A. Additional Preliminary Definitions

Combining the definitions of strong convexity in (8) and Bergman divergence in (9) results in an important property of the Bregman divergence associated with a σ-strongly convex function, i.e.,

$$D_r(x,y) \geq \frac{\sigma}{2} \parallel x-y \parallel^2 . \tag{19}$$

We assume that the Bregman divergence satisfies the following conditions (Bauschke & Borwein, 2001).

(a) (*Separate Convexity*) Let x and y_k, $k=1,\dots,K$, be arbitrary vectors in $\mathbb{R}^d$. Then, the following inequality holds:

$$D_r\Big(x, \sum_{k=1}^{K} \mu_k y_k\Big) \leq \sum_{k=1}^{K} \mu_k D_r(x, y_k), \tag{20}$$

where μ_k are non-negative constants with $\sum_k \mu_k = 1$.

(b) (*Lipschitz Continuity*) The Bregman divergence is Lipschitz continuous of the following form:

$$|D_r(x,z) - D_r(y,z)| \leq M \parallel x-y \parallel, \forall x,y,z \in \mathcal{X}, \tag{21}$$

where M is a positive constant.

Many commonly used functions satisfy (20), e.g., Euclidean distance and Kullback-Leibler divergence (refer to (Bauschke & Borwein, 2001) for further discussion and proof). Moreover, the condition in (21) is directly satisfied with a proper choice of Lipschitz continuous function $r(.)$ on the compact set $\mathcal{X}$.

B. Pseudocode of DABMD Algorithm

The psuedocode of DABMD operating over a heterogeneous network is given in Algorithm 1. Lines $3-9$ correspond to the any-batch computation step, in which every node i computes $b_{i,t}$ gradients during the fixed computation time interval, denoted by T_r. Line 10 corresponds to the update step, where nodes update their decisions using the mirror descent method. Line 11 represents the consensus averaging step, in which each node shares the local decision with its neighbors. Upon receiving the messages of neighboring nodes, every node i updates its estimate of the global minimizer $y_{i,t+1}$.

C. Proof of Lemma 1

In the proof of Lemma 1, we make use of another technical lemma. The following lemma presents an upper bound on the deviation of the local decisions from their approximate average.

Algorithm 1 DABMD algorithm

Input: initial points: $\{x_{i,0}, y_{i,0}\}$; step size α_t; time horizon T.
Output: sequence of decisions $\{x_{i,t}, y_{i,t} : 1 \leq t \leq T\}$.
1: **for** $t = 1, 2, \dots, T$ **do**
2: Initialize $b_{i,t} = 0, g_{i,t} = 0$
3: $T_0 = current\_time$
4: **while** $current\_time - T_0 \leq T_r$ **do**
5: Receive input $\omega_{i,t}^s$ sampled i.i.d from Ω
6: Compute gradient: $g_{i,t} \leftarrow g_{i,t} + \nabla f(y_{i,t}, \omega_{i,t}^s)$
7: $b_{i,t} \leftarrow b_{i,t} + 1$
8: **end while**
9: Normalize gradients: $g_{i,t} \leftarrow \frac{1}{b_{i,t}} g_{i,t}$
10: Update local decisions:
$x_{i,t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \Big\{ \langle x, g_{i,t} \rangle + \frac{1}{\alpha_t} D_r(x, y_{i,t}) \Big\}$
11: Average decisions by consensus iteration
single consensus iteration:
$y_{i,t+1} \leftarrow \sum_{j=1}^{n} P_{ij}^{t+1} x_{j,t+1}$
multiple consensus iteration:
$y_{i,t+1} \leftarrow \sum_{j=1}^{n} [P^{t+1}]_{ij}^{k} x_{j,t+1}$
12: **end for**

Lemma 7 *If every node uses a σ-strongly convex regularization function $r(.)$, we have*

$$\parallel x_{i,t+1} - y_{i,t} \parallel \leq \frac{L\alpha_t}{\sigma}, \forall i \in \mathcal{V}, \forall t \geq 0.$$

Proof. By applying the first-order optimality condition to the update (11), we get

$$\langle x - x_{i,t+1}, \alpha_t g_{i,t} + \nabla r(x_{i,t+1}) - \nabla r(y_{i,t}) \rangle \geq 0, \forall x \in \mathcal{X}. \tag{22}$$

Then, setting $x = y_{i,t}$ in (22) yields

$$\begin{aligned} &\langle y_{i,t} - x_{i,t+1}, \alpha_t g_{i,t} \rangle \\ &\quad \geq \langle x_{i,t+1} - y_{i,t}, \nabla r(x_{i,t+1}) - \nabla r(y_{i,t}) \rangle. \end{aligned} \tag{23}$$

Next, we exploit the strong convexity of regularizer $r(.)$, i.e.,

$$\begin{aligned} r(x_{i,t+1}) - r(y_{i,t}) - \nabla r(y_{i,t})^T (x_{i,t+1} - y_{i,t}) \\ \geq \frac{\sigma}{2} \parallel x_{i,t+1} - y_{i,t} \parallel^2 . \end{aligned}$$

Taking gradient with respect to $x_{i,t+1}$ yields

$$\nabla r(x_{i,t+1}) - \nabla r(y_{i,t}) \geq \sigma \parallel x_{i,t+1} - y_{i,t} \parallel .$$

Combining the above with (23), and using the Lipschitzness property, we obtain

$$\alpha_t L \parallel x_{i,t+1} - y_{i,t} \parallel \geq \sigma \parallel x_{i,t+1} - y_{i,t} \parallel^2 . \tag{24}$$

Dividing (24) by $\sigma \parallel x_{i,t+1} - y_{i,t} \parallel$ completes the proof. □

Now, we are ready to present an upperbound on the deviation of the local decisions from the exact average. So, we present the proof of Lemma 1.

Let $e_{i,t} = x_{i,t+1} - y_{i,t}$ denote the error between the local decision and local estimate between two consecutive time slots. Then, the update (12) can be recursively written as

$$\begin{aligned} x_{i,t+1} &= y_{i,t} + e_{i,t} = \sum_{j=1}^{n} P_{ij}^{t} x_j^t + e_{i,t} \\ &= \sum_{j=1}^{n} \Phi(t,0)_{ij} x_{j,0} + \sum_{s=1}^{t} \sum_{j=1}^{n} \Phi(t,s)_{ij} e_{j,s-1} + e_{i,t}. \end{aligned} \tag{25}$$

Averaging the above over the entire network yields

$$\begin{aligned} \bar{x}_{t+1} &= \frac{1}{n} \sum_{i=1}^{n} x_{i,t+1} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \Phi(t,0)_{ij} x_{j,0} \\ &\quad + \frac{1}{n} \sum_{s=1}^{t} \sum_{j=1}^{n} \sum_{i=1}^{n} \Phi(t,s)_{ij} e_{j,s-1} + \frac{1}{n} \sum_{i=1}^{n} e_{i,t} \\ &= \frac{1}{n} \sum_{j=1}^{n} x_{j,0} + \frac{1}{n} \sum_{s=1}^{t+1} \sum_{j=1}^{n} e_{j,s-1}, \end{aligned} \tag{26}$$

where in the last line, we have used the fact that $\Phi(t,0)$ and $\Phi(t,s)$ are doubly stochastic matrices, and thus $\sum_i \Phi(t,0) = \sum_i \Phi(t,s)_{ij} = 1$. By combining (25) and (26), we obtain

$$\begin{aligned} \parallel x_{i,t+1} - \bar{x}_{t+1} \parallel &\leq \sum_{j=1}^{n} \parallel \Phi(t,0)_{ij} - \frac{1}{n} \parallel \parallel x_{j,0} \parallel \\ &\quad + \sum_{s=1}^{t} \sum_{j=1}^{n} \parallel \Phi(t,s)_{ij} - \frac{1}{n} \parallel \parallel e_{j,s-1} \parallel \\ &\quad + \frac{1}{n} \sum_{j=1}^{n} \parallel e_{j,t} \parallel + \parallel e_{i,t} \parallel . \end{aligned}$$

We next use the result of Lemma 7 and (13) to bound the error and transition matrix errors, respectively. Noting that all initial local decisions are zero vectors, we have

$$\begin{aligned} &\parallel x_{i,t+1} - \bar{x}_{t+1} \parallel \\ &\quad \leq \sum_{s=1}^{t} \sum_{j=1}^{n} \gamma \Gamma^{t-s} \frac{L \alpha_{s-1}}{\sigma} + \frac{1}{n} \sum_{j=1}^{n} \frac{\alpha_t L}{\sigma} + \frac{\alpha_t L}{\sigma} \\ &\quad \leq \sum_{s=1}^{t} \frac{\alpha_0 L n}{\sigma} \gamma \Gamma^{t-s} + \frac{2 \alpha_t L}{\sigma} \\ &\quad \leq \frac{\alpha_0 L}{\sigma} \Big(2 + \frac{n\gamma}{1-\Gamma} \Big). \end{aligned} \tag{27}$$

where α_0 denotes the initial step size. To obtain the right-hand side above we used the fact that $\{\alpha_t\}$ is a non-increasing sequence. □

D. Proof of Theorem 2

D.1. Key Lemmas

The following two lemmas pave the way for our regret analysis provided in Theorem 2. Lemma 8 shows the impact of the dynamic minimizers on the regret bound.

Lemma 8 *For any non-increasing step size sequence $\{\alpha_t\}$ it holds that*

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{t=1}^{T} \Big(\frac{1}{\alpha_t} D_r(x_t^*, y_{i,t}) - \frac{1}{\alpha_t} D_r(x_t^*, x_{i,t+1}) \Big) \\ &\qquad \leq \frac{2nR^2}{\alpha_{T+1}} + \sum_{t=1}^{T} \frac{Mn \parallel x_{t+1}^* - x_t^* \parallel}{\alpha_{t+1}}, \end{aligned} \tag{28}$$

where $R^2 = \sup_{x,y \in \mathcal{X}} D_r(x,y)$.

Proof. We begin by adding and subtracting several terms as follows:

$$\begin{aligned} &\frac{1}{\alpha_t} D_r(x_t^*, y_{i,t}) - \frac{1}{\alpha_t} D_r(x_t^*, x_{i,t+1}) = \\ &\quad + \frac{1}{\alpha_t} D_r(x_t^*, y_{i,t}) - \frac{1}{\alpha_{t+1}} D_r(x_{t+1}^*, y_{i,t+1}) \\ &\quad + \frac{1}{\alpha_{t+1}} D_r(x_{t+1}^*, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x_t^*, y_{i,t+1}) \\ &\quad + \frac{1}{\alpha_{t+1}} D_r(x_t^*, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x_t^*, x_{i,t+1}) \\ &\quad + \frac{1}{\alpha_{t+1}} D_r(x_t^*, x_{i,t+1}) - \frac{1}{\alpha_t} D_r(x_t^*, x_{i,t+1}). \end{aligned} \tag{29}$$

We proceed by bounding every pair of terms on the right-hand side of (29). The first pair telescopes when summed over time t. For the second pair, by the Lipschitz condition on the Bregman divergence (21), we have

$$\begin{aligned} &\frac{1}{\alpha_{t+1}} D_r(x_{t+1}^*, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x_t^*, y_{i,t+1}) \\ &\qquad \leq \frac{M \parallel x_{t+1}^* - x_t^* \parallel}{\alpha_{t+1}}. \end{aligned} \tag{30}$$

Furthermore, using the separate convexity of the Bregman divergence given in (20), the third pair is bounded as fol-

lows:

$$\sum_{i=1}^{n}\Big(D_r(x_t^*,y_{i,t+1})-D_r(x_t^*,x_{i,t+1})\Big)$$
$$=\sum_{i=1}^{n}\left(D_r(x_t^*,\sum_{j=1}^{n}P_{ij}^t x_{j,t+1})-D_r(x_t^*,x_{i,t+1})\right)$$
$$\le\sum_{j=1}^{n}\left(\sum_{i=1}^{n}P_{ij}^t\right)D_r(x_t^*,x_{j,t+1})$$
$$-\sum_{i=1}^{n}D_r(x_t^*,x_{i,t+1})=0, \quad (31)$$

where the separate convexity of Bregman divergence and the doubly stochastic property of P^t are used in the last line. Summing (29) over time and nodes, we obtain

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{1}{\alpha_t}D_r(x_t^*,y_{i,t})-\frac{1}{\alpha_t}D_r(x_t^*,x_{i,t+1})\right)$$
$$\le\frac{nR^2}{\alpha_1}+\sum_{t=1}^{T}\frac{Mn\parallel x_{t+1}^*-x_t^*\parallel}{\alpha_{t+1}}$$
$$+nR^2\sum_{t=1}^{T}\left(\frac{1}{\alpha_{t+1}}-\frac{1}{\alpha_t}\right)$$
$$\le\frac{2nR^2}{\alpha_{T+1}}+\sum_{t=1}^{T}\frac{Mn\parallel x_{t+1}^*-x_t^*\parallel}{\alpha_{t+1}}, \quad (32)$$

where the fact that $\{\alpha_t\}$ is a positive and non-increasing sequence is used in the last line. □

Lemma 9 *Let* $b_{\min}=\min_{i,t}\{b_{i,t}\}$ *and* $b_{\max}=\max_{i,t}\{b_{i,t}\}$ *be the minimum and maximum of the mini-batch sizes across nodes and over time. The sequence* $y_{i,t}$ *generated by* (12) *satisfies*

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{E}\Big[f_{i,t}(y_{i,t})-f_{i,t}(x_t^*)\Big]$$
$$\le\sum_{t=1}^{T}\frac{V_bL^2n}{2\sigma b_{\min}}\alpha_t+\sum_{t=1}^{T}\frac{Mnb_{\max}\mathbb{E}\left[\parallel x_{t+1}^*-x_t^*\parallel\right]}{\alpha_{T+1}}$$
$$+\frac{2nb_{\max}R^2}{\alpha^{T+1}},$$

where $V_b=\mathbb{E}\left[b_{i,t}^2\right]$ *and the expectation is taken with respect to the variability in the minibatch sizes.*

Proof. We start by adding and subtracting $f_{i,t}(y_{i,t})$ as follows:

$$f_{i,t}(y_{i,t})-f_{i,t}(x_t^*)=\sum_{s=1}^{b_{i,t}}\big(f(y_{i,t},\omega_{i,t}^s)-f(x_t^*,\omega_{i,t}^s)\big)$$
$$\le b_{i,t}\langle g_{i,t},y_{i,t}-x_t^*\rangle$$
$$=b_{i,t}\langle g_{i,t},y_{i,t}-x_{i,t+1}+x_{i,t+1}-x_t^*\rangle$$
$$\le b_{i,t}L\parallel y_{i,t}-x_{i,t+1}\parallel+b_{i,t}\langle g_{i,t},x_{i,t+1}-x_t^*\rangle, \quad (33)$$

where we have used the convexity and Lipschitz continuity of $f_{i,t}(.)$ in the last line of (33). Next, we bound the last term of (33) as follows:

$$b_{i,t}\langle g_{i,t},x_{i,t+1}-x_t^*\rangle$$
$$\le\frac{b_{i,t}}{\alpha_t}\Big[D_r(x_t^*,y_{i,t})-D_r(x_t^*,x_{i,t})-D_r(x_{i,t+1},y_{i,t})\Big]$$
$$\le\frac{b_{i,t}}{\alpha_t}\Big[D_r(x_t^*,y_{i,t})-D_r(x_t^*,x_{i,t})\Big]$$
$$-\frac{\sigma b_{\min}}{2\alpha_t}\parallel y_{i,t}-x_{i,t+1}\parallel^2, \quad (34)$$

where we have used a simple algebra of Bregman divergences (see (Beck & Teboulle, 2003)) to derive the first inequality in (34). Also, the last line of (34) is obtained by the strong convexity of $r(.)$, as presented in (19).

Finally, by substituting (34) into (33), we get

$$f_{i,t}(y_{i,t})-f_{i,t}(x_t^*)$$
$$\le b_{i,t}L\parallel y_{i,t}-x_{i,t+1}\parallel-\frac{\sigma b_{\min}}{2\alpha_t}\parallel y_{i,t}-x_{i,t+1}\parallel^2$$
$$+\frac{b_{i,t}}{\alpha_t}\Big[D_r(x_t^*,y_{i,t})-D_r(x_t^*,x_{i,t})\Big]$$
$$\le\frac{b_{i,t}^2L^2\alpha_t}{2\sigma b_{\min}}+\frac{b_{\max}}{\alpha_t}\Big[D_r(x_t^*,y_{i,t})-D_r(x_t^*,x_{i,t})\Big], \quad (35)$$

where in the last line above we have used the fact that $cu-q\frac{u^2}{2}\le\frac{c^2}{2q}$, with $c=b_{i,t}L$ and $q=\sigma b_{\min}/\alpha_t$. We sum (35) across computing nodes and over time and take expectation, and apply Lemma 8 to the last term to achieve the result. □

D.2. Proof of the Theorem

Now, we are ready to present the proof of Theorem 2.

To bound the dynamic regret, we begin by adding and sub-

tracting $f_t(\bar{x}_t)$ as follows:

$$\begin{aligned}
&f_t(x_{i,t}) - f_t(x_t^*) \\
&\quad = f_t(x_{i,t}) - f_t(\bar{x}_t) + f_t(\bar{x}_t) - f_t(x_t^*) \\
&\quad = \sum_{j=1}^{n} \Big(f_{j,t}(x_{i,t}) - f_{j,t}(\bar{x}_t) \Big) \\
&\qquad + \sum_{i=1}^{n} \Big(f_{i,t}(\bar{x}_t) - f_{i,t}(y_{i,t}) + f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*) \Big) \\
&\quad \leq L \sum_{j=1}^{n} b_{j,t} \parallel x_{i,t} - \bar{x}_t \parallel + L \sum_{i=1}^{n} b_{i,t} \parallel y_{i,t} - \bar{x}_t \parallel \\
&\qquad + \sum_{i=1}^{n} \Big(f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*) \Big), \qquad (36)
\end{aligned}$$

where we have used the Lipschitz continuity of $f_{i,t}(.)$ to derive the last inequality in (36). We next need to bound the three terms in the last line of (36). The first term can be bounded using the result of Lemma 1. The second term is bounded as follows:

$$\begin{aligned}
L \sum_{i=1}^{n} b_{i,t} \parallel y_{i,t} - \bar{x}_t \parallel &= L \sum_{i=1}^{n} b_{i,t} \parallel \sum_{j=1}^{n} P_{ij}^t x_{j,t} - \bar{x}_t \parallel \\
&\leq L \sum_{i=1}^{n} b_{i,t} \sum_{j=1}^{n} P_{ij}^t \parallel x_{j,t} - \bar{x}_t \parallel \\
&\leq \frac{L^2}{\sigma} \alpha_0 \Big(2 + \frac{n\gamma}{1-\Gamma} \Big) \sum_{i=1}^{n} b_{i,t}, \qquad (37)
\end{aligned}$$

where we have used the result of Lemma 1 and doubly stochastic property of matrix P^t to obtain the last line. Substituting (37) into (36), taking expectation and summing over time, and combining that with the previous result in Lemma 9 completes the proof. □

E. Proof of Lemma 4

We use the following result on consensus averaging, which is presented in Lemma 1 of (Tsianos & Rabbat, 2016).

Let $\delta > 0$ be a given scalar and $\lambda_2(P^t)$ denote the second-largest eigen value of the doubly stochastic matrix P^t. If the number of consensus iterations satisfies

$$k \geq \frac{\log(\frac{1}{\delta} 2\sqrt{n} \max_j \parallel y_{j,t}^{(0)} - \bar{y}_t \parallel)}{1 - \lambda_2(P^t)}, \qquad (38)$$

the following bound holds on the output after k consensus iterations:

$$\parallel y_{i,t}^{(k)} - \bar{y}_t \parallel \leq \delta. \qquad (39)$$

Recall that the initial message $y_{i,t}^{(0)}$ is set to $x_{i,t}$, and thus the average $\bar{y}_t = \frac{1}{n} \sum_{i=1}^{n} y_{i,t}^{(0)}$ is equal to $\bar{x}_t$. Therefore, we can replace $\parallel y_{j,t}^{(0)} - \bar{y}_t \parallel$ by $\parallel x_{i,t} - \bar{x}_t \parallel$, and rewrite (38) as follows:

$$\delta \geq \frac{2\sqrt{n} \max_j \parallel y_{j,t}^{(0)} - \bar{y}_t \parallel}{\exp\Big[k \Big(1 - \lambda_2(P^t) \Big) \Big]}. \qquad (40)$$

Similar to Lemma 1, using the updates (16), the following bound can be established on the consensus error

$$\parallel x_{i,t} - \bar{x}_t \parallel \leq \frac{\alpha_0 L}{\sigma} \Big(2 + \frac{n\gamma^{(k)}}{1 - \Gamma^{(k)}} \Big). \qquad (41)$$

where $\gamma^{(k)}$ and $\Gamma^{(k)}$ are corresponding parameters of weight matrix $[P^t]^k$.

In addition, if D is an $n \times n$ doubly stochastic matrix, the second-largest eigen value is upper bounded by $\lambda_2(D) \leq 1 - n^{-3}$ (Landau & Odlyzko, 1981). Therefore, for every weight matrix P^t, we can conclude that

$$1 - \lambda_2(P^t) \leq n^{-3}. \qquad (42)$$

Finally, combining (40), (41), and (42), we can set $\delta = \frac{2\sqrt{n}\alpha_0 L}{\exp[kn^{-3}]\sigma} (2 + \frac{n\gamma^{(k)}}{1-\Gamma^{(k)}})$. This value of δ satisfies (38), and hence, we get

$$\parallel y_{i,t}^{(k)} - \bar{x}_t \parallel \leq \frac{2\sqrt{n}\alpha_0 L}{\exp[kn^{-3}]\sigma} \Big(2 + \frac{n\gamma^{(k)}}{1 - \Gamma^{(k)}} \Big). \qquad (43)$$

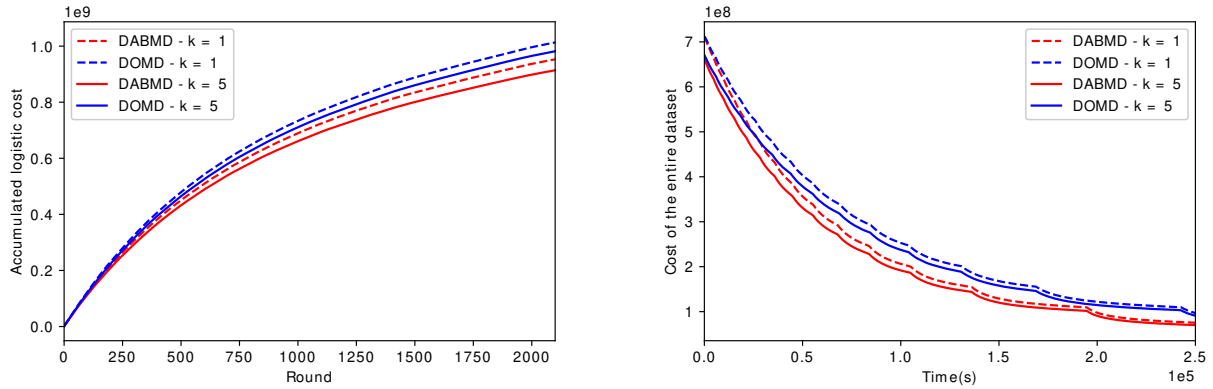
□

F. Proof of Theorem 5

To bound the dynamic regret, we begin by adding and subtracting $f_t(\bar{x}_t)$ as follows:

$$\begin{aligned}
&f_t(x_{i,t}) - f_t(x_t^*) \\
&\quad = f_t(x_{i,t}) - f_t(\bar{x}_t) + f_t(\bar{x}_t) - f_t(x_t^*) \\
&\quad = \sum_{j=1}^{n} \Big(f_{j,t}(x_{i,t}) - f_{j,t}(\bar{x}_t) \Big) \\
&\qquad + \sum_{i=1}^{n} \Big(f_{i,t}(\bar{x}_t) - f_{i,t}(y_{i,t}) + f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*) \Big) \\
&\quad \leq L \sum_{j=1}^{n} b_{j,t} \parallel x_{i,t} - \bar{x}_t \parallel + L \sum_{i=1}^{n} b_{i,t} \parallel y_{i,t} - \bar{x}_t \parallel \\
&\qquad + \sum_{i=1}^{n} \Big(f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*) \Big), \qquad (44)
\end{aligned}$$

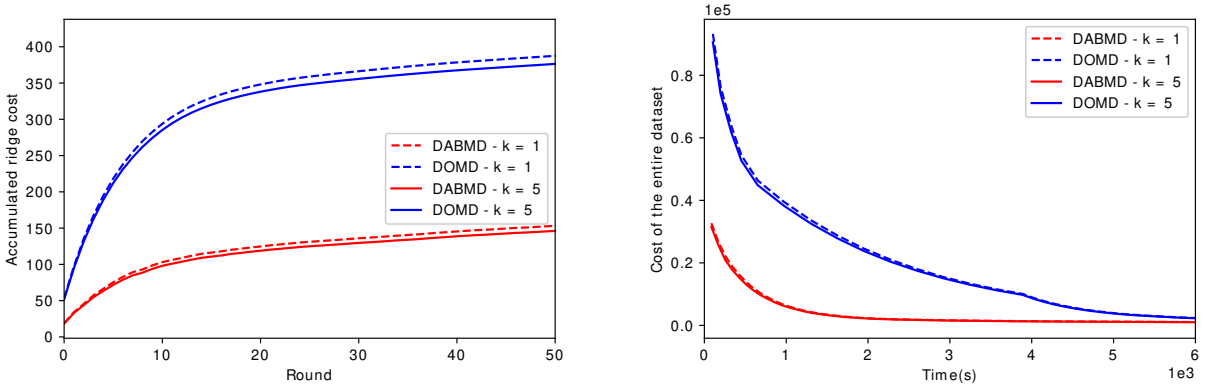
where we have used the Lipschitz continuity of $f_{i,t}(.)$. The first term on the right-hand side of (44) can be bounded by (41). The second term can be bounded using the result of



(a) Accumulated logistic cost of MNIST dataset.

(b) Logistic cost of MNIST dataset.

Figure 2. Performance comparison between DOMD and DABMD on MNIST dataset.



(a) Accumulated ridge cost of YouTube transcoding dataset.

(b) Ridge cost of YouTube transcoding dataset.

Figure 3. Performance comparison between DOMD and DABMD on YouTube transcoding dataset.

Lemma 4. By substituting (41) into (44), taking expectation and summing over time, and combining that with the previous result in Lemma 9, we obtain

$$\mathbb{E}[Reg_T^d] \leq \sum_{t=1}^{T} \frac{b_{\text{avg}} L^2 \alpha_0 n}{\sigma} \left(2 + \frac{n\gamma^{(k)}}{1-\Gamma^{(k)}}\right)\left(1 + \frac{2\sqrt{n}}{\exp[kn^{-3}]}\right) + \sum_{t=1}^{T} \frac{V_b L^2 \alpha_t}{2\sigma b_{\min}} + \sum_{t=1}^{T} \frac{Mnb_{\text{avg}}\mathbb{E}[\| x_{t+1}^* - x_t^* \|]}{\alpha_t} + \frac{2nR^2 b_{\text{avg}}}{\alpha_{T+1}}. \quad (45)$$

□

G. Additional Experiments

In this section, we present additional experiments regarding the performance of DABMD. Here we investigate the performance of DABMD with multiple consensus iterations using MNIST and YouTube transcoding datasets (Deneke et al., 2014). All system and experiment parameter settings are the same as those presented in Section 6, unless otherwise specified. In particular, we consider a time-varying network, similar to the one used in Yuan et. al. in ICLR'20, shown in Fig. 4. Note that the union graph of any two consecutive rounds is strongly connected.

G.1. Logistic Regression

We consider the MNIST logistic regression problem, where the computing nodes perform k iterations of consensus averaging.

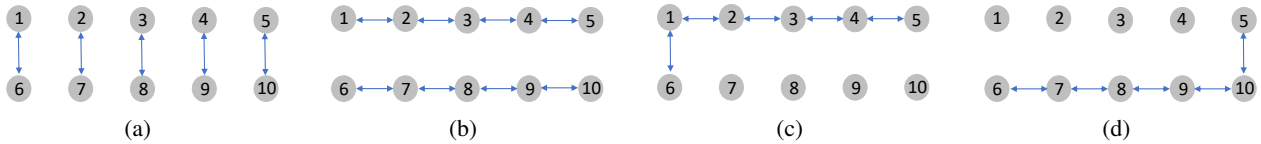


Figure 4. The network switches sequentially in a round robin manner among (a), (b), (c), and (d).

Fig. 2 illustrates the performance of DABMD versus DOMD on the MNIST logistic regression problem. We compare the results for $k = 1$ and $k = 5$ consensus averaging iterations. We observe that DABMD with $k = 5$ consensus averaging iterations incures up to 5% lower accumulated cost compared to the single consensus case. In addition, in both cases DABMD outperforms DOMD.

G.2. Ridge Regression

W also study the ridge regression problem on YouTube transcoding dataset under multiple consensus averaging settings.

We compare the performance of DABMD with DOMD for varying number of consensus iterations k in Fig. 3. We observe that the performance of DABMD improves with multiple consensus iterations, since computing nodes can more accurately approximate the global minimizer.