SCIENCE CHINA
Information Sciences

• LETTER •

April 2022, Vol. 65 149204:1–149204:2
https://doi.org/10.1007/s11432-020-3256-x

Achieving geometric convergence for distributed optimization with Barzilai-Borwein step sizes

Juan GAO[1], Xin-Wei LIU[2*], Yu-Hong DAI[3,4], Yakui HUANG[2] & Peng YANG[1]

[1]*School of Artificial Intelligence, Hebei University of Technology, Tianjin* 300401*, China;*
[2]*Institute of Mathematics, Hebei University of Technology, Tianjin* 300401*, China;*
[3]*LSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, China;*
[4]*School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received 25 November 2020/Revised 8 February 2021/Accepted 1 April 2021/Published online 9 December 2021

Citation Gao J, Liu X-W, Dai Y-H, et al. Achieving geometric convergence for distributed optimization with Barzilai-Borwein step sizes. Sci China Inf Sci, 2022, 65(4): 149204, https://doi.org/10.1007/s11432-020-3256-x

Dear editor,

We consider distributed optimization, which can be formulated to minimize the average of all local objective functions:

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where each local function $f_i : \mathbb{R}^p \to \mathbb{R}$ is L_i-smooth and μ_i-strongly convex and known only by agent i. All agents communicate over a time-invariant undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents, $\mathcal{E}$ is the collection of pairs $(i, j)(i, j \in \mathcal{V})$ such that agents i and j can exchange information with each other. Problem (1) has been extensively studied in machine learning and smart grids, e.g., [1, 2].

Distributed gradient methods have been a great success in solving the problem (1), primarily because of their low computational cost and easy implementation. To ensure convergence to the exact solution, distributed gradient descent (DGD) [3] should use a diminishing step size, which may result in a slow convergence rate. With a constant step size, DGD can be fast; however, it only converges to a neighborhood of the exact solution. Recently, distributed gradient methods have achieved significant improvements, which provide us new variants that geometrically converge to the exact solution for smooth and strongly convex functions [4–7]. Ref. [4] proposed an improved DGD, called EXTRA, which exploits the difference of two consecutive DGD iterates with different weight matrices to cancel the steady-state error. A variant of DGD [5], named NEAR-DGD$^+$, employs a multi-consensus inner loop strategy to DGD. NEAR-DGD$^+$ requires that the number of consensus steps is increasing at an appropriate rate, which leads to additional communications. Moreover, another variant of DGD [6, 7] is based on the dynamic average consensus approach, which replaces local gradients in DGD with tracking gradients. Using an adapt-then-combine strategy, these methods in [6,7] are capable of using uncoordinated constant step sizes. Note that the aforementioned methods neither address how to design an appropriate step size for each agent using its local information nor do they consider automatically computing the step sizes.

The centralized Barzilai-Borwein (BB) method is a simple and effective technique for selecting the step size and requires fewer storages and inexpensive computations [8, 9]. Moreover, the BB step size is automatically computed using gradient information. Recent years have witnessed the successful applications of the centralized BB method in image processing and machine learning [9].

Inspired by wonderful features and successful applications of the centralized BB method, we propose a distributed gradient method with Barzilai-Borwein step sizes (DGM-BB-C), which combines an adapt-then-combine variation of the dynamic average consensus approach [6, 7] with multi-consensus inner loops [5]. The primary contributions of our work are highlighted below. (i) Each agent can automatically compute the step size using its local gradient information. The step sizes of DGM-BB-C are not less than $\frac{1}{L_i}$, which provides a selection for a larger step size than previous studies. For example, Refs. [5–7] required that the step sizes are not greater than $\frac{1}{L}$ with $L = \max\{L_i\}$. (ii) By simultaneously using the dynamic average consensus approach and multi-consensus inner loops, DGM-BB-C can seek the exact optimum when the number of consensus steps stays constant, which results in fewer communications than NEAR-DGD$^+$ [5]. (iii) In contrast with existing methods [5–7], DGM-BB-C uses the Barzilai-Borwein step sizes and finite consensus steps, which leads to faster convergence. Under suitable conditions, we confirm that DGM-BB-C has geometric convergence to the optimal solution. We numerically show the superiority of DGM-BB-C compared with certain advanced methods and validate our theoretical discoveries.

Distributed Barzilai-Borwein step sizes. We now apply the centralized BB method [8] (described in Appendix A) to the distributed optimization. Note that the step size cannot be straightly computed using the centralized BB method because distributed optimization methods never compute the

[*] Corresponding author (email: mathlxw@hebut.edu.cn)

average gradient $\nabla f(x_k)$. Therefore, we should implement the centralized BB method in a distributed manner. Let x_k^i be the agent i's variable at iteration k and $\nabla f_i(x_k^i)$ be the gradient of f_i at x_k^i. The distributed BB step sizes are then described as follows:

$$\alpha_k^i = \min\left\{(\alpha_k^i)^{\rm BB}, t_k^i\right\}, \qquad (2)$$

where t_k^i is a local safeguarding parameter adopted by agent i, $(\alpha_k^i)^{\rm BB} = (\alpha_k^i)^{\rm BB1}$ or $(\alpha_k^i)^{\rm BB2}$, depending on the following formulae:

$$(\alpha_k^i)^{\rm BB1} = \frac{(s_k^i)^{\rm T} s_k^i}{(s_k^i)^{\rm T} z_k^i} \ \text{ or } \ (\alpha_k^i)^{\rm BB2} = \frac{(s_k^i)^{\rm T} z_k^i}{(z_k^i)^{\rm T} z_k^i}, \qquad (3)$$

where $s_k^i = x_k^i - x_{k-1}^i$ and $z_k^i = \nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)$ for $k \geqslant 1$.

DGM-BB-C algorithm. Now, we introduce the distributed BB step sizes into distributed optimization and propose a DGM-BB-C method in Algorithm 1. In our algorithm, each agent performs two steps: one is the local optimization step, and the other is the dynamic average consensus step. We use the adapt-then-combine strategy for the two steps because the BB step sizes are a type of uncoordinated constant step sizes. Thus, the adapt-then-combine scheme requires two rounds of communication; however, the other methods, such as DGD and EXTRA, require only one. Nevertheless, because the deviation of the two gradient estimates makes the algorithm not work well, we conduct multi-consensus inner loops to ensure estimated gradients $\nabla f_i(x_k^i)$ and y_k^i are as close to the average gradient $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_k^i)$ as possible. In particular, we use a multi-consensus inner loop strategy for the local optimization step and the dynamic average consensus step, respectively. Let R be a positive integer, which is the number of inner consensus iterations, and $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ be a weight matrix. Our algorithm DGM-BB-C is then described as Algorithm 1.

Algorithm 1 DGM-BB-C for undirected connected graphs

```
1: Initialization: for $i \in \mathcal{V}$, $x_0^i \in \mathbb{R}^p$, $y_0^i = \nabla f_i(x_0^i)$, $\alpha_0^i > 0$.
2: Local optimization: for $i \in \mathcal{V}$, compute
       $x_{k+1}^i(0) = x_k^i - \alpha_k^i y_k^i,$
       $x_{k+1}^i(r) = \sum_{j=1}^n w_{ij} x_{k+1}^j(r-1), \quad r = 1, 2, \ldots, R,$
   where $\alpha_k^i$ is computed by (2), and set $x_{k+1}^i = x_{k+1}^i(R)$.
3: Dynamic average consensus: for $i \in \mathcal{V}$, compute
       $y_{k+1}^i(0) = y_k^i + \nabla f_i(x_{k+1}^i) - \nabla f_i(x_k^i),$
       $y_{k+1}^i(r) = \sum_{j=1}^n w_{ij} y_{k+1}^j(r-1), \quad r = 1, 2, \ldots, R,$
   and set $y_{k+1}^i = y_{k+1}^i(R)$.
4: Set $k \to k+1$ and go to Step 2.
```

When $R = 1$ and $\alpha_k^i = \alpha^i$ (α^i is a constant with different values for different agent i), DGM-BB-C reduces to Aug-DGM [6]/ATC-DIGing [7].

Theorem 1. Under suitable assumptions, the sequence generated by DGM-BB-C exactly converges to the unique optimal solution at a geometric rate.

With the help of lemmas in Appendix B, the proof of Theorem 1 is derived in Appendix C.

Numerical experiments. We observe the behavior of distributed BB step sizes. The effects of initial step sizes and different values of R on DGM-BB-C are studied. We then compare the performance of DGM-BB-C with certain advanced methods, which shows that DGM-BB-C achieves better performance. Appendix D shows the experimental results and performance analysis.

Conclusion. By combining the adapt-then-combine variation of the dynamic average consensus approach and multi-consensus inner loops, we propose a DGM-BB-C method. Unlike the existing distributed gradient methods, our method automatically computes the step sizes for each agent. For smooth and strongly convex objective functions, we confirmed that DGM-BB-C geometrically converges to the optimal solution under appropriate assumptions. Numerical experiments demonstrated that DGM-BB-C achieves both the optimal computation and communication cost for distributed optimization. It has been numerically shown that the step sizes of DGM-BB-C are larger than those in previous studies. DGM-BB-C can then seek the exact solution both theoretically and empirically when the number of consensus steps stays constant. The connectivity of networks and the number of inner consensus iterations are inversely proportional to each other. Certain possible topics in the future are to extend our results to directed graphs and time-varying networks, respectively.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 11671116, 12071108, 11701137, 91630202) and Strategic Priority Research Program of CAS (Grant No. XDA27000000).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wang Y H, Lin P, Hong Y G. Distributed regression estimation with incomplete data in multi-agent networks. Sci China Inf Sci, 2018, 61: 092202

2 Yu W W, Li C J, Yu X H, et al. Economic power dispatch in smart grids: a framework for distributed optimization and consensus dynamics. Sci China Inf Sci, 2018, 61: 012204

3 Nedić A, Ozdaglar A. Distributed subgradient methods for multi-agent optimization. IEEE Trans Automat Contr, 2009, 54: 48–61

4 Shi W, Ling Q, Wu G, et al. EXTRA: an exact first-order algorithm for decentralized consensus optimization. SIAM J Optim, 2015, 25: 944–966

5 Berahas A S, Bollapragada R, Keskar N S, et al. Balancing communication and computation in distributed optimization. IEEE Trans Automat Contr, 2019, 64: 3141–3155

6 Xu J, Zhu S, Soh Y C, et al. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In: Proceedings of IEEE Conference on Decision and Control, Osaka, 2015. 2055–2060

7 Nedić A, Olshevsky A, Shi W, et al. Geometrically convergent distributed optimization with uncoordinated stepsizes. In: Proceedings of American Control Conference, Seattle, 2017. 3950–3955

8 Huang N, Dai Y H, Burdakov O. Stabilized Barzilai-Borwein method. J Comput Math, 2019, 37: 916–936

9 Tan C H, Ma S Q, Dai Y H, et al. Barzilai-Borwein step size for stochastic gradient descent. In: Proceedings of Annual Conference on Neural Information Processing Systems, Barcelona, 2016. 685–693