Research Article

Approximate Dual Averaging Method for Multiagent Saddle-Point Problems with Stochastic Subgradients

Deming Yuan and Yang Yang

College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Correspondence should be addressed to Deming Yuan; dmyuan1012@gmail.com

This paper considers the problem of solving the saddle-point problem over a network, which consists of multiple interacting agents. The global objective function of the problem is a combination of local convex-concave functions, each of which is only available to one agent. Our main focus is on the case where the projection steps are calculated approximately and the subgradients are corrupted by some stochastic noises. We propose an approximate version of the standard dual averaging method and show that the standard convergence rate is preserved, provided that the projection errors decrease at some appropriate rate and the noises are zero-mean and have bounded variance.

1. Introduction

The problem of solving optimization problems over a multi-agent network has attracted a lot of attention in recent years (see, e.g., [1–13]). The objective function of such problems is, in general, a sum of local objective functions, each of which is known to one specific agent only. Moreover, the estimates of all agents are restricted to lie in some convex set. Duo to the lack of a central coordinator, the methods that are developed to solve this problem have to be executed by individual agents through local interactions.

In this paper, we consider the multiagent saddle-point problem where the global objective function is given as a sum of local convex-concave functions, subject to some global constraint. We utilize the average consensus algorithm (see, e.g., [14–21]) as a mechanism to design a distributed method for solving this problem. The method is based on the standard dual averaging method (see, e.g., [1, 22]), and it can also be viewed as an approximate version of the distributed dual averaging method in [2]. Different from the distributed dual averaging methods in [1–4], which require that the projection steps have to be very accurately calculated, the proposed method assumes that they only have to be computed approximately. Moreover, the proposed method

also considers the case where the subgradients are corrupted by some stochastic noises.

Literature Review. In [9], the authors develop a general framework for solving convex optimization problem over a network of multiple agents. Based on the average consensus algorithms, they propose a subgradient-based method; the method is fully distributed, in the sense that each agent only needs to communicate with its neighbors. Different from the work [9], the authors in [1] propose a distributed method that is based on dual averaging of subgradients; in particular, the authors characterize the explicit convergence rate of the proposed method. The authors in [3] further study the effects of communication delays on the distributed dual averaging method. The work [4] utilizes the push-sum algorithm as a mechanism to design a distributed dual averaging method; the implementation of the method removes the need for the doubly stochastic communication matrices. In [2], the authors solve the saddle-point problem over a multiagent network; the objective function is given as a sum of multiple convex-concave functions. Based on the dual averaging method, the authors propose a distributed method and characterize its convergence rate.

The contribution of our work in this paper is mainly twofold. First, we propose an approximate dual averaging

method, and the implementation of the method does not need to calculate the projection steps accurately. We show how the projection errors affect the error bound of the method and conclude that the standard convergence rate is preserved when the errors decrease at some appropriate rate. Second, we further consider the case where the subgradients are corrupted by stochastic noises that are zero-mean and have bounded variance, and we also highlight the dependence of the error bound on the variance.

In contrast with the work [22], we solve the saddle-point problem over a multiagent network; in particular, we show that the standard convergence rate $O(1/\sqrt{T})$ (where T is the iteration counter) is preserved, even when the projection steps are computed approximately and the subgradients are corrupted by some stochastic noises. In contrast with the work [2], we propose an approximate version of the distributed dual averaging method and show that if the projection errors decrease at some appropriate rate, the standard convergence rate is preserved.

The remainder of this paper is organized as follows. Section 2 gives a formal statement of the multiagent saddle-point problem and the underlying network model. Section 3 presents the method and its main convergence results. Finally, we conclude with Section 4.

Notation and Terminology. We use $\mathbb{R}^d$ to denote the d-dimensional vector space. We denote the standard inner product on $\mathbb{R}^d$ by $\langle x, y \rangle = \sum_{i=1}^{d} x_i y_i$, for all $x, y \in \mathbb{R}^d$. Let $\mathcal{M}$ be a closed convex set in $\mathbb{R}^d$. We say $h(x)$ is a *proximal function* of the set $\mathcal{M}$ if it is continuous and strongly convex on $\mathcal{M}$ with respect to some norm $\| \cdot \|$; that is, for all $0 \leq \theta \leq 1$, $h(\theta x_1 + (1 - \theta)x_2) \leq \theta h(x_1) + (1 - \theta)h(x_2) - (\sigma/2)\theta(1 - \theta)\|x_1 - x_2\|^2$, for all $x_1, x_2 \in \mathcal{M}$, where σ is some positive scalar. We define the *proximal center* of the set $\mathcal{M}$ by $x_0 = \arg\min_{x \in \mathcal{M}} h(x)$. For $\mathbb{R}^n \times \mathbb{R}^m$, we introduce the following norm: $\|x\| := [\gamma \sigma_w \|w\|_2^2 + (1 - \gamma)\sigma_z \|z\|_2^2]^{1/2}$, where $\gamma \in (0, 1)$, $\| \cdot \|_2$ denotes the Eculidean norm, and σ_w and σ_z are the parameters that will be specified in the sequel. This implies the following dual norm of $\tau = (\tau_w, \tau_z) \in \mathbb{R}^n \times \mathbb{R}^m$: $\|\tau\|_* = [(\|\tau_w\|_2^2/\gamma\sigma_w) + (\|\tau_z\|_2^2/(1 - \gamma)\sigma_z)]^{1/2}$. A vector $\mathbf{G}(y_0)$ is called a *subgradient* of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $y_0 \in \mathbb{R}^d$ if, for all y, $f(y) \geq f(y_0) + \langle y - y_0, \mathbf{G}(y_0) \rangle$. The supergradient of a concave function can be defined accordingly.

2. Problem Setup

2.1. Communication Network Model. We consider a time-varying network with N agents. The network can be viewed as a directed graph with node set $\mathcal{V} = \{1, \ldots, N\}$ and time-varying link set. The information exchange at time k is modeled through using the communication matrix $P(k) = [p_{ij}(k)]_{i,j \in \mathcal{V}} \in \mathbb{R}^{N \times N}$, which induces the link set $\mathcal{E}(P(k))$; $\mathcal{E}(P(k))$ is the set of activated links at time k, defined as $\mathcal{E}(P(k)) = \{(j, i) \mid p_{ij}(k) > 0, i, j \in \mathcal{V}\}$. We represent the agents' connectivity at each time k by a directed graph $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(P(k)))$.

2.2. Multiagent Saddle-Point Problem. In this paper, we are interested in solving the following problem:

$$\min_{w \in \mathcal{W}} \max_{z \in \mathcal{Z}} \sum_{i=1}^{N} \mathsf{L}^i(w, z), \tag{1}$$

where $\mathcal{W}$ and $\mathcal{Z}$ are convex and compact sets in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and each L^i is a convex-concave function defined over $\mathcal{W} \times \mathcal{Z}$ known only by agent i.

We refer to a vector pair $(w^*, z^*) \in \mathcal{W} \times \mathcal{Z}$ as a *saddle point* of $\mathsf{L} := \sum_{i=1}^{N} \mathsf{L}^i$ over $\mathcal{W} \times \mathcal{Z}$ if

$$\mathsf{L}(w^*, z) \leq \mathsf{L}(w^*, z^*) \leq \mathsf{L}(w, z^*), \quad \forall w \in \mathcal{W}, \ z \in \mathcal{Z}. \tag{2}$$

Note that such a vector pair (w^*, z^*) is a solution to problem (1).

We now make some assumptions on problem (1). For the set $\mathcal{W}$, we assume that there exists a proximal function $h_w(w)$ with proximal center and convex parameter denoted by w_0 and σ_w, respectively. Without loss of generality, we assume that $h_w(w_0) = 0$. For the set $\mathcal{Z}$ we introduce the similar assumptions and notations; that is, $h_z(z_0) = 0$. Therefore, for $x = (w, z) \in \mathcal{X} := \mathcal{W} \times \mathcal{Z}$, it is natural to introduce a proximal function $h(x)$ of the set $\mathcal{X}$, given by

$$h(x) = \gamma h_w(w) + (1 - \gamma) h_z(z), \quad \gamma \in (0, 1). \tag{3}$$

It is easy to see that the proximal center of $\mathcal{X}$ is $x_0 = (w_0, z_0)$ and $h(x_0) = 0$. Furthermore, we denote $H := \max_{x \in \mathcal{X}} h(x)$.

3. Main Results

3.1. The Method and Assumptions. We now propose the method, which is based on the method in [2]. Specifically, each agent $i \in \mathcal{V}$ updates its estimates by setting ($k = 0, 1, \ldots$):

$$\varphi_{k+1}^i = \sum_{j=1}^{N} p_{ij}(k) \varphi_k^j + \mathbf{G}_k^i + \phi_k^i, \tag{4}$$

$$x_{k+1}^i = \pi_{\alpha_{k+1}}^{\xi_{k+1}}\left(-\varphi_{k+1}^i\right), \tag{5}$$

where $\varphi_0^i = 0 \in \mathbb{R}^n \times \mathbb{R}^m$, $x_0^i = x_0$, $\mathbf{G}_k^i = (\mathbf{G}_w^i(x_k^i), -\mathbf{G}_z^i(x_k^i))$ ($\mathbf{G}_w^i(x_k^i)$ and $\mathbf{G}_z^i(x_k^i)$ denote a subgradient of L^i with respect to w and a supergradient of L^i with respect to z at point x_k^i, respectively), $\phi_k^i \in \mathbb{R}^n \times \mathbb{R}^m$ is the *stochastic noise vector* in evaluating $\mathbf{G}_k^i$, $\{\alpha_{k+1}\}$ is a positive and nondecreasing sequence, $\pi_\alpha(\varphi) := \arg\min_{x \in \mathcal{X}} \{-\langle x, \varphi \rangle + \alpha h(x)\}$, and $\pi_\alpha^\xi(\varphi)$ satisfies the following two properties:

$$\text{(i) } \pi_\alpha^\xi(\varphi) \in \mathcal{X}, \qquad \text{(ii) } \left\| \pi_\alpha^\xi(\varphi) - \pi_\alpha(\varphi) \right\| \leq \xi, \tag{6}$$

where ξ is a positive scalar that represents the error in computing the next iterate by a projection defined by the proximal function h and parameter α. Note that $\pi_\alpha^\xi(\varphi)$ is not uniquely defined for each ξ.

In the paper, we make the following assumptions.

Assumption 1 (connectivity). For all $k \geq 0$, there exists a positive integer B such that the directed graph $(\mathcal{V}, \mathcal{E}(P(kB)) \cup \cdots \cup \mathcal{E}(P((k+1)B - 1)))$ is strongly connected.

Assumption 2 (weight matrix). For all $k \geq 0$, the communication matrix $P(k)$ satisfies the following properties: (i) $P(k)$ is doubly stochastic and (ii) there exists a positive scalar ζ such that $p_{ii}(k) \geq \zeta$ for all $i \in \mathcal{V}$. In addition, if $p_{ij}(k) > 0$, then $p_{ij}(k) \geq \zeta$ for all $i, j \in \mathcal{V}$.

Assumption 3 (bounded subgradients). We assume that the following inequalities hold for all $i \in \mathcal{V}$ and $x \in \mathcal{X}$:

$$\left\| \mathbf{G}_w^i(x) \right\|_2 \leq G_{\mathcal{W}}, \qquad \left\| \mathbf{G}_z^i(x) \right\|_2 \leq G_{\mathcal{Z}}, \tag{7}$$

where $G_{\mathcal{W}}$ and $G_{\mathcal{Z}}$ are positive scalars.

Assumption 4 (stochastic subgradient). We assume that the stochastic noise vector ϕ_k^i satisfies the following properties, for all $i \in \mathcal{V}$ and $k \geq 0$:

$$\text{(i)} \ \mathbb{E}\left[\phi_k^i\right] = 0, \qquad \text{(ii)} \ \mathbb{E}\left[\left\|\phi_k^i\right\|_*^2\right] \leq \Phi^2, \tag{8}$$

where Φ is some positive constant.

3.2. Convergence Results. We show convergence of the method (4) and (5) via local average pair $(\widehat{w}_t^v, \widehat{z}_t^v)$ defined at each agent $v \in \mathcal{V}$,

$$\widehat{w}_T^v = \frac{1}{T}\sum_{k=0}^{T-1} w_k^v, \qquad \widehat{z}_T^v = \frac{1}{T}\sum_{k=0}^{T-1} z_k^v, \tag{9}$$

where $T \geq 1$ is the iteration counter.

With the assumptions made in Section 3.1, we have the following main convergence result.

Theorem 5. *Under Assumptions 1, 2, 3, and 4, consider a sequence $\{x_{k+1}^i\}$ generated according to the method (4) and (5), with step and projection error sizes:*

$$\alpha_0 = C_\alpha, \qquad \alpha_k = C_\alpha \sqrt{k}, \quad \text{for } k \geq 1,$$
$$\xi_0 = 0, \qquad \xi_k = \frac{C_\xi}{\sqrt{k}}, \quad \text{for } k \geq 1, \tag{10}$$

where C_α and C_ξ are some positive scalars. Let $(w^*, z^*) \in \mathcal{W} \times \mathcal{Z}$ be a saddle point of $\mathsf{L}(w, z)$, and then, for each agent $v \in \mathcal{V}$ and all $T \geq 1$, we have

$$\mathbb{E}\left[\left|\mathsf{L}\left(\widehat{w}_T^v, \widehat{z}_T^v\right) - \mathsf{L}\left(w^*, z^*\right)\right|\right]$$
$$\leq \frac{1}{\sqrt{T}}\left[2R_\xi C_\xi + NHC_\alpha + 2R_\alpha C_\alpha\right], \tag{11}$$

where $R_\xi = N(5G_{\mathcal{Z}} + \Phi)$, $R_\alpha = (N/\sigma)((1/2)(G_{\mathcal{Z}} + \Phi)^2 + (G_{\mathcal{Z}} + \Phi)R + 4G_{\mathcal{Z}}R)$, $G_{\mathcal{X}} = [(G_{\mathcal{W}}^2/\gamma\sigma_w) + (G_{\mathcal{Z}}^2/(1-\gamma)\sigma_z)]^{1/2}$, $R = ((NB/V(1-V)) + 2)(G_{\mathcal{X}} + \Phi)$, and $V = 1 - (\zeta/4N^2)$.

Proof. See The Appendix. ☐

Remark 6. Theorem 5 represents the main convergence of the method (4) and (5), which shows that the function value $\mathsf{L}(\widehat{w}_t^v, \widehat{z}_t^v)$ converges to $\mathsf{L}(w^*, z^*)$ at rate $O(1/\sqrt{T})$ in expectation, for each $v \in \mathcal{V}$. It is easy to see that the error bound is an increasing function of the noise magnitude Φ. It is worth noting that, in method (4) and (5), we have considered the case where the subgradients are corrupted by stochastic noises that are zero-mean and have bounded variance, and moreover, the projection steps are calculated only approximately. In fact, the proposed method converges when the projection error ξ_k decreases as $O(1/k^a)$, where $a > 0$. However, for the case when $0 < a < 1/2$, the $O(1/\sqrt{T})$ convergence rate cannot be achieved.

Remark 7. As compared to the work [2], we show that the standard $O(1/\sqrt{T})$ convergence rate for the dual averaging method is preserved, under the assumption that the projection steps are only computed approximately, and the subgradients are corrupted by some stochastic noises as well. As compared to [23], the proposed method solves the saddle-point problem in a distributed setting, and the expected convergence rate is also established.

4. Conclusion

We have studied the problem of solving saddle-point problems over a multiagent network. The objective function is given as a sum of local convex-concave functions, subject to some global constraint. Based on the average consensus algorithm and the dual averaging method, we propose an approximate dual averaging method under the constraint that the projection steps are computed approximately and the subgradients are corrupted by stochastic noises. Finally, we have presented the main convergence results of the proposed method.

Appendix

Proof of Theorem 5

We provide three lemmas which will be used for the proof of Theorem 5.

Lemma A.1 (see [7]). *Let Assumptions 1 and 2 hold. Then*

$$\left|[\mathbf{P}(k, s)]_{ij} - \frac{1}{N}\right| \leq V^{\lceil(k-s+1)/B\rceil-2}, \tag{A.1}$$

where $\mathbf{P}(k, s) = P(s)P(s+1)\cdots P(k-1)P(k)$, for all $k \geq s \geq 0$.

Lemma A.2. *Under Assumptions 1, 2, 3, and 4, consider a sequence $\{\phi_{k+1}^i\}$ generated according to the method (4) and (5), and then, for all $k \geq 0$,*

$$\max_{v \in \mathcal{V}} \mathbb{E}\left[\left\|\overline{\varphi}_{k+1} - \varphi_{k+1}^v\right\|_*\right] \leq \left(\frac{NB}{V(1-V)} + 2\right)(G_{\mathcal{X}} + \Phi), \tag{A.2}$$

where $\overline{\varphi}_{k+1} = (1/N)\sum_{i=1}^{N} \varphi_{k+1}^i$.

Proof. We can compute the general evolution of φ_{k+1}^i as follows, by referring to (4):

$$\varphi_{k+1}^i = \sum_{j=1}^{N} p_{ij}(k)\,\varphi_k^j + \mathbf{G}_k^i + \phi_k^i$$

$$= \sum_{j=1}^{N} [\mathbf{P}(k,0)]_{ij}\varphi_0^j + \sum_{s=1}^{k}\sum_{j=1}^{N}[\mathbf{P}(k,s)]_{ij}\mathbf{G}_{s-1}^j \qquad \text{(A.3)}$$

$$+ \sum_{s=1}^{k}\sum_{j=1}^{N}[\mathbf{P}(k,s)]_{ij}\phi_{s-1}^j + \mathbf{G}_k^i + \phi_k^i.$$

In a similar way, for $\overline{\varphi}_k$, we have

$$\overline{\varphi}_{k+1} = \overline{\varphi}_0 + \sum_{s=1}^{k}\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_{s-1}^i + \sum_{s=1}^{k}\frac{1}{N}\sum_{i=1}^{N}\phi_{s-1}^i$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_k^i + \frac{1}{N}\sum_{i=1}^{N}\phi_k^i. \qquad \text{(A.4)}$$

Hence by noting that $\varphi_0^i = 0$ for all i it follows that

$$\left\|\overline{\varphi}_{k+1} - \varphi_{k+1}^i\right\|_* \le \sum_{s=1}^{k}\sum_{j=1}^{N}\left|[\mathbf{P}(k,s)]_{ij} - \frac{1}{N}\right|\left\|\mathbf{G}_{s-1}^j\right\|_*$$

$$+ \sum_{s=1}^{k}\sum_{j=1}^{N}\left|[\mathbf{P}(k,s)]_{ij} - \frac{1}{N}\right|\left\|\phi_{s-1}^j\right\|_* + \left\|\mathbf{G}_k^i\right\|_*$$

$$+ \left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_k^i\right\|_* + \left\|\phi_k^i\right\|_* + \left\|\frac{1}{N}\sum_{i=1}^{N}\phi_k^i\right\|_*. \qquad \text{(A.5)}$$

Note that $x_k^i \in \mathcal{X}$ (cf. (6)), for all $i \in \mathcal{V}$ and $k \ge 0$. Hence, we can use the definition of the dual norm and Assumption 3 to bound $\|\mathbf{G}_k^i\|_*$ as follows:

$$\left\|\mathbf{G}_k^i\right\|_* \le \left[\frac{G_{\mathcal{W}}^2}{\gamma\sigma_w} + \frac{G_{\mathcal{Z}}^2}{(1-\gamma)\sigma_z}\right]^{1/2} = G_{\mathcal{X}}. \qquad \text{(A.6)}$$

This, along with Lemma A.2, leads to the following estimate:

$$\mathbb{E}\left[\left\|\overline{\varphi}_{k+1} - \varphi_{k+1}^i\right\|_*\right] \le \sum_{s=1}^{k}\sum_{j=1}^{N}V^{\lceil(k-s+1)/B\rceil-2}\left(G_{\mathcal{X}} + \Phi\right) \qquad \text{(A.7)}$$

$$+ 2\left(G_{\mathcal{X}} + \Phi\right),$$

where we have used the inequality $\mathbb{E}[\|\phi_k^i\|_*] \le [\mathbb{E}[\|\phi_k^i\|_*^2]]^{1/2} \le \Phi$, according to Assumption 4. Hence, the desired result follows by using the inequality that $\sum_{s=1}^{k}\sum_{j=1}^{N}V^{\lceil(k-s+1)/B\rceil-2} \le NB/V(1-V)$. $\qquad\square$

Lemma A.3 (see [22]). *For function* $\Psi_\alpha(\varphi) := \max_{x\in\mathcal{X}}\{\langle x - x_0, \varphi\rangle - \alpha h(x)\}$, *where* x_0 *is the proximal center of* $\mathcal{X}$ *and* α *is some positive scalar, we have the following.*

(a) *Function* $\Psi_\alpha(\cdot)$ *is convex and differentiable, and its gradient satisfies* $\nabla\Psi_\alpha(\varphi) = \pi_\alpha(\varphi) - x_0$.

(b) $\|\nabla\Psi_\alpha(\varphi_1) - \nabla\Psi_\alpha(\varphi_2)\| \le (1/\sigma\alpha)\|\varphi_1 - \varphi_2\|_*$, *for all* $\varphi_1, \varphi_2 \in \mathbb{R}^n \times \mathbb{R}^m$.

(c) *Function* $\Psi_\alpha(\cdot)$ *satisfies* $\Psi_\alpha(\varphi + \Delta\varphi) \le \Psi_\alpha(\varphi) + \langle\nabla\Psi_\alpha(\varphi), \Delta\varphi\rangle + (1/2\sigma\alpha)\|\Delta\varphi\|_*^2$, *for all* $\varphi, \Delta\varphi \in \mathbb{R}^n \times \mathbb{R}^m$.

Proof of Theorem 5. First, we introduce the following gap sequence, for all $T \ge 1$:

$$\Lambda_T := \max_{x\in\mathcal{X}}\left\{\sum_{k=0}^{T-1}\sum_{i=1}^{N}\left\langle x_k^v - x, \mathbf{G}_k^i + \phi_k^i\right\rangle\right\}. \qquad \text{(A.8)}$$

Bounding $\mathbb{E}[\Lambda_T]$. It is easy to see that

$$\mathbb{E}[\Lambda_T] = \mathbb{E}\left[\max_{x\in\mathcal{X}}\left\{\sum_{k=0}^{T-1}\sum_{i=1}^{N}\left\langle x_k^v - x, \mathbf{G}_k^i + \mathbb{E}[\phi_k^i]\right\rangle\right\}\right]$$

$$= \mathbb{E}\left[\max_{x\in\mathcal{X}}\left\{\sum_{k=0}^{T-1}\sum_{i=1}^{N}\left\langle x_k^v - x, \mathbf{G}_k^i\right\rangle\right\}\right], \qquad \text{(A.9)}$$

where we have used Assumption 4; that is, $\mathbb{E}[\phi_k^i] = 0$. Breaking Λ_T into two parts, we have

$$\Lambda_T = \max_{x\in\mathcal{X}}\left\{\sum_{k=0}^{T-1}\sum_{i=1}^{N}\left\langle x_0 - x, \mathbf{G}_k^i + \phi_k^i\right\rangle\right\}$$

$$+ \sum_{k=0}^{T-1}\sum_{i=1}^{N}\left\langle x_k^v - x_0, \mathbf{G}_k^i + \phi_k^i\right\rangle. \qquad \text{(A.10)}$$

For the first term on the right-hand side of (A.10), we can follow an argument similar to that of the proof of Theorem 1 in [2] to provide the following bound:

$$N\Psi_{\alpha_T}\left(-\frac{1}{N}\sum_{k=0}^{T-1}\sum_{i=1}^{N}\left(\mathbf{G}_k^i + \phi_k^i\right)\right) + NH\alpha_T \qquad \text{(A.11)}$$

$$= N\Psi_{\alpha_T}\left(-\overline{\varphi}_T\right) + NH\alpha_T,$$

where we have used (A.4), while for the second term, we achieve this in the following way. By recalling the definition of $\overline{\varphi}_{k+1}$, we have

$$\overline{\varphi}_{k+1} = \frac{1}{N}\sum_{i=1}^{N}\varphi_{k+1}^i$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{N}p_{ij}(k)\right]\varphi_k^j + \frac{1}{N}\sum_{i=1}^{N}\mathbf{G}_k^i + \frac{1}{N}\sum_{i=1}^{N}\phi_k^i \qquad \text{(A.12)}$$

$$= \overline{\varphi}_k + \overline{\mathbf{G}}_k + \overline{\phi}_k,$$

where $\overline{\mathbf{G}}_k = (1/N)\sum_{i=1}^{N}\mathbf{G}_k^i$, $\overline{\phi}_k = (1/N)\sum_{i=1}^{N}\phi_k^i$, and the last equality follows from the fact that the weight matrix $P(k)$ is

double stochastic (cf. Assumption 2). Then, we investigate the sequence $\{\Psi_{\alpha_k}(-\overline{\varphi}_k)\}$; that is,

$$
\begin{aligned}
\Psi_{\alpha_{k+1}}\left(-\overline{\varphi}_{k+1}\right) &\leq \Psi_{\alpha_k}\left(-\overline{\varphi}_{k+1}\right) \\
&= \Psi_{\alpha_k}\left(-\overline{\varphi}_k - \overline{\mathbf{G}}_k - \overline{\phi}_k\right) \\
&\leq \Psi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \left\langle \nabla\Psi_{\alpha_k}\left(-\overline{\varphi}_k\right), \overline{\mathbf{G}}_k + \overline{\phi}_k \right\rangle \\
&\quad + \frac{1}{2\sigma\alpha_k}\left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_*^2.
\end{aligned}
\tag{A.13}
$$

It turns out that, for the term $\nabla\Psi_{\alpha_k}(-\overline{\varphi}_k)$, we have

$$
\begin{aligned}
\nabla\Psi_{\alpha_k}\left(-\overline{\varphi}_k\right) &= \pi_{\alpha_k}\left(-\overline{\varphi}_k\right) - x_0 \\
&= \pi_{\alpha_k}\left(-\varphi_k^v\right) - x_0 + \pi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \pi_{\alpha_k}\left(-\varphi_k^v\right) \\
&= \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^v\right) - x_0 + \pi_{\alpha_k}\left(-\varphi_k^v\right) - \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^v\right) \\
&\quad + \pi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \pi_{\alpha_k}\left(-\varphi_k^v\right),
\end{aligned}
\tag{A.14}
$$

where the first equality follows from Lemma A.3(a). Hence, we can bound the term $-\langle \nabla\Psi_{\alpha_k}(-\overline{\varphi}_k), \overline{\mathbf{G}}_k + \overline{\phi}_k \rangle$ as follows:

$$
\begin{aligned}
&- \left\langle \nabla\Psi_{\alpha_k}\left(-\overline{\varphi}_k\right), \overline{\mathbf{G}}_k + \overline{\phi}_k \right\rangle \\
&\leq - \left\langle \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^v\right) - x_0, \overline{\mathbf{G}}_k + \overline{\phi}_k \right\rangle \\
&\quad + \left\|\pi_{\alpha_k}\left(-\varphi_k^v\right) - \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^v\right)\right\| \left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_* \\
&\quad + \left\|\pi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \pi_{\alpha_k}\left(-\varphi_k^v\right)\right\| \left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_*.
\end{aligned}
\tag{A.15}
$$

By recalling property (ii) of the approximate projection (6) and Lemma A.3(b), we can further obtain

$$
\begin{aligned}
&- \left\langle \nabla\Psi_{\alpha_k}\left(-\overline{\varphi}_k\right), \overline{\mathbf{G}}_k + \overline{\phi}_k \right\rangle \\
&\leq - \left\langle x_k^v - x_0, \overline{\mathbf{G}}_k + \overline{\phi}_k \right\rangle + \xi_k \left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_* \\
&\quad + \frac{1}{\sigma\alpha_k}\left\|\overline{\varphi}_k - \varphi_k^v\right\|_* \left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_*,
\end{aligned}
\tag{A.16}
$$

where the equality $\pi_{\alpha_k}^{\xi_k}(-\varphi_k^v) = x_k^v$ was used, which holds for $k \geq 0$ (for $k = 0$, it is easy to verify that $\pi_{\alpha_k}^{\xi_k}(-\varphi_k^v) = \pi_{\alpha_k}(0) = x_0$). Substituting (A.16) into (A.13) and then taking the expectation, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle x_k^v - x_0, \overline{\mathbf{G}}_k + \overline{\phi}_k \right\rangle\right] \\
&\leq \Psi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \Psi_{\alpha_{k+1}}\left(-\overline{\varphi}_{k+1}\right) + \xi_k\mathbb{E}\left[\left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_*\right] \\
&\quad + \frac{1}{\sigma\alpha_k}\mathbb{E}\left[\left\|\overline{\varphi}_k - \varphi_k^v\right\|_*\right]\mathbb{E}\left[\left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_*\right] \\
&\quad + \frac{1}{2\sigma\alpha_k}\mathbb{E}\left[\left\|\overline{\mathbf{G}}_k + \overline{\phi}_k\right\|_*^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \Psi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \Psi_{\alpha_{k+1}}\left(-\overline{\varphi}_{k+1}\right) + \left(G_{\mathcal{X}} + \Phi\right)\xi_k \\
&\quad + \frac{G_{\mathcal{X}} + \Phi}{\sigma\alpha_k}\mathbb{E}\left[\left\|\overline{\varphi}_k - \varphi_k^v\right\|_*\right] + \frac{1}{2\sigma\alpha_k}\left(G_{\mathcal{X}} + \Phi\right)^2,
\end{aligned}
\tag{A.17}
$$

where we have used the bounds $\|\overline{\mathbf{G}}_k\|_* \leq G_{\mathcal{X}}$ and $\mathbb{E}[\|\overline{\phi}_k\|_*] \leq \Phi$. Hence, combining the inequalities (A.10), (A.11), and (A.17) yields

$$
\begin{aligned}
\mathbb{E}\left[\Lambda_T\right] &\leq N\sum_{k=0}^{T-1}\left[\Psi_{\alpha_k}\left(-\overline{\varphi}_k\right) - \Psi_{\alpha_{k+1}}\left(-\overline{\varphi}_{k+1}\right)\right] \\
&\quad + N\left(G_{\mathcal{X}} + \Phi\right)\sum_{k=0}^{T-1}\xi_k \\
&\quad + \frac{N\left(G_{\mathcal{X}} + \Phi\right)}{\sigma}\sum_{k=0}^{T-1}\frac{1}{\alpha_k}\mathbb{E}\left[\left\|\overline{\varphi}_k - \varphi_k^v\right\|_*\right] \\
&\quad + \frac{N(G_{\mathcal{X}} + \Phi)^2}{2\sigma}\sum_{k=0}^{T-1}\frac{1}{\alpha_k} \\
&\quad + N\Psi_{\alpha_T}\left(-\overline{\varphi}_T\right) + NH\alpha_T \\
&= N\left(G_{\mathcal{X}} + \Phi\right)\sum_{k=0}^{T-1}\xi_k + NH\alpha_T \\
&\quad + \frac{N(G_{\mathcal{X}} + \Phi)^2}{2\sigma}\sum_{k=0}^{T-1}\frac{1}{\alpha_k} \\
&\quad + \frac{N\left(G_{\mathcal{X}} + \Phi\right)}{\sigma}\sum_{k=0}^{T-1}\frac{1}{\alpha_k}\mathbb{E}\left[\left\|\overline{\varphi}_k - \varphi_k^v\right\|_*\right],
\end{aligned}
\tag{A.18}
$$

where we have used the fact that $\Psi_{\alpha_0}(-\overline{\varphi}_0) = 0$.

Bounding $\mathbb{E}[|\mathsf{L}(\widehat{w}_T^v, \widehat{z}_T^v) - \mathsf{L}(w^*, z^*)|]$. Following an argument similar to that of the proof of Theorem 1 in [2], we can arrive at

$$
\begin{aligned}
\frac{1}{T}\mathbb{E}\left[\Lambda_T\right] &\geq \mathbb{E}\left[\max_{z\in\mathcal{Z}}\left\{\frac{1}{T}\sum_{k=0}^{T-1}\mathsf{L}\left(w_k^v, z\right)\right\}\right] \\
&\quad - \mathbb{E}\left[\min_{w\in\mathcal{W}}\left\{\frac{1}{T}\sum_{k=0}^{T-1}\mathsf{L}\left(w, z_k^v\right)\right\}\right] \\
&\quad - \frac{2}{T}\sum_{k=0}^{T-1}\sum_{i=1}^{N}G_{\mathcal{X}}\mathbb{E}\left[\left\|x_k^i - x_k^v\right\|\right] \\
&\geq \mathbb{E}\left[\max_{z\in\mathcal{Z}}\mathsf{L}\left(\widehat{w}_T^v, z\right)\right] - \mathbb{E}\left[\min_{w\in\mathcal{W}}\mathsf{L}\left(w, \widehat{z}_T^v\right)\right] \\
&\quad - \frac{2G_{\mathcal{X}}}{T}\sum_{k=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\left[\left\|x_k^i - x_k^v\right\|\right],
\end{aligned}
\tag{A.19}
$$

where we have used the fact that L is a convex-concave function. By recalling the definition of a saddle point (2), we have

$$\max_{z \in \mathcal{Z}} L\left(\widehat{w}_T^\nu, z\right) \geq L\left(\widehat{w}_T^\nu, z^*\right) \geq L\left(w^*, z^*\right),$$

$$\min_{w \in \mathcal{W}} L\left(w, \widehat{z}_T^\nu\right) \leq L\left(w^*, \widehat{z}_T^\nu\right) \leq L\left(w^*, z^*\right),$$

(A.20)

which further implies

$$\left| L\left(\widehat{w}_T^\nu, \widehat{z}_T^\nu\right) - L\left(w^*, z^*\right) \right| \leq \max_{z \in \mathcal{Z}} L\left(\widehat{w}_T^\nu, z\right) - \min_{w \in \mathcal{W}} L\left(w, \widehat{z}_T^\nu\right).$$

(A.21)

By substituting the preceding inequality into (A.19) leads to

$$\mathbb{E}\left[\left| L\left(\widehat{w}_T^\nu, \widehat{z}_T^\nu\right) - L\left(w^*, z^*\right) \right|\right]$$

$$\leq \frac{1}{T}\mathbb{E}\left[\Lambda_T\right] + \frac{2G_{\mathcal{X}}}{T}\sum_{k=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\left[\left\| x_k^i - x_k^\nu \right\|\right].$$

(A.22)

It remains to bound $\max_{i,\nu \in \mathcal{V}}\mathbb{E}[\| x_k^i - x_k^\nu \|]$. We will use Lemma A.2 to achieve this. Using the update (5), we have

$$\left\| x_k^i - x_k^\nu \right\| = \left\| \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^i\right) - \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^\nu\right) \right\|$$

$$= \left\| \pi_{\alpha_k}\left(-\varphi_k^i\right) - \pi_{\alpha_k}\left(-\varphi_k^\nu\right) + \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^i\right) \right.$$

$$\left. -\pi_{\alpha_k}\left(-\varphi_k^i\right) + \pi_{\alpha_k}\left(-\varphi_k^\nu\right) - \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^\nu\right) \right\|$$

$$\leq \left\| \pi_{\alpha_k}\left(-\varphi_k^i\right) - \pi_{\alpha_k}\left(-\varphi_k^\nu\right) \right\|$$

$$+ \left\| \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^i\right) - \pi_{\alpha_k}\left(-\varphi_k^i\right) \right\|$$

$$+ \left\| \pi_{\alpha_k}\left(-\varphi_k^\nu\right) - \pi_{\alpha_k}^{\xi_k}\left(-\varphi_k^\nu\right) \right\|$$

$$\leq \frac{1}{\sigma\alpha_k}\left\| \varphi_k^i - \varphi_k^\nu \right\|_* + 2\xi_k,$$

(A.23)

where we have used Lemma A.3 and the property (ii) of the approximate projection (6). Using Lemma A.2 it follows that

$$\max_{i,\nu \in \mathcal{V}}\mathbb{E}\left[\left\| x_k^i - x_k^\nu \right\|\right] \leq \left(\frac{NB}{V(1-V)} + 2 \right)\left(G_{\mathcal{X}} + \Phi\right)\frac{2}{\sigma\alpha_k} + 2\xi_k$$

$$= \frac{2R}{\sigma} \cdot \frac{1}{\alpha_k} + 2\xi_k.$$

(A.24)

Combining the inequalities (A.18), (A.22), and (A.24) gives

$$\mathbb{E}\left[\left| L\left(\widehat{w}_T^\nu, \widehat{z}_T^\nu\right) - L\left(w^*, z^*\right) \right|\right]$$

$$\leq \left(N\left(G_{\mathcal{X}} + \Phi\right) + 4NG_{\mathcal{X}} \right)\frac{1}{T}\sum_{k=0}^{T-1}\xi_k + NH\frac{\alpha_T}{T}$$

$$+ \left(\frac{N(G_{\mathcal{X}} + \Phi)^2}{2\sigma} + \frac{N(G_{\mathcal{X}} + \Phi)R}{\sigma} + \frac{4NG_{\mathcal{X}}R}{\sigma} \right)$$

$$\times \frac{1}{T}\sum_{k=0}^{T-1}\frac{1}{\alpha_k}.$$

(A.25)

We are left to bound the terms $(1/T)\sum_{k=0}^{T-1}\xi_k$ and $(1/T)\sum_{k=0}^{T-1}(1/\alpha_k)$. By recalling the definition of the sequence $\{\xi_k\}$, we have

$$\frac{1}{T}\sum_{k=0}^{T-1}\xi_k = \frac{C_\xi}{T}\left[1 + \sum_{k=2}^{T-1}\frac{1}{\sqrt{k}} \right]$$

$$\leq \frac{C_\xi}{T}\left[1 + \int_1^{T-1}\frac{1}{\sqrt{u}}\mathrm{d}u \right]$$

$$= \frac{C_\xi}{T}\left[2\sqrt{T-1} - 1 \right]$$

$$\leq \frac{2C_\xi}{\sqrt{T}}.$$

(A.26)

In a similar way, we have $(1/T)\sum_{k=0}^{T-1}\alpha_k \leq 2C_\alpha/\sqrt{T}$; therefore, the desired result follows by substituting this and (A.26) into (A.25). The proof is complete. □

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

References

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

[2] D. Yuan, Q. Ma, and Z. Wang, "Dual averaging method for solving multi-agent saddle-point problems with quantized information," *Transactions of the Institute of Measurement and Control*, vol. 36, no. 1, pp. 38–46, 2014.

[3] K. I. Tsianos and M. G. Rabbat, "Distributed dual averaging for convex optimization under communication delays," in *Proceedings of the 2012 American Control Conference*, pp. 1067–1072, Montreal, Canada, June 2012.

[4] K. Tsianos, S. Lawlor, and M. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *Proceedings of the 51st IEEE Conference on Decision and Control*, pp. 5453–5458, Maui, Hawaii, USA, 2012.

[5] D. Yuan, S. Xu, H. Zhao, and L. Rong, "Distributed dual averaging method for multi-agent optimization with quantized communication," *Systems & Control Letters*, vol. 61, no. 11, pp. 1053–1061, 2012.

[6] J. Lu, C. Y. Tang, P. R. Regier, and T. D. Bow, "Gossip algorithms for convex consensus optimization over networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 12, pp. 2917–2923, 2011.

[7] A. Nedic and A. Ozdaglar, "Cooperative distributed multi-agent optimization," in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. C. Eldar, Eds., pp. 340–386, Cambridge University Press, Cambridge, UK, 2010.

[8] I. Matei and J. S. Baras, "Performance evaluation of the consensus-based distributed subgradient method under random communication topologies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 754–771, 2011.

[9] 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.

[10] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *Proceedings of the 47th IEEE Conference on Decision and Control (CDC '08)*, pp. 4185–4190, Cancún, Mexico, December 2008.

[11] 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*, 2014.

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

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

[14] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[15] H. Zhao, W. Ren, D. Yuan, and J. Chen, "Distributed discrete-time coordinated tracking with Markovian switching topologies," *Systems & Control Letters*, vol. 61, no. 7, pp. 766–772, 2012.

[16] J. Li, W. Ren, and S. Xu, "Distributed containment control with multiple dynamic leaders for double-integrator dynamics using only position measurements," *IEEE Transactions on Automatic Control*, vol. 57, no. 6, pp. 1553–1559, 2012.

[17] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[18] J. Lu, D. W. C. Ho, and J. Kurths, "Consensus over directed static networks with arbitrary finite communication delays," *Physical Review E*, vol. 80, no. 6, Article ID 066121, 2009.

[19] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[20] W. Ren and R. W. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 655–661, 2005.

[21] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[22] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Mathematical Programming*, vol. 120, no. 1, pp. 221–259, 2009.

[23] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, 2009.