A. Additional Preliminary Definitions

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

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

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

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

$$D_r\left( \sum_{k=1}^{K} \mu_k y_k \right) \leq \sum_{k=1}^{K} \mu_k D_r(x, y_k), \quad (20)$$

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

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

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

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 pseudocode 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 $\alpha_t$; time horizon $T$.

Output: sequence of decisions $\{x_{i,t}, y_{i,t} : 1 \leq t \leq T\}$.

1: for $t = 1, 2, \ldots, T$ do
2: Initialize $b_{i,t} = 0, g_{i,t} = 0$
3: $T_0 = \text{current time}$
4: while current time $- T_0 \leq T_r$ do
5: Receive input $\omega_{i,t}^n$ sampled i.i.d from $\Omega$
6: Compute gradient: $g_{i,t} \leftarrow g_{i,t} + \nabla f(y_{i,t}, \omega_{i,t}^n)$
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:
11: Average decisions by consensus iteration
   single consensus iteration:
   $y_{i,t+1} \leftarrow \frac{1}{n} \sum_{j=1}^{n} P_{ij} x_{j,t+1}$
   multiple consensus iteration:
   $y_{i,t+1} \leftarrow \frac{1}{n} \sum_{j=1}^{n} [P_{ij}]^{s} x_{j,t+1}$
12: end for

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

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

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

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

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

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

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

$$r(x_{i,t+1}) - r(y_{i,t}) - \nabla r(y_{i,t})^T (x_{i,t+1} - y_{i,t}) \geq \frac{\sigma}{2} \| x_{i,t+1} - y_{i,t} \|^2. \quad (24)$$

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

$$\nabla r(x_{i,t+1}) - \nabla r(y_{i,t}) \geq \sigma \| x_{i,t+1} - y_{i,t} \|. \quad (25)$$

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

$$\alpha_t L \| x_{i,t+1} - y_{i,t} \| \geq \sigma \| x_{i,t+1} - y_{i,t} \|^2. \quad (24)$$
Dividing (24) by $\| x_{i,t+1} - y_{i,t} \|$ 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

$$x_{i,t+1} = y_{i,t} + e_{i,t} = \sum_{j=1}^{n} p_{ij}^t x_{j}^t + e_{i,t}$$

$$= \frac{1}{n} \sum_{j=1}^{n} \Phi(t,0)_{ij} x_{j,0} + \sum_{s=1}^{t} \sum_{j=1}^{n} \Phi(t,s)_{ij} e_{j,s-1} + e_{i,t}.$$  

(25)

Averaging the above over the entire network yields

$$\bar{x}_{t+1} = \frac{1}{n} \sum_{i=1}^{n} x_{i,t+1} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \Phi(t,0)_{ij} x_{j,0} + \sum_{s=1}^{t} \sum_{j=1}^{n} \Phi(t,s)_{ij} e_{j,s-1} + \frac{1}{n} \sum_{i=1}^{n} e_{i,t}$$

$$= \frac{1}{n} \sum_{j=1}^{n} x_{j,0} + \frac{1}{n} \sum_{s=1}^{t} \sum_{j=1}^{n} e_{j,s-1},$$  

(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

$$\| x_{i,t+1} - \bar{x}_{t+1} \| \leq \frac{1}{n} \sum_{j=1}^{n} \| \Phi(t,0)_{ij} - \frac{1}{n} \| \| x_{j,0} \|$$

$$+ \frac{1}{n} \sum_{s=1}^{t} \sum_{j=1}^{n} \| \Phi(t,s)_{ij} - \frac{1}{n} \| \| e_{j,s-1} \|$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \| e_{i,t} \|.$$  

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

$$\frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t}) - \frac{1}{\alpha_t} D_r(x^*_t, x_{i,t+1})$$

$$\leq \frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t}) - \frac{1}{\alpha_{t+1}} D_r(x^*_{i+1}, y_{i,t+1})$$

$$+ \frac{1}{\alpha_{t+1}} D_r(x^*_{i+1}, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t+1})$$

$$+ \frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x^*_t, x_{i,t+1})$$

$$+ \frac{1}{\alpha_{t+1}} D_r(x^*_t, x_{i,t+1}) - \frac{1}{\alpha_{t}} D_r(x^*_t, x_{i,t+1}).$$  

(30)

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

$$\| x_{i,t+1} - \bar{x}_{t+1} \|$$

$$\leq \sum_{s=1}^{t} \sum_{j=1}^{n} \gamma^{t-s} L \alpha_{s-1} \sigma + \frac{1}{n} \sum_{j=1}^{n} \alpha_{t} L + \frac{\alpha_t L}{\sigma}$$

$$\leq \sum_{s=1}^{t} \frac{\alpha_{t} L n}{\sigma} \gamma^{t-s} + \frac{2 \alpha_t L}{\sigma}$$

$$\leq \frac{\alpha_0 L}{\sigma} \left( 2 + \frac{n \gamma}{1 - \Gamma} \right).$$  

(27)

where $\alpha_0$ denotes the initial step size. To obtain the right-hand side above we used the fact that $\{\alpha_t\}$ is a non-increasing sequence. 

\section*{D. Proof of Theorem 2}

\subsection*{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.

\begin{lemma}
For any non-increasing step size sequence $\{\alpha_t\}$ it holds that

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \left( \frac{1}{\alpha_t} D_r(x^*_t, y_{i,t}) - \frac{1}{\alpha_t} D_r(x^*_t, x_{i,t+1}) \right)$$

$$\leq \frac{2n R^2}{\alpha T+1} + \sum_{t=1}^{T} \frac{M n \| x^*_{t+1} - x^*_t \|}{\alpha_{t+1}}.$$  

(28)

\end{lemma}

\begin{proof}
We begin by adding and subtracting several terms as follows:

$$\frac{1}{\alpha_t} D_r(x^*_t, y_{i,t}) - \frac{1}{\alpha_t} D_r(x^*_t, x_{i,t+1})$$

$$+ \frac{1}{\alpha_{t+1}} D_r(x^*_{t+1}, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t+1})$$

$$+ \frac{1}{\alpha_{t+1}} D_r(x^*_{t+1}, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t+1})$$

$$+ \frac{1}{\alpha_{t+1}} D_r(x^*_t, y_{i,t+1}) - \frac{1}{\alpha_{t+1}} D_r(x^*_t, x_{i,t+1}).$$  

(29)

Furthermore, using the separate convexity of the Bregman divergence given in (20), the third pair is bounded as fol-
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
\begin{align}
\sum_{i=1}^{n} \sum_{t=1}^{T} \left[ \frac{1}{\alpha_t} D_r(x_{i,t}^*, y_{i,t}) \right]
&= \frac{nR^2}{\alpha_1} + \sum_{t=1}^{T} \frac{Mn}{\alpha_t+1} \| x_{i,t+1}^* - x_t^* \| \\
&\leq \frac{2nR^2}{\alpha_{T+1}} + \sum_{t=1}^{T} \frac{Mn}{\alpha_t+1} \| x_{i,t+1}^* - x_t^* \|,
\end{align}

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
\begin{align}
\sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{E} \left[ f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*) \right]
&\leq \sum_{i=1}^{T} \frac{V_b L^2n}{2\sigma b_{\min}} \alpha_t + \sum_{t=1}^{T} \frac{Mnb_{\max}^2}{\alpha_{T+1}} \| x_{i,t+1}^* - x_t^* \| \\
&\quad + \frac{2nb_{\max} R^2}{\alpha t+1},
\end{align}

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

**Proof.** We start by adding and subtracting $f_{i,t}(y_{i,t})$ as fol-

$$f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*) = \sum_{s=1}^{b_{i,t}} \left( f(y_{i,t}, \omega_{i,t}^s) - f(x_t^*) \right)$$

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

$$b_{i,t}(g_{i,t}, x_{i,t+1} - x_t^*)$$

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

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

$$f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*)$$

where in the last line above we have used the fact that
$cu - q_c^2 \leq \frac{c^2}{2q_c}$, 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-

$$f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*)$$

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

$$f_{i,t}(y_{i,t}) - f_{i,t}(x_t^*)$$

where in the last line above we have used the fact that
$cu - q_c^2 \leq \frac{c^2}{2q_c}$, 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-
We use the following result on consensus averaging, which is the average:

\[ \text{Recall that the initial message } y^{(0)}_{i,t} \text{ is set to } x_{i,t}, \text{ and thus the average } \bar{y}_t = \frac{1}{n} \sum_{t=1}^{n} y^{(0)}_{i,t} \text{ is equal to } \bar{x}_t. \]

Therefore, we can replace \( \| y^{(0)}_{i,t} - \bar{y}_t \| \) by \( \| x_{i,t} - \bar{x}_t \| \), and rewrite (38) as follows:

\[ \delta \geq \frac{2\sqrt{n} \max_j \| y^{(0)}_{j,t} - \bar{y}_t \|}{\exp \left[ k \left( 1 - \lambda_2(P^t) \right) \right]} \]  

(40)

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

\[ \| x_{i,t} - \bar{x}_t \| \leq \frac{\alpha_0 L}{\sigma} \left( 2 + \frac{n\gamma^{(k)}}{1 - \Gamma^{(k)}} \right). \]  

(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

\[ 1 - \lambda_2(P^t) \leq n^{-3}. \]  

(42)

Finally, combining (40), (41), and (42), we can set \( \delta = \frac{2\sqrt{n} \max_j \| y^{(0)}_{j,t} - \bar{y}_t \|}{\exp[kn^{-3} \max_j \| y^{(0)}_{j,t} - \bar{y}_t \|]} \). This value of \( \delta \) satisfies (38), and hence, we get

\[ \| y^{(k)}_{i,t} - \bar{x}_t \| \leq \frac{2\sqrt{n} \alpha_0 L}{\exp[kn^{-3} \sigma]} \left( 2 + \frac{n\gamma^{(k)}}{1 - \Gamma^{(k)}} \right). \]  

(43)

\[ \square \]

F. Proof of Theorem 5

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

\[ f_t(x_{i,t}) - f_t(x^*_t) = f_t(x_{i,t}) - f_t(\bar{x}_t) + f_t(\bar{x}_t) - f_t(x^*_t) \]

\[ = \sum_{j=1}^{n} \left( f_{j,t}(x_{i,t}) - f_{j,t}(\bar{x}_t) \right) \]

\[ + \sum_{i=1}^{n} \left( f_{i,t}(\bar{x}_t) - f_{i,t}(y_{i,t}) + f_{i,t}(y_{i,t}) - f_{i,t}(x^*_t) \right) \]

\[ \leq L \sum_{j=1}^{n} b_{j,t} \| x_{i,t} - \bar{x}_t \| + \sum_{i=1}^{n} b_{i,t} \| y_{i,t} - \bar{x}_t \| \]

\[ + \sum_{i=1}^{n} \left( f_{i,t}(y_{i,t}) - f_{i,t}(x^*_t) \right), \]  

(44)

where we have used the Lipschitz continuity of \( f_{i,t}(\cdot) \). 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
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

$$
\mathbb{E}[R_{t+1}] \leq 
\sum_{t=1}^{T} \frac{b_{t+1}}{\sigma} \frac{L^2 \alpha_0 n}{2} \left( 2 + \frac{n^2}{1 - \Gamma(k)} \right) \left( 1 + \frac{2\sqrt{n}}{\exp[kn^{-3}]} \right) + 
\sum_{t=1}^{T} V_t L^2 \alpha_t + 
\sum_{t=1}^{T} M n b_{t+1} \mathbb{E} \left[ \| x_{t+1} - x_t \| \right] \alpha_t + 
\frac{2nR^2 b_{t+1}}{\alpha_T}.
$$

\[45\]

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.
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 incurs up to 5% lower accumulated cost compared to the single consensus case. In addition, in both cases DABMD outperforms DOMD.

G.2. Ridge Regression

We 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.