Distributed Bregman-Distance Algorithms for Min-Max Optimization

Kunal Srivastava, Angelia Nedić and Dušan Stipanović

Abstract We consider a min-max optimization problem over a time-varying network of computational agents, where each agent in the network has its local convex cost function which is a private knowledge of the agent. The agents want to jointly minimize the maximum cost incurred by any agent in the network, while maintaining the privacy of their objective functions. To solve the problem, we consider subgradient algorithms where each agent computes its own estimates of an optimal point based on its own cost function, and it communicates these estimates to its neighbors in the network. The algorithms employ techniques from convex optimization, stochastic approximation and averaging protocols (typically used to ensure a proper information diffusion over a network), which allow time-varying network structure. We discuss two algorithms, one based on exact-penalty approach and the other based on primal-dual Lagrangian approach, where both approaches utilize Bregman-distance functions. We establish convergence of the algorithms (with probability one) for a diminishing step-size, and demonstrate the applicability of the algorithms by considering a power allocation problem in a cellular network.

1 Introduction

This work is motivated by coordinator-free distributed algorithms for optimization problems originating in [38, 5], which have recently seen a resurgence driven by wireless network applications. A canonical problem in coordinator-free distributed setting is the problem of reaching an agreement among the local decision variables

Kunal Srivastava
University of Illinois, Urbana, IL 61801, USA, e-mail: kkunal2@illinois.edu

Angelia Nedić
University of Illinois, Urbana, IL 61801, USA, e-mail: angelia@illinois.edu

Dušan Stipanović
University of Illinois, Urbana, IL 61801, USA, e-mail: dusan@illinois.edu

in a network of computational agents [39]. Protocols for achieving an agreement employ local averaging which is known to be robust to time-varying graph topology and noisy communication links [18]. These averaging protocols have led to a new class of distributed algorithms for parameter estimation [37], distributed optimization [29, 25, 23, 19, 21, 35] and control of multi-agent systems [16]. Recent work in agreement-based distributed optimization has focused on minimizing the sum of agents' local cost functions [25, 23, 35, 17, 19, 20, 32], which arises in wide variety of areas ranging from sensor networks [28] to distributed machine learning [1].

Optimizing the sum of local objective functions is a popular choice for problems arising in resource allocation and network utility maximization [34], as the objective function is a measure of fair resource allocation. An alternative notion of fairness is the min-max criterion [8], where the interest is in determining an optimal decision variable that minimizes the worst case loss incurred by any agent. However, the algorithms for solving distributed min-max problems over networks are in their infancy. To fill out this void, we have recently considered a distributed method for solving such a problem in [36], where we provided a subgradient algorithm based on the exact-penalty function approach. Here, we make a further progress in two different directions: (1) we present a distributed exact penalty algorithm that uses Bregman distances [10] as opposed to the standard Euclidean distance considered in [36]; and (2) we provide an alternative distributed primal-dual algorithm that also uses Bregman distances. We establish convergence of both algorithms for a diminishing step-size and time-varying network, under mild conditions on the network connectivity. We allow for stochastic errors in subgradient evaluations, which are assumed to be zero-mean and with uniformly bounded expected norm, as typically done in stochastic approximations or stochastic subgradient methods [13, 14, 27, 6, 9, 26].

The min-max optimization aspect addressed in this paper is novel with respect to the prior work on distributed optimization over a network, which is dealing exclusively with the minimization of the sum of agent objectives [29, 25, 23, 19, 21, 1, 35, 32], except for [36]. The network aspect of this paper sets it apart from the standard optimization problems with noisy (sub)-gradient evaluations [14, 27, 6, 9, 26].

The advantage of the proposed algorithms is their ability to solve a class min-max distributed problems for which currently there are no specific algorithms, except for [36]. The difficulty in developing such algorithms comes from the inherent distributed knowledge of the problem data in a network of agents. Thus, there is a need for an algorithm that has ability to "align" the iterates among the agents while simultaneously solving the network problem. Both of our proposed algorithms achieve this task.

Regarding the benefits of the proposed algorithms, the exact-penalty algorithm is simpler for implementation, but in principle it is a distributed subgradient approach due to the use of non-differentiable penalty functions. The primal-dual approach requires a larger number of variables than the penalty approach, but it preserves the smoothness of the problem (provided that the original agent objective functions f_i are smooth). In the presence of stochastic errors in (sub)gradient evaluations, both algorithms will have overall rate of the order of $1/\sqrt{k}$ in terms of the number of iterations k, which is typical for stochastic approximations [27, 26].

The rest of the chapter is organized as follows. In Section 2 we state our problem of interest and suitably reformulate it for the development of distributed algorithms. In Section 3 we present our distributed Bregman-distance based algorithm which utilizes the exact penalty function approach, and we establish the convergence of the algorithm. In Section 4 we develop an alternative algorithm which builds on the primal-dual approach of Arrow-Hurwicz-Uzawa [2], and we prove its convergence. This algorithm paves a way to handling the problems in which the network plays a min-max game against an exogenous signal, which is discussed in Section 5. In Section 6 we present an example of min-max power allocation in a cellular network and provide simulation results for both the exact penalty approach and the primal-dual approach. We conclude in Section 7.

Notation: The set of real numbers is denoted by $\mathbb{R}$, while non-negative real numbers are denoted by $\mathbb{R}_+$. All vectors are viewed as columns, where the j^{th} component of a vector x is denoted by x_j. We use the symbol $\langle x,y\rangle$ to denote the inner-product between two vectors x and y. We write $\mathbf{1}$ for the vector with each component equal to 1. A vector π is *stochastic* if $\pi_i \geq 0$ for all i and $\sum_i \pi_i = 1$.

For an $m \times m$ matrix A, we use A_{ij} or $[A]_{ij}$ to denote its entry in the i^{th} row and j^{th} column. An $m \times m$ matrix W is *stochastic* if $W_{ij} \geq 0$ for all i, j, and $W\mathbf{1} = \mathbf{1}$. A stochastic matrix W is *doubly stochastic* if it satisfies $\mathbf{1}^T W = \mathbf{1}$. Given a directed graph $G = (V,E)$, the link $(i, j) \in E$ is to be interpreted as the incoming edge from j to i. For a bidirectional graph G, we have $(i, j) \in E$ if and only if $(j,i) \in E$. We will sometimes denote the edge set of a graph G as $\mathscr{E}(G)$. We use the terms "agent" and "node" interchangeably. We say that agent j is a *neighbor* of agent i if $(i, j) \in E$, and we denote the set of all neighbors of agent i by N_i. When the edge set is time varying, we use $N_i(t)$ to denote the neighbors of agent i at time t. Given a logical statement $p(x)$ that is predicated on a variable x, we use $\mathbf{1}_{\{p(x)\}}$ to denote the indicator function which takes value 1 when $p(x)$ is $true$ and value 0 when $p(x)$ is $false$.

2 Problem Formulation

We consider a system of m computational agents, which is viewed as the node set $V = \{1,\ldots,m\}$. We assume that the time is discrete and use $k = 0, 1, \ldots$ to denote the time instances. The agents communicate with each other over a time-varying communication network. At any time k, the communications among the agents are represented by a directed graph $G(k) = (V,E(k))$ with an edge-set $E(k)$ that has a link $(i, j) \in E(k)$ if and only if agent i receives information from agent j at time k.

Let each agent i have a cost function f_i, which is known only to that agent. Consider a distributed multi-agent optimization problem subject to local agent communications, where the agents want to cooperatively solve the following problem:

$$\min_{x\in X} \max_{i\in V} f_i(x). \tag{1}$$

Here, each $f_i : \mathbb{R}^n \to \mathbb{R}$ is a *convex function*, representing a local objective function known only by agent i. The set $X \subseteq \mathbb{R}^n$ is a *closed and convex set* known by all agents. Throughout the paper, *we assume that the problem has a nonempty optimal set*, which is denoted by X^*.

We assume that $\mathbb{R}^n$ is equipped with some norm $\|\cdot\|$, with a corresponding dual norm $\|\cdot\|_*$. The goal is to develop a *distributed algorithm* for solving the constrained optimization problem in (1), while obeying the network connectivity structure and local information exchange among the neighboring agents.

The min-max-problem in (1) is a convex problem, as the function $f(x) = \max_{i\in V} f_i(x)$ is convex since point-wise maximum of convex functions preserves convexity ([4], Proposition 1.2.4, page 30). We are interested in the case when the agents' objective functions f_i are not necessarily differentiable. We also allow the local objective functions f_i to take the form of the following stochastic optimization:

$$f_i(x) = \mathbb{E}_{\omega_i}[F_i(x,\omega_i)] + \Omega_i(x), \tag{2}$$

where the expectation is taken with respect to the distribution of a random variable ω_i. The term $\Omega_i(x)$ is a regularization term that may be included to improve the generalization ability [15], or to enforce sparsity on solutions.

Min-max problem (1) does not lend itself to distributed computations that obey the local connectivity of the agents in the network. We find it useful to use the epigraph representation of the problem. In particular, we let $\eta \in \mathbb{R}$ and we re-cast problem (1) in an equivalent form:

$$\begin{array}{ll} \text{minimize} & \eta \\ \text{subject to} & f_i(x) \leq \eta \quad \text{for all } x \in X,\ \eta \in \mathbb{R}, \text{ and } i \in V, \end{array} \tag{3}$$

where x and η are variables. We use η^* to denote the optimal value of problem (3).

To distributedly solve the min-max problem, we provide two algorithms aimed at solving its epigraph formulation in (3). The first algorithm is based on an exact penalty function approach and the second algorithm is based on a primal-dual approach, where both algorithms employ Bregman-distance functions.

Before proceeding with the algorithmic development, we provide some basics of a Bregman-distance function as introduced by Bregman [10] (also, see for example [11]). Let $\mathbb{R}^n$ be equipped with some norm $\|\cdot\|$, whose dual norm is $\|x\|_* = \sup_{\|y\|\leq 1}\langle y,x\rangle$. Let $X \subseteq \mathbb{R}^n$ be a convex set and $\omega : X \to \mathbb{R}$ be a differentiable convex function over X. The function ω is *strongly convex* with a parameter $\sigma > 0$ (with respect to the norm $\|\cdot\|$), if it satisfies

$$\omega(y) - [\omega(x) + \langle \nabla\omega(x), y-x\rangle] \geq \frac{\sigma}{2}\|x-y\|^2 \quad \text{for all } x \in X^\circ \text{ and } y \in X,$$

where X° denotes the relative interior of the set X. Alternatively, ω is strongly convex over X with a parameter $\sigma > 0$ if there holds:

$$\langle \nabla\omega(x) - \nabla\omega(y), x-y\rangle \geq \sigma\|x-y\|^2 \quad \text{for all } x,y \in X^\circ.$$

Given a strongly convex function ω, we can define the *Bregman-distance* function $B : X \times X^\circ \to \mathbb{R}_+$ induced by ω:

$$B(y,x) = \omega(y) - [\omega(x) + \langle \nabla\omega(x), y - x\rangle].$$

This function is also referred to as the *prox-function* induced by ω. For the function B, by the convexity of ω we have

$$B(y,x) \geq 0 \qquad \text{for all } y \in X \text{ and } x \in X^\circ.$$

Furthermore, by the strong convexity of ω, for every $x \in X^\circ$, *the function $B(\cdot,x)$ is strongly convex over X with the same parameter σ*. The Bregman function $B(y,x)$ is used to define a nonlinear projection operator associated with the given set X, also known as the *prox-operator* [26], as follows:

$$P_X(d,x) = \operatorname{argmin}_{z\in X} \{\langle d, z - x\rangle + B(z,x)\}.$$

3 Exact Penalty Function Approach

We further transform the problem in (3) by penalizing the constraints to obtain the following problem:

$$\min_{x\in X,\eta\in\mathbb{R}} \ \eta + \sum_{i=1}^{m} r_i g_i(x,\eta), \tag{4}$$

where each g_i is a penalty function given by

$$g_i(x,\eta) = \max\{0, f_i(x) - \eta\}$$

and $r_i > 0$ is a penalty parameter for violating the constraint $f_i(x) \leq \eta$. Under certain conditions [3], the solutions of the penalized problem (4) are also the solutions of the constrained problem (3). Specifically, these conditions involve the Lagrangian dual problem associated with problem (3). We introduce the Lagrangian function:

$$L(x,\eta,\mu) = \eta + \sum_{i=1}^{m} \mu_i (f_i(x) - \eta), \tag{5}$$

where $\mu = (\mu_1, \ldots, \mu_m)'$ is the vector of dual variables satisfying $\mu_i \geq 0$ for all $i \in V$. The dual problem is

$$\max_{\mu\geq 0} q(\mu) \qquad \text{with} \quad q(\mu) = \inf_{x\in X,\eta\in\mathbb{R}} L(x,\eta,\mu). \tag{6}$$

It can be verified that the Slater condition is satisfied for problem (3) and, hence, there is no duality gap between the primal problem (3) and its dual (6). Furthermore, the set of dual optimal solutions is nonempty and bounded. The bound for the dual optimal variables can be found by rewriting the Lagrangian function (5) as follows:

$$L(x,\eta,\mu) = \left(1 - \sum_{i=1}^{m} \mu_i\right)\eta + \sum_{i=1}^{m} \mu_i f_i(x).$$

Thus, $\inf_{\eta\in\mathbb{R}} L(x,\eta,\mu) = -\infty$ when $\sum_{i=1}^{m}\mu_i \neq 1$, implying that $q(\mu) = -\infty$ whenever $\sum_{i=1}^{m}\mu_i \neq 1$. Therefore, the domain of the dual function q is the set of multipliers $\mu \geq 0$ such that $\sum_{i=1}^{m}\mu_i = 1$, showing that optimal multipliers μ_i^* must satisfy

$$\sum_{i=1}^{m} \mu_i^* = 1. \tag{7}$$

Hence, according to [3], when the penalty parameters satisfy $r_i > 1$ for all i, then the problems in (4) and (3) are equivalent.

Penalized problem (4) has a form suitable for distributed computations among the agents, as its objective can be written as $\sum_{i=1}^{m}(\eta/m + r_i g_i(x,\eta))$ and the function $\tilde{F}_i(x,\eta) = \eta/m + r_i g_i(x,\eta)$ can be interpreted as an objective function associated with agent i. In this setting, agent i is the only agent that knows the function f_i and, therefore, this agent is the only agent that deals with the dual variable μ_i associated with the constraint $f_i(x) \leq \eta$. Furthermore, observe that there is no need for any coordination of the penalty values r_i among the agents, as each agent just needs to choose its individual penalty $r_i > 1$.

3.1 Equivalence Between Epigraph and Penalty Formulations

We establish an important relation between the optimal solutions of the epigraph formulation (3) of the min-max problem and its penalized counterpart (4). The proof of this relation is basically along the lines of the work in [3], but somewhat shorter as it exploits the special structure of the epigraph formulation (3). Furthermore, the relation is important in our further development and it is not readily available.

To prove the result, we use the saddle-point theorem characterizing the optimal solutions of problem (3) and its dual problem (6), as given for example in [4], Proposition 6.2.4, page 360. The theorem is adjusted to the specific form of our Lagrangian function.

Theorem 1. *The pair (z^*,μ^*) with $z^* = (x^*,\eta^*) \in X\times\mathbb{R}$ and $\mu^* \geq 0$ is a saddle-point of the Lagrangian function $L(z,\mu)$ (i.e., primal-dual optimal pair) if and only if the following relation holds:*

$$L(z^*,\mu) \leq L(z^*,\mu^*) \leq L(z,\mu^*) \qquad \text{for all } z = (x,\eta) \in X\times\mathbb{R} \text{ and } \mu \geq 0.$$

Now, we have the following lemma.

Lemma 1. *Let $\eta^* = \min_{x\in X}\max_{i\in V} f_i(x)$ and $r_i > 1$ for all i. Then, for $g_i(x,\eta) = \max\{0, f_i(x) - \eta\}$ we have*

$$\sum_{i=1}^{m} r_i g_i(x,\eta) + \eta \geq \eta^* \qquad \textit{for all } x \in X \textit{ and } \eta \in \mathbb{R}.$$

Furthermore, the above inequality holds as equality if and only if $\eta = \eta^*$ *and* $x = x^*$ *for an optimal solution* x^* *of the problem* $\min_{x\in X} \max_i f_i(x)$.

Proof. Consider the definition of the Lagrangian in (5). For a given $x \in X$ and η, let us define the dual variables μ_i such that $\mu_i = r_i$ if $f_i(x) - \eta \geq 0$, and $\mu_i = 0$ if $f_i(x) - \eta < 0$, or compactly $\mu_i = r_i \mathbf{1}_{\{f_i(x)\geq\eta\}}$. Then, we have

$$\sum_{i=1}^{m} r_i \max\{f_i(x) - \eta, 0\} + \eta = \sum_{i=1}^{m} \mu_i (f_i(x) - \eta) + \eta = L(z,\mu).$$

Furthermore, we have $L(z^*,\mu^*) = \eta^*$. Thus, we need to prove that $L(z,\mu) - L(z^*,\mu^*) \geq 0$. For this we use Theorem 1, from which we obtain $-L(z^*,\mu^*) \geq -L(z,\mu^*)$, implying

$$\begin{aligned} L(z,\mu) - L(z^*,\mu^*) &\geq L(z,\mu) - L(z,\mu^*) = \sum_{i=1}^{m} (\mu_i - \mu_i^*)(f_i(x) - \eta) \\ &= \sum_{i=1}^{m} (r_i - \mu_i^*)\mathbf{1}_{\{f_i(x)\geq\eta\}}(f_i(x) - \eta) - \sum_{i=1}^{m} \mu_i^* \mathbf{1}_{\{f_i(x)<\eta\}}(f_i(x) - \eta) \geq 0, \end{aligned}$$

where we have used the decomposition $\mu_i^* = \mu_i^* \mathbf{1}_{\{f_i(x)\geq\eta\}} + \mu_i^* \mathbf{1}_{\{f_i(x)<\eta\}}$, and relations $r_i > 1$ and $1 \geq \mu_i^* \geq 0$ for all i (see (7)).

We next show that the preceding inequality holds as equality if and only if $\eta = \eta^*$ and $x = x^*$ for some $x^* \in X^*$. By the definition of min-max solution, if x^* solves the problem then we have $f_i(x^*) \leq \eta^*$ for all i, implying

$$\sum_{i=1}^{m} r_i \max\{f_i(x^*) - \eta^*, 0\} + \eta^* = \eta^*.$$

Thus, we just need to prove the "only if" part. Assume that for some $\bar{x} \in X$ and $\bar{\eta}$,

$$\sum_{i=1}^{m} r_i \max\{f_i(\bar{x}) - \bar{\eta}, 0\} + \bar{\eta} = \eta^*. \tag{8}$$

Since $\sum_{i=1}^{m} r_i \max\{f_i(\bar{x}) - \bar{\eta}, 0\} \geq 0$, it follows $\bar{\eta} \leq \eta^*$. Let us assume that $\bar{\eta} < \eta^*$. Then, for the equality to hold we must have $f_j(\bar{x}) > \bar{\eta}$ for some j. Thus, $f_{i^*}(\bar{x}) > \bar{\eta}$ for $i^* = \operatorname{argmax}_i f_i(\bar{x})$. By $\eta^* = \min_{x\in X} \max_{i\in V} f_i(x)$ we have $f_{i^*}(\bar{x}) \geq \eta^*$ implying $f_{i^*}(\bar{x}) - \bar{\eta} \geq \eta^* - \bar{\eta} > 0$. Since $r_{i^*} > 1$, it follows that $r_{i^*}(f_{i^*}(\bar{x}) - \bar{\eta}) > \eta^* - \bar{\eta}$. Therefore,

$$\sum_{i=1}^{m} r_i \max\{f_i(\bar{x}) - \bar{\eta}, 0\} + \bar{\eta} > \eta^*,$$

which contradicts (8). Hence, we must have $\bar{\eta} = \eta^*$ in (8). This further yields $\sum_{i=1}^m r_i \max\{f_i(\bar{x}) - \eta^*, 0\} = 0$, which by $r_i > 0$ implies $f_i(\bar{x}) \leq \eta^*$ for all i, thus showing that $\bar{x}$ is a min-max solution.

3.2 Penalty-Based Algorithm

Here, we present a distributed multi-agent algorithm for solving the penalty reformulation (4) of the min-max problem, where a penalty function $g_i(x,\eta)$ is associated with an agent i and $r_i > 1$ for all $i \in V$. Let x_k^j and η_k^j be the decision variables of agent j at time k, which are agent j estimates of an optimal solution x^* and the optimal value η^* of the problem, respectively. Recall that the agents' communications at time k are represented with a graph $G(k) = (V, E(k))$, where $(i,j) \in E(k)$ if agent i receives estimates $x^j(k)$ and $\eta^j(k)$ from agent j. To capture this information exchange, we let $N_i(k)$ denote the set of neighbors of agent i, i.e., $N_i(k) = \{j \in V \mid (i,j) \in E(k)\}$. Let us introduce a strongly convex function $\omega_x : X \to \mathbb{R}$, with a parameter $\sigma_x > 0$. We assume that $\mathbb{R}$ is equipped with square norm, i.e., $r \mapsto \frac{1}{2}r^2$, and we introduce a scalar function $\omega_\eta : \mathbb{R} \to \mathbb{R}$ that is strongly convex with respect to this norm, with a parameter $\sigma_\eta > 0$. Let us denote the Bregman-distance functions generated by these strongly convex functions by $B_x(\cdot,\cdot)$, and $B_\eta(\cdot,\cdot)$ respectively, i.e.,

$$B_x(y,u) = \omega_x(y) - [\omega_x(u) + \langle \nabla \omega_x(u), y-u\rangle],$$
$$B_\eta(\phi,\zeta) = \omega_\eta(\phi) - [\omega_\eta(\zeta) + \omega_\eta'(\zeta)(\phi-\zeta)],$$

where $\omega_\eta'(\zeta)$ denotes the derivative of ω_η at ζ. Upon receiving the estimates x_k^j and η_k^j from its neighbors, each agent i performs an intermittent adjustment of its estimates as follows:

$$\begin{bmatrix} \tilde{x}_k^i \\ \tilde{\eta}_k^i \end{bmatrix} = \sum_{j \in N_i(k)} w_{ij}(k) \begin{bmatrix} x_k^j \\ \eta_k^j \end{bmatrix}, \tag{9}$$

where $w_{ij}(k) \geq 0$ is a weight that agent i assigns to its neighbor $j \in N_i(k)$.

For a compact representation of relation (9), let $w_{ij}(k) = 0$ for all $j \notin N_i(k)$ and introduce a matrix W_k with entries $w_{ij}(k)$. With this notation, the intermittent adjustment in (9) can be written as follows:

$$z_k^i = \begin{bmatrix} \tilde{x}_k^i \\ \tilde{\eta}_k^i \end{bmatrix} = \sum_{j=1}^m [W_k]_{ij} \begin{bmatrix} x_k^j \\ \eta_k^j \end{bmatrix}. \tag{10}$$

After the intermittent adjustment, each agent i takes a step toward minimizing its own penalty function through an adjustment of the following form: for all $i \in V$,

$$x_{k+1}^i = \operatorname{argmin}_{y\in X}\left[\alpha_k r_i\langle \nabla_x g_i(\tilde{z}_k^i)+\varepsilon_k^i, y\rangle + B_x(y,\tilde{x}_k^i)\right],$$
$$\eta_{k+1}^i = \operatorname{argmin}_{s\in\mathbb{R}}\left[\alpha_k\left(\frac{1}{m}+r_i\nabla_\eta g_i(\tilde{z}_k^i)\right)s + B_\eta(s,\tilde{\eta}_k^i)\right], \tag{11}$$

where $r_i > 1$ and $\alpha_k > 0$ is a step size. The notation $\nabla_x g_i(x,\eta)$ denotes a subgradient of g with respect to x, i.e., the term $\nabla f_i(x)\mathbf{1}_{\{f_i(x)\geq\eta\}}$ where we use $\nabla f_i(x)$ to denote a subgradient of f_i at x. Similarly, $\nabla_\eta g_i(x,\eta)$ denotes the partial derivative of g with respect to η i.e., $\nabla_\eta g_i(x,\eta) = -\mathbf{1}_{\{f_i(x)\geq\eta\}}$.

If $\mathbb{R}^n$ is equipped with the Euclidean norm, and the Bregman distance functions are chosen as $\omega_x(y) = \frac{1}{2}\|y\|^2$ and $\omega_\eta(\zeta) = \frac{1}{2}\zeta^2$, then algorithm (10)–(11) reduces to the standard subgradient-projection method:

$$\begin{bmatrix} x_{k+1}^i \\ \eta_{k+1}^i \end{bmatrix} = \Pi_{X\times\mathbb{R}}\left[\tilde{z}_k^i - \alpha_k r_i\left(\nabla g_i(\tilde{z}_k^i) + \begin{bmatrix}\varepsilon_k^i \\ 0\end{bmatrix}\right)\right],$$

where Π_K stands for the Euclidean projection on a set K.

Let us now take a closer look at the first update relation in (11). This update involves taking a step along an erroneous subgradient of g_i at point $\tilde{z}_k^i$, i.e., the direction $\nabla_x g_i(\tilde{z}_k^i)+\varepsilon_k^i$ where ε_k^i is a subgradient error. The agent i objective function $g_i(x,\eta) = \max\{f_i(x)-\eta, 0\}$ is not differentiable at the point (x,η) where $f_i(x) - \eta = 0$. At such a point, a subgradient of the function g_i at (x,η) exists since each function f_i is assumed to be convex over the entire space ([4], Proposition 4.2.1). A subgradient of g at such a point is given by

$$\nabla g_i(x,\eta) = \begin{bmatrix}\nabla f_i(x) \\ -1\end{bmatrix}\mathbf{1}_{\{f_i(x)\geq\eta\}}, \tag{12}$$

where $\nabla f_i(x)$ denotes a subgradient of $f(x)$. Thus, the function g_i also has a nonempty subdifferential set at any point (x,η).

The subgradient error ε_k^i in algorithm (10)–(11) is assumed to be stochastic in order to address a general form of the objective function, as in (2), where the subgradient $\nabla f_i(x)$ is not readily available to us. We adopt a standard approach in stochastic optimization by using an unbiased estimate $\nabla f_i(x)+\bar{\varepsilon}_k^i$ of the subgradient, where $\bar{\varepsilon}_k^i$ is a zero mean random variable. Thus, in (11) we have

$$\varepsilon_k^i = \bar{\varepsilon}_k^i\mathbf{1}_{\{f_i(x_k^i)\geq\tilde{\eta}_k^i\}}.$$

The initial points $x_0^i \in X$ and η_0^i, for $i \in V$, may be selected randomly with a distribution independent of any other sources of randomness in the algorithm.

3.3 Assumptions

Our assumptions on the network are the same as, for example, those in [29, 32].

Assumption 1 *For the weight matrices and the communication graphs, we assume the following:*

(a)[Weights Rule] There exists a scalar $0 < \gamma < 1$ *such that* $[W_k]_{ii} \geq \gamma$ *for all* i *and* k*, and* $[W_k]_{ij} \geq \gamma$ *whenever* $[W_k]_{ij} > 0$*.*

(b)[Doubly Stochasticity] The matrix W_k *is doubly stochastic for all* k*, i.e.,* $W_k\mathbf{1} = \mathbf{1}$*, and* $\mathbf{1}'W_k = \mathbf{1}'$*.*

(c)[Connectedness] There exists an integer $Q \geq 1$ *such that the graph with the vertex set* V *and the edge set* $\cup_{\tau=kQ}^{(k+1)Q-1} E(\tau)$ *is strongly connected for every* k*.*

The assumptions ensure that agent's local variables are properly diffused over time-varying communication networks.

Next, we impose the following assumptions on the subgradients $\nabla f_i(x)$ and the errors, where we use $\partial f_i(x)$ to denote the set of all subgradients of f_i at x.

Assumption 2 *Let the following hold:*

(a) The subgradients of each f_i *are bounded over the set* X*, i.e., there is a scalar* $C > 0$ *such that* $\|\nabla f_i(x)\|_* \leq C$ *for all* $\nabla f_i(x) \in \partial f_i(x)$*, all* $x \in X$ *and all* $i \in V$*.*

(b) The subgradient errors $\bar{\varepsilon}_k^i$ *when conditioned on the point* $x = \tilde{x}_k^i$ *of the subgradient* $\nabla f_i(x)$ *evaluation are zero mean, i.e.,* $\mathbb{E}[\bar{\varepsilon}_k^i \mid \tilde{x}_k^i] = 0$ *for all* $i \in V$ *and* $k \geq 0$*, with probability 1.*

(c) There is a scalar $\nu > 0$ *such that* $\mathbb{E}[\|\bar{\varepsilon}_k^i\|_*^2 \mid \tilde{x}_k^i] \leq \nu^2$ *for all* $i \in V$ *and* $k \geq 0$*, with probability 1.*

Basically, under Assumptions 2-b and 2-c, the iterations $\{x_k^i\}$, $i \in V$, of the algorithm in (10)–(11) form a Markov process. In what follows, we will use F_k to denote the past iterates of the algorithm (11), i.e.,

$$F_k = \{x_t^i, \eta_t^i, i \in V, t = 0, 1, \ldots, k\} \qquad \text{for } k \geq 0.$$

Note that, given F_k, the iterates $\tilde{x}_k^i$ and $\tilde{\eta}_k^i$ in (10) are deterministic. In view of this, as a consequence of the subgradient norm and subgradient error boundedness (Assumptions 2-a and 2-c), it can be seen that with probability 1,

$$\mathbb{E}\left[\|\nabla f_i(x) + \bar{\varepsilon}_k^i\|_*^2 \mid F_k\right] \leq (C + \nu)^2 \qquad \text{for all } i \in V \text{ and } k \geq 0.$$

Also, as a result of Assumption 2-a, we have

$$\|\nabla_x g_i(\tilde{z}_k^i)\|_* = \|\nabla f_i(\tilde{x}_k^i)\|_* \mathbf{1}_{\{f_i(\tilde{x}_k^i) \geq \tilde{\eta}_k^i\}} \leq C \quad \text{for all } i \in V \text{ and } k \geq 0.$$

This and the zero-mean error assumption (Assumption 2-b) yield

$$\mathbb{E}[\varepsilon_k^i \,|\, F_k] = \mathbb{E}[\bar{\varepsilon}_k^i \mathbf{1}_{\{f_i(\tilde{x}_k^i) \geq \tilde{\eta}_k^i\}} \,|\, F_k] = \mathbb{E}\left[\bar{\varepsilon}_k^i \,|\, F_k\right] \mathbf{1}_{\{f_i(\tilde{x}_k^i) \geq \tilde{\eta}_k^i\}} = 0.$$

Similarly, as a result of Assumption 2-c we have with probability 1,

$$\mathbb{E}[\|\varepsilon_k^i\|_*^2 \,|\, F_k] \;=\; \mathbb{E}[\|\bar{\varepsilon}_k^i\|_*^2 \mathbf{1}_{\{f_i(\tilde{x}_k^i) \geq \tilde{\eta}_k^i\}} \,|\, F_k] \;=\; \mathbb{E}\left[\|\bar{\varepsilon}_k^i\|_*^2 \,|\, F_k\right] \mathbf{1}_{\{f_i(\tilde{x}_k^i) \geq \tilde{\eta}_k^i\}} \;\leq\; \nu^2.$$

This, in turn implies that with probability 1 for all $i \in V$ and $k \geq 0$,

$$\mathbb{E}\left[\|\nabla_x g_i(\tilde{z}_k^i) + \varepsilon_k^i\|_*^2 \mid F_k\right] \leq (C+\nu)^2. \tag{13}$$

Applying Jensen's inequality, we find that

$$\mathbb{E}\left[\|\nabla_x g_i(\tilde{z}_k^i) + \varepsilon_k^i\|_* \mid F_k\right] \leq C+\nu. \tag{14}$$

By the definition, a Bregman function is convex in its first variable. We further make the assumption on the choice of Bregman-distance functions that requires convexity with respect to the second variable. We depend on this assumption when showing the convergence of our algorithm.

Assumption 3 *Both Bregman-distance functions $B_x(y,z)$ and $B_\eta(\phi,\zeta)$ are convex in their second arguments z and ζ, respectively, for every fixed y and ϕ.*

3.4 Algorithm Convergence

In this section we prove the convergence of algorithm (10)–(11). We use techniques from Lyapunov analysis and the following generalization of the supermartingale convergence theorem, which is also known as the Robbins-Siegmund result as it originated in the work of Robbins and Siegmund [33].

Theorem 2. *Let F_t, $t = 0,1,2,\ldots$, be a filtration such that $F_t \subset F_{t+1}$ for $t \geq 0$. Let $\{X_t\}$, $\{Y_t\}$, $\{Z_t\}$ and $\{g_t\}$ be sequences of non-negativerandom variables that are adapted to the filtration F_t. Assume that for each t, we have with probability 1,*

$$\mathbb{E}[Y_{t+1}|F_t] \leq (1+g_t)Y_t - X_t + Z_t,$$

where $\sum_{t=0}^{\infty} Z_t < \infty$ and $\sum_{t=0}^{\infty} g_t < \infty$ with probability 1. Then, with probability 1, $\sum_{t=0}^{\infty} X_t < \infty$ and the sequence Y_t converges to a nonnegative random variable Y.

Our convergence analysis of algorithm (10)–(11) rests on Theorem 2. In order to use this theorem, we establish two main properties of the algorithm showing that the conditions of the theorem are satisfied. We develop these properties in forthcoming Lemmas 4 and 5. In the development, we make use of an alternative representation of the algorithm that relies on transition matrices, defined as follows:

$$\Phi(k,s) = W_k W_{k-1} \cdots W_s \qquad \text{for all } k,s \text{ with } k \geq s \geq 0. \tag{15}$$

We next state a result from [22] (Corollary 1) on the convergence properties of the matrix $\Phi(k,s)$.

Lemma 2. *Let Assumptions 1 hold. Then, we have $\left|[\Phi(k,s)]_{ij} - \frac{1}{m}\right| \leq \theta\beta^{k-s}$ for all $i,j \in V$ and all $k \geq s \geq 0$, with $\theta = \left(1 - \frac{\eta}{4m^2}\right)^{-2}$ and $\beta = \left(1 - \frac{\eta}{4m^2}\right)^{\frac{1}{Q}}$.*

We will also make use of the following result.

Lemma 3. *Let $\{\gamma_k\}$ be a non-negative scalar sequence such that $\sum_k \gamma_k < \infty$. Then, for any β with $0 < \beta < 1$, we have $\sum_{k=0}^{\infty} \left(\sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell\right) < \infty$.*

Proof. Let $\sum_{k=0}^{\infty} \gamma_k < \infty$. For any integer $M \geq 1$, we have $\sum_{k=0}^{M} \left(\sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell\right) = \sum_{\ell=0}^{M} \gamma_\ell \sum_{t=0}^{M-\ell} \beta^t \leq \sum_{\ell=0}^{M} \gamma_\ell \frac{1}{1-\beta}$, implying $\sum_{k=0}^{\infty} \left(\sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell\right) \leq \frac{1}{1-\beta} \sum_{\ell=0}^{\infty} \gamma_\ell < \infty$. □

In our analysis, we use auxiliary points, namely the instantaneous averages of the iterates x_k^i and η_k^i over $i \in V$, defined by

$$\hat{x}_k = \frac{1}{m}\sum_{i=1}^{m} x_k^i, \qquad \hat{\eta}_k = \frac{1}{m}\sum_{i=1}^{m} \eta_k^i \qquad \text{for all } k \geq 0.$$

We next provide an important result for these averages that we use to assert the convergence properties of our algorithm.

Lemma 4. *Let Assumptions 1 and 2 hold, and let $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$. Then, for the iterates of algorithm* (10)–(11) *we have $\sum_{k=0}^{\infty} \alpha_k \|\hat{x}_k - x_k^i\| < \infty$ and $\sum_{k=0}^{\infty} \alpha_k |\hat{\eta}_k - \eta_k^i| < \infty$ for all $i \in V$, with probability 1.*

Proof. Let us denote the noisy subgradient as $\tilde{d}_k^i = \nabla_x g_i(\tilde{z}_k^i) + \varepsilon_k^i$. Applying the optimality condition for (11), we get

$$\langle \alpha_k r_i \tilde{d}_k^i + \nabla\omega_x(x_{k+1}^i) - \nabla\omega_x(\tilde{x}_k^i), y - x_{k+1}^i\rangle \geq 0 \quad \text{for all } y \in X.$$

Since $\tilde{x}_k^i \in X$, by letting $y = \tilde{x}_k^i$ we have

$$\langle \alpha_k r_i \tilde{d}_k^i + \nabla\omega_x(x_{k+1}^i) - \nabla\omega_x(\tilde{x}_k^i), \tilde{x}_k^i - x_{k+1}^i\rangle \geq 0,$$

which implies

$$\alpha_k r_i \langle \tilde{d}_k^i, \tilde{x}_k^i - x_{k+1}^i\rangle \geq \langle \nabla\omega_x(\tilde{x}_k^i) - \nabla\omega_x(x_{k+1}^i), \tilde{x}_k^i - x_{k+1}^i\rangle \geq \sigma_x \|\tilde{x}_k^i - x_{k+1}^i\|^2,$$

where the last inequality follows by the strong convexity of ω_x. Using Hölder's inequality, we obtain

$$\alpha_k r_i \|\tilde{d}_k^i\|_* \|\tilde{x}_k^i - x_{k+1}^i\| \geq \sigma_x \|\tilde{x}_k^i - x_{k+1}^i\|^2.$$

Therefore $\|\tilde{x}_k^i - x_{k+1}^i\| \leq \alpha_k r_i \frac{\|\tilde{d}_k^i\|_*}{\sigma_x}$, and taking the conditional expectation yields

$$\mathbb{E}\left[\|\tilde{x}_k^i - x_{k+1}^i\| \,|\, F_k\right] \leq \alpha_k r_i \frac{\mathbb{E}\left[\|\tilde{d}_k^i\|_* \,|\, F_k\right]}{\sigma_x} \leq \alpha_k r_i \frac{C+\nu}{\sigma_x}, \tag{16}$$

where the last inequality follows from (14) under Assumptions 2-a and 2-c.

Let us now write the iterates as follows

$$x_{k+1}^i = \tilde{x}_k^i + e_k^i \qquad \text{with} \quad e_k^i = x_{k+1}^i - \tilde{x}_k^i. \tag{17}$$

By (16) and the iterated expectation rule, for e_k^i we obtain

$$\mathbb{E}[\|e_k^i\|] = \mathbb{E}\left[\mathbb{E}[\|e_k^i\| \mid F_k]\right] \leq \alpha_k r_i \frac{C+\nu}{\sigma_x} \qquad \text{for all } i \text{ and } k. \tag{18}$$

By taking the average of the relations in (17) over $i = 1, \ldots, m$ and using the doubly stochastic property of W_k, we can see that for all k,

$$\hat{x}_{k+1} = \hat{x}_k + \frac{1}{m}\sum_{i=1}^m e_k^i. \tag{19}$$

Now, we note that by (17) and the definition of $\tilde{x}_k^i$, we have $x_{k+1}^i = \sum_{j=1}^m [W_k]_{ij} x_k^j + e_k^i$. We then recursively use this equation to relate x_{k+1}^i and x_0^j for $j \in V$. We do so by using the matrices $\Phi(k,s)$ defined in (15) to obtain

$$x_{k+1}^i = \sum_{j=1}^m [\Phi(k,0)]_{ij} x_0^j + e_k^i + \sum_{\ell=0}^{k-1}\left(\sum_{j=1}^m [\Phi(k,\ell+1)]_{ij} e_\ell^j\right). \tag{20}$$

Similarly, we derive the recursive relation for $\hat{x}_{k+1}$ as given in (19), and obtain:

$$\hat{x}_{k+1} = \hat{x}_0 + \frac{1}{m}\sum_{\ell=0}^k \sum_{j=1}^m e_\ell^j = \frac{1}{m}\sum_{j=1}^m x_0^j + \frac{1}{m}\sum_{j=1}^m e_k^j + \frac{1}{m}\sum_{\ell=0}^{k-1}\sum_{j=1}^m e_\ell^j. \tag{21}$$

Then, from (20) and (21) we have

$$\begin{aligned}\|x_{k+1}^i - \hat{x}_{k+1}\| \leq &\left\|\sum_{j=1}^m \left([\Phi(k,0)]_{ij} - \frac{1}{m}\right) x_0^j\right\| + \left\|e_k^i - \frac{1}{m}\sum_{j=1}^m e_k^j\right\| \\ &+ \left\|\sum_{\ell=0}^{k-1}\sum_{j=1}^m \left([\Phi(k,\ell+1)]_{ij} - \frac{1}{m}\right) e_\ell^j\right\|.\end{aligned}$$

Therefore, for all $j \in V$ and all k,

$$\begin{aligned}\|x_{k+1}^i - \hat{x}_{k+1}\| \leq &\sum_{j=1}^m \left|[\Phi(k,0)]_{ij} - \frac{1}{m}\right| \|x_0^j\| + \frac{1}{m}\sum_{j\neq i}\|e_k^i - e_k^j\| \\ &+ \sum_{\ell=0}^{k-1}\sum_{j=1}^m \left|[\Phi(k,\ell+1)]_{ij} - \frac{1}{m}\right| \|e_\ell^j\|.\end{aligned}$$

At this point, we use the rate of convergence result from Lemma 2 to bound $\left|[\Phi(k,\ell)]_{ij} - \frac{1}{m}\right|$. By doing so, we obtain

$$\|x_{k+1}^i - \hat{x}_{k+1}\| \leq \theta\beta^k \sum_{j=1}^m \|x_0^j\| + \frac{1}{m}\sum_{j\neq i}\|e_k^i - e_k^j\| + \theta\sum_{\ell=0}^{k-1}\sum_{j=1}^m \beta^{k-(\ell+1)}\|e_\ell^j\|,$$

where $\beta < 1$ (see Lemma 2). By taking the expectation and using (18) we find that for all i and k,

$$\mathbb{E}[\|x_{k+1}^i - \hat{x}_{k+1}\|] \leq \theta\beta^k \sum_{j=1}^m \mathbb{E}[\|x_0^j\|] + \frac{2(m-1)c}{m}\alpha_k + m\theta c \sum_{\ell=0}^{k-1} \beta^{k-(\ell+1)}\alpha_\ell,$$

with $c = (\max_{i\in V} r_i)(C+\nu)/\sigma_x$. Next, we multiply the preceding relation with α_{k+1} and after using $ab \leq (a^2+b^2)/2$ for the terms $\alpha_{k+1}\alpha_k$ and $\alpha_{k+1}\alpha_\ell$, we obtain

$$\begin{aligned}\alpha_{k+1}\mathbb{E}[\|x_{k+1}^i - \hat{x}_{k+1}\|] \leq\ & \theta\alpha_{k+1}\beta^k \sum_{j=1}^m \mathbb{E}[\|x_0^j\|] + \frac{(m-1)c}{m}(\alpha_k^2 + \alpha_{k+1}^2) \\ & + m\frac{\theta c}{2}\sum_{\ell=0}^{k-1}\beta^{k-(\ell+1)}(\alpha_{k+1}^2 + \alpha_\ell^2),\end{aligned}$$

We observe that by $\sum_{k=0}^\infty \alpha_k^2 < \infty$ it follows that $\alpha_k \to 0$, and hence α_k is bounded. This and the fact $\beta < 1$ imply $\sum_{k=0}^\infty \alpha_{k+1}\beta^k < \infty$. The sum $\sum_{k=0}^\infty(\alpha_k^2 + \alpha_{k+1}^2)$ is obviously summable when $\sum_{k=0}^\infty \alpha_k^2 < \infty$. For the last term, by $\beta < 1$ we have

$$\sum_{k=1}^\infty \sum_{\ell=0}^{k-1} \beta^{k-(\ell+1)}(\alpha_{k+1}^2 + \alpha_\ell^2) \leq \sum_{k=1}^\infty \frac{\alpha_{k+1}^2}{1-\beta} + \sum_{k=1}^\infty \sum_{\ell=0}^{k-1}\beta^{k-(\ell+1)}\alpha_\ell^2.$$

The first sum is finite since $\sum_{k=0}^\infty \alpha_k^2 < \infty$, while the second sum is finite by Lemma 3. Thus, $\sum_{k=0}^\infty \alpha_{k+1}\mathbb{E}[\|x_{k+1}^i - \hat{x}_{k+1}\|] < \infty$ for all $i \in V$ and by the monotone convergence theorem [7], it follows $\mathbb{E}\left[\sum_{k=0}^\infty \alpha_{k+1}\|x_{k+1}^i - \hat{x}_{k+1}\|\right] < \infty$. When the expected value of a positive random variable is finite, then the variable must be finite with probability 1, so we must have $\sum_{k=0}^\infty \alpha_{k+1}\|x_{k+1}^i - \hat{x}_{k+1}\| < \infty$ for all $i \in V$, with probability 1.

A similar analysis proves $\sum_{k=0}^\infty \alpha_k|\hat{\eta}_k - \eta_k^i| < \infty$ for all $i \in V$, with probability 1. □

Lemma 4 provides one important property of the iterates generated by our algorithm. We next provide another important relation that in a way captures a descent property of the iterates in terms of the Lyapunov function given by the sum of the Bregman-distance functions B_x and B_η. For notational convenience, we define

$$z_k^i = \begin{bmatrix} x_k^i \\ \eta_k^i \end{bmatrix}, \quad \hat{z}_k = \begin{bmatrix} \hat{x}_k \\ \hat{\eta}_k \end{bmatrix} \qquad \text{for all } k \geq 0. \tag{22}$$

Also, we introduce the notation $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot\,|\,F_k]$.

We have the following relation for algorithm (10)–(11), under the assumption that the Bregman-distance functions B_x and B_η are convex in their second arguments.

Lemma 5. *Let Assumptions 1, 2 and 3 hold. Then, for algorithm* (10)–(11) *we have with probability 1 for any $x^* \in X^*$ and all $k \geq 0$,*

$$\sum_{i=1}^{m}\mathbb{E}_k\left[B_x(x^*,x_{k+1}^i)+B_\eta(\eta^*,\eta_{k+1}^i)\right]\le\sum_{i=1}^{m}\left(B_x(x^*,x_k^i)+B_\eta(\eta^*,\eta_k^i)\right)$$
$$-\alpha_k\left(\sum_{i=1}^{m}r_ig_i(\hat{z}_k)+\hat{\eta}_k-\eta^*\right)+\alpha_k\bar{r}\left(C\sum_{j=1}^{m}\|\hat{x}_k-x_k^j\|+\sum_{j=1}^{m}|\hat{\eta}_k-\eta_k^j|\right)$$
$$+\alpha_k^2m\left(\frac{\bar{r}^2(C+\nu)^2}{2\sigma_x}+\frac{\left(\frac{1}{m}+\bar{r}\right)^2}{2\sigma_\eta}\right),$$

where $\bar{r}=\max_{i\in V}r_i$.

Proof. By the definition of the Bregman function B_x, we have

$$B_x(x^*,x_{k+1}^i)-B_x(x^*,\tilde{x}_k^i)=w_x(\tilde{x}_k^i)-w_x(x_{k+1}^i)-\langle\nabla w_x(x_{k+1}^i),x^*-x_{k+1}^i\rangle$$
$$+\langle\nabla w_x(\tilde{x}_k^i),x^*-\tilde{x}_k^i\rangle.$$

Noting that $w_x(\tilde{x}_k^i)-w_x(x_{k+1}^i)=\langle\nabla w_x(\tilde{x}_k^i),\tilde{x}_k^i-x_{k+1}^i\rangle-B_x(x_{k+1}^i,\tilde{x}_k^i)$, we obtain

$$B_x(x^*,x_{k+1}^i)-B_x(x^*,\tilde{x}_k^i)=\langle\nabla w_x(\tilde{x}_k^i)-\nabla w_x(x_{k+1}^i),x^*-x_{k+1}^i\rangle-B_x(x_{k+1}^i,\tilde{x}_k^i). \tag{23}$$

Now, using the optimality condition for (11), since $x^*\in X$, we have $\langle\alpha_kr_i\tilde{d}_k^i+\nabla\omega_x(x_{k+1}^i)-\nabla\omega_x(\tilde{x}_k^i),x^*-x_{k+1}^i\rangle\ge 0$,where $\tilde{d}_k^i=\nabla_xg_i(\tilde{z}_k^i)+\varepsilon_k^i$. This implies

$$\langle\nabla\omega_x(\tilde{x}_k^i)-\nabla\omega_x(x_{k+1}^i),x^*-x_{k+1}^i\rangle\le\alpha_kr_i\langle\tilde{d}_k^i,x^*-x_{k+1}^i\rangle.$$

Upon substituting the preceding inequality in (23), we obtain

$$B_x(x^*,x_{k+1}^i)-B_x(x^*,\tilde{x}_k^i)\le\alpha_kr_i\langle\tilde{d}_k^i,x^*-x_{k+1}^i\rangle-B_x(x_{k+1}^i,\tilde{x}_k^i)$$
$$\le\alpha_kr_i\langle\tilde{d}_k^i,x^*-\tilde{x}_k^i\rangle+\alpha_kr_i\langle\tilde{d}_k^i,\tilde{x}_k^i-x_{k+1}^i\rangle-\frac{\sigma_x}{2}\|x_{k+1}^i-\tilde{x}_k^i\|^2. \tag{24}$$

where the last inequality follows from the strong convexity of the Bregman function, i.e., $B_x(x_{k+1}^i,\tilde{x}_k^i)\ge\frac{\sigma_x}{2}\|x_{k+1}^i-\tilde{x}_k^i\|^2$. Further, by Hölder's inequality we have $\alpha_kr_i\langle\tilde{d}_k^i,\tilde{x}_k^i-x_{k+1}^i\rangle\le\alpha_kr_i\|\tilde{d}_k^i\|_*\|\tilde{x}_k^i-x_{k+1}^i\|$. Using this and the scalar inequality $2ab\le a^2+b^2$, we obtain

$$\alpha_kr_i\langle\tilde{d}_k^i,\tilde{x}_k^i-x_{k+1}^i\rangle\le 2\,\frac{\alpha_kr_i}{\sqrt{2\sigma_x}}\|\tilde{d}_k^i\|_*\,\frac{\sqrt{\sigma_x}}{\sqrt{2}}\|\tilde{x}_k^i-x_{k+1}^i\|$$
$$\le\alpha_k^2r_i^2\,\frac{\|\tilde{d}_k^i\|_*^2}{2\sigma_x}+\frac{\sigma_x}{2}\|x_{k+1}^i-\tilde{x}_k^i\|^2. \tag{25}$$

Upon combining (25) and (24) we arrive at

$$B_x(x^*,x_{k+1}^i)-B_x(x^*,\tilde{x}_k^i)\le\alpha_kr_i\langle\tilde{d}_k^i,x^*-\tilde{x}_k^i\rangle+\alpha_k^2r_i^2\frac{\|\tilde{d}_k^i\|_*^2}{2\sigma_x}.$$

Since $\tilde{d}_k^i = \nabla_x g_i(\tilde{z}_k^i) + \varepsilon_k^i$, by taking the conditional expectation with respect to the history F_k, and by using (13) and $\mathbb{E}_k[\tilde{d}_k^i] = \nabla_x g_i(v_k^i)$, we obtain

$$\mathbb{E}_k\left[B_x(x^*, x_{k+1}^i)\right] \leq B_x(x^*, \tilde{x}_k^i) - \alpha_k r_i \langle \nabla_x g_i(\tilde{z}_k^i), \tilde{x}_k^i - x^* \rangle + \alpha_k^2 r_i^2 \frac{(C+\nu)^2}{2\sigma_x}. \quad (26)$$

Proceeding similarly, we can derive the following inequality for the iterates involving the min-max value estimates η_k^i,

$$\mathbb{E}_k\left[B_\eta(\eta^*, \eta_{k+1}^i)\right] \leq B_\eta(\eta^*, \tilde{\eta}_k^i) - \alpha_k \left(\frac{1}{m} + r_i \nabla_\eta g_i(\tilde{z}_k^i)\right)(\tilde{\eta}_k^i - \eta^*) + \alpha_k^2 \frac{\left(\frac{1}{m} + r_i\right)^2}{2\sigma_\eta}, \quad (27)$$

where we use the fact $|\nabla_\eta g_i(\tilde{z}_k^i)| \leq 1$. By the convexity of g_i and the subgradient property, we have

$$\nabla g_i(\tilde{z}_k^i)'(\tilde{z}_k^i - z^*) \geq g_i(\tilde{z}_k^i) - g_i(z^*) = g_i(\tilde{z}_k^i).$$

where the equality follows by $g_i(z^*) = \max\{0, f_i(x^*) - \eta^*\}$ and $f_i(x^*) - \eta^* \leq 0$ for all $i \in V$ and any optimal point $x^* \in X^*$. Upon adding the inequalities (26) and (27), and using the convexity of g_i, we obtain for any $x^* \in X^*$ and all $k \geq 0$ and $i \in V$,

$$\begin{aligned}\mathbb{E}_k\left[B_x(x^*, x_{k+1}^i) + B_\eta(\eta^*, \eta_{k+1}^i)\right] &\leq B_x(x^*, \tilde{x}_k^i) + B_\eta(\eta^*, \tilde{\eta}_k^i) - \alpha_k \frac{1}{m}(\tilde{\eta}_k^i - \eta^*) \\ &- \alpha_k r_i g_i(\tilde{z}_k^i) + \alpha_k^2 \left(\frac{r_i^2(C+\nu)^2}{2\sigma_x} + \frac{\left(\frac{1}{m} + r_i\right)^2}{2\sigma_\eta}\right). \quad (28)\end{aligned}$$

Now, by Assumption 3 on the convexity of the Bregman functions B_x and B_η and the doubly stochasticity of the weight matrix W_k (Assumption 1-b), we have

$$\sum_{i=1}^m B_x(x^*, \tilde{x}_k^i) = \sum_{i=1}^m B_x\left(x^*, \sum_{j=1}^m [W_k]_{ij} x_k^j\right) \leq \sum_{i=1}^m \sum_{j=1}^m [W_k]_{ij} B_x(x^*, x_k^j) = \sum_{j=1}^m B_x(x^*, x_k^j).$$

Similarly, we have $\sum_{i=1}^m B_\eta(\eta^*, \tilde{\eta}_k^i) \leq \sum_{j=1}^m B_\eta(\eta^*, \eta_k^j)$. Thus, by summing inequalities in (28) over $i \in V$ we obtain for any $x^* \in X^*$ and all $k \geq 0$,

$$\begin{aligned}&\sum_{i=1}^m \mathbb{E}_k\left[B_x(x^*, x_{k+1}^i) + B_\eta(\eta^*, \eta_{k+1}^i)\right] \leq \sum_{i=1}^m \left(B_x(x^*, x_k^i) + B_\eta(\eta^*, \eta_k^i)\right) \\ &- \alpha_k \frac{1}{m} \sum_{i=1}^m (\tilde{\eta}_k^i - \eta^*) - \alpha_k \sum_{i=1}^m r_i g_i(\tilde{z}_k^i) + \alpha_k^2 m \left(\frac{\bar{r}^2(C+\nu)^2}{2\sigma_x} + \frac{\left(\frac{1}{m} + \bar{r}\right)^2}{2\sigma_\eta}\right), \quad (29)\end{aligned}$$

where $\bar{r} = \max_{i \in V} r_i$.

From the definition of $\tilde{\eta}_k^i$ and the doubly stochasticity of the matrix W_k, we have $\sum_{i=1}^m \tilde{\eta}_k^i = \sum_{i=1}^m \sum_{j=1}^m [W_k]_{ij} \eta_k^j = \sum_{j=1}^m \eta_k^j = m\hat{\eta}_k$. Using this identity, together with

adding and subtracting the term $\alpha_k \sum_{i=1}^m r_i g_i(\hat{z}_k)$, from the preceding relation we obtain

$$\sum_{i=1}^m \mathbb{E}_k\left[B_x(x^*,x_{k+1}^i)+B_\eta(\eta^*,\eta_{k+1}^i)\right] \le \sum_{i=1}^m \left(B_x(x^*,x_k^i)+B_\eta(\eta^*,\eta_k^i)\right)$$
$$-\alpha_k\left(\sum_{i=1}^m r_i g_i(\hat{z}_k)+\hat{\eta}_k-\eta^*\right)+\alpha_k\sum_{i=1}^m r_i\left|g_i(\hat{z}_k)-g_i(\tilde{z}_k^i)\right|$$
$$+\alpha_k^2 m\left(\frac{\bar{r}^2(C+\nu)^2}{2\sigma_x}+\frac{\left(\frac{1}{m}+\bar{r}\right)^2}{2\sigma_\eta}\right), \tag{30}$$

where $\hat{z}_k$ is as defined in (22). Next, we consider the term $\sum_{i=1}^m r_i\left|g_i(\hat{z}_k)-g_i(\tilde{z}_k^i)\right|$. By the definition of g_i and relation $|\max\{a,0\}-\max\{b,0\}|\le|a-b|$ valid for any two scalars a and b, we have the following:

$$\left|g_i(\hat{z}_k)-g_i(\tilde{z}_k^i)\right| \le \left|f_i(\hat{x}_k)-f_i(\tilde{x}_k^i)-\hat{\eta}_k+\tilde{\eta}_k^i\right| \le C\|\hat{x}_k-\tilde{x}_k^i\|+\left|\hat{\eta}_k-\tilde{\eta}_k^i\right|,$$

where in the first inequality we use the definition of $\tilde{z}_k^i$ in (10), while in the last inequality we use the subgradient boundedness assumption for f_i. Further, by using the definition of the variables $\tilde{x}_k^i$ and $\tilde{\eta}_k^i$ in (10), the stochasticity of W_k and the convexity of the norm, we obtain

$$\left|g_i(\hat{z}_k)-g_i(\tilde{z}_k^i)\right| \le \sum_{j=1}^m [W_k]_{ij}\left[C\|\hat{x}_k-x_k^j\|+|\hat{\eta}_k-\eta_k^j|\right].$$

Therefore, by using the doubly stochasticity of W_k and $\bar{r}=\max r_i$, we obtain

$$\sum_{i=1}^m r_i\left|g_i(\hat{z}_k)-g_i(\tilde{z}_k^i)\right| \le \bar{r}C\sum_{j=1}^m\|\hat{x}_k-x_k^j\|+\bar{r}\sum_{j=1}^m|\hat{\eta}_k-\eta_k^j|.$$

Substituting this estimate back in (30) we get the desired result. □

We are now ready to prove our main convergence result. The result essentially states that under suitable conditions on the step size α_k, all the agents' estimates converge to a common optimal point. Moreover, the agents' estimates of the min-max value also converge to the optimal value of the problem.

Theorem 3. *Let Assumptions 1, 2, and 3 hold. Let the step sizes satisfy* $\sum_{k=0}^\infty \alpha_k=\infty$ *and* $\sum_{k=0}^\infty \alpha_k^2<\infty$. *Then, for all* $i\in V$, *the agents' iterates* x_k^i *and* η_k^i *generated by algorithm* (10)–(11) *are such that with probability 1 for all* $i\in V$*:*

(a) The decision variables x_k^i *converge to a common optimal (random) point* $x^*\in X^*$.
(b) The estimates η_k^i *converge to the optimal value* η^* *of the min-max problem.*

Proof. Our analysis is based on applying the Robbins-Siegmund result from Theorem 2 to the inequality derived in Lemma 5. By our assumption on the step sizes we have $\sum_{k=0}^\infty \alpha_k^2<\infty$, which trivially implies that

$$\sum_{k=0}^{\infty} \alpha_k^2 m \left(\frac{\bar{r}^2 (C+\nu)^2}{2\sigma_x} + \frac{\left(\frac{1}{m}+\bar{r}\right)^2}{2\sigma_\eta} \right) < \infty.$$

Furthermore, by virtue of Lemma 4 we have

$$\sum_{k=0}^{\infty} \alpha_k \bar{r} \sum_{j=1}^{m} \left(C\|\hat{x}_k - x_k^j\| + |\hat{\eta}_k - \eta_k^j| \right) < \infty. \tag{31}$$

In addition, by the definition $\hat{z}_k^i$ is a convex combination of $(x_k^i, \eta_k^i) \in X \times \mathbb{R}$ for all $i \in V$ and k, implying by Lemma 1 that $\sum_{i=1}^m r_i g_i(\hat{z}_k) + \hat{\eta}_k - \eta^* \geq 0$ for all $k \geq 0$. Thus, we can apply Lemma 2 to the relation of Lemma 5 and infer that with probability 1, $\sum_{i=1}^m \left(B_x(x^*, x_k^i) + B_\eta(\eta^*, \eta_k^i) \right)$ converges for every $x^* \in X^*$ and

$$\sum_{k=0}^{\infty} \alpha_k \left(\sum_{i=1}^{m} r_i g_i(\hat{z}_k) + \hat{\eta}_k - \eta^* \right) < \infty. \tag{32}$$

Now, since $\sum_{k=0}^{\infty} \alpha_k = \infty$, from (31)–(32) it follows that there exists a subsequence indexed by $\{k_\ell\}$ such that with probability 1,

$$\begin{aligned} &\lim_{\ell \to \infty} \left(\sum_{i=1}^{m} r_i g_i(\hat{z}_{k_\ell}) + \hat{\eta}_{k_\ell} - \eta^* \right) = 0, \\ &\lim_{\ell \to \infty} \|\hat{x}_{k_\ell} - x_{k_\ell}^j\| = 0 \ \text{ and } \ \lim_{\ell \to \infty} |\hat{\eta}_{k_\ell} - \eta_{k_\ell}^j| = 0 \ \text{ for all } j. \end{aligned} \tag{33}$$

Since $\sum_{i=1}^m \left(B_x(x^*, x_k^i) + B_\eta(\eta^*, \eta_k^i) \right)$ converges for every $x^* \in X^*$, the sequence $\sum_{i=1}^m \left(B_x(x^*, x_k^i) + B_\eta(\eta^*, \eta_k^i) \right)$ must be bounded for every $x^* \in X^*$ with probability 1. Note that from Assumption 3 on the convexity of Bregman functions B_x and B_η and their inherent strong convexity property we have

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \left(B_x(x^*, x_k^i) + B_\eta(\eta^*, \eta_k^i) \right) &\geq B_x(x^*, \hat{x}_k) + B_\eta(\eta^*, \hat{\eta}_k) \\ &\geq \frac{\sigma_x}{2} \|x^* - \hat{x}_k\|^2 + \frac{\sigma_\eta}{2} |\eta^* - \hat{\eta}_k|^2. \end{aligned}$$

Thus, the sequences $\{\hat{x}_k\}$ and $\{\hat{\eta}_k\}$ are also bounded with probability 1. Hence, along a further subsequence, which without loss of generality we can let it be indexed by the same index set $\{k_\ell, \ell = 1, 2, \ldots\}$, with probability 1 we have $\lim_{\ell \to \infty} \hat{x}_{k_\ell} = \bar{x}$ and $\lim_{\ell \to \infty} \hat{\eta}_{k\ell} = \bar{\eta}$, where $\bar{x} \in X$ since X is closed. Moreover, with probability 1 the limit points satisfy

$$\sum_{i=1}^{m} r_i g_i(\bar{x}) + \bar{\eta} - \eta^* = 0.$$

From this relation and Lemma 1 it follows that $\bar{x} \in X^*$ and $\bar{\eta} = \eta^*$ with probability 1.

In view of (33), and $\hat{x}_{k_\ell} \to \bar{x}$ and $\hat{\eta}_{k_\ell} \to \eta^*$, we further have $x^j_{k_\ell} \to \bar{x}$ and $\eta^j_{k_\ell} \to \bar{\eta}$ for all j with probability 1. Therefore, $\lim_{\ell\to\infty} \sum_{i=1}^m \left(B_x(\bar{x}, x^i_{k_\ell}) + B_\eta(\eta^*, \eta^i_{k_\ell})\right) = 0$. However, we have established that the sequence $\sum_{i=1}^m \left(B_x(x^*, x^i_k) + B_\eta(\eta^*, \eta^i_k)\right)$ converges with probability 1 for any $x^* \in X^*$, which in view of $\bar{x} \in X^*$, implies that with probability 1,

$$\lim_{k\to\infty} \sum_{i=1}^m \left(B_x(\bar{x}, x^i_k) + B_\eta(\eta^*, \eta^i_k)\right) = 0.$$

Finally, by the strong convexity of the Bregman-distance functions B_x and B_η, it follows that $x^j_k \to \bar{x}$ and $\eta^j_k \to \eta^*$ with probability 1 for all j. □

One may extend algorithm (10)–(11) to the case when each agent i uses a different Bregman function $B_{\eta,i}$ instead of a common one B_η. Following the same analysis, with a slight modification, it can be seen that the convergence result of Theorem 3 would be applicable to such a modified algorithm.

4 Primal-Dual Approach

In this section we present a distributed primal-dual algorithm which is motivated by the classical work of Arrow-Hurwicz-Uzawa [2]. Recently, a primal-dual method was studied in [24] for approximate solutions to saddle-point problems by considering standard Euclidean norm. The use of Bregman distance for a saddle-point problem was studied in [26]. The prior work is dealing with centralized problems, while here we consider a primal-dual method with Bregman distances for solving the *distributed min-max problem* in its epigraph formulation (3).

In order to apply primal-dual approach, we need to slightly modify the original epigraph formulation (3) of the min-max problem but without changing its set of optimal solutions. This is needed to ensure the convergence of the primal-dual algorithm. Specifically, we assume that we have a closed convex interval D such that $\eta^* \in D$, which is equivalent to having some (arbitrarily large) upper and lower estimates for η^*. Having such a set, we consider a modification of epigraph formulation (3), given by

$$\begin{array}{ll} \text{minimize} & \eta \\ \text{subject to} & f_i(x) \le \eta \quad \text{for all } x \in X,\ \eta \in D, \text{ and } i \in V. \end{array} \tag{34}$$

This problem has the same solutions as the original min-max problem since $\eta^* \in D$. The Lagrangian function associated with this problem is given by

$$\mathscr{L}(x, \eta, \mu) = \sum_{i=1}^m \mu_i \left(f_i(x) - \eta\right) + \eta \quad \text{for } x \in X, \eta \in D, \mu \ge 0.$$

We can restrict the domain of dual variables of the Lagrangian function to the unit interval, as we know that the dual optimal multipliers satisfy $\sum_{i=1}^{m} \mu_i^* = 1$, as seen in Section 3. Thus, we will consider the Lagrangian function with a restricted domain (yet large enough to contain all dual optimal solutions):

$$\mathscr{L}(x,\eta,\mu) = \sum_{i=1}^{m} \mu_i \left(f_i(x) - \eta\right) + \eta \quad \text{for } x \in X,\ \eta \in D,\ \mu \in I^m, \tag{35}$$

where I is the interval $[0,1]$ and I^m denotes the product of m copies of I. We are interested in determining a primal-dual optimal pair for problem (34) by an algorithm aimed at finding a saddle-point of the reduced Lagrangian (35).

4.1 Primal-Dual Algorithm

We are interested in distributed algorithm for computing a saddle-point of the Lagrangian (35). To accommodate distributed computations among m agents, we write the Lagrangian function as a sum of m functions, as follows:

$$L(x,\eta,\mu) = \sum_{i=1}^{m} \mu_i \left(f_i(x) - \eta\right) + \eta = \sum_{i=1}^{m} \left(\mu_i \left(f_i(x) - \eta\right) + \frac{1}{m}\eta\right). \tag{36}$$

Lagrangian-function component $L_i(x,\eta,\mu_i) = \mu_i \left(f_i(x) - \eta\right) + \frac{1}{m}\eta$ is assigned to agent $i \in V$ for processing without sharing the information about the function with any other agent.

The distributed algorithm will use the Bregman-distance functions B_x and B_η, as introduced in Section 3.2. In addition, we introduce another collection of Bregman-distance functions, one for each of the agents in order to handle the Lagrange multipliers for its constraint set $\{x \mid f_i(x) \leq \eta\}$. For this, for each $i \in V$, we let $B_{\mu,i}(\cdot,\cdot)$ be a Bregman-distance function associated with a strongly convex function $\omega_{\mu,i} : \mathbb{R} \to \mathbb{R}$ with parameter $\sigma_{\mu,i} > 0$.

The proposed primal-dual distributed algorithm for finding a saddle-point of the Lagrangian is as follows. At every iteration k, each agent i has estimates x_k^i, η_k^i and μ_i^k, respectively, for an optimal solution of the min-max problem, the optimal value and the Lagrangian multiplier associated with the constraint $\{x \mid f_i(x) \leq \eta\}$ that agent i is responsible for. Every agent firstly performs an intermittent adjustment of the variables x_k^i and η_k^i as in (10) to obtain the estimates $\tilde{x}_k^i$ and $\tilde{\eta}_i^k$. For convenience, we restate these updates:

$$\begin{bmatrix} \tilde{x}_k^i \\ \tilde{\eta}_k^i \end{bmatrix} = \sum_{j=1}^{m} [W_k]_{ij} \begin{bmatrix} x_k^j \\ \eta_k^j \end{bmatrix}. \tag{37}$$

This step ensures that agents locally align the variables that are coupling. We note that the agents do not have an intermittent adjustment for their multiplier estimates μ_k^i, as these variables do not couple.

Then, every agent $i \in V$ generates new iterates by taking step toward minimizing its own Lagrangian function $L_i(x,\eta,\mu_i) = \mu_i\,(f_i(x)-\eta) + \frac{1}{m}\eta$ with respect to (x,η) and maximizing it with respect to μ_i, in the following manner:

$$\begin{aligned} x_{k+1}^i &= \operatorname{argmin}_{y\in X}\left[\alpha_k \mu_k^i \langle d_k^i + \varepsilon_k^i, y\rangle + B_x(y,\tilde{x}_k^i)\right], \\ \eta_{k+1}^i &= \operatorname{argmin}_{s\in D}\left[\alpha_k (1/m-\mu_k^i)s + B_\eta(s,\tilde{\eta}_k^i)\right], \\ \mu_{k+1}^i &= \operatorname{argmin}_{\zeta\in I}\left[\alpha_k\left(\tilde{\eta}_k^i - f_i(\tilde{x}_k^i)\right)\zeta + B_{\mu,i}(\zeta,\mu_k^i)\right], \end{aligned} \tag{38}$$

where d_k^i is a subgradient of $f_i(x)$ evaluated at $\tilde{x}_k^i$ and ε_k is a random error in the subgradient evaluation. For each i, the initial values $x_0^i \in X$, $\eta_0^i \in D$, and $\mu_0^i \in I$ are random and independent from the stochastic errors ε_k.

4.2 Algorithm Convergence

The analysis is similar to that of algorithm (10)–(11) using exact-penalty approach. First, we state a lemma which relates the local iterates of the primal-dual algorithm to their respective average trajectory for the coupling variables x_k^i and η_k^i.

Lemma 6. *Let Assumptions 1 and 2 hold, and let the step sizes satisfy* $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$. *Then, for the instantaneous averages* $\hat{x}_k$ *and* $\hat{\eta}_k$ *of the iterates generated by algorithm* (37)–(38)*, we have* $\sum_{k=0}^{\infty}\alpha_k\|\hat{x}_k - x_k^i\| < \infty$ *and* $\sum_{k=0}^{\infty}\alpha_k|\hat{\eta}_k - \eta_k^i| < \infty$ *for all* $i \in V$*, with probability 1.*

Proof. The proof is similar to that of Lemma 4 and uses relations $\mu_k^i \in I$, $i \in V$.

As a Lyapunov function, we choose the composite function

$$\mathbf{B}(z,\mu,\mathbf{z}_k,\mu_k) = \sum_{i=1}^{m}\left[B_x(x,x_k^i) + B_\eta(\eta,\eta_k^i) + B_{\mu,i}(\mu_i,\mu_k^i)\right], \tag{39}$$

where $\mathbf{z}_k = (x_k^1,\ldots,x_k^m,\eta_k^1,\ldots,\eta_k^m)$ and $\mu_k = (\mu_k^1,\ldots,\mu_k^m)$. We next establish a descent-type relation for the expected value of the Lyapunov function, which requires an appropriate *sigma*-field. We define the σ-field as follows:

$$F_k = \{x_t^i, \eta_t^i, \mu_t^i, i \in V, t = 0,1,\ldots,k\} \qquad \text{for } k \geq 0.$$

From now on, we abbreviate the conditional expectation notation by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot\,|\,F_k]$. We have the following result that will play the key role in the convergence analysis.

Lemma 7. *Let* $D \subset \mathbb{R}$ *be a convex compact set such that* $\eta^* \in D$. *Let Assumptions 1, 2 and 3 hold. Then, for algorithm* (37)–(38) *the following relation holds with probability 1, for all* $k \geq 0$*, all* $z^* = (x^*,\eta^*)$ *with* $x^* \in X^*$*, and all optimal multipliers* μ^**,*

$$\mathbb{E}_k[\boldsymbol{B}(z^*,\mu^*,z_{k+1},\mu_{k+1})] \leq \boldsymbol{B}(z^*,\mu^*,z_k,\mu_k) + 2\alpha_k \sum_{j=1}^{m} \left(C\|x_k^j - \hat{x}_k\| + |\eta_k^j - \hat{\eta}_k| \right)$$
$$- \alpha_k \left(\mathscr{L}(\hat{x}_k,\hat{\eta}_k,\mu^*) - \mathscr{L}(x^*,\eta^*,\mu_k) \right)$$
$$+ \alpha_k^2 \left(m\frac{(C+\nu)^2}{2\sigma_x} + \sum_{i=1}^{m} \frac{(1/m-\mu_k^i)^2}{2\sigma_\eta} + \sum_{i=1}^{m} \frac{(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)^2}{2\sigma_{\mu,i}} \right).$$

Proof. Let $z=(x^*,\eta^*)\in X^*\times D$ and $\mu^*\in I^m$ be arbitrary optimal primal-dual pairs, where I^m is the product of m copies of the interval $I=[0,1]$. Using the optimality conditions for (38) and proceeding as in the proof of Lemma 5, we obtain with probability 1,

$$\mathbb{E}_k\left[B_x(x^*,x_{k+1}^i)\right] \leq B_x(x^*,\tilde{x}_k^i) - \alpha_k\mu_k^i\langle \mathbb{E}_k[d_k^i+\varepsilon_k^i],\tilde{x}_k^i - x^*\rangle + \alpha_k^2\frac{(C+\nu)^2}{2\sigma_x}.$$

Now, since $\mathbb{E}_k[\varepsilon_k^i]=0$, and d_k^i is a subgradient of $f_i(x)$ at $\tilde{x}_k^i$ it follows that

$$\mathbb{E}_k\left[B_x(x^*,x_{k+1}^i)\right] \leq B_x(x^*,\tilde{x}_k^i) - \alpha_k\mu_k^i(f_i(\tilde{x}_k^i)-f_i(x^*)) + \alpha_k^2\frac{(C+\nu)^2}{2\sigma_x}. \tag{40}$$

A similar analysis gives the following inequality for the iterates η_k^i,

$$B_\eta(\eta^*,\eta_{k+1}^i) \leq B_\eta(\eta^*,\tilde{\eta}_k^i) - \alpha_k(1/m-\mu_k^i)(\tilde{\eta}_k^i-\eta^*) + \alpha_k^2\frac{(1/m-\mu_k^i)^2}{2\sigma_\eta}, \tag{41}$$

and the inequality for the multiplier iterates μ_k^i,

$$B_{\mu,i}(\mu_i^*,\mu_{k+1}^i) \leq B_{\mu,i}(\mu_i^*,\mu_k^i) + \alpha_k(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)(\mu_k^i-\mu_i^*) + \alpha_k^2\frac{(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)^2}{2\sigma_{\mu,i}}. \tag{42}$$

Summing equations (40)–(42), and then summing the resulting relation over all $i\in V$, we have with probability 1,

$$\mathbb{E}_k[\mathbf{B}(z^*,\mu^*,\mathbf{z}_{k+1},\mu_{k+1})] \leq \mathbf{B}(z^*,\mu^*,\tilde{\mathbf{z}}_k,\mu_k) + \alpha_k^2 K \tag{43}$$
$$- \alpha_k \sum_{i=1}^{m} \left(\underbrace{\mu_k^i(f_i(\tilde{x}_k^i)-f_i(x^*))}_{\text{Term 1}} + \underbrace{(1/m-\mu_k^i)(\tilde{\eta}_k^i-\eta^*)}_{\text{Term 2}} - \underbrace{(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)(\mu_k^i-\mu_i^*)}_{\text{Term 3}} \right),$$

where we use notation (39) for the Lyapunov function, $\tilde{\mathbf{z}}_k=(\tilde{x}_k^1,\ldots,\tilde{x}_k^m,\tilde{\eta}_k^1,\ldots,\tilde{\eta}_k^m)$, and $K = m\frac{(C+\nu)^2}{2\sigma_x} + \sum_{i=1}^m \frac{(1/m-\mu_k^i)^2}{2\sigma_\eta} + \sum_{i=1}^m \frac{(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)^2}{2\sigma_{\mu,i}}$. We now estimate the identified terms in the preceding relation by adding and subtracting $f(\hat{x}_k)$ or $\hat{\eta}_k$. We have

$$f_i(x_k^i)-f_i(x^*) = (f_i(\hat{x}_k)-f_i(x^*)) + (f_i(\tilde{x}_k^i)-f_i(\hat{x}_k)) \geq f_i(\hat{x}_k)-f_i(x^*) - C\|\tilde{x}_k^i-\hat{x}_k\|,$$

where the inequality follows from the convexity of f_i and subgradient boundedness. Since $\mu_k^i \in I = [0,1]$ for all $i \in V$, it follows

$$\text{Term } 1 \geq \mu_k^i(f_i(\hat{x}_k) - f_i(x^*)) - C\|\tilde{x}_k^i - \hat{x}_k\|. \tag{44}$$

For the second term, we have

$$\begin{aligned}(1/m - \mu_k^i)(\tilde{\eta}_k^i - \eta^*) &= (1/m - \mu_k^i)(\hat{\eta}_k - \eta^*) + (1/m - \mu_k^i)(\tilde{\eta}_k^i - \hat{\eta}_k) \\ &\geq (1/m - \mu_k^i)(\hat{\eta}_k - \eta^*) - |1/m - \mu_k^i|\,|\tilde{\eta}_k^i - \hat{\eta}_k|.\end{aligned}$$

Using $\mu_k^i \in I = [0,1]$ for all $i \in V$, we obtain

$$\text{Term } 2 \geq (1/m - \mu_k^i)(\hat{\eta}_k - \eta^*) - |\tilde{\eta}_k^i - \hat{\eta}_k|. \tag{45}$$

For the third term, we write

$$\begin{aligned}(f_i(\tilde{x}_k^i) - \tilde{\eta}_k^i)(\mu_k^i - \mu_i^*) &= (f_i(\hat{x}_k) - \hat{\eta}_k)(\mu_k^i - \mu_i^*) \\ &\quad + \big((f_i(\tilde{x}_k^i) - f_i(\hat{x}_k)) - (\tilde{\eta}_k^i - \hat{\eta}_k)\big)(\mu_k^i - \mu_i^*) \\ &\geq (f_i(\hat{x}_k) - \hat{\eta}_k)(\mu_k^i - \mu_i^*) - \big(C\|\tilde{x}_k^i - \hat{x}_k\| + |\tilde{\eta}_k - \hat{\eta}_k|\big)\,|\mu_k^i - \mu_i^*|,\end{aligned}$$

where the inequality follows by the convexity of f_i and the subgradient boundedness. Again, since $\mu_k^i, \mu \in I$, we see that $|\mu_k^i - \mu_i| \leq 1$, implying

$$\text{Term } 3 \geq (f_i(\hat{x}_k) - \hat{\eta}_k)(\mu_k^i - \mu_i^*) - \big(C\|\tilde{x}_k^i - \hat{x}_k\| + |\tilde{\eta}_k - \hat{\eta}_k|\big). \tag{46}$$

Substituting estimates (44)–(46) back in relation (43), we obtain

$$\begin{aligned}\mathbb{E}_k[\mathbf{B}(z^*, \mu^*, \mathbf{z}_{k+1}, \mu_{k+1})] &\leq \mathbf{B}(z^*, \mu^*, \tilde{\mathbf{z}}_k, \mu_k) + \alpha_k^2 K + 2\alpha_k \sum_{i=1}^m \big(C\|\tilde{x}_k^i - \hat{x}_k\| + |\tilde{\eta}_k^i - \hat{\eta}_k|\big) \\ &- \alpha_k \sum_{i=1}^m \big(\mu_k^i(f_i(\hat{x}_k) - f_i(x^*)) + (1/m - \mu_k^i)(\hat{\eta}_k - \eta^*) - (f_i(\hat{x}_k) - \hat{\eta}_k)(\mu_k^i - \mu_i^*)\big).\end{aligned} \tag{47}$$

By definition in (10), the estimates $\tilde{x}_k^i$ and $\tilde{\eta}_k^i$ are convex combinations of x_k^j and η_k^j, respectively. Since under our assumption the Bregman-distance functions B_x and B_η are convex in the second argument, it follows

$$B_x(x^*, \tilde{x}_k^i) \leq \sum_{j=1}^m [W_k]_{ij} B_x(x^*, x_k^j), \qquad B_\eta(\eta^*, \tilde{\eta}_k^i) \leq \sum_{j=1}^m [W_k]_{ij} B_\eta(\eta^*, \eta_k^j).$$

By summing these relations over $i \in V$ and using the doubly stochasticity of the weight matrix W_k, we obtain

$$\sum_{i=1}^m B_x(x^*, \tilde{x}_k^i) \leq \sum_{j=1}^m B_x(x^*, x_k^j), \qquad \sum_{i=1}^m B_\eta(\eta^*, \tilde{\eta}_k^i) \leq \sum_{j=1}^m B_\eta(\eta^*, \eta_k^j).$$

Using the definition of $\mathbf{B}$ (see (39)) and these relations, from (47) we have

$$\mathbb{E}_k[\mathbf{B}(z^*,\mu^*,\mathbf{z}_{k+1},\mu_{k+1})] \leq \mathbf{B}(z^*,\mu^*,\mathbf{z}_k,\mu_k) + \alpha_k^2 K + 2\alpha_k \sum_{i=1}^m \left(C\|\tilde{x}_k^i - \hat{x}_k\| + |\tilde{\eta}_k^i - \hat{\eta}_k|\right)$$
$$- \alpha_k \underbrace{\sum_{i=1}^m \left(\mu_k^i (f_i(\hat{x}_k) - f_i(x^*)) + (1/m - \mu_k^i)(\hat{\eta}_k - \eta^*) - (f_i(\hat{x}_k) - \hat{\eta}_k)(\mu_k^i - \mu_i^*)\right)}_{\text{Term}}. \tag{48}$$

Now, we consider the identified term in (48). We note that

$$\sum_{i=1}^m \mu_k^i (f_i(\hat{x}_k) - f_i(x^*)) + (1/m - \mu_k^i)(\hat{\eta}_k - \eta^*) = \mathscr{L}(\hat{x}_k, \hat{\eta}_k, \mu_k) - \mathscr{L}(x^*, \eta^*, \mu_k),$$

$$\sum_{i=1}^m (f_i(\hat{x}_k) - \hat{\eta}_k)(\mu_k^i - \mu_i^*) = \mathscr{L}(\hat{x}_k, \hat{\eta}_k, \mu_k) - \mathscr{L}(\hat{x}_k, \hat{\eta}_k, \mu^*),$$

which imply

$$\text{Term} = \mathscr{L}(\hat{x}_k, \hat{\eta}_k, \mu^*) - \mathscr{L}(\hat{x}^*, \eta^*, \mu_k). \tag{49}$$

Furthermore, from convexity of the norm and the absolute value functions, since W_k is doubly stochastic, and $\tilde{x}_k^i = \sum_{j=1}^m x_k^j$ and $\tilde{\eta}_k^i = \sum_{j=1}^m x_k^j$ it follows that

$$\sum_{i=1}^m \left(C\|\tilde{x}_k^i - \hat{x}_k\| + |\tilde{\eta}_k^i - \hat{\eta}_k|\right) \leq \sum_{j=1}^m \left(C\|x_k^j - \hat{x}_k\| + |\eta_k^j - \hat{\eta}_k|\right). \tag{50}$$

Using (49)–(50) and $K = m\frac{(C+\nu)^2}{2\sigma_x} + \sum_{i=1}^m \frac{(1/m-\mu_k^i)^2}{2\sigma_\eta} + \sum_{i=1}^m \frac{(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)^2}{2\sigma_{\mu,i}}$, from (48) we obtain the desired relation. □

In order to connect the limiting vector $(\bar{x}, \bar{\eta}, \bar{\mu})$ of the iterates generated by the primal-dual algorithm to the solutions of problem (34), we will invoke the necessary and sufficient Karush-Khun-Tucker (KKT) optimality conditions. These conditions are stated below for convenience, as adjusted to our problem.

Theorem 4. *The vector $(\bar{x}, \bar{\eta}, \bar{\mu})$ is a primal-dual optimal vector if and only if $f_i(\bar{x}) \leq \bar{\eta}$ and $\bar{\mu} \in I^m$ and the following two conditions are satisfied*

$$(\bar{x}, \bar{\eta}, \bar{\mu}) \in \text{argmin}_{(x,\eta)\in X\times D} \mathscr{L}(x, \eta, \bar{\mu}) \qquad \bar{\mu} \in \text{argmax}_{\mu\in I^m} \mathscr{L}(\bar{x}, \bar{\eta}, \mu).$$

We are now in position to assert the convergence property of the algorithm for a diminishing step size.

Theorem 5. *Let Assumptions 1, 2, and 3 hold, except for subgradient norm boundedness of Assumption 2-b. Assume that X and D are compact convex sets, and that min-max problem* (3) *has a unique optimal solution x^*. Let the step sizes satisfy $\sum_{k=0}^\infty \alpha_k = \infty$ and $\sum_{k=0}^\infty \alpha_k^2 < \infty$. Then, the agents' iterates x_k^i, η_k^i, μ_k^i, generated by algorithm* (37)–(38) *are such that with probability 1 for all $i \in V$:*

(a) The estimates x_k^i converge to the optimal point x^*.
(b) The estimates η_k^i converge to the optimal value η^* of the min-max problem.
(c) The dual iterates μ_k^i converge to a (random) optimal dual variable μ_i^*.

Proof. The proof proceeds by applying the Robbins-Siegmund result (Theorem 2) to the relation of Lemma 7. By our assumption on the step sizes we have $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, which immediately yields $\sum_{k=0}^{\infty}\alpha_k^2 m\frac{(C+\nu)^2}{2\sigma_x} < \infty$. Since, the projection step in algorithm (38) constrains the dual variables μ_k^i to lie in the interval $[0,1]$, it follows that $\sum_{k=0}^{\infty}\alpha_k^2\sum_{i=1}^{m}\frac{(1/m-\mu_k^i)^2}{2\sigma_\eta} < \infty$ with probability 1. Moreover, as the constraint set X and the set D are compact, and f_i are continuous, the term $(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)^2$ is uniformly bounded for all i and k, implying that $\sum_{k=0}^{\infty}\alpha_k^2\sum_{i=1}^{m}\frac{(f_i(\tilde{x}_k^i)-\tilde{\eta}_k^i)^2}{2\sigma_{\mu,i}} < \infty$ with probability 1. From Lemma 6 we have $\sum_{k=0}^{\infty}\alpha_k\left(C\sum_{j=1}^{m}\|\hat{x}_k - x_k^j\| + |\hat{\eta}_k - \eta_k^j|\right) < \infty$ with probability 1. As $\sum_{k=0}^{\infty}\alpha_k = \infty$, it follows that with probability 1,

$$\lim_{k\to\infty}\sum_{j=1}^{m}\left(\|\hat{x}_k - x_k^j\| + |\hat{\eta}_k - \eta_k^j|\right) = 0.$$

By the saddle-point Theorem 1 we have $\mathscr{L}(\hat{x}_k,\hat{\eta}_k,\mu^*) - \mathscr{L}(x^*,\eta^*,\mu_k) \geq 0$. Thus, all the conditions of Theorem 2 are satisfied.

By applying Theorem 2 we infer that $\mathbf{B}(z^*,\mu^*,\mathbf{z}_k,\mu_k)$ converges for every dual-optimal μ^* with probability 1, and that $\sum_{k=0}^{\infty}\alpha_k\left(\mathscr{L}(\hat{x}_k,\hat{\eta}_k,\mu^*) - L(x^*,\eta^*,\mu_k)\right) < \infty$ also holds with probability 1. Since $\sum_{k=0}^{\infty}\alpha_k = \infty$, it follows that with probability 1,

$$\lim_{k\to\infty}\left(\mathscr{L}(\hat{x}_k,\hat{\eta}_k,\mu^*) - L(x^*,\eta^*,\mu_k)\right) = 0.$$

As the sequences $\{x_k^i\}$, $\eta_k^i\}$, $\{\mu_k^i\}$ are bounded, they have accumulation points in the sets X, D and $I=[0,1]$, respectively. We can use Cantor diagonalization-type argument to select a subsequence $\{k_\ell\}$ along which the following relations hold with probability 1:

$$\lim_{\ell\to\infty}(\hat{x}_{k_\ell},\hat{\eta}_{k_\ell},\mu_{k_\ell}) = (\bar{x},\bar{\eta},\bar{\mu}) \qquad \text{with} \quad (\bar{x},\bar{\eta},\bar{\mu}) \in X\times D\times I^m, \tag{51}$$

$$\lim_{\ell\to\infty}\left(\mathscr{L}(\hat{x}_{k_\ell},\hat{\eta}_{k_\ell},\mu^*) - \mathscr{L}(x^*,\eta^*,\mu_{k_\ell})\right) = 0, \tag{52}$$

$$\lim_{\ell\to\infty}\|\hat{x}_{k_\ell} - x_{k_\ell}^j\| = 0, \qquad \lim_{\ell\to\infty}|\hat{\eta}_{k_\ell} - \eta_{k_\ell}^j| = 0 \qquad \text{for all } j\in V. \tag{53}$$

We now examine the consequences of these relations. Since (x^*,η^*,μ^*) is a saddle-point of the Lagrangian, it follows that $\mathscr{L}(\bar{x},\bar{\eta},\mu^*) \geq \mathscr{L}(x^*,\eta^*,\mu^*) \geq \mathscr{L}(x^*,\eta^*,\bar{\mu})$, implying that both inequalities hold as equalities. Therefore,

$$(\bar{x},\bar{\eta}) \in \operatorname{argmin}_{(x,\eta)\in X\times D}\mathscr{L}(x,\eta,\mu^*), \qquad \bar{\mu} \in \operatorname{argmax}_{\mu\in I^m}\mathscr{L}(x^*,\eta^*,\mu).$$

Since (x^*, η^*) is the unique solution of problem (34) (by our assumption), the preceding relations together with the KKT conditions (Theorem 4) imply that $(\bar{x}, \bar{\eta}) = (x^*, \eta^*)$ and that $\bar{\mu}$ is an optimal dual multiplier.

It remains to show that the whole sequences converge to the desired optimal points. From (53) it follows that $\lim_{\ell\to\infty} \|x^j_{k_\ell} - x^*\| = 0$ and $\lim_{\ell\to\infty} |\eta^j_{k_\ell} - \eta^*| = 0$ with probability 1 for all j, implying $\lim_{\ell\to\infty} \mathbf{B}(z^*, \bar{\mu}, \mathbf{z}_{k_\ell}, \mu_{k_\ell}) = 0$. However, we have established that the sequence $\mathbf{B}(z^*, \mu^*, \mathbf{z}_k, \mu_k)$ converges with probability 1 for any optimal dual vector μ^*, so it converges with $\mu^* = \bar{\mu}$. Thus, it follows $\lim_{k\to\infty} \mathbf{B}(z^*, \bar{\mu}, \mathbf{z}_k, \mu_k) = 0$ with probability 1. Owing to the strong convexity of the Bregman functions we have

$$\mathbf{B}(z^*, \bar{\mu}, \mathbf{z}_k, \mu_k) \geq \sum_{i=1}^{m} \Big(\frac{\sigma_x}{2} \|x^* - x^i_k\|^2 + \frac{\sigma_\eta}{2} |\eta^* - \eta^i_k|^2 + \frac{\sigma_{\mu,i}}{2} |\bar{\mu}_i - \mu^i_k|^2 \Big),$$

which yields $x^i_k \to x^*$, $\eta^i_k \to \eta^*$ and $\mu^i_k \to \bar{\mu}_i$ with probability 1 for all $i \in V$. □

5 Min-Max Game Against Exogenous Player

In this section we consider a different formulation than the one discussed so far. We consider the case when the network of cooperative agents need to solve a min-max game against an exogenous player. The exogenous player is a malicious agent/nature which adversely affects the cost of each agent. Let us denote the action of the adversarial agent by ξ. We also require that the feasible set of allowable actions ξ is a compact set denoted by Θ. The objective is to solve:

$$\min_{x\in X} \max_{\xi\in\Theta} \sum_{i=1}^{m} f_i(x, \xi). \tag{54}$$

This can be thought of as the robust version of the problem considered in [35, 29], where the optimization problem of the form $\min_{x\in X} \sum_{i=1}^{m} \mathbb{E}_\xi [f_i(x, \xi)]$ was considered. In certain cases when it is desired to model the unknown signal ξ as lying in an uncertainty set Θ the robust version of problem (54) is more suitable. The problem (54) could alternatively be thought of as a zero sum game between the exogenous player and the network. To guarantee the existence of a min-max optimal solution to (54) we impose the following assumption.

Assumption 4 *Let the following hold:*

(a) The functions f_i are continuous over some open set containing $X \times \Theta$.
(b) The cost functions $f_i(x, \xi)$ are convex in x for every fixed value of $\xi \in \Theta$, and concave in ξ for every fixed $x \in X$.
(c) The constraint sets X and Θ are convex and compact.

Under Assumption 4 the min-max problem (54) admits a solution set $X^* \times \Theta^*$ such that for any $x^* \in X^*$ and $\xi^* \in \Theta^*$, we have the saddle-point property [4]:

$$\sum_{i=1}^{m} f_i(x^*,\xi) \le \sum_{i=1}^{m} f_i(x^*,\xi^*) \le \sum_{i=1}^{m} f_i(x,\xi^*) \quad \text{for all } x \in X \text{ and } \xi \in \Theta. \tag{55}$$

Let $B_x(\cdot,\cdot)$ and $B_\xi(\cdot,\cdot)$ be Bregman-distance functions for the sets X and Θ, respectively. We propose the following algorithm for min-max problem (54): at each iteration, at first, the agents update using the estimates x_k^j and ξ_k^i and obtain intermittent estimates

$$\begin{bmatrix} \tilde{x}_k^i \\ \tilde{\xi}_k^i \end{bmatrix} = \sum_{j=1}^{m} [W_k]_{ij} \begin{bmatrix} x_k^j \\ \xi_k^j \end{bmatrix}, \tag{56}$$

where W_k is a weight matrix as in (10). Using these intermittent adjustment, every agent updates according to the following rules:

$$\begin{aligned} x_{k+1}^i &= \operatorname{argmin}_{y\in X} \left[\alpha_k \langle \nabla_x f_i(\tilde{x}_k^i, \tilde{\xi}_k^i), y\rangle + B_x(y,\tilde{x}_k^i) \right], \\ \xi_{k+1}^i &= \operatorname{argmin}_{\zeta\in\Theta} \left[-\alpha_k \langle \nabla_\xi f_i(\tilde{x}_k^i, \tilde{\xi}_k^i), \zeta\rangle + B_\xi(\zeta,\tilde{\xi}_k^i) \right], \end{aligned} \tag{57}$$

where ∇_x and ∇_ξ denote the partial derivative operators with respect to the variables x and ξ, respectively. The initial points x_0^i and ξ_0^i satisfy $x_0^i \in X$ and $\xi_0^i \in \Theta$ for all i. It is also assumed that the constraint sets X and Θ are common knowledge for all agents $i \in V$.

The analysis of the algorithm follows along lines similar to that of the primal-dual algorithm (37)–(38). This can be seen in light of the fact that the primal-dual algorithm computes a saddle-point of the Lagrangian function in (5), whereas algorithm (56)–(57) computes a saddle-point of problem (54). A major difference between the algorithms is the fact that in (38) the agents update their own local dual variables μ_k^i, whereas in the algorithm (57) the agents update the whole vector ξ which is coupling the agents. Note that there is no stochasticity in the current formulation unless we consider a stochastic model of the network. The final result is that asymptotically the agents estimates x_k^i and ξ_k^i converge to a common min-max optimal pair (x^*,ξ^*). We formalize the statement in the following theorem.

Theorem 6. *Let Assumption 1 and 4 hold. Assume that problem* (54) *has an optimal set of the form* $\{x^*\} \times \Theta^*$. *Moreover, assume that the Bregman-distance functions* $B_x(y,v)$ *and* $B_\xi(\zeta,\phi)$ *are convex in their second arguments* v *and* ϕ, *respectively for every fixed* y *and* ζ. *If the step sizes* α_k *in algorithm* (56)–(57) *are chosen to satisfy* $\sum_{k=0}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$, *then the local variables* (x_k^i, ξ_k^i) *converge to a common saddle-point solution* (x^*,ξ^*) *of the min-max problem* (54), *for all* $i \in V$.

Proof. The proof is similar to the proof of Theorem 5, with the Lyapunov function $\sum_{i=1}^{m} \left(B_x(x^*,x_k^i) + B_\xi(\xi^*,\xi_k^i) \right)$ for an arbitrary saddle-point solution (x^*,ξ^*). □

6 Uplink Power Control

In this section we show the suitability of our algorithms in (10)–(11) and (37)–(38) to achieve a min-max fair allocation of utility in a cellular network. We will keep our discussion brief and refer the readers to [12] for a general discussion on the power allocation problem. We will be using the formulation discussed in [30].

There are m mobile users (MU) in neighboring cells communicating with their respective base stations (BS) using a common wireless channel. Let p_i denote the power used by MU i to communicate with its base station. Due to the shared nature of the wireless medium the total received SINR at BS i is given by

$$\gamma_i(\bar{p},\bar{h}_i) = \frac{p_i h_{i,i}^2}{\sigma_i^2 + \sum_{j\neq i} p_j h_{i,j}^2},$$

where $h_{i,j}$ is the channel coefficient between MU j and BS i, and σ_i^2 is the receiver noise variance. The vector containing power variables p_i is denoted $\bar{p}$ and the vector of channel coefficients at BS i is denoted $\bar{h}_i$. The power variables are non-negative and constrained to a maximum value of p_t, i.e., $0 \leq p_i \leq p_t$ for all i.

Let $U_i(\gamma_i(\bar{p},\bar{h}_i))$ be the utility derived by BS i and $V(p_i)$ be a cost function penalizing excessive power. We are interested in finding an allocation that minimizes the worst case loss to any agent i, which amounts to solving the following problem:

$$\min_{\bar{p}\in\Pi} \max_{i\in V} \left[V(p_i) - U_i(\gamma_i(\bar{p},\bar{h}_i))\right],$$

where $\Pi = \{\bar{p} \in \mathbb{R}^m \mid 0 \leq p_i \leq p_t \text{ for all } i\}$ and p_t is the maximum power. We consider the logarithmic utility function $U_i(u) = \log(u)$ for $u > 0$. Using the transformation $p_i = e^{x_i}$, it can be shown that the preceding problem can be cast in the form of (1), with a cost function for each base station i given by:

$$f_i(x) = \log\left(\sigma_i^2 h_{i,i}^{-2} e^{-x_i} + \sum_{j\neq i} h_{i,i}^{-2} h_{j,i}^2 e^{x_j - x_i}\right) + V(e^{x_i}),$$

and $X = \{x \in \mathbb{R}^m \mid x_i \leq \log(p_t) \text{ for all } i\}$.

We have considered a cellular network of 16 square cells of the same size ($m = 16$). The connectivity network for the BSs is shown in Figure 1. Within each cell, the MU is randomly located (with a uniform distribution over the cell) and the base station is located at the center of the cell. The channel coefficient $h_{i,j}$ is assumed to decay as the fourth power of the distance between the MU j and the BS i. The shadow fading is assumed to be log-normal with variance 0.1. The receiver noise variance σ_i^2 is taken to be 0.01. The cost of the power is modeled as $V(p_i) = 10^{-3} p_i$.

In the simulations, there are no stochastic errors, i.e., all gradients and subgradients are evaluated without stochastic errors. Four algorithms are used, namely, the standard (centralized) gradient descent algorithm (applied to the penalized problem), a centralized primal-dual algorithm (that computes a saddle point of the min-

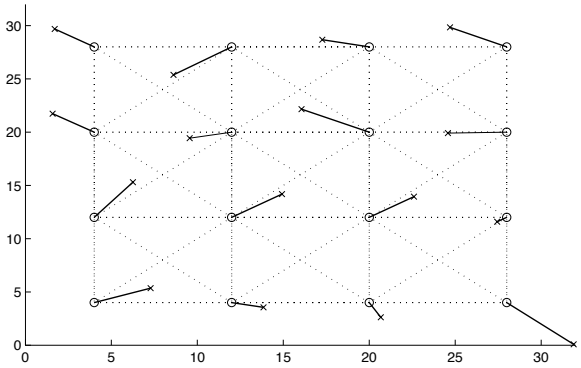

Fig. 1 The circles denote the BSs. The dotted lines denote the communication links between adjacent BSs. The crosses denote the MUs. The bold lines connect each MU to its respective BS.

max problem), the distributed exact penalty (10)–(11), and the distributed primal-dual (37)–(38). The standard (centralized) gradient descent algorithm is used to determine the optimal min-max allocation, while the centralized primal-dual is used to get a sense of dual optimal variables. In the distributed algorithms, the weight matrices W_k are all equal since the connectivity graph is static, i.e., $W_k = W$ for all k. The weights W_{ij} are given by:

$$W_{ij} = \frac{1}{\max\{|N_i|, |N_j|\}} \quad \text{if } i \text{ and } j \text{ are neighbors,}$$

and otherwise $W_{ij} = 0$, where $|N_i|$ denotes the cardinality of the neighbor set N_i (which includes agent i itself). The stepsize is $\alpha_k = \frac{10}{k}$ in the centralized methods, $\alpha_k = \frac{50}{k^{0.65}}$ in the distributed exact penalty method, and $\alpha_k = \frac{4}{k^{0.6}}$ in the distributed primal-dual method. The penalty parameter r_i is 1.3 for all i in algorithm (10)–(11). The Bregman-distance generating functions are the Euclidean norms. Each algorithm is run for 4000 iterations.

Figure 2 shows the behavior of algorithms (10)–(11) and (37)–(38). As seen in the figure, both centralized algorithms perform the best (as they have the whole knowledge of the problem information) and they have a similar behavior. The distributed algorithms are slightly worse than the centralized, which is expected due to their "decentralized" incomplete knowledge of the problem. Of the two distributed algorithms, the primal-dual algorithm is worse, as it often assigns much larger allocations than the distributed exact penalty method. Primal-dual algorithms (in absence of strong convexity of the primal) are known to be highly sensitive to the choices of dual variables, which may be reflected in the results we see in Figure 2, where the primal-dual algorithms use the initial values $\mu_i = \frac{1}{m}$ for the multipliers.

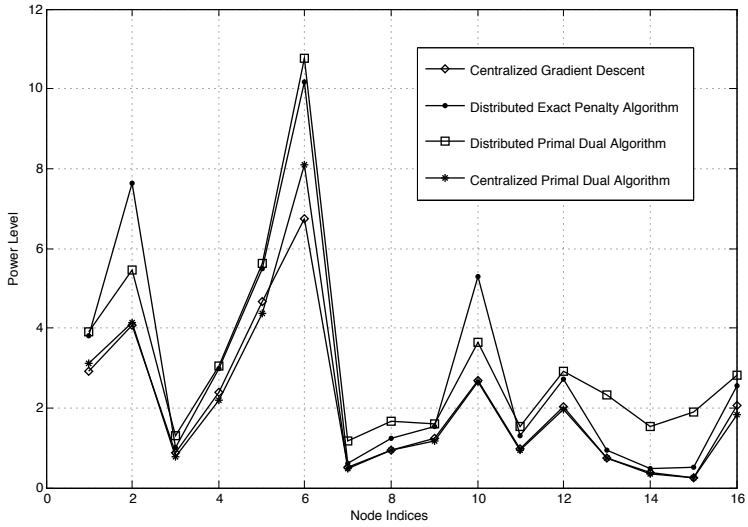

Fig. 2 The final iterate values after 4000 iterations of the algorithms. The plot shows the allocations achieved by Centralized Gradient Descent, Distributed Exact Penalty Algorithm, Distributed Primal-Dual Algorithm, and Centralized Primal-Dual Algorithm

7 Conclusion

We presented distributed algorithms for solving stochastic min-max optimization problems in networks. We developed two algorithms based on Bregman-distance functions. The first algorithm uses a non-differentiable penalty function to translate the min-max problem to a format which is suitable for distributed algorithms. The second algorithm is based on the primal-dual iterative update scheme. In both of these algorithms we allow the presence of stochastic subgradient noise. We provided conditions on the dynamic network under which we can guarantee almost sure convergent behavior of the algorithms. We illustrated the applicability of the algorithms on a power allocation problem in a cellular network.

The effectiveness of these algorithms is highly dependable on the underlying connectivity structure of the agent network. For future work, we plan to investigate error bounds for the proposed algorithms, which will capture the scalability of the algorithms with the number m of agents. Based on our prior work [31], we know that these algorithms can scale at best in the order of $m^{3/2}$ when the sum of the functions f_i is to be minimized and the $\sum_{i=1}^{m} f_i$ is strongly convex. We believe that this bound is also achievable by the proposed algorithms for a class of functions f_i, such as linear for example. Such results will also provide better insights into the practical short-term behavior of these algorithms.

Acknowledgements This work has been supported by the NSF Grant CMMI-0742538.

References

1. A. Agarwal, J. Duchi, and M. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 2011. To appear.
2. K.J. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-Linear Programming*. Stanford University Press, Stanford, CA, 1958.
3. D.P. Bertsekas. Necessary and sufficient conditions for a penalty method to be exact. *Mathematical Programming*, 9:87–99, 1975.
4. D.P. Bertsekas, A. Nedić, and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA, USA, 2003.
5. D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.
6. D.P. Bertsekas and J.N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. OPTIM.*, 10(3):627–642, 2000.
7. P. Billingsley. *Probability and Measure*. John Wiley and Sons, 1979.
8. H. Boche, M. Wiczanowski, and S. Stanczak. Unifying view on min-max fairness and utility optimization in cellular networks. In *Wireless Communications and Networking Conference, 2005 IEEE*, volume 3, pages 1280 – 1285 Vol. 3, march 2005.
9. V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.
10. L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.
11. Y. A. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.
12. M. Chiang, P. Hande, T. Lan, and W.C. Tan. Power Control in Wireless Cellular Networks. *Found. Trends Netw.*, 2(4):381–533, 2008.
13. Y. Ermoliev. Stochastic quasi-gradient methods and their application to system optimization. *Stochastics*, 9(1):1–36, 1983.
14. Y. Ermoliev. Stochastic quazigradient methods. In *Numerical Techniques for Stochastic Optimization*, pages 141–186. Springer-Verlag, N.Y., 1988.
15. T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.
16. A. Jadbabaie, J. Lin, and S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48:988–1001, 2003.
17. B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, Aug 2009.
18. S. Kar and J.M.F. Moura. Distributed consensus algorithms in sensor networks with imperfect communication: link failures and channel noise. *IEEE Tran. Signal Process.*, 57(1):355–369, 2009.
19. I. Lobel and A. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306, june 2011.
20. D. Mosk-Aoyama, T. Roughgarden, and D. Shah. Fully distributed algorithms for convex optimization problems. In *DISC '07: Proceedings of the 21st international symposium on Distributed Computing*, pages 492–493, Berlin, Heidelberg, 2007. Springer-Verlag.
21. A. Nedić. Asynchronous broadcast-based convex optimization over a network. *IEEE Transactions on Automatic Control*, 56(6):1337–1351, 2011.
22. A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis. Distributed subgradient algorithms and quantization effects. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 4177–4184, 2008.
23. A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

24. A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.
25. A. Nedić, A. Ozdaglar, and P. A. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55:922–938, 2010.
26. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, 2008.
27. B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., New York, 1987.
28. M. Rabbat and R. D. Nowak. Distributed optimization in sensor networks. In *IPSN*, pages 20–27, 2004.
29. S. S. Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545, 2010.
30. S. S. Ram, V. V. Veeravalli, and A. Nedić. Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM*, pages 3001–3005, 2009.
31. S. Sundhar Ram, A. Nedić, and V.V. Veeravalli. Asynchronous gossip algorithms for stochastic optimization: Constant stepsize analysis. In M. Diehl, F. Glineur, E. Jarlebring, and W. Michiels, editors, *Recent Advances in Optimization and its Applications in Engineering*, pages 51–60. Volume of the 14th Belgian-French-German Conference on Optimization (BFG), 2010.
32. S.S. Ram, A. Nedić, and V.V. Veeravalli. A new class of distributed optimization algorithms: Application to regression of distributed data. *Optimization Methods and Software*, 27(1):71–88, 2012.
33. H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In J.S. Rustagi, editor, *Proceedings of a Symposium on Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.
34. R. Srikant. *The Mathematics of Internet Congestion Control*. Birkhäuser, Boston, 2003.
35. K. Srivastava and A. Nedić. Distributed asynchronous constrained stochastic optimization. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):772 –790, aug. 2011.
36. K. Srivastava, A. Nedić, and D. Stipanović. Distributed min-max optimization in networks. In *17th International Conference on Digital Signal Processing*, 2011.
37. S.S. Stanković, M.S. Stanković, and D.M. Stipanović. Decentralized parameter estimation by consensus based stochastic approximation. In *Decision and Control, 2007 46th IEEE Conference on*, pages 1535 –1540, dec. 2007.
38. J. N. Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, Boston, 1984.
39. J. N. Tsitsiklis and M. Athans. Convergence and asymptotic agreement in distributed decision problems. *IEEE Trans. Automat. Control*, 29:42–50, 1984.