More details for rebuttal

1 Double Squeeze

In this part, we follow the notations of [57] to derive explicit upper bounds for server/worker compression errors $\mathbb{E}[\|\boldsymbol{\delta}_t\|]$ and $\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|]$ which, combined with Corollary 2, [57], lead to the rate in our Table 1. We consider compressors Q (server) and Q_i (worker i) utilized are δ-contractive. We use $\boldsymbol{g}_t^{(i)}$ to indicate local (stochastic) gradient.

1.1 $\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|] = O(G/\delta)$

By δ-contraction and the Cauchy-Schwartz inequality, we have for any $\rho > 0$ that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|^2] =& \mathbb{E}[\|\boldsymbol{v}_t^{(i)} - Q_i(\boldsymbol{v}_t^{(i)})\|^2] \\
\leq& (1-\delta)\mathbb{E}[\|\boldsymbol{v}_t^{(i)}\|^2] = (1-\delta)\mathbb{E}[\|\boldsymbol{g}_t^{(i)} + \boldsymbol{\delta}_{t-1}^{(i)}\|^2] \\
\leq& (1+\rho)(1-\delta)\mathbb{E}[\|\boldsymbol{g}_t^{(i)}\|^2] + (1+1/\rho)(1-\delta)\mathbb{E}[\|\boldsymbol{\delta}_{t-1}^{(i)}\|^2]
\end{aligned} \tag{1}$$

Iterating (1) for $t, t-1, \ldots, 0$ and noting $\boldsymbol{\delta}_0^{(i)} = 0$, we reach

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|^2] \leq& (1+\rho)(1-\delta)\sum_{s=1}^{t}(1+1/\rho)^{t-s}(1-\delta)^{t-s}\mathbb{E}[\|\boldsymbol{g}_s^{(i)}\|^2] \\
\leq& \frac{(1+\rho)(1-\delta)G^2}{1-(1+1/\rho)(1-\delta)}
\end{aligned} \tag{2}$$

where the last inequality holds because $\mathbb{E}[\|\boldsymbol{g}_s^{(i)}\|^2] \leq G^2$ for all $1 \leq s \leq t$. Here one must choose $\rho = \Omega(1/\delta)$ to avoid the explosion of the upper bound, which leads to $\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|^2] = O(G^2/\delta^2)$. Therefore, by Jessen's inequality, we have $\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|] \leq \sqrt{\mathbb{E}[\|\boldsymbol{\delta}_t^{(i)}\|^2]} = O(G/\delta)$.

1.2 $\mathbb{E}[\|\boldsymbol{\delta}_t\|] = O(G/\delta^2)$

By δ-contraction and the Cauchy-Schwartz inequality, we have for any $\rho > 0$ that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\delta}_t\|^2] =& \mathbb{E}[\|\boldsymbol{v}_t - Q(\boldsymbol{v}_t)\|^2] \\
\leq& (1-\delta)\mathbb{E}[\|\boldsymbol{v}_t\|^2] = (1-\delta)\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{v}_t^{(i)}) + \boldsymbol{\delta}_{t-1}\|^2] \\
\leq& (1-\delta)(1+\rho)\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{v}_t^{(i)})\|^2] + (1+1/\rho)(1-\delta)\mathbb{E}[\|\boldsymbol{\delta}_{t-1}\|^2].
\end{aligned} \tag{3}$$

By the Cauchy-Schwartz inequality and δ-contraction, we have

$$\begin{aligned}
\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{v}_t^{(i)})\|^2] \leq& 2\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} Q_i(\boldsymbol{v}_t^{(i)}) - \boldsymbol{v}_t^{(i)}\|^2] + 2\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_t^{(i)}\|^2] \\
\leq& \frac{2-\delta}{n}\sum_{i=1}^{n}\mathbb{E}[\|\boldsymbol{v}_t^{(i)}\|^2] = O(G^2/\delta^2)
\end{aligned} \tag{4}$$

where the last identity is because the upper bound of $\mathbb{E}[\|\boldsymbol{v}_t^{(i)}\|^2]$ is revealed by the derivations in (1) and (2). Taking $\rho = \frac{2(1-\delta)}{\delta} = O(1/\delta)$, we reach an inequality taking a form like

$$\mathbb{E}[\|\boldsymbol{\delta}_t\|^2] \le (1-\delta/2)\mathbb{E}[\|\boldsymbol{\delta}_{t-1}\|^2] + O(G^2/\delta^3). \tag{5}$$

Iterating (5) similarly, we easily reach $\mathbb{E}[\|\boldsymbol{\delta}_t\|^2] = O(G^2/\delta^4)$ and thus $\mathbb{E}[\|\boldsymbol{\delta}_t\|] = O(G/\delta^2)$.

2 MEM-SGD

In this part, we show the rate for MEM-SGD in the non-convex setup. The main recursion of MEM-SGD is

$$\mathbf{p}_t^i = \eta\nabla f_i(\mathbf{x}_t, \xi_t^i) + \mathbf{e}_t^i, \quad \mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{n}\sum_{i=1}^n Q_i(\mathbf{p}_t^i), \quad \mathbf{e}_{t+1}^i = \mathbf{p}_t^i - Q_i(\mathbf{p}_t^i).$$

The key steps of our derivation are listed as follows.

1. Following the similar argument to (1) and (2), we can bound the compression error as $\mathbb{E}[\|\mathbf{e}_t^i\|^2] = O(\eta^2 G^2/\delta^2)$. Note that here η^2 appears since compression is conducted after the step size η multiplied.
2. The recursion formula of MEM-SGD is $\mathbf{y}_{t+1} = \mathbf{y}_t - \frac{\eta}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t, \xi_t^i)$ with $\mathbf{y}_t \triangleq \mathbf{x}_t - \frac{1}{n}\sum_{i=1}^n \mathbf{e}_t^i$. In fact, one easily check that

$$\begin{aligned}
\mathbf{x}_{t+1} =& \mathbf{x}_t - \frac{1}{n}\sum_{i=1}^n Q_i(\mathbf{p}_t^i) \\
=& \mathbf{x}_t - \frac{1}{n}\sum_{i=1}^n (\mathbf{p}_t^i - \mathbf{e}_{t+1}^i) = \mathbf{x}_t - \frac{1}{n}\sum_{i=1}^n (\eta\nabla f_i(\mathbf{x}_t, \xi_t^i) + \mathbf{e}_t^i - \mathbf{e}_{t+1}^i) \\
=& \mathbf{y}_t - \frac{\eta}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t, \xi_t^i) + \frac{1}{n}\sum_{i=1}^n \mathbf{e}_{t+1}^i.
\end{aligned}$$

3. Following eqn. (33), Lemma 7 in our submitted manuscript, one has

$$\mathbb{E}[f(\mathbf{y}_{t+1})] - \mathbb{E}[f(\mathbf{y}_t)] \le 2\eta L^2 \mathbb{E}[\|\mathbf{y}_t - \mathbf{x}_t\|^2] - \frac{\eta(1-\eta L)}{2}\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] + \frac{\eta^2 L\sigma^2}{2n}. \tag{6}$$

   Setting $\eta \le \frac{1}{2L}$ such that $\frac{\eta(1-\eta L)}{2} \ge \frac{\eta}{4}$ and rearranging (6), we have

$$\mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \le \frac{4(\mathbb{E}[f(\mathbf{y}_t)] - \mathbb{E}[f(\mathbf{y}_{t+1})])}{\eta} + \frac{2\eta L\sigma^2}{n} + 8L^2\mathbb{E}[\|\mathbf{y}_t - \mathbf{x}_t\|^2]. \tag{7}$$

4. By the definition of $\mathbf{y}_t$ and step 1., we have

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{x}_t\|^2] \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{e}_t^i\|^2] = O(\eta^2 G^2/\delta^2). \tag{8}$$

   Averaging (7) with (8) plugged into, we reach

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\mathbf{y}_t)\|^2] \le O\left(\frac{\mathbb{E}[f(\mathbf{x}_0)] - f^\star}{\eta} + \frac{\eta L\sigma^2}{n} + \frac{\eta^2 L^2 G^2}{\delta^2}\right). \tag{9}$$

   Setting step size $\eta = \frac{1}{2L + (\frac{L\sigma^2}{n\Delta})^{\frac{1}{2}} + (\frac{L^2G^2}{\delta^2\Delta})^{\frac{1}{3}}}$ leads to the rate we listed in Table 1.

3 Updates on Linear Regression & Logistic Regression

In this part, we provide the experimental results of linear regression and logistic regression with a longer period. It is observed that all methods converge to a "SGD oscillation stage".

We also supplement the results of EF21-SGD. For EF21-SGD, we conduct the same number of gradient queries and of communication rounds for a fair comparison. It is observer that EF21-SGD performs significantly worse than others.

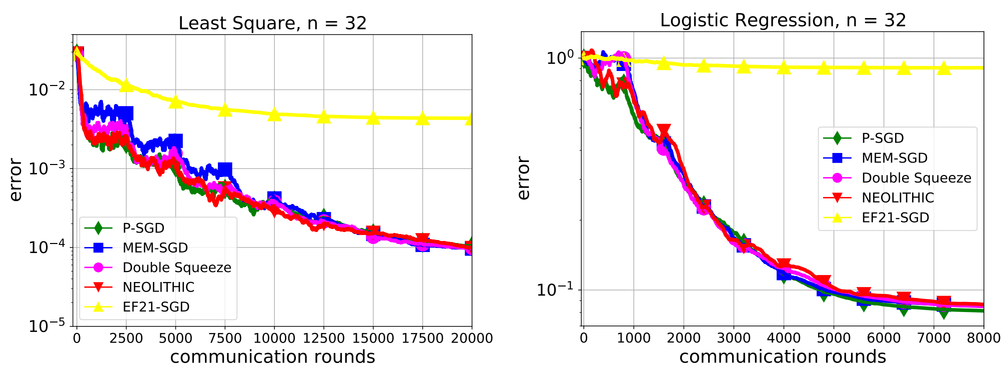

Figure 1: Convergence results on synthetic problems in terms of the mean-square error $(\mathbb{E})\|x - x^\star\|^2$ versus communication rounds.