Q-linear Convergence of Distributed Optimization with Barzilai-Borwein Step Sizes.

Iyanuoluwa Emiola and Chinwendu Enyioha

***Abstract*—The growth in sizes of large-scale systems and data in machine learning have made distributed optimization a naturally appealing technique to solve decision problems in different contexts. In such methods, each agent iteratively carries out computations on its local objective using information received from its neighbors, and shares relevant information with neighboring agents. Though gradient-based methods are widely used because of their simplicity, they are known to have slow convergence rates. On the other hand, though Newton-type methods have better convergence properties, though they are not as applicable because of the enormous computation and memory requirements. In this work, we introduce a distributed quasi-Newton method with Barzilai-Borwein step-sizes. We prove a Q-linear convergence to the optimal solution, present conditions under which the algorithm is superlinearly convergent and validate our results via numerical simulations.**

I. Introduction

Solving distributed optimization problems have different applications in different areas including distributed learning [1, 2]. These problems are formulated in such a manner that agents in a network have to coordinate with other agents connected to them in a network to achieve a desired goal. In the distributed optimization problem below:

$$\min_x f(x) = \sum_{i=1}^{n} f_i(x), \tag{1}$$

where n represents the total number of agents in the network and the objective $f_i(\cdot)$ is only known by agent i.

Typically, solutions to optimization or decision making problems over a network are approached using a gradient descent methods in dual domain or subgradient descent methods in the case of non-differentiable cost functions [3–5]. A major advantage of first-order methods, is the fact that their structure easily enables local computation for decision making in a distributed manner, where agents do not need global information to solve their decision problems via local computation. This simplicity of the computations associated with gradient methods make them amenable to distribute and parallelize in large-scale, computation-intensive problems such as machine learning [6]. However, applying these gradient-based algorithms to large-scale problems face several challenges and become impractical due to their well-known slow convergence rates [7, 8]. To address the slow convergence rates of first order methods, second-order (Newton-type) methods have been proposed [9, 10]. Even though Newton-type methods result in quadratic convergence, they also have a significant computational overhead from the need to invert and store the Hessian of the cost function. This computational burden makes them not suitable for large-scale systems and difficult to distribute, despite the efforts at adapting them for distributed implementation [11, 12].

Quasi-Newton methods have been introduced as a way to leverage the fast convergence properties of second-order methods using the computation architecture of first-order methods. Quasi-Newton methods propose ways of incorporating the curvature information of the objective function from the second-order methods into the first-order approaches. Some examples include the BFGS [13], alongside its variations such as the low-memory BFGS [14], the Barzilai-Borwein (BB) [15] and the DFP (Davidon - Fletcher - Powell method) [16], with different assumptions made on the cost function.

A. Contributions

In this work, we present a fully distributed algorithm for solving an unconstrained optimization problem using uncoordinated BB step-sizes and obtain *Q-linear* convergence when the cost function in Problem (1) is strongly convex. Furthermore, we show that for a class of objective functions, superlinear rate of convergence can be obtained. The results in this work are related to those in [17–20]. However, the difference between our work and those in [17, 18] is that the authors based their approaches on the Adapt-Then-Combine (ATC) strategy [21] and obtained geometric convergence rate, while our BB method has Q-linear convergence attributes specifically using different bound possibilities of the BB step sizes. Also, the authors in [19] obtained a R-linear convergence rate, but the variations in the BB step sizes are not explored in the manner shown in our work. In addition, our paper explores the condition for superlinear convergence on a special class of convex quadratic function which is different from the minimization of the ratio of successive gradients method in which superlinear convergence was attained in [20].

The remainder of this section presents an overview of existing literature on convergence rates in distributed optimization and key results on convergence rates for distributed optimization using quasi-newton methods. In Section II, the problem considered and convergence results for the centralized BB are presented. Section III presents the distributed BB problem and algorithm and Section IV presents the convergence analysis of the D-BB algorithm – the main results of this paper. The obtained results are illustrated and compared with other methods via numerical simulations in Section VI.

B. Literature Review

This paper builds on earlier work on Distributed Gradient Descent (DGD) methods [22, 23], [17, 18] as well

The authors are with the Electrical & Computer Engineering Department at the University of Central Florida, Orlando FL 32816, USA. Emails: iemiola@knights.ucf.edu, cenyioha@ucf.edu

as distributed Barzilai-Borwein methods [19, 20] where the authors analyze two-dimensional convex-quadratic functions. More recent efforts in [24], which took an adapt-then-combine strategy for agreement updates, obtained a geometric rate of convergence.

In this paper, we propose a fully distributed algorithm that converges Q-linearly to the optimal solution. Furthermore, we show that a particular class of objective functions admit superlinear convergence with the modified Barzilai-Borwein algorithm presented in the paper. The approach taken in this paper is applicable to strongly convex functions and we analyze the centralized and distributed cases where computation of the step-sizes are done in an uncoordinated manner. In our approach, agents locally carry out computations, exchange information with neighboring agents to reach an agreement and use information obtained from other agents to continue the iterative process.

C. Notation

Vectors and matrices are represented by boldface lower and upper case letters, respectively. We denote the set of positive and negative reals as $\mathbb{R}_+$ and $\mathbb{R}_-$ respectively, a vector or matrix transpose as $(\cdot)^T$, and the L2-norm of a vector by $||\cdot||$. The gradient of a function $f(\cdot)$ is denoted $\nabla f(\cdot)$.

II. Problem formulation

We consider the problem of the form below over a network of agents where their objective is to

$$\min_x f(x) = \sum_{i=1}^{n} f_i(x), \tag{2}$$

In Problem (2), f is strongly convex and smooth. Each agent i in the network has access to a f_i a component of f and the agents collectively seek to optimize $f(x)$ by locally optimizing $f_i(x)$ iteratively.

The communication graph of the multi-agent network is represented by an undirected weighted Graph $G = (\mathcal{V}, \mathcal{E})$ in which $\mathcal{V} = 1, 2, ...n$ is the set of nodes (agents) and $\mathcal{E} = (i, j)$ is the set of edges such that agents i, j are connected in the edge set, where $j \neq i$. The neighbors of agent i is represented by the set $N_i = \{j : \ (i, j) \in \mathcal{E}\}$. Symmetry of the underlying graph implies that agents i and j for which $(i, j) \in \mathcal{E}$ means that information flows in both directions between both agents.

The common approach to solve Problem (2) is to use first-order methods, which involves updating the variable $x(k)$ iteratively using the gradient of the cost function with the following equation:

$$x(k+1) = x(k) - \alpha \nabla f(x(k)). \tag{3}$$

It is well known that with an appropriate choice of the step size α, the sequence $\{x(k)\}$ generated from Equation (3) converges to x^*.

A. Barzilai-Borwein Quasi-Newton Method

The Barzilai-Borwein method differs from other quasi-Newton methods because it only uses one step size for the iteration as opposed to other quasi-Newton method that need approximations for the inverse of the hessian, thus, increasing the computation overhead. Problem (2) is solved using the iterative scheme summarized in Algorithm 1, where a step-size $\alpha(k)$ is computed in the gradient descent method (3) so that $\alpha(k)\nabla f(x(k))$ approximates the $(\nabla^2 f(x(k)))^{-1}\nabla f(x(k))$ term in the Newton update $x(k+1) = x(k) - \nabla f(x)(\nabla^2 f(x))^{-1}$. Let $s(k-1) \triangleq x(k) - x(k-1)$, and $y(k-1) = \nabla f(x(k)) - \nabla f(x(k-1))$. The first BB step size is given by:

$$\alpha_1(k) = \frac{s(k-1)^T s(k-1)}{s(k-1)^T y(k-1)}. \tag{4}$$

Similarly, the second step size, $\alpha_2(k)$ is given by:

$$\alpha_2(k) = \frac{s(k-1)^T y(k-1)}{y(k-1)^T y(k-1)}. \tag{5}$$

In general, there is flexibility in the choice to use $\alpha_1(k)$ or $\alpha_2(k)$ [25]. In addition, both step sizes can be alternated within the same algorithm after a considerable amount of iterations to facilitate convergence. The procedure is summarized in Algorithm 1.

Algorithm 1 Algorithm for Centralized BB

Initialize: $\alpha_1(0), x(0), \nabla f(x(0)), \varepsilon$.
while $\|\nabla f(x(k))\| \geq \varepsilon$ **do**
  Compute
$$\alpha_1(k) = \frac{s(k-1))^T s(k-1)}{s(k-1)^T y(k-1)}$$
  ▷ α_2 in Equation (5) may also be used
  Update $x(k+1) = x(k) - \alpha_1(k)\nabla f(x(k))$
end while

Before proceeding with the distributed BB algorithm and its convergence analyses, we first present a convergence analysis of the centralized case, where the following assumptions are made about Problem (2) and Algorithm 1.

Assumption 1. *The objective function $f(x)$ in Problem (2) is strongly convex and twice differentiable. This implies that for $x, y \in \mathbb{R}^{np}$, there exists $\mu > 0$ such that:*

$$f_i(x) \geq f_i(y) + \nabla f_i(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2.$$

Assumption 2. *The gradient of the objective function ∇f is Lipschitz continuous. This implies that for all x and y, there exists $L > 0$ such that:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|.$$

B. Convergence Analysis of Centralized BB

The convergence rate of the centralized BB is just a special case of the distributed BB with the number of agents $n = 1$. Therefore, the focus is on the distributed case and we note that the strong convexity of the cost function is assumed.

III. Distributed Barzilai-Borwein Quasi-Newton Method

In this section we present a distributed solution to Problem (2), where Assumptions 1, and 2 hold. In our proposed distributed algorithm, each agent in the network keeps a local copy of the decision variable $x_i(k)$ and a local gradient $\nabla f_i(x_i(k))$ and updates them at each time-step using locally computed step sizes $\alpha_i(k)$. The step size computation is similar to the centralized case. Using the local variables $x_i(k)$ and local gradient variables $\nabla f_i(x_i(k))$, each agent computes

$$s_i(k-1) = x_i(k) - x_i(k-1), \tag{6}$$

$$y_i(k-1) = \nabla f_i(x_i(k)) - \nabla f_i(x_i(k-1)), \tag{7}$$

and computes $\alpha_i(k)$ in a manner that ensures

$$(\alpha_i(k)^{-1} I) s_i(k-1) \approx y_i(k-1). \tag{8}$$

Using the expressions in (6) and (7), we obtain the distributed form of the step size for each agent i, which is given by:

$$\alpha_i(k) = \frac{(s_i(k-1))^T s_i(k-1)}{(s_i(k-1))^T y_i(k-1)}. \tag{9}$$

To distribute the computations locally at each time step k, each agent i uses the following update scheme $x_i \in \mathbb{R}^n$:

$$x_i(k+1) = x_i(k) - \alpha_i(k) \nabla f_i(x_i(k)). \tag{10}$$

To ensure all agents converge to the optimal solution, each agent carries an iterative local computation step and the interaction with neighbors lead to a consensus step. Each agent takes a weighted average of the information received from its neighbors, to compute its next update. With this protocol, the local update at each agent i is given by:

$$x_i(k+1) = \sum_{j \in N_i \cup i} w_{ij} x_j(k) - \alpha_i(k) \nabla f_i(x_i(k)). \tag{11}$$

where w_{ij} are weights attached by agent i to agent j's estimate.
Given that $W = [w_{ij}]$, and if we let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{np}$ be the concatenation of local variables x_i, I_p is the identity matrix of dimension p, $\otimes$ represents the Kronecker operation of matrix product and $W = [w_{ij}]$ is the doubly stochastic weight matrix that satisfies: $W \otimes I_p \in \mathbb{R}^{np \times np}$. We can re-write Equation (11) as follows:

$$X(k+1) = (W \otimes I_p) X(k) - \alpha_i \nabla f(X(k)). \tag{12}$$

where X is the concatenation of local x_i, and $\nabla f(X(k)) \in \mathbb{R}^{np}$ is the concatenated gradients. Similarly we can denote the average of local estimates to be $\overline{x}(k)$, the average gradient of the agents evaluated at their local estimates to be $g(x(k))$, the average of the Lipschitz constants for agents to be L, and the average of the strong convexity parameters for agents to be μ. We also denote the average of the step size of agents to be $\overline{\alpha_i}$. Because W is doubly stochastic, it has one eigenvalue $\lambda = 1$ and the other eigenvalues satisfy $0 < \lambda < 1$.

In the distributed implementation of the BB algorithm, neighbors compute their local step sizes using local information and exchange decision estimates with their neighbors over the communication network. The process is summarized in Algorithm 2.

A. Algorithm for Distributed BB

Algorithm 2 Algorithm for Distributed BB

Initialize: $\alpha_i(0), x_i(0), \nabla f_i(x_i(0))$
while $\|g(x(k))\| \geq \varepsilon$ **do**
  Compute
    $s_i(k-1)$ using (6),
    $y_i(k-1)$ using (7),
    $\alpha_i(k+1)$ using (9).
  Local update in equation (11)
  Communicate updates $x_i(k+1)$ with neighbors.
end while

IV. Convergence Analysis of Distributed BB

We examine convergence of Algorithm 2 to the optimal point based on the local estimates.

A. Distributed BB Convergence Analysis with the First Step-Size

We present the main result of this section in Theorem 1; and later prove it via a series of Lemmas in the rest of the section.

Theorem 1. *Consider Algorithm 2 for Problem* (2) *and let Assumptions 1 and 2 hold. In addition, let the Lipschitz continuity constant for $\nabla f_i(\cdot)$ and strong convexity parameter for $f_i(\cdot)$ satisfy $\mu_i \leq L_i$ for each agent i. If $\alpha_i(k)$ in Equation* (9) *is such that $1/L_i \leq \alpha_i(k) \leq 2/(\mu_i + L_i)$, the iterates of each agent i generated from Algorithm 1 converge Q-Linearly to the optimal point x^*; that is*

$$\|x_i(k) - x^*\| \leq \|x_i(k) - \overline{x}(k)\| + \|\overline{x}(k) - x^*\|. \tag{13}$$

and we obtain a Q-Linear convergence after the agents reach consensus on the average value. Moreover, each local estimate $x_i(k)$ converges to the neighborhood of the optimal solution, x^* based on the step size α_i.

To prove the main result in Theorem 1 we will take a two-step approach. First, we upper bound the norm of the difference between the individual agent iterates and the average of the agents' iterates in Lemma 1. Next, we show that the average of the agents' iterates converges Q-linearly to the optimal solution in Lemma 2.

Lemma 1. *Consider Algorithm 2 with BB step size α_i in Equation* (9) *for Problem 2 and suppose Assumptions 1 and 2 hold; and G be the upper bound of the gradients, then the norm of the difference between each local agent's estimate and the average agents' estimate is bounded by:*

$$\|x_i(k) - \overline{x}(k)\| \leq G \left(\sum_{m=0}^{k-1} \alpha_i^2(m) \right)^{\frac{1}{2}} \left(\sum_{m=0}^{k-1} \lambda^{2(k-1-m)} \right)^{\frac{1}{2}},$$

where

$$\left(\sum_{m=0}^{k}\alpha_i^2(m)\right)^{\frac{1}{2}} \leq \frac{\sqrt{k}}{\mu},$$

and

$$\left(\sum_{m=0}^{k-1}\lambda^{2(k-1-m)}\right)^{\frac{1}{2}} \leq \left(\frac{1}{1-\lambda^2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{1-\lambda^2}} \triangleq Q_3.$$

Moreover if

$$\sum_{m=0}^{k-1}\alpha_i^2(m) \leq \frac{1}{G^2\sum_{m=0}^{k-1}\lambda^{2(k-1-m)}},$$

then each local agent's estimates converges Q-linearly to its average; that is $\|x_i(k)-\overline{x}(k)\| \leq 1$.

Proof. The proof is presented in Appendix B □

After obtaining the norm of the difference between local estimates and the consensus average estimates, we now examine the convergence attribute of the average estimates $\bar{x}(k)$ to the optimal solution x^*. Before proceeding, we state an important lemma that leads to the convergence behavior of average estimates to the optimal point.

Lemma 2. *Consider Algorithm 2 for Problem* (2) *and let Assumptions 1 and 2 hold. In addition, let the Lipschitz continuity constant for* $\nabla f(\cdot)$ *and strong convexity parameter for* $f(\cdot)$ *satisfy* $\mu \leq L$. *If the average of the first distributed BB step size* $\overline{\alpha_i}$ *is such that* $1/L \leq \overline{\alpha_i} \leq 2/(\mu+L)$, *For finite values of* i *and* k, *and bounded gradients, the consensus average estimate converges Q-Linearly to the optimal point. Moreover, the local agent estimates converge to the neighborhood of the optimal point,* x^*.

Proof. Let $\overline{x}(k) = \frac{1}{n}\sum_{i=1}^{n} x_i(k)$, $g(k) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i\left(x_i(k)\right)$, and $\overline{\alpha_i}(k) = \frac{1}{n}\sum_{i=1}^{n}\alpha_i(k)$.

From equation 10, we first consider $\|\overline{x}(k+1)-x^*\|^2$ to obtain bounds for convergence. First, we let $g(k)$ be the average of gradient at local estimates and we let $\overline{\alpha_i}$ be the average of the agents step sizes corresponding to the average of the iterates.

$$\|\overline{x}(k+1)-x^*\| = \|\overline{x}(k)-x^*-\overline{\alpha_i}g(k)\|.$$

By squaring both sides and evaluating the right hand side, we have:

$$\begin{aligned}\|x(k)-x^*-\overline{\alpha_i}(k)g(k)\|^2 = &\|x(k)-x^*\|^2+\overline{\alpha_i}^2(k)\|g(k)\|^2 \\ &- 2\left(x(k)-x^*\right)^T\left(\overline{\alpha_i}(k)g(k)\right).\end{aligned} \tag{14}$$

Using the fact that for all vectors $a,\ b,\ 2a^Tb \leq \|a\|^2+\|b\|^2$, we obtain the relationship:

$$2\left(\overline{x}(k)-x^*\right)^T\left(g(k)\right) \leq \|g(k)\|^2+\|\overline{x}(k)-x^*\|^2.$$

Just as we did for the centralized case, μ and L are strong convexity and Lipschitz parameters respectively and c_1, c_2 are given by $c_1 = 2/(\mu+L)$ and $c_2 = 2\mu L/(\mu+L)$. We now obtain:

$$\begin{aligned}\|\overline{x}(k+1)-x^*-\overline{\alpha_i}(k)g(k)\|^2 \leq &\|\overline{x}(k)-x^*\|^2+\overline{\alpha_i}^2(k)\|g(k)\|^2 \\ &- \overline{\alpha_i}(k)c_1\|g(k)\|^2 \\ &- \overline{\alpha_i}(k)c_2\|\overline{x}(k)-x^*\|^2, \\ \leq &(1-\overline{\alpha_i}(k)c_2)\|\overline{x}(k)-x^*\|^2 \\ &+ (\overline{\alpha_i}^2(k)-\overline{\alpha_i}(k)c_1)\|g(k)\|^2, \\ \leq &(1-\overline{\alpha_i}(k)c_2)\|\overline{x}(k)-x^*\|^2.\end{aligned} \tag{15}$$

We note that the last inequality is due to Theorem 2.1.12 from Chapter 2 of [26]. We also note that in the previous inequality, the term that contains the distributed form of the step size, $(\overline{\alpha_i}^2(k)-\overline{\alpha_i}(k)c_1)\|g(k)\|^2 \leq 0$ provided $\overline{\alpha_i}(k) \leq c_1$. We show that the step size $\overline{\alpha_i}(k) = c_1$ is within the range of the BB step size bounds below:

Corollary 1. *Let* L *and* μ *be the Lipschitz and strong convexity parameters respectively with* $\mu \leq L$. *The range of the average of the distributed BB step size* $\overline{\alpha_i}(k)$ *is given by:*

$$\frac{1}{L} \leq \overline{\alpha_i}(k) \leq \frac{2}{\mu+L} \leq \frac{1}{\mu}. \tag{16}$$

where the condition $\overline{\alpha_i}(k) \leq \frac{1}{\mu}$ is assumed. See Appendix A for details.

Therefore the distributed BB convergence using the first BB step size can be analysed as:

$$\begin{aligned}\|\overline{x}(k+1)-x^*\|^2 &\leq (1-\overline{\alpha_i}(k)c_2)\|\overline{x}(k)-x^*\|^2, \\ \frac{\|\overline{x}(k+1)-x^*\|^2}{\|\overline{x}(k)-x^*\|^2} &\leq 1-\overline{\alpha_i}(k)c_2, \\ \text{and}\quad \frac{\|\overline{x}(k+1)-x^*\|}{\|\overline{x}(k)-x^*\|} &\leq (1-\overline{\alpha_i}(k)c_2)^{\frac{1}{2}}.\end{aligned}$$

We will now bound: $(1-\overline{\alpha_i}(k)c_2)^{\frac{1}{2}}$. The distributed form of the first Barzilai-Borwein step size $\alpha_i(k)$ is given by:

$$\alpha_i(k) = \frac{\|s_i(k-1)\|^2}{\left[x_i(k)-x_i(k-1)\right]^T\left[\nabla f_i\left(x_i(k)\right)-\nabla f_i\left(x_i(k-1)\right)\right]}$$

By using Lipschitz continuity of $\nabla f(\cdot)$ with L as the Lipschitz constant, we obtain the lower bound of distributed form of the first BB step size as:

$$\alpha_i(k) > \frac{\|x_i(k)-x_i(k-1)\|^2}{L\|x_i(k)-x_i(k-1)\|^2} = \frac{1}{L}.$$

We know that $\overline{\alpha_i}(k) = \frac{1}{n}\sum_{i=1}^{n}\alpha_i(k)$, and it follows that $n\overline{\alpha_i}(k) = \sum_{i=1}^{n}\alpha_i(k)$. But we know that $\alpha_i(k) > \frac{1}{L}$, and as a fact, $\alpha_i(k) < \sum_{i=1}^{n}\alpha_i(k)$. Therefore we obtain the relationship:

$$\frac{1}{L} < \alpha_i(k) < \sum_{i=1}^{n}\alpha_i(k). \tag{17}$$

From equation (17), $n\overline{\alpha_i}(k) = \sum_{i=1}^{n}\alpha_i(k) > \frac{1}{L}$ and we obtain the fact that $\overline{\alpha_i}(k) > \frac{1}{nL}$.

If $\overline{\alpha_i}(k)$ and c_2 are positive and $\overline{\alpha_i}(k) > 1/nL$, then $-\overline{\alpha_i}(k)c_2 < -c_2/nL$. Since $\overline{\alpha_i}(k) > 1/nL$, it implies that $0 < 1 - \overline{\alpha_i}(k)c_2 < 1 - c_2/nL$, Therefore,

$$0 < (1 - \overline{\alpha_i}(k)c_2)^{\frac{1}{2}} < \left(1 - \frac{c_2}{nL}\right)^{\frac{1}{2}}.$$

We will now show that $c_2/nL < 1$. If $c_2 = 2\mu L/(\mu + L)$, then it implies that $c_2/nL = 2\mu/n(\mu+L)$. If $\mu \leq L$, then we have $\mu + \mu \leq L + \mu$ and we obtain that $2\mu/n(\mu + L) \leq 1$ for all positive values of n. Therefore,

$$\lim_{k\to\infty} \quad \frac{\|\overline{x}(k+1) - x^*\|}{\|\overline{x}(k) - x^*\|} \leq \left(1 - \frac{c_2}{nL}\right)^{\frac{1}{2}} \leq 1,$$

So the average of the estimates converges Q-linearly to the optimal point, x^*. □

V. On Superlinear Convergence of Algorithm 1

In the preceding sections, we obtained Q-linear convergence for Algorithms 1 and 2, when the cost function being minimized is strongly convex with Lipschitz-continuous gradients. A consequence of Lemmas 2 is that under certain conditions, Algorithm 2 converges superlinearly to the optimal solution x^*.

Corollary 2. *Consider Problem 2. If the objective function $f(\cdot)$ being minimized is such that its strong convexity parameter is equal to the Lipschiz gradient, then the iterates generated by Algorithm 1 converges superlinearly to the optimal solution.*

Example 1. *An example of a function that meets the conditions required for super linear convergence in Corollary 2 is the quadratic function:*

$$f(x) = 0.5x^T x. \tag{18}$$

in a distributed manner.

Superlinear convergence can be obtained for strongly convex quadratic functions and two-dimensional strictly convex quadratic functions [20]. However, we illustrate an example of a strongly convex quadratic function that yield a superlinear convergence behavior. The strong convexity parameter of the cost function in (18) is equal to the Lipschitz parameter of its gradient; that is, $L = \mu = 1$. By using this function on the upper bound obtained for rates of convergence as seen in Lemma 2, we have:

$$\lim_{k\to\infty} \quad \frac{\|x(k+1) - x^*\|}{\|x(k) - x^*\|} < \left(1 - \frac{c_2}{L}\right)^{\frac{1}{2}},$$

where $c_2 = 2\mu L/(\mu + L)$. To obtain superlinear convergence rate, if $L = \mu$, then we obtain:

$$\left(1 - \frac{c_2}{L}\right) = \left(1 - \frac{2\mu}{2\mu}\right) = 0.$$

and we obtain:

$$\lim_{k\to\infty} \quad \frac{\|x(k+1) - x^*\|}{\|x(k) - x^*\|} < 0,$$

and we obtain a superlinear convergence rate.

VI. Numerical Experiments

We show some simulations for results in Lemma II-B (the centralized case) and the results in Theorem 1 (the distributed case). For the centralized, we consider the least square cost function:

$$f(x) = \|Ax - b\|_2^2, \tag{19}$$

where the matrix A is positive-definite so that $f(x)$ is strongly convex with parameter μ and $\mu \leq L$, where L is the Lipschitz parameter of the gradient of $f(x)$. We run the simulations for 50 iterations and compare the BB step size in (4) with the gradient method using a decaying step size of $\frac{1}{k}$. In Figure 1, the label "gradient" is the curve obtained when a step size of $\frac{1}{k}$ is used using a gradient descent algorithm while the label "BB" is the the curve obtained when the actual BB step size in equation (4) is used. The results of the simulation of Algorithm 1 are summarized in Figure 1 where the errors as a function of time are illustrated. As can be observed, the error converges to zero indicating the iterates $x(k)$ are approaching the optimal solution x^*.

For the distributed case, we consider the following objective function, which is separable per agent:

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} x^T A_i x + b_i^T x, \tag{20}$$

where $A_i \in \mathbb{R}^{p\times p}$, $b_i \in \mathbb{R}^p$ are used by each agent i for their own computation, and n is the number of agents in the network and its dimension is $m = 10$. In equation (20), the gradient function is given by:

$$\nabla f(x) = \frac{1}{2}(A_i + A_i^T)x + b_i. \tag{21}$$

Suppose there are 100 nodes in the network and the matrix W is a positive, symmetric, random doubly stochastic matrix. Our simulations aim to compare different step sizes with the distributed Barzilai-Bowein in equations (9). Specifically, we use the following step sizes of $\alpha_i = \frac{1}{L_i}$, $\alpha_i = \frac{1}{\mu_i}$, and $\alpha_i = \frac{2}{L_i+\mu_i}$ according to Lemma 2 and the BB step size as seen in equation (9). In Figure 2, the label $c_1 = \frac{2}{\mu+L}$ bound curve (the circular curve) is the step size according to convergence result in Lemma 2. The label BB step size in figure 2 (the curve beneath all other curves) is the actual BB step size in equation (9) and the labels BB-Upperbound step size curve (the one in asterisk) and BB-Lowerbound step size curve (triangular) are the lower and upper bounds of the BB step sizes ($\alpha_i = \frac{1}{L_i}$, $\alpha_i = \frac{1}{\mu_i}$) respectively. Our simulations affirm that the step size $\alpha_i = \frac{2}{L_i+\mu_i}$ lies in between the lower and upper bounds of the BB step size and also agrees with the theoretical result in Corollary 1. We apply these step sizes to the iteration shown in equation (11) to compare the rates at which each step size converges to the optimal point. Convergence is attained for the three step sizes used and we run the simulations for 20 iterations to compare convergence speeds. We compare convergence rates by first initializing x, A and b as zeros between time step $k = 1$ to the total number of iterations $T = 20$. We also initialize the values of L, μ and the three forms of the step sizes as zeros. We plot the

error curve for the three step sizes indicated by the expression $\|\overline{x}(k+1)-x^*\|$ and compare the results. Our Numerical results are shown graphically below in Figures 1 and 2:

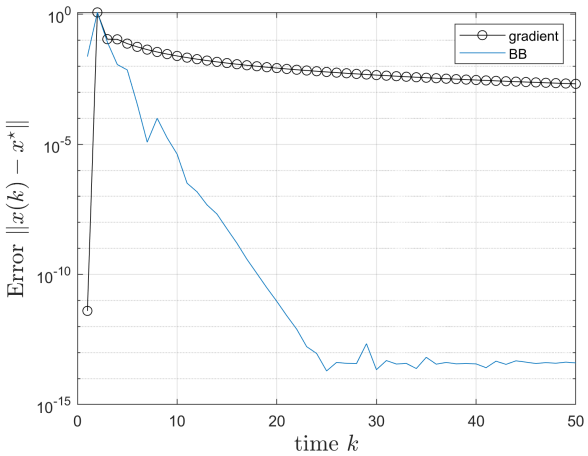

Fig. 1: Centralized Simulations for 50 iterations

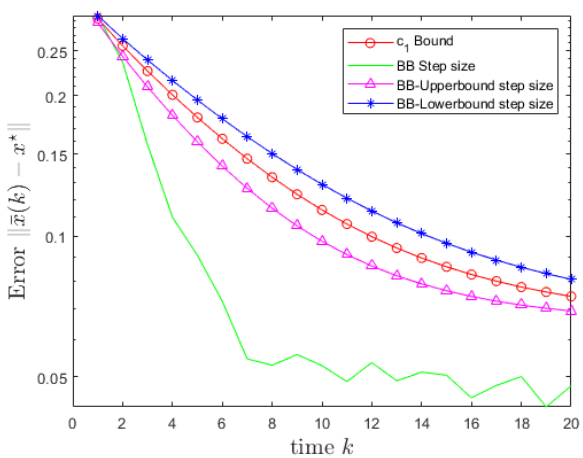

Fig. 2: Distributed Simulations for 20 iterations

VII. Conclusions

We examined the convergence attributes of an unconstrained problem and we obtained Q-linear convergence for both the centralized and distributed case by assuming strong convexity on the objective function. However, because this is a consensus-type scenario, the Q-linear convergence rate holds when the agents in a network locally agree to an average, $\overline{x}(k)$ for both the centralized and distributed cases. We also compared the speed of convergence for the gradient method with the BB method by using the two Barzilai-Borwein step sizes, and the BB method converges to the optimal point at a faster than the gradient method. In section VI, the faster convergence of the BB method is evident because of its minimal error in getting to the optimal point than in the gradient method after some iterations.

References

[1] Y. Zeng and R. Zhang, "Energy-efficient uav communication with trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3747–3760, 2017.

[2] Y. Arjevani and O. Shamir, "Communication complexity of distributed convex learning and optimization," in *Advances in neural information processing systems*, 2015, pp. 1756–1764.

[3] S. Magnússon, C. Enyioha, N. Li, and C. Fischione, "Practical coding schemes for bandwidth limited one-way communication resource allocation," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 221–226.

[4] J. Lei, H.-F. Chen, and H.-T. Fang, "Primal–dual algorithm for distributed constrained optimization," *Systems & Control Letters*, vol. 96, pp. 110–117, 2016.

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

[6] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 2737–2745.

[7] J.-L. Goffin, "On convergence rates of subgradient optimization methods," *Mathematical programming*, vol. 13, no. 1, pp. 329–347, 1977.

[8] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in neural information processing systems*, 2011, pp. 1458–1466.

[9] N. N. Schraudolph, J. Yu, and S. Günter, "A stochastic quasi-newton method for online convex optimization," in *Artificial intelligence and statistics*, 2007, pp. 436–443.

[10] A. F. Izmailov and M. V. Solodov, "Newton-type methods for optimization problems without constraint qualifications," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 210–228, 2004.

[11] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[12] N. Zheng and J.-F. Yin, "Convergence of accelerated modulus-based matrix splitting iteration methods for linear complementarity problem with an h+-matrix," *Journal of Computational and Applied Mathematics*, vol. 260, pp. 281–293, 2014.

[13] M. Eisen, A. Mokhtari, and A. Ribeiro, "Decentralized quasi-newton methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2613–2628, 2017.

[14] A. Mokhtari and A. Ribeiro, "Global convergence of online limited memory bfgs," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3151–3181, 2015.

[15] Y.-H. Dai and R. Fletcher, "Projected barzilai-borwein methods for large-scale box-constrained quadratic programming," *Numerische Mathematik*, vol. 100, no. 1, pp. 21–47, 2005.

[16] D. Pu, "Convergence of the dfp algorithm without exact line search," *Journal of optimization theory and applications*, vol. 112, no. 1, pp. 187–211, 2002.

[17] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[18] A. Nedić, A. Olshevsky, W. Shi, and C. A. Uribe, "Geometrically convergent distributed optimization with uncoordinated step-sizes," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 3950–3955.

[19] Y.-H. Dai and L.-Z. Liao, "R-linear convergence of the barzilai and borwein gradient method," *IMA Journal of Numerical Analysis*, vol. 22, no. 1, pp. 1–10, 2002.

[20] Y.-H. Dai, "A new analysis on the barzilai-borwein gradient method," *Journal of the operations Research Society of China*, vol. 1, no. 2, pp. 187–198, 2013.

[21] A. H. Sayed, "Diffusion adaptation over networks," in *Academic*

Press Library in Signal Processing. Elsevier, 2014, vol. 3, pp. 323–453.
[22] A. S. Berahas, R. Bollapragada, N. S. Keskar, and E. Wei, "Balancing communication and computation in distributed optimization," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3141–3155, 2018.
[23] D. Jakovetić, J. M. Moura, and J. Xavier, "Linear convergence rate of a class of distributed augmented lagrangian algorithms," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 922–936, 2014.
[24] J. Gao, X. Liu, Y.-H. Dai, Y. Huang, and P. Yang, "Geometric convergence for distributed optimization with barzilai-borwein step sizes," *arXiv preprint arXiv:1907.07852*, 2019.
[25] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA journal of numerical analysis*, vol. 8, no. 1, pp. 141–148, 1988.
[26] Y. Nesterov, "Introductory lectures on convex programming volume i: Basic course," *Lecture notes*, vol. 3, no. 4, p. 5, 1998.

APPENDIX

A. Proof of Corollary 1

Proof. We establish equation (16) in a manner that the range of step size bounds below holds:

$$\frac{1}{L} \leq \alpha_i(k) \leq \frac{2}{\mu + L} \leq \frac{1}{\mu}.$$

We also note that the step size range also applied to both the centralized and distributed form of the step sizes. First, according to [24], we first start by noting that the BB step size can be upper and lower bounded according to:

$$\frac{1}{L} \leq \alpha_i(k) \leq \frac{1}{\mu}.$$

To include $2/(\mu + L)$ between the first distributed step size, $\alpha_i(k)$ and $1/\mu$, we prove that $2/(\mu + L) \leq 1/\mu$ and $2/(\mu + L) \geq 1/L$.

To prove that $2/(\mu+L) \leq 1/\mu$, we show that $\frac{1}{\mu} - \frac{2}{\mu+L} > 0$.

$$\frac{1}{\mu} - \frac{2}{\mu + L} = \frac{L - \mu}{\mu(\mu + L)}.$$

We know that $L \geq \mu$ and L and μ are positive, then we obtain that $1/\mu - 2/(\mu + L) > 0$.

Now we will prove that $2/(\mu + L) \geq 1/L$. below:

$$\frac{2}{\mu + L} - \frac{1}{L} = \frac{L - \mu}{L(L + \mu)}.$$

Since $L \geq \mu$ and L and μ are both positive, we have that $2/(\mu + L) \geq 1/L$. Therefore we obtain the bounds:

$$\frac{1}{L} \leq \alpha_i(k) \leq \frac{2}{\mu + L} \leq \frac{1}{\mu}.$$

So we conclude that the step size $\alpha_i(k) = 2/(\mu + L)$ lies between the lower $(1/L)$ and lower bounds $(1/\mu)$ of the BB step sizes. The fact that $\overline{\alpha_i}(k) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i(k)$ and $\overline{\alpha_i}(k) \leq \frac{1}{\mu}$ hold completes the proof. □

B. Proof of Lemma 1

Proof. We first consider the first step size α_i as expressed in equation (9), and for notation simplicity, we denote $Z = W \otimes I_p \in \mathbb{R}^{np \times np}$. Then the distributed iteration at time step k is given by:

$$X(k + 1) = ZX(k) - \alpha_i(k)\nabla f(X(k)). \tag{22}$$

where $\otimes$ is the kronecker product. From the definitions of $\overline{x}(k)$ as the average of local estimates, we have the expression: $\overline{x}(k) = \frac{1}{n}\sum_{i=1}^{n} x_i(k)$. Likewise, from the definition of $g(x(k))$ as the average of local gradients, we obtain the relationship: $g(x(k)) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i\left(x_i(k)\right)$. If we solve for $X(k)$ in equation (22) where α_i is the BB step size, we obtain the expression:

$$X(k) = -\sum_{m=0}^{k-1} \alpha_1(m)\left(W^{(k-1-m)} \otimes I\right)\nabla f\left(x(m)\right). \tag{23}$$

Suppose $\overline{X}(k)$ is the average of all concatenated $x_i(k)$, then we obtain

$$\overline{X}(k) = \frac{1}{n}\left(\left(1_n 1_n^T\right) \otimes I\right) x(k). \tag{24}$$

Also, we know that the following relationship:

$$\|x_i(k) - \overline{x}(k)\| \leq \|X(k) - \overline{X}(k)\|. \tag{25}$$

where the distributed form of the first BB step size is given in α_i as expressed in equation (9). From equations (23) and (24), we obtain:

$$\begin{aligned}
\|x_i(k) - \overline{x}(k)\| \leq & \|X(k) - \overline{X}(k)\|, \\
= & \|X(k) - \frac{1}{n}\left(\left(1_n 1_n^T\right) \otimes I\right) X(k)\|, \\
= & \|-\sum_{m=0}^{k-1} \alpha_i(m)(W^{(k-1-m)} \otimes I)\nabla f(x(m)) \\
& + \frac{1}{n}((1_n 1_n^T) \otimes I) \sum_{m=0}^{k-1} (\alpha_i(m) \\
& (W^{(k-1-m)} \otimes I)\nabla f(x(m)))\|, \\
= & \|-\sum_{m=0}^{k-1} (\alpha_i(m)(W^{(k-1-m)} \otimes I)\nabla f(x(m))) \\
& + \sum_{m=0}^{k-1} \frac{\alpha_i(m)}{n}((1_n 1_n^T) \otimes I)\nabla f(x(m))\|.
\end{aligned} \tag{26}$$

Because W is doubly stochastic, then we have the relationship:

$$\begin{aligned}
\|x_i(k) - \overline{x}(k)\| \leq & -\sum_{m=0}^{k-1} \|\alpha_i(m)((W^{(k-1-m)} - \frac{1}{n} 1_n 1_n^T) \\
& \otimes I)\nabla f(x(m)).\|
\end{aligned}$$

If $\alpha_i(k) = \frac{s_i(k-1)^T s_i(k-1)}{s_i(k-1)^T y_i(k-1)}$, then $\alpha_i(m) = \frac{s_i(m-1)^T s_i(m-1)}{s_i(m-1)^T y_i(m-1)}$. Therefore we have the expression:

$$\begin{aligned}
\|x_i(k) - \overline{x}(k)\| \quad \leq \quad & -\sum_{m=0}^{k-1} \frac{\|s_i(m-1)\|^2}{s_i(m-1)^T y_i(m-1)}. \\
& \|W^{(k-1-m)} - \frac{1}{n} 1_n 1_n^T\| \|\nabla f(x(m))\|.
\end{aligned}$$

So we obtain that the expression:
$\|x_i(k) - \overline{x}(k)\|$.

$$\leq \sum_{m=0}^{k-1} \frac{\|s_i(m-1)\|^2}{s_i(m-1)^T y_i(m-1)} \lambda^{(k-1-m)} \|\nabla f(x(m)).\|$$

If $\nabla f(x(m))$ is bounded meaning that $\|\nabla f(x(m)\| \leq G$ where G is positive, then we have:

$$\|x_i(k) - \overline{x}(k)\| \leq \sum_{m=0}^{k-1} \frac{G\|s_i(m-1)\|^2}{s_i(m-1)^T y_i(m-1)} \lambda^{(k-1-m)}. \quad (27)$$

The eigenvalues λ of the weight matrix W satisfies the bounds, $0 < \lambda \leq 1$. From equation (27) and by Cauchy-Schwarz on sums, we obtain:

$$\|x_i(k) - \overline{x}(k)\| \leq G \left(\sum_{m=0}^{k-1} \alpha_i^2(m) \right)^{\frac{1}{2}} \left(\sum_{m=0}^{k-1} \lambda^{2(k-1-m)} \right)^{\frac{1}{2}}. \quad (28)$$

We know that the BB step size is upper bounded such that $\alpha_i < 1/\mu$.
Then by squaring both sides, $\alpha_i^2 \leq \frac{1}{\mu^2}$ and we obtain the relationship $\sum_{m=0}^{k} \alpha_i^2(m) \leq \frac{k}{\mu^2}$ and we obtain the result $\left(\sum_{m=0}^{k} \alpha_i^2(m) \right)^{\frac{1}{2}} \leq \frac{\sqrt{k}}{\mu}$.
In equation (28),

$$\|x_i(k) - \overline{x}(k)\| \leq G \left(\sum_{m=0}^{k-1} \alpha_i^2(m) \right)^{\frac{1}{2}} \left(\sum_{m=0}^{k-1} \lambda^{2(k-1-m)} \right)^{\frac{1}{2}}, \quad (29)$$

where

$$\left(\sum_{m=0}^{k} \alpha_i^2(m) \right)^{\frac{1}{2}} \leq \frac{\sqrt{k}}{\mu},$$

and

$$\left(\sum_{m=0}^{k-1} \lambda^{2(k-1-m)} \right)^{\frac{1}{2}} \leq \left(\frac{1}{1-\lambda^2} \right)^{\frac{1}{2}} = \frac{1}{\sqrt{1-\lambda^2}} \triangleq \mathrm{Q}_3.$$

Moreover if

$$\sum_{m=0}^{k-1} \alpha_i^2(m) \leq \frac{1}{G^2 \sum_{m=0}^{k-1} \lambda^{2(k-1-m)}},$$

then each local agent's estimates converges Q-linearly to its average; that is $\|x_i(k) - \overline{x}(k)\| \leq 1$.

□