A Proof of Auxiliary Lemmas

Lemma 1. *For the sequence defined in (9), we have*

$$Z_{t+1} - Z_t = \alpha \frac{\beta_1}{1-\beta_1} \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot \left(\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}\right) - \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}. \tag{6}$$

Proof: By update rule of Algorithm 2, we first have

$$\begin{aligned}
\overline{X}_{t+1} =& \frac{1}{N} \sum_{i=1}^{N} x_{t+1,i} \\
=& \frac{1}{N} \sum_{i=1}^{N} \left(x_{t+0.5,i} - \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}} \right) \\
=& \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} W_{ij} x_{t,j} - \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}} \right) \\
\overset{(i)}{=}& \left(\frac{1}{N} \sum_{j=1}^{N} x_{t,j} \right) - \frac{1}{N} \sum_{i=1}^{N} \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}} \\
=& \overline{X}_t - \frac{1}{N} \sum_{i=1}^{N} \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}},
\end{aligned}$$

where (i) is due to an interchange of summation and $\sum_{i=1} W_{ij} = 1$. Then, we have

$$\begin{aligned}
Z_{t+1} - Z_t =& \overline{X}_{t+1} - \overline{X}_t + \frac{\beta_1}{1-\beta_1}(\overline{X}_{t+1} - \overline{X}_t) - \frac{\beta_1}{1-\beta_1}(\overline{X}_{t+1} - \overline{X}_t) \\
=& \frac{1}{1-\beta_1}(\overline{X}_{t+1} - \overline{X}_t) - \frac{\beta_1}{1-\beta_1}(\overline{X}_{t+1} - \overline{X}_t) \\
=& \frac{1}{1-\beta_1}\left(-\frac{1}{N} \sum_{i=1}^{N} \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}} \right) - \frac{\beta_1}{1-\beta_1}\left(-\frac{1}{N} \sum_{i=1}^{N} \alpha \frac{m_{t-1,i}}{\sqrt{u_{t-1,i}}} \right) \\
=& \frac{1}{1-\beta_1}\left(-\frac{1}{N} \sum_{i=1}^{N} \alpha \frac{\beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}}{\sqrt{u_{t,i}}} \right) - \frac{\beta_1}{1-\beta_1}\left(-\frac{1}{N} \sum_{i=1}^{N} \alpha \frac{m_{t-1,i}}{\sqrt{u_{t-1,i}}} \right) \\
=& \alpha \frac{\beta_1}{1-\beta_1} \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot \left(\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}\right) - \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}},
\end{aligned}$$

which is the desired result. □

Lemma 2. *Given a set of numbers $a_1, \cdots, a_n$ and denote their mean to be $\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$. Define $b_i(r) \triangleq \max(a_i, r)$ and $\bar{b}(r) = \frac{1}{n}\sum_{i=1}^{n} b_i(r)$. For any r and r' with $r' \geq r$ we have*

$$\sum_{i=1}^{n} |b_i(r) - \bar{b}(r)| \geq \sum_{i=1}^{n} |b_i(r') - \bar{b}(r')| \tag{7}$$

and when $r \leq \min_{i\in[n]} a_i$, we have

$$\sum_{i=1}^{n} |b_i(r) - \bar{b}(r)| = \sum_{i=1}^{n} |a_i - \bar{a}|. \tag{8}$$

Proof: Without loss of generality, assume $a_i \leq a_j$ when $i < j$, i.e. a_i is a non-decreasing sequence. Define

$$h(r) = \sum_{i=1}^{n} |b_i(r) - \bar{b}(r)| = \sum_{i=1}^{n} |\max(a_i, r) - \frac{1}{n}\sum_{j=1}^{n} \max(a_j, r)|.$$

We need to prove that h is a non-increasing function of r. First, it is easy to see that h is a continuous function of r with non-differentiable points $r = a_i, i \in [n]$, thus h is a piece-wise linear function.

Next, we will prove that $h(r)$ is non-increasing in each piece. Define $l(r)$ to be the largest index with $a(l(r)) < r$, and $s(r)$ to be the largest index with $a_{s(r)} < \bar{b}(r)$. Note that we have for $i \leq l(r)$, $b_i(r) = r$ and for $i \leq s(r)$ $b_i(r) - \bar{b}(r) \leq 0$ since a_i is a non-decreasing sequence. Therefore, we have

$$h(r) = \sum_{i=1}^{l(r)} (\bar{b}(r) - r) + \sum_{i=l(r)+1}^{s(r)} (\bar{b}(r) - a_i) + \sum_{i=s(r)+1}^{n} (a_i - \bar{b}(r))$$

and

$$\bar{b}(r) = \frac{1}{n}\left(l(r)r + \sum_{i=l(r)+1}^{n} a_i \right).$$

Taking derivative of the above form, we know the derivative of $h(r)$ at differentiable points is

$$\begin{aligned} h'(r) =& l(r)(\frac{l(r)}{n} - 1) + (s(r) - l(r))\frac{l(r)}{n} - (n - s(r))\frac{l(r)}{n} \\ =& \frac{l(r)}{n}((l(r) - n) + (s(r) - l(r)) - (n - s(r)))\,. \end{aligned}$$

Since we have $s(r) \leq n$ we know $(l(r) - n) + (s(r) - l(r)) - (n - s(r)) \leq 0$ and thus

$$h'(r) \leq 0\,,$$

which means $h(r)$ is non-increasing in each piece. Combining with the fact that $h(r)$ is continuous, (7) is proven. When $r \leq a(i)$, we have $b(i) = \max(a_i, r) = r$, for all $r \in [n]$ and $\bar{b}(r) = \frac{1}{n}\sum_{i=1}^{n} a_i = \bar{a}$ which proves (8). $\square$

B Proof of Theorem 2

To prove convergence of the algorithm, we first define an auxiliary sequence

$$Z_t = \overline{X}_t + \frac{\beta_1}{1 - \beta_1}(\overline{X}_t - \overline{X}_{t-1})\,, \tag{9}$$

with $\overline{X}_0 \triangleq \overline{X}_1$. Since $\mathbb{E}[g_{t,i}] = \nabla f(x_{t,i})$ and $u_{t,i}$ is a function of $G_{1:t-1}$ (which denotes $G_1, G_2, \cdots, G_{t-1}$), we have

$$\mathbb{E}_{G_t|G_{1:t-1}}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\,.$$

Assuming smoothness (A1) we have

$$f(Z_{t+1}) \leq f(Z_t) + \langle \nabla f(Z_t), Z_{t+1} - Z_t \rangle + \frac{L}{2}\|Z_{t+1} - Z_t\|^2\,.$$

Using Lemma 1 into the above inequality and take expectation over G_t given $G_{1:t-1}$, we have

$$\begin{aligned} &\mathbb{E}_{G_t|G_{1:t-1}}[f(Z_{t+1})] \\ \leq& f(Z_t) - \alpha \left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N} \frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}} \right\rangle + \frac{L}{2}\mathbb{E}_{G_t|G_{1:t-1}}\left[\|Z_{t+1} - Z_t\|^2\right] \\ &+ \alpha \frac{\beta_1}{1 - \beta_1}\mathbb{E}_{G_t|G_{1:t-1}}\left[\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N} m_{t-1,i} \odot (\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}) \right\rangle\right]. \end{aligned}$$

Then take expectation over $G_{1:t-1}$ and rearrange, we have

$$
\begin{aligned}
&\alpha\mathbb{E}\left[\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\rangle\right] && (10)\\
\leq&\mathbb{E}[f(Z_t)] - \mathbb{E}[f(Z_{t+1})] + \frac{L}{2}\mathbb{E}\left[\|Z_{t+1}-Z_t\|^2\right]\\
&+\alpha\frac{\beta_1}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N} m_{t-1,i}\odot\left(\frac{1}{\sqrt{u_{t-1,i}}}-\frac{1}{\sqrt{u_{t,i}}}\right)\right\rangle\right]. && (11)
\end{aligned}
$$

In addition, we have

$$
\begin{aligned}
&\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\rangle\\
=&\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\rangle + \left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})\odot\left(\frac{1}{\sqrt{u_{t,i}}}-\frac{1}{\sqrt{\overline{U}_t}}\right)\right\rangle && (12)
\end{aligned}
$$

and the first term on RHS of the equality can be lower bounded as

$$
\begin{aligned}
&\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\rangle\\
=&\frac{1}{2}\left\|\frac{\nabla f(Z_t)}{\overline{U}_t^{1/4}}\right\|^2 + \frac{1}{2}\left\|\frac{\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})}{\overline{U}_t^{1/4}}\right\|^2 - \frac{1}{2}\left\|\frac{\nabla f(Z_t)-\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})}{\overline{U}_t^{1/4}}\right\|^2\\
\geq&\frac{1}{4}\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 + \frac{1}{4}\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 - \frac{1}{2}\left\|\frac{\nabla f(Z_t)-\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})}{\overline{U}_t^{1/4}}\right\|^2\\
&-\frac{1}{2}\left\|\frac{\nabla f(Z_t)-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 - \frac{1}{2}\left\|\frac{\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\\
\geq&\frac{1}{2}\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 - \frac{3}{2}\left\|\frac{\nabla f(Z_t)-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 - \frac{3}{2}\left\|\frac{\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2, && (13)
\end{aligned}
$$

where the inequalities are all due to Cauchy-Schwartz. Substituting (13) and (12) into (10), we get

$$
\begin{aligned}
\frac{1}{2}\alpha\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] \leq&\mathbb{E}[f(Z_t)] - \mathbb{E}[f(Z_{t+1})] + \frac{L}{2}\mathbb{E}\left[\|Z_{t+1}-Z_t\|^2\right]\\
&+\alpha\frac{\beta_1}{1-\beta_1}\mathbb{E}\left[\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N} m_{t-1,i}\odot\left(\frac{1}{\sqrt{u_{t-1,i}}}-\frac{1}{\sqrt{u_{t,i}}}\right)\right\rangle\right]\\
&-\alpha\mathbb{E}\left[\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})\odot\left(\frac{1}{\sqrt{u_{t,i}}}-\frac{1}{\sqrt{\overline{U}_t}}\right)\right\rangle\right]\\
&+\frac{3}{2}\alpha\mathbb{E}\left[\left\|\frac{\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 + \left\|\frac{\nabla f(Z_t)-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right].
\end{aligned}
$$

Then sum over the above inequality from $t=1$ to T and divide both sides by $T\alpha/2$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] \\
\leq &\frac{2}{T\alpha}(\mathbb{E}[f(Z_1)]-\mathbb{E}[f(Z_{T+1})])+\frac{L}{T\alpha}\sum_{t=1}^{T}\mathbb{E}\left[\|Z_{t+1}-Z_t\|^2\right] \\
&+\frac{2}{T}\frac{\beta_1}{1-\beta_1}\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\left\langle\nabla f(Z_t),\frac{1}{N}\sum_{i=1}^{N}m_{t-1,i}\odot(\frac{1}{\sqrt{u_{t-1,i}}}-\frac{1}{\sqrt{u_{t,i}}})\right\rangle\right]}_{D_1} \\
&+\frac{2}{T}\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\left\langle\nabla f(Z_t),\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})\odot\left(\frac{1}{\sqrt{\overline{U}_t}}-\frac{1}{\sqrt{u_{t,i}}}\right)\right\rangle\right]}_{D_2} \\
&+\frac{3}{T}\underbrace{\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2+\left\|\frac{\nabla f(Z_t)-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right]}_{D_3}.
\end{aligned}
\tag{14}
$$

Now we need to upper bound all the terms on RHS of the above inequality to get the convergence rate. For the terms composing D_3 in (14), we can upper bound them by

$$
\begin{aligned}
\left\|\frac{\nabla f(Z_t)-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 &\leq \frac{1}{\min_{j\in[d]}[\overline{U}_t^{1/2}]_j}\left\|\nabla f(Z_t)-\nabla f(\overline{X}_t)\right\|^2 \\
&\leq L\frac{1}{\min_{j\in[d]}[\overline{U}_t^{1/2}]_j}\underbrace{\left\|Z_t-\overline{X}_t\right\|^2}_{D_4}
\end{aligned}
\tag{15}
$$

and

$$
\begin{aligned}
\left\|\frac{\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})-\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2 \leq &\frac{1}{\min_{j\in[d]}[\overline{U}_t^{1/2}]_j}\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(x_{t,i})-\nabla f(\overline{X}_t)\right\|^2 \\
\leq &L\frac{1}{\min_{j\in[d]}[\overline{U}_t^{1/2}]_j}\frac{1}{N}\underbrace{\sum_{i=1}^{N}\left\|x_{t,i}-\overline{X}_t\right\|^2}_{D_5},
\end{aligned}
\tag{16}
$$

using Jensen's inequality, Lipschitz continuity of f_i, and the fact that $f=\frac{1}{N}\sum_{i=1}^{N}f_i$. Next we need to bound D_4 and D_5. Recall the update rule of X_t, we have

$$
X_t=X_{t-1}W-\alpha\frac{M_{t-1}}{\sqrt{U_{t-1}}}=X_1W^{t-1}-\alpha\sum_{k=0}^{t-2}\frac{M_{t-k-1}}{\sqrt{U_{t-k-1}}}W^k\,,
\tag{17}
$$

where we define $W^0=\mathbf{I}$. Since W is a symmetric matrix, we can decompose it as $W=Q\Lambda Q^T$ where Q is a orthonormal matrix and Λ is a diagonal matrix whose diagonal elements correspond to eigenvalues of W in an descending order, i.e. $\Lambda_{ii}=\lambda_i$ with λ_i being ith largest eigenvalue of W. In addition, because W is a doubly stochastic matrix, we know $\lambda_1=1$ and $q_1=\frac{\mathbf{1}_N}{\sqrt{N}}$. With eigen-decomposition of W, we can rewrite D_5 as

$$
\sum_{i=1}^{N}\left\|x_{t,i}-\overline{X}_t\right\|^2=\|X_t-\overline{X}_t\mathbf{1}_N^T\|_F^2=\|X_tQQ^T-X_t\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\|_F^2=\sum_{l=2}^{N}\|X_tq_l\|^2\,.
\tag{18}
$$

In addition, we can rewrite (17) as

$$X_t = X_1 W^{t-1} - \alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{\sqrt{U_{t-k-1}}} W^k = X_1 - \alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{\sqrt{U_{t-k-1}}} Q\Lambda^k Q^T \,, \tag{19}$$

where the last equality is because $x_{1,i} = x_{1,j}$, for all i, j and thus $X_1 W = X_1$. Then we have when $l > 1$,

$$X_t q_l = (X_1 - \alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{\sqrt{U_{t-k-1}}} Q\Lambda^k Q^T) q_l = -\alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{\sqrt{U_{t-k-1}}} q_l \lambda_l^k \,, \tag{20}$$

since Q is orthonormal and $X_1 q_l = x_{1,1} \mathbf{1}_N^T q_l = x_{1,1} \sqrt{N} q_1^T q_l = 0$, for all $l \neq 1$.

Combining (18) and (20), we have

$$\begin{aligned} D_5 = \sum_{i=1}^{N} \left\| x_{t,i} - \overline{X}_t \right\|^2 &= \sum_{l=2}^{N} \| X_t q_l \|^2 \\ &= \sum_{l=2}^{N} \alpha^2 \left\| \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{\sqrt{U_{t-k-1}}} \lambda_l^k q_l \right\|^2 \\ &\leq \alpha^2 \left(\frac{1}{1-\lambda} \right)^2 N d G_\infty^2 \frac{1}{\epsilon} \,, \end{aligned} \tag{21}$$

where the last inequality follows from the fact that $g_{t,i} \leq G_\infty$, $\|q_l\| = 1$, and $|\lambda_l| \leq \lambda < 1$. Now let us turn to D_4, it can be rewritten as

$$\begin{aligned} \left\| Z_t - \overline{X}_t \right\|^2 = \left\| \frac{\beta_1}{1-\beta_1} (\overline{X}_t - \overline{X}_{t-1}) \right\|^2 &= \left(\frac{\beta_1}{1-\beta_1} \right)^2 \alpha^2 \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{m_{t-1,i}}{\sqrt{u_{t-1,i}}} \right\|^2 \\ &\leq \left(\frac{\beta_1}{1-\beta_1} \right)^2 \alpha^2 d \frac{G_\infty^2}{\epsilon} \,. \end{aligned} \tag{22}$$

Now we know both D_4 and D_5 are in the order of $\mathcal{O}\alpha^2)$ and thus D_3 is in the order of $\mathcal{O}\alpha^2)$. Next we will bound D_2 and D_1. Define $G_1 \triangleq \max_{t \in [T]} \max_{i \in [N]} \|\nabla f_i(x_{t,i})\|_\infty$, $G_2 \triangleq \max_{t \in [T]} \|\nabla f(Z_t)\|_\infty$, $G_3 \triangleq \max_{t \in [T]} \max_{i \in [N]} \|g_{t,i}\|_\infty$ and $G_\infty = \max(G_1, G_2, G_3)$. Then we have

$$\begin{aligned} D_2 &= \sum_{t=1}^{T} \mathbb{E} \left[\left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \odot \left(\frac{1}{\sqrt{\overline{U}_t}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right\rangle \right] \\ &\leq \sum_{t=1}^{T} \mathbb{E} \left[G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \frac{1}{\sqrt{[\overline{U}_t]_j}} - \frac{1}{\sqrt{[u_{t,i}]_j}} \right| \right] \\ &= \sum_{t=1}^{T} \mathbb{E} \left[G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \frac{1}{\sqrt{[\overline{U}_t]_j}} - \frac{1}{\sqrt{[u_{t,i}]_j}} \right| \frac{\sqrt{[\overline{U}_t]_j} + \sqrt{[u_{t,i}]_j}}{\sqrt{[\overline{U}_t]_j} + \sqrt{[u_{t,i}]_j}} \right] \\ &= \sum_{t=1}^{T} \mathbb{E} \left[G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \frac{[\overline{U}_t]_j - [u_{t,i}]_j}{[\overline{U}_t]_j \sqrt{[u_{t,i}]_j} + \sqrt{[\overline{U}_t]_j} [u_{t,i}]_j} \right| \right] \\ &\leq \mathbb{E} \underbrace{\left[\sum_{t=1}^{T} G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \frac{[\overline{U}_t]_j - [u_{t,i}]_j}{2\epsilon^{1.5}} \right| \right]}_{D_6} , \end{aligned} \tag{23}$$

where the last inequality is due to $[u_{t,i}]_j \geq \epsilon$, for all t, i, j. To simplify notations, define $\|A\|_{abs} = \sum_{i,j} |A_{ij}|$ to be the entry-wise L_1 norm of a matrix A, then we obtain

$$
\begin{aligned}
D_6 \leq \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \|\overline{U}_t \mathbf{1}^T - U_t\|_{abs} \leq & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \|\overline{\tilde{U}}_t \mathbf{1}^T - \tilde{U}_t\|_{abs} \\
= & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \|\tilde{U}_t \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T - \tilde{U}_t Q Q^T\|_{abs} \\
= & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \| - \sum_{l=2}^{N} \tilde{U}_t q_l q_l^T\|_{abs} \, ,
\end{aligned}
$$

where the second inequality is due to Lemma 2, introduced Section A, and the fact that $U_t = \max(\tilde{U}_t, \epsilon)$ (element-wise max operator). Recall from update rule of U_t, by defining $\hat{V}_{-1} \triangleq \hat{V}_0$ and $U_0 \triangleq U_{1/2}$, we have for all $t \geq 0$, $\tilde{U}_{t+1} = (\tilde{U}_t - \hat{V}_{t-1} + \hat{V}_t)W$. Thus, we obtain

$$
\tilde{U}_t = \tilde{U}_0 W^t + \sum_{k=1}^{t} (-\hat{V}_{t-1-k} + \hat{V}_{t-k}) W^k = \tilde{U}_0 + \sum_{k=1}^{t} (-\hat{V}_{t-1-k} + \hat{V}_{t-k}) Q \Lambda^k Q^T \, .
$$

Then we further obtain when $l \neq 1$,

$$
\tilde{U}_t q_l = (\tilde{U}_0 + \sum_{k=1}^{t} (-\hat{V}_{t-1-k} + \hat{V}_{t-k}) Q \Lambda^k Q^T) q_l = \sum_{k=1}^{t} (-\hat{V}_{t-1-k} + \hat{V}_{t-k}) q_l \lambda_l^k \, ,
$$

where the last equality is due to the definition $\tilde{U}_0 \triangleq U_{1/2} = \epsilon \mathbf{1_d} \mathbf{1}_N^T = \sqrt{N} \epsilon \mathbf{1_d} \mathbf{1}_N^T$ (recall that $q_1 = \frac{1}{\sqrt{N}} \mathbf{1}_N^T$) and $q_i^T q_j = 0$ when $i \neq j$. Note that by definition of $\| \cdot \|_{abs}$, we have for all A, B, $\|A + B\|_{abs} \leq \|A\|_{abs} + \|B\|_{abs}$, then

$$
\begin{aligned}
D_6 \leq & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \| - \sum_{l=2}^{N} \tilde{U}_t q_l q_l^T\|_{abs} \\
= & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \| - \sum_{k=1}^{t} (-\hat{V}_{t-1-k} + \hat{V}_{t-k}) \sum_{l=2}^{N} q_l \lambda_l^k q_l^T\|_{abs} \\
\leq & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \sum_{k=1}^{t} \sum_{j=1}^{d} \| \sum_{l=2}^{N} q_l \lambda_l^k q_l^T \|_1 \| (-\hat{V}_{t-1-k} + \hat{V}_{t-k})^T e_j \|_1 \\
\leq & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \sum_{k=1}^{t} \sum_{j=1}^{d} \sqrt{N} \| \sum_{l=2}^{N} q_l \lambda_l^k q_l^T \|_2 \| (-\hat{V}_{t-1-k} + \hat{V}_{t-k})^T e_j \|_1 \\
\leq & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \sum_{k=1}^{t} \sum_{j=1}^{d} \| (-\hat{V}_{t-1-k} + \hat{V}_{t-k})^T e_j \|_1 \sqrt{N} \lambda^k \\
= & \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon^{1.5}} \sum_{k=1}^{t} \| (-\hat{V}_{t-1-k} + \hat{V}_{t-k}) \|_{abs} \sqrt{N} \lambda^k \\
= & \frac{G_\infty^2}{N} \frac{1}{2\epsilon^{1.5}} \sum_{o=0}^{T-1} \sum_{t=o+1}^{T} \| (-\hat{V}_{o-1} + \hat{V}_o) \|_{abs} \sqrt{N} \lambda^{t-o} \\
\leq & \frac{G_\infty^2}{\sqrt{N}} \frac{1}{2\epsilon^{1.5}} \sum_{o=0}^{T-1} \frac{\lambda}{1-\lambda} \| (-\hat{V}_{o-1} + \hat{V}_o) \|_{abs} \, ,
\end{aligned}
\tag{24}
$$

where $\lambda = \max(|\lambda_2|, |\lambda_N|)$. Combining (23) and (24), we have

$$
D_2 \leq \frac{G_\infty^2}{\sqrt{N}} \frac{1}{2\epsilon^{1.5}} \frac{\lambda}{1-\lambda} \mathbb{E} \left[\sum_{o=0}^{T-1} \| (-\hat{V}_{o-1} + \hat{V}_o) \|_{abs} \right] .
$$

Now we need to bound D_1, we have

$$
\begin{aligned}
D_1 =& \sum_{t=1}^{T} \mathbb{E}\left[\left\langle \nabla f(Z_t), \frac{1}{N}\sum_{i=1}^{N} m_{t-1,i} \odot (\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}) \right\rangle\right] \\
\leq& \sum_{t=1}^{T} \mathbb{E}\left[G_\infty^2 \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\left|\frac{1}{\sqrt{[u_{t-1,i}]_j}} - \frac{1}{\sqrt{[u_{t,i}]_j}}\right|\right] \\
=& \sum_{t=1}^{T} \mathbb{E}\left[G_\infty^2 \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\left|\left(\frac{1}{\sqrt{[u_{t-1,i}]_j}} - \frac{1}{\sqrt{[u_{t,i}]_j}}\right)\frac{\sqrt{[u_{t,i}]_j}+\sqrt{[u_{t-1,i}]_j}}{\sqrt{[u_{t,i}]_j}+\sqrt{[u_{t-1,i}]_j}}\right|\right] \\
\leq& \sum_{t=1}^{T} \mathbb{E}\left[G_\infty^2 \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\left|\frac{1}{2\epsilon^{1.5}}\left([u_{t-1,i}]_j - [u_{t,i}]_j\right)\right|\right] \\
\overset{(a)}{\leq}& \sum_{t=1}^{T} \mathbb{E}\left[G_\infty^2 \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\frac{1}{2\epsilon^{1.5}}\left|([\tilde{u}_{t-1,i}]_j - [\tilde{u}_{t,i}]_j)\right|\right] \\
=& G_\infty^2 \frac{1}{2\epsilon^{1.5}}\frac{1}{N}\mathbb{E}\left[\sum_{t=1}^{T}\|\tilde{U}_{t-1} - \tilde{U}_t\|_{abs}\right],
\end{aligned}
\tag{25}
$$

where (a) is due to $[\tilde{u}_{t-1,i}]_j = \max([u_{t-1,i}]_j, \epsilon)$ and the function $\max(\cdot, \epsilon)$ is 1-Lipschitz. In addition, by update rule of U_t, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\|\tilde{U}_{t-1} - \tilde{U}_t\|_{abs} \\
=& \sum_{t=1}^{T}\|\tilde{U}_{t-1} - (\tilde{U}_{t-1} - \hat{V}_{t-2} + \hat{V}_{t-1})W\|_{abs} \\
=& \sum_{t=1}^{T}\|\tilde{U}_{t-1}(QQ^T - Q\Lambda Q^T) + (-\hat{V}_{t-2} + \hat{V}_{t-1})W\|_{abs} \\
=& \sum_{t=1}^{T}\|\tilde{U}_{t-1}(\sum_{l=2}^{N} q_l(1-\lambda_l)q_l^T) + (-\hat{V}_{t-2} + \hat{V}_{t-1})W\|_{abs} \\
\leq& \sum_{t=1}^{T}\|\sum_{k=1}^{t-1}(-\hat{V}_{t-2-k} + \hat{V}_{t-1-k})\sum_{l=2}^{N} q_l\lambda_l^k(1-\lambda_l)q_l^T\|_{abs} + \sum_{t=1}^{T}\|(-\hat{V}_{t-2} + \hat{V}_{t-1})W\|_{abs} \\
\leq& \sum_{t=1}^{T}\left(\sum_{k=1}^{t-1}\| - \hat{V}_{t-2-k} + \hat{V}_{t-1-k}\|_{abs}\sqrt{N}\lambda^k\right) + \sum_{t=1}^{T}\|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs} \\
=& \sum_{t=1}^{T}\left(\sum_{o=1}^{t-1}\| - \hat{V}_{o-2} + \hat{V}_{o-1}\|_{abs}\sqrt{N}\lambda^{t-o}\right) + \sum_{t=1}^{T}\|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs} \\
=& \sum_{o=1}^{T-1}\sum_{t=o+1}^{T}\left(\| - \hat{V}_{o-2} + \hat{V}_{o-1}\|_{abs}\sqrt{N}\lambda^{t-o}\right) + \sum_{t=1}^{T}\|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs} \\
\leq& \sum_{o=1}^{T-1}\frac{\lambda}{1-\lambda}\left(\| - \hat{V}_{o-2} + \hat{V}_{o-1}\|_{abs}\sqrt{N}\right) + \sum_{t=1}^{T}\|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs} \\
\leq& \frac{1}{1-\lambda}\sum_{t=1}^{T}\|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs}\sqrt{N}.
\end{aligned}
\tag{26}
$$

Combining (25) and (26), we have

$$D_1 \leq G_\infty^2 \frac{1}{2\epsilon^{1.5}} \frac{1}{N} \mathbb{E}\left[\frac{1}{1-\lambda} \sum_{t=1}^{T} \|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs} \sqrt{N}\right]. \tag{27}$$

What remains is to bound $\sum_{t=1}^{T} \mathbb{E}\left[\|Z_{t+1} - Z_t\|^2\right]$. By update rule of Z_t, we have

$$\begin{aligned}
&\|Z_{t+1} - Z_t\|^2 \\
=&\left\|\alpha \frac{\beta_1}{1-\beta_1} \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot \left(\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}\right) - \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2 \\
\leq& 2\alpha^2 \left\|\frac{\beta_1}{1-\beta_1} \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot \left(\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}\right)\right\|^2 + 2\alpha^2 \left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2 \\
\leq& 2\alpha^2 \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{1}{\sqrt{\epsilon}} \left|\frac{1}{\sqrt{[u_{t-1,i}]_j}} - \frac{1}{\sqrt{[u_{t,i}]_j}}\right| + 2\alpha^2 \left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2 \\
\leq& 2\alpha^2 \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{1}{\sqrt{\epsilon}} \left|\frac{[u_{t,i}]_j - [u_{t-1,i}]_j}{2\epsilon^{1.5}}\right| + 2\alpha^2 \left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2 \\
\leq& 2\alpha^2 \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{1}{2\epsilon^2} |[\tilde{u}_{t,i}]_j - [\tilde{u}_{t-1,i}]_j| + 2\alpha^2 \left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2 \\
=& 2\alpha^2 \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_\infty^2 \frac{1}{N} \frac{1}{2\epsilon^2} \|\tilde{U}_t - \tilde{U}_{t-1}\|_{abs} + 2\alpha^2 \left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2,
\end{aligned} \tag{28}$$

where the last inequality is again due to the definition that $[\tilde{u}_{t,i}]_j = \max([u_{t,i}]_j, \epsilon)$ and the fact that $\max(\cdot, \epsilon)$ is 1-Lipschitz. Then, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\|Z_{t+1} - Z_t\|^2] \\
\leq& 2\alpha^2 \left(\frac{\beta_1}{1-\beta_1}\right)^2 G_\infty^2 \frac{1}{N} \frac{1}{2\epsilon^2} \mathbb{E}\left[\sum_{t=1}^{T} \|\tilde{U}_t - \tilde{U}_{t-1}\|_{abs}\right] + 2\alpha^2 \sum_{t=1}^{T} \mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right] \\
\leq& \alpha^2 \left(\frac{\beta_1}{1-\beta_1}\right)^2 \frac{G_\infty^2}{\sqrt{N}} \frac{1}{\epsilon^2} \frac{1}{1-\lambda} \mathbb{E}\left[\sum_{t=1}^{T} \|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{abs}\right] + 2\alpha^2 \sum_{t=1}^{T} \mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right],
\end{aligned}$$

where the last inequality is due to (26).

We now bound the last term on RHS of the above inequality. A trivial bound can be

$$\sum_{t=1}^{T} \left\|\frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2 \leq \sum_{t=1}^{T} d G_\infty^2 \frac{1}{\epsilon},$$

due to $\|g_{t,i}\| \leq G_\infty$ and $[u_{t,i}]_j \geq \epsilon$, for all j (verified from update rule of $u_{t,i}$ and the assumption that $[v_{t,i}]_j \geq \epsilon$, for all i). However, the above bound is independent of N, to get a better bound, we need a more involved analysis to show its dependency on N. To do this, we first notice that

$$\begin{aligned}
&\mathbb{E}_{G_t|G_{1:t-1}}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right] \\
=&\mathbb{E}_{G_t|G_{1:t-1}}\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\langle\frac{\nabla f_i(x_{t,i})+\xi_{t,i}}{\sqrt{u_{t,i}}},\frac{\nabla f_j(x_{t,j})+\xi_{t,j}}{\sqrt{u_{t,j}}}\right\rangle\right] \\
\overset{(a)}{=}&\mathbb{E}_{G_t|G_{1:t-1}}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2\right]+\mathbb{E}_{G_t|G_{1:t-1}}\left[\frac{1}{N^2}\sum_{i=1}^{N}\left\|\frac{\xi_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right] \\
\overset{(b)}{=}&\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{l=1}^{d}\frac{\mathbb{E}_{G_t|G_{1:t-1}}[[\xi_{t,i}]_l^2]}{[u_{t,i}]_l} \\
\overset{(c)}{\le}&\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2+\frac{d}{N}\frac{\sigma^2}{\epsilon}\,,
\end{aligned}$$

where (a) is due to $\mathbb{E}_{G_t|G_{1:t-1}}[\xi_{t,i}]=0$ and $\xi_{t,i}$ is independent of $x_{t,j}$, $u_{t,j}$ for all j, and ξ_j, for all $j\neq i$, (b) comes from the fact that $x_{t,i}$, $u_{t,i}$ are fixed given $G_{1:t}$, (c) is due to $\mathbb{E}_{G_t|G_{1:t-1}}[[\xi_{t,i}]_l^2\le\sigma^2$ and $[u_{t.i}]_l\ge\epsilon$ by definition. Then we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right]=&\mathbb{E}_{G_{1:t-1}}\left[\mathbb{E}_{G_t|G_{1:t-1}}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right]\right] \\
\le&\mathbb{E}_{G_{1:t-1}}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2+\frac{d}{N}\frac{\sigma^2}{\epsilon}\right] \\
=&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2\right]+\frac{d}{N}\frac{\sigma^2}{\epsilon}\,.
\end{aligned}\tag{29}$$

In traditional analysis of SGD-like distributed algorithms, the term corresponding to $\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2\right]$ will be merged with the first order descent when the stepsize is chosen to be small enough. However, in our case, the term cannot be merged because it is different from the first order descent in our algorithm. A brute-force upper bound is possible but this will lead to a worse convergence rate in terms of N. Thus, we need a more detailed analysis for the term in the following.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2\right] \\
=&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}+\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})\odot\left(\frac{1}{\sqrt{u_{t,i}}}-\frac{1}{\sqrt{\overline{U}_t}}\right)\right\|^2\right] \\
\le&2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right]+2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{t,i})\odot\left(\frac{1}{\sqrt{u_{t,i}}}-\frac{1}{\sqrt{\overline{U}_t}}\right)\right\|^2\right] \\
\le&2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right]+2\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(x_{t,i})\odot\left(\frac{1}{\sqrt{u_{t,i}}}-\frac{1}{\sqrt{\overline{U}_t}}\right)\right\|^2\right]
\end{aligned}$$

$$\leq 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right] + 2\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}G_\infty^2\frac{1}{\sqrt{\epsilon}}\left\|\frac{1}{\sqrt{u_{t,i}}} - \frac{1}{\sqrt{\overline{U}_t}}\right\|_1\right].$$

Summing over T, we have

$$\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}}\right\|^2\right]$$
$$\leq 2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right] + 2\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}G_\infty^2\frac{1}{\sqrt{\epsilon}}\left\|\frac{1}{\sqrt{u_{t,i}}} - \frac{1}{\sqrt{\overline{U}_t}}\right\|_1\right]. \tag{30}$$

For the last term on RHS of (30), we can bound it similarly as what we did for D_2 from (23) to (24), which yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}G_\infty^2\frac{1}{\sqrt{\epsilon}}\left\|\frac{1}{\sqrt{u_{t,i}}} - \frac{1}{\sqrt{\overline{U}_t}}\right\|_1\right] &\leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}G_\infty^2\frac{1}{\sqrt{\epsilon}}\frac{1}{2\epsilon^{1.5}}\left\|u_{t,i} - \overline{U}_t\right\|_1\right] \\
&= \sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{N}G_\infty^2\frac{1}{2\epsilon^2}\left\|\overline{U}_t\mathbf{1}^T - U_t\right\|_{abs}\right] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{N}G_\infty^2\frac{1}{2\epsilon^2}\| - \sum_{l=2}^{N}\tilde{U}_t q_l q_l^T\|_{abs}\right] \\
&\leq \frac{1}{\sqrt{N}}G_\infty^2\frac{1}{2\epsilon^2}\mathbb{E}\left[\sum_{o=0}^{T-1}\frac{\lambda}{1-\lambda}\|(-\hat{V}_{o-1} + \hat{V}_o)\|_{abs}\right].
\end{aligned} \tag{31}$$

Further, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right] \\
\leq& 2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(\overline{X}_t)}{\sqrt{\overline{U}_t}}\right\|^2\right] + 2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(\overline{X}_t) - \nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right] \\
=& 2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\sqrt{\overline{U}_t}}\right\|^2\right] + 2\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(\overline{X}_t) - \nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right]
\end{aligned}$$

and the last term on RHS of the above inequality can be bounded following similar procedures from (16) to (21), as what we did for D_3. Completing the procedures yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla f_i(\overline{X}_t) - \nabla f_i(x_{t,i})}{\sqrt{\overline{U}_t}}\right\|^2\right] &\leq \sum_{t=1}^{T}\mathbb{E}\left[L\frac{1}{\epsilon}\frac{1}{N}\sum_{i=1}^{N}\left\|x_{t,i} - \overline{X}_t\right\|^2\right] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[L\frac{1}{\epsilon}\frac{1}{N}\alpha^2\left(\frac{1}{1-\lambda}\right)NdG_\infty^2\frac{1}{\epsilon}\right] \\
&= TL\frac{1}{\epsilon^2}\alpha^2\left(\frac{1}{1-\lambda}\right)dG_\infty^2.
\end{aligned} \tag{32}$$

Finally, combining (29) to (32), we get

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{g_{t,i}}{\sqrt{u_{t,i}}}\right\|^2\right] \leq & 4\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\sqrt{\overline{U}_t}}\right\|^2\right] + 4TL\frac{1}{\epsilon^2}\alpha^2\left(\frac{1}{1-\lambda}\right)dG_\infty^2 \\ & + 2\frac{1}{\sqrt{N}}G_\infty^2\frac{1}{2\epsilon^2}\mathbb{E}\left[\sum_{o=0}^{T-1}\frac{\lambda}{1-\lambda}\|(-\hat{V}_{o-1}+\hat{V}_o)\|_{abs}\right] + T\frac{d}{N}\frac{\sigma^2}{\epsilon} \\ \leq & 4\frac{1}{\sqrt{\epsilon}}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] + 4TL\frac{1}{\epsilon^2}\alpha^2\left(\frac{1}{1-\lambda}\right)dG_\infty^2 \\ & + 2\frac{1}{\sqrt{N}}G_\infty^2\frac{1}{2\epsilon^2}\mathbb{E}\left[\sum_{o=0}^{T-1}\frac{\lambda}{1-\lambda}\|(-\hat{V}_{o-1}+\hat{V}_o)\|_{abs}\right] + T\frac{d}{N}\frac{\sigma^2}{\epsilon}.\end{aligned}$$

where the last inequality is due to each element of $\overline{U}_t$ is lower bounded by ϵ by definition.

Combining all above, we obtain

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] \\ \leq & \frac{2}{T\alpha}(\mathbb{E}[f(Z_1)]-\mathbb{E}[f(Z_{T+1})]) \\ & + \frac{L}{T}\alpha\left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{G_\infty^2}{\sqrt{N}}\frac{1}{\epsilon^2}\frac{1}{1-\lambda}\mathbb{E}\left[\mathcal{V}_T\right] \\ & + \frac{8L}{T}\alpha\frac{1}{\sqrt{\epsilon}}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] + 8L^2\alpha\frac{1}{\epsilon^2}\alpha^2\left(\frac{1}{1-\lambda}\right)dG_\infty^2 \\ & + \frac{4L}{T}\alpha\frac{1}{\sqrt{N}}G_\infty^2\frac{1}{2\epsilon^2}\mathbb{E}\left[\sum_{o=0}^{T-1}\frac{\lambda}{1-\lambda}\|(-\hat{V}_{o-1}+\hat{V}_o)\|_{abs}\right] + 2L\alpha\frac{d}{N}\frac{\sigma^2}{\epsilon} \\ & + \frac{2}{T}\frac{\beta_1}{1-\beta_1}G_\infty^2\frac{1}{2\epsilon^{1.5}}\frac{1}{\sqrt{N}}\mathbb{E}\left[\frac{1}{1-\lambda}\mathcal{V}_T\right] \\ & + \frac{2}{T}\frac{G_\infty^2}{\sqrt{N}}\frac{1}{2\epsilon^{1.5}}\frac{\lambda}{1-\lambda}\mathbb{E}\left[\mathcal{V}_T\right] \\ & + \frac{3}{T}\left(\sum_{t=1}^{T}L\left(\frac{1}{1-\lambda}\right)^2\alpha^2 dG_\infty^2\frac{1}{\epsilon^{1.5}} + \sum_{t=1}^{T}L\left(\frac{\beta_1}{1-\beta_1}\right)^2\alpha^2 d\frac{G_\infty^2}{\epsilon^{1.5}}\right) \\ = & \frac{2}{T\alpha}(\mathbb{E}[f(Z_1)]-\mathbb{E}[f(Z_{T+1})]) + 2L\alpha\frac{d}{N}\frac{\sigma^2}{\epsilon} + 8L\alpha\frac{1}{\sqrt{\epsilon}}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] \\ & + 3\alpha^2 d\left(\left(\frac{\beta_1}{1-\beta_1}\right)^2+\left(\frac{1}{1-\lambda}\right)^2\right)L\frac{G_\infty^2}{\epsilon^{1.5}} + 8\alpha^3L^2\left(\frac{1}{1-\lambda}\right)d\frac{G_\infty^2}{\epsilon^2} \\ & + \frac{1}{T\epsilon^{1.5}}\frac{G_\infty^2}{\sqrt{N}}\frac{1}{1-\lambda}\left(L\alpha\left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{1}{\epsilon^{0.5}} + \lambda + \frac{\beta_1}{1-\beta_1} + 2L\alpha\frac{1}{\epsilon^{0.5}}\lambda\right)\mathbb{E}\left[\mathcal{V}_T\right].\end{aligned} \tag{33}$$

where $\mathcal{V}_T := \sum_{t=1}^{T}\|(-\hat{V}_{t-2}+\hat{V}_{t-1})\|_{abs}$. Set $\alpha = \frac{1}{\sqrt{dT}}$ and when $\alpha \leq \frac{\epsilon^{0.5}}{16L}$, we further have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] \\ \leq & \frac{4}{T\alpha}(\mathbb{E}[f(Z_1)]-\mathbb{E}[f(Z_{T+1})]) + 4L\alpha\frac{d}{N}\frac{\sigma^2}{\epsilon}\end{aligned}$$

$$
\begin{aligned}
&+6\alpha^2 d\left(\left(\frac{\beta_1}{1-\beta_1}\right)^2+\left(\frac{1}{1-\lambda}\right)^2\right)L\frac{G_\infty^2}{\epsilon^{1.5}}+16\alpha^3L^2\left(\frac{1}{1-\lambda}\right)d\frac{G_\infty^2}{\epsilon^2}\\
&+\frac{2}{T\epsilon^{1.5}}\frac{G_\infty^2}{\sqrt{N}}\frac{1}{1-\lambda}\left(L\alpha\left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{1}{\epsilon^{0.5}}+\lambda+\frac{\beta_1}{1-\beta_1}+2L\alpha\frac{1}{\epsilon^{0.5}}\lambda\right)\mathbb{E}\left[\mathcal{V}_T\right]\\
=&\frac{4\sqrt{d}}{\sqrt{T}}(\mathbb{E}[f(Z_1)]-\mathbb{E}[f(Z_{T+1})])+4L\frac{\sqrt{d}}{\sqrt{T}}\frac{1}{N}\frac{\sigma^2}{\epsilon}\\
&+6\frac{1}{T}\left(\left(\frac{\beta_1}{1-\beta_1}\right)^2+\left(\frac{1}{1-\lambda}\right)^2\right)L\frac{G_\infty^2}{\epsilon^{1.5}}+16\frac{1}{T^{1.5}d^{0.5}}L^2\left(\frac{1}{1-\lambda}\right)\frac{G_\infty^2}{\epsilon^2}\\
&+\frac{2}{T\epsilon^{1.5}}\frac{G_\infty^2}{\sqrt{N}}\frac{1}{1-\lambda}\left(\frac{L}{\sqrt{Td}}\left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{1}{\epsilon^{0.5}}+\lambda+\frac{\beta_1}{1-\beta_1}+2\frac{L}{\sqrt{Td}}\frac{1}{\epsilon^{0.5}}\lambda\right)\mathbb{E}\left[\mathcal{V}_T\right]\\
\leq& C_1\frac{\sqrt{d}}{\sqrt{T}}\left(\mathbb{E}[f(Z_1)]-\min_z f(z)+\frac{\sigma^2}{N}\right)+\frac{1}{T}C_2+\frac{1}{T^{1.5}d^{0.5}}C_3\\
&+\left(\frac{1}{TN^{0.5}}C_4+\frac{1}{T^{1.5}d^{0.5}N^{0.5}}C_5\right)\mathbb{E}\left[\mathcal{V}_T\right],
\end{aligned}
$$

where the first inequality is obtained by moving the term $8L\alpha\frac{1}{\sqrt{\epsilon}}\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right]$ on the RHS of (33) to the LHS to cancel it using the assumption $8L\alpha\frac{1}{\sqrt{\epsilon}}\leq\frac{1}{2}$ followed by multiplying both sides by 2. The constants introduced in the last step are defined as following

$$
\begin{aligned}
C_1=&\max(4,4L/\epsilon),\\
C_2=&6\left(\left(\frac{\beta_1}{1-\beta_1}\right)^2+\left(\frac{1}{1-\lambda}\right)^2\right)L\frac{G_\infty^2}{\epsilon^{1.5}},\\
C_3=&16L^2\left(\frac{1}{1-\lambda}\right)\frac{G_\infty^2}{\epsilon^2},\\
C_4=&\frac{2}{\epsilon^{1.5}}\frac{1}{1-\lambda}\left(\lambda+\frac{\beta_1}{1-\beta_1}\right)G_\infty^2,\\
C_5=&\frac{2}{\epsilon^2}\frac{1}{1-\lambda}L\left(\frac{\beta_1}{1-\beta_1}\right)^2G_\infty^2+\frac{4}{\epsilon^2}\frac{\lambda}{1-\lambda}LG_\infty^2.
\end{aligned}
$$

Substituting into $Z_1=\overline{X}_1$ completes the proof. □

C Proof of Theorem 3

Under some assumptions stated in Theorem 2, we have that

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right]\leq& C_1\frac{\sqrt{d}}{\sqrt{T}}\left(\mathbb{E}[f(\overline{X}_1)]-\min_z f(z)]+\frac{\sigma^2}{N}\right)+\frac{1}{T}C_2+\frac{1}{T^{1.5}d^{0.5}}C_3\\
&+\left(\frac{1}{TN^{0.5}}C_4+\frac{1}{T^{1.5}d^{0.5}N^{0.5}}C_5\right)\mathbb{E}\left[\sum_{t=1}^T\|(-\hat{V}_{t-2}+\hat{V}_{t-1})\|_{abs}\right],
\end{aligned}
\tag{34}
$$

where $\|\cdot\|_{abs}$ denotes the entry-wise L_1 norm of a matrix (i.e $\|A\|_{abs}=\sum_{i,j}|A_{ij}|$) and C_1,C_2,C_3,C_4,C_5 are defined in Theorem 2.

Since Algorithm 3 is a special case of 2, building on result of Theorem 2, we just need to characterize the growth speed of $\mathbb{E}\left[\sum_{t=1}^T\|(-\hat{V}_{t-2}+\hat{V}_{t-1})\|_{abs}\right]$ to prove convergence of Algorithm 3. By the update rule of Algorithm 3, we know $\hat{V}_t$ is non decreasing and thus

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\|(-\hat{V}_{t-2}+\hat{V}_{t-1})\|_{abs}\right] =&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{d}|-[\hat{v}_{t-2,i}]_j+[\hat{v}_{t-1,i}]_j|\right]\\ =&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{d}(-[\hat{v}_{t-2,i}]_j+[\hat{v}_{t-1,i}]_j)\right]\\ =&\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{d}(-[\hat{v}_{-1,i}]_j+[\hat{v}_{T-1,i}]_j)\right]\\ =&\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{d}(-[\hat{v}_{0,i}]_j+[\hat{v}_{T-1,i}]_j)\right],\end{aligned}$$

where the last equality is because we defined $\hat{V}_{-1} \triangleq \hat{V}_0$ previously.

Further, because $\|g_{t,i}\|_\infty \le G_\infty$ for all t, i and $v_{t,i}$ is a exponential moving average of $g_{k,i}^2, k = 1, 2, \cdots, t$, we know $|[v_{t,i}]_j| \le G_\infty^2$, for all t, i, j. In addition, by update rule of $\hat{V}_t$, we also know each element of $\hat{V}_t$ also cannot be greater than G_∞^2, i.e. $|[\hat{v}_{t,i}]_j| \le G_\infty^2$, for all t, i, j. Given the fact that $[\hat{v}_{0,i}]_j \ge 0$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\|(-\hat{V}_{t-2}+\hat{V}_{t-1})\|_{abs}\right] = \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{d}(-[\hat{v}_{0,i}]_j+[\hat{v}_{T-1,i}]_j)\right] \le \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{d}G_\infty^2\right] = NdG_\infty^2\,.$$

Substituting the above into (34), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\frac{\nabla f(\overline{X}_t)}{\overline{U}_t^{1/4}}\right\|^2\right] \le& C_1\frac{\sqrt{d}}{\sqrt{T}}\left(\mathbb{E}[f(\overline{X}_1)]-\min_z f(z)+\frac{\sigma^2}{N}\right)+\frac{1}{T}C_2+\frac{1}{T^{1.5}d^{0.5}}C_3\\ &+\frac{d}{T}C_4\sqrt{N}G_\infty^2+\frac{\sqrt{d}}{T^{1.5}}C_5\sqrt{N}G_\infty^2\\ =&C_1'\frac{\sqrt{d}}{\sqrt{T}}\left(\mathbb{E}[f(\overline{X}_1)]-\min_z f(z)+\frac{\sigma^2}{N}\right)+\frac{1}{T}C_2'+\frac{1}{T^{1.5}d^{0.5}}C_3'\\ &+\frac{d}{T}\sqrt{N}C_4'+\frac{\sqrt{d}}{T^{1.5}}\sqrt{N}C_5',\end{aligned} \tag{35}$$

where we have

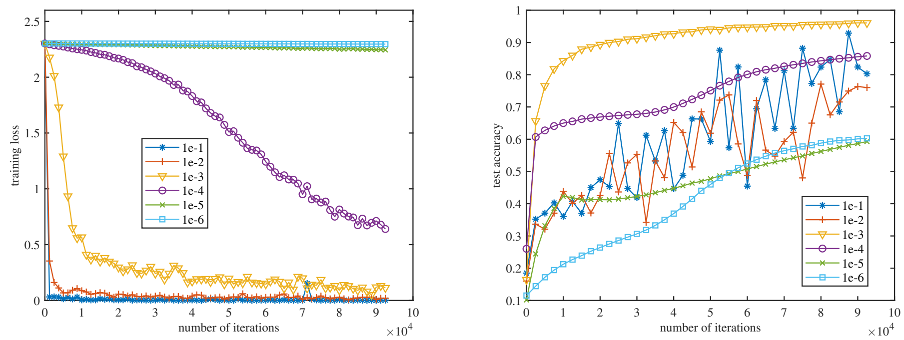
$$C_1' = C_1 \quad C_2' = C_2 \quad C_3' = C_3 \quad C_4' = C_4G_\infty^2 \quad C_5' = C_5G_\infty^2\,. \tag{36}$$

and we conclude the proof. □

D Additional Experiments and Details

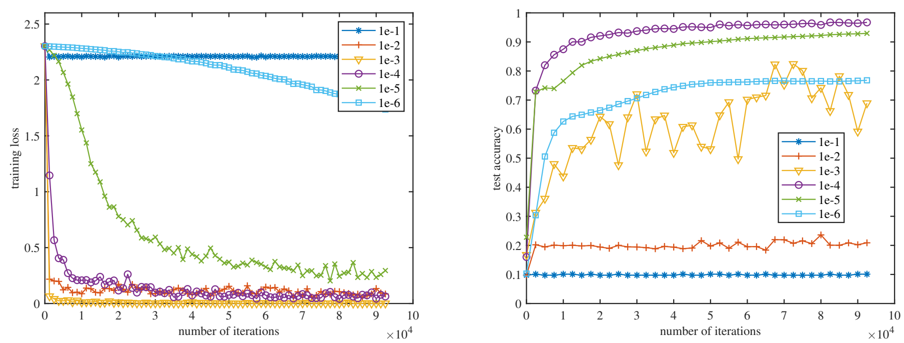
In this section, we compare the training loss and testing accuracy of different algorithms, namely Decentralized Stochastic Gradient Descent (DGD), Decentralized Adam (DADAM) and our proposed Decentralized AMSGrad, with different stepsizes on heterogeneous data distribution. We use 5 nodes and the heterogeneous data distribution is created by assigning each node with data of only two labels. Note that there are no overlapping labels between different nodes. For all algorithms, we compare stepsizes in the grid $[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$.

Figure 2 shows the training loss and test accuracy for DGD algorithm. We observe that the stepsize 10^{-3} works best for DGD in terms of test accuracy and 10^{-1} works best in terms of training loss. This difference is caused by the inconsistency among the value of parameters on different nodes when the stepsize is large. The training loss is calculated as the average of the loss value of different local models evaluated on their local training batch. Thus, while the training loss is small at a particular node, the test accuracy will be low when evaluating data with labels not seen by the node (recall that each node contains data with different labels since we are in the heterogeneous setting).



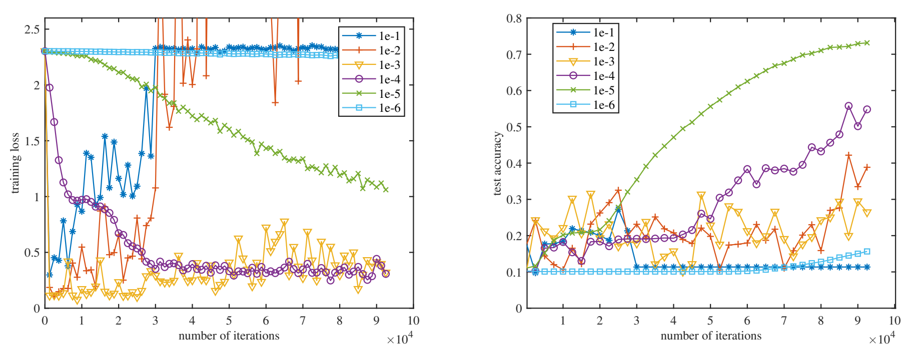
(a) Training loss (b) Test accuracy

Figure 2: Performance comparison of different stepsizes for DGD



(a) Training loss (b) Test accuracy

Figure 3: Performance comparison of different stepsizes for decentralized AMSGrad



(a) Training loss (b) Test accuracy

Figure 4: Performance comparison of different stepsizes for DADAM

Figure 3 shows the performance of decentralized AMSGrad with different stepsizes. We see that its best performance is better than the one of DGD and the performance is more stable (the test performance is less sensitive to stepsize tuning).

Figure 4 displays the performance of Decentralized Adam algorithm. As expected, the performance of DADAM is not as good as DGD or decentralized AMSGrad. Its divergence characteristic, highlighted Section 2.3, coupled with the heterogeneity in the data amplify its non-convergence issue in our experiments. From the experiments above, we can see the advantages of decentralized AMSGrad in terms of both performance and ease of parameter tuning, and the importance of ensuring the theoretical convergence of any newly proposed methods in the presented setting.