Supplementary Material for: On Distributed Adaptive Optimization with Gradient Compression

The supplementary material of this paper is organized in three main parts. Section A contains additional content and discussion such as the algorithmic formulation of the single-machine COMP-AMS and QADAM. Section B includes the proof of the main theoretical result. Section C contains more details on the experiments.

A Additional content

A.1 Extension to the Single-Machine Setting

In Corollary 1 we obtain the convergence rate of COMP-AMS in the single machine setting. Such setting has been fully considered in detail for SGD [33]. For clarity, we provide in this subsection the formulation of our method in the single-worker setting, see Algorithm 3. Here, the computations, of the stochastic gradient and the various moment estimates, are all performed on a single-machine and the data is stored in this same worker.

Algorithm 3 COMP-AMS for a single-machine

1: **Input**: parameter β_1, β_2, learning rate η_t.
2: Initialize: central server parameter $\theta_1 \in \Theta \subseteq \mathbb{R}^d$; $e_1 = 0$ the error accumulator; sparsity parameter k; $m_0 = 0$, $v_0 = 0$, $\hat{v}_0 = 0$
3: **for** $t = 1$ to T **do**
4: Compute stochastic gradient $g_t := g_{t,i_t}$ at θ_t for randomly sampled index i_t among the available observations indices
5: Compute $\tilde{g}_t = \mathcal{C}(g_t + e_t)$
6: Update the error $e_{t+1} = e_t + g_t - \tilde{g}_t$
7: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)\tilde{g}_t$
8: $v_t = \beta_2 v_{t-1} + (1 - \beta_2)\tilde{g}_t^2$
9: $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$
10: Update the model $\theta_{t+1} = \theta_t - \eta_t \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$
11: **end for**

A.2 QADAM Method

The closely related work to ours, QADAM discussed in [13], is presented in Algorithm 4. Note that, the original method also compresses the model parameters in the server-to-worker communication, so we adapt it to one-way compression (only for the gradients) as our COMP-AMS. Here, $Q(\cdot)$ is a uniform quantization function that represents the effective update ratio $m/\sqrt{v}$ using low bits. It is formally defined as

$$Q_b(g) = \|g\|_\infty \tilde{Q}_b(g/\|g\|_\infty),$$

where $\tilde{Q}_b(x) = \arg\min_{y \in M_b^d} \|y - x\|_2$, with $M_b := \{-1, -\frac{2^{b-1}-2}{2^{b-1}-1}, ..., 0, ..., \frac{2^{b-1}-2}{2^{b-1}-1}, 1\}$. As we can see, QADAM does not contain the $\hat{v}_t$ term, and needs local moment estimations $m_{t,i}$ and $v_{t,i}$, for $i = 1, ..., n$ on each worker. As discussed in the main paper, this costs substantially more memory and space when training large deep learning models.

Algorithm 4 QADAM [13]

Input: parameters β_1, β_2, learning rate η_t.
Initialize: central server parameter $\theta_1 \in \Theta \subseteq \mathbb{R}^d$; $e_{1,i} = 0$ the error accumulator for each worker; sparsity parameter k; n local workers; local moment estimate $m_{0,i} = 0$, $v_{0,i} = 0$
for $t = 1$ to T **do**
    parallel for worker $i \in [n]$ **do**:
        Receive model parameter θ_t from central server
        Compute stochastic gradient $g_{t,i}$ at θ_t
        $m_{t,i} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}$
        $v_{t,i} = \beta_2 v_{t-1,i} + (1-\beta_2) g_{t,i}^2$
        $a_{t,i} = \frac{m_{t,i}}{\sqrt{v_{t,i}+\epsilon}}$
        Compute $\tilde{a}_{t,i} = Q(a_{t,i} + e_{t,i})$
        Update the error $e_{t+1,i} = e_{t,i} + a_{t,i} - \tilde{a}_{t,i}$
        Send $\tilde{a}_{t,i}$ back to central server
    end parallel
    Central server do:
    $\bar{a}_t = \frac{1}{n}\sum_{i=1}^n \tilde{a}_{t,i}$
    Update the global model $\theta_{t+1} = \theta_t - \eta_t \bar{a}_t$
end for

B Proof of the Convergence Result

B.1 Proof of Theorem 1

Theorem. *Denote* $C_0 = \sqrt{\frac{4(1+q^2)^3}{(1-q^2)^2} G^2 + \epsilon}$, $C_1 = \frac{\beta_1}{1-\beta_1} + \frac{2q}{1-q^2}$. *Under Assumption 1 to Assumption 4, with* $\eta_t = \eta \le \frac{\epsilon}{3C_0\sqrt{2L\max\{2L, C_2\}}}$, *for any* $T > 0$, COMP-AMS *satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\theta_t)\|^2] \le 2C_0\Big(\frac{\mathbb{E}[f(\theta_1) - f(\theta^*)]}{T\eta} + \frac{\eta L\sigma^2}{n\epsilon} + \frac{3\eta^2 L C_0 C_1 \sigma^2}{\epsilon^2} + \frac{12\eta^2 q^2 L C_0 \sigma_g^2}{(1-q^2)^2\epsilon^2} + \frac{(1+C_1)G^2 d}{T\sqrt{\epsilon}} + \frac{\eta(1+2C_1)C_1 L G^2 d}{T\epsilon}\Big).$$

Proof. We first clarify some notations. At time t, let the full-precision gradient of the j-th worker be $g_{t,j}$, the error accumulator be $e_{t,j}$, and the compressed gradient be $\tilde{g}_{t,j} = \mathcal{C}(g_{t,j} + e_{t,j})$. Denote $\bar{g}_t = \frac{1}{n}\sum_{j=1}^N g_{t,j}$, $\bar{\tilde{g}}_t = \frac{1}{n}\sum_{j=1}^N \tilde{g}_{t,j}$ and $\bar{e}_t = \frac{1}{n}\sum_{j=1}^n e_{t,j}$. The second moment computed by the compressed gradients is denoted as $v_t = \beta_2 v_{t-1} + (1-\beta_2)\bar{\tilde{g}}_t^2$, and $\hat{v}_t = \max\{\hat{v}_{t-1}, v_t\}$. Also, the first order moving average sequence

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\bar{\tilde{g}}_t \quad \text{and} \quad m'_t = \beta_1 m'_{t-1} + (1-\beta_1)\bar{g}_t.$$

By construction we have $m'_t = (1-\beta_1)\sum_{i=1}^t \beta_1^{t-i}\bar{g}_i$.

Denote the following auxiliary sequences,

$$\mathcal{E}_{t+1} := (1-\beta_1)\sum_{\tau=1}^{t+1}\beta_1^{t+1-\tau}\bar{e}_\tau$$

$$\theta'_{t+1} := \theta_{t+1} - \eta\frac{\mathcal{E}_{t+1}}{\sqrt{\hat{v}_t + \epsilon}}.$$

Then,

$$
\begin{aligned}
\theta'_{t+1} &= \theta_{t+1} - \eta \frac{\mathcal{E}_{t+1}}{\sqrt{\hat{v}_t + \epsilon}} \\
&= \theta_t - \eta \frac{(1-\beta_1)\sum_{\tau=1}^{t} \beta_1^{t-\tau} \overline{\tilde{g}}_\tau + (1-\beta_1)\sum_{\tau=1}^{t+1} \beta_1^{t+1-\tau} \bar{e}_\tau}{\sqrt{\hat{v}_t + \epsilon}} \\
&= \theta_t - \eta \frac{(1-\beta_1)\sum_{\tau=1}^{t} \beta_1^{t-\tau} (\overline{\tilde{g}}_\tau + \bar{e}_{\tau+1}) + (1-\beta)\beta_1^t \bar{e}_1}{\sqrt{\hat{v}_t + \epsilon}} \\
&= \theta_t - \eta \frac{(1-\beta_1)\sum_{\tau=1}^{t} \beta_1^{t-\tau} \bar{e}_\tau}{\sqrt{\hat{v}_t + \epsilon}} - \eta \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} \\
&= \theta_t - \eta \frac{\mathcal{E}_t}{\sqrt{\hat{v}_{t-1} + \epsilon}} - \eta \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta \Big(\frac{1}{\sqrt{\hat{v}_{t-1} + \epsilon}} - \frac{1}{\sqrt{\hat{v}_t + \epsilon}}\Big)\mathcal{E}_t \\
&\overset{(a)}{=} \theta'_t - \eta \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta \Big(\frac{1}{\sqrt{\hat{v}_{t-1} + \epsilon}} - \frac{1}{\sqrt{\hat{v}_t + \epsilon}}\Big)\mathcal{E}_t \\
&:= \theta'_t - \eta a'_t + \eta D_t \mathcal{E}_t,
\end{aligned}
$$

where (a) uses the fact that for every $j \in [n]$, $\tilde{g}_{t,j} + e_{t+1,j} = g_{t,j} + e_{t,j}$, and $e_{t,1} = 0$ at initialization. Further define the virtual iterates:

$$
x_{t+1} := \theta'_{t+1} - \eta \frac{\beta_1}{1-\beta_1} a'_t = \theta'_{t+1} - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}},
$$

which follows the recurrence:

$$
\begin{aligned}
x_{t+1} &= \theta'_{t+1} - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} \\
&= \theta'_t - \eta \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta D_t \mathcal{E}_t \\
&= \theta'_t - \eta \frac{\beta_1 m'_{t-1} + (1-\beta_1)\bar{g}_t + \frac{\beta_1^2}{1-\beta_1} m'_{t-1} + \beta_1 \bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta D_t \mathcal{E}_t \\
&= \theta'_t - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_{t-1}}{\sqrt{\hat{v}_t + \epsilon}} - \eta \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta D_t \mathcal{E}_t \\
&= x_t - \eta \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + \eta D_t \mathcal{E}_t.
\end{aligned}
$$

When summing over $t = 1, ..., T$, the difference sequence D_t satisfies the bounds of Lemma 5.

By Assumption 2 we have

$$
f(x_{t+1}) \le f(x_t) - \eta \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2.
$$

Taking expectation w.r.t. the randomness at time t, we obtain

$$
\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - f(x_t) \\
&\le -\eta \mathbb{E}[\langle \nabla f(x_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} \rangle] + \eta \mathbb{E}[\langle \nabla f(x_t), \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t \mathcal{E}_t \rangle] \\
&\qquad\qquad + \frac{\eta^2 L}{2} \mathbb{E}[\| \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} - \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} - D_t \mathcal{E}_t \|^2] \\
&= \underbrace{-\eta \mathbb{E}[\langle \nabla f(\theta_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} \rangle]}_{I} + \underbrace{\eta \mathbb{E}[\langle \nabla f(x_t), \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t \mathcal{E}_t \rangle]}_{II} \\
&\quad + \underbrace{\frac{\eta^2 L}{2} \mathbb{E}[\| \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} - \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} - D_t \mathcal{E}_t \|^2]}_{III} + \underbrace{\eta \mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(x_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_t + \epsilon}} \rangle]}_{IV},
\end{aligned} \tag{3}
$$

Bounding term I. We have

$$\begin{aligned} I &= -\eta\mathbb{E}[\langle\nabla f(\theta_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}] - \eta\mathbb{E}[\langle\nabla f(\theta_t), (\frac{1}{\sqrt{\hat{v}_t+\epsilon}} - \frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}})\bar{g}_t\rangle] \\ &\leq -\eta\mathbb{E}[\langle\nabla f(\theta_t), \frac{\nabla f(\theta_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}] + \eta G^2\mathbb{E}[\|D_t\|]. \\ &\leq -\frac{\eta}{\sqrt{\frac{4(1+q^2)^3}{(1-q^2)^2}G^2+\epsilon}}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \eta G^2\mathbb{E}[\|D_t\|_1], \end{aligned} \tag{4}$$

where we use Assumption 3, Lemma 4 and the fact that l_2 norm is no larger than l_1 norm.

Bounding term II. It holds that

$$\begin{aligned} II &\leq \eta(\mathbb{E}[\langle\nabla f(\theta_t), \frac{\beta_1}{1-\beta_1}D_t m'_{t-1} + D_t\mathcal{E}_t\rangle] + \mathbb{E}[\langle\nabla f(x_t) - \nabla f(\theta_t), \frac{\beta_1}{1-\beta_1}D_t m'_{t-1} + D_t\mathcal{E}_t\rangle]) \\ &\leq \eta\mathbb{E}[\|\nabla f(\theta_t)\|\|\frac{\beta_1}{1-\beta_1}D_t m'_{t-1} + D_t\mathcal{E}_t\|] + \eta^2\ L\mathbb{E}[\|\frac{\frac{\beta_1}{1-\beta_1}m'_{t-1}+\mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\|\|\frac{\beta_1}{1-\beta_1}D_t m'_{t-1} + D_t\mathcal{E}_t\|] \\ &\leq \eta C_1 G^2\mathbb{E}[\|D_t\|_1] + \frac{\eta^2 C_1^2 L G^2}{\sqrt{\epsilon}}\mathbb{E}[\|D_t\|_1], \end{aligned} \tag{5}$$

where $C_1 := \frac{\beta_1}{1-\beta_1} + \frac{2q}{1-q^2}$. The second inequality is because of smoothness of $f(\theta)$, and the last inequality is due to Lemma 2, Assumption 3 and the property of norms.

Bounding term III. This term can be bounded as follows:

$$\begin{aligned} III &\leq \eta^2 L\mathbb{E}[\|\frac{\bar{g}_t}{\sqrt{\hat{v}_t+\epsilon}}\|^2] + \eta^2 L\mathbb{E}[\|\frac{\beta_1}{1-\beta_1}D_t m'_{t-1} - D_t\mathcal{E}_t\|^2]] \\ &\leq \frac{\eta^2 L}{\epsilon}\mathbb{E}[\|\frac{1}{n}\sum_{j=1}^{i} g_{t,j} - \nabla f(\theta_t) + \nabla f(\theta_t)\|^2] + \eta^2 L\mathbb{E}[\|D_t(\frac{\beta_1}{1-\beta_1}m'_{t-1} - \mathcal{E}_t)\|^2] \\ &\overset{(a)}{\leq} \frac{\eta^2 L}{\epsilon}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta^2 L\sigma^2}{n\epsilon} + \eta^2 C_1^2 L G^2\mathbb{E}[\|D_t\|^2], \end{aligned} \tag{6}$$

where (a) follows from $\nabla f(\theta_t) = \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\theta_t)$ and Assumption 4 that $g_{t,j}$ is unbiased of $\nabla f_j(\theta_t)$ and has bounded variance σ^2.

Bounding term IV. We have

$$\begin{aligned} IV &= \eta\mathbb{E}[\langle\nabla f(\theta_t) - \nabla f(x_t), \frac{\bar{g}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\rangle] + \eta\mathbb{E}[\langle\nabla f(\theta_t) - \nabla f(x_t), (\frac{1}{\sqrt{\hat{v}_t+\epsilon}} - \frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}})\bar{g}_t\rangle] \\ &\leq \eta\mathbb{E}[\langle\nabla f(\theta_t) - \nabla f(x_t), \frac{\nabla f(\theta_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}\rangle] + \eta^2 L\mathbb{E}[\|\frac{\frac{\beta_1}{1-\beta_1}m'_{t-1}+\mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\|\|D_t g_t\|] \\ &\overset{(a)}{\leq} \frac{\eta\rho}{2\epsilon}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta}{2\rho}\mathbb{E}[\|\nabla f(\theta_t) - \nabla f(x_t)\|^2] + \frac{\eta^2 C_1 L G^2}{\sqrt{\epsilon}}\mathbb{E}[\|D_t\|] \\ &\overset{(b)}{\leq} \frac{\eta\rho}{2\epsilon}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{\eta^3 L}{2\rho}\mathbb{E}[\|\frac{\frac{\beta_1}{1-\beta_1}m'_{t-1}+\mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\|^2] + \frac{\eta^2 C_1 L G^2}{\sqrt{\epsilon}}\mathbb{E}[\|D_t\|_1], \end{aligned} \tag{7}$$

where (a) is due to Young's inequality and (b) is based on Assumption 2.

Regarding the second term in (7), by Lemma 3 and Lemma 1, summing over $t = 1, ..., T$ we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\eta^3 L}{2\rho} \mathbb{E}[\| \frac{\frac{\beta_1}{1-\beta_1} m'_{t-1} + \mathcal{E}_t}{\sqrt{\hat{v}_{t-1} + \epsilon}} \|^2] \\
&\leq \sum_{t=1}^{T} \frac{\eta^3 L}{2\rho\epsilon} \mathbb{E}[\| \frac{\beta_1}{1-\beta_1} m'_{t-1} + \mathcal{E}_t \|^2] \\
&\leq \sum_{t=1}^{T} \frac{\eta^3 L}{\rho\epsilon} \Big[\frac{\beta_1^2}{(1-\beta_1)^2} \mathbb{E}[\|m'_t\|^2] + \mathbb{E}[\|\mathcal{E}_t\|^2] \Big] \\
&\leq \frac{T\eta^3 \beta_1^2 L \sigma^2}{\rho(1-\beta_1)^2 \epsilon} + \frac{\eta^3 \beta_1^2 L}{\rho(1-\beta_1)^2 \epsilon} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \\
&\qquad\qquad + \frac{4T\eta^3 q^2 L}{\rho(1-q^2)^2 \epsilon} (\sigma^2 + \sigma_g^2) + \frac{4\eta^3 q^2 L}{\rho(1-q^2)^2 \epsilon} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \\
&= \frac{T\eta^3 L C_2 \sigma^2}{\rho\epsilon} + \frac{4T\eta^3 q^2 L \sigma_g^2}{\rho(1-q^2)^2 \epsilon} + \frac{\eta^3 L C_2}{\rho\epsilon} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2], \qquad (8)
\end{aligned}
$$

with $C_2 := \frac{\beta_1^2}{(1-\beta_1)^2} + \frac{4q^2}{(1-q^2)^2}$. Now integrating (4), (5), (6), (7) and (8) into (3), taking the telescoping summation over $t = 1, ..., T$, we obtain

$$
\begin{aligned}
&\mathbb{E}[f(x_{T+1}) - f(x_1)] \\
&\leq (-\frac{\eta}{C_0} + \frac{\eta^2 L}{\epsilon} + \frac{\eta\rho}{2\epsilon} + \frac{\eta^3 L C_2}{\rho\epsilon}) \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{T\eta^2 L \sigma^2}{n\epsilon} + \frac{T\eta^3 L C_2 \sigma^2}{\rho\epsilon} + \frac{4T\eta^3 q^2 L \sigma_g^2}{\rho(1-q^2)^2 \epsilon} \\
&\qquad + (\eta(1+C_1)G^2 + \frac{\eta^2(1+C_1)C_1 L G^2}{\sqrt{\epsilon}}) \sum_{t=1}^{T} \mathbb{E}[\|D_t\|_1] + \eta^2 C_1^2 L G^2 \sum_{t=1}^{T} \mathbb{E}[\|D_t\|^2.
\end{aligned}
$$

with $C_0 := \sqrt{\frac{4(1+q^2)^3}{(1-q^2)^2} G^2 + \epsilon}$. Setting $\eta \leq \frac{\epsilon}{3C_0\sqrt{2L\max\{2L, C_2\}}}$ and choosing $\rho = \frac{\epsilon}{3C_0}$, we obtain

$$
\begin{aligned}
&\mathbb{E}[f(x_{T+1}) - f(x_1)] \\
&\leq -\frac{\eta}{2C_0} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{T\eta^2 L \sigma^2}{n\epsilon} + \frac{3T\eta^3 L C_0 C_2 \sigma^2}{\epsilon^2} + \frac{12T\eta^3 q^2 L C_0 \sigma_g^2}{(1-q^2)^2 \epsilon^2} \\
&\qquad + \frac{\eta(1+C_1)G^2 d}{\sqrt{\epsilon}} + \frac{\eta^2(1+2C_1)C_1 L G^2 d}{\epsilon}.
\end{aligned}
$$

where the last inequality follows from Lemma 5. Re-arranging terms, we get that

$$
\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2] &\leq 2C_0 \Big(\frac{\mathbb{E}[f(x_1) - f(x_{T+1})]}{T\eta} + \frac{\eta L \sigma^2}{n\epsilon} + \frac{3\eta^2 L C_0 C_2 \sigma^2}{\epsilon^2} \\
&\qquad + \frac{12\eta^2 q^2 L C_0 \sigma_g^2}{(1-q^2)^2 \epsilon^2} + \frac{(1+C_1)G^2 d}{T\sqrt{\epsilon}} + \frac{\eta(1+2C_1)C_1 L G^2 d}{T\epsilon} \Big) \\
&\leq 2C_0 \Big(\frac{\mathbb{E}[f(\theta_1) - f(\theta^*)]}{T\eta} + \frac{\eta L \sigma^2}{n\epsilon} + \frac{3\eta^2 L C_0 C_1 \sigma^2}{\epsilon^2} \\
&\qquad + \frac{12\eta^2 q^2 L C_0 \sigma_g^2}{(1-q^2)^2 \epsilon^2} + \frac{(1+C_1)G^2 d}{T\sqrt{\epsilon}} + \frac{\eta(1+2C_1)C_1 L G^2 d}{T\epsilon} \Big),
\end{aligned}
$$

where $C_0 = \sqrt{\frac{4(1+q^2)^3}{(1-q^2)^2} G^2 + \epsilon}$, $C_1 = \frac{\beta_1}{1-\beta_1} + \frac{2q}{1-q^2}$. The last inequality is because $\theta'_1 = \theta_1$, $\theta^* := \arg\min_\theta f(\theta)$ and the fact that $C_2 \leq C_1$. This completes the proof. $\square$

Proofs of Corollary 2 and Corollary 1 follow naturally from the above.

B.2 Intermidiary Lemmas

Lemma 1. *Under Assumption 1 to Assumption 4 we have:*

$$\sum_{t=1}^{T}\mathbb{E}\|\bar{m}_t'\|^2 \leq T\sigma^2 + \sum_{\tau=1}^{t}\mathbb{E}[\|\nabla f(\theta_t)\|^2].$$

Proof. Firstly, the expected squared norm of average stochastic gradient can be bounded by

$$\begin{aligned}\mathbb{E}[\|\bar{g}_t^2\|] &= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} g_{t,i} - \nabla f(\theta_t) + \nabla f(\theta_t)\|^2] \\ &= \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}(g_{t,i} - \nabla f_i(\theta_t))\|^2] + \mathbb{E}[\|\nabla f(\theta_t)\|^2] \\ &\leq \sigma^2 + \mathbb{E}[\|\nabla f(\theta_t)\|^2],\end{aligned}$$

where we use Assumption 4 that $g_{t,i}$ is unbiased and has bounded variance. Let $\bar{g}_{t,i}$ denote the i-th coordinate of $\bar{g}_t$. By the updating rule of COMP-AMS

$$\begin{aligned}\mathbb{E}[\|\bar{m}_t'\|^2] &= \mathbb{E}[\|(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{g}_\tau\|^2] \\ &\leq (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}[(\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{g}_{\tau,i})^2] \\ &\overset{(a)}{\leq} (1-\beta_1)^2\sum_{i=1}^{d}\mathbb{E}[(\sum_{\tau=1}^{t}\beta_1^{t-\tau})(\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{g}_{\tau,i}^2)] \\ &\leq (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\mathbb{E}[\|\bar{g}_\tau\|^2] \\ &\leq \sigma^2 + (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\mathbb{E}[\|\nabla f(\theta_t)\|^2],\end{aligned}$$

where (a) is due to Cauchy-Schwartz inequality. Summing over $t = 1, ..., T$, we obtain

$$\sum_{t=1}^{T}\mathbb{E}\|\bar{m}_t'\|^2 \leq T\sigma^2 + \sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\theta_t)\|^2].$$

This completes the proof.

$\square$

Lemma 2. *Under Assumption 4, we have for $\forall t$ and each local worker $\forall i \in [n]$,*

$$\begin{aligned}&\|e_{t,i}\|^2 \leq \frac{4q^2}{(1-q^2)^2}G^2, \\ &\mathbb{E}[\|e_{t+1,i}\|^2] \leq \frac{4q^2}{(1-q^2)^2}\sigma^2 + \frac{2q^2}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}[\|\nabla f_i(\theta_\tau)\|^2].\end{aligned}$$

Proof. We start by using Assumption 1 and Young's inequality to get

$$\begin{aligned}\|e_{t+1,i}\|^2 &= \|g_{t,i} + e_{t,i} - \mathcal{C}(g_{t,i} + e_{t,i})\|^2 \\ &\leq q^2\|g_{t,i} + e_{t,i}\|^2 \\ &\leq q^2(1+\rho)\|e_{t,i}\|^2 + q^2(1+\frac{1}{\rho})\|g_{t,i}\|^2 \\ &\leq \frac{1+q^2}{2}\|e_{t,i}\|^2 + \frac{2q^2}{1-q^2}\|g_{t,i}\|^2, \qquad (9)\end{aligned}$$

by choosing $\rho = \frac{1-q^2}{2q^2}$. Now by recursion and the initialization $e_{1,i} = 0$, we have

$$\mathbb{E}[\|e_{t+1,i}\|^2] \leq \frac{2q^2}{1-q^2} \sum_{\tau=1}^{t} (\frac{1+q^2}{2})^{t-\tau} \mathbb{E}[\|g_{\tau,i}\|^2]$$
$$\leq \frac{4q^2}{(1-q^2)^2}\sigma^2 + \frac{2q^2}{1-q^2} \sum_{\tau=1}^{t} (\frac{1+q^2}{2})^{t-\tau} \mathbb{E}[\|\nabla f_i(\theta_\tau)\|^2],$$

which proves the second argument. Meanwhile, the absolute bound $\|e_{t,i}\|^2 \leq \frac{4q^2}{(1-q^2)^2} G^2$ follows directly from (9). □

Lemma 3. *For the moving average error sequence $\mathcal{E}_t$, it holds that*

$$\sum_{t=1}^{T} \mathbb{E}[\|\mathcal{E}_t\|^2] \leq \frac{4Tq^2}{(1-q^2)^2}(\sigma^2 + \sigma_g^2) + \frac{4q^2}{(1-q^2)^2} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2].$$

Proof. Let $\bar{e}_{t,i}$ be the j-th coordinate of $\bar{e}_t$. Denote $K_{t,i} := \sum_{\tau=1}^{t} (\frac{1+q^2}{2})^{t-\tau} \mathbb{E}[\|\nabla f_i(\theta_\tau)\|^2]$ and $K_{t,i} = 0, \forall i \in [n]$. Using the same technique as in the proof of Lemma 1, we have

$$\mathbb{E}[\|\mathcal{E}_t\|^2] = \mathbb{E}[\|(1-\beta_1)\sum_{\tau=1}^{t} \beta_1^{t-\tau} \bar{e}_\tau\|^2]$$
$$\leq (1-\beta_1)^2 \sum_{j=1}^{d} \mathbb{E}[(\sum_{\tau=1}^{t} \beta_1^{t-\tau} \bar{e}_{\tau,j})^2]$$
$$\overset{(a)}{\leq} (1-\beta_1)^2 \sum_{j=1}^{d} \mathbb{E}[(\sum_{\tau=1}^{t} \beta_1^{t-\tau})(\sum_{\tau=1}^{t} \beta_1^{t-\tau} \bar{e}_{\tau,j}^2)]$$
$$\leq (1-\beta_1) \sum_{\tau=1}^{t} \beta_1^{t-\tau} \mathbb{E}[\|\bar{e}_\tau\|^2]$$
$$\leq (1-\beta_1) \sum_{\tau=1}^{t} \beta_1^{t-\tau} \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \|e_{\tau,i}\|^2]$$
$$\overset{(b)}{\leq} \frac{4q^2}{(1-q^2)^2}\sigma^2 + \frac{2q^2(1-\beta_1)}{(1-q^2)} \sum_{\tau=1}^{t} \beta_1^{t-\tau} (\frac{1}{n}\sum_{i=1}^{n} K_{\tau,i}),$$

where (a) is due to Cauchy-Schwartz and (b) is a result of Lemma 2. Summing over $t = 1, ..., T$ and using the technique of geometric series summation leads to

$$\sum_{t=1}^{T} \mathbb{E}[\|\mathcal{E}_t\|^2] = \frac{4Tq^2}{(1-q^2)^2}\sigma^2 + \frac{2q^2(1-\beta_1)}{(1-q^2)} \sum_{t=1}^{T}\sum_{\tau=1}^{t} \beta_1^{t-\tau} (\frac{1}{n}\sum_{i=1}^{n} K_{\tau,i})$$
$$\leq \frac{4Tq^2}{(1-q^2)^2}\sigma^2 + \frac{2q^2}{(1-q^2)} \sum_{t=1}^{T}\sum_{\tau=1}^{t} (\frac{1+q^2}{2})^{t-\tau} \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\theta_\tau)\|^2]$$
$$\leq \frac{4Tq^2}{(1-q^2)^2}\sigma^2 + \frac{4q^2}{(1-q^2)^2} \sum_{t=1}^{T} \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\theta_t)\|^2]$$
$$\overset{(a)}{\leq} \frac{4Tq^2}{(1-q^2)^2}\sigma^2 + \frac{4q^2}{(1-q^2)^2} \sum_{t=1}^{T} \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta_t)\|^2 + \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\theta_t) - \nabla f(\theta_t)\|^2]$$
$$\leq \frac{4Tq^2}{(1-q^2)^2}(\sigma^2 + \sigma_g^2) + \frac{4q^2}{(1-q^2)^2} \sum_{t=1}^{T} \mathbb{E}[\|\nabla f(\theta_t)\|^2],$$

where (a) is derived by the variance decomposition and the last inequality holds due to Assumption 4. The desired result is obtained.

$\square$

Lemma 4. *It holds that* $\forall t \in [T]$, $\forall i \in [d]$, $\hat{v}_{t,i} \leq \frac{4(1+q^2)^3}{(1-q^2)^2} G^2$.

Proof. For any t, by Lemma 2 and Assumption 3 we have

$$\begin{aligned}
\|\tilde{g}_t\|^2 &= \|\mathcal{C}(g_t + e_t)\|^2 \\
&\leq \|\mathcal{C}(g_t + e_t) - (g_t + e_t) + (g_t + e_t)\|^2 \\
&\leq 2(q^2 + 1)\|g_t + e_t\|^2 \\
&\leq 4(q^2 + 1)(G^2 + \frac{4q^2}{(1-q^2)^2} G^2) \\
&= \frac{4(1+q^2)^3}{(1-q^2)^2} G^2.
\end{aligned}$$

It's then easy to show by the updating rule of $\hat{v}_t$,

$$\hat{v}_{t,i} = (1 - \beta_2) \sum_{\tau=1}^{t} \beta_2^{t-\tau} \tilde{g}_{t,i}^2 \leq \frac{4(1+q^2)^3}{(1-q^2)^2} G^2.$$

$\square$

Lemma 5. *Let* $D_t := \frac{1}{\sqrt{\hat{v}_{t-1} + \epsilon}} - \frac{1}{\sqrt{\hat{v}_t + \epsilon}}$ *be defined as above. Then,*

$$\sum_{t=1}^{T} \|D_t\|_1 \leq \frac{d}{\sqrt{\epsilon}}, \quad \sum_{t=1}^{T} \|D_t\|^2 \leq \frac{d}{\epsilon}.$$

Proof. By the updating rule of COMP-AMS, $\hat{v}_{t-1} \leq \hat{v}_t$ for $\forall t$. Therefore, by the initialization $\hat{v}_0 = 0$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \|D_t\|_1 &= \sum_{t=1}^{T} \sum_{i=1}^{d} (\frac{1}{\sqrt{\hat{v}_{t-1,i} + \epsilon}} - \frac{1}{\sqrt{\hat{v}_{t,i} + \epsilon}}) \\
&= \sum_{i=1}^{d} (\frac{1}{\sqrt{\hat{v}_{0,i} + \epsilon}} - \frac{1}{\sqrt{\hat{v}_{T,i} + \epsilon}}) \\
&\leq \frac{d}{\sqrt{\epsilon}}.
\end{aligned}$$

For the sum of squared l_2 norm, note the fact that for $a \geq b > 0$, it holds that

$$(a - b)^2 \leq (a - b)(a + b) = a^2 - b^2.$$

Thus,

$$\begin{aligned}
\sum_{t=1}^{T} \|D_t\|^2 &= \sum_{t=1}^{T} \sum_{i=1}^{d} (\frac{1}{\sqrt{\hat{v}_{t-1,i} + \epsilon}} - \frac{1}{\sqrt{\hat{v}_{t,i} + \epsilon}})^2 \\
&\leq \sum_{t=1}^{T} \sum_{i=1}^{d} (\frac{1}{\hat{v}_{t-1,i} + \epsilon} - \frac{1}{\hat{v}_{t,i} + \epsilon}) \\
&\leq \frac{d}{\epsilon},
\end{aligned}$$

which gives the desired result. $\square$

C Model Architecture of the Experiments

In Figure 4, we provide the detailed description of the data and model architectures used in our numerical study. MNIST [37] is a popular hand-written letter recognition dataset, where each training sample is a 28×28 black and white image belonging to a class (digits 0-9). CIFAR-10 [36] is a benchmark image classification dataset consisting of natural images from 10 classes. The image size is a $3 \times 32 \times 32$. In IMDB dataset [41], each sample is a movie review, and the task is to classify the reviews as positive or negative. The reviews are tokenized by words and transformed into integer vectors. We threshold at 300 for the length of each review. Zero-padding is applied to reviews that have less than 300 words. All our experiments are trained on a Linux server equipped with four Nvidia Tesla V100 cards. We use two Convolutional Neural Networks (CNN) for MNIST and CIFAR-10. For IMDB dataset, we use a LSTM network. For all three models, ReLu activation is adopted. For LSTM, each input movie review is a 300-dimensional vector, and the embedding layer embeds top 1000 most frequent words into 32-dimensional vectors. 64 LSTM cells are used, where the last hidden state is connected to two fully connected layers before the output.

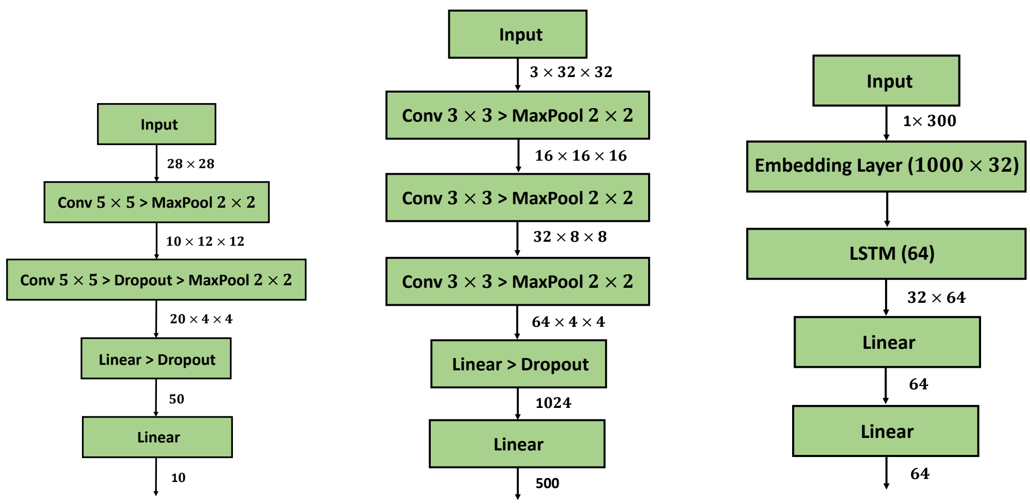

Figure 4: Model architectures used in the experiments. Left: MNIST + CNN. Middle: CIFAR-10 + CNN. Right: IMDB + LSTM. In the last figure, the penultimate linear layer takes the last hidden state (64-dim vector) as the input.