Distributed Mirror Descent over Directed Graphs

Chenguang Xi†, Qiong Wu‡, and Usman A. Khan†

Abstract

In this paper, we propose Distributed Mirror Descent (DMD) algorithm for constrained convex optimization problems on a (strongly-)connected multi-agent network. We assume that each agent has a private objective function and a constraint set. The proposed DMD algorithm employs a locally designed Bregman distance function at each agent, and thus can be viewed as a generalization of the well-known Distributed Projected Subgradient (DPS) methods, which use identical Euclidean distances at the agents. At each iteration of the DMD, each agent optimizes its own objective adjusted with the Bregman distance function while exchanging state information with its neighbors. To further generalize DMD, we consider the case where the agent communication follows a directed graph and it may not be possible to design doubly-stochastic weight matrices. In other words, we restrict the corresponding weight matrices to be row-stochastic instead of doubly-stochastic. We study the convergence of DMD in two cases: (i) when the constraint sets at the agents are the same; and, (ii) when the constraint sets at the agents are different. By partially following the spirit of our proof, it can be shown that a class of consensus-based distributed optimization algorithms, restricted to doubly-stochastic matrices, remain convergent with stochastic matrices.

Index Terms

Distributed convex optimization; multi-agent network; mirror descent; projected subgradient; directed graphs.

†C. Xi and U. A. Khan are with the Department of Electrical and Computer Engineering, Tufts University, 161 College Ave, Medford, MA 02155; chenguang.xi@tufts.edu, khan@ece.tufts.edu. This work has been partially supported by an NSF Career Award # CCF-1350264.

‡Q. Wu is with the Department of Mathematics, Tufts University, 503 Boston Ave, Medford, MA 02155; qiong.wu@tufts.edu.
I. INTRODUCTION

Distributed computation and optimization, [1, 2], has received significant recent interest in many areas, e.g. multi-agent networks, [3], model predictive control, [4], cognitive networks, [5], source localization, [6], resource scheduling, [7], and message routing, [8]. The related problems, in general, can be posed as the minimization of a sum of objective functions with constraints. In this paper, we focus on the constrained convex optimization on a connected network of \( m \) agents. The global objective is to cooperatively minimize the sum, \( \sum_{i=1}^{m} f_i(x) \), where \( f_i : \mathbb{R}^p \rightarrow \mathbb{R} \) is a private objective function available only to agent \( i \). The values of agent \( i \) are constrained to lie in a private closed convex set, \( X_i \).

There has been a considerable work on related distributed optimization problems. Of particular interest is the Distributed Projected Subgradient (DPS) method, see, e.g. [9], whose convergence rate and fault tolerance are well analyzed in related literature [10–15]. The DPS method proved to be simple and efficient, and is widely used in large-scale distributed optimization [6, 8, 16–18]. This is due to the fact that DPS is a first-order algorithm requiring only the calculation of subgradients and projections. In some applications such as large-scale learning, these first order methods are preferable to higher order approaches as e.g. in [19]. Despite benefiting from the simplicity of first-order, it is sometimes challenging to generate projections for certain objective functions and constraint sets, yielding inefficient updates. One example is the entropy-based loss function with the constraint set being the unit simplex, [20].

In this paper, we propose a first-order generalization to DPS that we refer to as the Distributed Mirror Descent (DMD) algorithm. This generalization is motivated by the mirror descent method proposed originally by Nemirovski and Yudin, [21]. Mirror descent methods generalize the classical first-order gradient descent by using a Bregman divergence instead of the Euclidean distances. Additionally, DMD can be viewed as a proximal algorithm, [20], where the proximal function used is the Bregman divergence, [22]. Many widely used distance measures turn out to be special cases of the Bregman divergence, e.g. the Euclidean distance and the Kullback Liebler (KL) divergence, [23].

Compared to the gradient descent, mirror descent is more efficient in high dimensions, e.g. in reconstructing 3D medical images from Positron Emission Tomography (PET), where mirror descent is well-suited for the corresponding optimization problem with millions of variables, [24]. It is shown in [24] that mirror descent with Bregman divergence defined by the \( p \)-norm function...
can outperform projected subgradient methods by a factor $\frac{p}{\log p}$, where $p$ is the dimensionality of the space. In addition, mirror descent allows the flexibility to generate efficient projections by carefully choosing the Bregman divergence. For example, for entropy-based loss function minimization with constraint set being the unit simplex, [20], a KL-based Bregman divergence is more appropriate with which the mirror descent becomes an exponentiated gradient update algorithm, [25]; in contrast to the additive updates in gradient descent. Other advantages of mirror descent appear in reinforcement learning, [26], where the performance is similar to the traditional techniques, [27], but the complexity is linear in the number of features, while traditional methods require near cubic complexity. Recent work, [28], in online learning has explored the applications of mirror descent in developing sparse methods for regularized classification and regression.

Our work on DMD generalizes the DPS methods [9–14] towards solving distributed optimization problems, where we employ a local Bregman divergence at each agent instead of the same global Euclidean distance at all agents. To further generalize the DMD, we note that the existing work assumes the inter-agent communication to follow an undirected graph, which leads to doubly-stochastic weight matrices. In contrast, we consider the case where the agent communication graph is directed. In particular, we do not assume the weight matrices to be doubly-stochastic but only row-stochastic. Clearly, a directed topology has broader applications in contrast to the undirected graphs and may further result in reduced communication cost and simplified topology design.

Recent work has considered distributed algorithms, [29–33], for directed graphs by combining gradient descent and push-sum protocol [31]. The related algorithms are well-suited for solving either unconstrained problems [29, 30], or problems with identical constraints [31–33]. However, these algorithms require the agents to not only exchange their own states but also some additional auxiliary states, which increases the communication cost. Specifically, the work in [29, 30] solves unconstrained optimization in time-varying networks and the implementation requires every agent to know its out-degree, which may not be possible e.g. in a broadcast-based directed graph. Similarly, the work in [31–33] focuses on identical constraints but fixed topologies and the implementation requires the knowledge of the graph or of the number of agents.

In this paper, we analyze DMD in two cases: (i) when the constraint sets at the agents are identical, the results are applicable to directed, time-varying networks; and, (ii) when the constraints at the agents are different, the results are applicable to fixed but directed topologies. In both cases, our results do not require the knowledge of the topology or the out-degree of the
agents. Compared to the existing work in [29–33] on directed graphs, the proposed DMD, in spirit, is similar to the DPS methods, [9–14], and requires the agents to exchange their states with their neighbors while no auxiliary states are introduced. By partially following the spirit of our proofs, one can show that existing algorithms on consensus-based distributed optimization, [9–14], remain convergent in the directed graphs. In particular, it can be shown that the DPS methods, [9–14], restricted to doubly-stochastic matrices, remain convergent with stochastic matrices.

The remainder of the paper is organized as follows. We provide the problem formulation, the proposed DMD algorithm, and the corresponding assumptions in Section II. In Section III, we show two key results that are useful to the develop the subsequent convergence analyses. The convergence behavior of the DMD is studied in Section IV, where we consider two cases: (i) when the constraint sets at the agents are identical; and, (ii) when the constraint sets at the agents are different. Section V contains concluding remarks, and the Appendix recapitulates some existing results, which will be frequently used in this paper.

**Notation** We use lowercase bold letters to denote vectors and uppercase italic letters to denote matrices. We denote by \([x]_i\), the \(i\)th component of a vector \(x\), and by \([A]_{ij}\) the \((i,j)\)th element of a matrix, \(A\). A vector with all elements equal to one is represented by \(1\). The inner product of two vectors \(x\) and \(y\) is \(\langle x, y \rangle\). We use \(\|x\|\) to denote the standard Euclidean norm. For a function \(f : \mathbb{R}^p \to (-\infty, \infty]\), we denote the domain of \(f\) by \(\text{dom}(f)\), where \(\text{dom}(f) = \{x \in \mathbb{R}^p | f(x) < \infty\}\). For any function \(f\), we write \(f \in C^\zeta\) if the first \(\zeta\) derivatives \(f^{(1)}, f^{(2)}, \cdots, f^{(\zeta)}\), all exist and are continuous. Finally, for two matrices, \(X, Y\), we use \(X \succeq Y\) to represent the matrix \(X - Y\) is positive semi-definite.

**II. Problem Formulation**

Consider a time-varying network of \(m\) agents communicating over a directed graph, \(G_k = (\mathcal{V}, \mathcal{E}_k)\), where \(\mathcal{V}\) is the set of agents, and \(\mathcal{E}_k\) is the collection of ordered pairs, \((i, j), i, j \in \mathcal{V}\), such that agent \(j\) can send information to agent \(i\), at time \(k\). Define \(\mathcal{N}^{\text{in}}_i(k)\) to be the collection of in-neighbors, i.e. the set of agents that can send information to agent \(i\) at time \(k\). Similarly, \(\mathcal{N}^{\text{out}}_i(k)\) is defined as the out-neighborhood of agent \(i\) at time \(k\), i.e. the set of agents that can receive information from agent \(i\) at time \(k\). We focus on solving a constrained convex optimization problem that is distributed over the network.
In particular, the network of agents cooperatively solve the following constrained problem:

\[
(P1) : \quad \text{minimize} \quad f(x) = \sum_{i=1}^{m} f_i(x)
\]
subject to \( x \in X = \bigcap_{i=1}^{m} X_i \),

where each \( f_i : \mathbb{R}^p \to \mathbb{R} \) is convex, not necessarily differentiable, representing the local objective function at agent \( i \), and each \( X_i \subseteq \mathbb{R}^p \) is a closed convex set, representing the local constraint at agent \( i \). The intersection, \( X \), of the constraint sets is assumed to be nonempty. We use \( f^* \) to denote the optimal value of the problem, and \( X^* \) to denote the solution set of the problem. Formally, we have

\[
f^* = \min_{x \in X} f(x), \quad X^* = \left\{ x \in X \mid f(x) = \min_{y \in X} f(y) \right\}.
\]

Assuming each local function, \( f_i \), is known only to agent \( i \), the goal is to solve Problem (P1) using a distributed algorithm in which agents in the network do not share their private functions with each other, but only exchange the iterative states with their immediate neighbors.

A. Distributed Mirror Descent (DMD) Algorithm

We consider the DMD algorithm to iteratively solve Problem (P1). The iterative algorithm makes use of the Bregman divergence, which is defined in Definition 1.

**Definition 1** (Bregman divergence [22]). Given a strongly convex differentiable function, \( \mu : \mathbb{R}^p \to \mathbb{R} \), we call, \( B_{\mu}(x, y) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R} \), a Bregman divergence between \( x \) and \( y \) based on a distance generating function, \( \mu \), such that

\[
B_{\mu}(x, y) = \mu(x) - \mu(y) - \langle x - y, \nabla \mu(y) \rangle,
\]

where \( \nabla \) is the gradient.

Due to strong convexity of the distance generating function, \( \mu \), it can be seen that the Bregman divergence between any two vectors, \( B_{\mu}(x, y) \), is nonnegative and \( B_{\mu}(x, y) = 0 \) if and only if \( x = y \). Thus, Bregman divergence can be a “metric” between any two vectors. In particular, a special case is when the Bregman divergence is the squared Euclidean distance, i.e. \( B_{\mu}(x, y) = \|x - y\|^2 \), when \( \mu(x) = x^\top x, \mu(y) = y^\top y, \langle x - y, \nabla \mu(y) \rangle = \langle x - y, 2y \rangle = 2x^\top y - 2y^\top y \), where \((\cdot)^\top\) denotes the transpose.
Towards the iterative DMD, let $x^k_i$ be the state at agent $i$ and time $k$. At $(k+1)$th iteration, agent $i$ receives the states $x^k_j$ from its neighbors, $j \in \mathcal{N}_i^m(k)$, computes a weighted average of these states, and performs a local optimization according to the (sub)gradient of its objective function, $f_i$. In particular, agent $i$ generates the following sequence:

$$v^k_i = \sum_{j \in \mathcal{N}_i^m(k)} w^k_{ij} x^k_j,$$

(1a)

$$x^{k+1}_i = \arg\min_{x \in \mathcal{X}_i} \left\{ \langle x, d^k_i \rangle + \frac{1}{\alpha^k_i} B_{\mu_i}(x, v^k_i) \right\},$$

(1b)

where $w^k_{ij}$ is the weight assigned by agent $i$ to agent $j$ at time $k$, $d^k_i$ represents the subgradient of $f_i$ at $v^k_i$, $\alpha^k_i > 0$ is the stepsize at agent $i$, and $\mu_i$ is a distance generating function locally designed by agent $i$ to calculate its Bregman divergence. The local design of Bregman divergence, $B_{\mu_i}$, at each agent $i$, will be discussed later in Assumption A4.

We refer to the iterations in Eq. (1) as the Distributed Mirror Descent (DMD) algorithm, which consists of two steps: the consensus step, Eq. (1a), and the optimization step, Eq. (1b). We may write DMD equivalently as follows:

$$v^k_i = \sum_{j \in \mathcal{N}_i^m(k)} w^k_{ij} x^k_j,$$

(2a)

$$x^{k+1}_i = v^k_i + e^k_i,$$

(2b)

$$e^k_i = \arg\min_{x \in \mathcal{X}_i} \left\{ \langle x, d^k_i \rangle + \frac{1}{\alpha^k_i} B_{\mu_i}(x, v^k_i) \right\} - v^k_i,$$

(2c)

where $e^k_i$ is a perturbed subgradient. Compared to the nonlinear representation, Eq. (1b), of agent states, $x^k_i$, we are able to capture the optimization step in a linear fashion, which will be exploited to recursively represent $x^k_i$ in Section III. The nonlinear effect occurs only in the representation of the perturbed subgradient, $e^k_i$, which can be bounded with the properties of Bregman divergence.

B. Contributions

As discussed in Introduction, recall that the primary advantage of Bregman-based distributed optimization lies in: (i) improved performance in high dimensions; (ii) efficient projections generation; and, (iii) applications in large-scale online learning. The proposed DMD, Eq. (1), can be viewed as a generalization of Distributed Projected Subgradient (DPS) method [9], employing a general Bregman divergence at each agent instead of using identical Euclidean
distances at all of the agents. Our goal in this paper is to: (i) prove the convergence of DMD; and, (ii) show that doubly stochasticity is not required (by both DMD and existing consensus-based optimization algorithms, [9–14]). The convergence proof of DMD is divided into two cases. The first case assumes that the constraints at all of the agents are identical, while the second case covers the non-identical constraints. Existing work over directed graphs, e.g. in [29–33], is restricted to identical constraints or unconstrained problems; and further require the knowledge of the entire topology or the out-degree at each agent, none of which are assumed here. It is further noteworthy that the case of non-identical constraints over directed graphs has not been considered in the existing literature.

C. Assumptions

We now formulate the assumptions, which are commonly used in the related literature, [9–14]. The first assumption is to make sure that every agent sufficiently communicates with each other during the algorithm such that each individual objective function influences the states at all agents.

Assumption A1 (Network Connectivity). Let \( \mathcal{E}_k \) be the edge set of the multi-agent network, \( \mathcal{G}_k \), at time \( k \), then there exists an \( L_1 \geq 1 \) such that the graph, \( (\mathcal{V}, \bigcup_{l=0}^{L_1-1} \mathcal{E}_{k+l}) \), is strongly-connected, \( \forall k \geq 0 \).

We now describe a weighting rule for agents reaching a “consensus” after iteratively exchanging information. The rule is applicable to directed graphs where any agent calculates a weighted average from its neighbors.

Assumption A2 (Non-doubly Stochasticity). For all \( i \in \mathcal{V} \) and all \( k \geq 0 \):

(a) There exists a scalar \( \eta \), \( 0 < \eta < 1 \), such that \( w_{ij}^k \geq \eta \) when \( j \in \mathcal{N}_i^m(k) \), and \( w_{ij}^k = 0 \) otherwise.

(b) \( \sum_{j=1}^{m} w_{ij}^k = 1 \).

Assumptions [A2](a) ensures that each agent gives significant weights locally to all of its neighbors as well as itself. Assumption [A2](b) states that each agent receives and calculates a weighted average of the neighboring agent states. The collection of weights, \( W(k) = \{w_{ij}^k\} \), satisfying Assumption [A2] forms a non-doubly stochastic matrix at any time \( k \), i.e. \( W(k) \) is row-stochastic and \( W(k) \neq W(k)^\top \). Note that in a “consensus” problem, this weighting rule ensures that
every agent converges to the same limit, which is a weighted average of the agents’ initial states. Moreover, “average consensus” is achieved when \( w_{ij}^k = w_{ji}^k, \forall i, j, k \), see [34–39] for additional information. Similarly, in related literature, [9–15], on distributed optimization, the weight matrices are assumed to be doubly-stochastic such that the influence of each agent is “equal” in the long run. In this paper, we will show the convergence of the DMD with row stochastic weight matrices. This enables DMD to be applicable to directed graphs.

We next assume the following on the constraint sets.

**Assumption A3 (Compactness).** For each agent \( i \), the constraint set, \( \mathcal{X}_i \), is convex and compact, and the function, \( f_i \), is convex over \( \text{dom}(f_i) \), which contains an open set covering \( \mathcal{X} \).

This assumption implies two things: (i) the optimal value, \( f^* \), is finite and the optimal set, \( \mathcal{X}^* \), is nonempty by the Weierstrass theorem; and, (ii) the subgradients of \( f_i \) at all points, \( x \in \mathcal{X}_i \), are bounded since \( \mathcal{X}_i \) is bounded, i.e. there exists \( D \in \mathbb{R}_{\geq 0} \) such that the subgradients, \( d_i \), of \( f_i \) satisfy

\[
\|d_i\| \leq D, \quad \forall i. \tag{3}
\]

This assumption is satisfied for example when each function, \( f_i \), is defined and convex over \( \mathbb{R}^p \).

We now discuss the rules for the distance generating function, \( \mu \), on which the Bregman divergence is based.

**Assumption A4 (Separate Convexity).** Each agent, \( i \), locally designs its own distance generating function, \( \mu_i \), such that

(a) \( \mu_i \) is continuous and \( \sigma \)-strongly convex, \( \sigma \in \mathbb{R}^+ \);

(b) \( \mu_i \) has Lipschitz continuous gradients, i.e., \( \|\nabla \mu_i(x) - \nabla \mu_i(y)\| \leq L\|x - y\| \);

(c) \( \mu_i \in C^3 \) with \( \nabla^2 \mu_i \succeq 0 \);

(d) Denote \( H_i = \nabla^2 \mu_i \), then \( H(y) + \nabla H(y)(y - x) \succeq 0, \forall x, y \in \text{dom}(\mu_i) \).

Assumption [A4(a)] establishes the relationship between the Bregman divergence and the Euclidean distance. In particular,

\[
B_{\mu_i}(x, y) \geq \frac{\sigma}{2}\|x - y\|^2, \quad \forall x, y \in \text{dom}(\mu_i), \sigma \in \mathbb{R}^+. \tag{4}
\]

Assumptions [A4(b)] bounds the gradients of the distance-generating function, while [A4(c)] and [A4(d)] ensure separate convexity of \( B_{\mu_i}(x, y) \) (see Lemma [8]). Assumption [A4] is satisfied, for example, when each distance generating function is defined as, but not limited to, a quadratic function.
Finally, we assume the following on the step-sizes.

**Assumption A5 (Step sizes).** *In the DMD Algorithm, the non-negative step-sizes are diminishing and satisfy the persistence conditions for all $i$. In particular,*

\[
0 \leq \alpha_i^k, \quad \sum_{k=0}^{\infty} \alpha_i^k = \infty, \quad \sum_{k=0}^{\infty} (\alpha_i^k)^2 < \infty.
\]

### III. Basic Relations

The convergence analysis of DMD is based on two critical relations that capture the decrease in $\|x_i^k - \hat{x}_i^k\|$ for all $i$, and $f(\hat{x}_i^k) - f(x^*)$, as DMD progresses, where $\hat{x}_i^k$ is a convex combination of agents’ states, $x_i^k$, i.e. $\hat{x}_i^k = \sum_{i=1}^{m} \theta_i x_i^k$, $\sum_{i=1}^{m} \theta_i = 1$, $0 \leq \theta_i$. The decrease in the first relation, $\|x_i^k - \hat{x}_i^k\|$, reflects the consensus properties of DMD. In other words, any two agents sequences, $\{x_i^k\}$ and $\{x_j^k\}$, accumulate to the same point, $\hat{x}_i^k$. We provide an upper bound of the consensus relation in Lemma 1. The convergence of DMD to the optimal solution of Problem (P1) is reflected by showing the decrease in $\|\hat{x}_i^k - x^*\|$ or $f(\hat{x}_i^k) - f(x^*)$, both of which provide a metric between the accumulation point, $\hat{x}_i^k$, and the optima, $x^*$. In Lemma 2 we capture the decrease in values, $B_{\mu_i}(x^*, x_i^{k+1})$, with respect to the Bregman divergence. An upper bound of the relation, $f(\hat{x}_i^k) - f(x^*)$, is provided in Lemma 3. In Section IV we will improve these bounds and show the convergence of DMD.

#### A. Consensus

In distributed systems, where all agents aim to reach the same limit but iterate locally, it is important to measure the disagreement as the algorithm progresses. To this aim, we use the transition matrix defined as follows. Let $W(k)$ be the matrix collecting the weights following Assumption A2, i.e. $[W(k)]_{ij} = w_{ij}^k$. Define, for all $k$, $r$, with $k \geq r$,

\[
\Phi(k, r) = W(k)W(k-1)\cdots W(r).
\]

(5)

The above matrix, $\Phi(k, r)$, is the transition matrix, which records the weight matrices history from time $r$ to $k$. With the help of Transition Matrix Convergence (See Lemma 9 in the Appendix), we now quantify the agent disagreement of DMD in time. We consider this disagreement with respect to some common point, $\hat{x}_i^k$, defined as some convex combination of all agent states at time $k$:

\[
\hat{x}_i^k = \sum_{i=1}^{m} \theta_i x_i^k,
\]

(6)
where $\theta_i \geq 0$ for all $i$, and $\sum_{i=1}^{m} \theta_i = 1$. We will give an upper bound on the agent disagreement, measured by the sequence, \( \{ \|x^k_i - \hat{x}^k_i\| \} \), for all $i$. The next lemma provides this upper bound in terms of the transition matrices, $\Phi(k, r)$, defined in Eq. (5), the initial states of agent, $x^0_i$, and the perturbed subgradient, $e^k_i$, derived in Eq. (2c).

**Lemma 1.** Let Assumptions $\mathcal{A1}$ and $\mathcal{A2}$ hold. Let $\{x^k_i\}$ be the sequence over $k$ generated by the DMD algorithm, Eq. (2), and $\hat{x}^k$ be the weighted sum of agent states as defined in Eq. (6). Then, for any $i, k \geq 1$, and $0 \leq \gamma < 1$

\[
\|x^k_i - \hat{x}^k_i\| \leq 2\Gamma \gamma^{k-1} \sum_{j=1}^{m} \|x^0_j\| + \sum_{r=1}^{k-1} \sum_{j=1}^{m} 2\Gamma \gamma^{k-1-r} \|e^{r-1}_j\| + (1 - \theta_i) \|e^{k-1}_i\| + \sum_{j \neq i} \theta_j \|e^{k-1}_j\|.
\]

**Proof.** For any $k \geq 1$, we write Eq. (2b) recursively such that the agent states are written in terms of the initial states, $x^0_i$, the perturbed subgradient, $\{e^r_j\}$, and the transition matrices, $\{\Phi(k - 1, r)\}$:

\[
x^k_i = \sum_{j=1}^{m} [\Phi(k - 1, 0)]_{ij} x^0_j + \sum_{r=1}^{k-1} \sum_{j=1}^{m} [\Phi(k - 1, r)]_{ij} e^{r-1}_j + e^{k-1}_i.
\]

Since $\hat{x}^k$ is given by Eq. (6), we represent $\hat{x}^k$ as

\[
\hat{x}^k = \sum_{j=1}^{m} \sum_{i=1}^{m} \theta_i [\Phi(k - 1, 0)]_{ij} x^0_j + \sum_{r=1}^{k-1} \sum_{j=1}^{m} \sum_{i=1}^{m} \theta_i [\Phi(k - 1, r)]_{ij} e^{r-1}_j + \sum_{i=1}^{m} \theta_i e^{k-1}_i.
\]

Denote $\psi_j(k, r)$ as the convex combination of the elements of the $j$th column of the transition matrix, i.e. $\psi_j(k, r) = \sum_{i=1}^{m} \theta_i [\Phi(k, r)]_{ij}$. The difference between the preceding two relations equals:

\[
x^k_i - \hat{x}^k = \sum_{j=1}^{m} \left[ [\Phi(k - 1, 0)]_{ij} - \psi_j(k - 1, 0) \right] x^0_j + \sum_{r=1}^{k-1} \sum_{j=1}^{m} \left[ [\Phi(k - 1, r)]_{ij} - \psi_j(k - 1, r) \right] e^{r-1}_j + (1 - \theta_i) e^{k-1}_i - \sum_{j \neq i} \theta_j e^{k-1}_j.
\]
Taking the norm on both sides of the equation above, we get
\[
\|x_i^k - \hat{x}_i^k\| \leq \sum_{j=1}^{m} \left\| [\Phi(k - 1, 0)]_{ij} - \psi_j(k - 1, 0) \right\| \|x_0\| \\
+ \sum_{r=1}^{k-1} \sum_{j=1}^{m} \left\| [\Phi(k - 1, r)]_{ij} - \psi_j(k - 1, r) \right\| \|e_j^{r-1}\| \\
+ (1 - \theta_i) \|e_i^{k-1}\| + \sum_{j \neq i} \theta_j \|e_j^{k-1}\|.
\]

The term \(\| [\Phi(k, r)]_{ij} - \psi_j(k, r) \|\) can be bounded as follows:
\[
\| [\Phi(k, r)]_{ij} - \psi_j(k, r) \| \leq \sum_{t=1}^{m} \theta_t \| [\Phi(k, r)]_{ij} - [\Phi(k, r)]_{tj} \|,
\]
\[
\leq \sum_{t=1}^{m} \theta_t \left[ \| [\Phi(k, r)]_{ij} - \phi_j(r) \| \\
+ \| [\Phi(k, r)]_{tj} - \phi_j(r) \| \right],
\]
\[
\leq 2\Gamma \gamma^{(k-r)},
\]
where we use the convexity of \(\| [\Phi(k, r)]_{ij} - \psi_j(k, r) \|\) in the first inequality, and the convergence result of transition matrix in the last inequality (see Lemma 9 in the Appendix). The proof follows by combining the preceding two inequalities. \(\square\)

In prior work, [9–14], where weight matrices, \(W(k)\), are assumed to be doubly-stochastic for all \(k\), agent disagreement is measured between agent states, \(x_i^k\), and the average, \(\bar{x}_i^{k} = \frac{1}{m} \sum_{i=1}^{m} x_i^k\), i.e. \(\|x_i^k - \bar{x}^k\|\). In Lemma 1, where we restrict the weight matrices to be stochastic, we extend the result of agent consensus by measuring the disagreement between agent states and any linear-convex combination, i.e. \(\|x_i^k - \bar{x}^k\|\).

B. Optimality

We next show how the accumulation point, \(\hat{x}_i^k\), approaches the optima, \(x^*\), as DMD progresses. To this aim, Lemma 2 shows how each agent state, \(x_i^k\), approaches \(x^*\) by capturing the decrease in values, \(B_{\mu_i}(x^*, x_i^{k+1})\), with respect to Bregman divergence. In Lemma 3 we quantify the gap between the accumulation point, \(\hat{x}_i^k\), and the optima, \(x^*\), with respect to objective value, \(f(\hat{x}_i^k) - f(x^*)\). In the analysis of Lemmas 2 and 3 we write DMD, Eq. (1b), equivalently as
\[
x_i^{k+1} = \arg\min_{x \in \mathcal{X}_i} \{ B_{\mu_i}(x, \hat{x}_i^k) \},
\]
for some $\tilde{x}_i^k \in \mathbb{R}^p$, where $\tilde{x}_i^k$ is any point in $\mathbb{R}^p$ such that $x_i^{k+1}$ is the projection of $\tilde{x}_i^k$ on $X_i$. Note that the projection is in the sense of minimizing the Bregman divergence and for any such projection, we have

$$\nabla \mu_i (\tilde{x}_i^k) = \nabla \mu_i (v_i^k) - \alpha_i^k d_i^k.$$  

Eq. (8) can be verified by letting Eq. (1b) equal to Eq. (7) and using the definition of Bregman divergence.

**Lemma 2.** Let Assumptions A3 and A4 hold. For all $i$, let $\{x_i^k\}, \{v_i^k\}$, be the sequences (over $k$) generated by the DMD, Eqs. (2), and $x^* \in X_i$ be an optimal solution of the Problem (P1), then the following holds for all $i$ and $k \geq 0$:

$$B_{\mu_i}(x^*, x_i^{k+1}) \leq B_{\mu_i}(x^*, \tilde{x}_i^k) - B_{\mu_i}(x_i^{k+1}, \tilde{x}_i^k) + \alpha_i^k \|x^* - x_i^{k+1}\|.$$  

**Proof.** Since $x^* \in X_i$ for any $i$, we apply the non-expansive property (see Lemma 6 in the Appendix) of Bregman divergence with $x_i^{k+1}$ defined by Eq. (7):

$$B_{\mu_i}(x_i^{k+1}, \tilde{x}_i^k) \leq B_{\mu_i}(x^*, \tilde{x}_i^k) - B_{\mu_i}(x_i^{k+1}, \tilde{x}_i^k).$$  

(9)

Considering the three-points identity of the Bregman divergence (see Lemma 7 in Appendix), we get

$$B_{\mu_i}(x_i^{k+1}, \tilde{x}_i^k) = B_{\mu_i}(x_i^{k+1}, v_i^k) - B_{\mu_i}(\tilde{x}_i^k, v_i^k) + \langle \nabla \mu_i(v_i^k) - \nabla \mu_i(\tilde{x}_i^k), x_i^{k+1} - \tilde{x}_i^k \rangle,$$

$$B_{\mu_i}(x^*, \tilde{x}_i^k) = B_{\mu_i}(x^*, v_i^k) - B_{\mu_i}(\tilde{x}_i^k, v_i^k) + \langle \nabla \mu_i(v_i^k) - \nabla \mu_i(\tilde{x}_i^k), x^* - \tilde{x}_i^k \rangle.$$

By substituting the preceding two relations into Eq. (9) and rearranging the terms, it follows that

$$B_{\mu_i}(x^*, x_i^{k+1}) = B_{\mu_i}(x^*, v_i^k) - B_{\mu_i}(x_i^{k+1}, v_i^k) + \langle \nabla \mu_i(v_i^k) - \nabla \mu_i(\tilde{x}_i^k), x^* - x_i^{k+1} \rangle.$$

The lemma follows by noting that $\tilde{x}_i^k$ satisfies Eq. (8) and the subgradient is bounded by $D$ (Eq. (3)).
In Section IV, we will use the result of Lemma 2 to show the boundedness of the perturbed subgradient, \( e_k \), and prove the convergence of the Bregman divergence, \( \lim_{k \to \infty} B_{\mu_i}(x^*, x_k) \), between agents’ states and the optima. We now compare the difference in objective function value between agents’ states, \( x_k^i \), and the optima, \( x^* \), in the following lemma.

**Lemma 3.** Let Assumptions A3 and A4 hold. Let \( \{x_k^i\} \) and \( \{v_k^i\} \) be the sequences (over \( k \)) generated by the DMD, Eqs. (2), \( \hat{x}_k \) be defined as in Eq. (6), and \( x^* \in X^* \) be an optimal solution of Problem (P1), then the following hold:

(a) For all \( k \geq 0 \) and every agent \( i \),

\[
\alpha_k^i \left( f_i(v_k^i) - f_i(x^*) \right) \leq B_{\mu_i}(x^*, v_k^i) - B_{\mu_i}(x^*, x_{k+1}^i) + \left( \frac{\alpha_k^i}{2} \right)^2 D^2.
\]

(b) When step sizes are the same for all \( i \), i.e. \( \alpha_k^i = \alpha_k^*, \forall i \), we have, for all \( k \geq 0 \),

\[
\alpha_k \left( f(\hat{x}_k) - f(x^*) \right) \leq \frac{m \alpha_k^2 D^2}{2\sigma} + \sum_{i=1}^{m} \alpha_k D \left\| v_k^i - \hat{x}_k \right\| + \sum_{i=1}^{m} B_{\mu_i}(x^*, v_k^i) - \sum_{i=1}^{m} B_{\mu_i}(x^*, x_{k+1}^i).
\]

**Proof.** From the DMD Algorithm, Eq. (1b), and the definition of Bregman divergence, we have, for any \( x \in X \) and \( i \),

\[
\langle x - x_{k+1}^i, \nabla_{\mu_i}(v_k^i) - \nabla_{\mu_i}(x_{k+1}^i) - \alpha_k^i d_k^i \rangle \leq 0.
\]

Thus, in particular when \( x = x^* \), we obtain

\[
\langle x^* - x_{k+1}^i, \nabla_{\mu_i}(v_k^i) - \nabla_{\mu_i}(x_{k+1}^i) - \alpha_k^i d_k^i \rangle \leq 0. \tag{10}
\]

Using the subgradient inequality for the convex function \( f_i \), it follows that for all \( i \),

\[
\alpha_k^i \left( f_i(v_k^i) - f_i(x^*) \right) \leq \alpha_k^i \left( v_k^i - x^*, d_k^i \right) = \langle x^* - x_{k+1}^i, \nabla_{\mu_i}(v_k^i) - \nabla_{\mu_i}(x_{k+1}^i) - \alpha_k^i d_k^i \rangle \tag{11a}
\]

\[
+ \langle x^* - x_{k+1}^i, \nabla_{\mu_i}(x_{k+1}^i) - \nabla_{\mu_i}(v_k^i) \rangle \tag{11b}
\]

\[
+ \langle v_k^i - x_{k+1}^i, \alpha_k^i d_k^i \rangle. \tag{11c}
\]
We now analyze the three items on the RHS of Eq. (11). Directly from Eq. (10), the term (11a) is not greater than zero. We represent (11b) using the three-points identity (see Lemma 7 in Appendix) of Bregman divergence, i.e.

\[
\langle x^* - x^{k+1}, \nabla \mu_i(x^{k+1}) - \nabla \mu_i(v^k_i) \rangle = B_{\mu_i}(x^*, v^k_i) - B_{\mu_i}(x^*, x^{k+1}) - B_{\mu_i}(x^{k+1}, v^k_i).
\]

Following from \( \langle a, b \rangle \leq \frac{\sigma}{2} \|a\|^2 + \frac{1}{2\sigma} \|b\|^2, \forall a, b \in \mathbb{R}^p, \sigma \in \mathbb{R}^+, \) Eq. (11c) is bounded by

\[
\langle v^k_i - x^{k+1}, \alpha_i^k d_i^k \rangle \leq \frac{\sigma}{2} \|v^k_i - x^{k+1}\|^2 + \frac{1}{2\sigma} (\alpha_i^k)^2 \|d_i^k\|^2.
\]

Therefore, Eq. (11) is equivalent to

\[
\alpha_i^k \left( f_i(v^k_i) - f_i(x^*) \right) \leq B_{\mu_i}(x^*, v^k_i) - B_{\mu_i}(x^*, x^{k+1}) - B_{\mu_i}(x^{k+1}, v^k_i) + \frac{\sigma}{2} \|v^k_i - x^{k+1}\|^2 + \frac{1}{2\sigma} (\alpha_i^k)^2 \|d_i^k\|^2.
\]

Since \( B_{\mu_i}(x, y) \geq \frac{\sigma}{2} \|x - y\|^2, \forall i, x, y, \) due to convexity of the distance generating function, \( \mu_i, \) see Eq. (4), it follows that

\[
\alpha_i^k \left( f_i(v^k_i) - f_i(x^*) \right) \leq B_{\mu_i}(x^*, v^k_i) - B_{\mu_i}(x^*, x^{k+1}) + \frac{1}{2\sigma} (\alpha_i^k)^2 \|d_i^k\|^2. \tag{12}
\]

Considering the boundedness of the subgradient (Eq. (3)), we obtain the desired result, (a), in the lemma’s statement.

We now consider statement (b) in the lemma. When \( \alpha_i^k = \alpha_k \) for all agents, adding and subtracting \( \alpha_k f_i(\hat{x}_k) \) in the RHS of Eq. (12) imply that

\[
\alpha_k \left( f_i(\hat{x}^k) - f_i(x^*) \right) \leq B_{\mu_i}(x^*, v^k_i) - B_{\mu_i}(x^*, x^{k+1}) + \frac{\alpha_k^2}{2\sigma} \|d_i^k\|^2 + \alpha_k \left( f_i(\hat{x}^k) - f_i(v^k_i) \right). \tag{13}
\]

We use the first order properties of a convex function:

\[
f_i(\hat{x}^k) - f_i(v^k_i) \leq \|f_i(v^k_i) - f_i(\hat{x}^k)\| \leq \|d_i^k\| \|v^k_i - \hat{x}^k\|;
\]

Eq. (13) now becomes

\[
\alpha_k \left( f_i(\hat{x}^k) - f_i(x^*) \right) \leq B_{\mu_i}(x^*, v^k_i) - B_{\mu_i}(x^*, x^{k+1}) + \frac{\alpha_k^2}{2\sigma} D^2 + \alpha_k D \|v^k_i - \hat{x}^k\|.
\]

December 18, 2014
The proof follows by summing over \( i = 1, \cdots, m \). \( \square \)

In this section, we provide three relations bounding the agent state, \( x_k^i \). Lemma 1 shows the consensus properties of DMD by capturing the decrease in \( \|x_k^i - \hat{x}^k\| \), for all \( i \). Lemmas 2 and 3 quantify the distance between the agent states and optimal solution of Problem (P1). Relying on these three lemmas, we show the upper bounds provided in Lemmas 1, 2, and 3 go to zero as \( k \to \infty \), in Section IV.

IV. CONVERGENCE OF DISTRIBUTED MIRROR DESCENT

We now prove the convergence of DMD using Lemmas 1, 2 and 3. To outline the main idea of the proof, we note that Lemma 1 provides an upper bound on the distance, \( \|x_k^i - \hat{x}^k\| \), between each agent state and the accumulation point. Lemma 2 provides an upper bound on the Bregman divergence, \( B_{\mu_i}(x^*, x_k^i) \), while Lemma 3 provides an upper bound on \( f(\hat{x}^k) - f(x^*) \). To show the convergence of DMD to an optimal solution, it remains to relate the accumulation point, \( \hat{x}^k \), to the optimal solution, \( x^* \), of Problem (P1). We will show that as \( k \to \infty \), the value of the objective function at the accumulation point, \( f(\hat{x}^k) \), converges to the optimal value, \( f^* \).

A special case when the agents have identical constraints, i.e. \( X_i = X, \forall i \), is discussed first in this section. Following that, we consider the case when the constraint sets, \( X_i \)'s, are different convex (compact) sets.

A. Convergence with Identical Constraints

In proving the convergence of the DMD algorithm when the agents have identical constraints, our assumptions are the same as those in the existing literature, [9–14], except that we restrict the weight matrices to be row-stochastic instead of doubly-stochastic. In particular, given Assumptions A1A5, we assume that the step size at each agent is the same over time, i.e. \( \alpha_i^k = \alpha_k, \forall i, k \). We prove DMD to converge in time-varying graphs without any additional knowledge of network or agents. We start with an upper bound on the norm of the perturbed subgradient, \( \|e_i^k\| \), in the following lemma.

**Lemma 4.** Let Assumptions A3 and A4 hold. Let \( \{e_i^k\} \) be the sequence (over \( k \)) generated by the DMD, Eqs. (2). Then for all \( i \) and \( k \geq 0 \), \( e_i^k \) satisfies:

\[
\|e_i^k\| \leq \sqrt{2D} \frac{\alpha_i^k}{\sigma}.
\]
Proof. From the definition of $e^k_i$ in Eq. (2b) and the strong convexity, Eq. (4), of the Bregman divergence, we have
\[
\|e^k_i\|^2 = \|x^{k+1}_i - v^k_i\|^2 \leq \frac{2}{\sigma} B_{\mu_i}(v^k_i, x^{k+1}_i).
\tag{14}
\]
Since $v^k_i$ (see Eq. (2a)) is a linear-convex combination of agent states at time $k$ and each agent state lies in the same constraint set, $\mathcal{X}$, it follows that $v^k_i \in \mathcal{X}, \forall i$. Therefore, we are able to apply the non-expansive property (see Lemma 6 in the Appendix) of Bregman divergence, i.e.
\[
B_{\mu_i}(v^k_i, x^{k+1}_i) \leq B_{\mu_i}(v^k_i, \hat{x}^k_i),
\tag{15}
\]
see Eq. (7) for $\hat{x}_i$. For any convex function, $\mu_i$, it is always true that
\[
\mu_i(v^k_i) - \mu_i(\hat{x}^k_i) \leq \langle \nabla \mu_i(v^k_i), v^k_i - \hat{x}^k_i \rangle.
\]
Therefore,
\[
B_{\mu_i}(v^k_i, \hat{x}^k_i) \leq \langle \nabla \mu_i(\hat{x}^k_i) - \nabla \mu_i(v^k_i), \hat{x}^k_i - v^k_i \rangle,
\]
\[
\leq \frac{1}{\sigma} \| \nabla \mu_i(\hat{x}^k_i) - \nabla \mu_i(v^k_i) \|^2,
\]
\[
\leq \frac{D^2}{\sigma} (\alpha^k_i)^2,
\tag{16}
\]
and the lemma follows by Assumption A4 (a) and Eq. (8).

Using the above lemma, the following result improves Lemma 1 showing that any two sequences, $\{x^k_i\}$ and $\{x^k_j\}$, generated by DMD have the same limit accumulation.

**Proposition 1.** Let Assumptions A1-A6 hold. Let $\{x^k_i\}$ be the sequence (over $k$) generated by the DMD, Eqs. (2), and $\hat{x}^k$ be given by Eq. (6). Assume that $\alpha^k_i = \alpha_k, \forall i$. Then, for any $i$:
\[
\sum_{k=1}^{\infty} \alpha_k \| x^k_i - \hat{x}^k \| < \infty.
\]

**Proof.** Adopting the boundedness of the perturbed subgradient (Lemma 4) in the result of Lemma 1 we get
\[
\sum_{k=1}^{n} \alpha_k \| x^k_i - \hat{x}^k \| \leq 2 \Gamma \left( \sum_{j=1}^{m} \| x^0_j \| \right) \sum_{k=1}^{n} \alpha_k \gamma^{(k-1)} + \frac{2\sqrt{2}m \Gamma \Delta}{\sigma} \sum_{k=1}^{n} \alpha_k \gamma^{(k-1)} + \frac{2\sqrt{2}D}{\sigma} \sum_{k=0}^{n-1} \alpha_k^2.
\tag{17}
\]
With the basic inequality \( ab \leq \frac{1}{2}(a^2 + b^2) \), \( a, b \in \mathbb{R} \), we have:

\[
2 \sum_{k=1}^{n} \alpha_k \gamma^{k-1} \leq \sum_{k=1}^{n} \left( \alpha_k^2 + \gamma^{2(k-1)} \right) \leq \sum_{k=1}^{n} \alpha_k^2 + \frac{1}{1 - \gamma^2};
\]

\[
\sum_{k=1}^{n} \sum_{r=1}^{k-1} \gamma^{(k-1-r)} \alpha_k \alpha_r \leq \frac{1}{2} \sum_{k=1}^{n} \alpha_k^2 \sum_{r=1}^{k-1} \gamma^{(k-1-r)}
\]

\[+ \frac{1}{2} \sum_{r=1}^{n-1} \alpha_r^2 \sum_{k=r+1}^{n} \gamma^{(k-1-r)} \leq \frac{1}{1 - \gamma} \sum_{k=1}^{n} \alpha_k^2. \tag{18}\]

The lemma follows by using the preceding relations in Eq. (17) along with \( \sum_{k=0}^{n} \alpha_k^2 < \infty \) (Assumption A5) as \( n \to \infty \).

Since \( \sum_{k=0}^{n} \alpha_k = \infty \), the result of Proposition 1 reveals that all agents converge to the same point as \( k \to \infty \). With the help of the Proposition 1, we now present our main convergence result under the weighting rules of non-doubly stochastic updates. In this analysis of Proposition 1, we will use the separate convexity of Bregman divergence (see Lemma 8 in the Appendix). In particular, we have the Bregman divergence being convex on the second variable with Assumption A4.

**Theorem 1.** Let Assumptions A7, A5 hold. Let \( \{x_k^i\} \) be the sequence (over \( k \)) generated by the DMD, Eqs. (2), \( \hat{x}^k \) be the accumulation point given by Eq. (6), and \( f^* \in \mathcal{X}^* \) be an optimal solution of the Problem (P1). Let \( \alpha_k^i = \alpha_k, \forall i \), then

\[
\lim_{k \to \infty} f(\hat{x}^k_i) = f^*, \quad \forall i.
\]

**Proof.** According to the separate convexity (see Lemma 8(b) in the Appendix) of the Bregman divergence, we have

\[
B_{\mu_i}(x^*, v_i^k) \leq \sum_{j=1}^{m} w_{ij}^k B_{\mu_i}(x^*, x_j^k).
\]

Substituting the above into the result of Lemma 3(b), we get

\[
\alpha_k \left( f(\hat{x}^k) - f^* \right) \leq \sum_{j=1}^{m} \left( \sum_{i=1}^{m} w_{ij}^k \right) \alpha_k D \|x_j^k - \hat{x}^k\|
\]

\[+ \sum_{j=1}^{m} \sum_{i=1}^{m} w_{ij}^k B_{\mu_i}(x^*, x_j^k) - \sum_{j=1}^{m} B_{\mu_j}(x^*, x_j^{k+1}) + \frac{mD^2}{2\sigma} \alpha_k \]

December 18, 2014 DRAFT
Note that $\sum_i w_{ij}^k \leq m$ and sum the preceding relation over $k$:

$$
\sum_{k=0}^N \alpha_k \left( f(\hat{x}^k) - f^* \right) \leq mD \sum_{j=1}^m \sum_{k=0}^N \alpha_k \|x_j^k - \hat{x}^k\|
$$

$$
\sum_{k=1}^N \left( \sum_{i=1}^m \sum_{j=1}^m w_{ij}^k B_{\mu_i}(x^*, x_{ij}^k) - \sum_{j=1}^m B_{\mu_j}(x^*, x_j^k) \right)
$$

$$
+ \frac{mD^2}{2\sigma} \sum_{k=0}^N \alpha_k^2,
$$

(19)

$$
:= s_1 + s_2 + s_3 + s_4,
$$

where $s_1, s_2, s_3, s_4$ denote each of RHS terms in Eq. (19). We have, for any $N$: $s_1 < \infty$ by Prop. 1, $s_2 < \infty$ by Assumption A3 and A4, and, $s_4 < \infty$ by Assumption A5.

We now show that $s_3 < \infty$ for any $N > 0$. Denote $y^k, \bar{y}^k$ as the first and second terms in $s_3$, i.e.

$$
y^k = \sum_{j=1}^m \sum_{i=1}^m w_{ij}^k B_{\mu_i}(x^*, x_{ij}^k), \quad \bar{y}^k = \sum_{j=1}^m B_{\mu_j}(x^*, x_j^k).
$$

Note that $y^k$ is a variable in terms of the weights, $w_{ij}^k$s; we choose suitable weights such that $y^k_{\text{max}}, y^k_{\text{min}}$ is the maximum and minimum of $y^k$ for fixed $k$, i.e.

$$
y^k_{\text{max}} = \max_{w_{ij}^k} \left\{ \sum_{j=1}^m \sum_{i=1}^m w_{ij}^k B_{\mu_i}(x^*, x_{ij}^k) \right\},
$$

$$
y^k_{\text{min}} = \min_{w_{ij}^k} \left\{ \sum_{j=1}^m \sum_{i=1}^m w_{ij}^k B_{\mu_i}(x^*, x_{ij}^k) \right\}.
$$

Since $\hat{x}^k$ is a linear-convex combination of agent states at time $k$, with all states $x_i^k \in \mathcal{X}$, it follows that $\hat{x}^k \in \mathcal{X}$. Therefore, $f(\hat{x}^k) \geq f^*$ for all $k$. In particular, we have $\sum_{k=0}^N \alpha_k \left( f(\hat{x}^k) - f^* \right) \geq 0$ for any $N$, which reveals that $s_3 > -\infty$ for any $N$, i.e. for all $N$ and $y^k$,

$$
\sum_{k=1}^N (y^k - \bar{y}^k) > -\infty.
$$

Specially, when $y^k = y^k_{\text{min}}$, we obtain,

$$
\sum_{k=1}^N (y^k_{\text{min}} - \bar{y}^k) > -\infty.
$$
Since \( x_i^k \in \mathcal{X} \) for all \( i \) and \( k \), and \( \mathcal{X} \) is compact (Assumption [A3]), it follows that the sequence \( \{ x_i^k \} \) is (element-wise) finite for all \( i \) and \( k \). Combining with Assumption [A4] we note that \( y_{\text{max}}^k - \bar{y}^k \) and \( \bar{y}^k - y_{\text{min}}^k \) are finite for all \( k \), because the Bregman divergence is finite. Besides, due to the fact that \( y_{\text{max}}^k - \bar{y}^k \) and \( \bar{y}^k - y_{\text{min}}^k \) are always positive for non doubly-stochastic matrices, it follows that for any \( k \), there always exists some bounded positive constant, \( R \), such that
\[
y^k - \bar{y}^k \leq y_{\text{max}}^k - \bar{y}^k \leq R(\bar{y}^k - y_{\text{min}}^k).
\]
Summing the preceding relation over \( k = 1, \cdots, N \), we obtain
\[
s_3 = \sum_{k=1}^{N} (y^k - \bar{y}^k) \leq R \sum_{k=1}^{N} (\bar{y}^k - y_{\text{min}}^k) < \infty.
\]
Finally, it follows from Eq. (19) that
\[
\sum_{k=0}^{N} \alpha_k (f(\hat{x}^k) - f^*) < \infty, \quad \forall N.
\]
The theorem follows by letting \( N \to \infty \) and noting that \( \sum_{k=0}^{\infty} \alpha_k = \infty \) and \( (f(\hat{x}^k) - f^*) \geq 0 \), for all \( k \).

In the existing literature, [9–14], Distributed Projected Subgradient (DPS) method assumes the weight matrices to be doubly-stochastic, i.e. \( \sum_{i=1}^{m} w_{ij}^k = 1, \forall j \), which simplifies the proof of Theorem 1. In particular, if we let the weight matrices to be doubly-stochastic, \( s_3 \) in Eq. (19) is 0. This also reveals the fact that each agent “contributes equally” in optimizing the Problem (P1). When we restrict the weigh matrices to be row-stochastic, \( s_3 \) in Eq. (19) does not vanish. Since it is a summation of an infinite numbers terms, where some can be positive and others negative; bounding \( s_3 \) is non-trivial. The spirit of the proof is that if \( s_3 \) is larger than negative infinity, it should be less than positive infinity due to compactness of the constraint sets at all of the agents.

Theorem 1 shows the convergence of DMD in a time-varying graph when the agents possess identical constraints. The assumption used in the proof is the same as, e.g. in [9–14], with no additional knowledge on either the graph topology (required e.g. in [31–33]), or out-degree of the agents (required e.g. in [29, 30]). Since DPS method is a special case of DMD, i.e. when the Bregman divergence is Euclidean squared, we note that DPS methods may also be extended to directed graphs.

---

1 We restrict to non doubly-stochastic updates, because \( y_{\text{max}}^k = \bar{y}^k = y_{\text{min}}^k \) otherwise, and thus \( s_3 = 0 \) in Eq. (19).
B. Convergence Analysis with Different Constraints

We now provide convergence analysis for the case when the constraint sets, \( X_i \)'s, are different. We show that when the constraints are different, the agent states \( x^k_i, \forall i \), converge to an optimal solution of Problem (P1) under some conditions, which can be realized in a distributed multi-agent network. In particular, we prove the convergence of DMD in fixed topologies, i.e. \( \mathcal{E}_k = \mathcal{E} \), and adopt the following assumption on designing the distance generating function.

**Assumption A6.** *The distance generating function (in the Bregman divergence) is identical at all agents, i.e. \( \mu_i = \mu, \forall i \).*

We emphasize that even though Assumption [A6] is more restrictive, it is not uncommon in related literature. For example, DPS methods require each agent to use the same Bregman divergence, i.e. Euclidean squared. Under this assumption, we show that the perturbed subgradient, \( e^k_i \), converges to zero for all \( i \), in the following lemma.

**Lemma 5.** Let Assumptions [A1]-[A6] hold. Let \( \{x^k_i\} \) and \( \{e^k_i\} \) be the sequences (over \( k \)) generated by the DMD, Eqs. (2), then the perturbed subgradient, \( e^k_i \), converges to 0 for all \( i \), i.e.

\[
\lim_{k \to \infty} \|e^k_i\| = 0, \quad \forall i.
\]

**Proof.** Under the Assumption [A6] we adopt the separate convexity (see Lemma 8(b) in the Appendix) of Bregman divergence in the result of Lemma 2

\[
B_\mu(x^*, x_i^{k+1}) \leq \sum_{j=1}^{m} w^k_{ij} B_\mu(x^*, x_j^k) - B_\mu(x_i^{k+1}, v^k_i) + \alpha^k_i D \|x^* - x_i^{k+1}\|.
\] (20)

We consider a weighted sum of the preceding relation, Eq. (20), as follows:

\[
\sum_{j=1}^{m} \pi^k_j B_\mu(x^*, x_j^{k+1}) \leq \sum_{j=1}^{m} \left( \sum_{i=1}^{m} \pi^k_i w^k_{ij} \right) B_\mu(x^*, x_j^k) - \sum_{j=1}^{m} \pi^k_j B_\mu(x_j^{k+1}, v^k_j) + D \sum_{j=1}^{m} \alpha^k_i \pi^k_j \|x^* - x_j^{k+1}\|.
\] (21)

where \( \pi^k = [\cdots, \pi^k_j, \cdots] \) is the left eigenvector of row-stochastic matrix, \( W(k) \), satisfying \( \pi^k_j = \sum_{i=1}^{m} \pi^k_i w^k_{ij} \). Since \( W(k) \) is row-stochastic, we know that \( \sum_{j=1}^{m} \pi^k_j = 1 \).
Considering the compactness of the constraint sets (Assumption \text{A3}) and the continuity of Bregman divergence (Assumption \text{A4}), it is true that the sequence \( \{x_i^k\} \) is finite for all \( i, k \), and therefore the sequence, \( \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^k) \), is finite. Since the step-size is diminishing, i.e. \( \alpha_i^k \to 0, \forall i \), we note that \( \lim_{k \to \infty} \sum_{j=1}^{m} \alpha_j^k \pi_j^k \|x^* - x_j^{k+1}\| \) exists. By dropping \( \sum_{j=1}^{m} \pi_j^k B_{\mu}(x_j^{k+1}, v_j^k) \) in Eq. (21), we get
\[
\limsup_{k \to \infty} \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^{k+1}) \leq \liminf_{k \to \infty} \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^k) + \lim_{k \to \infty} \left( D \sum_{j=1}^{m} \alpha_j^k \pi_j^k \|x^* - x_j^{k+1}\| \right).
\]
Since the second term in the preceding relation is zero, this implies that \( \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^k) \) is convergent. Therefore, by rearranging Eq. (21), and letting \( k \to \infty \), we obtain
\[
\limsup_{k \to \infty} \sum_{j=1}^{m} \pi_j^k B_{\mu}(x_j^{k+1}, v_j^k) \leq \lim_{k \to \infty} \left( \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^k) - \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^{k+1}) \right) + \lim_{k \to \infty} \left( D \sum_{j=1}^{m} \alpha_j^k \pi_j^k \|x^* - x_j^{k+1}\| \right). \tag{22}
\]
Since the first term on the RHS of the Eq. (22) is zero by the convergence of the sequence \( \sum_{j=1}^{m} \pi_j^k B_{\mu}(x^*, x_j^k) \), and the second term equals to zero by \( \lim_{k \to \infty} \alpha_k = 0 \), we have \( \sum_{j=1}^{m} \pi_j^k B_{\mu}(x_j^{k+1}, v_j^k) \) converges to zero when \( k \to \infty \). Moreover, we adopt the convexity property of the distance generating function, Eq. (4), such that the perturbed subgradient of any agent, \( e_i^k \), satisfies
\[
\lim_{k \to \infty} \sum_{j=1}^{m} \pi_j^k \|e_j^k\|^2 = \lim_{k \to \infty} \sum_{j=1}^{m} \pi_j^k \|x_j^{k+1} - v_j^k\|^2, \leq \lim_{k \to \infty} \frac{2}{\sigma} \sum_{j=1}^{m} \pi_j^k B_{\mu}(x_j^{k+1}, v_j^k) = 0,
\]
from which we obtain the desired result.

Using the fact that the perturbed subgradient, \( e_i^k \), converges to zero for all \( i \), we next refine the upper bound provided in the result of Lemma 1 in Section III in the following proposition.

**Proposition 2.** Let Assumptions \text{A1}-\text{A6} hold. Let \( \{x_i^k\} \) be the sequence over \( k \) generated by DMD, Eqs. (2), and \( \hat{x}^k \) be given in Eq. (6). Then, for all \( i \), we have:
\[
\lim_{k \to \infty} \|x_i^k - \hat{x}^k\| = 0.
\]
Proof. Consider Lemma II when $k \to \infty$,

\[
\lim_{k \to \infty} \|x^k - \hat{x}^k\| \leq 2\Gamma \sum_{j=1}^{m} \|x^0_j\| \lim_{k \to \infty} \gamma^{k-1} \\
+ 2\Gamma \sum_{j=1}^{m} \lim_{k \to \infty} \left( \sum_{r=1}^{k-1} \gamma^{k-1-r} \|e^{r-1}_j\| \right) \\
+ \lim_{k \to \infty} \left( \|e^{k-1}_i\| + \sum_{j=1}^{m} \theta_j \|e^{k-1}_j\| \right),
\]

where the first term on the RHS is zero, the second term is zero by infinite summability (see Lemma 10 in the Appendix), and the third term equals zero by $\lim_{k \to \infty} \|e^k_i\| = 0$ (Lemma 5).

We now make two additional assumptions on weighting rules and step-sizes, which are crucial to the main result when the constraint sets are different. Since the weight matrices are not doubly-stochastic, each agent has an unequal contribution; to make this contribution the same as the DMD progresses, we allow each agent to design its own step-sizes (in a distributed manner) that results in balancing the agent contributions. In particular, each agent that “contributes” less chooses a larger step-size, while the agents that “contribute” more choose a lower step-size.

Assumption A7. Assume the graph to be fixed, i.e. $\mathcal{E}_k = \mathcal{E}$. For any agent $i$, it assigns equal weights to its in-neighbors, i.e. $w_{ij} = 1/|\mathcal{N}_i^{in}|$, $\forall j \in \mathcal{N}_i^{in}$.

It is obvious that the weight matrices following Assumption A7 are row-stochastic, satisfying Assumption A2.

Assumption A8. Each agent $i$ at time $k$ designs its step-size as $\alpha_i^k = \frac{1}{|\mathcal{N}_i^{in}|} \alpha_k$, where $\alpha_k$ satisfies Assumption A5 i.e. $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

Assumptions requiring the knowledge of out-degrees can be founded in [29, 30], which we do not require. Clearly, in-degrees are already known to each agent. The next theorem provides the convergence result with different constraint sets.

Theorem 2. Let Assumptions A7, A8 hold. Let $\{x^k_i\}$ be the sequence (over $k$) generated by the DMD, Eqs. (2), $\hat{x}^k$ be given by Eq. (6), and $f^* \in \mathcal{X}^*$ be an optimal solution of the Problem (P1). Then,

$$
\lim_{k \to \infty} f(\hat{x}^k) = f^*.
$$
Proof. By adopting the separate convexity of Bregman divergence (Lemma 8(b)) into the result of Lemma 3(a) in Section III, it follows that
\begin{equation}
\alpha^k_i \left( f_i(v^k_i) - f_i(x^*) \right) \leq \sum_{j=1}^m w_{ij} B_\mu(x^*, x^k_j) - B_\mu(x^*, x^{k+1}_i) + \frac{D^2}{2\sigma} \left( \alpha^k_i \right)^2. \tag{23}
\end{equation}

We consider a weighted sum over agents of the preceding relation, Eq. (23), with agent \( i \) weighting \( \lambda_i = |N^\text{in}_i| \):
\begin{equation}
\sum_{j=1}^m \lambda_j \alpha^k_j \left( f_j(v^k_j) - f_j(x^*) \right) \leq \sum_{j=1}^m \left( \sum_{i=1}^m \lambda_i w_{ij} \right) B_\mu(x^*, x^k_j) - \sum_{j=1}^m \lambda_j B_\mu(x^*, x^{k+1}_j) + \frac{D^2}{2\sigma} \sum_{j=1}^m \lambda_j \left( \alpha^k_j \right)^2. \tag{24}
\end{equation}

Apply the weight rule satisfying Assumption A7 and the step-size satisfying Assumption A8 in Eq. (24), and note that \( \lambda_j \alpha^k_j = \alpha^k \), and \( \sum_{i=1}^m \lambda_i w_{ij} = |N^\text{out}_j| \). Therefore, Eq. (24) becomes
\begin{equation}
\sum_{j=1}^m \alpha^k \left( f_j(v^k_j) - f_j(x^*) \right) \leq \sum_{j=1}^m |N^\text{out}_j| B_\mu(x^*, x^k_j) - \sum_{j=1}^m |N^\text{in}_j| B_\mu(x^*, x^{k+1}_j) + \frac{D^2}{2\sigma} \sum_{j=1}^m \frac{1}{|N^\text{in}_j|} \alpha^k. \tag{25}
\end{equation}

Considering the Three-points Identity (see Lemma 7 in the Appendix), we have
\begin{equation}
B_\mu(x^*, x^k_j) = B_\mu(x^*, \tilde{x}^k) + B_\mu(\tilde{x}^k, x^k_j) - \langle \nabla \mu(x^k_j), x^* - \tilde{x}^k \rangle. \tag{26}
\end{equation}

From Proposition 2, we know that all agent accumulate to the same point as \( k \to \infty \), which means for any \( \epsilon \), there exists some \( K \) such that for \( k > K \), \( B_\mu(\tilde{x}^k, x^k_j) < \epsilon, \forall i \). So Eq. (26) becomes
\begin{equation}
B_\mu(x^*, x^k_j) \leq B_\mu(x^*, \tilde{x}^k) + \epsilon + L \| x^k_j - \tilde{x}^k \| \| x^* - \tilde{x}^k \| = B_\mu(x^*, \tilde{x}^k) + \epsilon + L \| x^* - \tilde{x}^k \|, \tag{27}
\end{equation}

where \( L \) is the Lipschitz constant for \( \mu \). Similarly, we get
\begin{equation}
-B_\mu(x^*, x^{k+1}_j) \geq -B_\mu(x^*, \tilde{x}^{k+1}) - \epsilon - L \| x^* - \tilde{x}^{k+1} \|. \tag{28}
\end{equation}
Substitute Eqs. (27) and (28) into Eq. (25) and note that for any graph \( \sum_{j=1}^{m} |\mathcal{N}_{j}^{\text{in}}| = \sum_{j=1}^{m} |\mathcal{N}_{j}^{\text{out}}| \), we obtain

\[
\sum_{j=1}^{m} \alpha_{k} \left( f_{j}(v_{j}^{k}) - f_{j}(x^{*}) \right) \leq \frac{D^{2}}{2\sigma} \sum_{j=1}^{m} \frac{1}{|\mathcal{N}_{j}^{\text{in}}|} \alpha_{k}^{2} \\
+ \sum_{j=1}^{m} |\mathcal{N}_{j}^{\text{in}}| \left( B_{\mu}(x^{*}, \hat{x}^{k}) - B_{\mu}(x^{*}, \hat{x}^{k+1}) \right) \\
+ \sum_{j=1}^{m} |\mathcal{N}_{j}^{\text{in}}| \left( L \epsilon \left( \|x^{*} - \hat{x}^{k}\| - \|x^{*} - \hat{x}^{k+1}\| \right) \right). \tag{29}
\]

We show that the preceding relation, Eq. (29), implies

\[
\liminf_{k \to \infty} \sum_{i=1}^{m} f_{i}(v_{i}^{k}) \leq f^{*}, \tag{30}
\]

by contradiction. Suppose that Eq. (30) is not true, i.e. \( \liminf_{k \to \infty} \sum_{i=1}^{m} f_{i}(v_{i}^{k}) > f^{*} \), then there exists some \( K \) and \( \xi > 0 \) such that for all \( k > K \), we have for all \( i \),

\[
\sum_{i=1}^{m} f_{i}(v_{i}^{k}) > f^{*} + \xi.
\]

Summing the relation, Eq. (29), from time \( K \) to \( N \), we get

\[
\sum_{k=K}^{N} \alpha_{k} \xi < \sum_{k=K}^{N} \alpha_{k} \left( \sum_{j=1}^{m} f_{j}(v_{j}^{k}) - f^{*} \right) \leq \sum_{j=1}^{m} |\mathcal{N}_{j}^{\text{in}}| \left( B_{\mu}(x^{*}, \hat{x}^{K}) - B_{\mu}(x^{*}, \hat{x}^{N+1}) \right) \\
+ L \epsilon \sum_{j=1}^{m} |\mathcal{N}_{j}^{\text{in}}| \left( \|x^{*} - \hat{x}^{K}\| - \|x^{*} - \hat{x}^{N+1}\| \right) \\
+ \frac{D^{2}}{2\sigma} \sum_{j=1}^{m} \frac{1}{|\mathcal{N}_{j}^{\text{in}}|} \sum_{k=K}^{N} \alpha_{k}^{2}. \tag{31}
\]

When \( N \to \infty \), the LHS of Eq. (31) goes to infinity while the RHS of Eq. (31) is finite with Assumption A8 therefore, we reach a contradiction; hence, Eq. (30) is true. Also considering the stochasticity of weighting matrix \( W \), Proposition 2 reveals that

\[
\lim_{k \to \infty} \|v_{i}^{k} - \hat{x}^{k}\| \leq \lim_{k \to \infty} \sum_{j=1}^{m} w_{ij} \|x_{i}^{k} - \hat{x}^{k}\| = 0. \tag{32}
\]

Combining Eq. (30) and (32), we obtain

\[
\liminf_{k \to \infty} f(\hat{x}^{k}) \leq f^{*}.
\]
Note that the proof of Lemma 5 shows that the sequence \( \left\{ \sum_{j=1}^{m} \pi_j B_{\mu}(x^*, x_j^k) \right\} \) is convergent, which implies that \( \left\{ \| \hat{x}^k - x^* \| \right\} \) is convergent given Proposition 2. Therefore, \( \hat{x}^k \) must have a limit point, i.e.

\[
\liminf_{k \to \infty} f(\hat{x}^k) = f^*.
\]

Using the continuity of \( f \), this implies that one of the limit points of \( \hat{x}^k \) must belong to the optimal set, \( X^* \); denote this limit point by \( x^* \). Since the sequence \( \left\{ \| \hat{x}^k - x^* \| \right\} \) is convergent, it follows that \( \hat{x}^k \) has a unique limit point, thus completing the proof.

We explain the spirit of proof of Theorem 2 as follows. The objective of Problem (P1) is to minimize a sum of private objective functions, i.e. \( f = \sum_{i=1}^{m} f_i \), whose subgradient is also the sum of each private function’s subgradient. This reveals that in the long run of the DMD algorithm, all agent should “contribute equally” their subgradient information to the network. When weight matrices are doubly stochastic, this is achieved due to the fact that the column sum of doubly stochastic matrices are same. On the other hand, when the weight matrices are row-stochastic, each agent contributes differently. We force all agents to contribute equally by setting their step-size differently.

V. Conclusions

In this paper, we implement a distributed optimization algorithm to minimize a sum of convex functions over directed graphs, that we refer to as Distributed Mirror Descent (DMD). DMD generalizes the distributed projected subgradient methods by using Bregman divergence instead of a global Euclidean squared distance. Our convergence proof is based on the communication described by a directed graph. We establish the convergence of the algorithm in two cases: (i) when the constraint sets of agents are the same; (ii) when the constraint sets of agents are different. When the constraint sets are assumed to be the same, each agent designs its own local Bregman divergence. The results are applicable to time-varying networks, requiring the same knowledge as distributed optimization algorithms proposed in previous literature for undirected graphs. When the constraint sets are different for each agent, the Bregman divergence is required to be global. The results are applicable to fixed topologies and the underlying algorithm is fully distributed. By partially following the spirit of our proof, it can be shown that a class of existing consensus-based optimization algorithms, restricted to doubly-stochastic matrices, remain convergent with non-doubly stochastic matrices.
A. Preliminaries

The proof in this paper relies on some existing results that we present in the following for reference.

**Non-Expansive Property**: For all \(i\) and \(x \in \mathbb{R}^p\), define \(P_i[x]\) as a point in agent \(i\)'s constraint set satisfying \(B_{\mu_i}(P_i[x], x) = \min_{y \in X_i} B_{\mu_i}(y, x)\).

**Lemma 6**. (Bregman [22]) Let Assumption \([A4]\) hold and choose some \(z \in X\). For any \(i\) and \(x \in \mathbb{R}^p\), it follows that
\[
B_{\mu_i}(P_i[x], x) \leq B_{\mu_i}(z, x) - B_{\mu_i}(z, P_i[x]).
\]

**Three-points Identity**: The Bregman divergence satisfies a simple identity, which appears to be a generalization of Euclidean distance.

**Lemma 7**. (Chen and Teboulle [41]) Let \(\mu\) be a distance generating function satisfying the Assumption \([A4]\) and \(B_{\mu}\) be the Bregman divergence based on \(\mu\). Then for any three points \(x, y, z \in \text{dom}(\mu)\) the following identity holds:
\[
B_{\mu}(z, x) + B_{\mu}(x, y) - B_{\mu}(z, y) = \langle \nabla \mu(y) - \nabla \mu(x), z - x \rangle.
\]

**Separate convexity**: It is obvious that the Bregman divergence is convex in the first variable with the convexity of distance generating function. The following result provides the condition of Bregman divergence convexity in the second variable.

**Lemma 8**. (Bauschke and Borwein [42]) The Bregman divergence is separately convex for all \(i\) if and only if Assumption \([A4]\) holds. In particular, separate convexity means, for any \(i\) and \(\sum_{i=1}^m \theta_i = 1\),
\[
(a) \quad B_{\mu_i}(\sum_{i=1}^m \theta_i x_i, y) \leq \sum_{i=1}^m \theta_i B_{\mu_i}(x_i, y), \quad \forall x_i, y;
\]
\[
(b) \quad B_{\mu_i}(x, \sum_{i=1}^m \theta_i y_i) \leq \sum_{i=1}^m \theta_i B_{\mu_i}(x, y_i), \quad \forall x, y_i.
\]

**Transition Matrix Convergence**: We use the following lemma, which states a result on the convergence of the transition matrix, \(\Phi(k, r)\), defined in Eq. (5).

**Lemma 9**. (Nedic et al. [43]) Let Assumptions \([A7]\) and \([A2]\) hold. Then:
\[
(a) \quad \text{The limit } \Phi(r) = \lim_{k \to \infty} \Phi(k, r) \text{ exists for each } r.
\]
(b) The limit matrix $\Phi(r)$ has identical rows and the rows are stochastic, i.e.

$$\Phi(r) = 1\phi(r)^T,$$

where $\phi(r) \in \mathbb{R}^m$ is a stochastic vector for each $r$.

(c) For every $j \in \{1, ..., m\}$ and all $r$, the entries $[\Phi(k,r)]_{ij}$ and $\phi_j(r)$ satisfy

$$|[\Phi(k,r)]_{ij} - \phi_j(r)| \leq \Gamma \gamma^{k-r}, \quad \forall i,$$

where $\Gamma = (1 - \frac{n}{4m^2})^{-2}$ and $\gamma = (1 - \frac{n}{4m^2})^{\frac{1}{L_1}}$.

Infinite Summability: We consider infinite summability of products of positive scalar sequences with certain properties.

**Lemma 10.** (Lobel et al. [10]) Let $\{\beta_l\}$ and $\{\gamma_k\}$ be positive scalar sequences, such that $\sum_{l=0}^{\infty} \beta_l < \infty$ and $\lim_{k \to \infty} \gamma_k = 0$. Then,

$$\lim_{k \to \infty} \sum_{l=0}^{k} \beta_{k-l} \gamma_l = 0.$$

**References**

[1] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*, Ph.D. thesis, Dept. Elect. Eng. Comp. Sci., Massachusetts Institute of Technology, 1984.

[2] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.

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

[4] I. Necoara and J. A. K. Suykens, “Application of a smoothing technique to decomposition in convex optimization,” *IEEE Transactions on Automatic Control*, vol. 53, no. 11, pp. 2674–2679, Dec. 2008.

[5] G. Mateos, J. A. Bazerque, and G. B. Giannakis, “Distributed sparse linear regression,” *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[6] M. Rabbat and R. Nowak, “Distributed optimization in sensor networks,” in *Third International Symposium on Information Processing in Sensor Networks*, Berkeley, CA, Apr. 2004, pp. 20–27.
[7] L. Chunlin and L. Layuan, “A distributed multiple dimensional qos constrained resource scheduling optimization policy in computational grid,” Journal of Computer and System Sciences, vol. 72, no. 4, pp. 706 – 726, 2006.

[8] G. Neglia, G. Reina, and S. Alouf, “Distributed gradient optimization for epidemic routing: A preliminary evaluation,” in 2nd IFIP in IEEE Wireless Days, Paris, Dec. 2009, pp. 1–6.

[9] A. Nedic, A. Ozdaglar, and P. A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” IEEE Transactions on Automatic Control, vol. 55, no. 4, pp. 922–938, Apr. 2010.

[10] I. Lobel, A. Ozdaglar, and D. Feijer, “Distributed multi-agent optimization with state-dependent communication,” Mathematical Programming, vol. 129, no. 2, pp. 255–284, 2011.

[11] I. Lobel and A. Ozdaglar, “Distributed subgradient methods for convex optimization over random networks,” IEEE Transactions on Automatic Control, vol. 56, no. 6, pp. 1291–1306, Jun. 2011.

[12] S. S. Ram, A. Nedic, and V. V. Veeravalli, “Distributed subgradient projection algorithm for convex optimization,” in IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, Apr. 2009, pp. 3653–3656.

[13] S. S. Ram, A. Nedic, and V. V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” Journal of Optimization Theory and Applications, vol. 147, no. 3, pp. 516–545, 2010.

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

[15] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, “Subgradient methods and consensus algorithms for solving convex optimization problems,” in 47th IEEE Conference on Decision and Control, Cancun, Mexico, Dec. 2008, pp. 4185–4190.

[16] S. S. Ram, V. V. Veeravalli, and A. Nedic, “Distributed non-autonomous power control through distributed convex optimization,” in IEEE INFOCOM, Rio de Janeiro, Apr. 2009, pp. 3001–3005.

[17] G. C. Calafiore, L. Carlone, and M. Wei, “A distributed gradient method for localization of formations using relative range measurements,” in IEEE International Symposium on Computer-Aided Control System Design, Yokohama, Japan, Sep. 2010, pp. 1146–1151.
[18] C. Jianshu and A. H. Sayed, “Diffusion adaptation strategies for distributed optimization and learning over networks,” *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[19] L. Bottou and O. Bousquet, “The Tradeoffs of Large Scale Learning,” in *Advances in Neural Information Processing Systems*. 2007, pp. 161–168, MIT Press.

[20] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” *Operations Research Letters*, vol. 31, no. 3, pp. 167 – 175, 2003.

[21] A. Nemirovksi and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley Press, 1983.

[22] L. M. 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*, vol. 7, no. 3, pp. 200 – 217, 1967.

[23] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with bregman divergences,” *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[24] A. Ben-Tal, T. Margalit, and A. Nemirovski, “The ordered subsets mirror descent optimization method with applications to tomography,” *SIAM Journal on Optimization*, vol. 12, no. 1, pp. 79–108, 2001.

[25] J. Kivinen and M. K. Warmuth, “Exponentiated gradient versus gradient descent for linear predictors,” *Information and Computation*, vol. 132, no. 1, pp. 1 – 63, 1997.

[26] S. Mahadevan and B. Liu, “Sparse q-learning with mirror descent,” *CoRR*, vol. abs/1210.4893, 2012.

[27] J. Z. Kolter and A. Y. Ng, “Regularization and feature selection in least-squares temporal difference learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML ’09, pp. 521–528, ACM.

[28] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Maching Learning Research*, vol. 12, pp. 2121–2159, 2011.

[29] A. Nedic and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” *IEEE Transactions on Automatic Control*, vol. PP, no. 99, pp. 1–1, 2014.

[30] A. Nedic and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” in *IEEE 52nd Annual Conference on Decision and Control*, Florence, Italy, Dec. 2013, pp. 6855–6860.
[31] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, “Push-sum distributed dual averaging for convex optimization,” in *IEEE 51st Annual Conference on Decision and Control*, Maui, Hawaii, Dec. 2012, pp. 5453–5458.

[32] K. I. Tsianos, *The role of the Network in Distributed Optimization Algorithms: Convergence Rates, Scalability, Communication/Computation Tradeoffs and Communication Delays*, Ph.D. thesis, Dept. Elect. Comp. Eng. McGill University, 2013.

[33] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, “Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning,” in *50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Oct. 2012, pp. 1543–1550.

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

[35] C. W. Reynolds, “Flocks, herds and schools: A distributed behavioral model,” in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1987, SIGGRAPH ’87, pp. 25–34, ACM.

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

[37] R. Olfati-Saber and R. M. Murray, “Consensus protocols for networks of dynamic agents,” in *Proceedings of the 2003 American Control Conference*, Denver, Colorado, Jun. 2003, vol. 2, pp. 951–956.

[38] R. Olfati-Saber, J. A. Fax, and R. M. Murray, “Consensus and cooperation in networked multi-agent systems,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[39] L. Xiao, S. Boyd, and S. J. Kim, “Distributed average consensus with least-mean-square deviation,” *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33 – 46, 2007.

[40] Walter Rudin, *Principles of Mathematical Analysis*, McGraw-Hill Science/Engineering/Math, 1976.

[41] G. Chen and M. Teboulle, “Convergence analysis of a proximal-like minimization algorithm using bregman functions,” *SIAM Journal on Optimization*, vol. 3, no. 3, pp. 538–543, 1993.

[42] H. H. Bauschke and J. M. Borwein, “Joint and separate convexity of the bregman distance,”
in *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, D. Butnariu, Y. Censor, and S. Reich, Eds., vol. 8 of *Studies in Computational Mathematics*, pp. 23 – 36. Elsevier, 2001.

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