SCIENCE CHINA
Information Sciences

• **Supplementary File** •

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*

Appendix A Centralized BB Method

The iterative format of the centralized BB method [8] for solving (1) takes the following form:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{A1}$$

where $\alpha_k = \min\left\{\alpha_k^{BB}, t_k\right\}$, $t_k > 0$ is a safeguarding parameter and $\alpha_k^{BB} = \alpha_k^{BB1}$ or $\alpha_k^{BB} = \alpha_k^{BB2}$, depending on the following formulae

$$\alpha_k^{BB1} = \frac{s_k^T s_k}{s_k^T z_k} \tag{A2}$$

or

$$\alpha_k^{BB2} = \frac{s_k^T z_k}{z_k^T z_k}, \tag{A3}$$

where $s_k = x_k - x_{k-1}$ and $z_k = \nabla f(x_k) - \nabla f(x_{k-1})$ for $k \geqslant 1$. Note that t_k can be a constant or a variable. For $t_k = +\infty$, it reduces to the original BB method proposed by Barzilai and Borwein. When t_k depends on the gradient values at x_k, [8] proves the global convergence of the centralized BB method for strongly convex functions with Lipschitz gradients.

Appendix B Several Important Lemmas

In this section, three lemmas are given, which are helpful for convergence analysis of DGM-BB-C. Firstly, we make the following three standard assumptions.

Assumption 1 (Smoothness)**.** Every function f_i is differentiable and has Lipschitz continuous gradient, i.e., for any $x, y \in \mathbb{R}^p$, there exists a positive constant L_i such that

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leqslant L_i \|x - y\|.$$

Assumption 2 (Strong convexity)**.** Every function f_i is strongly convex, i.e., for any $x, y \in \mathbb{R}^p$, there exists a positive constant μ_i such that

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

Constants L_i and μ_i satisfy $\mu_i \leqslant L_i$ for each agent i. It is obtained immediately from Assumption 2 that problem (1) has a unique optimal solution denoted by $x^* \in \mathbb{R}^p$. In our analysis, we will denote $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$ and $\overline{\mu} = \frac{1}{n}\sum_{i=1}^n \mu_i$ as the Lipschitz constant of $\nabla f(x)$ and the strongly convex constant of $f(x)$, respectively. We also denote $L = \max\{L_i\}$ and $\mu = \min\{\mu_i\}$.

Assumption 3. The graph $\mathcal{G}$ is connected and the nonnegative weight matrix $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ is doubly stochastic. In addition, $w_{ii} > 0$ for some $i \in \mathcal{V}$.

Let δ denote the spectral norm of the matrix $W - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is the n-dimensional column vector of all ones. By Assumption 3, we have $\delta < 1$.

For the sake of analysis, we let

$$\mathbf{x}_k = [x_k^1, x_k^2, ..., x_k^n]^T \in \mathbb{R}^{n\times p},\ \mathbf{y}_k = [y_k^1, y_k^2, ..., y_k^n]^T \in \mathbb{R}^{n\times p},\ \nabla\mathbf{f}(\mathbf{x}_k) = [\nabla f_1(x_k^1), \nabla f_2(x_k^2), ..., \nabla f_n(x_k^n)]^T \in \mathbb{R}^{n\times p},$$

$$\overline{x}_k = \frac{1}{n}\mathbf{1}^T\mathbf{x}_k \in \mathbb{R}^{1\times p}, \overline{y}_k = \frac{1}{n}\mathbf{1}^T\mathbf{y}_k \in \mathbb{R}^{1\times p},\ \nabla f(\overline{x}_k) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\overline{x}_k) \in \mathbb{R}^{1\times p}.$$

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

We denote I as the $n \times n$ identity matrix and $\rho(\cdot)$ as the spectral radius of a square matrix. Let $\|\cdot\|$ denote Euclidean norm for vectors, and Frobenius norm for matrices.

Based on the notations above, the DGM-BB-C method can be rewritten as the following compact matrix form:

$$\mathbf{x}_{k+1} = W^R[\mathbf{x}_k - D_k\mathbf{y}_k], \tag{B1}$$
$$\mathbf{y}_{k+1} = W^R[\mathbf{y}_k + \nabla\mathbf{f}(\mathbf{x}_{k+1}) - \nabla\mathbf{f}(\mathbf{x}_k)], \tag{B2}$$

where D_k is a diagonal matrix and $[D_k]^{ii} = \alpha_k^i$ is computed by (2).

Next, we derive the range of the distributed BB step sizes in DGM-BB-C.

Lemma 1. Under Assumptions 1-2, for all $k \geqslant 0$ and $i \in \mathcal{V}$, the BB step size α_k^i in DGM-BB-C satisfies

$$\frac{1}{L_i} \leqslant \alpha_k^i \leqslant \min\left\{\frac{1}{\mu_i}, t_k^i\right\}. \tag{B3}$$

Proof. Firstly, we give the proof of bounds on the BB step size $(\alpha_k^i)^{BB1}$. By the strong convexity of f_i, we obtain

$$(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)) \geqslant \mu_i\|x_k^i - x_{k-1}^i\|^2.$$

Thus, we have the upper bound on $(\alpha_k^i)^{BB1}$ since

$$\begin{aligned}(\alpha_k^i)^{BB1} &= \frac{(x_k^i - x_{k-1}^i)^T(x_k^i - x_{k-1}^i)}{(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i))} \\ &\leqslant \frac{\|x_k^i - x_{k-1}^i\|^2}{\mu_i\|x_k^i - x_{k-1}^i\|^2} \\ &= \frac{1}{\mu_i}.\end{aligned} \tag{B4}$$

Using the Cauchy inequality and the L_i-Lipschitz continuity of $\nabla f_i(x)$, one can get that $(\alpha_k^i)^{BB1}$ is uniformly lower bounded due to

$$\begin{aligned}(\alpha_k^i)^{BB1} &= \frac{(x_k^i - x_{k-1}^i)^T(x_k^i - x_{k-1}^i)}{(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i))} \\ &\geqslant \frac{\|x_k^i - x_{k-1}^i\|^2}{\|x_k^i - x_{k-1}^i\|\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|} \\ &\geqslant \frac{\|x_k^i - x_{k-1}^i\|^2}{L_i\|x_k^i - x_{k-1}^i\|^2} \\ &= \frac{1}{L_i}.\end{aligned} \tag{B5}$$

Now, we give the proof of bounds on the BB step size $(\alpha_k^i)^{BB2}$. By the L_i-Lipschitz continuity of $\nabla f_i(x)$, we get

$$(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)) \geqslant \frac{1}{L_i}\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2. \tag{B6}$$

Thus, $(\alpha_k^i)^{BB2}$ is uniformly lower bounded due to

$$\begin{aligned}(\alpha_k^i)^{BB2} &= \frac{(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i))}{\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2} \\ &\geqslant \frac{\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2}{L_i\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2} \\ &= \frac{1}{L_i}.\end{aligned} \tag{B7}$$

Then, by the strong convexity of f_i, we obtain

$$(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)) \leqslant \frac{1}{\mu_i}\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2. \tag{B8}$$

Thus, $(\alpha_k^i)^{BB2}$ is uniformly upper bounded due to

$$\begin{aligned}(\alpha_k^i)^{BB2} &= \frac{(x_k^i - x_{k-1}^i)^T(\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i))}{\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2} \\ &\leqslant \frac{\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2}{\mu_i\|\nabla f_i(x_k^i) - \nabla f_i(x_{k-1}^i)\|^2} \\ &= \frac{1}{\mu_i}.\end{aligned} \tag{B9}$$

The desired result then follows by using the definition of α_k^i.

Note that in terms of the strong convexity of f_i, one has $(s_k^i)^T z_k^i > 0$. Then, by the Cauchy inequality, it is easy to get $(\alpha_k^i)^{BB2} \leqslant \frac{\|s_k^i\|}{\|z_k^i\|} \leqslant (\alpha_k^i)^{BB1}$, which means that $(\alpha_k^i)^{BB1}$ is a longer step size while $(\alpha_k^i)^{BB2}$ is a shorter one.

We define $\alpha_{\max} = \max\limits_{k\geqslant 0}\{\alpha_k^i\}$ and $t_{\max} = \max\limits_{k\geqslant 0}\{t_k^i\}$. It follows from Lemma 1 that

$$\frac{1}{L} \leqslant \alpha_{\max} \leqslant \min\left\{\frac{1}{\mu}, t_{\max}\right\}. \tag{B10}$$

In the next lemma, we establish bounds on $\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|$, $\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|$ and $\|\overline{x}_{k+1} - (x^*)^T\|$ in terms of linear combinations of their past values.

Lemma 2. For all k, if $\frac{1}{n}\sum_{i=1}^n t_k^i \leqslant \frac{2}{\overline{L}} - \frac{\overline{\mu}}{aL\overline{L}}$, where $a > 1$, the following linear time invariant system inequality holds:

$$\mathbf{v}_{k+1} \leqslant G^\alpha \mathbf{v}_k, \ \forall k \tag{B11}$$

where $\mathbf{v}_k \in \mathbb{R}^3, G^\alpha \in \mathbb{R}^{3\times 3}$ are defined as

$$\mathbf{v}_k = \begin{bmatrix} \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| \\ \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| \\ \|\overline{x}_k - (x^*)^T\| \end{bmatrix},$$

$$G^\alpha = \begin{bmatrix} \delta^R + \delta^R L\alpha_{\max} & \delta^R\alpha_{\max} & \delta^R L\sqrt{n}\alpha_{\max} \\ 2\delta^R L + \delta^R L^2\alpha_{\max} & \delta^R + \delta^R L\alpha_{\max} & \delta^R L^2\sqrt{n}\alpha_{\max} \\ \frac{L}{\sqrt{n}}\alpha_{\max} & \frac{1}{\sqrt{n}}\alpha_{\max} & 1 - \frac{\overline{\mu}}{aL} \end{bmatrix},$$

where R is the number of inner consensus iterations, and δ is the spectral norm of the matrix $W - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Proof. Step 1: Bound $\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|$. From (B1), it follows that

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\| &= \|(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)W^R[\mathbf{x}_k - D_k\mathbf{y}_k]\| \\ &\leqslant \|(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)W^R\mathbf{x}_k\| + \|(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)W^R D_k\mathbf{y}_k\|. \\ &\leqslant \delta^R\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \delta^R\alpha_{\max}\|\mathbf{y}_k\|. \end{aligned} \tag{B12}$$

Taking the average of (B2) over i and doing it recursively, we have that $\overline{y}_k = \frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}_k)$ for all $k \geqslant 0$. By [10, Lemma 8 c)] and due to $\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^*) = 0$, we have

$$\begin{aligned} \|\mathbf{y}_k\| &\leqslant \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| + \|\mathbf{1}[\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}_k) - \nabla f(\overline{x}_k)]\| + \|\mathbf{1}[\nabla f(\overline{x}_k) - \frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^*)]\| \\ &\leqslant \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| + L\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + L\sqrt{n}\|\overline{x}_k - (x^*)^T\|. \end{aligned} \tag{B13}$$

Substituting (B13) into (B12) yields

$$\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\| \leqslant (\delta^R + \delta^R L\alpha_{\max})\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \delta^R\alpha_{\max}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| + \delta^R L\sqrt{n}\alpha_{\max}\|\overline{x}_k - (x^*)^T\|. \tag{B14}$$

Step 2: Bound $\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|$. Noticing the similarity between (B1) and (B2), it follows from (B2) that

$$\begin{aligned} \|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\| &\leqslant \delta^R\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| + \delta^R\|\nabla\mathbf{f}(\mathbf{x}_{k+1}) - \nabla\mathbf{f}(\mathbf{x}_k)\| \\ &\leqslant \delta^R\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| + \delta^R L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|, \end{aligned} \tag{B15}$$

where we have used [10, Lemma 8 a)] in the last inequality. Let us look into the last term in (B15), we have

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| &= \|W^R[\mathbf{x}_k - D_k\mathbf{y}_k] - \mathbf{x}_k\| \\ &= \|(W^R - I)[\mathbf{x}_k - \mathbf{1}\overline{x}_k] - W^R D_k\mathbf{y}_k\| \\ &\leqslant 2\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \alpha_{\max}\|\mathbf{y}_k\|. \end{aligned} \tag{B16}$$

By combining (B13), (B15) with (B16), one has

$$\begin{aligned} \|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\| \leqslant\ & (2\delta^R L + \delta^R L^2\alpha_{\max})\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + (\delta^R + \delta^R L\alpha_{\max})\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| \\ & + \delta^R L^2\sqrt{n}\alpha_{\max}\|\overline{x}_k - (x^*)^T\|. \end{aligned} \tag{B17}$$

Step 3: Bound $\|\overline{x}_{k+1} - (x^*)^T\|$. Taking the average of (B1) over i gives us

$$\overline{x}_{k+1} = \overline{x}_k - \frac{1}{n}\mathbf{1}^T D_k\mathbf{y}_k. \tag{B18}$$

It follows from (B18) that

$$\|\overline{x}_{k+1} - (x^*)^T\| = \|\overline{x}_k - \frac{1}{n}\mathbf{1}^T D_k\mathbf{y}_k + \frac{1}{n}\mathbf{1}^T(D_k - D_k)\mathbf{1}\overline{y}_k - (x^*)^T\|$$

$$\begin{aligned}&\leqslant \|\overline{x}_k-\frac{1}{n}\mathbf{1}^TD_k\mathbf{1}\overline{y}_k-(x^*)^T\|+\|\frac{1}{n}\mathbf{1}^TD_k(\mathbf{y}_k-\mathbf{1}\overline{y}_k)\|\\&\leqslant \|\overline{x}_k-\frac{1}{n}\mathbf{1}^TD_k\mathbf{1}\overline{y}_k-(x^*)^T\|+\frac{\alpha_{\max}}{\sqrt{n}}\|\mathbf{y}_k-\mathbf{1}\overline{y}_k\|.\end{aligned}\tag{B19}$$

Considering the first term in (B19), by using [10, Lemma 8 c)] and the fact that $\overline{y}_k=\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}_k)$, we have

$$\begin{aligned}&\|\overline{x}_k-\frac{1}{n}\mathbf{1}^TD_k\mathbf{1}\overline{y}_k-(x^*)^T\|\\&\leqslant \|\overline{x}_k-(\frac{1}{n}\sum_{i=1}^n\alpha_k^i)\nabla f(\overline{x}_k)-(x^*)^T\|+\frac{1}{n}\sum_{i=1}^n\alpha_k^i\|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}_k)-\nabla f(\overline{x}_k)\|\\&\leqslant \|\overline{x}_k-(\frac{1}{n}\sum_{i=1}^n\alpha_k^i)\nabla f(\overline{x}_k)-(x^*)^T\|+\alpha_{\max}\frac{L}{\sqrt{n}}\|\mathbf{x}_k-\mathbf{1}\overline{x}_k\|.\end{aligned}\tag{B20}$$

If $\frac{1}{n}\sum_{i=1}^n\alpha_k^i<\frac{2}{L}$, by [10, Lemma 10], we have

$$\|\overline{x}_k-(\frac{1}{n}\sum_{i=1}^n\alpha_k^i)\nabla f(\overline{x}_k)-(x^*)^T\|\leqslant\lambda\|\overline{x}_k-(x^*)^T\|,\tag{B21}$$

where $\lambda=\max\{|1-\frac{\overline{\mu}}{n}\sum_{i=1}^n\alpha_k^i|,|1-\frac{\overline{L}}{n}\sum_{i=1}^n\alpha_k^i|\}$. From Lemma 1, it follows that $\frac{1}{L}\leqslant\frac{1}{n}\sum_{i=1}^n\alpha_k^i\leqslant\frac{1}{n}\sum_{i=1}^nt_k^i$. If $\frac{1}{n}\sum_{i=1}^nt_k^i\leqslant\frac{2}{\overline{L}}-\frac{\overline{\mu}}{aL\overline{L}}$, where $a>1$ is a constant, then $\lambda\leqslant1-\frac{\overline{\mu}}{aL}$. Thus, (B20) and (B21) together yield

$$\|\overline{x}_k-\frac{1}{n}\mathbf{1}^TD_k\mathbf{1}\overline{y}_k-(x^*)^T\|\leqslant(1-\frac{\overline{\mu}}{aL})\|\overline{x}_k-(x^*)^T\|+\frac{L\alpha_{\max}}{\sqrt{n}}\|\mathbf{x}_k-\mathbf{1}\overline{x}_k\|.\tag{B22}$$

By substituting (B22) into (B19), one gets

$$\|\overline{x}_{k+1}-(x^*)^T\|\leqslant\frac{L\alpha_{\max}}{\sqrt{n}}\|\mathbf{x}_k-\mathbf{1}\overline{x}_k\|+\frac{\alpha_{\max}}{\sqrt{n}}\|\mathbf{y}_k-\mathbf{1}\overline{y}_k\|+(1-\frac{\overline{\mu}}{aL})\|\overline{x}_k-(x^*)^T\|.\tag{B23}$$

The proof is completed.

Note that a linear iterative relation between $\mathbf{v}_{k+1}$ and $\mathbf{v}_k$ with matrix G^α is established in $(B11)$. If $\rho(G^\alpha)<1$, then $(G^\alpha)^k$ converges linearly to 0 at rate $O(\rho(G^\alpha)^k)$, in which case $\|\mathbf{v}_k\|$ also converges linearly to 0 at rate $O(\rho(G^\alpha)^k)$.

The next lemma shows that when the largest step size $\alpha_{\max}$ satisfies $(B10)$, with the appropriate lower bound on R, the spectral radius of G^α is less than 1.

Lemma 3. Suppose that Assumptions 1-3 hold. Consider the matrix G^α defined in $(B11)$ with $\alpha_{\max}$ satisfying $(B10)$. If $R\geqslant\lceil\frac{\ln(3a+5)+3\ln\kappa}{-\ln\delta}\rceil+1$, where $\kappa=\frac{L}{\mu}$, then $\rho(G^\alpha)<1$.

Proof. In light of [11, Corollary 8.1.29], we derive the lower bound on R and a positive vector $c=[c_1,c_2,c_3]^T$ from

$$G^\alpha c<c,\tag{B24}$$

which is equivalent to the following set of inequalities

$$\begin{cases}(\delta^RLc_1+\delta^Rc_2+\delta^RL\sqrt{n}c_3)\alpha_{\max}<c_1(1-\delta^R),\\ \delta^RL(Lc_1+c_2+L\sqrt{n}c_3)\alpha_{\max}<(1-\delta^R)c_2-2\delta^RLc_1,\\ (\frac{Lc_1}{\sqrt{n}}+\frac{c_2}{\sqrt{n}})\alpha_{\max}<\frac{c_3\overline{\mu}}{aL}.\end{cases}\tag{B25}$$

Since the right hand side of the second inequality in $(B25)$ has to be positive, we obtain

$$\delta^R<\frac{c_2}{2Lc_1+c_2}.\tag{B26}$$

It follows from $(B25)$ that

$$\alpha_{\max}<\widehat{\alpha},\tag{B27}$$

where

$$\widehat{\alpha}\equiv\min\left\{\frac{(1-\delta^R)c_1}{\delta^RC},\frac{(1-\delta^R)c_2-2\delta^RLc_1}{\delta^RLC},\frac{\overline{\mu}\sqrt{n}c_3}{aL(Lc_1+c_2)}\right\},$$

$C=Lc_1+c_2+L\sqrt{n}c_3$, c_1,c_2,c_3 are positive constants and the range of δ^R is given in $(B26)$. Since the largest step size $\alpha_{\max}$ satisfies (B10), we require $\widehat{\alpha}>\frac{1}{\mu}$. Then, by further using $(B26)$, we get

$$\delta^R<\min\left\{\frac{\mu c_1}{C+\mu c_1},\frac{\mu c_2}{LC+2L\mu c_1+\mu c_2}\right\},\tag{B28}$$

and

$$c_3>\frac{aL(Lc_1+c_2)}{\mu\overline{\mu}\sqrt{n}}.\tag{B29}$$

We choose $c_2=Lc_1$, $c_3=\frac{3aL(Lc_1+c_2)}{2\mu^2\sqrt{n}}$. Then, from $(B28)$, we have $\delta^R<\frac{1}{\kappa(3a\kappa^2+2)+3}$ with $\kappa=\frac{L}{\mu}$. Taking the natural logarithm on both sides of the above relation obtains

$$R>\frac{\ln[\kappa(3a\kappa^2+2)+3]}{-\ln\delta}.$$

Due to $\kappa\geqslant1$, the above condition is fulfilled if $R>\frac{\ln(3a+5)\kappa^3}{-\ln\delta}$. The proof is completed.

Remark 1. Note that the lower bound on R may not be computable due to the global information, e.g., μ, L, δ. In practice, we need to manually optimize R to achieve the best performance. Because R is a positive integer and is in general not large, selecting a suitable R is not difficult. The derivation of the lower bound on R is based on $\frac{1}{\mu}$, while each α_k^i is generated automatically in practice, so $\alpha_{\max}$ may be smaller than $\frac{1}{\mu}$ required theoretically. This also leads to the fact that R required in practice may be smaller than the lower bound on R derived theoretically. Moreover, the lower bound on R is a function on δ that reflects the connectivity of the network. When the network is better connected, fewer inner consensus iterations are required. Numerical experiments also demonstrate these observations.

Appendix C Proof of Theorem 1

The proof of Theorem 1 is based on these results of Lemmas 1-3. When the largest step size $\alpha_{\max}$ satisfies $(B10)$, the proof of Theorem 1 is derived under suitable assumptions that include Assumptions 1-3, $\frac{1}{n}\sum_{i=1}^n t_k^i \leqslant \frac{2}{L} - \frac{\overline{\mu}}{aL\overline{L}}$ and $R \geqslant \lceil \frac{\ln(3a+5)+3\ln\kappa}{-\ln\delta} \rceil + 1$.

Proof. Applying (B11) recursively, we get

$$\mathbf{v}_k \leqslant (G^\alpha)^k \mathbf{v}_0. \tag{C1}$$

Taking the norm on both sides of the above relation gives

$$\|\mathbf{v}_k\| \leqslant \rho(G^\alpha)^k \|\mathbf{v}_0\|. \tag{C2}$$

Denote $v_k = \sum_{l=0}^{k-1} \|\mathbf{v}_l\|$, then $(C2)$ can be written as

$$\|\mathbf{v}_k\| = v_{k+1} - v_k \leqslant \rho(G^\alpha)^k \|\mathbf{v}_0\|, \tag{C3}$$

which implies that $v_{k+1} \leqslant v_k + \rho(G^\alpha)^k \|\mathbf{v}_0\|$. By Lemma 3, one has that v_k converges and therefore is bounded. For any $\gamma \in (\rho(G^\alpha), 1)$, it follows from $(C3)$ that

$$\lim_{k\to\infty} \frac{\|\mathbf{v}_k\|}{\gamma^k} \leqslant \lim_{k\to\infty} \frac{\rho(G^\alpha)^k \|\mathbf{v}_0\|}{\gamma^k} \leqslant \|\mathbf{v}_0\|. \tag{C4}$$

Therefore, $\|\mathbf{v}_k\| = O(\gamma^k)$. That is, there exists some positive constant T such that, for all k

$$\|\mathbf{v}_k\| \leqslant T(\rho(G^\alpha) + \xi)^k, \tag{C5}$$

where ξ is a arbitrarily small positive constant. Moreover,

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}^*\| &\leqslant \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \|\mathbf{1}\overline{x}_k - \mathbf{x}^*\| \\ &\leqslant \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \sqrt{n}\|\overline{x}_k - (x^*)^T\| \\ &\leqslant (1+\sqrt{n})\|\mathbf{v}_k\|, \end{aligned} \tag{C6}$$

where $\mathbf{x}^* = \mathbf{1}(x^*)^T$.

By combining $(C5)$ with $(C6)$, we obtain that

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leqslant (1+\sqrt{n})T(\rho(G^\alpha) + \xi)^k. \tag{C7}$$

Thus, the sequence $\{\mathbf{x}_k\}$ generated by DGM-BB-C converges exactly to the unique optimal solution $\mathbf{x}^*$ at a geometric rate.

Remark 2. Theorem 1 shows that the geometric convergence rate of DGM-BB-C is ensured when each BB step size α_k^i has such a lower bound $\frac{1}{L_i}$. The proposed algorithm provides a possible selection for a larger step size than previous works [5–7], where the step sizes of these methods are not greater than $\frac{1}{L}$ theoretically. The selection for larger step sizes has been empirically verified.

Appendix D Numerical Experiments

In this section, we analyze the performance of DGM-BB-C and illustrate our theoretical findings. We use the Metropolis constant edge weight matrix W [12].

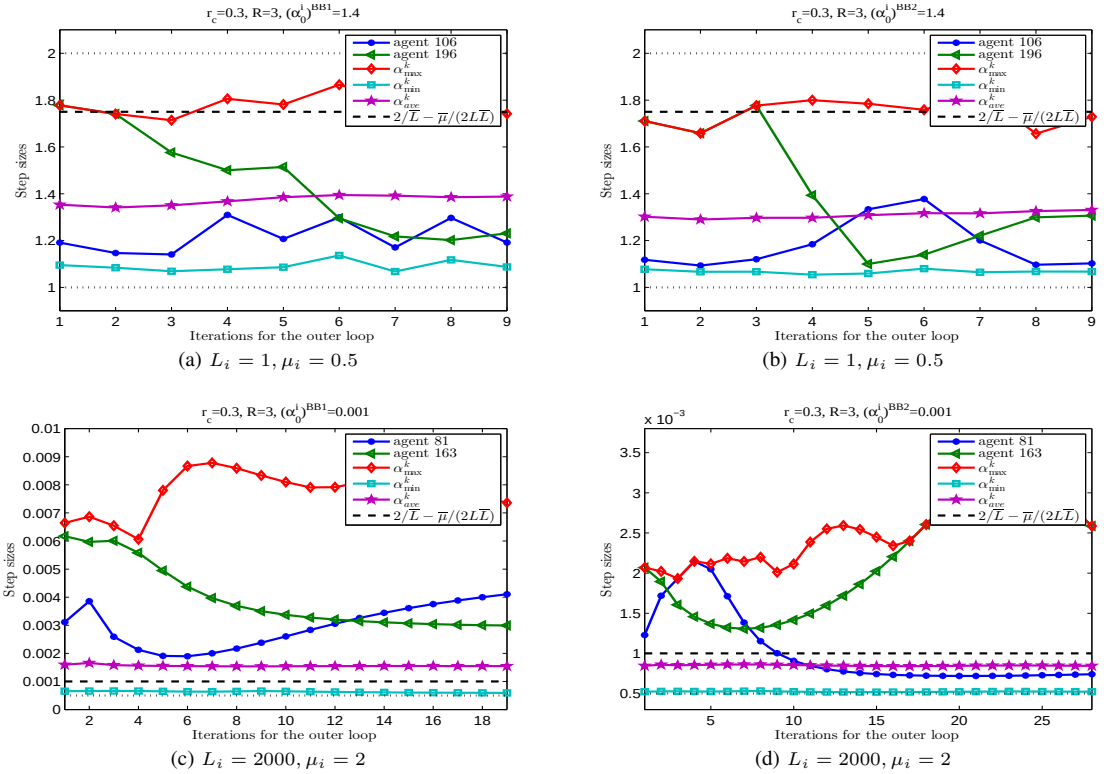
Appendix D.1 Distributed Least Squares

We consider a distributed sensing problem for solving an unknown signal $x \in \mathbb{R}^p$ [13]. Each agent $i \in \{1, 2, ..., n\}$ holds its own measurement equation, $y_i = M_i x + e_i$, where $y_i \in \mathbb{R}^{m_i}$ and $M_i \in \mathbb{R}^{m_i \times p}$ are measured data, $e_i \in \mathbb{R}^{m_i}$ is unknown noise. We apply the least squares loss and solve

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\|M_i x - y_i\|_2^2. \tag{D1}$$

In our experiment, the network is generated by using the Erdős-Rényi model with connectivity ratio r_c [14]. We set $n = 200, m_i = 20, p = 10$ and initialize $x_0^i = 0$ for every $i \in \mathcal{V}$.

To observe the behavior of the distributed BB step sizes in DGM-BB-C, when $t_k^i = +\infty$, DGM-BB-C with $(\alpha_k^i)^{BB}$ computed by (3) is tested on random network with $r_c = 0.3$. We set $L_i = 1, \mu_i = 0.5$ and $L_i = 2000, \mu_i = 2$, respectively. The parameter a is set to 2. We plot the maximum, minimum and average value of all each agents' α_k^i at each iteration, respectively. Additionally, we randomly choose two agents and follow the behavior of their step sizes. Figure D1 shows that the change of the BB step sizes is non-monotone and the BB step size is different for each agent. It can be seen from Figure D1 that all step sizes $(\alpha_k^i)^{BB1}$ and $(\alpha_k^i)^{BB2}$ are not less than the lower bound $\frac{1}{L_i}$ required theoretically. For different L_i and μ_i, the step sizes $(\alpha_k^i)^{BB2}$ are much smaller



(a) $L_i = 1, \mu_i = 0.5$ (b) $L_i = 1, \mu_i = 0.5$

(c) $L_i = 2000, \mu_i = 2$ (d) $L_i = 2000, \mu_i = 2$

Figure D1 The behaviour of distributed BB step sizes versus the number of iterations for the outer loop on random network with $r_c = 0.3$.

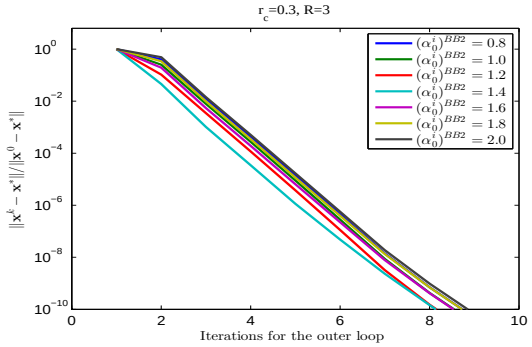


Figure D2 The performance of DGM-BB-C with different $(\alpha_0^i)^{BB2}$.

than the upper bound $\frac{1}{\mu_i}$ required theoretically and $\frac{1}{n}\sum_{i=1}^n(\alpha_k^i)^{BB2} \leqslant \frac{2}{L} - \frac{\mu}{2L\overline{L}}$ is satisfied, while $\frac{1}{n}\sum_{i=1}^n(\alpha_k^i)^{BB1} \leqslant \frac{2}{L} - \frac{\mu}{2L\overline{L}}$ is violated for $L_i = 2000, \mu_i = 2$. It means that DGM-BB-C with $(\alpha_k^i)^{BB1}$ needs to set safeguarding parameter to satisfy the requirement that $\frac{1}{n}\sum_{i=1}^n t_k^i \leqslant \frac{2}{L} - \frac{\mu}{2L\overline{L}}$, while DGM-BB-C with $(\alpha_k^i)^{BB2}$ does not. The numerical results presented below refer to the $(\alpha_k^i)^{BB2}$. DGM-BB-C is tested for $L_i = 1$ and $\mu_i = 0.5$. The safeguarding parameter t_k^i is set to 10^{20} for each agent.

Figure D2 shows the performance of DGM-BB-C with different initial step size $(\alpha_0^i)^{BB2}$ for each i when $r_c = 0.3$. We can see that DGM-BB-C is not sensitive to the choice of $(\alpha_0^i)^{BB2}$. There are similar results with different r_c.

From Figure D3, we can notice that DGM-BB-C with $R > 1$ is faster than it with $R = 1$, which illustrates the need for an additional consensus iteration. However, as R increases, the performance of DGM-BB-C does not improve much. Therefore, in our algorithm, we choose $R = 4$ for the less well-connected network, instead of $R = 3$ for the well-connected network, which matches the theory. Figure D3 also indicates that we manage to tune the proper R for DGM-BB-C in practice. In addition, since there is only one gradient computation at each iteration, Figure D3 can reflect that increasing the number of consensus iterations R properly can reduce the gradient computations (local computations) cost.

To show that the proposed algorithm is more effective, we compare the convergence rate of DGM-BB-C with several distributed algorithms, including DGD [3], NEAR-DGD$^+$ [5], EXTRA [4], DIGing [15], ATC-DIGing [7]. We also analyze the convergence rate of DGM-C, which is a practical variant of DGM-BB-C. DGM-C uses the same step size rule as that of ATC-DIGing instead of using the BB step sizes. For ATC-DIGing, we firstly tune an optimized identical step size by hand for all agents, which is $\frac{1}{L}$ in our experiment and then perturb it by random variables satisfying the uniform distribution over interval $(0.6, 1.2)$. For DGM-BB-C, we set $(\alpha_0^i)^{BB2} = 1.4$. The step sizes employed in other algorithms are hand-optimized.

Figures D4 and D5 plot the relative error $\|\mathbf{x}_k - \mathbf{x}^*\|/\|\mathbf{x}_0 - \mathbf{x}^*\|$ with respect to: (i) iterations for the outer loop, (ii) number of

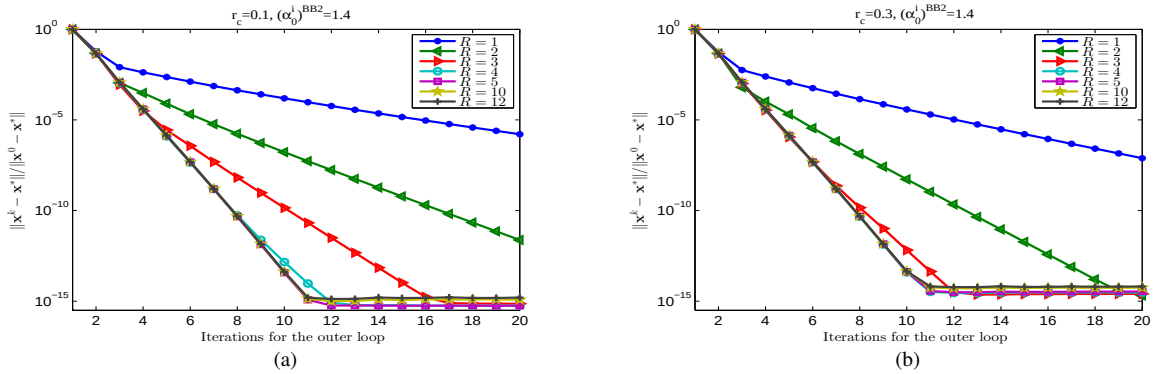


Figure D3 The performance of DGM-BB-C with different R on random network with different r_c. (a) $r_c = 0.1$; (b) $r_c = 0.3$.

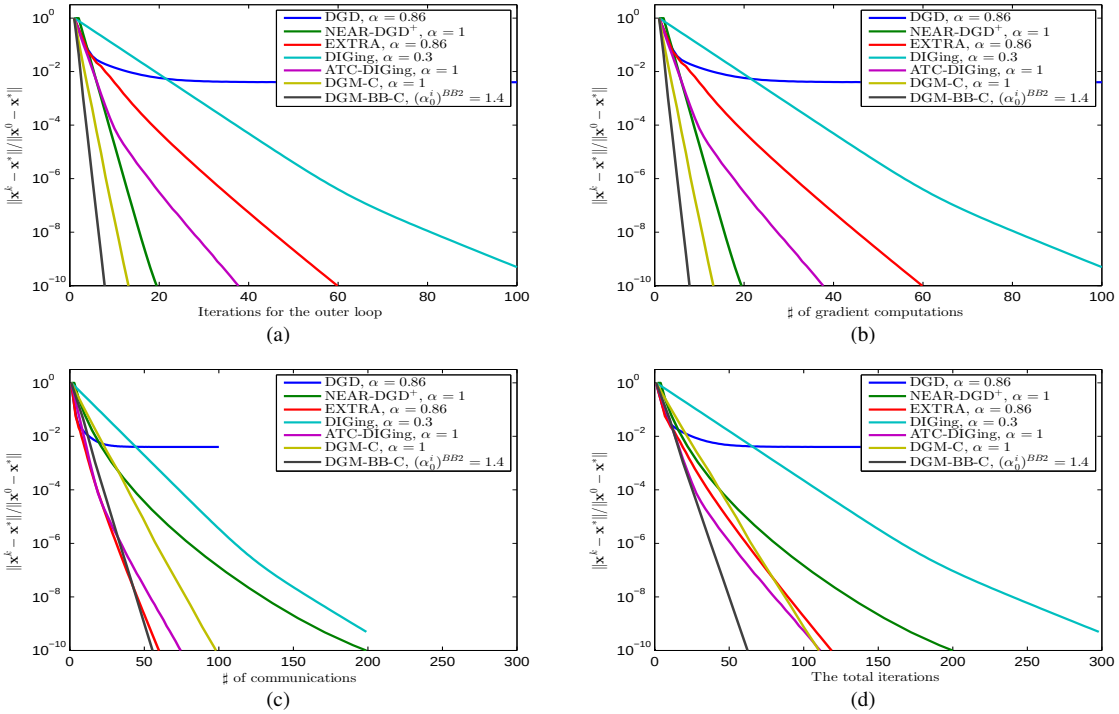


Figure D4 Convergence rates comparison of different distributed algorithms with $r_c = 0.1$.

gradient computations, (iii) number of communications, (iv) the total iterations. The total iterations represent the total number of iterations for the outer loop and the inner consensus loop. There is only one gradient computation at each iteration for the outer loop, and the number of iterations for the inner consensus loop is essentially the number of communications. So, these curves measured according to the total iterations can reflect the total cost of computations and communications. DGM-C performs better than ATC-DIGing, which suggests that doing inner loops of consensus iterations can improve the performance of the algorithm. It follows from Figures D4 and D5 that DGM-BB-C has the smallest number of iterations, gradient computations, and communications, and the lowest adaptive cost to reach an ϵ-optimal solution because it allows to use larger step sizes and does not need to increase the number of consensus steps in practice. NEAR-DGD$^+$ seeks to the exact solution by increasing the number of consensus steps linearly with the iteration number, whereas DGM-BB-C can converge to the exact solution when the number of consensus steps stays constant.

Appendix D.2 Distributed Logistic Regression

We next consider a distributed logistic regression problem over a 4-regular graph

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}\log(1+\exp(-(M_{ij}^T x)y_{ij})) + \frac{\nu}{2}\|x\|^2,$$

where M_{ij} is the feature vector and $y_{ij} \in \{-1, +1\}$ is the class label, and $\nu > 0$ is a weighting parameter. We set $\nu = 0.01$. In our experiments, we test the heart-scale dataset [16], with 270 data points, distributed uniformly over 10 agents. Each data point has a feature vector of size 13. The benchmark x^* is pre-computed with a centralized method. DGM-BB-C with $(\alpha_k^i)^{BB2}$ is tested

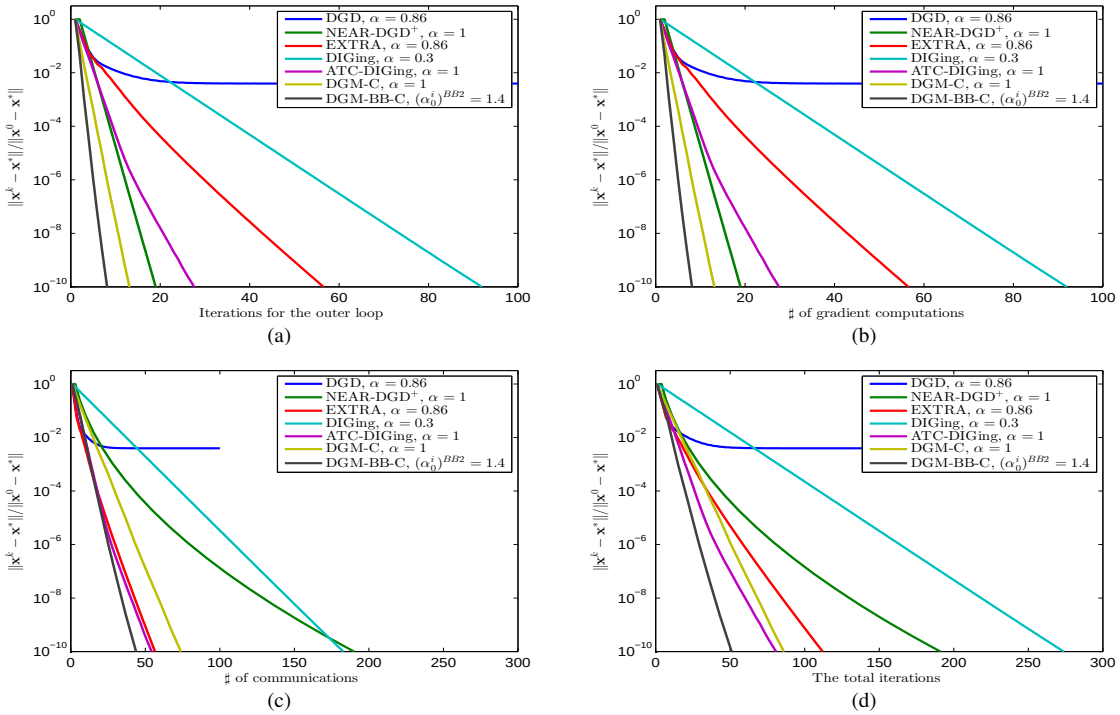


Figure D5 Convergence rates comparison of different distributed algorithms with $r_c = 0.3$.

and the safeguarding parameter t_k^i is set to 10^{20} for each agent. It follows from Figure D6 that among these compared algorithms, DGM-BB-C performs best according to the above four performances measured.

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 Autom Control, 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 Autom Control, 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, Japan, 2015. 2055–2060
7 Nedić A, Olshevsky A, Shi W, et al. Geometrically convergent distributed optimization with uncoordinated step-sizes. In: Proceedings of American Control Conference, Seattle, WA, USA, 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
10 Qu G, Li N. Harnessing smoothness to accelerate distributed optimization. IEEE Trans Control Netw Syst, 2018, 5: 1245–1260
11 Horn R A, Johnson C R. Matrix analysis. Cambridge: Cambridge University Press, 2012
12 Boyd S, Diaconis P, Xiao L. Fastest mixing markov chain on a graph. SIAM Rev, 2004, 46: 667–689
13 Li Z, Shi W, Yan M. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. IEEE Trans Signal Process, 2019, 67: 4494–4506
14 Erdos P, Renyi A. On random graphs i. Publ Math Debrecen, 1959, 6: 290–297
15 Nedic A, Olshevsky A, Shi W. Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J Optim, 2017, 27: 2597–2633
16 Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM T Intel Syst Tec, 2011, 2: 1–27

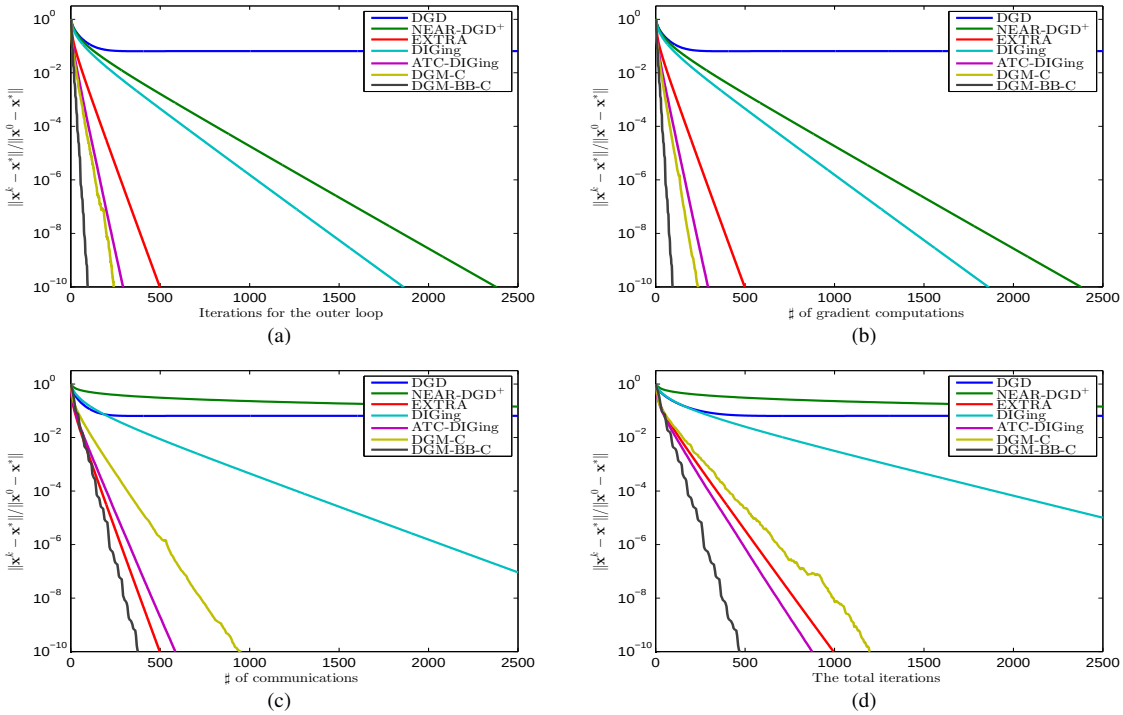


Figure D6 Convergence rates comparison of different distributed algorithms on heart-scale dataset.