BRAM: Communication-Efficient 1-bit Adaptive Optimizer for Practical Distributed DNN Training

Anonymous Author(s)
Affiliation
Address
email

A Theoretical Analysis for Algorithm 1

In practice, we implement BRAM in a non-parameter-server model to further reduce the communication overhead, but the data exchange is essentially equivalent to that in a parameter-server prototype. Hence, we provide the theoretical analysis for BRAM in a parameter-server model as shown in Algorithm 1.

According to Algorithm 1, the update $\bar{u}_t$ can be recursively formulated as

$$\begin{aligned}
\bar{u}_t =& \mathcal{Q}\left(\frac{1}{n}\sum_{i=1}^{n} u_t^{(i)} + \bar{e}_t\right) \\
=& \frac{1}{n}\sum_{i=1}^{n} u_t^{(i)} + \bar{e}_t - \bar{e}_{t+1} \\
=& \frac{1}{n}\sum_{i=1}^{n} \mathcal{Q}\left(\frac{m_t^{(i)}}{b_t^{(i)}} + e_t^{(i)}\right) + \bar{e}_t - \bar{e}_{t+1} \\
=& \frac{1}{n}\sum_{i=1}^{n} \left(\frac{m_t^{(i)}}{b_t^{(i)}} + e_t^{(i)} - e_{t+1}^{(i)}\right) + \bar{e}_t - \bar{e}_{t+1} \\
=& \frac{1}{n}\sum_{i=1}^{n} \frac{m_t^{(i)}}{b_t^{(i)}} + \frac{1}{n}\sum_{i=1}^{n}\left(e_t^{(i)} - e_{t+1}^{(i)}\right) + \bar{e}_t - \bar{e}_{t+1}
\end{aligned} \tag{1}$$

Denote

$$g_t \triangleq \frac{1}{n}\sum_{i=1}^{n} g_t^{(i)}, \tag{2}$$

$$m_t \triangleq \frac{1}{n}\sum_{i=1}^{n} m_t^{(i)} = \beta m_{t-1} + (1-\beta) g_t, \tag{3}$$

$$b_t \triangleq \frac{1}{n}\sum_{i=1}^{n} b_t^{(i)}, \tag{4}$$

$$\delta_t \triangleq \frac{1}{n}\sum_{i=1}^{n} \frac{m_t^{(i)}}{b_t^{(i)}} - \frac{m_t}{b_t}, \tag{5}$$

$$e_t \triangleq \frac{1}{n}\sum_{i=1}^{n} e_t^{(i)} + \bar{e}_t \tag{6}$$

$$\tag{7}$$

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Hence, the updating rule can be summarized as

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t \bar{u}_t \\ &= x_t - \alpha_t \left(\frac{m_t}{b_t} + \delta_t + e_t - e_{t+1} \right) \end{aligned} \tag{8}$$

A.1 Auxiliary Lemmas

Lemma 1. *Let $u_t = \frac{m_t}{b_t}$, the element-wise quantization function is defined in Eq.(5) can be reformulated as*

$$\mathcal{Q}\left((u_t)_j\right) = \begin{cases} 1, & \text{with probability } p = \frac{(u_t)_j + 1}{2} \\ -1, & \text{with probability } 1 - p \end{cases} \quad (j \in \{1, 2, ..., d\}, \quad -1 \leq (u_t)_j \leq 1). \tag{9}$$

We have $e_t = u_t - \mathcal{Q}(u_t)$, and then the following holds true

$$\mathbb{E}[e_t] = 0, \ \ \mathbb{E}\left[\|e_t\|^2\right] \leq d. \tag{10}$$

Proof. From Eq.(9), we know

$$\begin{aligned} \mathbb{E}\left[(e_t)_j\right] &= \mathbb{E}\left[u_t - \mathcal{Q}(u_t)\right] \\ &= \frac{1}{2}\left((u_t)_j + 1\right)\left((u_t)_j - 1\right) + (1 - \frac{1}{2}((u_t)_j + 1))((u_t)_j + 1) = 0, \end{aligned} \tag{11}$$

and,

$$\begin{aligned} \mathbb{E}\left[(e_t)_j^2\right] &= \mathbb{E}\left[\left((u_t)_j - \mathcal{Q}\left((u_t)_j\right)\right)^2\right] \\ &= \frac{1}{2}\left((u_t)_j + 1\right)\left((u_t)_j - 1\right)^2 + (1 - \frac{1}{2}((u_t)_j + 1))((u_t)_j + 1)^2 \\ &= 1 - \left((u_t)_j\right)^2 \leq 1. \end{aligned} \tag{12}$$

Hence,

$$\mathbb{E}[e_t] = 0, \ \ \mathbb{E}\left[\|e_t\|^2\right] \leq d. \tag{13}$$

Lemma 2. *Let $x_0 = x_1$ and $\alpha_0 = \alpha_1$ in Algorithm 1, defining the sequence*

$$z_1 = x_1 + \alpha_1(\delta_1 - e_1) \tag{14}$$

$$z_t = x_t + \frac{\beta}{1-\beta}(x_t - x_{t-1}) + \frac{\alpha_{t-1}}{1-\beta}(\delta_{t-1} + \beta e_{t-1} - e_t), \forall t \geq 2. \tag{15}$$

Then the following equality will hold, i.e.,

$$z_{t+1} = z_t + \frac{\beta}{1-\beta}\left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1} - \alpha_t \frac{g_t}{b_t} - \alpha_{t-1}\delta_{t-1} - (\alpha_t - \alpha_{t-1})e_t. \tag{16}$$

Proof. For $t = 1$, we have

$$\begin{aligned} z_2 - z_1 &= x_2 + \frac{\beta}{1-\beta}(x_2 - x_1) + \frac{\alpha_1}{1-\beta}(\delta_1 + \beta e_1 - e_2) - (x_1 + \alpha_1(\delta_1 - e_1)) \\ &= (\frac{\beta}{1-\beta} + 1)(x_2 - x_1) + \frac{\alpha_1}{1-\beta}(\delta_1 + \beta e_1 - e_2) - \alpha_1(\delta_1 - e_1) \\ &= -\frac{\alpha_1}{1-\beta}\left(\frac{(1-\beta)g_1}{b_1} + \delta_1 + e_1 - e_2\right) + \frac{\alpha_1}{1-\beta}(\delta_1 + \beta e_1 - e_2) - \alpha_1(\delta_1 - e_1) \\ &= -\alpha_1 \frac{g_1}{b_1} - \alpha_0 \delta_1 \end{aligned} \tag{17}$$

where the second equality follows the updating rule in Eq.(8).

For $t \geq 2$, following the updating rule in Eq.(8), we have

$$\begin{aligned} x_{t+1} - x_t + \alpha_t(\delta_t + e_t - e_{t+1}) =& -\alpha_t \frac{m_t}{b_t} \\ =& -\alpha_t \frac{\beta m_{t-1} + (1-\beta) g_t}{b_t} \\ =& \beta \left(x_t - x_{t-1} + \alpha_{t-1}(\delta_t + e_{t-1} - e_t) \right) \\ & + \beta \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t} \right) \odot m_{t-1} - (1-\beta)\alpha_t \frac{g_t}{b_t} \end{aligned} \tag{18}$$

We know $x_{t+1} - x_t + \alpha_t(e_t - e_{t+1}) = (1-\beta)(x_{t+1} + -\alpha_t(e_{t+1} - \delta_t)) - (1-\beta)(x_t - \alpha_t e_t) + \beta(x_{t+1} - x_t + \alpha_t(\delta_t + e_t - e_{t+1}))$, so Eq. (18) can be rearranged as

$$\begin{aligned} & (1-\beta)(x_{t+1} + \alpha_t(\delta_t - e_{t+1})) + \beta(x_{t+1} - x_t + \alpha_t(\delta_t + e_t - e_{t+1})) \\ =& (1-\beta)(x_t - \alpha_t e_t) + \beta \left(x_t - x_{t-1} + \alpha_{t-1}(\delta_{t-1} + e_{t-1} - e_t) \right) \\ & + \beta \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t} \right) \odot m_{t-1} - (1-\beta)\alpha_t \frac{g_t}{b_t} \end{aligned} \tag{19}$$

Divided both sides by $1-\beta$, we obtain

$$\begin{aligned} & x_{t+1} + \alpha_t(\delta_t - e_{t+1}) + \frac{\beta}{1-\beta}(x_{t+1} - x_t + \alpha_t(\delta_t + e_t - e_{t+1})) \\ =& x_t + \alpha_{t-1}(\delta_{t-1} - e_t) + \frac{\beta}{1-\beta} \left(x_t - x_{t-1} + \alpha_{t-1}(\delta_{t-1} + e_{t-1} - e_t) \right) \\ & + \frac{\beta}{1-\beta} \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t} \right) \odot m_{t-1} \\ & - \alpha_t \frac{g_t}{b_t} - \alpha_{t-1}\delta_{t-1} - (\alpha_t - \alpha_{t-1}) e_t \end{aligned} \tag{20}$$

Rearranging Eq. (20), we have

$$\begin{aligned} & x_{t+1} + \frac{\beta}{1-\beta}(x_{t+1} - x_t) + \frac{\alpha_t}{1-\beta}(\delta_t + \beta e_t - e_{t+1}) \\ =& x_t + \frac{\beta}{1-\beta}(x_t - x_{t-1}) + \frac{\alpha_{t-1}}{1-\beta}(\delta_{t-1} + \beta e_{t-1} - e_t) \\ & + \frac{\beta}{1-\beta} \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t} \right) \odot m_{t-1} \\ & - \alpha_t \frac{g_t}{b_t} - \alpha_{t-1}\delta_{t-1} - (\alpha_t - \alpha_{t-1}) e_t \end{aligned} \tag{21}$$

Define the sequence

$$z_t = x_t + \frac{\beta}{1-\beta}(x_t - x_{t-1}) + \frac{\alpha_{t-1}}{1-\beta}(\delta_{t-1} + \beta e_{t-1} - e_t) \tag{22}$$

We finally obtain

$$z_{t+1} = z_t + \frac{\beta}{1-\beta} \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t} \right) \odot m_{t-1} - \alpha_t \frac{g_t}{b_t} - \alpha_{t-1}\delta_{t-1} - (\alpha_t - \alpha_{t-1}) e_t. \tag{23}$$

Recalling $x_1 = x_0$ and $\alpha_1 = \alpha_0$, we have $\frac{\alpha_1}{b_1} = \frac{\alpha_0}{b_0}$. Then, combining Eq.(17) and Eq.(23), we obtain the conclusion.

A.2 Proof of Theorem 1

Proof. By the the gradient Lipschitz continuous in Assumption 2 and Lemma 2, we obtain

$$
\begin{aligned}
\mathbb{E}[f(z_{t+1}) - f(z_t)] \leq & \mathbb{E}\langle \nabla f(z_t), z_{t+1} - z_t\rangle + \frac{L}{2}\mathbb{E}\|z_{t+1} - z_t\|^2 \\
=& \mathbb{E}\left[\frac{\beta}{1-\beta}\langle \nabla f(z_t), \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1}\rangle\right] - \mathbb{E}\left[\langle \nabla f(z_t), \alpha_t \frac{g_t}{b_t}\rangle\right] \\
& - \mathbb{E}\left[\langle \nabla f(z_t), \alpha_{t-1}\delta_{t-1}\rangle\right] - \mathbb{E}\left[\langle \nabla f(z_t), (\alpha_t - \alpha_t)e_{t-1}\rangle\right] \\
& + \mathbb{E}\left[\frac{L}{2}\left\|\frac{\beta}{1-\beta}\left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1} - \alpha_t\frac{g_t}{b_t} - \alpha_{t-1}\delta_{t-1} - (\alpha_t - \alpha_{t-1})e_{t-1}\right\|^2\right] \\
=& \mathbb{E}\left[\frac{\beta}{1-\beta}\langle \nabla f(z_t), \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1}\rangle\right] - \mathbb{E}\left[\langle \nabla f(z_t), \alpha_t \frac{g_t}{b_t}\rangle\right] \\
& + \mathbb{E}\left[\frac{L}{2}\left\|\frac{\beta}{1-\beta}\left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1} - \alpha_t\frac{g_t}{b_t} - \alpha_{t-1}\delta_{t-1} - (\alpha_t - \alpha_{t-1})e_{t-1}\right\|^2\right] \\
\leq & \mathbb{E}\left[\frac{\beta}{1-\beta}\langle \nabla f(z_t), \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1}\rangle\right] - \mathbb{E}\left[\langle \nabla f(z_t), \alpha_t \frac{g_t}{b_t}\rangle\right] \\
& + L\mathbb{E}\left[\left\|\frac{\beta}{1-\beta}\left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1}\right\|^2\right] + L\mathbb{E}\left[\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
& + \frac{L}{2}\mathbb{E}\left[\|\alpha_{t-1}\delta_{t-1}\|^2\right] + \frac{L}{2}\mathbb{E}\left[\|(\alpha_{t-1} - \alpha_t)e_t\|^2\right]
\end{aligned}
\tag{24}
$$

where the second equality holds due to $\mathbb{E}\left[\delta_{t-1}\right] = 0$ and $\mathbb{E}\left[e_{t-1}\right] = 0$. The last inequality holds owing to $\mathbb{E}[\|a+b\|^2] = \mathbb{E}[\|a\|^2] + \mathbb{E}[\|b\|^2]$ if $\mathbb{E}[a] = 0$ or $\mathbb{E}[b] = 0$, and $\mathbb{E}[\|a+b\|^2] \leq 2\mathbb{E}[\|a\|^2] + 2\mathbb{E}[\|b\|^2]$ if $\mathbb{E}[a] \neq 0$ and $\mathbb{E}[b] \neq 0$.

Taking telescope sum from 1 to T on the both sides of Eq.(24) , we then have

$$
\begin{aligned}
\mathbb{E}[f(z_T) - f(z_1)] \leq & \underbrace{\frac{\beta}{1-\beta}\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla f(z_t), \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1}\rangle\right]}_{T_1} \underbrace{-\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla f(z_t), \alpha_t \frac{g_t}{b_t}\rangle\right]}_{T_2} \\
& + \underbrace{L\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\beta}{1-\beta}\left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{t-1}\right\|^2\right]}_{T_3} \\
& + \underbrace{L\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] + \frac{L}{2}\mathbb{E}[\sum_{t=1}^{T}\|\alpha_{t-1}\delta_{t-1}\|^2] + \frac{L}{2}\mathbb{E}[\sum_{t=1}^{T}\|(\alpha_{t-1} - \alpha_t)e_t\|^2]}_{T_4}
\end{aligned}
\tag{25}
$$

Now we focus on bounding T_1 below. From Assumption 4, we know $\|g_t\| \leq G$ $(t = 1, 2, ..., T)$ and $\|\nabla f(z_t)\| \leq G$. Due to $m_t = \beta m_{t-1} + (1-\beta)g_t$ and $m_1 = g_1$, it is easy to obtain $\|m_t\| \leq G$ by complete induction.

Since $\|\nabla f(z_t)\| \leq G$ and $\|m_t\| \leq G$, we have

$$
\begin{aligned}
T_1 &= \frac{\beta}{1-\beta}\mathbb{E}\left[\sum_{i=1}^{T}\langle \nabla f(z_i), \left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right) \odot m_{i-1}\rangle\right] \\
&\overset{(i)}{\leq} \frac{\beta}{1-\beta}\mathbb{E}\left[\sum_{i=1}^{T}\|\nabla f(z_t)\|\|m_t\|\left\|\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right\|_1\right] \\
&\overset{(ii)}{\leq} \frac{\beta}{1-\beta}G^2\mathbb{E}\left[\sum_{i=1}^{T}\left\|\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right\|_1\right] \\
&\overset{(iii)}{=} \frac{\beta}{1-\beta}G^2\mathbb{E}\left[\left\|\sum_{i=1}^{T}\left(\frac{\alpha_{t-1}}{b_{t-1}} - \frac{\alpha_t}{b_t}\right)\right\|_1\right] \\
&\leq \frac{\beta}{1-\beta}G^2\mathbb{E}\left[\left\|\frac{\alpha_0}{b_0}\right\|_1\right] \\
&\overset{(iv)}{\leq} \frac{\alpha_0\beta d}{(1-\beta)\rho}G^2,
\end{aligned}
\tag{26}
$$

where (i) holds sice $\|a \odot b\| \leq \|a\| \max_j |(b)_j| \leq \|a\|\|b\|_1$, (ii) holds due to $\|\nabla f(z_t)\| \leq G$ and $\|m_t\| \leq G$, (iii) holds because $\frac{\alpha_{t-1}}{(b_{t-1})_j} - \frac{\alpha_t}{(b_t)_j} \geq 0$ for any $j \in [1, 2, ..., d]$, (iv) holds due to $\min_j (b_t)_j \geq \rho > 0$ for any $j \in [1, 2, ..., d]$.

Let us turn to bound T_2,

$$
\begin{aligned}
T_2 &= -\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla f(z_t), \alpha_t \frac{g_t}{b_t}\rangle\right] \\
&= \underbrace{-\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla f(z_t) - f(x_t), \alpha_t \frac{g_t}{b_t}\rangle\right]}_{T_5} \underbrace{-\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla f(x_t), \alpha_t \frac{g_t}{b_t}\rangle\right]}_{T_6}
\end{aligned}
\tag{27}
$$

We now analyze T_5 below,

$$
\begin{aligned}
T_5 &= -\mathbb{E}\left[\sum_{t=1}^{T}\langle \nabla f(z_t) - f(x_t), \alpha_t \frac{g_t}{b_t}\rangle\right] \\
&\overset{(i)}{\leq} \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|f(z_t) - f(x_t)\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&\overset{(ii)}{\leq} \frac{L^2}{2}\mathbb{E}\left[\sum_{t=1}^{T}\|z_t - x_t\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&\overset{(iii)}{=} \frac{L^2}{2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\beta}{1-\beta}(x_t - x_{t-1}) + \frac{\alpha_{t-1}}{1-\beta}\left(\delta_{t-1} + \beta e_{t-1} - e_t\right)\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&\overset{(iv)}{\leq} \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|x_t - x_{t-1}\|^2\right] + \frac{L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}\delta_{t-1}\|^2\right] \\
&\quad + \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_{t-1}\|^2\right] + \frac{L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_t\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&\overset{(v)}{=} \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_{t-1}^2\left\|\left(\frac{m_{t-1}}{b_{t-1}} + \delta_{t-1} + e_{t-1} - e_t\right)\right\|^2\right] + \frac{L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}\delta_{t-1}\|^2\right] \\
&\quad + \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_{t-1}\|^2\right] + \frac{L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_t\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&= \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_{t-1}m_{t-1}}{b_{t-1}}\right\|^2\right] + \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}\delta_{t-1}\|^2\right] \\
&\quad + \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_{t-1}\|^2\right] + \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_t\|^2\right] + \frac{L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}\delta_{t-1}\|^2\right] \\
&\quad + \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_{t-1}\|^2\right] + \frac{L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\|\alpha_{t-1}e_t\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&= \frac{\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_{t-1}^2\left\|\frac{m_{t-1}}{b_{t-1}}\right\|^2\right] + \frac{(1+\beta^2)L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_{t-1}^2\|\delta_{t-1}\|^2\right] \\
&\quad + \frac{2\beta^2 L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_{t-1}^2\|e_{t-1}\|^2\right] + \frac{(1+\beta^2)L^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_{t-1}^2\|e_t\|^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right] \\
&\overset{(vi)}{\leq} \left(\frac{\beta^2 L^2 d}{(1-\beta)^2} + \frac{4(1+\beta^2)L^2 d}{(1-\beta)^2} + \frac{2\beta^2 L^2 d}{(1-\beta)^2} + \frac{(1+\beta^2)L^2 d}{(1-\beta)^2} + \frac{G^2}{2\rho^2}\right)\sum_{t=1}^{T}\alpha_{t-1}^2 \\
&= \left(\frac{(8\beta^2 + 10\beta + 5)L^2 d}{(1-\beta)^2} + \frac{G^2}{2\rho^2}\right)\sum_{t=1}^{T}\alpha_{t-1}^2
\end{aligned}
\tag{28}
$$

where (i) holds by following $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|a\|^2$, (ii) holds due to Assumption 1, (iii) holds due to Assumption 1 owing to Eq.(15), (iii) holds since $\mathbb{E}[\|a+b\|^2] = \mathbb{E}[\|a\|^2] + \mathbb{E}[\|b\|^2]$ if $\mathbb{E}[a] = 0$ or $\mathbb{E}[b] = 0$, (v) holds resulting from the updating rule in Eq. (8), (vi) holds due to $\left|\frac{(m_t)_j}{(b_t)_j}\right| \leq 1$, $|(\delta)_j| \leq 2$ (the definition of $\delta_t$in Eq. (5)), $\mathbb{E}[\|e_t\|^2] \leq d$ in Lemma 1, $\|g_t\| \leq G$ in Assumption 2 and $\min_j (b_t)_j \geq \rho > 0$.

We then bound T_6

$$
\begin{aligned}
T_6 =& -\mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla f(x_t), \alpha_t\frac{g_t}{b_t}\rangle\right] \\
=& -\mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla f(x_t), \alpha_t\frac{\nabla f(x_t)}{b_t}\rangle\right] - \mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla f(x_t), \alpha_t\frac{g_t-\nabla f(x_t)}{b_t}\rangle\right] \\
\overset{(i)}{\leq}& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + \mathbb{E}\left[\sum_{t=1}^{T}\langle\nabla f(x_t), \alpha_t\frac{\nabla f(x_t)-g_t}{b_t}\rangle\right] \\
=& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + \mathbb{E}\left[\langle\nabla f(x_1), \alpha_1\frac{\nabla f(x_1)-g_1}{b_1}\rangle\right] \\
&+ \mathbb{E}\left[\sum_{t=2}^{T}\langle\nabla f(x_t), \nabla(f(x_t)-g_t)\odot\left(\frac{\alpha_t}{b_t}-\frac{\alpha_{t-1}}{b_{t-1}}\right)\rangle\right] + \mathbb{E}\left[\sum_{t=2}^{T}\langle\nabla f(x_t), \alpha_{t-1}\frac{\nabla f(x_t)-g_t}{b_{t-1}}\rangle\right] \\
\overset{(ii)}{=}& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + \mathbb{E}\left[\langle\nabla f(x_1), \alpha_1\frac{\nabla f(x_1)-g_1}{b_1}\rangle\right] \\
&+ \mathbb{E}\left[\sum_{t=2}^{T}\langle\nabla f(x_t), (\nabla f(x_t)-g_t)\odot\left(\frac{\alpha_t}{b_t}-\frac{\alpha_{t-1}}{b_{t-1}}\right)\rangle\right] \\
\overset{(iii)}{\leq}& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + \mathbb{E}\left[\|\nabla f(x_1)\|\|\nabla f(x_1)-g_1\|\left\|\frac{\alpha_1}{b_1}\right\|_1\right] \\
&+ \mathbb{E}\left[\sum_{t=2}^{T}\|\nabla f(x_t)\|\|\nabla f(x_t)-g_t\|\left\|\frac{\alpha_t}{b_t}-\frac{\alpha_{t-1}}{b_{t-1}}\right\|_1\right] \\
\overset{(iv)}{\leq}& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + 2G^2\mathbb{E}\left[\left\|\frac{\alpha_1}{b_1}\right\|_1 + \sum_{t=2}^{T}\left\|\frac{\alpha_t}{b_t}-\frac{\alpha_{t-1}}{b_{t-1}}\right\|_1\right] \\
\overset{(v)}{=}& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + 2G^2\mathbb{E}\left[\left\|\frac{\alpha_1}{b_1} + \sum_{t=2}^{T}\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right\|_1\right], \\
=& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + 4G^2\mathbb{E}\left[\left\|\frac{\alpha_1}{b_1}\right\|_1\right], \\
\overset{(vi)}{\leq}& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] + \frac{4G^2\alpha_1 d}{\rho}
\end{aligned}
\tag{29}
$$

where (i) holds due to $\max_j(b_t)_j \leq \|b_t\| \leq G$, (ii) holds owing to $\mathbb{E}[\nabla f(x_t) - g_t] = 0$ in Assumption 2 and g_t, b_{t-1} are independent, (iii) holds sice $\|a \odot b\| \leq \|a\| \max_j |(b)_j| \leq \|a\|\|b\|_1$, (iv) holds resulting from $\|\nabla f(x_t)\| \leq G$ and $\|\nabla f(x_t) - g_t\| \leq \|\nabla f(x_t)\| + \|g_t\| \leq 2G$, and (v) holds because $\frac{\alpha_{t-1}}{(b_{t-1})_j} - \frac{\alpha_t}{(b_t)_j} \geq 0$ for any $j \in [1, 2, ..., d]$, (vi) holds due to $\min_j(b_t)_j \geq \rho > 0$ for any $j \in [1, 2, ..., d]$.

Then, we pay attention to T_3,

$$
\begin{aligned}
T_3 &= L\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\beta}{1-\beta}\left(\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right)\odot m_{t-1}\right\|^2\right] \\
&\overset{(i)}{\leq} \frac{\beta^2 L}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right\|^2\|m_{t-1}\|^2\right] \\
&\overset{(ii)}{\leq} \frac{\beta^2 L G^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right\|^2\right] \\
&\overset{(iii)}{\leq} \frac{\beta^2 L G^2}{(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\max_j\left|\frac{\alpha_{t-1}}{(b_{t-1})_j}-\frac{\alpha_t}{(b_t)_j}\right|\left\|\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right\|_1\right] \\
&\overset{(iv)}{\leq} \frac{\alpha_0\beta^2 L G^2}{\rho(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\max_j\left(\frac{\alpha_{t-1}}{(b_{t-1})_j}\right)\left\|\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right\|_1\right] \\
&\overset{(v)}{\leq} \frac{\alpha_0\beta^2 L G^2}{\rho(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_{t-1}}{b_{t-1}}-\frac{\alpha_t}{b_t}\right\|_1\right] \\
&\overset{(vi)}{\leq} \frac{\alpha_0\beta^2 L G^2}{\rho(1-\beta)^2}\mathbb{E}\left[\sum_{t=1}^{T}\left\|\frac{\alpha_{t-1}}{b_{t-1}}\right\|_1-\left\|\frac{\alpha_t}{b_t}\right\|_1\right] \\
&\overset{(vii)}{\leq} \frac{\alpha_0\beta^2 L G^2}{\rho(1-\beta)^2}\mathbb{E}\left[\left\|\frac{\alpha_0}{b_0}\right\|_1-\left\|\frac{\alpha_T}{b_T}\right\|_1\right] \\
&\overset{(viii)}{\leq} \frac{\alpha_0^2\beta^2 L G^2 d}{\rho^2(1-\beta)^2},
\end{aligned}
\tag{30}
$$

where (i) holds due to $\|a\odot b\|\leq\|a\|\|b\|$, (ii) holds owing to $\|m_{t-1}\|\leq G$, (ii) holds due to $\|a\|^2\leq\max_j|(a)_j|\|a\|_1$, (iv) holds due to $\frac{\alpha_{t-1}}{(b_{t-1})_j}-\frac{\alpha_t}{(b_t)_j}\geq 0$ and $\frac{\alpha_t}{(b_t)_j}>0$ for any $j\in[1,2,...,d]$, (v) holds resulting from $\min_j(b_t)_j\geq\rho>0$ for any j and α_t is non-increasing, (vi) holds resulting from $\frac{\alpha_{t-1}}{(b_{t-1})_j}-\frac{\alpha_t}{(b_t)_j}\geq 0$ for any $j\in[1,2,...,d]$, (vii) holds due to telescoping sum, and $(viii)$ holds due to $\min_j(b_t)_j\geq\rho>0$ for any $j\in[1,2,...,d]$..

Now we turn attention to T_4,

$$
\begin{aligned}
T_4 &= L\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t^2\left\|\frac{g_t}{b_t}\right\|^2\right]+\frac{L}{2}\mathbb{E}[\sum_{t=1}^{T}\|\alpha_{t-1}\delta_{t-1}\|^2]+\frac{L}{2}\mathbb{E}[\sum_{t=1}^{T}\|(\alpha_{t-1}-\alpha_t)e_t\|^2] \\
&\leq \left(L\frac{G^2}{\rho^2}+2dL\right)\sum_{t=1}^{T}\alpha_t^2+\frac{dL}{2}\sum_{t=1}^{T}(\alpha_{t-1}-\alpha_t)^2,
\end{aligned}
\tag{31}
$$

where the inequality holds owing to $\|m_{t-1}\|\leq G$ and $\min_j(b_t)_j\geq\rho>0$, $\|(\delta_{t-1})_j\|\leq 2$, and $\mathbb{E}[\|e_t\|^2]\leq d$.

Combining Eq.(25-31), we can obtain

$$
\begin{aligned}
\mathbb{E}[f(z_T)-f(z_1)] \leq & \frac{\alpha_0\beta d}{(1-\beta)\rho}G^2+\left(\frac{(8\beta^2+10\beta+5)L^2 d}{(1-\beta)^2}+\frac{G^2}{2\rho^2}\right)\sum_{t=1}^{T}\alpha_{t-1}^2 \\
& -\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right]+\frac{4G^2\alpha_1 d}{\rho}+\frac{\alpha_0^2\beta^2 L G^2 d}{\rho^2(1-\beta)^2} \\
& +\left(L\frac{G^2}{\rho^2}+2dL\right)\sum_{t=1}^{T}\alpha_t^2+\frac{dL}{2}\sum_{t=1}^{T}(\alpha_{t-1}-\alpha_t)^2.
\end{aligned}
\tag{32}
$$

Reformulating Eq.(32), we then have

$$\begin{aligned}\frac{1}{G}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] \leq & \mathbb{E}[f(z_1)-f(z_T)] \\ & + \left(\frac{(8\beta^2+10\beta+5)L^2d}{(1-\beta)^2}+\frac{G^2(1+L)}{2\rho^2}+2dL\right)\sum_{t=1}^{T}\alpha_{t-1}^2 \\ & + \frac{dL}{2}\sum_{t=1}^{T}(\alpha_{t-1}-\alpha_t)^2 \\ & + \frac{\alpha_0\beta d}{(1-\beta)\rho}G^2+\frac{4G^2\alpha_1 d}{\rho}+\frac{\alpha_0^2\beta^2LG^2d}{\rho^2(1-\beta)^2}\end{aligned} \tag{33}$$

It is known the learning rate saftifies $\alpha_t = \frac{c}{\sqrt{t}}, \forall t \geq 1$ and $\alpha_0 = \alpha_1 = c$. Utilizing non-increasing α_t and Cauchy-Schwarz inequality, we know $\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(x_t)\|^2\right] \geq T\alpha_T\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|\right]^2 = \frac{\sqrt{T}}{c}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|\right]^2$. $\sum_{t=1}^{T}\alpha_{t-1}^2 = \sum_{t=1}^{T}\frac{c^2}{t} \leq c^2(1+\int_1^{T-1}\frac{1}{t}dt) \leq c^2(1+\log T)$, and $\sum_{t=1}^{T}(\alpha_{t-1}-\alpha_t)^2 = \sum_{t=2}^{T}(\alpha_{t-1}-\alpha_t)^2 \leq \sum_{t=2}^{T}\frac{c^2}{4(t-1)^3} \leq \frac{c^2}{4}(1+\int_1^{T-2}t^{-3}dt) = \frac{c^2}{4}(\frac{3}{2}-\frac{1}{2(T-2)}) \leq \frac{3c^2}{8}$, we further have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|\right]^2 \leq \frac{C_1}{\sqrt{T}}+\frac{C_2(1+\log T)}{\sqrt{T}}, \tag{34}$$

where we define

$$C_1 = cG\left(\mathbb{E}[f(z_1)-f^*]+\frac{3c^2dL}{16}+\frac{\beta cdG^2}{(1-\beta)\rho}+\frac{4cdG^2}{\rho}+\frac{c^2\beta^2LG^2d}{\rho^2(1-\beta)^2}\right), \tag{35}$$

$$C_2 = c^3G\left(\frac{(8\beta^2+10\beta+5)L^2d}{(1-\beta)^2}+\frac{G^2(1+L)}{2\rho^2}+2dL\right). \tag{36}$$

B Experiments for Comparing Vanilla SGD, SGDM, Adam, BRAM and SoftSignSGD

To address the bottleneck in communication during distributed training, numerous gradient compression algorithms have been proposed, aiming to reduce the communication volume. Most of these algorithms can be reduced to Vanilla SGD without momentum if compression is not performed. Generally speaking, the epoch-wise convergence rate and inference performance a compressed algorithms is upper bounded by its uncompressed counterpart. In the experiments, we conducted empirical experiments to evaluate the training and inference performance of of Vanilla SGD, SGDM, Adam, BRAM and its uncompressed version in training typical CNN-base, LSTM-base and Transformer-base DNNs.

Algorithm 1. SoftSignSGD

Input: model parameter x_0, x_1 , the momentum $m_0^{(i)} = 0$, $b_0^{(i)} = 0$, the exponential moving average factor β, the learning rate sequence $\{\alpha_t\}$
for $t = 1, ..., T$ **do**
  Randomly sample ξ_t and compute the gradient: $g_t = \nabla f(x_t; \xi_t)$
  Update the momentum m_t: $m_t = \beta m_{t-1} + (1-\beta)g_t$
  Update the momentum b_t: $b_t = \beta b_{t-1} + (1-\beta)|g_t|$
  Update the model parameter x_{t+1}: $x_{t+1} = x_t - \alpha_t \frac{m_t}{b_t}$
end for

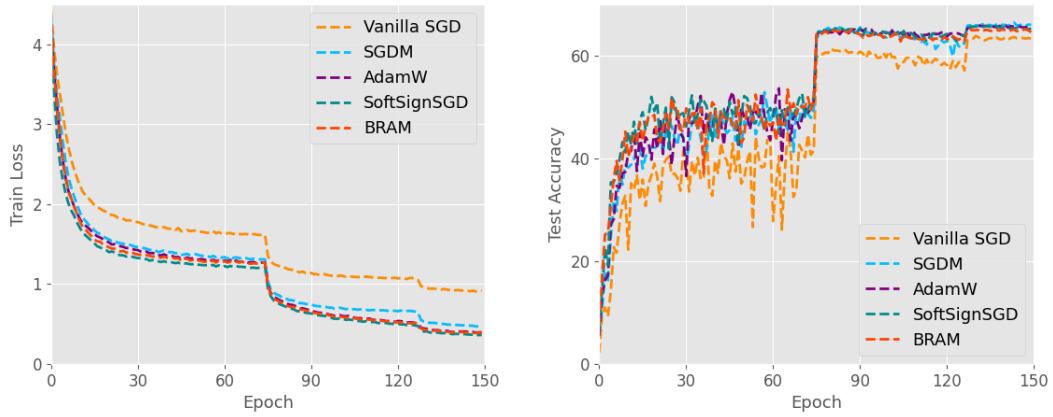
We refer to the uncompressed BRAM as SoftSignSGD. The implementation details for SoftSignSGD are presented in Algorithm 1. When comparing SoftSignSGD to Adam, there are two key differences. First, instead of using the square root of the exponential moving average of the squared gradient,

denoted as $\sqrt{v_t} = \sqrt{(1-\beta_2)v_{t-1} + (1-\beta_2)g_t^2}$, SoftSignSGD utilizes the exponential moving average of the absolute gradient, represented as $b_t = (1-\beta)b_{t-1} + |g_t|$. Second, in SoftSignSGD, the exponential moving factors for both the numerator m_t and the denominator b_t are the same. These differences ensure that each element of the updating amount in SoftSignSGD satisfies the condition $-1 \leq (\frac{m_t}{b_t})_j \leq 1$.

B.1 Experimental Results for training VGG16

We evaluated the performance of five optimization algorithms: Vanilla SGD, SGDM, AdamW, BRAM and SoftSignSGD, for training VGG-16 on CIFAR100. Each batch consisted of a set of 128 examples sampled with replacement. For SGDM, we set the momentum parameter β to 0.9, while for SoftSignSGD and BRAM, it was set to 0.95. For AdamW, the parameters β_1 and β_2 were set to 0.9 and 0.999, respectively. The weight decay was uniformly set to 0.0005 for Vanilla SGD and SGDM, and 0.05 for AdamW, BRAM and SoftSignSGD. To simplify the tuning process and ensure fair comparisons, we initialized the learning rates at 0.1 for Vanilla SGD and SGDM, and 0.005 for AdamW, BRAM and SoftSignSGD. We divided the learning rates by 10 after 75 and 130 epochs, and terminated the training after 150 epochs.

Figure 1 visually demonstrates that Vanilla SGD exhibits slower convergence speed and lower test accuracy compared to SGDM. In contrast, both BRAM and SoftSignSGD show comparable training and inference performance to the commonly used SGDM and AdamW. This observation suggests that BRAM outperforms existing gradient compression algorithms when training CNN-based VGG-16 models.



(a) Train Loss (b) Test Accuracy

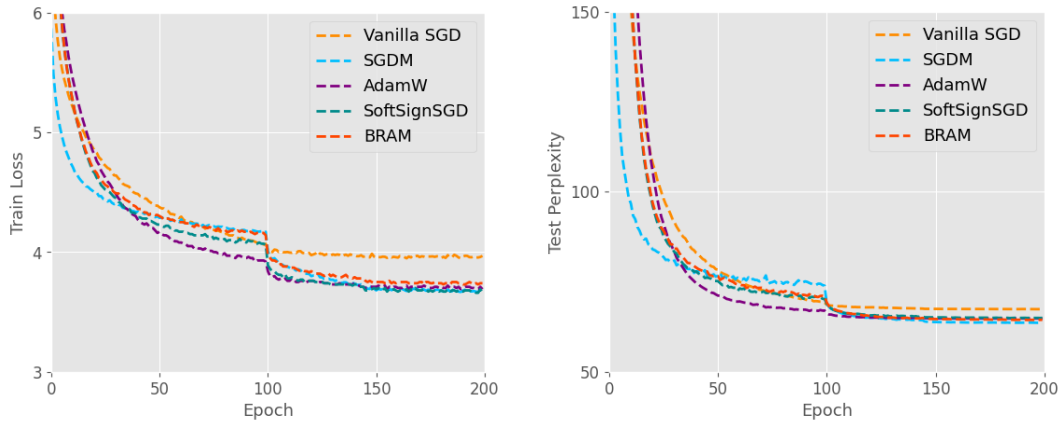
Figure 1: Training loss and test accuracy for VGG-16 on CIFAR100.

B.2 Experimental Results for training LSTM

We conducted experiments to train a 3-layer LSTM model on the Penn TreeBank dataset to evaluate the performance of five optimization algorithms: Vanilla SGD, SGDM, AdamW, BRAM and SoftSignSGD. Our implementations were built upon the code provided in the AdaBelief paper[1], and we used the default experimental settings for SGDM and AdamW. For Vanilla SGD, we used the experimental settings of SGDM with the exception that we set the momentum parameter β to 0. For BRAM and SoftSignSGD, we adopted the experimental settings of AdamW, except that we set the momentum parameter β to 0.99.

As visually illustrated in Figure 2, Vanilla SGD is still less effective than SGDM in terms of the convergence speed and the test accuracy, while the training and inference performance of SoftSignSGD and BRAM are comparative to common-used SGDM and AdamW. It indicates the BRAM is superior to exiting gradient compression algorithms for training LSTM.

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



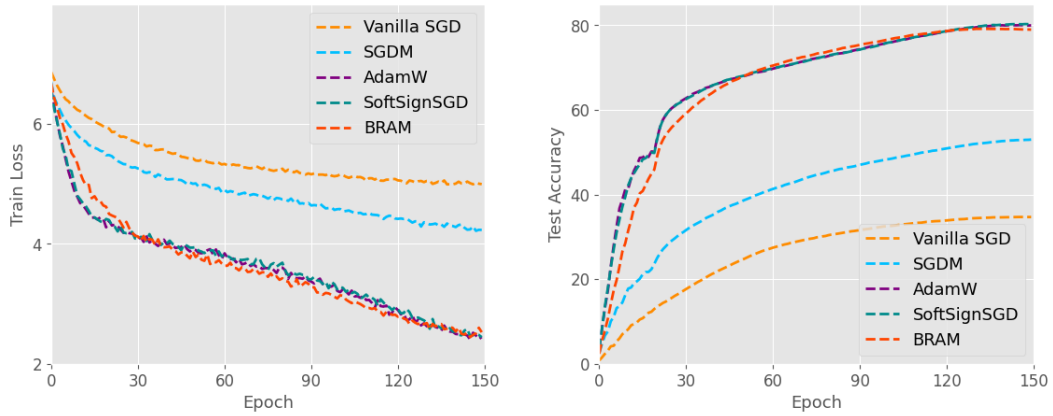
(a) Train Loss (b) Test Perplexity

Figure 2: Training loss and test perplexity (the lower, the better) for 3-layer LSTM on Penn TreeBank.

B.3 Experimental Results for training ViT

We train ViT-B with Vanilla SGD, SGDM, AdamW, SoftSignSGD and BRAM on the ILSVRC2012 with 32 GPUs (4 nodes). We use the Pytorch official implementation for ViT [2]. For AdamW, SoftSignSGD and BRAM, we followed the recommended experimental settings, with the exception that we set the momentum parameter β to 0.95 for SoftSignSGD and BRAM. As for Vanilla SGD and SGDM, we set the basic learning rate to 0.1 and the weight decay to 0.001, while keeping other settings the same as AdamW. Instead of the default 300 epochs, we uniformly set the total number of epochs to 150 for all optimizers

As visually illustrated in Figure 3, Vanilla SGD is still less effective than SGDM in terms of the convergence speed and the test accuracy, while the training and inference performance of SoftSignSGD and BRAM are comparative to common-used SGDM and AdamW. Notably, the performance of SGD-type optimizers are substantially inferior to that of adaptive optimizers.



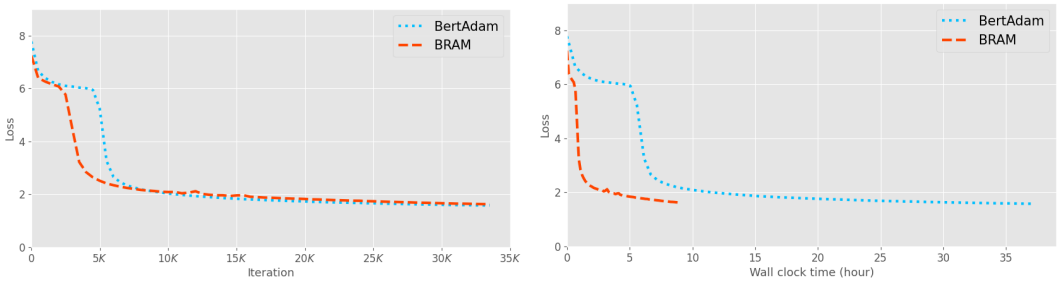
(a) Train Loss (b) Test Accuracy

Figure 3: Training loss and test accuracy for ViT-B-16 on ILSVRC2012.

C Experimental Results for pre-training BERT-Base

We employed BertAdam and BRAM to pre-train BERT-Base on Wikipedia using 64 GPUs (8 nodes). The sequence length was set to 512, and the batch size per GPU was set to 16. The training process

[2] https://github.com/pytorch/vision/tree/main/references/classification



(a) Iteration-wise, BERT-Base, batch size=16 × 64 (b) Time-wise, BERT-Base, batch size=16 × 64

Figure 4: Iteration-wise and time-wise convergence speed for pre-training BERT-Base with 16 samples per GPU with 64 GPUs.

consisted of $37,000$ iterations. The learning rate started at $4 \times 10^{-}4$ and linearly increased in the first $12,500$ iterations, after which it linearly decreased to 0 for the remaining iterations. ForBertAdam, the parameter values $[\beta_1, \beta_2]$ were set to $[0.9, 0.999]$, and for BRAM, the momentum parameter beta was set to 0.9.

As depicted in Figure 4,BRAM demonstrates a comparable iteration-wise convergence rate to BertAdam. However, in terms of time-wise convergence, BRAM achieves a 4.2x faster convergence rate compared to BertAdam.

D Experiments with InfiniBand connections

Table 1: System throughput (samples/s) of SGDM, 1-bit Adam and BRAM for training ResNet-50 on ILSVRC2012 with 10Gbps Ethernet and 200Gbps InfiniBand.

#GPUs	Optimizer	Ethernet (10Gbps)			InfiniBand (200Gbps)		
		Throughput (samples/s)	Speedup	Scale Efficiency	Throughput (samples/s)	Speedup	Scale Efficiency
8	SGDM	3693	1.00×	100%	3693	1.00×	100%
	1-bit Adam	3243	0.83×	100%	3243	0.83×	100%
	BRAM	3462	0.94×	100%	3462	0.94×	100%
16	SGDM	2959	1.00×	40.1%	4673	1.00×	63.2%
	1-bit Adam	4715	1.60×	72.7%	5708	1.22×	88.0%
	BRAM	6015	2.03×	86.9%	6784	1.45×	97.9%
32	SGDM	4270	1.00×	28.9%	9063	1.00×	61.3%
	1-bit Adam	7268	1.70×	56.0%	10249	1.13×	79.0%
	BRAM	9416	2.21×	68.0%	12131	1.34×	87.6%
32	SGDM	6189	1.00×	20.9%	16608	1.00×	56.2%
	1-bit Adam	5546	0.89×	21.3%	16920	1.02×	65.2%
	BRAM	15253	2.47×	55.1%	19956	1.21×	72.1%

To further evaluate the communication efficiency of SGDM/Adam, SoftSignSGD and BRAM with high bandwidth connections, we implement experiments for training ResNet-50 and BERT-Base with distributed nodes connected with 200Gbps InfiniBand. All the experimental settings are the same as we perform experiments with Ethernet in Subsection 5.1, and the experimental results are listed in Table 1 and Table 2.

As shown in Table 1 and Table 2, compared with the baseline SGDM/Adam, BRAM can still reach up to $1.45\times$ speedup for ResNet-50 on ILSVRC2012 and $2.85\times$ speedup for BERT-Base on SQuAD 1.1, although the speed advantage is not so obvious as that with lower-bandwidth Ethernet connections. An interesting phenomenon is that the system throughput of BRAM with 10Gbps Ethernet can match that of SGDM/Adam with 200Gbps InfiniBand.

The experimental results in Table 1 and Table 2 also show that as the number of GPUs is increasing, the scale efficiency of SGDM/Adam, SoftSignSGD and BRAM becomes lower. The reason for this phenomenon can be summarized in the following. When the number of GPUs doubles, the number of communication trips also multiplies. We take the communication scheme *All-Reduce* for example.

Table 2: System throughput (samples/s) of BertAdam, 1-bit Adam and BRAM for fine tuning BERT-Base on SQuAD 1.1 with 10Gbps Ethernet and 200Gbps InfiniBand.

#GPUs	Optimizer	Ethernet (10Gbps)			InfiniBand (200Gbps)		
		Throughput (samples/s)	Speedup	Scale Efficiency	Throughput (samples/s)	Speedup	Scale Efficiency
8	BertAdam	413	1.00×	100%	413	1.00×	100%
	1-bit Adam	358	0.87×	100%	358	0.83×	100%
	BRAM	412	1.00×	100%	412	0.94×	100%
16	BertAdam	84	1.00×	10.1%	272	1.00×	32.9%
	1-bit Adam	213	2.54×	29.7%	522	1.92×	72.9%
	BRAM	431	5.13×	52.3%	776	2.85×	94.1%
32	BertAdam	119	1.00×	7.20%	543	1.00×	32.8%
	1-bit Adam	274	2.30×	19.1%	903	1.66×	63.1%
	BRAM	730	6.13×	44.2%	1365	2.51×	82.9%
32	BertAdam	158	1.00×	4.78%	998	1.00×	30.2%
	1-bit Adam	252	1.59×	8.80%	1496	1.50×	52.2%
	BRAM	990	6.26×	30.0%	2299	2.30×	69.8%

If the number of GPUs is n, each GPU requires $2(n-1)$ trips across the network confections. When the number is non-trivial, the computation time of the communication primitives may exceed the time of the pure communication itself and dominate the overall communication time, since the total communication overhead does not change with the number of GPUs. Notably, *All-reduce* is more efficient than *All-to-All* which is the core of our *Hierarchical-1-bit-All-Reduce*. Hence, as shown in in Table 1 and Table 2, the scale efficiency of BRAM decreases more quickly than SGDM/Adam with the number of GPUs growing.

E Discussion

The original paper on 1-bit Adam reports a significant speed advantage (up to $3.8\times$) for 1-bit Adam compared to full-precision Adam, with the advantage becoming more prominent as the number of GPUs increases. However, in our experiments, we did not observe clear speed advantages for 1-bit Adam over the original Adam. In fact, when running on 64 GPUs, 1-bit Adam was not only slower than the original Adam, but its throughput rate was even lower than that on 32 GPUs. There are several reasons for this phenomenon. First, the speedup of 1-bit Adam is obtained by comparing the throughput of the compression phase with that of the warm-up phase. However, in our experiments, we evaluated the overall average throughput of both the warm-up phase and the compression phase for 1-bit Adam. Second, the baseline Adam did not run with system-level efficient *DDP*. Third, the authors of 1-bit Adam customized highly efficient communication primitives specifically for their optimizer, whereas we utilized off-the-shelf communication primitives in PyTorch for all the optimizers to ensure fairness.

As shown in Figure 4, as the number of GPUs increases, the communication time for BRAM also grows superlinearly. One of the reasons for this is that the communication primitive *All-to-All* accounts for an increasing portion of the communication time. However, the native *All-to-All* in Step (iii) of the *Hierarchical-1-bit-All-Reduce* is not less efficient than the native *All-Reduce*. Therefore, we plan to further optimize the *All-to-All* and *All-Gather* primitives to accelerate BRAM.

When training large-scale DNNs, the mixed-precision technique is commonly used to reduce memory consumption, allowing for larger model sizes. While optimizers still utilize full-precision states and computations, which typically contribute to 33-75% of the total memory footprint, BRAM does not require full-precision states or computations. Moreover, due to the random quantization of updates to 1 or -1, BRAM can leverage lower precision than FP16 gradients to estimate the update. Therefore, BRAM shows promise for applications that focus on reducing memory usage, as highlighted in recent research on 8-bit optimizers via block-wise quantization (Tim Dettmers et al., ICLR 2022).