r+1}-\lambda^r\|^2$. Taking expectation and utilizing (19) we reach

$$\mathbb{E}\|Ax^{r+1}\|^2 = \frac{1}{\rho^2}\mathbb{E}\|\lambda^{r+1}-\lambda^r\|^2 \le \frac{9L_0^2(Q+4)^2}{\rho^2\sigma_{\min}J^2} + \frac{6L_1^2}{\rho^2\sigma_{\min}}\mathbb{E}\|x^r-x^{r-1}\|^2 + \frac{3\|L^+\|}{\sigma_{\min}}\mathbb{E}\|w^r\|_{L^+}^2. \quad (41)$$

Summing (40) and (41), we have the following bound for the stationarity gap

$$\Phi_\mu^{r+1} \le \beta_1\mathbb{E}\|x^{r+1}-x^r\|^2 + \beta_2\mathbb{E}\|x^r-x^{r-1}\|^2 + \beta_3\mathbb{E}\|w^r\|_{L^+}^2 + \beta_4\frac{L_0^2(Q+4)^2}{J^2}, \quad (42)$$

where $\beta_1,\beta_2,\beta_3,\beta_4$ are positive constants given by

$$\beta_1 = 4{L_1}^2 + 2\rho^2\|L^+\|^2,\ \beta_2 = \frac{6{L_1}^2}{\rho^2\sigma_{min}},$$
$$\beta_3 = \frac{3\|L^+\|}{\sigma_{min}},\ \beta_4 = \frac{9+4\rho^2\sigma_{\min}}{\rho\sigma_{\min}}.$$

Summing both sides of (42) over T iterations, we get

$$\sum_{r=1}^{T}\Phi_\mu^{r+1} \le \sum_{r=1}^{T-1}(\beta_1+\beta_2)\mathbb{E}\|x^{r+1}-x^r\|^2 + \sum_{r=1}^{T}\beta_3\mathbb{E}\|w^r\|_{L^+}^2 + \beta_2\mathbb{E}\|x^1-x^0\|^2 + \beta_1\mathbb{E}\|x^{T+1}-x^T\|^2 + T\beta_4\frac{L_0^2(Q+4)^2}{J^2}. \quad (43)$$

Applying Lemma 2 and summing both sides of (21) over T iterations, we obtain

$$\mathbb{E}\big[P^1-P^{T+1}\big] \ge \sum_{r=1}^{T-1}\alpha_1\mathbb{E}\|x^{r+1}-x^r\|^2 + \sum_{r=1}^{T}\alpha_2\mathbb{E}\|w^r\|_{L^+}^2 + \alpha_1\mathbb{E}\|x^{T+1}-x^\top\|^2 - T\alpha_3\frac{L_0^2(Q+4)^2}{J^2}. \quad (44)$$

Let us set $\tau = \frac{\max(\beta_1+\beta_2,\beta_3)}{\min(\alpha_1,\alpha_2)}$. Combining the two inequalities (43) and (44) and utilizing the fact that $\mathbb{E}[P^{T+1}]$ is lower bounded by $\underline{P}$, we arrive at the following inequality

$$\sum_{r=1}^{T}\Phi_\mu^{r+1} \le \tau\mathbb{E}[P^1-\underline{P}] + \beta_2\mathbb{E}\|x^1-x^0\|^2 + T\big(\tau\alpha_3+\beta_4\big)\frac{L_0^2(Q+4)^2}{J^2}. \quad (45)$$

Because u is a uniform random number from $\{0,1,\cdots,T-1\}$ we have

$$\mathbb{E}_u[\Phi_\mu^u] = \frac{1}{T}\sum_{r=1}^{T}\Phi_\mu^{r+1}. \quad (46)$$

Combining (45) and (46) implies the following

$$\mathbb{E}_u[\Phi_\mu^u] \le \frac{\tau\mathbb{E}[P^1-\underline{P}]+\beta_2\mathbb{E}\|x^1-x^0\|^2}{T} + \big(\tau\alpha_3+\beta_4\big)\frac{L_0^2(Q+4)^2}{J^2}.$$

The proof is complete.

REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[2] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[3] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *Automatic Control, IEEE Transactions on*, vol. 54, no. 11, pp. 2506–2517, 2009.

[4] S. Lee and M. M. Zavlanos, "Approximate projections for decentralized optimization with sdp constraints," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec 2016, pp. 1030–1035.

[5] M. Gurbuzbalaban, A. Ozdaglar, and P. Parrilo, "On the convergence rate of incremental aggregated gradient algorithms," *arXiv preprint arXiv:1506.02081*, 2015.

[6] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2014.

[7] Nikolaos Chatzipanagiotis, Darinka Dentcheva, and Michael M Zavlanos, "An augmented lagrangian method for distributed optimization," *Mathematical Programming*, vol. 152, no. 1-2, pp. 405–434, 2015.

[8] S. Lee, N. Chatzipanagiotis, and M. M. Zavlanos, "Complexity certification of a distributed augmented lagrangian method," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 827–834, March 2018.

[9] D. Hajinezhad, M. Hong, T. Zhao, and Z. Wang, "NESTT: A nonconvex primal-dual splitting method for distributed and stochastic optimization," in *Advances in Neural Information Processing Systems 29*, pp. 3215–3223. 2016.

[10] D. Hajinezhad and M. Hong, "Nonconvex alternating direction method of multipliers for distributed sparse principal component analysis," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015.

[11] M. Hong, D. Hajinezhad, and M-M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, vol. 70, pp. 1529–1538.

[12] D. Hajinezhad and Q. Shi, "Alternating direction method of multipliers for a class of nonconvex bilinear optimization: convergence analysis and applications," *Journal of Global Optimization*, pp. 1–28, 2018.

[13] N. Chatzipanagiotis and M. M. Zavlanos, "On the convergence of a distributed augmented lagrangian method for nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4405–4420, Sept 2017.

[14] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, pp. 1–40, 2011.

[15] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[16] D. Yuan and D.W.C Ho, "Randomized gradient-free method for multiagent optimization over time-varying networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1342–1347, 2015.

[17] D. Yuan, S. Xu, and J. Lu, "Gradient-free method for distributed multi-agent optimization via push-sum algorithms," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 10, pp. 1569–1580, 2015.

[18] Deming Yuan, Daniel WC Ho, and Shengyuan Xu, "Zeroth-order method for distributed optimization with approximate projections," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 2, pp. 284–294, 2016.

[19] A. Antoniadis, I. Gijbels, and M. Nikolova, "Penalized likelihood regression for generalized linear models with non-quadratic penalties," *Annals of the Institute of Statistical Mathematics*, vol. 63, no. 3, pp. 585–615, 2009.