APPENDIX

A PROOF OF THEOREM 1

Proof. Denote $\hat{\alpha}_t = \alpha_t \frac{\sqrt{1-\beta_2^t}}{1-\beta_{1,t}^t}$ and $\boldsymbol{v}_t = \max(\sqrt{\boldsymbol{b}_t}, \delta\sqrt{1-\beta_2^t})$, then

$$\begin{aligned}\|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_{t+1} - \boldsymbol{w}^*)\|^2 &= \|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t - \hat{\alpha}_t\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t} - \boldsymbol{w}^*)\|^2 = \|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t - \boldsymbol{w}^*)\|^2 - 2\hat{\alpha}_t\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{m}_t\right\rangle + \hat{\alpha}_t^2\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\|^2 \\ &= \|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t - \boldsymbol{w}^*)\|^2 + \hat{\alpha}_t^2\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\|^2 - 2\hat{\alpha}_t\beta_{1,t}\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{m}_{t-1}\right\rangle - 2\hat{\alpha}_t(1-\beta_{1,t})\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{g}_t\right\rangle.\end{aligned} \tag{6}$$

Rearranging Equation 6, we have

$$\begin{aligned}\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{g}_t\right\rangle =& \frac{1}{1-\beta_{1,t}}\left[\frac{1}{2\hat{\alpha}_t}(\|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t - \boldsymbol{w}^*)\|^2 - \|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_{t+1} - \boldsymbol{w}^*)\|^2) - \beta_{1,t}\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{m}_{t-1}\right\rangle + \frac{\hat{\alpha}_t}{2}\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\|^2\right] \\ \leq& \frac{1}{1-\beta_1}\Bigg[\frac{1}{2\hat{\alpha}_t}(\|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t - \boldsymbol{w}^*)\|^2 - \|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_{t+1} - \boldsymbol{w}^*)\|^2) + \frac{\beta_{1,t}}{2\hat{\alpha}_t}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 \\ &+ \frac{\beta_{1,t}\hat{\alpha}_t}{2}\|\boldsymbol{m}_{t-1}\|^2 + \frac{\hat{\alpha}_t}{2}\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\|^2\Bigg],\end{aligned}$$

where the first inequality follows from Cauchy-Schwartz inequality and $ab \leq \frac{1}{2}(a^2+b^2)$. Hence, the regret

$$\begin{aligned}\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^*)) \leq& \sum_{t=1}^{T}\left\langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{g}_t\right\rangle \\ \leq& \frac{1}{1-\beta_1}\sum_{t=1}^{T}\Bigg[\frac{1}{2\hat{\alpha}_t}(\|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t - \boldsymbol{w}^*)\|^2 - \|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_{t+1} - \boldsymbol{w}^*)\|^2) + \frac{\beta_{1,t}}{2\hat{\alpha}_t}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 \\ &+ \frac{\beta_{1,t}\hat{\alpha}_t}{2}\|\boldsymbol{m}_{t-1}\|^2 + \frac{\hat{\alpha}_t}{2}\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\|^2\Bigg],\end{aligned} \tag{7}$$

where the first inequality follows from the convexity of $f_t(\boldsymbol{w})$. For further bounding Equation 7, we need the following lemmas.

Lemma 1. *For the parameter settings and conditions assumed in Theorem 1, we have*

$$\hat{\alpha}_t > \hat{\alpha}_{t+1},\ t \in [T],$$

where $\hat{\alpha}_t = \alpha_t \frac{\sqrt{1-\beta_2^t}}{1-\beta_{1,t}^t}$.

Proof. Since $\frac{1}{1-\beta_{1,t}^t}$ is non-increasing, we only need to prove $\phi(t) = \alpha_t\sqrt{1-\beta_2^t}$ is decreasing. Since

$$\phi'(t) = -\frac{\alpha}{2}t^{-\frac{3}{2}}(1-\beta_2^t)^{\frac{1}{2}} - \frac{\alpha}{2}t^{\frac{1}{2}}(1-\beta_2^t)^{-\frac{1}{2}}\beta_2^{t-1} < 0,$$

hence we complete the proof. □

Lemma 2. *For the parameter settings and conditions assumed in Theorem 1, we have*

$$\sum_{t=1}^{T}\hat{\alpha}_t\|\boldsymbol{m}_t\|^2 < \frac{2n\alpha G_\infty^2}{(1-\beta_1)^3}\sqrt{T}.$$

Proof. From Equation 4, we have

$$\begin{aligned}\hat{\alpha}_t\|\boldsymbol{m}_t\|^2 &= \hat{\alpha}_t\|\sum_{i=1}^{t}(1-\beta_{1,t-i+1})\boldsymbol{g}_{t-i+1}\prod_{j=1}^{i-1}\beta_{1,t-j+1}\|^2 \leq \hat{\alpha}_t\|\sum_{i=1}^{t}\boldsymbol{g}_{t-i+1}\beta_1^{i-1}\|^2 \\ &= \hat{\alpha}_t\sum_{j=1}^{n}\left(\sum_{i=1}^{t}g_{t-i+1,j}\beta_1^{i-1}\right)^2 \leq \hat{\alpha}_t\sum_{j=1}^{n}\left(\sum_{i=1}^{t}g_{t-i+1,j}^2\beta_1^{i-1}\right)\left(\sum_{i=1}^{t}\beta_1^{i-1}\right) \\ &< \frac{\alpha}{\sqrt{t}}\frac{1}{1-\beta_{1,t}}\frac{nG_\infty^2}{(1-\beta_1)^2} \leq \frac{n\alpha G_\infty^2}{(1-\beta_1)^3}\frac{1}{\sqrt{t}},\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality. Therefore,

$$\sum_{t=1}^{T}\hat{\alpha}_t\|\boldsymbol{m}_t\|^2 < \frac{n\alpha G_\infty^2}{(1-\beta_1)^3}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} < \frac{2n\alpha G_\infty^2\sqrt{T}}{(1-\beta_1)^3},$$

where the last inequality follows from

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} &= 1+\int_2^3\frac{1}{\sqrt{2}}ds+\cdots+\int_{T-1}^{T}\frac{1}{\sqrt{T}}ds \\ &< 1+\int_2^3\frac{1}{\sqrt{s-1}}ds+\cdots+\int_{T-1}^{T}\frac{1}{\sqrt{s-1}}ds \\ &= 1+\int_2^T\frac{1}{\sqrt{s-1}}ds = 2\sqrt{T-1}-1 < 2\sqrt{T}.\end{aligned}$$

This completes the proof. □

Lemma 3. *For the parameter settings and conditions assumed in Theorem 1, we have*

$$\|\sqrt{\boldsymbol{v}_t}\|^2 < \frac{n(2G_\infty+\delta)}{(1-\beta_1)^2},\ t\in[T],$$

where $\boldsymbol{v}_t = \max(\sqrt{\boldsymbol{b}_t}, \delta\sqrt{1-\beta_2^t})$.

Proof.

$$\begin{aligned}\|\boldsymbol{m}_t\|_\infty &= \|\sum_{i=1}^{t}(1-\beta_{1,t-i+1})\boldsymbol{g}_{t-i+1}\prod_{j=1}^{i-1}\beta_{1,t-j+1}\|_\infty \leq \|\sum_{i=1}^{t}\boldsymbol{g}_{t-i+1}\beta_1^{i-1}\|_\infty \leq \frac{G_\infty}{1-\beta_1}, \\ \|\boldsymbol{s}_t\|_\infty &\leq \begin{cases}\frac{\|\boldsymbol{m}_1\|_\infty}{1-\beta_{1,t}} \leq \frac{G_\infty}{(1-\beta_1)^2} < \frac{2G_\infty}{(1-\beta_1)^2} & t=1, \\ \frac{\|\boldsymbol{m}_t\|_\infty}{1-\beta_{1,t}^t} + \frac{\|\boldsymbol{m}_{t-1}\|_\infty}{1-\beta_{1,t}^{t-1}} \leq \frac{2G_\infty}{(1-\beta_1)^2} & t>1,\end{cases} \\ \|\boldsymbol{b}_t\|_\infty &= \|(1-\beta_2)\sum_{i=1}^{t}\boldsymbol{s}_{t-i+1}^2\beta_2^{i-1}\|_\infty \leq \frac{4G_\infty^2}{(1-\beta_1)^4}, \\ \|\sqrt{\boldsymbol{v}_t}\|^2 &= \sum_{i=1}^{n}v_{t,i} < n(\|\sqrt{\boldsymbol{b}_t}\|_\infty+\delta) \leq \frac{n(2G_\infty+\delta)}{(1-\beta_1)^2}.\end{aligned}$$

□

Now we return to the proof of Theorem 1. Let $\hat{\alpha}_0 := \hat{\alpha}_1$. By Lemma 1, Lemma 2, Lemma 3 and Equation 7, we have

$$\begin{aligned}
\sum_{t=1}^{T}(f_t(\boldsymbol{w}_t)-f_t(\boldsymbol{w}^*)) \leq & \frac{1}{1-\beta_1}\Bigg[\frac{1}{2\hat{\alpha}_1}\|\sqrt{\boldsymbol{v}_1}(\boldsymbol{w}_1-\boldsymbol{w}^*)\|^2+\sum_{t=2}^{T}\left(\frac{1}{2\hat{\alpha}_t}\|\sqrt{\boldsymbol{v}_t}(\boldsymbol{w}_t-\boldsymbol{w}^*)\|^2-\frac{1}{2\hat{\alpha}_{t-1}}\|\sqrt{\boldsymbol{v}_{t-1}}(\boldsymbol{w}_t-\boldsymbol{w}^*)\|^2\right)\\
&+\sum_{t=1}^{T}\frac{\beta_{1,t}}{2\hat{\alpha}_t}\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2+\sum_{t=1}^{T}\left(\frac{\hat{\alpha}_{t-1}}{2}\|\boldsymbol{m}_{t-1}\|^2+\frac{\hat{\alpha}_t}{2}\|\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t}}\|^2\right)\Bigg]\\
\leq & \frac{1}{1-\beta_1}\Bigg[\frac{D_\infty^2}{2\hat{\alpha}_1}\|\sqrt{\boldsymbol{v}_1}\|^2+\sum_{t=2}^{T}D_\infty^2\left(\frac{\|\sqrt{\boldsymbol{v}_t}\|^2}{2\hat{\alpha}_t}-\frac{\|\sqrt{\boldsymbol{v}_{t-1}}\|^2}{2\hat{\alpha}_{t-1}}\right)+\sum_{t=1}^{T}\frac{\beta_{1,t}}{2\hat{\alpha}_t}nD_\infty^2\\
&+\sum_{t=1}^{T}\hat{\alpha}_t\left(\frac{1}{2}+\frac{1}{2\delta\sqrt{1-\beta_2}}\right)\|\boldsymbol{m}_t\|^2\Bigg]\\
= & \frac{1}{1-\beta_1}\left[D_\infty^2\frac{\|\sqrt{\boldsymbol{v}_T}\|^2}{2\alpha_T}+\sum_{t=1}^{T}\frac{\beta_{1,t}}{2\hat{\alpha}_t}nD_\infty^2+\sum_{t=1}^{T}\hat{\alpha}_t\left(\frac{1}{2}+\frac{1}{2\delta\sqrt{1-\beta_2}}\right)\|\boldsymbol{m}_t\|^2\right]\\
< & \frac{1}{1-\beta_1}\left[\frac{n(2G_\infty+\delta)D_\infty^2}{2\alpha\sqrt{1-\beta_2}(1-\beta_1)^2}\sqrt{T}+\sum_{t=1}^{T}\frac{\beta_{1,t}}{2\hat{\alpha}_t}nD_\infty^2+\frac{n\alpha G_\infty^2}{(1-\beta_1)^3}\left(1+\frac{1}{\delta\sqrt{1-\beta_2}}\right)\sqrt{T}\right].
\end{aligned}$$

This completes the proof. □

B Proof of Corollary 1

Proof. Since $\beta_{1,t}=\beta_1/t$, we have

$$\sum_{t=1}^{T}\frac{\beta_{1,t}}{2\hat{\alpha}_t}=\sum_{t=1}^{T}\frac{(1-\beta_{1,t}^t)\sqrt{t}\beta_{1,t}}{2\alpha\sqrt{1-\beta_2^t}}<\sum_{t=1}^{T}\frac{\sqrt{t}\beta_{1,t}}{2\alpha\sqrt{1-\beta_2}}=\frac{\beta_1}{2\alpha\sqrt{1-\beta_2}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}<\frac{\beta_1}{\alpha\sqrt{1-\beta_2}}\sqrt{T}.$$

This completes the proof. □

C Proof of Theorem 2

Proof. Denote $\hat{\alpha}_t=\alpha_t\frac{\sqrt{1-\beta_2^t}}{1-\beta_{1,t}^t}$ and $\boldsymbol{v}_t=\max(\sqrt{\boldsymbol{b}_t},\delta\sqrt{1-\beta_2^t})$. By assumptions 2, 4, Lemma 1 and Lemma 3, $\forall t\in[T]$, we have

$$\|\boldsymbol{m}_t\|_\infty\leq\frac{G_\infty}{1-\beta_1},\quad \|\boldsymbol{v}_t\|_\infty\leq\frac{2G_\infty}{(1-\beta_1)^2},\quad \frac{\hat{\alpha}_t}{v_{t,i}}>\frac{\hat{\alpha}_{t+1}}{v_{t+1,i}},\ \forall i\in[n]. \tag{8}$$

Following Yang et al. (2016); Chen et al. (2019); Zhou et al. (2018), we define an auxiliary sequence $\{\boldsymbol{u}_t\}$: $\forall t\geq 2$,

$$\boldsymbol{u}_t=\boldsymbol{w}_t+\frac{\beta_{1,t}}{1-\beta_{1,t}}(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})=\frac{1}{1-\beta_{1,t}}\boldsymbol{w}_t-\frac{\beta_{1,t}}{1-\beta_{1,t}}\boldsymbol{w}_{t-1}, \tag{9}$$

hence, we have

$$\begin{aligned}
\boldsymbol{u}_{t+1}-\boldsymbol{u}_t=&\left(\frac{1}{1-\beta_{1,t+1}}-\frac{1}{1-\beta_{1,t}}\right)\boldsymbol{w}_{t+1}-\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}-\frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\boldsymbol{w}_t\\
&+\frac{1}{1-\beta_{1,t}}(\boldsymbol{w}_{t+1}-\boldsymbol{w}_t)-\frac{\beta_{1,t}}{1-\beta_{1,t}}(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})\\
=&\left(\frac{1}{1-\beta_{1,t+1}}-\frac{1}{1-\beta_{1,t}}\right)(\boldsymbol{w}_t-\hat{\alpha}_t\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t})-\left(\frac{\beta_{1,t+1}}{1-\beta_{1,t+1}}-\frac{\beta_{1,t}}{1-\beta_{1,t}}\right)\boldsymbol{w}_t\\
&-\frac{\hat{\alpha}_t}{1-\beta_{1,t}}\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t}+\frac{\beta_{1,t}\hat{\alpha}_{t-1}}{1-\beta_{1,t}}\frac{\boldsymbol{m}_{t-1}}{\boldsymbol{v}_{t-1}}\\
=&\left(\frac{1}{1-\beta_{1,t}}-\frac{1}{1-\beta_{1,t+1}}\right)\hat{\alpha}_t\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t}-\frac{\hat{\alpha}_t}{1-\beta_{1,t}}\left(\beta_{1,t}\frac{\boldsymbol{m}_{t-1}}{\boldsymbol{v}_t}+(1-\beta_{1,t})\frac{\boldsymbol{g}_t}{\boldsymbol{v}_t}\right)+\frac{\beta_{1,t}\hat{\alpha}_{t-1}}{1-\beta_{1,t}}\frac{\boldsymbol{m}_{t-1}}{\boldsymbol{v}_{t-1}}\\
=&\left(\frac{1}{1-\beta_{1,t}}-\frac{1}{1-\beta_{1,t+1}}\right)\hat{\alpha}_t\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t}+\frac{\beta_{1,t}}{1-\beta_{1,t}}\left(\frac{\hat{\alpha}_{t-1}}{\boldsymbol{v}_{t-1}}-\frac{\hat{\alpha}_t}{\boldsymbol{v}_t}\right)\boldsymbol{m}_{t-1}-\hat{\alpha}_t\frac{\boldsymbol{g}_t}{\boldsymbol{v}_t}.
\end{aligned}\tag{10}$$

By assumption 1 and Equation 10, we have

$$\begin{aligned}
f(\boldsymbol{u}_{t+1})\leq& f(\boldsymbol{u}_t)+\langle\nabla f(\boldsymbol{u}_t),\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\rangle+\frac{L}{2}\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|^2\\
=&f(\boldsymbol{u}_t)+\langle\nabla f(\boldsymbol{w}_t),\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\rangle+\langle\nabla f(\boldsymbol{u}_t)-\nabla f(\boldsymbol{w}_t),\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\rangle+\frac{L}{2}\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|^2\\
=&f(\boldsymbol{u}_t)+\left\langle\nabla f(\boldsymbol{w}_t),\left(\frac{1}{1-\beta_{1,t}}-\frac{1}{1-\beta_{1,t+1}}\right)\hat{\alpha}_t\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t}\right\rangle+\frac{\beta_{1,t}}{1-\beta_{1,t}}\left\langle\nabla f(\boldsymbol{w}_t),\left(\frac{\hat{\alpha}_{t-1}}{\boldsymbol{v}_{t-1}}-\frac{\hat{\alpha}_t}{\boldsymbol{v}_t}\right)\boldsymbol{m}_{t-1}\right\rangle\\
&-\hat{\alpha}_t\left\langle\nabla f(\boldsymbol{w}_t),\frac{\boldsymbol{g}_t}{\boldsymbol{v}_t}\right\rangle+\langle\nabla f(\boldsymbol{u}_t)-\nabla f(\boldsymbol{w}_t),\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\rangle+\frac{L}{2}\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|^2.
\end{aligned}\tag{11}$$

Rearranging Equation 11 and taking expectation both sides, by assumption 3 and Equation 8, we get

$$\begin{aligned}
\frac{(1-\beta_1)^2\hat{\alpha}_t}{2G_\infty}\mathbf{E}[\|\nabla f(\boldsymbol{w}_t)\|^2]\leq&\hat{\alpha}_t\mathbf{E}\left[\left\langle\nabla f(\boldsymbol{w}_t),\frac{\nabla f(\boldsymbol{w}_t)}{\boldsymbol{v}_t}\right\rangle\right]\\
\leq&\mathbf{E}[f(\boldsymbol{u}_t)-f(\boldsymbol{u}_{t+1})]+\underbrace{\mathbf{E}\left[\left\langle\nabla f(\boldsymbol{w}_t),\left(\frac{1}{1-\beta_{1,t}}-\frac{1}{1-\beta_{1,t+1}}\right)\hat{\alpha}_t\frac{\boldsymbol{m}_t}{\boldsymbol{v}_t}\right\rangle\right]}_{P_1}\\
&+\frac{\beta_{1,t}}{1-\beta_{1,t}}\underbrace{\mathbf{E}\left[\left\langle\nabla f(\boldsymbol{w}_t),\left(\frac{\hat{\alpha}_{t-1}}{\boldsymbol{v}_{t-1}}-\frac{\hat{\alpha}_t}{\boldsymbol{v}_t}\right)\boldsymbol{m}_{t-1}\right\rangle\right]}_{P_2}\\
&+\underbrace{\mathbf{E}\left[\langle\nabla f(\boldsymbol{u}_t)-\nabla f(\boldsymbol{w}_t),\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\rangle\right]}_{P_3}+\frac{L}{2}\underbrace{\mathbf{E}\left[\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|^2\right]}_{P_4}.
\end{aligned}\tag{12}$$

For further bounding Equation 12, we need the following lemma.

Lemma 4. *For the sequence $\{\boldsymbol{u}_t\}$ defined as Equation 9, $\forall t\geq 2$, we have*

$$\begin{aligned}
\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|&\leq\frac{\sqrt{n}G_\infty}{\delta}\left(\frac{\hat{\alpha}_t\beta_{1,t}}{(1-\beta_1)^3}+\frac{\hat{\alpha}_{t-1}\beta_{1,t}}{(1-\beta_1)^2}+\hat{\alpha}_t\right),\\
\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|^2&\leq\frac{3nG_\infty^2}{\delta^2}\left(\frac{\hat{\alpha}_t^2\beta_{1,t}^2}{(1-\beta_1)^6}+\frac{\hat{\alpha}_{t-1}^2\beta_{1,t}^2}{(1-\beta_1)^4}+\hat{\alpha}_t^2\right).
\end{aligned}$$

Proof. Since $\forall t \in [T], \forall i \in [n]$, $1/(1-\beta_{1,t}) \geq 1/(1-\beta_{1,t+1})$, $\boldsymbol{v}_t \geq \delta\sqrt{1-\beta_2}$, $\hat{\alpha}_{t-1}/v_{t-1,i} > \hat{\alpha}_t/v_{t,i}$. By Equation 10, we have

$$\begin{aligned}
\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\| &\leq \hat{\alpha}_t \frac{\sqrt{n}G_\infty}{(1-\beta_1)\delta\sqrt{1-\beta_2}} \left(\frac{\beta_{1,t} - \beta_{1,t+1}}{(1-\beta_{1,t})(1-\beta_{1,t+1})} \right) + \frac{\beta_{1,t}}{1-\beta_{1,t}} \hat{\alpha}_{t-1} \frac{\sqrt{n}G_\infty}{(1-\beta_1)\delta\sqrt{1-\beta_2}} + \hat{\alpha}_t \frac{\sqrt{n}G_\infty}{\delta\sqrt{1-\beta_2}} \\
&\leq \frac{\sqrt{n}G_\infty}{\delta\sqrt{1-\beta_2}} \left(\frac{\hat{\alpha}_t \beta_{1,t}}{(1-\beta_1)^3} + \frac{\hat{\alpha}_{t-1}\beta_{1,t}}{(1-\beta_1)^2} + \hat{\alpha}_t \right), \\
\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|^2 &\leq \frac{3nG_\infty^2}{\delta^2(1-\beta_2)} \left(\frac{\hat{\alpha}_t^2 \beta_{1,t}^2}{(1-\beta_1)^6} + \frac{\hat{\alpha}_{t-1}^2 \beta_{1,t}^2}{(1-\beta_1)^4} + \hat{\alpha}_t^2 \right),
\end{aligned}$$

where the last inequality follows from Cauchy-Schwartz inequality. This completes the proof. □

Now we bound P_1, P_2, P_3 and P_4 of Equation 12 separately. By assumptions 1, 2, Equation 8 and Lemma 4, we have

$$\begin{aligned}
P_1 &\leq \hat{\alpha}_t \left(\frac{1}{1-\beta_{1,t}} - \frac{1}{1-\beta_{1,t+1}} \right) \mathbf{E} \left[\|\nabla f(\boldsymbol{w}_t)\| \| \frac{\boldsymbol{m}_t}{\boldsymbol{v}_t} \| \right] \\
&\leq \hat{\alpha}_t \left(\frac{\beta_{1,t} - \beta_{1,t+1}}{(1-\beta_{1,t})(1-\beta_{1,t+1})} \right) \frac{nG_\infty^2}{\delta\sqrt{1-\beta_2}(1-\beta_1)} \leq \frac{nG_\infty^2}{(1-\beta_1)^3\delta\sqrt{1-\beta_2}} \hat{\alpha}_t(\beta_{1,t} - \beta_{1,t+1}), \\
P_2 &\leq \mathbf{E} \left[\|\nabla f(\boldsymbol{w}_t)\| \| \left(\frac{\hat{\alpha}_{t-1}}{\boldsymbol{v}_{t-1}} - \frac{\hat{\alpha}_t}{\boldsymbol{v}_t} \right) \boldsymbol{m}_{t-1} \| \right] \leq \frac{G_\infty^2}{1-\beta_1} \| \frac{\hat{\alpha}_{t-1}}{\boldsymbol{v}_{t-1}} - \frac{\hat{\alpha}_t}{\boldsymbol{v}_t} \| \leq \frac{G_\infty^2}{1-\beta_1} \sum_{i=1}^n (\frac{\hat{\alpha}_{t-1}}{v_{t-1,i}} - \frac{\hat{\alpha}_t}{v_{t,i}}), \\
P_3 &\leq \mathbf{E} \left[\|\nabla f(\boldsymbol{u}_t) - \nabla f(\boldsymbol{w}_t)\| \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\| \right] \leq L\mathbf{E} \left[\|\boldsymbol{u}_t - \boldsymbol{w}_t\| \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\| \right] \\
&= L\hat{\alpha}_{t-1} \frac{\beta_{1,t}}{1-\beta_{1,t}} \mathbf{E} \left[\| \frac{\boldsymbol{m}_{t-1}}{\boldsymbol{v}_{t-1}} \| \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\| \right] \leq \frac{LnG_\infty^2}{(1-\beta_1)^2\delta^2(1-\beta_2)} \left(\frac{\hat{\alpha}_{t-1}\hat{\alpha}_t\beta_{1,t}^2}{(1-\beta_1)^3} + \frac{\hat{\alpha}_{t-1}^2\beta_{1,t}^2}{(1-\beta_1)^2} + \hat{\alpha}_{t-1}\hat{\alpha}_t\beta_{1,t} \right) \\
&< \frac{LnG_\infty^2}{(1-\beta_1)^2\delta^2(1-\beta_2)} \left(\frac{\hat{\alpha}_{t-1}^2}{(1-\beta_1)^3} + \frac{\hat{\alpha}_{t-1}^2}{(1-\beta_1)^2} + \hat{\alpha}_{t-1}^2 \right) < \frac{3LnG_\infty^2}{(1-\beta_1)^5\delta^2(1-\beta_2)} \hat{\alpha}_{t-1}^2, \\
P_4 &\leq \frac{3nG_\infty^2}{\delta^2(1-\beta_2)} \left(\frac{\hat{\alpha}_t^2\beta_{1,t}^2}{(1-\beta_1)^6} + \frac{\hat{\alpha}_{t-1}^2\beta_{1,t}^2}{(1-\beta_1)^4} + \hat{\alpha}_t^2 \right) < \frac{3nG_\infty^2}{\delta^2(1-\beta_2)} \left(\frac{\hat{\alpha}_t^2}{(1-\beta_1)^6} + \frac{\hat{\alpha}_{t-1}^2}{(1-\beta_1)^4} + \hat{\alpha}_t^2 \right) \\
&< \frac{9nG_\infty^2}{(1-\beta_1)^6\delta^2(1-\beta_2)} \hat{\alpha}_{t-1}^2.
\end{aligned} \tag{13}$$

Replacing P_1, P_2, P_3 and P_4 of Equation 12 with Equation 13 and telescoping Equation 12 for $t = 2$ to T, we have

$$\begin{aligned}
&\sum_{t=2}^T \frac{(1-\beta_1)^2\hat{\alpha}_t}{2G_\infty} \mathbf{E} \left[\|\nabla f(\boldsymbol{w}_t)\|^2 \right] < \mathbf{E} \left[f(\boldsymbol{u}_2) - f(\boldsymbol{u}_{T+1}) \right] + \frac{nG_\infty^2}{(1-\beta_1)^3\delta\sqrt{1-\beta_2}} \sum_{t=2}^T \hat{\alpha}_t(\beta_{1,t} - \beta_{1,t+1}) \\
&+ \frac{\beta_1 G_\infty^2}{(1-\beta_1)^2} \sum_{i=1}^n \left(\frac{\hat{\alpha}_1}{v_{1,i}} - \frac{\hat{\alpha}_T}{v_{T,i}} \right) + \frac{3LnG_\infty^2}{(1-\beta_1)^5\delta^2(1-\beta_2)} \sum_{t=2}^T \hat{\alpha}_{t-1}^2 + \frac{9LnG_\infty^2}{2(1-\beta_1)^6\delta^2(1-\beta_2)} \sum_{t=2}^T \hat{\alpha}_{t-1}^2 \\
<&\mathbf{E} \left[f(\boldsymbol{u}_2) \right] - f(\boldsymbol{w}^*) + \frac{nG_\infty^2}{(1-\beta_1)^3\delta\sqrt{1-\beta_2}} \sum_{t=1}^T \hat{\alpha}_t(\beta_{1,t} - \beta_{1,t+1}) + \frac{\alpha\beta_1 nG_\infty^2}{(1-\beta_1)^3\delta} \\
&+ \frac{15LnG_\infty^2}{2(1-\beta_1)^6\delta^2(1-\beta_2)} \sum_{t=1}^T \hat{\alpha}_t^2.
\end{aligned} \tag{14}$$

Since

$$
\begin{aligned}
\sum_{t=2}^{T}\hat{\alpha}_t =& \sum_{t=2}^{T}\frac{\alpha}{\sqrt{t}}\frac{\sqrt{1-\beta_2^t}}{1-\beta_{1,t}^t} \geq \alpha\sqrt{1-\beta_2}\sum_{t=2}^{T}\frac{1}{\sqrt{t}} \\
=& \alpha\sqrt{1-\beta_2}\left(\int_2^3 \frac{1}{\sqrt{2}}ds + \cdots + \int_{T-1}^{T}\frac{1}{\sqrt{T}}ds\right) > \alpha\sqrt{1-\beta_2}\int_2^T \frac{1}{\sqrt{s}}ds \\
=& 2\alpha\sqrt{1-\beta_2}\left(\sqrt{T}-\sqrt{2}\right), \\
\sum_{t=1}^{T}\hat{\alpha}_t^2 =& \sum_{t=1}^{T}\frac{\alpha^2}{t}\frac{1-\beta_2^t}{(1-\beta_{1,t})^2} \leq \frac{\alpha^2}{(1-\beta_1)^2}\sum_{t=1}^{T}\frac{1}{t} \\
=& \frac{\alpha^2}{(1-\beta_1)^2}\left(1+\int_2^3 \frac{1}{2}ds + \cdots + \int_{T-1}^{T}\frac{1}{T}ds\right) < \frac{\alpha^2}{(1-\beta_1)^2}\left(1+\int_2^3 \frac{1}{s-1}ds\right) \\
=& \frac{\alpha^2}{(1-\beta_1)^2}\left(\log(T-1)+1\right) < \frac{\alpha^2}{(1-\beta_1)^2}\left(\log T+1\right), \\
\mathbf{E}\left[f(\boldsymbol{u}_2)\right] \leq& f(\boldsymbol{w}_1) + \mathbf{E}\left[\langle\nabla f(\boldsymbol{w}_1), \boldsymbol{u}_2-\boldsymbol{w}_1\rangle\right] + \frac{L}{2}\mathbf{E}\left[\|\boldsymbol{u}_2-\boldsymbol{w}_1\|^2\right] \\
=& f(\boldsymbol{w}_1) - \frac{\hat{\alpha}_1}{1-\beta_{1,2}}\mathbf{E}\left[\left\langle\nabla f(\boldsymbol{w}_1), \frac{\boldsymbol{m}_1}{\boldsymbol{v}_1}\right\rangle\right] + \frac{L\hat{\alpha}_1^2}{2(1-\beta_{1,2})^2}\mathbf{E}\left[\|\frac{\boldsymbol{m}_1}{\boldsymbol{v}_1}\|^2\right] \\
\leq& f(\boldsymbol{w}_1) + \frac{\alpha\sqrt{1-\beta_2}}{(1-\beta_1)^2}\mathbf{E}\left[\|\nabla f(\boldsymbol{w}_1)\|\|\frac{\boldsymbol{m}_1}{\boldsymbol{v}_1}\|\right] + \frac{L\alpha^2(1-\beta_2)}{2(1-\beta_1)^4}\mathbf{E}\left[\|\frac{\boldsymbol{m}_1}{\boldsymbol{v}_1}\|^2\right] \\
\leq& f(\boldsymbol{w}_1) + \frac{\alpha n G_\infty^2}{(1-\beta_1)^2\delta} + \frac{L\alpha^2 n G_\infty^2}{2(1-\beta_1)^4\delta^2} \leq f(\boldsymbol{w}_1) + \frac{nG_\infty^2\alpha}{2(1-\beta_1)^4\delta^2}(2\delta+L\alpha),
\end{aligned}
\tag{15}
$$

substituting Equation 15 into Equation 14, we have

$$
\begin{aligned}
&\min_{t\in[T]}\mathbf{E}\left[\|\nabla f(\boldsymbol{w}_t)\|^2\right] \\
<& \frac{2G_\infty}{(1-\beta_1)^2\sum_{t=2}^{T}\hat{\alpha}_t}\left(\mathbf{E}\left[f(\boldsymbol{u}_2)\right] - f(\boldsymbol{w}^*) + \frac{nG_\infty^2}{(1-\beta_1)^3\delta\sqrt{1-\beta_2}}\sum_{t=1}^{T}\hat{\alpha}_t(\beta_{1,t}-\beta_{1,t+1})\right. \\
&\left. + \frac{\alpha\beta_1 nG_\infty^2}{(1-\beta_1)^3\delta} + \frac{15LnG_\infty^2}{2(1-\beta_1)^6\delta^2(1-\beta_2)}\sum_{t=1}^{T}\hat{\alpha}_t^2\right) \\
<& \frac{G_\infty}{\alpha(1-\beta_1)^2(1-\beta_2)^2(\sqrt{T}-\sqrt{2})}\left(f(\boldsymbol{w}_1) - f(\boldsymbol{w}^*) + \frac{nG_\infty^2\alpha}{2(1-\beta_1)^4\delta^2}(2\delta+L\alpha)\right. \\
&\left. + \frac{nG_\infty^2}{(1-\beta_1)^3\delta}\sum_{t=1}^{T}\hat{\alpha}_t(\beta_{1,t}-\beta_{1,t+1}) + \frac{\alpha\beta_1 nG_\infty^2}{(1-\beta_1)^3\delta} + \frac{15LnG_\infty^2\alpha^2}{2(1-\beta_1)^8\delta^2}(\log T+1)\right) \\
\leq& \frac{G_\infty}{\alpha(1-\beta_1)^2(1-\beta_2)^2}\left(f(\boldsymbol{w}_1) - f(\boldsymbol{w}^*) + \frac{nG_\infty^2\alpha}{(1-\beta_1)^8\delta^2}(\delta+8L\alpha) + \frac{\alpha\beta_1 nG_\infty^2}{(1-\beta_1)^3\delta}\right)\frac{1}{\sqrt{T}-\sqrt{2}} \\
&+ \frac{15LnG_\infty^3\alpha}{2(1-\beta_2)^2(1-\beta_1)^{10}\delta^2}\frac{\log T}{\sqrt{T}-\sqrt{2}} + \frac{nG_\infty^3}{\alpha(1-\beta_1)^5(1-\beta_2)^2\delta}\frac{\sum_{t=1}^{T}\hat{\alpha}_t(\beta_{1,t}-\beta_{1,t+1})}{\sqrt{T}-\sqrt{2}}.
\end{aligned}
\tag{16}
$$

This completes the proof. □

D PROOF OF COROLLARY 2

Proof. Since $\beta_{1,t} = \beta_1/\sqrt{t}$, we have

$$\sum_{t=1}^{T} \hat{\alpha}_t(\beta_{1,t} - \beta_{1,t+1}) \leq \sum_{t=1}^{T} \hat{\alpha}_t \beta_{1,t} = \sum_{t=1}^{T} \frac{\alpha}{\sqrt{t}} \frac{\sqrt{1-\beta_2^t}}{1-\beta_{1,t}^t} \beta_{1,t} < \frac{\alpha}{1-\beta_1} \sum_{t=1}^{T} \frac{1}{t} < \frac{\alpha}{1-\beta_1}(\log T + 1). \tag{17}$$

Substituting Equation 17 into Equation 16, we have

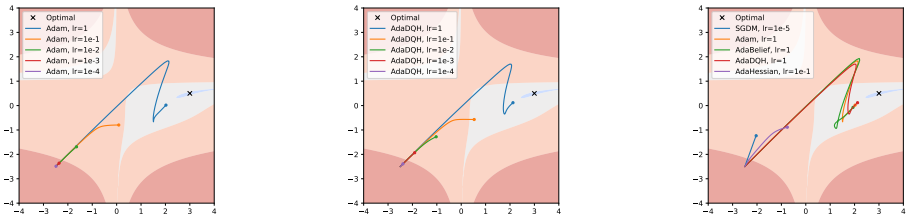
$$\min_{t\in[T]} \mathbf{E}\left[\|\nabla f(\boldsymbol{w}_t)\|^2\right] < \frac{G_\infty}{\alpha(1-\beta_1)^2(1-\beta_2)^2}\left(f(\boldsymbol{w}_1) - f(\boldsymbol{w}^*) + \frac{nG_\infty^2\alpha}{(1-\beta_1)^8\delta^2}(2\delta + 8L\alpha)\right.$$
$$\left. + \frac{\alpha\beta_1 nG_\infty^2}{(1-\beta_1)^3\delta}\right)\frac{1}{\sqrt{T}-\sqrt{2}} + \frac{nG_\infty^3}{(1-\beta_2)^2(1-\beta_1)^{10}\delta^2}\left(\frac{15}{2}L\alpha + \delta\right)\frac{\log T}{\sqrt{T}-\sqrt{2}}.$$

This completes the proof. □

E DETAILS OF EXPERIMENTS

E.1 NUMERICAL EXPERIMENTS

We use the same reasonable learning rate across optimizers in numerical experiments, mainly because larger learning rate usually gives an extra edge to the rate of convergence, as shown in Figures 5a and 5b. Another observation is that larger learning rate makes training unstable. So we search for the largest lr each optimizer can get in {1e-5, 1e-4, ..., 1, 10} for Beale function. The optimization trajectory is shown in Figure 5c, and AdaDQH has a slightly better performance. Regardless, we believe the learning rate choice in Section 2.3 is a more appropriate representation.



(a) Adam (b) AdaDQH (c) Comparison

Figure 5: Optimization trajectories on Beale function using various learning rates.

E.2 CONFIGURATION OF OPTIMIZERS

Here, we give the details of the hyperparameters of optimizers on different tasks. We reuse most of the configurations reported in the literature Yao et al. (2020); Zhuang et al. (2020).

CV

- SGD/Adam/AdamW: We adopt the same experimental setup in Yao et al. (2020). For SGD, the initial learning rate is 0.1 and the momentum is set to 0.9. For Adam, the initial learning rate is set to 0.001 and the epsilon is set to 1e-8. For AdamW, the initial learning rate is set to 0.005 and the epsilon is set to 1e-8.
- AdaBelief: We explore the best learning rate for ResNet20/32 on Cifar10 and ResNet18 on ImageNet, respectively. Finally, the initial learning rate is set to be 0.01 for ResNet20 on Cifar10 and 0.005 for ResNet32/ResNet18 on Cifar10/ImageNet. The epsilon is set to 1e-16.
- AdaHessian: We use as much as possible the recommended configuration from Yao et al. (2020). The block size and the Hessian power are both 1. The initial learning rate is 0.15 when training on Cifar10, whereas it causes the training to diverge on ImageNet, hence we search for the learning

rate among {1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1} and choose the best which is 0.0001. The epsilon is set to 1e-4.

- AdaDQH: We conduct a grid search of δ and the learning rate. The choice of δ is among {1e-8, 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2} and the search range for the learning rate is from 1e-4 to 1e-2. Finally, the chosen learning rate and δ is 0.007 and 1e-2 for Cifar10 task and 0.0004 and 1e-5 for ImageNet task.

The weight decay for all optimizers is set to 0.0005 on Cifar10 and 0.0001 on ImageNet. $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are for all adaptive optimizers.

NLP

- SGD/Adam/AdamW: For the NMT task, we report the results of SGD from Yao et al. (2020), while setting learning rate as 5e-4 and epsilon as 1e-8 for Adam. For the LM task, we follow the settings from Zhuang et al. (2020), setting learning rate to 30 for SGD and 0.001 for Adam/AdamW while epsilon to 1e-12 for Adam/AdamW when training 1-layer LSTM. For 2,3-layer LSTM, learning rate 0.01 and epsilon 1e-8 are used for Adam/AdamW.
- AdaBelief: For the NMT task we use the recommended configuration from the latest implementation[2] for transformer, which sets learning rate as 5e-4 and epsilon as 1e-16. We adopt the same LSTM experimental setup for the LM task and maintain the optimal settings provided by Zhuang et al. (2020).
- AdaHessian: For the NMT task, we adopt the same experimental setup as in the official implementation[3]. For LM task, we search the learning rate among {1e-3, 1e-2, 0.1, 1} and hessian power among {0.5, 1, 2}, and finally select 0.1/0.5 for learning rate and hessian power for 1-layer LSTM, as well as 1.0/0.5 for 2,3-layer LSTM. Note that AdaHessian appears to overfit when using learning rate 1.0. Accordingly, we also try to decay its learning rate in the 50/90 epoch, but it achieves a similar PPL.
- AdaDQH: For the NMT task, we search learning rate among {5e-5, 1e-4, 5e-4, 1e-3} and δ among {1e-16, 1e-14, 1e-12, 1e-10, 1e-8}. We report the best result with learning rate 5e-5 and δ as 1e-14 for AdaDQH. As for LM task, we search learning rate among {1e-4, 5e-4, 1e-3, 5e-3, 1e-2} and δ from 1e-16 to 1e-4, and the best settings for learning rate/δ are 5e-4/1e-10 and 1e-3/1e-5 for 1-layer LSTM and 2,3-layer LSTM, respectively.

The weight decay is set to 1e-4/1.2e-6 for all optimizers in the NMT/LM task, respectively. For adaptive optimizers, we set (β_1, β_2) to (0.9, 0.98) in the NMT task and (0.9, 0.999) in the LM task.

RecSys It is noteworthy that we reimplement the optimizers for training on our internal distributed system.

- SGD: We search for the learning rate among {1e-4, 1e-3, 1e-2, 0.1, 1} and choose the best results, which are 0.1 for Avazu task and 1e-3 for Criteo task.
- Adam/AdaBelief/AdaHessian: We search the learning rate among {1e-5, 1e-4, 1e-3, 1e-2} and choose the best results which are 1e-4 for Avazu task and 1e-3 for Criteo task. The epsilon of Adam, AdaBelief, and AdaHessian is set to 1e-8, 1e-16, and 1e-8 respectively. The block size and the Hessian power of AdaHessian are both 1.
- AdaDQH: We conduct a grid search of the learning rate and epsilon. The choice of the learning rate is among {1e-5, 1e-4, 1e-3} and δ is among {1e-12, 1e-10, 1e-8, 1e-6, 1e-4, 1e-2}. The best settings for the learning rate/δ are 1e-4/1e-4 for Avazu task and 1e-4/1e-8 for Criteo task.

$\beta_1 = 0.9$ and $\beta_2 = 0.999$ are for all adaptive optimizers.

E.3 AdaDQH with AMSGrad condition

Algorithm 2 gives the AdaDQH optimizer with AMSGrad condition. The AMSGrad condition is usually used to ensure the convergence. We empirically prove that adding AMSGrad condition to

[2]https://github.com/juntang-zhuang/Adabelief-Optimizer

[3]https://github.com/amirgholami/adahessian

AdaDQH will slightly degenerate AdaDQH's performance, as listed in Table 7. For AdaDQH with AMSGrad condition, we re-search for the learning rate among {0.001, 0.003, 0.005, 0.007, 0.01} and for δ among {1e-2, 1e-4, 1e-6, 1e-8}, and find the best learning rate and δ are 0.007 and 1e-2 respectively, which are the same as AdaDQH without AMSGrad condition.

Algorithm 2 AdaDQH with AMSGrad condition

Input: parameters β_1, β_2, δ, $\boldsymbol{w}_1 \in \mathbb{R}^n$, step size α_t, initialize $\boldsymbol{m}_0 = \boldsymbol{0}, \boldsymbol{b}_0 = \boldsymbol{0}$
for $t = 1$ **to** T **do**
  $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{w}_t)$
  $\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
  $\boldsymbol{s}_t = \begin{cases} \boldsymbol{m}_1/(1-\beta_1) & t = 1 \\ \boldsymbol{m}_t/(1-\beta_1^t) - \boldsymbol{m}_{t-1}/(1-\beta_1^{t-1}) & t > 1 \end{cases}$
  $\boldsymbol{b}_t \leftarrow \beta_2 \boldsymbol{b}_{t-1} + (1-\beta_2)\boldsymbol{s}_t^2$
  $\boldsymbol{b}_t \leftarrow \max(\boldsymbol{b}_t, \boldsymbol{b}_{t-1})$ // AMSGrad condtion
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha_t \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \frac{\boldsymbol{m}_t}{\max(\sqrt{\boldsymbol{b}_t}, \delta\sqrt{1-\beta_2^t})}$
end for

Table 7: Top-1 accuracy for AdaDQH without and with AMSGrad condition when trained with ResNet20 on Cifar10.

Optimizer	AdaDQH	AdaDQH + AMSGrad
Accuracy	$92.35 \pm .24$	92.25 ± 0.11

E.4 Discussion of ImageNet experiment

We compare our experimental results on ImageNet with other articles (Yao et al., 2020; Zhuang et al., 2020; Chen et al., 2020). AdaHessian can achieve 70.08% top-1 accuracy in Yao et al. (2020) while we report $69.94 \pm 0.09\%$. We note that we have searched for the best learning rate of AdaHessian, since using the recommended learning rate of 0.15 will cause the training to diverge, which may account for this difference. Our reported top-1 accuracy of AdaBelief is $69.93 \pm 0.09\%$, which is slightly lower than what is reported in Zhuang et al. (2020), i.e. 70.08%. The differences in learning rate scheduler and weight decay rate may account for this discrepancy. Our reported top-1 accuracy of SGD is $69.85 \pm 0.04\%$, significantly lower than 70.23% reported in Chen et al. (2020). We find that the differences in training epochs, learning rate scheduler and weight decay rate are the main reason. We also run the experiment using the same configuration as in Chen et al. (2020), and AdaDQH can achieve 70.45% accuracy at lr = 4e-4 and δ = 1e-5, which is still better than the 70.23% result reported in Chen et al. (2020).

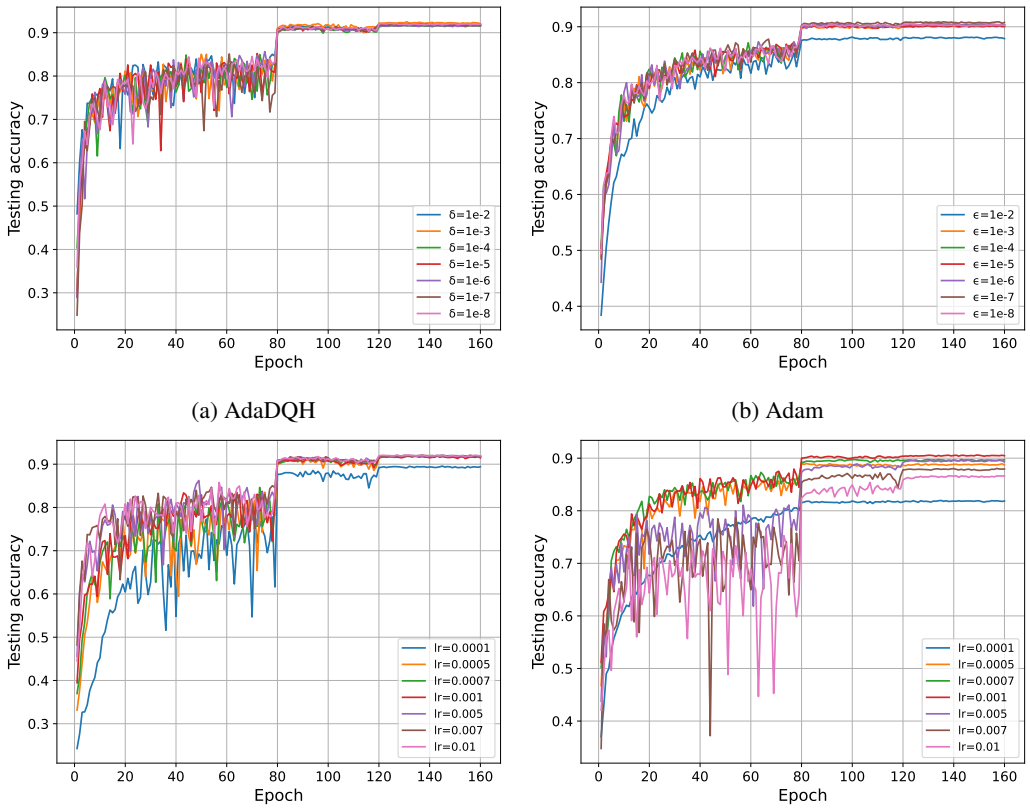
E.5 AdaDQH for ResNet18 on Cifar10

Since the SOTA accuracy [4] for ResNet18 on Cifar10 is 95.55% using SGD optimizer with a cosine learning rate schedule, we also test the performance of AdaDQH for ResNet18 on Cifar10. We find AdaDQH can achieve 95.79% accuracy at lr = 0.001 and δ = 1e-2 when using the same training configuration as the experiment of SGD in Moreau et al. (2022), which exceeds the current SOTA result.

E.6 Robustness to hyperparameters

We test the performance of AdaDQH and Adam with respect to δ/ϵ and learning rate. The experiments are performed with ResNet20 on Cifar10 and the results are shown in Figure 6. Compared to Adam, AdaDQH shows better robustness to the change of hyperparameters.

[4] https://paperswithcode.com/sota/stochastic-optimization-on-cifar-10-resnet-18



(a) AdaDQH (b) Adam

(c) AdaDQH (d) Adam

Figure 6: Testing accuracy of ResNet20 on Cifar10, trained with AdaDQH and Adam using different δ/ϵ and learning rate. For (a) and (b), we choose learning rate as 0.007 and 0.001, respectively. For (c) and (d), we set δ/ϵ to be 1e-2 and 1e-8, respectively.