Stochastic Bregman Parallel Direction Method of Multipliers for Distributed Optimization

Yue Yu and Behçet Açıkmeşe

***Abstract*— Bregman parallel direction method of multipliers (BPDMM) efficiently solves distributed optimization over a network, which arises in a wide spectrum of collaborative multi-agent learning applications. In this paper, we generalize BPDMM to stochastic BPDMM, where each iteration only solves local optimization on a randomly selected subset of nodes rather than all the nodes in the network. Such generalization reduce the need for computational resources and allows applications to larger scale networks. We establish both the global convergence and the $O(1/T)$ iteration complexity of stochastic BPDMM. We demonstrate our results via numerical examples.**

I. Introduction

Distributed optimization over a connected undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as follows

$$\begin{array}{ll} \underset{x \in \mathcal{X}^{|\mathcal{V}|}}{\text{minimize}} & \sum\limits_{i \in \mathcal{V}} f_i(x_i) \\ \text{subject to} & x_i = x_j, \ \ \forall \{i,j\} \in \mathcal{E} \end{array} \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a closed convex set, $\mathcal{X}^{\mathcal{V}}$ is the Cartesian product of $|\mathcal{V}|$ copies of $\mathcal{X}$, each f_i is a convex function accessible by node i only. The global optimality is achieved by local optimization on each node and efficient communication between neighboring nodes. In addition to classical applications such as formation control [1], distributed tracking [2] and estimation [3], [4], problem (1) also arises in collaborative learning scenarios [5], [6], where problem (1) represents distributed learning from data collected by multiple agents.

There has been an increasing interest in applying multiplier methods to solve problem (1) [7], [8], [9]. At each iteration of such methods, every primal variable is updated by optimizing a quadratic augmented Lagrangian; every dual variable is updated by numerically integrating local disagreement. Recently, Bregman parallel direction method of multipliers (PDMM) generalized the quadratic augmentation in local optimization to Bregman augmentation, which better exploits the structure of constraint set $\mathcal{X}$, and hence leads to significant improvement in convergence speed [10], [11].

The authors are with the Department of Aeronautics and Astronautics, University of Washington, Seattle, WA, 98195; emails: {yueyu,behcet}@uw.edu

One challenge in implementing multiplier methods for problem (1) is that a local optimization problem needs to be solved on every node in parallel at each iteration, which requires demanding computational resources when applied to large scale networks. A popular approach to address this challenge is stochastic multiplier methods [12], [13], [14], which combine multiplier methods with the idea of stochastic block coordinate descent [15], [16]. At each iteration, stochastic multiplier methods only solve local optimization problems on, rather than all the nodes, a randomly selected subset of nodes. Such algorithms guarantee global convergence to optimum in expectation via proper choice of algorithm parameters. However, to our best knowledge, all existing stochastic multiplier methods use quadratic augmentation. In other words, there is no stochastic extension to Bregman augmentation based multiplier methods.

In this paper, we close this gap in the literature by proposing stochastic BPDMM, which combines the benefits of BPDMM and stochastic multiplier methods. Compared with BPDMM [11], it only requires solving local optimization on a randomly selected subset of nodes, which allows application to larger scale networks; compared with existing stochastic multiplier methods [12], [13], [14], it extends quadratic augmented Lagrangian to Bregman augmented Lagrangian, which improves the convergence speed by better exploiting constraints structure. We establish the global convergence and $O(1/T)$ iteration complexity of stochastic BPDMM, and demonstrate its effectiveness and efficiency via numerical examples.

The rest of the paper is organized as follows. Section II covers necessary background and reformulates problem (1) with consensus constraints. Section III develops the stochastic BPDMM, whose convergence proof is established in Section IV. Section V presents numerical examples and demonstrates the advantages of stochastic BPDMM over prior work. Section VI concludes and comments on future directions.

II. Preliminaries and Background

A. Notation

Let $\mathbb{R}$ ($\mathbb{R}_+$) denote the set of (nonnegative) real numbers, $\mathbb{R}^n$ ($\mathbb{R}^n_+$) the set of n-dimensional (elementwise nonnegative) vectors. Let $\geq$ ($\leq$) denote elementwise inequality when applied to vectors and matrices. Let $\langle \cdot, \cdot \rangle$ denote the dot product. Let $I_n \in \mathbb{R}^{n\times n}$ denote the n-dimensional identity matrix, $\mathbf{1}_n \in \mathbb{R}^n$ the n-dimensional vector of all 1s. Given matrix $A \in \mathbb{R}^{n\times n}$, let A_{ij} denote its (i,j) entry; $A^\top$ denotes its transpose. Let $\otimes$ denote the Kronecker product.

B. Subgradients

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then $g \in \mathbb{R}^n$ is a subgradient of f at $u \in \mathbb{R}^n$ if and only if for any $v \in \mathbb{R}^n$ one has

$$f(v) - f(u) \geq \langle g, v - u \rangle . \tag{2}$$

We denote $\partial f(u)$ the set of subgradients of f at u. An important case of subdifferential is the case of indicator function of a non-empty convex set $\mathcal{X}$ defined as $\delta_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and ∞ otherwise. We will use the following results.

Lemma 1. *[17, Theorem 27.4] Given a closed convex set $\mathcal{X} \subseteq \mathbb{R}^n$ and closed, convex, proper function $f : \mathbb{R}^n \to \mathbb{R}$, then $u^\star = \operatorname{argmin}_{u\in\mathcal{X}} f(u)$ if and only if $0 \in \partial(f + \delta_{\mathcal{X}})(u^\star)$.*

C. Mirror maps and Bregman divergence

Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a convex open set. We say that $\phi : \mathcal{D} \to \mathbb{R}$ is a *mirror map* [18, p.298] if it satisfies: 1) ϕ is differentiable and strictly convex, 2) $\nabla\phi$ takes all possible values, and 3) $\nabla\phi$ diverges on the boundary of the closure of $\mathcal{D}$, *i.e.*, $\lim_{u\to\partial\bar{\mathcal{D}}} \|\nabla\phi(u)\| = \infty$, where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$. The Bregman divergence $B_\phi : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_+$ is defined as [19, Sec. 2.1]

$$B_\phi(u, v) = \phi(u) - \phi(v) - \langle \nabla\phi(v), u - v \rangle . \tag{3}$$

Note that $B_\phi(u, v) \geq 0$ and $B_\phi(u, v) = 0$ only if $u = v$. B_ϕ also satisfy the following three-point identity,

$$\begin{aligned} &\langle \nabla\phi(u) - \nabla\phi(v), w - u \rangle \\ =&B_\phi(w, v) - B_\phi(w, u) - B_\phi(u, v). \end{aligned} \tag{4}$$

D. Graphs and distibuted optimization

An undirected connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ contains a vertex set $\mathcal{V} = \{1, 2, \ldots, m\}$ and an edge set $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ such that $(i, j) \in \mathcal{E}$ if and only if $(j, i) \in \mathcal{E}$ for all $i, j \in \mathcal{V}$. Denote $\mathcal{N}(i)$ the set of neighbors of node i such that $j \in \mathcal{N}(i)$ if $(i, j) \in \mathcal{E}$.

Consider a symmetric stochastic matrix $P \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ defined on the graph $\mathcal{G}$ such that $P_{ij} > 0$ implies that $j \in \mathcal{N}(i)$. Such a matrix P can be constructed, for example, by the graph Laplacian [1, Proposition 3.18]. If P is irreducible [20, Lem. 8.4.1], then 1 is a simple eigenvalue of P with eigenvectors spanned by $\mathbf{1}_{|\mathcal{V}|}$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the underlying graph over which problem (1) is defined. A common approach to solve problem is to create local copies of the design variable $\{x_1, x_2, \ldots, x_{|\mathcal{V}|}\}$ and impose the consensus constraints: $x_i = x_j$ for all $(i, j) \in \mathcal{E}$ [21], [22]. Many different consensus constraints have been proposed [7], [23], [24], [25]. In this paper, we consider consensus constraints of the form:

$$(P \otimes I_n)x = x, \tag{5}$$

where $x = [x_1^\top, x_2^\top, \ldots, x_{|\mathcal{V}|}^\top]^\top$, P is a symmetric, stochastic and irreducible matrix defined on $\mathcal{G}$. We will focus on the following reformulation of problem (1),

$$\begin{array}{ll} \underset{x\in\mathcal{X}^{|\mathcal{V}|}}{\text{minimize}} & \sum\limits_{i\in\mathcal{V}} f_i(x_i) \\ \text{subject to} & (P \otimes I_n)x = x. \end{array} \tag{6}$$

III. Stochastic Bregman Parallel Direction Method of Multipliers

In this section, we first review BPDMM in Algorithm 1, then combine it with the stochastic node update in [13] and propose sBPDMM in Algorithm 2.

BPDMM [11] solves problem (6) with Algorithm 1, which combines the idea of PDMM [8] and Bregman augmented Lagrangian [10]. Each iteration of the algorithm include the following steps:

(a) *Mirror averaging* Step (8a) computes a nodal mirror average of neighboring nodes' variables, and can be further decomposed as follows:

$$\nabla\Phi(z^t) = (P \otimes I_n)\nabla\Phi(x^t) \tag{7a}$$

$$y^t = \underset{y\in\mathcal{X}^{|\mathcal{V}|}}{\operatorname{argmin}} \; B_\Phi(y, z^t) \tag{7b}$$

where $\Phi(x) = \sum_{i\in\mathcal{V}} \phi(x_i)$. Therefore this step is equivalent to first apply $\nabla\Phi$ to x^t, then run an average step, followed by $(\nabla\Phi)^{-1}$, and finally a projection step. See Fig. 1 for an illustration.

(b) *Local optimization* Step (8b) optimizes a nodal augmented Lagrangian. In particular, the Bregman divergence term in the objective of (8b) augments the nodal Lagrangian by penalizing the difference from the nodal mirror average.

(c) *Disagreement integration* Step (9) is a discrete integration of the disagreement between neighboring nodes. Such integration is equivalent to a spring dynamics among neighboring nodes and improves the

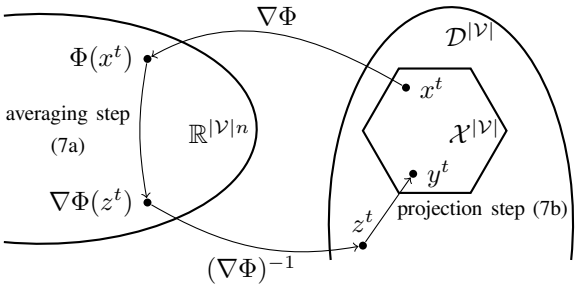

Fig. 1. Mirror averaging

disturbance rejection performance of the algorithm. See [26], [27] for a detailed discussion.

Both mirror averaging step (8a) and disagreement integration step (9) have close-form update when the constraint set $\mathcal{X}$ is structured, *e.g.*, $\mathcal{X}$ is $\mathbb{R}^n$ or the probability simplex [11]. On the other hand, the local optimization step (8b) typically requires an iterative algorithm itself, *e.g.*, mirror descent method [28]. Hence the main computational effort of implementing Algorithm 1 is caused by the local optimization step (8b). At each iteration, Algorithm 1 requires at least $|\mathcal{V}|$ processors, one assigned to each node, to solve optimization (8b) in parallel. Such requirements are computationally demanding for large scale networks.

Algorithm 1 BPDMM

Input: Parameters: $\tau, \rho > 0$; initial point $x^0 \in (\mathcal{X} \cap \mathcal{D})^{|\mathcal{V}|}, \mu^0 \in \mathbb{R}^{|\mathcal{V}|n}$.

for all $t = 0, 1, 2, \ldots$ **do**

$$y_i^t = \operatorname*{argmin}_{y_i \in \mathcal{X}} \sum_{j \in \mathcal{N}(i)} P_{ij} B_\phi(y_i, x_j^t), \ \ \forall i \in \mathcal{V} \quad (8a)$$

$$x_i^{t+1} = \operatorname*{argmin}_{x_i \in \mathcal{X}} f_i(x_i) + \langle x_i, \mu_i^t - \sum_{j \in \mathcal{N}(i)} P_{ij} \mu_j^t \rangle + \rho B_\phi(x_i, y_i^t), \ \ \forall i \in \mathcal{V} \quad (8b)$$

$$\mu_i^{t+1} = \mu_i^t + \tau x_i^{t+1} - \tau \sum_{j \in \mathcal{N}(i)} P_{ij} x_j^{t+1}, \ \ \forall i \in \mathcal{V} \quad (9)$$

end for

In order to address this challenge, we propose Algorithm 2, which uses a stochastic node update [12], [13], [14]. Compared with Algorithm 1, each iteration of Algorithm 2 only execute local optimization step on a set of randomly selected nodes, which requires less number of processors running in parallel. This flexibility reduce the requirements on the total computation power of the network, and allows BPDMM to be applicable much larger scale networks.

Algorithm 2 stochastic BPDMM

Input: Parameters: $\tau, \rho > 0$; initial point $x^0 \in (\mathcal{X} \cap \mathcal{D})^{|\mathcal{V}|}, \mu^0 \in \mathbb{R}^{|\mathcal{V}|n}$.

for all $t = 0, 1, 2, \ldots$ **do**

Randomly select a subset of nodes $\mathcal{S}_{t+1} \subset \mathcal{V}$.

$$y_i^t = \operatorname*{argmin}_{y_i \in \mathcal{X}} \sum_{j \in \mathcal{N}(i)} P_{ij} B_\phi(y_i, x_j^t), \ \ \forall i \in \mathcal{S}_{t+1} \quad (10a)$$

$$x_i^{t+1} = \operatorname*{argmin}_{x_i \in \mathcal{X}} f_i(x_i) + \langle x_i, \mu_i^t - \sum_{j \in \mathcal{N}(i)} P_{ij} \mu_j^t \rangle + \rho B_\phi(x_i, y_i^t), \ \ \forall i \in \mathcal{S}_{t+1} \quad (10b)$$

$$x_i^{t+1} = x_i^t, \ \ \forall i \in \mathcal{V} \setminus \mathcal{S}_{t+1} \quad (10c)$$

$$\mu_i^{t+1} = \mu_i^t + \tau x_i^{t+1} - \tau \sum_{j \in \mathcal{N}(i)} P_{ij} x_j^{t+1}, \ \ \forall i \in \mathcal{V} \quad (11)$$

end for

Although the generalization from Algorithm 1 to Algorithm 2 seems straightforward, the generalization in the corresponding convergence proof requires more careful treatment. In particular, the convergence proof of Algorithm 1 in [11] hinges on a monotonically non-increasing non-negative Lyapunov function for full primal update in (8) with carefully chosen algorithm parameters. In order to generalize such proof to Algoritjm 2, we need to answer the following questions:

- How to find a monotonically non-increasing non-negative Lyapunov function for stochastic partial primal update in (10)?
- How does the randomly selected node set $\mathcal{S}_{t+1}$ affect the choice of algorithm parameters?

In the sequel, we aim to answer theses questions and establish the convergence proof of Algorithm 2.

IV. Convergence

In this section, we prove the global convergence as well as the $O(1/T)$ iteration complexity of Algorithm 2. All detailed proof in this section can be found in the Appendix.

We first group our assumptions in Assumption 1.

Assumption 1. *(a) Function $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are closed, proper and convex for all $i \in \mathcal{V}$.*

(b) Set $\mathcal{X} \subset \mathbb{R}^n$ is closed and convex. There exists a saddle point $(x^\star, \mu^\star)$ such that $x_i^\star \in \mathcal{X}$ and

$$\sum_{j\in\mathcal{V}} P_{ij} x_j^\star = x_i^\star \tag{12a}$$

$$-\mu_i^\star + \sum_{j\in\mathcal{V}} P_{ij}\mu_j^\star \in \partial(f_i + \delta_{\mathcal{X}})(x_i^\star), \tag{12b}$$

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

(c) Function $\phi : \mathcal{D} \to \mathbb{R}$ is a mirror map, where $\mathcal{D}$ is a open convex set such that $\mathcal{X}$ is included in its closure. In addition, function ϕ is α-strongly convex with respect to l_p-norm, i.e., for any $u, v \in \mathcal{X}$,

$$B_\phi(u, v) \geq \frac{\alpha}{2}\|u - v\|_p^2. \tag{13}$$

(d) Matrix P is symmetric, stochastic, irreducible and positive semi-definite.

(e) At each iteration $t+1$, we assume $|\mathcal{S}_{t+1}|/|\mathcal{V}| = \omega, 0 < \omega < 1$.

Now we start to construct the convergence proof of Algorithm 2 under Assumption 1. The optimality condition of (10b) is that for all $i \in \mathcal{S}_{t+1}$,

$$\begin{aligned} &-\mu_i^t + \sum_{j\in\mathcal{V}} P_{ij}\mu_j^t - \rho\left(\nabla\phi(x_i^{t+1}) - \nabla\phi(y_i^t)\right) \\ &\in \partial(f_i + \delta_{\mathcal{X}})(x_i^{t+1}) \end{aligned} \tag{14}$$

Define the residuals of optimality conditions (14) at iteration t as

$$\begin{aligned} R(t+1) \coloneqq\ & \omega(L(x^t, \mu^\star) - L(x^\star, \mu^\star)) \\ &+ \rho \sum_{i\in\mathcal{S}_{t+1}} B_\phi(x_i^{t+1}, y_i^t) + \frac{\gamma\rho}{2}\left\|((I_{|\mathcal{V}|} - P)\otimes I_n)x^t\right\|_2^2, \end{aligned} \tag{15}$$

where $\gamma > 0$ and Lagrangian $L(x, \mu)$ is defined as

$$L(x, \mu) = \sum_{i\in\mathcal{V}}(f_i + \delta_{\mathcal{X}})(x_i) + \langle \mu, ((I_{|\mathcal{V}|} - P)\otimes I_n)x\rangle. \tag{16}$$

Using (12) and (2) we can show the following

$$L(x^t, \mu^\star) - L(x^\star, \mu^\star) \geq 0 \tag{17}$$

Hence $L(x^t, \mu^\star) - L(x^\star, \mu^\star)$ defines a *running duality gap* that measures distance to optimality [8]. Notice that given x^t, $R(t+1)$ is a random variable only depends on $\mathcal{S}_{t+1}$ and $\mathbb{E}_{\mathcal{S}_{t+1}}[R(t+1)] = 0$ implies that $L(x^t, \mu^\star) = L(x^\star, \mu^\star)$ and $x_i^t = x_j^t$ for all $i, j \in \mathcal{V}$, *i.e.*, both optimality and consensus are achieved.

In order to show $\mathbb{E}_{\mathcal{S}_{t+1}}[R(t+1)] = 0$, we define the following Lyapunov function of Algorithm 2

$$\begin{aligned} V(t) \coloneqq & H(x^t, \mu^t) + \frac{\omega}{2\tau}\left\|\mu^\star - \mu^{t-1}\right\|_2^2 \\ &+ \rho\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t). \end{aligned} \tag{18}$$

where

$$H(x^t, \mu^t) = L(x^t, \mu^t) - L(x^\star, \mu^\star) - \tau\left\|Q\otimes I_n)x^t\right\|_2^2 \tag{19}$$

with $Q = I_{|\mathcal{V}|} - P$ and $\mu^{-1} \coloneqq \mu^0 - \tau((I_{|\mathcal{V}|} - P)\otimes I_n)x^0$.

Compared with the one used in [11], the Lyapunov function $V(t)$ defined by (18) contains a generalized Lagrangian $H(x^t, \mu^t)$, which renders the positive definiteness of $V(t)$ unclear. The following lemma shows that $V(t)$ is indeed positive definite, and lower bounded by a Bregman divergence to the optimum.

Lemma 2. *Suppose Assumption 1 holds, if*

$$\tau \leq \frac{\rho\,(\omega\alpha\sigma - \gamma)}{2 - \omega}, \quad 0 < \gamma < \omega\alpha\sigma, \tag{20}$$

where $\sigma = \min\{1, n^{\frac{2}{p}-1}\}$, p and α are defined in (13)*, then the Lyapunov function defined in* (18) *satisfy*

$$V(t) \geq \frac{(1-\omega)\omega\alpha\sigma\rho + \gamma\rho}{(2-\omega)\omega\alpha\sigma}\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t). \tag{21}$$

The sketch of the proof is as follows. Use equation (12b) and (11) we can show

$$H(x^t, \mu^t) \geq -\tfrac{\omega}{2\tau}\left\|\mu^{t-1} - \mu^\star\right\|_2^2 - \tfrac{1}{2\omega\tau}\left\|\mu^t - \mu^{t-1}\right\|_2^2.$$

In addition, equation (11) and Assumption 1, particularly assumptions on function ϕ and matrix P, ensures that

$$-\tfrac{1}{2\omega\tau}\left\|\mu^t - \mu^{t-1}\right\|_2^2 + \tfrac{\tau}{2\omega\sigma}\textstyle\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t) \geq 0.$$

Substitute these two inequalities into (18), use (13) we can show $V(t) \geq (\rho - \frac{\tau}{\omega\alpha\sigma})\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t)$, which, due to the assumption in (20), finally reduces to (21). Then positive definiteness of $V(t)$ follows from the positive definiteness of Bregman divergence and the fact $\frac{(1-\omega)\omega\alpha\sigma\rho+\gamma\rho}{(2-\omega)\omega\alpha\sigma} > 0$ when $0 < \omega < 1$.

Notice that $V(t)$ is a random variable whose value depends on the realization of $\mathcal{S}_{1:t}$, which is the history of selected node sets, *i.e.*, $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_t\}$. The following theorem shows that the expected value of $V(t)$ conditioned on $\mathcal{S}_{1:t}$, *i.e.*, $\mathbb{E}_{\mathcal{S}_{1:t}}[V(t)]$ is monotonically non-increasing with respect to t.

Theorem 1 (Global convergence)**.** *Suppose that Assumption 1 . Let the sequence $\{y^t, x^t, \mu^t\}$ be generated by Algorithm 2. Let $R(t+1)$ and $V(t)$ be defined as in* (15) *and* (18)*, respectively. If $\rho, \tau, \gamma, \omega$ satisfy* (20)*, then we have the following monotonicity relation*

$$\mathbb{E}_{\mathcal{S}_{1:t}}[V(t)] - \mathbb{E}_{\mathcal{S}_{1:t+1}}[V(t+1)] \geq \mathbb{E}_{\mathcal{S}_{1:t+1}}[R(t+1)].$$

The sketch of the proof is as follows. We substitute the subgradient in (14) into (2) and obtain an inequality. Use three point property (4) we can split the right hand side of this inequality into three parts, each contributes to $R(t+1), V(t)$ and $V(t+1)$, respectively. Taking the expectation over realization of $\mathcal{S}_{t+1}$ conditioned on the value of x^t, we obtain the following relation

$$\mathbb{E}_{\mathcal{S}_{t+1}}[R(t+1)] \leq V(t) - \mathbb{E}_{\mathcal{S}_{t+1}}[V(t+1)], \qquad (22)$$

where assumptions in Assumption 1 and (20) ensures that all intermediate terms cancel each other. Taking the expectation over the realization of $\mathcal{S}_{1:t}$ on both sides of (22), we reach the inequality in Theorem 1.

Summing the inequality in Theorem 1 from the case of $t = 0$ to $t = T - 1$ we have

$$\sum_{t=1}^{T} \mathbb{E}_{\mathcal{S}_{1:t}}[R(t)] \leq V(0). \qquad (23)$$

Since $\mathbb{E}_{\mathcal{S}_{1:t}}[R(t)] \geq 0$ for all t, inequality (23) implies that $\mathbb{E}_{\mathcal{S}_{1:t}}[R(t)] \to 0$ as $T \to \infty$, which establishes the global convergence of Algorithm 2. In addition, if we apply Jensen's inequality to (23), we obtain the following corollary, which shows the the $O(1/T)$ iteration complexity of Algorithm 2 in an ergodic sense.

Corollary 1 (Iteration complexity)**.** *Suppose that Assumption 1 holds. Let the sequence $\{y^t, x^t, \mu^t\}$ be generated by Algorithm 2. Let $V(t)$ be defined as in* (18)*, $\overline{x}^T = \frac{1}{T}\sum_{t=0}^{T-1} x^t$. If $\rho, \tau, \gamma, \omega$ satisfy* (20)*, then*

$$\mathbb{E}_{\mathcal{S}_{1:T}}\left[L(\overline{x}^T, \mu^\star) - L(x^\star, \mu^\star)\right] \leq \frac{V(0)}{\omega T}$$
$$\mathbb{E}_{\mathcal{S}_{1:T}}\left[\frac{1}{2}\left\|((I_{|\mathcal{V}|} - P) \otimes I_n)\overline{x}^T\right\|_2^2\right] \leq \frac{V(0)}{\gamma\rho T}$$

The bound on running duality gap was used in [8].

V. Numerical examples

In this section, we demonstrate the effectiveness and efficiency of Algorithm 2 via numerical examples.

Consider the an instance of problem (1) where $f_i(x_i) = \langle c_i, x_i\rangle$ and $\mathcal{X} = \{u \in \mathbb{R}^n_+ | \, \|u\|_1 = 1\}$ is the probability simplex, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a undirected connected communication graph. Such optimizaton can model, for example, multi-agent decision making, where c_i is the cost of agent i for choosing policy x_i.

We generate an instance of this optimization where entries of $c_1, \ldots, c_{|\mathcal{V}|} \in \mathbb{R}^{100}$ are sampled from standard normal distribution. $\mathcal{G}$ is a randomly generated with $|\mathcal{V}| = 100$ and edge probability 0.2 [1, p. 90]. Matrix P is obtained by minimizing its second largest eigenvalue (in this case, $\lambda_2(P) = 0.4786$) while preserving graph adjacency constraints. We choose the following parameters in Algorithm 2:

- $\phi(u) = \sum_{k=1}^n u[k] \ln u[k]$, where $u[k]$ denotes the k-th element of vector u. Then assumption in (13) is satisfied by $\alpha = 1, p = 1$ (see Remark 1 in [10]).
- $\rho = 1, \tau = \omega/(4 - 2\omega)$. Notice that assumptions in (20) are satisfied with $\gamma = \omega/2$.

With these assumptions, the mirror averaging step (10a) and local optimization step (10b) reduces to the following (see Section 4.3 in [18] for details)

$$y_i^t = \text{Proj}\left[\prod_{j\in\mathcal{N}(i)} (x_j^t)^{P_{ij}}\right] \qquad (24a)$$
$$x_i^t = \text{Proj}\left[y_i^t \exp\frac{-c_i - \mu_i + \sum_{j\in\mathcal{N}(i)} P_{ij}\mu_j}{\rho}\right] \qquad (24b)$$

where multiplication, power and exponential operation on vectors are all elementwise, and $\text{Proj}[u] = u/\|u\|_1$ for all $u \in \mathbb{R}^n$. Update (24) amounts to elementwise operation that allows massive parallel implementation.

We demonstrate the convergence performance of Algorithm 2 in Fig. 2 and Fig. 3, where f^t and $f^\star$ are the objective function value achieved at iteration t and, respectively, optimality. In particular, Fig. 2 shows that as ω increases, the convergence of Algorithm 2 becomes faster and less oscillating, which is because more nodes get updated at each iteration. Fig. 3 shows that when we choose ϕ as negative entropy function rather than quadratic function, the convergence speed is improved dramatically. This is because compared with quadratic function, negative entropy function exploits the structure of probability simplex much better. Such improvement demonstrates the advantage of Algorithm 2 over stochastic multiplier methods based on quadratic augmentation [12], [13], [14].

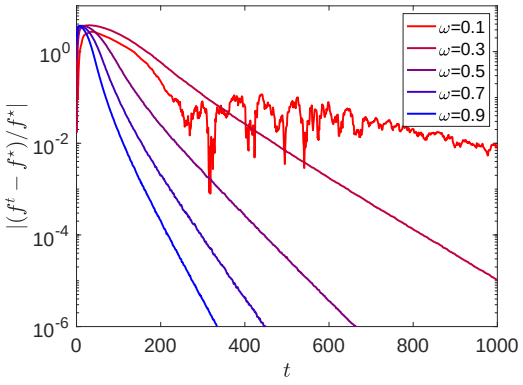

Fig. 2. Comparison of different ω values

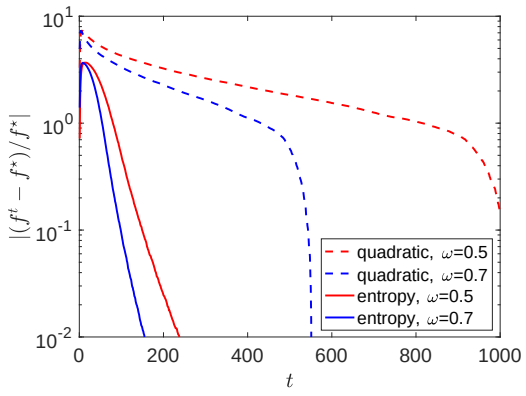

Fig. 3. Comparison of different ϕ function

VI. Conclusions

In this paper, we generalize BPDMM [11] to stochastic BPDMM, where each iteration only solves local optimization on a randomly selected subset of nodes rather than all the nodes in the network. Such generalization requires less number of processors running in parallel, hence allows application to much larger scale networks. Future directions include generalization to directed and time varying networks.

References

[1] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.

[2] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 17–29, 2002.

[3] B. Açıkmeşe, M. Mandić, and J. L. Speyer, "Decentralized observers with consensus filters for distributed discrete-time linear systems," *Automatica*, vol. 50, no. 4, pp. 1037–1052, 2014.

[4] V. Lesser, C. L. Ortiz Jr, and M. Tambe, *Distributed Sensor Networks: A Multiagent Perspective*. Springer Science & Business Media, 2012, vol. 9.

[5] B. Gholami, S. Yoon, and V. Pavlovic, "Decentralized approximate bayesian inference for distributed sensor network." in *AAAI Conf. Artificial Intell.*, 2016, pp. 1582–1588.

[6] A. Yahya, A. Li, M. Kalakrishnan, Y. Chebotar, and S. Levine, "Collective robot reinforcement learning with distributed asynchronous guided policy search," in *Int. Conf. Intell. Robots Syst.* IEEE, 2017, pp. 79–86.

[7] E. Wei and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *Proc. IEEE Conf. Decision Control*, 2012, pp. 5445–5450.

[8] D. Meng, M. Fazel, and M. Mesbahi, "Proximal alternating direction method of multipliers for distributed optimization on weighted graphs," in *Proc. IEEE Conf. Decision Control*, 2015, pp. 1396–1401.

[9] W. Deng, M.-J. Lai, Z. Peng, and W. Yin, "Parallel multi-block ADMM with $O(1/k)$ convergence," *J. Sci. Comput.*, vol. 71, no. 2, pp. 712–736, 2017.

[10] H. Wang and A. Banerjee, "Bregman alternating direction method of multipliers," in *Adv. Neural Inform. Process. Syst.*, 2014, pp. 2816–2824.

[11] Y. Yu, B. Açıkmeşe, and M. Mesbahi, "Bregman parallel direction method of multipliers for distributed optimization via mirror averaging," *IEEE Control Syst. Lett.*, vol. 2, no. 2, pp. 302–306, 2018.

[12] E. Wei and A. Ozdaglar, "On the $O(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers," in *Proc. Global Conf. Signal Inform. Process.* IEEE, 2013, pp. 551–554.

[13] H. Wang, A. Banerjee, and Z.-Q. Luo, "Parallel direction method of multipliers," in *Adv. Neural Inform. Process. Syst.*, 2014, pp. 181–189.

[14] Z. Zhu and A. J. Storkey, "Stochastic parallel block coordinate descent for large-scale saddle point problems." in *Proc. AAAI Conf. Artificial Intell.*, 2016, pp. 2429–2437.

[15] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM J. Optim.*, vol. 22, no. 2, pp. 341–362, 2012.

[16] P. Richtárik and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Math. Prog.*, vol. 144, no. 1-2, pp. 1–38, 2014.

[17] R. T. Rockafellar, *Convex analysis*. Princeton University Press, 2015.

[18] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, 2015.

[19] Y. Censor, S. A. Zenios, *et al.*, *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1997.

[20] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[21] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall Englewood Cliffs, NJ, 1989, vol. 23.

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

[23] D. Jakovetic, J. M. Moura, and J. Xavier, "Distributed augmented lagrangian algorithms: convergence rate," in *Proc. IEEE Global Conf. Signal Inform. Process.*, 2013, pp. 563–566.

[24] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers," in *Proc. IEEE Conf. Decision Control*, 2013, pp. 3671–3676.

[25] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization." *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, 2014.

[26] J. Wang and N. Elia, "Control approach to distributed optimization," in *Proc. Allerton Conf. Commun. Control Comput.* IEEE, 2010, pp. 557–561.

[27] Y. Yu, B. Açıkmeşe, and M. Mesbahi, "Mass-spring-damper network for distributed averaging and optimization," *arXiv preprint arXiv:1808.01999 [math. OC]*, 2018.

[28] A. S. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization.* Wiley, 1983.

APPENDIX

For notation simplicity, we let $Q \coloneqq I_{|\mathcal{V}|} - P$. Suppose Assumption 1 holds, then the nullspace of $I_{|\mathcal{V}|} - P$ is spanned by $\mathbf{1}_{|\mathcal{V}|}$ In addition, Assumption (1) and update rule (10) ensure that

$$(Q \otimes I_n)x^\star = 0 \tag{25a}$$
$$\delta_{\mathcal{X}}(x_i^t) = \delta_{\mathcal{X}}(x_i^\star) = 0, \ \ \forall i \in \mathcal{V} \tag{25b}$$

for all t. We will need the following lemmas.

Lemma 3. *Let*

$$y_i^t = \operatorname*{argmin}_{y_i \in \mathcal{X}} \sum_{j \in \mathcal{N}(i)} P_{ij} B_\phi(y_i, x_j^t), \tag{26}$$

for all $i \in \mathcal{V}$. Then for any $u \in \mathcal{X}$,

$$\sum_{i \in \mathcal{V}} \big(B_\phi(u, x_i^t) - B_\phi(u, y_i^t)\big) \geq \sum_{i,j \in \mathcal{V}} P_{ij} B_\phi(y_i^t, x_j^t) \tag{27}$$

Proof. Equation (26) holds if and only if: for any $u \in \mathcal{X}$,

$$\sum_{j \in \mathcal{V}} P_{ij} \langle \nabla\phi(y_i^{(t)}) - \nabla\phi(x_j^{(t)}), u - y_i^{(t)} \rangle \geq 0$$

Using three point property (4), we have

$$\begin{aligned} &\sum_{j \in \mathcal{V}} P_{ij} B_\phi(u, x_j^{(t)}) - \sum_{j \in \mathcal{V}} P_{ij} B_\phi(u, y_i^{(t)}) \\ \geq &\sum_{j \in \mathcal{V}} P_{ij} B_\phi(y_i^{(t)}, x_j^{(t)}) \end{aligned} \tag{28}$$

Summing (28) over all $i \in \mathcal{V}$ completes the proof. □

Lemma 4. *Suppose Assumption 1 holds. Then*

$$\sigma \left\|(Q \otimes I_n)u\right\|_2^2 \leq \sum_{i,j \in \mathcal{V}} P_{ij} \left\|u_i - v_j\right\|_p^2 \tag{29}$$

for all $u, v \in \mathcal{X}^{|\mathcal{V}|}$, where $\|\cdot\|_p$ denote l_p norm and $\sigma = \min\{1, n^{\frac{2}{p}-1}\}$.

Proof. First, observe that if P is symmetric, stochastic, irreducible and positive semi-definite, $P - P^2$ is positive semi-definite [20, Theorem 8.4.4]. Since $P\mathbf{1}_{|\mathcal{V}|} = P^\top \mathbf{1}_{|\mathcal{V}|} = \mathbf{1}_{|\mathcal{V}|}$, we can show the following

$$\begin{aligned} &\sum_{i,j \in \mathcal{V}} P_{ij} \left\| \sum_{k \in \mathcal{V}} P_{ik} u_k - u_j \right\|_2^2 \\ = &\|u\|_2^2 - \|(P \otimes I_n)u\|_2^2 \\ \geq &\|u\|_2^2 - \|(P \otimes I_n)u\|_2^2 - 2\langle u, ((P - P^2) \otimes I_n)u \rangle \\ = &\|(Q \otimes I_n)u\|_2^2 \end{aligned}$$

Hence (29) holds due to the fact that

$$\sum_{k \in \mathcal{V}} P_{ik} u_k = \operatorname*{argmin}_{w \in \mathcal{X}} \sum_{j \in \mathcal{V}} P_{ij} \left\|w - u_j\right\|_2^2,$$

for all $i \in \mathcal{V}$, and that $\|w\|_2^2 \leq 1/\sigma \|w\|_p^2$ for all $w \in \mathbb{R}^n$ where $\sigma = \min\{1, n^{\frac{2}{p}-1}\}$. □

A. Lemma 2

Proof. Using (25a) and (16) we can show that

$$\begin{aligned} &L(x^t, \mu^t) - L(x^\star, \mu^\star) \\ = &\sum_{i \in \mathcal{V}} \big((f_i + \delta_{\mathcal{X}})(x_i^t) - (f_i + \delta_{\mathcal{X}})(x_i^\star)\big) \\ &+ \langle \mu^t, (Q \otimes I_n)x^t \rangle \overset{(12b)}{\geq} \langle \mu^t - \mu^\star, (Q \otimes I_n)x^t \rangle \end{aligned} \tag{30}$$

Substitute (30) into (19) we have

$$\begin{aligned} &H(x^t, \mu^t) \\ \geq &\langle \mu^t - \mu^\star, (Q \otimes I_n)x^t \rangle - \tau \left\|(Q \otimes I_n)x^t\right\|_2^2 \\ \overset{(11)}{=} &\frac{1}{\tau} \langle \mu^{t-1} - \mu^\star, \mu^t - \mu^{t-1} \rangle \\ \geq &-\frac{\omega}{2\tau} \left\|\mu^{t-1} - \mu^\star\right\|_2^2 - \frac{1}{2\omega\tau} \left\|\mu^t - \mu^{t-1}\right\|_2^2 \end{aligned} \tag{31}$$

where the last step is due to $2\langle a, b \rangle \geq -\|a\|_2^2 - \|b\|_2^2$. Therefore, substitute (31) into (18) we have

$$\begin{aligned} V(t) \geq &\rho \sum_{i \in \mathcal{V}} B_\phi(x_i^\star, x_i^t) - \frac{1}{2\omega\tau} \left\|\mu^t - \mu^{t-1}\right\|_2^2 \\ \overset{(13)}{\geq} &\left(\rho - \frac{\tau}{\omega\alpha\sigma}\right) \sum_{i \in \mathcal{V}} B_\phi(x_i^\star, x_i^t) \\ &+ \frac{\tau}{2\omega\sigma} \sum_{i \in \mathcal{V}} \left\|x_i^t - x_i^\star\right\|_p^2 - \frac{1}{2\omega\tau} \left\|\mu^t - \mu^{t-1}\right\|_2^2 \\ \overset{(20)}{\geq} &\frac{(1-\omega)\omega\alpha\sigma\rho + \gamma\rho}{(2-\omega)\omega\alpha\sigma} \sum_{i \in \mathcal{V}} B_\phi(x_i^\star, x_i^t) \\ &+ \frac{\tau}{2\omega\sigma} \left(\sum_{i \in \mathcal{V}} \left\|x_i^t - x_i^\star\right\|_p^2 - \frac{\sigma}{\tau^2} \left\|\mu^t - \mu^{t-1}\right\|_2^2 \right) \end{aligned} \tag{32}$$

Since $x_i^\star = x_j^\star$ for all $i, j \in \mathcal{V}$, we have

$$\begin{aligned} 0 \overset{(29)}{\leq} &\sum_{i,j \in \mathcal{V}} P_{ij} \left\|x_i^t - x_j^\star\right\|_p^2 - \sigma \left\|(Q \otimes I_n)x^t\right\|_2^2 \\ = &\sum_{i,j \in \mathcal{V}} P_{ij} \left\|x_i^t - x_i^\star\right\|_p^2 - \sigma \left\|(Q \otimes I_n)x^t\right\|_2^2 \\ \overset{(11)}{=} &\sum_{i \in \mathcal{V}} \left\|x_i^t - x_i^\star\right\|_p^2 - \frac{\sigma}{\tau^2} \left\|\mu^t - \mu^{t-1}\right\|_2^2 \end{aligned}$$

Substitute the above inequality into (32) we obtain (21). □

B. Theorem 1

Proof. Let q_i be the i-th column of Q. Since $f+\delta_{\mathcal{X}}$ is convex, the subgradient in (14) satisfy the following

$$\begin{aligned}&\sum_{i\in\mathcal{S}_{t+1}} f_i(x_i^{t+1}) - \sum_{i\in\mathcal{S}_{t+1}} f_i(x_i^\star)\\ \le&\sum_{i\in\mathcal{S}_{t+1}} \langle -\mu^t, (q_i\otimes I_n)(x_i^{t+1}-x_i^\star)\rangle\\ &+\rho\sum_{i\in\mathcal{S}_{t+1}}\langle \nabla\phi(x_i^{t+1})-\nabla\phi(y_i^t), x_i^\star - x_i^{t+1}\rangle,\end{aligned} \tag{33}$$

where we use (25b).

The first term on the RHS of (33) can be rewritten as

$$\begin{aligned}&\sum_{i\in\mathcal{S}_{t+1}} \langle -\mu^t, (q_i\otimes I_n)(x_i^{t+1}-x_i^\star)\rangle\\ \overset{(10c)}{=}&\sum_{i\in\mathcal{S}_{t+1}} \langle -\mu^t, (q_i\otimes I_n)(x_i^{t}-x_i^\star)\rangle\\ &+\langle \mu^t, (Q\otimes I_n)x^{t+1}\rangle - \langle \mu^t, (Q\otimes I_n)x^{t+1}\rangle\\ \overset{(11)}{=}&-\sum_{i\in\mathcal{S}_{t+1}} \langle \mu^t, (q_i\otimes I_n)(x_i^{t}-x_i^\star)\rangle + \langle \mu^t, (Q\otimes I_n)x^{t}\rangle\\ &-\langle \mu^{t+1}, (Q\otimes I_n)x^{t+1}\rangle + \tau\left\|(Q\otimes I_n)x^{t+1}\right\|_2^2\end{aligned} \tag{34}$$

To simplify the second term on the RHS of (33), notice that

$$\begin{aligned}&\sum_{i\in\mathcal{S}_{t+1}} \langle \nabla\phi(x_i^{t+1})-\nabla\phi(y_i^t), x_i^\star - x_i^{t+1}\rangle\\ \overset{(4)}{=}&\sum_{i\in\mathcal{S}_{t+1}} \left(B_\phi(x_i^\star, y_i^t) - B_\phi(x_i^\star, x_i^{t+1}) - B_\phi(x_i^{t+1}, y_i^t)\right)\\ \overset{(10c)}{=}&\sum_{i\in\mathcal{S}_{t+1}} \left(B_\phi(x_i^\star, y_i^t) - B_\phi(x_i^\star, x_i^{t})\right) + \sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t)\\ &-\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^{t+1}) - \sum_{i\in\mathcal{S}_{t+1}} B_\phi(x_i^{t+1}, y_i^t)\end{aligned} \tag{35}$$

Substitute (34) and (35) into (33), we have

$$\begin{aligned}&\sum_{i\in\mathcal{S}_{t+1}} f_i(x_i^{t+1}) - \sum_{i\in\mathcal{S}_{t+1}} f_i(x_i^\star)\\ \le&-\sum_{i\in\mathcal{S}_{t+1}} \langle \mu^t, (q_i\otimes I_n)(x_i^{t}-x_i^\star)\rangle + \langle \mu^t, (Q\otimes I_n)x^{t}\rangle\\ &-\langle \mu^{t+1}, (Q\otimes I_n)x^{t+1}\rangle + \tau\left\|(Q\otimes I_n)x^{t+1}\right\|_2^2\\ &+\rho\sum_{i\in\mathcal{S}_{t+1}} \left(B_\phi(x_i^\star, y_i^t) - B_\phi(x_i^\star, x_i^{t})\right)\\ &+\rho\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t) - \rho\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^{t+1})\\ &-\rho\sum_{i\in\mathcal{S}_{t+1}} B_\phi(x_i^{t+1}, y_i^t)\end{aligned} \tag{36}$$

In addition, notice that

$$\begin{aligned}&\sum_{i\in\mathcal{S}_{t+1}} \left(f_i(x_i^t) - f_i(x_i^\star)\right) \overset{(10c)}{=}\\ &\sum_{i\in\mathcal{S}_{t+1}} \left(f_i(x_i^{t+1}) - f_i(x_i^\star)\right) + \sum_{i\in\mathcal{V}} \left(f_i(x_i^t) - f_i(x_i^{t+1})\right)\end{aligned} \tag{37}$$

Substitute (36) into (37), we have

$$\begin{aligned}&\sum_{i\in\mathcal{S}_{t+1}} \left(f_i(x_i^t) - f_i(x_i^\star)\right)\\ \le&H(x^t,\mu^t) - H(x^{t+1},\mu^{t+1}) + \tau\left\|(Q\otimes I_n)x^t\right\|_2^2\\ &-\sum_{i\in\mathcal{S}_{t+1}} \langle \mu^t, (q_i\otimes I_n)(x_i^{t}-x_i^\star)\rangle\\ &+\rho\sum_{i\in\mathcal{S}_{t+1}} \left(B_\phi(x_i^\star, y_i^t) - B_\phi(x_i^\star, x_i^{t})\right)\\ &+\rho\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t) - \rho\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^{t+1})\\ &-\rho\sum_{i\in\mathcal{S}_{t+1}} B_\phi(x_i^{t+1}, y_i^t)\end{aligned} \tag{38}$$

where we use the definition in (15).

Taking the expectation of (38) over $\mathcal{S}_{t+1}$ conditioned on x^t, we have the following

$$\begin{aligned}&\omega\sum_{i\in\mathcal{V}} \left(f_i(x_i^t) - f_i(x_i^\star)\right)\\ \le&H(x^t,\mu^t) - \mathbb{E}_{\mathcal{S}_{t+1}}\left[H(x^{t+1},\mu^{t+1})\right]\\ &+\tau\left\|(Q\otimes I_n)x^t\right\|_2^2 - \omega\langle \mu^t, (Q\otimes I_n)x^t\rangle\\ &+\rho\omega\sum_{i\in\mathcal{V}} \left(B_\phi(x_i^\star, y_i^t) - B_\phi(x_i^\star, x_i^{t})\right)\\ &+\rho\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^t) - \rho\mathbb{E}_{\mathcal{S}_{t+1}}\left[\sum_{i\in\mathcal{V}} B_\phi(x_i^\star, x_i^{t+1})\right]\\ &-\rho\mathbb{E}_{\mathcal{S}_{t+1}}\left[\sum_{i\in\mathcal{S}_{t+1}} B_\phi(x_i^{t+1}, y_i^t)\right]\end{aligned} \tag{39}$$

where we use (25a). Here we assume y_i^t is computed as in (8a) for all nodes in $\mathcal{V}$, even though Algorithm 1 only require computation on nodes in $\mathcal{S}_{t+1}$. Substitute (25b) into (16) we have

$$\begin{aligned}&\sum_{i\in\mathcal{V}}(f_i(x_i^t) - f_i(x_i^\star))\\ =&L(x^t,\mu^\star) - L(x^\star,\mu^\star) - \langle \mu^\star, (Q\otimes I_n)x^t\rangle\end{aligned} \tag{40}$$

Combine (39) and (40) we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_{t+1}}\left[R(t+1)\right]\\
\leq& H(x^t,\mu^t)-\mathbb{E}_{\mathcal{S}_{t+1}}\left[H(x^{t+1},\mu^{t+1})\right]\\
&-\omega\langle\mu^t-\mu^\star,(Q\otimes I_n)x^t\rangle\\
&+\rho\omega\sum_{i\in\mathcal{V}}\left(B_\phi(x_i^\star,y_i^t)-B_\phi(x_i^\star,x_i^t)\right)\\
&+\rho\sum_{i\in\mathcal{V}}B_\phi(x_i^\star,x_i^t)-\rho\mathbb{E}_{\mathcal{S}_{t+1}}\left[\sum_{i\in\mathcal{V}}B_\phi(x_i^\star,x_i^{t+1})\right]\\
&+(\tau+\frac{\rho\gamma}{2})\left\|(Q\otimes I_n)x^t\right\|_2^2
\end{aligned}\tag{41}$$

Using (4) and (11) we can show

$$\begin{aligned}
&-\langle\mu^t-\mu^\star,(Q\otimes I_n)x^t\rangle=\frac{1}{2\tau}\left\|\mu^\star-\mu^{t-1}\right\|_2^2\\
&-\frac{1}{2\tau}\left\|\mu^\star-\mu^t\right\|_2^2-\frac{\tau}{2}\left\|(Q\otimes I_n)x^t\right\|_2^2
\end{aligned}\tag{42}$$

Substitue (42) into (41), use the definition in (18) we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_{t+1}}\left[R(t+1)\right]\\
\leq& V(t)-\mathbb{E}_{\mathcal{S}_{t+1}}\left[V(t+1)\right]\\
&+\rho\omega\sum_{i\in\mathcal{V}}\left(B_\phi(x_i^\star,y_i^t)-B_\phi(x_i^\star,x_i^t)\right)\\
&+\left(\tau+\frac{\rho\gamma}{2}-\frac{\omega\tau}{2}\right)\left\|(Q\otimes I_n)x^t\right\|_2^2
\end{aligned}\tag{43}$$

Since

$$\begin{aligned}
&\sum_{i\in\mathcal{V}}(B_\phi(x_i^\star,y_i^t)-B_\phi(x_i^\star,x_i^t))\\
\overset{(27)}{\leq}&-\sum_{i,j\in\mathcal{V}}P_{ij}B_\phi(y_i^t,x_j^t)\overset{(13)}{\leq}-\frac{\alpha}{2}\sum_{i,j\in\mathcal{V}}P_{ij}\left\|y_i^t-x_i^t\right\|_p^2\\
\overset{(29)}{\leq}&-\frac{\alpha\sigma}{2}\left\|(Q\otimes I_n)x^t\right\|_2^2
\end{aligned}\tag{44}$$

Substitute (42) into (41) we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_{t+1}}\left[R(t+1)\right]\\
\leq& V(t)-\mathbb{E}_{\mathcal{S}_{t+1}}\left[V(t+1)\right]\\
&+\frac{(2-\omega)\tau+\rho(\gamma-\omega\alpha\sigma)}{2}\left\|(Q\otimes I_n)x^t\right\|_2^2\\
\overset{(20)}{\leq}&V(t)-\mathbb{E}_{\mathcal{S}_{t+1}}\left[V(t+1)\right].
\end{aligned}\tag{45}$$

Taking the expectation of (41) over realization of $\mathcal{S}_{1:t}$ we obtain the desired results. □