Convex Optimisation

Local SGD convergence

June 29, 2022

1 Introduction

Mini-batch Stochastic Gradient Descent (SGD) has achieved great success in large-scale distributed training. It can theoretically reach a linear speedup with respect to the number of devices. However, the linear speedup can hardly be realized due to the network delay or bandwidth limit in a real-world scenario [2]. For example, the development of Gboard requires the participation of hundreds or thousands of phone users. If the model is trained such that it needs to communicate across phones at every iteration, the communication cost can be a bottleneck. Therefore, we analyze mini-batch LocalSGD that has been demonstrated to be efficient in reducing communication frequency in this study. Rather than synchronize the models over different devices at every iteration, it only averages the models once in a while (H). This report will study if Local SGD can achieve the same accuracy as mini-batch SGD by using Hx communications.

2 Method

We consider the finite-sum convex optimization problem $f : \mathbb{R}^d \to \mathbb{R}$ of the form:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad x^* := \text{argmin}_{x\in\mathbb{R}^d} f(x), \quad f^* := f(x^*) \tag{1}$$

where f is L-smooth and μ-strongly convex. We consider K parallel mini-batch SGD sequences with minibatch size of b that are synchronized (by averaging) after at most every H iterations. When $H = 1$, the synchronization is performed every iteration and so is called (parallel) mini-batch SGD. The comparison is shown in Fig. 1.

We initialize all the available devices (K) with the same model at the beginning of the training. We then optimize the model for H iterations with a mini-batch size of b on each device. Once all the devices finish training, we can obtain an aggregated model by simply averaging the weights (single communication step). This aggregated model is then used as the initialization for the next communication rounds (see Fig. 1(a)). We repeat this process until we achieve a certain level of accuracy. Compared to mini-batch SGD (see Fig. 1(b)), local SGD allows us to communicate Hx less rounds.

Algorithm 1 Local SGD

Initialise variables $x_0^k = x_0$ for workers $k \in [K]$
procedure MODEL UPDATING
  for $t = 0 \to T-1$ **do**
    parallel for $k \in [K]$ **do**
    sample i_t^k uniformly in [n]
    if t+1 $\in \mathcal{I}_T$ **then**
      $x_{t+1}^k \leftarrow \frac{1}{K}\sum_{k=1}^{K}(x_t^k - \eta_t \nabla f_{i_t^k}(x_t^k))$ ▷ *global synchronization*
    else
      $x_{t+1}^k \leftarrow x_t^k - \eta_t \nabla f_{i_t^k}(x_t^k)$ ▷ *local update*
    end if
    end parallel for
  end for
end procedure

The algorithm is shown in Algorithm 1 where K represents the number of devices, $\{x_t^k\}_{t=0}^{T-1}$

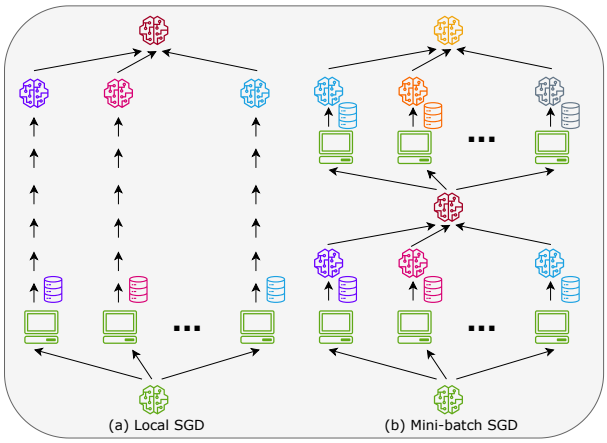

Figure 1: Differences between local SGD and (parallel) mini-batch SGD. With local SGD, we only let the models/devices communicate once a model has been optimised H steps on each of the devices. With mini-batch SGD, multiple devices need to communicate at every iteration

is a sequence of model updates on device k, η_t is the learning rate at iteration t, and $\mathcal{I}_T$ is a sequences of communication signals. If $\mathcal{I}_T = [T]$, then all the devices communicate at every iteration (*mini-batch SGD*). If $\mathcal{I}_T = \{T\}$, all the devices only communicate once at the end of the training (*one-shot averaging*). To demonstrate the influence of each of the parameters on convergence, we show convergence:

Theorem 1 *Let f be L-smooth and μ-strongly convex function, $\mathbb{E}_i||\nabla f_i(x_t^k) - \nabla f(x_t^k)|| \leq \sigma^2, \mathbb{E}||\nabla f_i(x_t^k)|| \leq G^2$, for t = 0, 1, ..., T-1 where $\{x_t^k\}_{t=0}^T$ for $k \in [K]$ are generated according to Algorithm 1 with gap($\mathcal{I}_T \leq H$) and for step size $\eta_t = \frac{4}{\mu(a+t)}$ with shift parameter $a > max\{16\kappa, H\}$ for $\kappa = \frac{L}{\mu}$, b=1. Then*

$$\mathbb{E}(f(\hat{x}_T) - f^*) \leq \frac{\mu a^3}{2S_T}||x_0 - x^*||^2 + \frac{4T(T+2a)}{\mu K S_T}\sigma^2 + \frac{256T}{\mu^2 S_T}G^2H^2L \tag{2}$$

where $\hat{x}_T = \frac{1}{KS_T}\sum_{k=1}^K \sum_{t=0}^{T-1} w_i x_t^k$, for $w_t = (a+t)^2$ and $S_T = \sum_{t=0}^{T-1} w_t \geq \frac{1}{3}T^3$

To carry out the proof of the convergence rate, we first define:

$$\bar{x}_0 = x_0, \quad \bar{x}_t = \frac{1}{K}\sum_{k=1}^K x_t^k, \quad g_t := \frac{1}{K}\sum_{k=1}^K \nabla f_{i_t^k}(x_t^k), \quad \bar{g}_t := \frac{1}{K}\sum_{k=1}^K \nabla f(x_t^k) \tag{3}$$

where the sequence $\{x_t^k\}_{t\geq 0}$ is defined as:

$$x_{t+1}^k := \begin{cases} x_t^k - \eta_t \nabla f_{i_t^k}(x_t^k) & if\ t+1 \notin \mathcal{I}_T \\ \frac{1}{K}\sum_{k=1}^K (x_t^k - \eta_t \nabla f_{i_t^k}(x_t^k)) & if\ t+1 \in \mathcal{I}_T \end{cases} \tag{4}$$

If $\mathcal{I}_t = [T]$, then $\bar{x}_t = x_t^k, \quad \forall t \geq 0$. Besides, $\bar{x}_t = x_t^k, \quad \forall t \in \mathcal{I}_T$. With the above definition, we can derive $\bar{x}_{t+1} = \bar{x}_t - \eta_t g_t = \frac{1}{K}\sum_{k=1}^K (x_t^k - \eta_t \nabla f_{i_t^k}(x_t^k)) = \frac{1}{K}\sum_{k=1}^K x_{t+1}^k = \bar{x}_{t+1}$. Note, we do not need to compute $\bar{x}$ and $\bar{g}$ explicitly, it is only a tool that we use for proving the convergence rate. To demonstrate the convergence rate, we need to find the bound for $f(x_t) - f^*$ where x^* is the optimal solution.

2.1 Convergence proof

Lemma 2 *Let $\{x_t\}_{t\geq 0}$ and $\{\bar{x}_t\}_{t\geq 0}$ for $k \in [K]$ be defined as in Eq. 4 and Eq. 3 and let f be L-smooth and μ-strongly convex and $\eta_t \leq \frac{1}{4L}$, then*

$$\mathbb{E}||\bar{x}_{t+1} - x^*||^2 \leq (1-\mu\eta_t)\mathbb{E}||\bar{x}_t - x^*||^2 + \eta_t^2\mathbb{E}||g_t - \bar{g}_t||^2 - \frac{1}{2}\eta_t\mathbb{E}(f(\bar{x}_t) - f^*) + 2\eta_t\frac{L}{K}\sum_{k=1}^{K}\mathbb{E}||\bar{x}_t - x_t^k||^2 \quad (5)$$

$$\begin{aligned} ||\bar{x}_{t+1} - x^*||^2 &= ||\bar{x}_t - \eta_t g_t - x^*||^2 \qquad \leftarrow \text{Eq. 3} \\ &= ||\bar{x}_t - \eta_t g_t - \eta_t\bar{g}_t - x^* + \eta_t\bar{g}_t||^2 \\ &= ||\bar{x}_t - \eta_t\bar{g}_t - x^*||^2 + \eta_t^2||g_t - \bar{g}_t||^2 - 2(\bar{x}_t - \eta_t\bar{g}_t - x^*)(g_t - \bar{g}_t) \end{aligned} \quad (6)$$

$$||\bar{x}_t - \eta_t\bar{g}_t - x^*||^2 = ||\bar{x}_t - x^*||^2 - 2\eta_t\langle\bar{x}_t - x^*, \bar{g}_t\rangle + \eta_t^2||\bar{g}_t||^2 \quad (7)$$

$$\begin{aligned} -2\eta_t\langle\bar{x}_t - x^*, \bar{g}_t\rangle &= -2\frac{\eta_t}{K}\sum_{k=1}^{K}\langle\bar{x}_t - x^*, \nabla f(x_k^t)\rangle \\ &= -2\frac{\eta_t}{K}\sum_{k=1}^{K}\langle\bar{x}_t - x_t^k + x_t^k - x^*, \nabla f(x_k^t)\rangle \\ &= -2\frac{\eta_t}{K}\sum_{k=1}^{K}\langle\bar{x}_t - x_t^k, \nabla f(x_k^t)\rangle - 2\frac{\eta_t}{K}\sum_{k=1}^{K}\langle x_t^k - x^*, \nabla f(x_k^t)\rangle \end{aligned} \quad (8)$$

$$-2\frac{\eta_t}{K}\sum_{k=1}^{K}\langle x_t^k - x^*, \nabla f(x_k^t)\rangle \leq 2\frac{\eta_t}{K}\sum_{k=1}^{K} f^* - f(x_t^k) - \frac{\mu}{2}||x^* - x_t^k||^2 \quad \mu - \text{strongly convex}^1 \quad (9)$$

As for the first term, we know that $2\langle a, b\rangle \leq \gamma||a||^2 + \frac{1}{\gamma}||b||^2 \quad \forall\gamma > 0$. Therefore:

$$\begin{aligned} -2\frac{\eta_t}{K}\sum_{k=1}^{K}\langle\bar{x}_t - x_t^k, \nabla f(x_k^t)\rangle &= \frac{\eta_t}{K}\sum_{k=1}^{K}2\langle x_t^k - \bar{x}_t, \nabla f(x_k^t)\rangle \\ &\leq \frac{\eta_t}{K}\sum_{k=1}^{K}(2L||x_t^k - \bar{x}_t||^2 + \frac{1}{2L}||\nabla f(x_k^t)||^2) \qquad \gamma = 2L \\ &\leq \frac{\eta_t}{K}\sum_{k=1}^{K}(2L||x_t^k - \bar{x}_t||^2 + \frac{1}{2L}2L(f(x_t^k) - f*)) \quad L - \text{smooth}^2 \\ &= \frac{\eta_t}{K}\sum_{k=1}^{K}(2L||x_t^k - \bar{x}_t||^2 + f(x_t^k) - f^*) \end{aligned} \quad (10)$$

[1] $f^* \geq f(x) + \nabla f(x)^T(x^* - x) + \frac{\mu}{2}||x^* - x||^2, \forall x \in \mathbb{R}^d$

Therefore:

$$-2\eta_t\langle \bar{x}_t - x^*, \bar{g}_t\rangle \leq 2\frac{\eta_t L}{K}\sum_{k=1}^{K}||x_t^k - \bar{x}_t||^2 + \frac{\eta_t}{K}\sum_{k=1}^{K}(f^* - f(x_k^t)) - \frac{\eta_t\mu}{K}\sum_{k=1}^{K}||x^* - x_t^k||^2 \quad (11)$$

Then we need to bound $\eta_t^2||\bar{g}_t||^2$,

$$\begin{aligned}\eta_t^2||\bar{g}_t||^2 &= \eta_t^2||\sum_{k=1}^{K}\nabla f(x_t^k))||^2 \\ &\leq \frac{\eta_t^2 K}{K^2}\sum_{k=1}^{K}||\nabla f(x_k^t)||^2 \quad \leftarrow ||\sum_i^n a_i||^2 \leq n\sum_{i=1}^{n}||a_i||^2 \quad (12)\\ &\leq \frac{\eta_t^2}{K}\sum_{k=1}^{K}2L(f(x_k^t) - f^*) \quad L-\text{smooth}\end{aligned}$$

Therefore by combining Eq. 11, Eq. 12, and the definition of convex[3], Eq. 7 can be written as:

$$\begin{aligned}||\bar{x}_t - \eta_t\bar{g}_t - x^*||^2 &= ||\bar{x}_t - x^*||^2 - 2\eta_t\langle \bar{x}_t - x^*, \bar{g}_t\rangle + \eta_t^2||\bar{g}_t||^2, \qquad \eta_t \leq \frac{1}{4L}\\ &\leq ||\bar{x}_t - x^*||^2 + 2\frac{\eta_t L}{K}\sum_{k=1}^{K}||x_t^k - \bar{x}_t||^2 - \frac{\eta_t}{2K}\sum_{k=1}^{K}(f(x_t^k) - f^*) - \frac{\eta_t\mu}{K}\sum_{k=1}^{K}||x^* - x_t^k||^2\\ &\leq ||\bar{x}_t - x^*||^2 + 2\frac{\eta_t L}{K}\sum_{k=1}^{K}||x_t^k - \bar{x}_t||^2 - \frac{\eta_t}{2K}\sum_{k=1}^{K}(f(x_t^k) - f^*) - \eta_t\mu||x^* - \bar{x}_t||^2\\ &\leq ||\bar{x}_t - x^*||^2 + 2\frac{\eta_t L}{K}\sum_{k=1}^{K}||x_t^k - \bar{x}_t||^2 - \frac{\eta_t}{2}(f(\bar{x}_t) - f^*) - \eta_t\mu||x^* - \bar{x}_t||^2\end{aligned}$$

Therefore,

$$\mathbb{E}||\bar{x}_{t+1} - x^*||^2 \leq (1-\eta_t\mu)\mathbb{E}||\bar{x}_t - x^*||^2 + 2\frac{\eta_t L}{K}\sum_{k=1}^{K}\mathbb{E}||x_t^k - \bar{x}_t||^2 - \frac{\eta_t}{2}\mathbb{E}f(\bar{x}_t) - f^* + \eta_t^2\mathbb{E}||g_t - \bar{g}_t||^2 \quad (13)$$

The term $\mathbb{E}(f\bar{x}_t) - f^*$ depends on $\mathbb{E}||\bar{x}_{t+1} - x^*||^2$, $\mathbb{E}||\bar{x}_t - x^*||^2$, and $\mathbb{E}||g_t - \bar{g}_t||^2$. The first two terms have the recurrence relationship such that if we repeat Eq. 13 t times, we can basically get a relation between $\mathbb{E}||\bar{x}_{t+1} - x^*||^2$ and $\mathbb{E}||\bar{x}_0 - x^*||^2$. Therefore, we only need to bound $\mathbb{E}||x_t^k - \bar{x}_t||^2$ and $\mathbb{E}||g_t - \bar{g}_t||^2$.

Lemma 3 *Let* $\sigma^2 \geq \mathbb{E}_i||\nabla f_i(x_t^k) - \nabla f(x_t^k)||^2$ *for* $k \in [K], t \in [T]$, *then* $\mathbb{E}||g_t - \bar{g}_t||^2 \leq \frac{\sigma^2}{K}$

[2] $||\nabla f(x) - \nabla f^*||_2^2 \leq 2L(f(x) - f^*) \leq L^2||x - x^*||_2^2, \forall x \in \mathbb{R}^d$

[3] $\frac{1}{K}\sum_{k=1}^{K}(f(x_t^k) - f^*) = \frac{1}{K}\sum_k f(x_t^k) - f^* \geq f(\frac{1}{K}\sum_{k=1}^{K}x_t^k) - f^*$ convexity

Proof:

$$\begin{aligned}\mathbb{E}||g_t - \bar{g}_t||^2 &= \mathbb{E}||\frac{1}{K}\sum_{k=1}^{K}\nabla f_{i_t^k}(x_t^k) - \frac{1}{K}\sum_{k=1}^{K}\nabla f(x_k^t)||^2 \\ &= \mathbb{E}||\frac{1}{K}\sum_{k=1}^{K}(\nabla f_{i_t^k}(x_t^k) - \nabla f(x_t^k)||^2 \\ &= \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}||\nabla f_{i_t^k}(x_t^k) - \nabla f(x_t^k)||^2 \leq \frac{1}{K^2}\sum_{k=1}^{K}\sigma^2 = \frac{\sigma^2}{K}\end{aligned} \tag{14}$$

where $\mathbb{E}||g_t - \bar{g}_t|| = 0$, so $\mathbb{E}||g_t - \bar{g}_t||^2$ can be considered as the variance. As K devices calculate the gradients in parallel without interfering (K independent variables), so $\text{Var}(\sum_{k=1}^{K} x_k) = \sum_{k=1}^{K}\text{Var}(x_k)$

Lemma 4 *If gap$(\mathcal{I}_T) \leq H$[4]and sequence of decreasing positive stepsize $\{\eta_t\}_{\{t\geq 0\}}$ satisfying $\eta_t \leq 2\eta_{t+H}$ for all $t \geq 0$, then*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}||\bar{x}_t - x_t^k||^2 \leq 2\eta_t^2 G^2 H^2 \tag{15}$$

where G^2 is a constant such that $\mathbb{E}_i||\nabla f_i(x_k^t)||^2 \leq G^2$ for all $k \in [K], t \in [T]$

Proof:

As the gap$(\mathcal{I}_T) \leq h$, there is an index $t_0, t - t_0 \leq H$ such that $\bar{x}_{t_0} = x_{t_0}^k$ for $k \in [K]$, we know that:

$$\mathbb{E}||x_t^k - x_{t_0}|| = \frac{1}{K}\sum_{k=1}^{K}||x_t^k - x_{t_0}|| = \frac{1}{K}\sum_{k=1}^{K}x_t^k - x_{t_0} = \bar{x}_t - x_{t_0} \tag{16a}$$

$$\mathbb{E}(x - \mathbb{E}(x))^2 = \mathbb{E}(x)^2 - (\mathbb{E}(x))^2 \tag{16b}$$

$$\mathbb{E}||\bar{x}_t - x_t^k||^2 = \mathbb{E}||x_t^k - x_{t_0} + x_{t_0} - \bar{x}_t||^2 = \mathbb{E}||x_t^k - x_{t_0}||^2 - (\mathbb{E}||x_t^k - x_{t_0}||)^2 \leq \mathbb{E}||x_t^k - x_{t_0}||^2 \tag{17}$$

$$x_{t_1}^k = x_{t_0}^k - \eta_{t_0}\nabla f_{i_{t_0}^k}(x_{t_0}^k) \tag{18a}$$

$$x_{t_2}^k = x_{t_1}^k - \eta_{t_1}\nabla f_{i_{t_1}^k}(x_{t_1}^k) \tag{18b}$$

$$\cdots$$

$$x_t^k = x_{t-1}^k - \eta_{t-1}\nabla f_{i_{t-1}^k}(x_{t-1}^k) \tag{18c}$$

Therefore,

$$\begin{aligned}x_t^k &= x_{t_0}^k - \eta_{t_0}\nabla f_{i_{t_0}^k}(x_{t_0}^k) - \eta_{t_1}\nabla f_{i_{t_1}^k}(x_{t_1}^k) - \ldots - \eta_{t-1}\nabla f_{i_{t-1}^k}(x_{t-1}^k) \\ &\geq x_{t_0}^k - \eta_{t_0}\sum_{h=t_0}^{t-1}\nabla f_{i_h^k}(x_h^k), \qquad \eta_{t_0} \geq \eta_{t_1} \geq \ldots \geq \eta_{t-1}\end{aligned} \tag{19a}$$

[4]*The gap of a set $\mathcal{P} := \{p_0, \ldots, p_t\}$ of $t+1$ integers, $p_i \leq p_{i+1}$ for $i = 0, \ldots, t-1$, is defined as:* $\text{gap}(\mathcal{P}) := \max_{i=0}^{t-1} p_{i+1} - p_i$

$$x_{t_0}^k - x_t^k \leq \eta_{t_0} \sum_{h=t_0}^{t-1} \nabla f_{i_h^k}(x_h^k) \tag{19b}$$

$$\begin{aligned}
\mathbb{E}||x_{t_0} - x_t^k||^2 &\leq \eta_{t_0}^2 \mathbb{E}|| \sum_{h=t_0}^{t-1} \nabla f_{i_h^k}(x_h^k)||^2 \\
&\leq \eta_{t_0}^2 \mathbb{E}(t - t_0) \sum_{h=t_0}^{t-1} ||\nabla f_{i_h^k}(x_h^k)||^2 \\
&\leq \eta_{t_0}^2 H \mathbb{E} \sum_{h=t_0}^{t-1} ||\nabla f_{i_h^k}(x_h^k)||^2 \quad , t - t_0 \leq H \\
&= \eta_{t_0}^2 H \sum_{h=t_0}^{t-1} \mathbb{E}||\nabla f_{i_h^k}(x_h^k)||^2 \\
&\leq \eta_{t_0}^2 H (t - t_0) G^2 \\
&\leq \eta_{t_0}^2 H^2 G^2 \\
&\leq 4\eta_t^2 H^2 G^2 \qquad \eta_{t_0} \leq 2\eta_t
\end{aligned} \tag{20}$$

By plugging Eq. 15 and Eq. 20, we can obtain:

$$\mathbb{E}||\bar{x}_{t+1} - x^*||^2 \leq (1 - \eta_t \mu)\mathbb{E}||\bar{x}_t - x^*||^2 + \frac{\eta_t^2}{K}\sigma^2 - \frac{1}{2}\eta_t \mathbb{E}(f(\bar{x}_t) - f^*) + 8\eta_t^3 H^2 G^2 L \tag{21}$$

Based on Lemma 3.4 from [3] (see Appendix), we can obtain Theorem 1. As there are many constant terms, we can simplify it as:

1. $\frac{\mu a^3}{2S_T}||x_0 - x^*||^2 \leq \frac{\mu a^3}{2S_T}\frac{4G^2}{\mu^2} \sim \frac{a^3}{\mu T^3}G^2 \quad \leftarrow$ Lemma 1 from [1] (see Appendix)
2. $\frac{4T(T+2a)}{\mu K S_T}\sigma^2 \sim \frac{T^2+Ta}{\mu K T^3}\sigma^2 \sim (\frac{1}{\mu K T} + \frac{\kappa+H}{\mu K T^2})\sigma^2$
3. $\frac{256T}{\mu^2 S_T}G^2 H^2 L \sim \frac{H^2 L}{\mu^2 T^2}G^2 \sim \frac{H^2\kappa}{\mu T^2}G^2$

Therefore,

Corollary 4.1 *Let $\hat{x}_T$ be defined as in Theorem 1 for parameter $a = max(16\kappa, H)$, then*[5]

$$\mathbb{E}f(\hat{x}_T) - f^* = O(\frac{1}{\mu K T} + \frac{\kappa + H}{\mu K T^2})\sigma^2 + O(\frac{\kappa H^2}{\mu T^2} + \frac{\kappa^3 + H^3}{\mu T^3})G^2 \tag{22}$$

Given Eq. 22, we observe that:

1. When $\sigma > 0$, then the first term $\frac{1}{KbT}$ dominates the convergence at rate $O(\frac{1}{KbT})$. It indicates that local SGD achieves a linear speedup w.r.t. the number of workers K and mini-batch size b
2. When we look at the second term, we can achieve a linear speedup $O(\frac{1}{TbK})$ if $H \sim \sqrt{\frac{T}{Kb}}$. It means that we can reduce the number of communications by a factor of $O(\sqrt{\frac{T}{Kb}})$

We next show some numerical results of the convergence.

[5]So far, we have assumed that each worker only computes a single stochastic gradeint. In mini-batch local SGD, each worker computes a mini-batch of size b in each iteration, which reduces the variance by a factor of b. Therefore, we replace σ^2 by $\frac{\sigma^2}{b}$ when we use local SGD with a mini-batch size of b.

3 Experimental results

We consider the logistic regression problem:

$$f(x) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(1+\exp(-(a_i x + b))), \quad \text{where} \quad a_i \in \mathbb{R}^{784}, y_i \in \{0,1\}, b \in \mathbb{R} \tag{23}$$

where a_i and y_i are corresponding to a single data sample and its label. We perform a binary classification task using the MNIST and Shape datasets (see some examples in Fig. 2). There are 100000 training images (50000 images for MNIST and 50000 images for Shape) and 10000 testing images (5000 images for MNIST and 5000 images for Shape). We uniformly distribute the images across devices such that each device sees a similar number of Shape and MNIST images. We use the same initialized model as the starting point of the optimization on each device. After each communication round, we extract the models from all the devices and simply average them to get the aggregated model. The number of iterations that are required to achieve a specific level of accuracy on the test dataset using the aggregated model is then compared with the centralized learning mini-batch SGD experiment ($H = 1, T = 1$) for measuring the speedup.

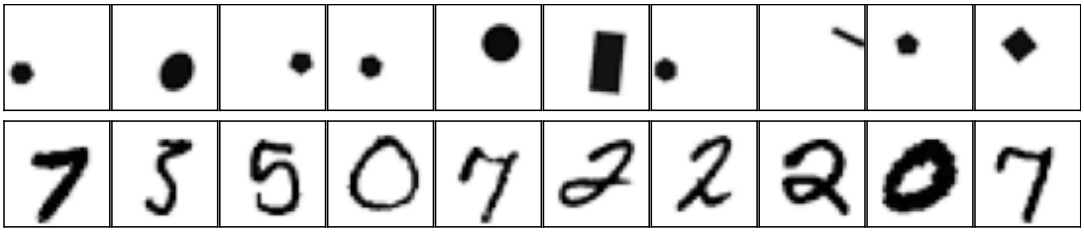

Figure 2: Some example images from the dataset (with shape as class 0 and digit as class 1)

We measure the speedup following Eq. 24:

$$\text{speedup} = \frac{T(\text{centralised} \to \epsilon)}{T(\text{LocalSGD} \to \epsilon)}, \quad \epsilon = \sum_{i=1}^{n} y_i \log(1+\exp(-a_i x + b)) \tag{24}$$

Where T is the number of gradient computation steps and ϵ is the prediction loss. We do not take into account the communication cost in this study as this ratio ρ is fairly low in the powerful simulated virtual environment[6]and is difficult to estimate. Therefore, ρ is set to be zero in the theoretical speedup $S(K) = \frac{K}{(\frac{1}{2}+\frac{1}{2}\sqrt{1+\epsilon(1+H+H^2K)})(1+2\rho\frac{K-1}{H})}$ [2]. Due to the time limit, we only experiment with $K = 2, 4, 8, 16$, and $H = 1, 2, 4, 8, 16, 32$. The result can be seen in Fig. 3. Note, the choice of ϵ can highly influence the measured speedup. However, with the current ϵ (manually chosen), and experimental setup($\rho = 0$),

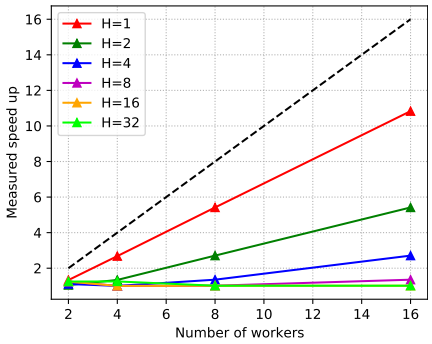

Figure 3: Measured speed up of LocalSGD with mini-batch size 256 for different number of workers and computation steps on each device ($\epsilon = 0.22$, $\rho = 0$, H is epochs)

we observe that the smaller the H is, the bigger speedup we can obtain. This is reasonable as we omit the communication cost. However, in a real-world scenario where $\rho > 0$, it may require more communication steps when H is small and thus reduce the speedup.

4 Conclusion

We derived the theoretical convergence rate of Local SGD with K devices and H computation steps on each device for parallel computing. We performed a simple numerical experiment on the MNIST and Shape dataset to demonstrate the benefit of using Local SGD. We observed Local SGD requires fewer communication steps to achieve a certain level of accuracy.

References

[1] Ohad Shamir. Making gradient descent optimal for strongly convex stochastic optimization. *CoRR*, abs/1109.5647, 2011.

[2] Sebastian U. Stich. Local SGD converges fast and communicates little. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[3] Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. *CoRR*, abs/1809.07599, 2018.

5 Appendix

Lemma 5 *Let* $\{a_t\}_{t\geq 0}$, $a_t \geq 0$, $\{e_t\}_{t\geq 0}$, $e_t \geq 0$ *be sequences satisfying:*

$$a_{t+1} \leq (1 - \frac{\mu\eta_t}{2})a_t + \eta_t^2 A + \eta_t^3 B - \eta_t e_t \tag{25}$$

for $\eta_t = \frac{8}{\mu(a+t)}$ *and constants* $A, B \geq 0$, $\mu > 0$, $a > 1$, *Then:*

$$\frac{1}{S_T}\sum_{t=0}^{T-1} w_t e_t \leq \frac{\mu A^3}{8S_T}a_0 + \frac{4T(T+2a)}{\mu S_T}A + \frac{64T}{\mu^2 S_T}B \tag{26}$$

for $w_t = (a+t)^2$ *and* $S_T := \sum_{t=0}^{T-1} w_t = \frac{T}{6}(2T^2 + 6aT - 3T + 6a^2 - 6a + 1) \geq \frac{1}{3}T^3$

Lemma 6 *Suppose f is* μ*-strongly convex and* L*-smooth function with respect to* x^* *over a convex set and* $\mathbb{E}_i||\nabla f_i(x_k^t)||^2 \leq G^2$, *then if we pick* $\eta_t = \frac{1}{\mu}$

$$\mathbb{E}||x_0 - x^*||^2 \leq \frac{4G^2}{\mu^2} \tag{27}$$

[6] We use DTU compute GPU cluster with Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz