Appendix: On the Convergence of Decentralized Adaptive Gradient Methods

The main purpose of this appendix is to give thorough and detailed proofs for our convergence analysis described in the main paper. After having established several important Lemmas in Section A, we provide a proof for our main Theorem, namely Theorem 2, in Section B. Section C and Section D correspond to the proofs for the extension and application of Theorem 2 to the AMSGrad and AdaGrad algorithms used as prototypes of our general class of decentralized adaptive gradient methods. Section E contains additional numerical runs for more empirical insights on our scheme.

Appendix A. Proof of Auxiliary Lemmas

Similarly to Yan et al. (2018); Chen et al. (2019) with SGD (with momentum) and centralized adaptive gradient methods, define the following auxiliary sequence:

$$Z_t = X_t + \frac{\beta_1}{1 - \beta_1} (X_t - X_{t-1}),$$

with $X_0 \triangleq X_1$. Such an auxiliary sequence can help us deal with the bias brought by the momentum and simplifies the convergence analysis.

**Lemma 1** For the sequence defined in (4), we have

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

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

$$X_{t+1} = \frac{1}{N} \sum_{i=1}^{N} x_{t+1,i} = \frac{1}{N} \sum_{i=1}^{N} \left( x_{t+0.5,i} - \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}} \right) = \frac{1}{N} \sum_{i=1}^{N} \left( \sum_{j=1}^{N} W_{ij} x_{t,j} - \alpha \frac{m_{t,i}}{\sqrt{u_{t,i}}} \right),$$

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

$$Z_{t+1} - Z_t = \frac{1}{N} \sum_{i=1}^{N} x_{t+1,i} - \frac{1}{N} \sum_{i=1}^{N} x_{t,i} = \frac{1 - \beta_1}{1 - \beta_1} (X_{t+1} - X_t) - \frac{\beta_1}{1 - \beta_1} (X_{t+1} - X_t)$$

$$= \frac{1}{1 - \beta_1} \left( \frac{1}{N} \sum_{i=1}^{N} m_{t,i} \odot \left( \frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right) - \frac{\beta_1}{1 - \beta_1} \left( \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot \left( \frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right)$$

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

which is the desired result. $\square$
Lemma 2 Given a set of numbers \( a_1, \ldots, a_n \) 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 
\( b(r) = \frac{1}{n} \sum_{i=1}^{n} b_i(r) \). For any \( r \) and \( r' \) with \( r' \geq r \) we have
\[
\sum_{i=1}^{n} |b_i(r) - \bar{b}(r)| \geq \sum_{i=1}^{n} |b_i(r') - \bar{b}(r')| \tag{5}
\]
and when \( r \leq \min_{i \in [n]} a_i \), we have
\[
\sum_{i=1}^{n} |b_i(r) - \bar{b}(r)| = \sum_{i=1}^{n} |a_i - \bar{a}|. \tag{6}
\]

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 \( a_s(r) < b(r) \). 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
\[
h'(r) = l(r) \left( \frac{l(r)}{n} - 1 \right) + (s(r) - l(r)) \frac{l(r)}{n} - (n - s(r)) \frac{l(r)}{n}
\]
\[
= \frac{l(r)}{n} \left( (l(r) - n) + (s(r) - l(r)) - (n - s(r)) \right).
\]
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, (5) is proven. When \( r \leq \min_{i \in [n]} a_i \), 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 (6). \( \square \)
Appendix B. Proof of Theorem 2

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

$$Z_t = \bar{X}_t + \frac{\beta_1}{1 - \beta_1} (X_t - \bar{X}_{t-1}),$$

with $\bar{X}_0 \triangleq \bar{X}_1$. 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 $G_1, G_2, \ldots, G_{t-1}$), we have

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

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

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

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

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

$$\alpha \mathbb{E} \left[ \langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \rangle \right] \leq \mathbb{E}[f(Z_t)] - \mathbb{E}[f(Z_{t+1})] + \frac{L}{2} \mathbb{E} \left[ \|Z_{t+1} - Z_t\|^2 \right]$$

$$+ \alpha \frac{\beta_1}{1 - \beta_1} \mathbb{E} \left[ \langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot (\frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}}) \rangle \right].$$

In addition, we have

$$\left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\rangle$$

$$= \left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\rangle + \left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \odot \left( \frac{1}{\sqrt{u_{t,i}}} - \frac{1}{\sqrt{U_t}} \right) \right\rangle.$$
and the first term on RHS of the equality can be lower bounded as
\[
\left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\rangle \\
= \frac{1}{2} \left\| \frac{\nabla f(Z_t)}{U_t^{1/4}} \right\|^2 + \frac{1}{2} \left\| \frac{\nabla f(Z_t) - \nabla f(X_t)}{U_t^{1/4}} \right\|^2 - \frac{1}{2} \left\| \frac{\nabla f(Z_t) - \nabla f(X_t)}{U_t^{1/4}} \right\|^2
\]

where the inequalities are all due to Cauchy-Schwartz. Substituting (10) and (9) into (8), we obtain
\[
\frac{1}{2} \alpha \mathbb{E} \left[ \left\| \frac{\nabla f(X_t)}{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{\mu_{t-1,i}}} - \frac{1}{\sqrt{\mu_{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{\mu_{t,i}}} - \frac{1}{\sqrt{U_t}} \right) \right\rangle \right]

+ 3 \alpha \mathbb{E} \left[ \left\| \nabla f(Z_t) - \nabla f(X_t) \right\|^2 \right],
\]
Decentralized Adaptive Gradient Methods

Summing over the above inequality from \( t = 1 \) to \( T \) and dividing both sides by \( T \alpha / 2 \), yields

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(\mathbf{X}_t) / U_t^{1/4} \right\|^2 \right] \leq \frac{2}{T \alpha} (\mathbb{E}[f(Z_1)] - \mathbb{E}[f(Z_{T+1})]) + \frac{L}{T \alpha} \sum_{t=1}^{T} \mathbb{E} \left[ \|Z_{t+1} - Z_t\|^2 \right]
\]

\[
+ \frac{2}{T (1 - \beta_1)} \sum_{t=1}^{T} \mathbb{E} \left[ \left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} m_{t-1,i} \odot \left( \frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right\rangle \right]
\]

\[
+ \frac{2}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \odot \left( \frac{1}{\sqrt{U_t}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right\rangle \right]
\]

\[
+ \frac{3}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) - \nabla f(\mathbf{X}_t) \right\|^2 \left\| \nabla f(Z_t) - \nabla f(\mathbf{X}_t) \right\|^2 \right]. \tag{11}
\]

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 (11), we can upper bound them by

\[
\left\| \frac{\nabla f(Z_t) - \nabla f(\mathbf{X}_t)}{U_t^{1/4}} \right\|^2 \leq \frac{1}{\min_j [d] [U_t^{1/2}]} \left\| \nabla f(Z_t) - \nabla f(\mathbf{X}_t) \right\|^2
\]

\[
\leq \frac{L}{\min_j [d] [U_t^{1/2}]} \left\| Z_t - \mathbf{X}_t \right\|^2 \tag{12}
\]

and

\[
\left\| \frac{\sum_{i=1}^{N} \nabla f_i(x_{t,i}) - \nabla f(\mathbf{X}_t)}{U_t^{1/4}} \right\|^2 \leq \frac{1}{\min_j [d] [U_t^{1/2}]} \frac{1}{N} \sum_{i=1}^{N} \left\| \nabla f_i(x_{t,i}) - \nabla f(\mathbf{X}_t) \right\|^2
\]

\[
\leq L \frac{1}{\min_j [d] [U_t^{1/2}]} \frac{1}{N} \sum_{i=1}^{N} \left\| x_{t,i} - \mathbf{X}_t \right\|^2, \tag{13}
\]

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

\[
X_t = X_{t-1} W - \alpha \frac{M_{t-1}}{\sqrt{U_{t-1}}} = X_1 W^{t-1} - \alpha \sum_{k=0}^{t-2} M_{t-k-1} / \sqrt{U_{t-k-1}} W^k,
\]

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 an orthonormal matrix and \( \Lambda \) is a diagonal matrix whose diagonal elements correspond to eigenvalues of \( W \) in an descending order, i.e. \( \Lambda_{ii} = \lambda_i \) with \( \lambda_i \) being \( i \)th largest eigenvalue of
$W$. In addition, because $W$ is a doubly stochastic matrix, we know $\lambda_1 = 1$ and $q_1 = \frac{1}{\sqrt{N}}$. With eigen-decomposition of $W$, we can rewrite $D_5$ as

$$\sum_{i=1}^{N} \| x_{t,i} - \bar{X}_t \|^2 = \| X_t - \bar{X}_t 1^T_N \|^2 = \| X_t Q Q^T - X_t \frac{1}{N} 1_N 1^T_N \|^2 = \sum_{l=2}^{N} \| X_t q_l \|^2. \quad (15)$$

In addition, we can rewrite (14) as

$$X_t = X_1 W^{t-1} - \alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{U_{t-k-1}} W^k = X_1 - \alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{U_{t-k-1}} Q \Lambda^k Q^T, \quad (16)$$

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 $l > 1$,

$$X_t q_l = (X_1 - \alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{U_{t-k-1}} Q \Lambda^k Q^T) q_l = -\alpha \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{U_{t-k-1}} q_l \lambda^k, \quad (17)$$

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

Combining (15) and (17), we have

$$D_5 = \sum_{i=1}^{N} \| x_{t,i} - \bar{X}_t \|^2 = \sum_{l=2}^{N} \| X_t q_l \|^2 = \sum_{l=2}^{N} \alpha^2 \left\| \sum_{k=0}^{t-2} \frac{M_{t-k-1}}{U_{t-k-1}} \lambda^k q_l \right\|^2 \leq \alpha^2 \left( \frac{1}{1 - \lambda} \right)^2 N \lambda G^2 \frac{1}{\epsilon}, \quad (18)$$

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

$$\| Z_t - \bar{X}_t \|^2 = \left\| \frac{\beta_1}{1 - \beta_1} (X_t - X_{t-1}) \right\|^2 = \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \alpha^2 \left\| \sum_{i=1}^{N} \frac{m_{t-1,i}}{U_{t-1,i}} \right\|^2 \leq \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \alpha^2 d G^2 \frac{1}{\epsilon}. \quad (19)$$

Now we know both $D_4$ and $D_5$ are in the order of $O(\alpha^2)$ and thus $D_3$ is in the order of $O(\alpha^2)$. 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

\[ D_2 = \sum_{t=1}^{T} \mathbb{E} \left[ \langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \odot \left( \frac{1}{\sqrt{U_t}} - \frac{1}{\sqrt{u_{t,i}}} \right) \rangle \right] \]

\[ \leq \sum_{t=1}^{T} \mathbb{E} \left[ G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{1}{\sqrt{|U_t|_j}} - \frac{1}{\sqrt{|u_{t,i}|_j}} \right] \]

\[ = \sum_{t=1}^{T} \mathbb{E} \left[ G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left( \frac{[U_t]_j - [u_{t,i}]_j}{[U_t]_j \sqrt{|u_{t,i}|_j} + [U_t]_j [u_{t,i}]_j} \right) \right] \]

\[ \leq \sum_{t=1}^{T} \mathbb{E} \left[ \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{|U_t|_j - |u_{t,i}|_j}{2\epsilon 1.5} \right] \]

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

\[ D_6 \leq \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon 1.5} \| \tilde{U}_t 1^T - U_t \|_{abs} \leq \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon 1.5} \| \tilde{U}_t 1^T - U_t \|_{abs} \]

\[ = \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon 1.5} \| \tilde{U}_t \frac{1}{N} 1_N 1_N^T - \tilde{U}_t \tilde{Q} \tilde{Q}^T \|_{abs} \]

\[ = \frac{G_\infty^2}{N} \sum_{t=1}^{T} \frac{1}{2\epsilon 1.5} \| - \frac{N}{2} \tilde{U}_t q_t q_t^T \|_{abs} , \]

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 \( \tilde{V}_{t-1} \triangleq \tilde{V}_0 \) and \( \tilde{U}_0 \triangleq U_1/2 \), we have for all \( t \geq 0 \), \( \tilde{U}_{t+1} = (\tilde{U}_t - \tilde{V}_{t-1} + \tilde{V}_t) W \). Thus, we obtain

\[ \tilde{U}_t = \tilde{U}_0 W^t + \sum_{k=1}^{t} (-\tilde{V}_{t-1-k} + \tilde{V}_{t-k}) W^k = \tilde{U}_0 + \sum_{k=1}^{t} (-\tilde{V}_{t-1-k} + \tilde{V}_{t-k}) Q A^k Q^T. \]

Then we further obtain when \( l \neq 1 \),

\[ \tilde{U}_t q_l = (\tilde{U}_0 + \sum_{k=1}^{t} (-\tilde{V}_{t-1-k} + \tilde{V}_{t-k}) Q A^k Q^T) q_l = \sum_{k=1}^{t} (-\tilde{V}_{t-1-k} + \tilde{V}_{t-k}) q_l A^k \]

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

\[
D_6 \leq \frac{G_2^2}{N} \sum_{t=1}^{T} \frac{1}{2e^{1.5}} \| - \sum_{l=2}^{N} \hat{U}_t q_l q_i^T \|_{abs}
= \frac{G_2^2}{N} \sum_{t=1}^{T} \frac{1}{2e^{1.5}} \| - \sum_{k=1}^{t} (\hat{V}_{t-1-k} + \hat{V}_{t-k}) \sum_{l=2}^{N} q_l \lambda_l^k q_i^T \|_{abs}
\leq \frac{G_2^2}{N} \sum_{t=1}^{T} \frac{1}{2e^{1.5}} \sum_{k=1}^{t} \sum_{j=1}^{d} \| \sum_{l=2}^{N} q_l \lambda_l^k q_i^T \|_1 \| (\hat{V}_{t-1-k} + \hat{V}_{t-k}) e_j \|_1
\leq \frac{G_2^2}{N} \sum_{t=1}^{T} \frac{1}{2e^{1.5}} \sum_{k=1}^{t} \sum_{j=1}^{d} \sqrt{N} \| \sum_{l=2}^{N} q_l \lambda_l^k q_i^T \|_2 \| (\hat{V}_{t-1-k} + \hat{V}_{t-k}) e_j \|_1
\leq \frac{G_2^2}{N} \sum_{t=1}^{T} \frac{1}{2e^{1.5}} \sum_{k=1}^{t} \| (\hat{V}_{t-1-k} + \hat{V}_{t-k}) \|_{abs} \sqrt{N} \lambda_k
\]

\[
= \frac{G_2^2}{N} \sum_{t=1}^{T} \frac{1}{2e^{1.5}} \sum_{k=1}^{t} \| (\hat{V}_{t-1-k} + \hat{V}_{t-k}) \|_{abs} \sqrt{N} \lambda_k
\]

\[
= \frac{G_2^2}{N} \frac{1}{2e^{1.5}} \sum_{o=0}^{T-1} \sum_{t=o+1}^{T} \| (\hat{V}_{o-1} + \hat{V}_{o}) \|_{abs} \sqrt{N} \lambda^{t-o}
\leq \frac{G_2^2}{\sqrt{N}} \frac{1}{2e^{1.5}} \sum_{o=0}^{T-1} \frac{\lambda}{1 - \lambda} \| (\hat{V}_{o-1} + \hat{V}_{o}) \|_{abs}, \quad (21)
\]

where \( \lambda = \max(|\lambda_2|, |\lambda_N|) \). Combining (20) and (21), we have

\[
D_2 \leq \frac{G_2^2}{\sqrt{N}} \frac{1}{2e^{1.5}} \frac{\lambda}{1 - \lambda} \left[ \sum_{o=0}^{T-1} \| (\hat{V}_{o-1} + \hat{V}_{o}) \|_{abs} \right].
\]
Now we need to bound $D_1$, we have

$$D_1 = \sum_{t=1}^{T} \mathbb{E} \left[ \left\langle \nabla f(Z_t), \frac{1}{N} \sum_{i=1}^{N} s_{t-1,i} \odot \left( \frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right\rangle \right]$$

$$\leq \sum_{t=1}^{T} \mathbb{E} \left[ G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left( \frac{1}{\sqrt{|u_{t-1,i}|}} - \frac{1}{\sqrt{|u_{t,i}|}} \right) \sqrt{|u_{t,i}|} + \sqrt{|u_{t-1,i}|} \right]$$

$$= \sum_{t=1}^{T} \mathbb{E} \left[ G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left( \frac{1}{2 e^{1.5}} \left( |u_{t-1,i,j} - u_{t,i,j}| \right) \right) \right]$$

$$\leq \sum_{t=1}^{T} \mathbb{E} \left[ G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{1}{2 e^{1.5}} \left( |\tilde{u}_{t-1,i,j} - \tilde{u}_{t,i,j}| \right) \right]$$

$$= G_\infty^2 \frac{1}{2 e^{1.5}} \frac{1}{N} \mathbb{E} \left[ \sum_{t=1}^{T} \left\| \tilde{U}_{t-1} - \tilde{U}_t \right\|_{\text{abs}} \right], \quad (22)$$
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

$$
\sum_{t=1}^{T} \|\tilde{U}_{t-1} - \tilde{U}_t\|_{abs} = \sum_{t=1}^{T} \|\tilde{U}_{t-1} - (\tilde{U}_{t-1} - \tilde{V}_{t-2} + \tilde{V}_{t-1})W\|_{abs}
$$

$$
= \sum_{t=1}^{T} \|\tilde{U}_{t-1}(QQ^T - QAQ^T) + (-\tilde{V}_{t-2} + \tilde{V}_{t-1})W\|_{abs}
$$

$$
= \sum_{t=1}^{T} \|\tilde{U}_{t-1}(\sum_{l=2}^{N} q_l(1 - \lambda_t)q_l^T) + (-\tilde{V}_{t-2} + \tilde{V}_{t-1})W\|_{abs}
$$

$$
\leq \sum_{t=1}^{T} \|\sum_{k=1}^{t-1} (-\tilde{V}_{t-2-k} + \tilde{V}_{t-1-k})\sum_{l=2}^{N} q_l \lambda^k_l (1 - \lambda_l)q_l^T\|_{abs} + \sum_{t=1}^{T} \|(-\tilde{V}_{t-2} + \tilde{V}_{t-1})W\|_{abs}
$$

$$
\leq \sum_{t=1}^{T} \left\{ \sum_{k=1}^{t-1} \| - \tilde{V}_{t-2-k} + \tilde{V}_{t-1-k}\|_{abs} \sqrt{N} \lambda^k_l \right\} + \sum_{t=1}^{T} \|(-\tilde{V}_{t-2} + \tilde{V}_{t-1})\|_{abs}
$$

$$
= \sum_{t=1}^{T} \left\{ \sum_{o=1}^{t-1} \| - \tilde{V}_{o-2} + \tilde{V}_{o-1}\|_{abs} \sqrt{N} \lambda^{t-o} \right\} + \sum_{t=1}^{T} \|(-\tilde{V}_{t-2} + \tilde{V}_{t-1})\|_{abs}
$$

$$
\leq \sum_{o=1}^{T} \frac{1}{1 - \lambda} \left\{ \| - \tilde{V}_{o-2} + \tilde{V}_{o-1}\|_{abs} \sqrt{N} \right\} + \sum_{t=1}^{T} \|(-\tilde{V}_{t-2} + \tilde{V}_{t-1})\|_{abs}
$$

$$
\leq \frac{1}{1 - \lambda} \sum_{t=1}^{T} \|(-\tilde{V}_{t-2} + \tilde{V}_{t-1})\|_{abs} \sqrt{N}.
$$

(23)

Combining (22) and (23), we have

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

(24)
What remains is to bound \( \sum_{t=1}^{T} \mathbb{E} [\|Z_{t+1} - Z_{t}\|^2] \). By update rule of \( Z_t \), we have

\[
\|Z_{t+1} - Z_t\|^2 = \left\| \frac{\beta_1}{1 - \beta_1} N \sum_{i=1}^{N} m_{t-1,i} \odot \left( \frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}} \right) - \frac{\alpha}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \\
\leq 2\alpha^2 \left\| \frac{\beta_1}{1 - \beta_1} N \sum_{i=1}^{N} m_{t-1,i} \odot \left( \frac{1}{\sqrt{u_{t-1,i}}} - \frac{1}{\sqrt{u_{t,i}}} \right) \right\|^2 + 2\alpha^2 \left\| \frac{\alpha}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \\
\leq 2\alpha^2 \left( \frac{\beta_1}{1 - \beta_1} \right)^2 G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{1}{\sqrt{\epsilon}} \left| \frac{1}{\sqrt{u_{t-1,i,j}}} - \frac{1}{\sqrt{u_{t,i,j}}} \right| + 2\alpha^2 \left\| \frac{\alpha}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \\
\leq 2\alpha^2 \left( \frac{\beta_1}{1 - \beta_1} \right)^2 G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \left[ |\tilde{u}_{t,i,j} - [u_{t-1,i,j}| + 2\alpha^2 \left\| \frac{\alpha}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \\
= 2\alpha^2 \left( \frac{\beta_1}{1 - \beta_1} \right)^2 G_\infty^2 \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2\epsilon^2} \|\tilde{U}_t - \hat{U}_{t-1}\|_{\text{abs}} + 2\alpha^2 \left\| \frac{\alpha}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2, \quad (25)\]

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

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

where the last inequality is due to (23).

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

\[
\sum_{t=1}^{T} \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \leq \sum_{t=1}^{T} dG_\infty^2 \frac{1}{\epsilon},
\]

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 \( [u_{t,i}] \geq \epsilon \), 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

\[
\mathbb{E}_{G_t|G_{1:t-1}} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} g_{t,i} \right\|^2 \right] = \mathbb{E}_{G_t|G_{1:t-1}} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{\nabla f_i(x_{t,i}) + \xi_{t,i}}{\sqrt{u_{t,i}}} \right) \frac{\nabla f_j(x_{t,j}) + \xi_{t,j}}{\sqrt{u_{t,j}}} \right\|^2 \right]
\]

\[
 \overset{(a)}{=} \mathbb{E}_{G_t|G_{1:t-1}} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 \right] + \mathbb{E}_{G_t|G_{1:t-1}} \left[ \frac{1}{N^2} \sum_{i=1}^{N} \left\| \frac{\xi_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \right]
\]

\[
 \overset{(b)}{=} \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{l=1}^{d} \mathbb{E}_{G_t|G_{1:t-1}} \left[ \xi_{t,i}^2 \right]_{[u_{t,i}]l}
\]

\[
 \overset{(c)}{<} \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 + \frac{d \sigma^2}{N \epsilon},
\]

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 $\xi_j$, 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 $\mathbb{E}_{G_t|G_{1:t-1}}[\xi_{t,i}^2] \leq \sigma^2$ and $[u_{t,i}]_l \geq \epsilon$ by definition. Then we have

\[
\mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \right] = \mathbb{E}_{G_{1:t-1}} \left[ \mathbb{E}_{G_t|G_{1:t-1}} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \right] \right]
\]

\[
 \leq \mathbb{E}_{G_{1:t-1}} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 + \frac{d \sigma^2}{N \epsilon} \right]
\]

\[
= \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}} \right\|^2 \right] + \frac{d \sigma^2}{N \epsilon}.
\]

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.
For the last term on RHS of (27), we can bound it similarly as what we did for $D_2$ from (20) to (21), which yields

$$\sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{\nabla f_i(x_{t,i})}{\sqrt{u_{t,i}}} \right\|^2 \right]$$

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

Summing over $T$, we have

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

For the last term on RHS of (27), we can bound it similarly as what we did for $D_2$ from (20) to (21), which yields

$$\sum_{t=1}^{T} \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^{N} G_{\infty}^2 \frac{1}{\sqrt{\epsilon}} \left\| \frac{1}{\sqrt{u_{t,i}}} - \frac{1}{\sqrt{U_t}} \right\|_1 \right] \leq \sum_{t=1}^{T} \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^{N} G_{\infty}^2 \frac{1}{\sqrt{\epsilon}} \right\| u_{t,i} - U_t \|_1 \right] = \sum_{t=1}^{T} \mathbb{E} \left[ \frac{1}{N} G_{\infty}^2 \frac{1}{\sqrt{\epsilon}} \right\| U_t^T - U_t \|_{abs} \right]$$

$$\leq \sum_{t=1}^{T} \mathbb{E} \left[ \frac{1}{N} G_{\infty}^2 \frac{1}{\sqrt{\epsilon}} \right\| \hat{U}_t q T \|_{abs} \right] + \sum_{t=2}^{T} \mathbb{E} \left[ \frac{1}{N} G_{\infty}^2 \frac{1}{\sqrt{\epsilon}} \right\| U_t^T - U_t \|_{abs} \right].$$

(28)
Further, we have
\[
\sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 \right] \\
\leq 2 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 \right] + 2 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) - \nabla f_i(x_{t,i}) \right\|^2 \right] \\
= 2 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(x_{t,i}) \right\|^2 \right] + 2 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) - \nabla f_i(x_{t,i}) \right\|^2 \right]
\]
and the last term on RHS of the above inequality can be bounded following similar procedures from (13) to (18), as what we did for \( D_3 \). Completing the procedures yields
\[
\sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{t,i}) \right\|^2 \right] \leq \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(x_{t,i}) \right\|^2 \right] + 2T L \frac{1}{\epsilon} \alpha^2 \left( \frac{1}{1 - \lambda} \right) dG_\infty^2 \frac{1}{\epsilon} \| \nabla f(x_{t,i}) \|_{\infty}^2
\]
(30)

Finally, combining (26) to (30), we get
\[
\sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{g_{t,i}}{\sqrt{u_{t,i}}} \right\|^2 \right] \leq 4 \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(x_{t,i}) \right\|^2 \right] + 4TL \frac{1}{\epsilon} \alpha^2 \left( \frac{1}{1 - \lambda} \right) dG_\infty^2 \\
+ 2 \frac{1}{\sqrt{N}} G_\infty^2 \frac{1}{2\epsilon} \mathbb{E} \left[ \sum_{o=0}^{T-1} \frac{\lambda}{1 - \lambda} \| (\hat{V}_{o-1} + \hat{V}_o) \|_{abs} \right] + T \frac{d}{N} \frac{\sigma^2}{\epsilon}
\]
Combining all above, we obtain

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(X_t) \right\| U_t^{-1/4} \right]^2 \right]
\leq \frac{2}{T \alpha} (E[f(Z_1)] - E[f(Z_{T+1})])
+ \frac{L}{T \alpha} \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \frac{\sqrt{T}}{\epsilon^2} \frac{1}{1 - \lambda} \mathbb{E} [\mathcal{V}_T]
+ \frac{8L}{T \alpha} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(X_t) \right\| U_t^{-1/2} \right]^2
+ 8L^2 \frac{1}{\epsilon^2} \alpha^2 \left( \frac{1}{1 - \lambda} \right) dG_\infty^2
+ \frac{4L}{T \alpha} \frac{1}{\sqrt{T}} \frac{\sqrt{T}}{\epsilon} \mathbb{E} \left[ \sum_{o=0}^{T-1} \frac{\lambda}{1 - \lambda} \|(-\hat{V}_{o-1} + \hat{V}_o)\|_{\text{abs}} \right]
+ \frac{2}{T} \frac{\beta_1}{1 - \beta_1} G_\infty^2 \frac{1}{\epsilon} \frac{1}{\sqrt{T}} \mathbb{E} \left[ \frac{1}{1 - \lambda} \mathcal{V}_T \right]
+ \frac{2}{T \sqrt{T}} \frac{G_\infty^2}{2 \epsilon^{1.5}} \frac{1}{1 - \lambda} \mathbb{E} [\mathcal{V}_T]
+ \frac{3}{T} \left( \sum_{t=1}^{T} L \left( \frac{1}{1 - \lambda} \right)^2 \alpha^2 dG_\infty^2 \frac{1}{\epsilon^{1.5}} + \sum_{t=1}^{T} L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \alpha^2 \frac{dG_\infty^2}{\epsilon^{1.5}} \right)
\leq \frac{2}{T \alpha} (E[f(Z_1)] - E[f(Z_{T+1})])
+ \frac{L}{T \alpha} \frac{d \sigma^2}{N \epsilon} + 8L \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(X_t) \right\| U_t^{-1/2} \right]^2
+ 3\alpha^2 \frac{d}{\epsilon} \left( \left( \frac{\beta_1}{1 - \beta_1} \right)^2 + \left( \frac{1}{1 - \lambda} \right)^2 \right) \frac{L G_\infty^2}{\epsilon^{1.5}} + 8\alpha^3 L^2 \left( \frac{1}{1 - \lambda} \right) \frac{dG_\infty^2}{\epsilon^2}
+ \frac{1}{T \epsilon^{1.5}} \frac{G_\infty^2}{\sqrt{N}} \frac{1}{1 - \lambda} \left( L \alpha \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \frac{1}{\epsilon^{0.5}} + \lambda + \frac{\beta_1}{1 - \beta_1} + 2L \alpha \frac{1}{\epsilon^{0.5} \lambda} \right) \mathbb{E} [\mathcal{V}_T].
\]
\[ V := \sum_{t=1}^{T} \|(\hat{V}_{t-2} + \hat{V}_{t-1})\|_{\text{abs}}. \] Set \( \alpha = \frac{1}{\sqrt{T}} \) and when \( \alpha \leq \frac{e^5}{16T} \), we further have

\[
\begin{align*}
\frac{2}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(\bar{X}_t) \right\|^{2} \right] \\
\leq \frac{4}{T^\alpha} (\mathbb{E}[f(Z_1)] - \mathbb{E}[f(Z_{T+1})]) + 4L\alpha \frac{\sigma d}{N^\epsilon} + 16L\alpha \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(\bar{X}_t) \right\|^{2} \right] \\
+ 6\alpha^2 d \left( \left( \frac{\beta_1}{1 - \beta_1} \right)^2 + \left( \frac{1}{1 - \lambda} \right)^2 \right) L \frac{G_{\infty}^2}{\epsilon^{1.5}} + 16\alpha^3 L^2 \left( \frac{1}{1 - \lambda} \right) \frac{G_{\infty}^2}{\epsilon^2} \\
+ \frac{2}{T e^{1.5}} \frac{G_{\infty}^2}{\sqrt{N}} \frac{1}{1 - \lambda} \left( L \alpha \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \frac{1}{\epsilon^{0.5}} + \frac{\beta_1}{1 - \beta_1} + 2L\alpha \frac{1}{\epsilon^{0.5}} \lambda \right) \mathbb{E} [V_T] \\
\leq \frac{4}{T^\alpha} (\mathbb{E}[f(Z_1)] - \min_x f(x)) + 4L\alpha \frac{\sigma d}{N^\epsilon} + 8L\alpha \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(\bar{X}_t) \right\|^{2} \right] \\
+ 6\alpha^2 d \left( \left( \frac{\beta_1}{1 - \beta_1} \right)^2 + \left( \frac{1}{1 - \lambda} \right)^2 \right) L \frac{G_{\infty}^2}{\epsilon^{1.5}} + 16\alpha^3 dL^2 \left( \frac{1}{1 - \lambda} \right) \frac{G_{\infty}^2}{\epsilon^2} \\
+ \frac{2}{T e^{1.5}} \frac{G_{\infty}^2}{\sqrt{N}} \frac{1}{1 - \lambda} \left( L \alpha \left( \frac{\beta_1}{1 - \beta_1} \right)^2 \frac{1}{\epsilon^{0.5}} + \frac{\beta_1}{1 - \beta_1} + 2L\alpha \frac{1}{\epsilon^{0.5}} \lambda \right) \mathbb{E} [V_T] \\
\leq C_1 \left( \frac{1}{T^\alpha} (\mathbb{E}[f(Z_1)] - \min_x f(x)) + \alpha \frac{d\sigma^2}{N^\epsilon} \right) + C_2\alpha^2 d + C_3\alpha^3 d + \frac{1}{TV} (C_4 + C_5 \alpha) \mathbb{E} [V_T] \\
+ 16L\alpha \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(\bar{X}_t) \right\|^{2} \right].
\end{align*}
\]

(32)

where the first inequality is obtained by moving the term \( 8L\alpha \frac{1}{\sqrt{T}} T \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(\bar{X}_t) \right\|^{2} \right] \) on the RHS of (31) to the LHS to cancel it using the assumption \( 8L\alpha \frac{1}{\sqrt{T}} \leq \frac{1}{2} \) followed by multiplying both sides by \( 2 \). The constants introduced in the last step are defined as following

\[
C_1 = \max(4, 4L/\epsilon),
\]

\[
C_2 = 6 \left( \left( \frac{\beta_1}{1 - \beta_1} \right)^2 + \left( \frac{1}{1 - \lambda} \right)^2 \right) L \frac{G_{\infty}^2}{\epsilon^{1.5}},
\]

\[
C_3 = 16L^2 \left( \frac{1}{1 - \lambda} \right) \frac{G_{\infty}^2}{\epsilon^2},
\]

\[
C_4 = \frac{2}{\epsilon^{1.5}} \frac{1}{1 - \lambda} \left( \lambda + \frac{\beta_1}{1 - \beta_1} \right) G_{\infty}^2,
\]

\[
C_5 = \frac{2}{\epsilon^2} \frac{1}{1 - \lambda} L \left( \frac{\beta_1}{1 - \beta_1} \right)^2 G_{\infty}^2 + \frac{4}{\epsilon^2} \frac{\lambda}{1 - \lambda} LG_{\infty}^2.
\]
Setting $\alpha \leq \frac{\epsilon}{10L}$, we can cancel last term on the RHS of (31) with LHS, then substituting into $Z_1 = X_1$ completes the proof. \qed
Appendix C. Proof of Theorem 3

Under some assumptions stated in Corollary 1, we have that

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(X_t) \right\|_{U_t}^{1/4} \right]^2 \leq C_1 \sqrt{d} \sqrt{T/N} \left( (\mathbb{E}[f(Z_1)]) - \min_x f(x) + \sigma^2 \right) + C_2 \frac{N}{T} + C_3 \frac{N^{1.5}}{T^{1.5}d^{0.5}}
\]

\[+ \left( C_4 \frac{1}{T^{1/2}} + C_5 \frac{1}{T^{1.5}d^{0.5}} \right) \mathbb{E} \left[ \sum_{t=1}^{T} \left\| (-\hat{V}_{t-2} + \hat{V}_{t-1}) \right\|_{abs} \right] \tag{33}
\]

where \( \left\| \cdot \right\|_{abs} \) denotes the entry-wise \( L_1 \) norm of a matrix (i.e \( \left\| A \right\|_{abs} = \sum_{i,j} |A_{ij}| \)) 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 \( \mathbb{E} \left[ \sum_{t=1}^{T} \left\| (-\hat{V}_{t-2} + \hat{V}_{t-1}) \right\|_{abs} \right] \) to prove convergence of Algorithm 3.

By the update rule of Algorithm 3, we know \( \hat{V}_t \) is non decreasing and thus

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

\[= \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d} (-[\hat{V}_{t-2,i}]_j + [\hat{V}_{t-1,i}]_j) \right]
\]

\[= \mathbb{E} \left[ \sum_{i=1}^{N} \sum_{j=1}^{d} (-[\hat{V}_0,i]_j + [\hat{V}_{T-1,i}]_j) \right]
\]

where the last equality is because we defined \( \hat{V}_{-1} \equiv \hat{V}_0 \) previously. Furthermore, because \( \|g_{t,i}\|_{\infty} \leq G_{\infty} \) for all \( t, i \) and \( \nu_{t,i} \) is an exponential moving average of \( g_{k,i}^2, k = 1, 2, \cdots, t \), we know \( |[\nu_{t,i}]_j| \leq G_{\infty}^2 \), 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 \( G_{\infty}^2 \), i.e. \( |[\hat{V}_t,i]_j| \leq G_{\infty}^2 \), for all \( t, i, j \). Given the fact that \( |[\hat{V}_0,i]_j| \geq 0 \), we have

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \left\| (-\hat{V}_{t-2} + \hat{V}_{t-1}) \right\|_{abs} \right] \leq \mathbb{E} \left[ \sum_{i=1}^{N} \sum_{j=1}^{d} (-[\hat{V}_0,i]_j + [\hat{V}_{T-1,i}]_j) \right] \leq \mathbb{E} \left[ \sum_{i=1}^{N} \sum_{j=1}^{d} G_{\infty}^2 \right] = NdG_{\infty}^2.
\]
Substituting the above into (33), we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \nabla f(X_t) \right\|_{U_t^{1/4}}^2 \right] \leq C_1 \frac{\sqrt{d}}{\sqrt{TN}} \left( (\mathbb{E}[f(Z_1)] - \min_x f(x)) + \sigma^2 \right) + C_2 \frac{N}{T} + C_3 \frac{N^{1.5}}{T^{1.5}d^{0.5}} \\
+ \left( C_4 \frac{1}{T \sqrt{N}} + C_5 \frac{1}{T^{1.5}d^{0.5}} \right) NdG^2_\infty
\]

\[
= C_1' \frac{\sqrt{d}}{\sqrt{TN}} \left( (\mathbb{E}[f(Z_1)] - \min_x f(x)) + \sigma^2 \right) + C_2' \frac{N}{T} + C_3' \frac{N^{1.5}}{T^{1.5}d^{0.5}} \\
+ C_4' \frac{\sqrt{N}d}{T} + C_5' \frac{Nd^{0.5}}{T^{1.5}},
\]

where we have

\[
C_1' = C_1 \quad C_2' = C_2 \quad C_3' = C_3 \quad C_4' = C_4 G^2_\infty \quad C_5' = C_5 G^2_\infty.
\]

and we conclude the proof. \qed
Appendix D. Proof of Theorem 4

The proof follows the same flow as that of Theorem 3. Under assumptions stated in Corollary 1, set \( \alpha = \sqrt{N}/\sqrt{T d} \), we have that

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{\nabla f(\overline{X}_t)}{U_t^{1/4}} \right\|^2 \right] \leq C_1 \frac{\sqrt{d}}{\sqrt{T N}} \left( (\mathbb{E}[f(Z_1)] - \min_x f(x)) + \sigma^2 \right) + C_2 \frac{N}{T} + C_3 \frac{N^{1.5}}{T^{1.5}d^{0.5}}
\]

\[+ \left( C_4 \frac{1}{T \sqrt{N}} + C_5 \frac{1}{T^{1.5}d^{0.5}} \right) \mathbb{E} \left[ \sum_{t=1}^{T} \|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{\text{abs}} \right], \quad (36)
\]

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

Again, Since decentralized AdaGrad is a special case of 2, we can apply Corollary 1 and what we need is to upper bound \( \mathbb{E} \left[ \sum_{t=1}^{T} \|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{\text{abs}} \right] \) derive convergence rate. By the update rule of decentralized AdaGrad, we have \( \hat{v}_{t,i} = \frac{1}{t} (\sum_{k=1}^{t} g^2_{k,i}) \) for \( t \geq 1 \) and \( \hat{v}_{0,i} = \epsilon 1 \). Then we have for \( t \geq 3 \),

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \|(-\hat{V}_{t-2} + \hat{V}_{t-1})\|_{\text{abs}} \right]
\]

\[= \mathbb{E} \left[ \sum_{t=3}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \hat{v}_{t-2,i,j} + \hat{v}_{t-1,i,j} \right| \right]
\]

\[\leq \mathbb{E} \left[ \sum_{t=3}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \frac{1}{t-2} \left( \sum_{k=1}^{t-2} g^2_{k,i,j} \right) + \frac{1}{t-1} \left( \sum_{k=1}^{t-1} g^2_{k,i,j} \right) \right| \right] + N d (G^2 - \epsilon)
\]

\[\leq \mathbb{E} \left[ \sum_{t=3}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \frac{1}{t-1} - \frac{1}{t-2} \right| \left( \sum_{k=1}^{t-2} g^2_{k,i,j} \right) + \frac{1}{t-1} \left( \sum_{k=1}^{t-1} g^2_{k,i,j} \right) \right] + N d G^2_{\infty}
\]

\[\leq \mathbb{E} \left[ \sum_{t=3}^{T} \sum_{i=1}^{N} \sum_{j=1}^{d} \max \left( \frac{1}{(t-1)(t-2)} \left( \sum_{k=1}^{t-2} g^2_{k,i,j} \right), \frac{1}{t-1} \left( g^2_{t-1,i,j} \right) \right) \right] + N d G^2_{\infty}
\]

\[\leq \mathbb{E} \left[ N d \sum_{t=3}^{T} \frac{G^2_{\infty}}{t-1} \right] + N d G^2_{\infty}
\]

\[\leq N d G^2_{\infty} \log(T) + N d G^2_{\infty}
\]

\[= N d G^2_{\infty} (\log(T) + 1)
\]

where the first equality is because we defined \( \hat{V}_{t-1} \triangleq \hat{V}_0 \) previously and \( \|g_{k,i}\|_{\infty} \leq G_{\infty} \) by assumption.
Substituting the above into (36), we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left\| \frac{\nabla f(X_t)}{U_t^{1/4}} \right\|^2 \right] \leq C_1 \frac{\sqrt{d}}{\sqrt{T}N} \left( (\mathbb{E}[f(Z_1)] - \min_x f(x)) + \sigma^2 \right) + C_2 \frac{N}{T} + C_3 \frac{N^{1.5}}{T^{1.5}d^{0.5}}
\]

\[
+ \left( C_4 \frac{1}{T \sqrt{N}} + C_5 \frac{1}{T^{1.5}d^{0.5}} \right) N d G_\infty^2 (\log(T) + 1)
\]

\[
= C_1' \frac{\sqrt{d}}{\sqrt{T}N} \left( (\mathbb{E}[f(Z_1)] - \min_x f(x)) + \sigma^2 \right) + C_2' \frac{N}{T} + C_3' \frac{N^{1.5}}{T^{1.5}d^{0.5}}
\]

\[
+ C_4' \frac{d \sqrt{N}(\log(T) + 1)}{T} + C_5' \frac{(\log(T) + 1)N \sqrt{d}}{T^{1.5}},
\]

where we have

\[
C_1' = C_1, \quad C_2' = C_2, \quad C_3' = C_3, \quad C_4' = C_4 G_\infty^2, \quad C_5' = C_5 G_\infty^2.
\]

and we conclude the proof. \(\square\)

**Appendix E. Additional Experiments and Details**

In this section, we compare the training loss and testing accuracy of different algorithms, namely Decentralized Stochastic Gradient Descent (D-PSGD), Decentralized ADAM (DADAM) and our proposed Decentralized AMSGrad, with different stepizes 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 stepizes in the grid \([10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]\).

Figure 3 shows the training loss and test accuracy for D-PSGD algorithm. We observe that the stepsize \(10^{-3}\) works best for D-PSGD 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).

Figure 4: Training loss and Testing accuracy for different stepsizes for AMSGrad

Figure 5: Training loss and Testing accuracy for different stepsizes for DADAM

Figure 4 shows the performance of decentralized AMSGrad with different stepsizes. We see that
its best performance is better than the one of D-PSGD. Its performance is also more stable in the
sense that the test performance is less sensitive to stepsize tuning according to our experiments.

Figure 5 displays the performance of Decentralized ADAM algorithm. As expected, the perform-
ance of DADAM is not as good as D-PSGD or decentralized AMSGrad. Its divergence characteris-
tic, highlighted Section 3.1, coupled with the heterogeneity in the data amplify its non-convergence
issue in our experiments. From the experiments above, we can see the benefits of decentralized
AMSGrad both in terms of performance and ease of parameter tuning, and the importance of ensuring
the theoretical convergence of any newly proposed methods in the presented setting.