Asynchronous SGD

November 27, 2024

1 Previous algorithm

1.1 Assumptions

Assumption 1. Local functions f_i are differentiable and L-smooth for some positive constant L, namely,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 2. Stochastic gradients $g_i(x) = \nabla f_i(x, \xi)$ are unbiased estimators of $\nabla f_i(x)$, *i.e.*,

$$\mathbb{E}_{\xi \sim \mathcal{D}_i}\left[\nabla f_i(x, \xi)\right] = \nabla f_i(x), \quad \forall x \in \mathbb{R}^d,$$

and have bounded variance $\sigma^2 \geq 0$, namely,

$$\mathbb{E}_{\xi \sim \mathcal{D}_i}\left[\|\nabla f_i(x, \xi) - \nabla f_i(x)\|^2\right] \leq \sigma^2, \quad \forall x \in \mathbb{R}^d.$$

Next, we also assume that the bounded function heterogeneity assumption holds since in general case it is not possible to derive any convergence guarantees for asynchronous algorithms.

Assumption 3. Local gradients $\nabla f_i(x)$ satisfy bounded heterogeneity condition for some $\zeta^2 \geq 0$, *i.e.*,

$$\|\nabla f_i(x) - \nabla f(x)\|^2 \leq \zeta^2, \quad \forall x \in \mathbb{R}^d.$$

Assumption 4. Local functions $f_i(x)$ are G- Lipschitz, i. e. for some positive constant G they satisfy

$$|f_i(x) - f_i(y)| \leq G\|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

Note that, in the case of differentiable f_i, this assumption implies that local gradients are bounded, i.e., for all $x \in \mathbb{R}^d$ $\|\nabla f_i(x)\| \leq G$.

1.2 Notations

Definition 0. Corresponding delays: $\tau_t, \tilde{\tau}_t \geq 0$, then

$$\pi_t := t - \tau_t, \quad \alpha_t := t - \tilde{\tau}_t.$$

Definition 1. Let $\{\tau_t\}_{t=0}^{T-1}$ be the delays of all applied gradients.
The average and maximum delays are defined as follows:

$$\tau_{\text{avg}} := \frac{1}{|\mathcal{A}_{T+1}|}\left(\sum_{t=0}^{T-1}\tau_t + \sum_{(i,j)\in\mathcal{A}_{T+1}\setminus\mathcal{R}_T} T - j\right), \quad \tau_{\max} := \max\left\{\max_{0\leq t<T}\tau_t, \max_{(i,j)\in\mathcal{A}_{T+1}\setminus\mathcal{R}_T} T - j\right\}.$$

Definition 2. The maximum number of active jobs or concurrency is defined as

$$\tau_C := \max_{0\leq t\leq T}|\mathcal{A}_{t+1}\setminus\mathcal{R}_t|.$$

Definition 3.

$$\widetilde{x}_0 = x_0,\ \widetilde{x}_{t+1} = \begin{cases}\widetilde{x}_t - \gamma\nabla f(x_t) & \text{if } t+1 \neq 0 \mod \tau,\\ x_{t+1} & \text{if } t+1 = 0 \mod \tau.\end{cases}$$

where $\tau = \Theta(\frac{1}{L\gamma})$.

1.3 Pure Asynchronous SGD

1.3.1 Algorithm

Algorithm 1 Pure Asynchronous SGD

Input: initial point x_0, stepsize γ, set of assigned jobs $\mathcal{A}_0 = \varnothing$, $\mathcal{A}_1 = \{(i, 0) \colon i \in [n]\}$, set of received jobs $\mathcal{R}_0 = \varnothing$

1: **for** $t = 0, 1, 2, \cdots, T-1$ **do**

2: once worker i_t finishes a job $(i_t, \pi_t) \in \mathcal{A}_{t+1}$(computing $g_{i_t}(x_{\pi_t})$), it sends $g_{i_t}(x_{\pi_t})$ to the server

3: server updates the current model $x_{t+1} = x_t - \gamma g_{i_t}(x_{\pi_t})$ and the set $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \{(i_t, \pi_t)\}$

4: server assigns worker i_t to compute a gradient $g_{i_t}(x_{t+1})$

5: server updates the set $\mathcal{A}_{t+2} = \mathcal{A}_{t+1} \cup \{(i_t, t+1)\}$

6: **end for**

1.3.2 Analysis

Proposition 1. Let Assumptions 1,2 and 3 hold. Let the stepsize γ satisfy inequalities

$$20L\gamma\sqrt{\tau_{\max}\tau_C} \leq 1, \quad 6L\gamma \leq 1$$

Let $\tau = \lfloor \frac{1}{20L\gamma} \rfloor$. Then the iterates of Algorithm 2 satisfy

$$\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{F_0}{\gamma T} + L\gamma\sigma^2 + \zeta^2\right),$$

where $\hat{x}_t$ is chosen uniformly at random from $\{x_0, \ldots, x_{T-1}\}$ and $F_0 := f(x_0) - f^*$.

Moreover, if we tune the stepsize, then the iterates of pure asynchronous SGD satisfy

$$\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{LF_0\sqrt{\tau_{\max}\tau_C}}{T} + \left(\frac{LF_0\sigma^2}{T}\right)^{1/2} + \zeta^2\right)$$

Proof.

$$A := \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - x_{\pi_t}\|^2\right]$$

$$B := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right]$$

$$C := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] \quad \text{(introduced later)}$$

$$D := \sum_{t=0}^{T-1} \mathbb{E}\left[\|x_t - \widetilde{x}_t\|^2\right]$$

We want to estimate B.

we consider a descent inequality for the virtual iterates $\tilde{x}_t$:

$$\widetilde{x}_0 = x_0, \ \widetilde{x}_{t+1} = \begin{cases} \widetilde{x}_t - \gamma \nabla f(x_t) & \text{if } t+1 \neq 0 \mod \tau, \\ x_{t+1} & \text{if } t+1 = 0 \mod \tau. \end{cases}$$

(Equation 22) We can get:

$$\begin{aligned} \mathbb{E}\left[f(\widetilde{x}_{t+1})\right] \leq \mathbb{E}\left[f(\widetilde{x}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] - \frac{\gamma}{3}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{L^2\gamma}{2}\mathbb{E}\left[\|\widetilde{x}_t - x_t\|^2\right] \\ + \left(\frac{1}{160L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] + 41L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right]\right)\xi_t, \quad \forall t \geq 0. \quad (*) \end{aligned}$$

where

$$\Delta_t^m := \sum_{j=r(t)}^{m} (\nabla f(x_j) - g_{i_j}(x_{\pi_j})), \quad \xi_t = \begin{cases} 0, & \text{if } t+1 \neq 0 \mod \tau, \\ 1, & \text{if } t+1 = 0 \mod \tau. \end{cases}$$

There are some strange terms here, so let's estimate them.

(Lemma C.1.)

$$\mathbb{E}\left[\|\Delta_t^m\|^2\right] \leq 4\tau^2\zeta^2 + 4L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] + 8L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{r(t)}\|^2\right] + \tau\sigma^2.$$

$$\begin{aligned} \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{r(t)}\|^2\right] \leq \frac{25}{3}\gamma^2\tau^3\zeta^2 + \frac{25}{3}\gamma^2L^2\tau^2 \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] \\ + \frac{25}{12}\gamma^2\tau^2 \sum_{j=r(t)}^{m} \mathbb{E}\left[\|\nabla f(x_j)\|^2\right] + \frac{25}{12}\gamma^2\tau^2\sigma^2. \end{aligned}$$

$$\mathbb{E}\left[\|\Delta_t^m\|^2\right] \leq \frac{25}{6}\tau^2\zeta^2 + \frac{25}{6}L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{\pi_j}\|^2\right] + \frac{\tau}{24} \sum_{j=r(t)}^{m} \mathbb{E}\left[\|\nabla f(x_j)\|^2\right] + \frac{25}{24}\tau\sigma^2.$$

Below we estimate the two terms associated with ξ_t.

First term(Equation 23):

$$\sum_{t=0}^{T-1} \frac{1}{160L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right]\xi_t \leq \frac{\gamma}{1600}B + \frac{\gamma}{2}C$$

Second term: (Equation 24)

By(Lemma C.3)

$$A \leq \frac{1}{132L^2}B + \frac{\zeta^2 T}{132L^2} + \frac{\gamma T\sigma^2}{5L}$$

$$\begin{aligned} L\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t^t\|^2\xi_t \leq & \frac{25}{6}L\gamma^2\tau\zeta^2 T + \frac{25}{6}\gamma^2L^3\tau A + \frac{1}{24}L\gamma^2\tau B + \frac{25}{24}L\gamma^2\sigma^2 T \\ \leq & \frac{25}{6}L\gamma^2\tau\zeta^2 T + \frac{25}{6}\gamma^2L^3\tau\left(\frac{1}{132L^2}B + \frac{\zeta^2 T}{132L^2} + \frac{\gamma T\sigma^2}{5L}\right) + \frac{1}{24}L\gamma^2\tau B + \frac{25}{24}L\gamma^2\sigma^2 T \\ \leq & 5L\gamma^2\tau\zeta^2 T + \frac{1}{10}\gamma^2 L\tau B + 2L\gamma^2\sigma^2 T \end{aligned}$$

Now we're just left with the D terms.

$$
\begin{aligned}
D &= \gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|\Delta_t^{t-1}\|^2\right] \\
&\leq \frac{\zeta^2 T}{200L^2} + \frac{1}{200}A + \frac{\gamma}{L}T\sigma^2 \\
&\leq \frac{\zeta^2 T}{200L^2} + \frac{1}{200}\left(\frac{1}{132L^2}B + \frac{\zeta^2 T}{132L^2} + \frac{\gamma T\sigma^2}{5L}\right) + \frac{\gamma}{L}T\sigma^2 \\
&\leq \frac{\zeta^2 T}{100L^2} + \frac{1}{20000L^2}B + \frac{2\gamma}{L}T\sigma^2
\end{aligned}
$$

At last, plug the two terms back and sum it up from 0 to $T-1$,

$$
\begin{aligned}
\mathbb{E}\left[f(\widetilde{x}_T) - f(\widetilde{x}_0)\right] \leq & -\frac{\gamma}{2}C - \frac{\gamma}{3}B + \frac{L^2\gamma}{2}D \\
& + \frac{1}{160L}\sum_{t=0}^{T-1}\xi_t \mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right] + 41L\gamma^2\sum_{t=0}^{T-1}\xi_t\mathbb{E}\left[\|\Delta_t^t\|^2\right] \\
\leq & -\frac{\gamma}{2}C - \frac{\gamma}{3}B \\
& + \frac{L^2\gamma}{2}\left(\frac{\zeta^2 T}{100L^2} + \frac{1}{20000L^2}B + \frac{2\gamma}{L}T\sigma^2\right) \\
& + \frac{\gamma}{1600}B + \frac{\gamma}{2}C \\
& + 124L\gamma^2\tau\zeta^2 T + \gamma^2 L\tau B + 82L\gamma^2\sigma^2 T \\
\leq & -\frac{\gamma}{4}B + 7\gamma T\zeta^2 + 83L\gamma^2\sigma^2 T
\end{aligned}
$$

Let $F_0 := f(x_0) - f^*$, the final rate

$$
\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{F_0}{\gamma T} + L\gamma\sigma^2 + \zeta^2\right).
$$

Since $\gamma \leq \dfrac{1}{L\sqrt{\tau_{\max}\tau_C}}$,

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] &\leq \mathcal{O}\left(\frac{F_0}{T}\sqrt{L\tau_{\max}\tau_C} + L\sigma^2\left(\frac{F_0}{L\sigma^2 T}\right)^{1/2} + \zeta^2\right) \\
&= \mathcal{O}\left(\frac{LF_0\sqrt{\tau_{\max}\tau_C}}{T} + \left(\frac{LF_0\sigma^2}{T}\right)^{1/2} + \zeta^2\right)
\end{aligned}
$$

1.4 Random Asynchronous SGD

1.4.1 Algorithm

Algorithm 2 Random Asynchronous SGD

Input: initial point x_0, stepsize γ, set of assigned jobs $\mathcal{A}_0 = \varnothing$, $\mathcal{A}_1 = \{(i, 0) \colon i \in [n]\}$, set of received jobs $\mathcal{R}_0 = \varnothing$

for $t = 0, 1, 2, \cdots, T-1$ **do**
- once worker i_t finishes a job $(i_t, \pi_t) \in \mathcal{A}_{t+1}$(computing $g_{i_t}(x_{\pi_t})$), it sends $g_{i_t}(x_{\pi_t})$ to the server
- server updates the current model $x_{t+1} = x_t - \gamma g_{i_t}(x_{\pi_t})$ and the set $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \{(i_t, \pi_t)\}$
- server assigns worker $k_{t+1} \sim Uni[1, \cdots, n]$ to compute a gradient $g_{k_{t+1}}(x_{t+1})$
- server updates the set $\mathcal{A}_{t+2} = \mathcal{A}_{t+1} \cup \{(k_{t+1}, t+1)\}$

end for

1.4.2 Analysis

Proposition D. 1. Let Assumptions 1, 2, 3, and 4 hold. Let the stepsize satisfy $30L\tau_C\gamma \leq 1$, and $\tau = \lfloor \frac{1}{30L\gamma} \rfloor$. Then the iterates of Algorithm 2 satisfy

$$\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{F_1}{\gamma T} + L\gamma\sigma^2 + L\gamma\zeta^2 + L^2\tau_C^2\gamma^2G^2\right),$$

where $\hat{x}_t$ is chosen uniformly at random from $\{x_1, \ldots, x_T\}$ and $F_1 = f(y_1) - f^*$. Moreover, if we tune the stepsize, then the iterates of random asynchronous SGD satisfy

$$\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{LF_1\tau_C}{T} + \left(\frac{LF_1\sigma^2}{T}\right)^{1/2} + \left(\frac{LF_1\zeta^2}{T}\right)^{1/2} + \left(\frac{F_1L\tau_CG}{T}\right)^{2/3}\right).$$

Proof.

$$y_0 = x_0, \quad y_{t+1} = y_t - \gamma \sum_{(i,j)\in\mathcal{A}_{t+1}\setminus\mathcal{A}_t} g_i(x_j) \overset{t\geq 0}{=} y_t - \gamma g_{k_t}(x_t).$$

The purpose of this step is to reduce the upper bound when estimating the term about ζ in the case of random assigning process given G.

$$\widetilde{y}_1 = y_1, \quad \widetilde{y}_{t+1} = \begin{cases} \widetilde{y}_t - \gamma\nabla f(x_t) & \text{if } t \neq 0 \mod \tau, \\ y_{t+1} & \text{if } t = 0 \mod \tau. \end{cases}$$

(Equation 45)

$$\begin{aligned}\mathbb{E}\left[f(\widetilde{y}_{t+1})\right] \leq \quad & \mathbb{E}\left[f(\widetilde{y}_t)\right] - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\widetilde{y}_t)\|^2\right] - \frac{\gamma}{3}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + L^2\gamma\mathbb{E}\left[\|\widetilde{y}_t - y_t\|^2\right] \\ & + L^2\gamma\mathbb{E}\left[\|y_t - x_t\|^2\right] + \left(\frac{1}{240L}\mathbb{E}\left[\|\nabla f(\widetilde{y}_t)\|^2\right] + 61L\gamma^2\mathbb{E}\left[\|\Delta_t^t\|^2\right]\right)\xi_t\end{aligned}$$

where

$$\Delta_t^m := \sum_{j=r(t)}^{m} (\nabla f(x_j) - g_{k_j}(x_j)), \quad \xi_t = \begin{cases} 0, & \text{if } t \neq 0 \mod \tau, \\ 1, & \text{if } t = 0 \mod \tau. \end{cases}$$

$$B := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right]$$

$$C := \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\widetilde{y}_t)\|^2\right]$$

$$D := \sum_{t=0}^{T-1} \mathbb{E}\left[\|y_t - \widetilde{y}_t\|^2\right]$$

If given G, we have:
(Lemma D.1.&D.2.)

$$\mathbb{E}\left[\|y_t - x_t\|^2\right] \le \gamma^2(\tau_C - 1)^2 G^2 + (\tau_C - 1)\gamma^2\sigma^2.$$

(Lemma D.3.)

$$\mathbb{E}\left[\|\Delta_t^m\|^2\right] \le 3\tau\zeta^2 + 6L^2\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|x_j - x_{r(t)}\|^2\right] + \tau\sigma^2.$$

To prevent x_{π_t}, the second term should be split to $(x_j - y_j) + (y_j - y_{r(t)}) + (y_{r(t)} - x_{r(t)})$. Then, in the same way as Pure, we can get:

$$\mathbb{E}\left[\|\Delta_t^m\|^2\right] \le 4\tau\zeta^2 + \frac{2}{25}\left[(\tau_C - 1)^2 G^2 + (\tau_C - 1)\sigma^2\right] + \frac{1}{24}\tau \sum_{j=r(t)}^{m} \mathbb{E}\left[\|\nabla f(x_j)\|^2\right] + 2\tau\sigma^2$$

Below we estimate the two terms associated with ξ_t.

First term: (Equation 46)

$$\sum_{t=0}^{T-1} \frac{1}{240L}\mathbb{E}\left[\|\nabla f(\widetilde{x}_t)\|^2\right]\xi_t \le \frac{\gamma}{3600}B + \frac{\gamma}{2}C$$

Second term: (Equation 47)

$$L\gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\|\Delta_t^t\|^2 \xi_t \le 4L\gamma^2\zeta^2 T + \frac{1}{24}L\gamma^2\tau B + 2L\gamma^2\sigma^2 T + 5L^2\gamma^3 T\left[(\tau_C - 1)^2 G^2 + (\tau_C - 1)\sigma^2\right]$$

Now we're just left with the D terms.

$$\begin{aligned} D &= \gamma^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\|\Delta_t^{t-1}\|^2\right] \\ &\le 4\gamma^2\tau\zeta^2 T + \frac{1}{24}\gamma^2\tau^2 B + 4\gamma^2\tau\sigma^2 T + \frac{2}{25}\gamma^2 T\left[(\tau_C - 1)^2 G^2 + (\tau_C - 1)\sigma^2\right] \end{aligned}$$

At last, plug the two terms back and sum it up from 0 to $T-1$,

$$
\begin{aligned}
\mathbb{E}\left[f(\widetilde{y}_{T+1}) - f(\widetilde{y}_1)\right] \leq & -\frac{\gamma}{2}C - \frac{\gamma}{3}B + L^2\gamma D + L^2\gamma\sum_{t=1}^{T}\mathbb{E}\left[\|x_t - y_t\|^2\right] \\
& + \frac{1}{240L}\sum_{t=0}^{T-1}\xi_t\mathbb{E}\left[\|\nabla f(\widetilde{y}_t)\|^2\right] + 61L\gamma^2\sum_{t=0}^{T-1}\xi_t\mathbb{E}\left[\|\Delta_t^t\|^2\right] \\
\leq & -\frac{\gamma}{2}C - \frac{\gamma}{3}B \\
& + L^2\gamma\left(4\gamma^2\tau\zeta^2 T + \frac{1}{24}\gamma^2\tau^2 B + 4\gamma^2\tau\sigma^2 T + \frac{2}{25}\gamma^2 T\left[(\tau_C - 1)^2 G^2 + (\tau_C - 1)\sigma^2\right]\right) \\
& + \frac{\gamma}{3600}B + \frac{\gamma}{2}C \\
& + 4L\gamma^2\zeta^2 T + \frac{1}{24}L\gamma^2\tau B + 2L\gamma^2\sigma^2 T + 5L^2\gamma^3 T\left[(\tau_C - 1)^2 G^2 + (\tau_C - 1)\sigma^2\right] \\
\leq & -\frac{\gamma}{4}B + 5L\gamma^2 T\zeta^2 + 3L\gamma^2\sigma^2 T + 6L^2\gamma^3 T\left[(\tau_C - 1)^2 G^2 + (\tau_C - 1)\sigma^2\right]
\end{aligned}
$$

Let $F_1 := f(y_1) - f^*$, the final rate

$$
\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \leq \mathcal{O}\left(\frac{F_1}{\gamma T} + L\gamma\sigma^2 + L\gamma\zeta^2 + L^2\tau_C^2\gamma^2 G^2\right). \quad (30L\tau_C\gamma \leq 1)
$$