Asynchronous Distributed Semi-Stochastic Gradient Optimization

Appendix

Proof of Theorem 3

Before moving to the proof of Theorem 3, we need the following key lemmas.

Lemma 1 *Let i be a random sample drawn from $\mathcal{D}_p$ owned by worker p, then*

$$\mathbb{E}_i\|\nabla f_i(w) - \nabla f_i(w^*)\|^2 \leq 2L\left(F_p(w) - F_p(w^*) - \langle \nabla F_p(w^*), w - w^*\rangle\right)$$

Proof 1 *Similar to (Johnson and Zhang 2013), consider*

$$h_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^*\rangle.$$

It is easy to see that $h_i(w^*) = 0$ and $\nabla h_i(w^*) = 0$, thus

$$\begin{aligned} h_i(w^*) \leq & \min_\eta h_i(w - \eta\nabla h_i(w)) \\ \leq & \min_\eta \left(h_i(w) - \langle \nabla h_i(w), \eta\nabla h_i(w)\rangle + \frac{L\eta^2}{2}\|\nabla g_i(w)\|^2\right) \\ = & h_i(w) - \frac{1}{2L}\|\nabla h_i(w)\|^2, \end{aligned}$$

thus,

$$\|\nabla f_i(w) - \nabla f_i(w^*)\|^2 \leq 2L\left(f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^*\rangle\right).$$

Next, by taking expectation w.r.t $i \in \mathcal{D}_p$, we reach the inequality.

Lemma 2 *At a specific stage s and for a worker p, let*

$$g_i^* = \nabla f_i(w^*) - \nabla f_i(\tilde{w}^s) + \nabla F(\tilde{w}^s), i \in \mathcal{D}_p,$$

then the following inequality holds,

$$\mathbb{E}_p\left[\mathbb{E}_i\|g_i^*\|^2\right] \leq \ 2L\left[F(\tilde{w}^s) - F(w^*)\right],$$

Proof 2

$$\begin{aligned} \mathbb{E}_i\|g_i^*\|^2 = & \mathbb{E}_i\|\nabla f_i(\tilde{w}^s) - \nabla f_i(w^*) - \nabla F(\tilde{w}^s)\|^2 \\ = & \mathbb{E}_i\|\nabla f_i(\tilde{w}^s) - \nabla f_i(w^*)\|^2 - 2\mathbb{E}_i\langle \nabla F(\tilde{w}^s), \nabla f_i(\tilde{w}^s) - \nabla f_i(w^*)\rangle + \|\nabla F(\tilde{w}^s)\|^2 \\ \leq & 2L\left(F_p(\tilde{w}^s) - F_p(w^*) - \langle \nabla F_p(w^*), \tilde{w}^s - w^*\rangle\right) \\ & - 2\langle \nabla F(\tilde{w}^s), \nabla F_p(\tilde{w}^s) - \nabla F_p(w^*)\rangle + \|\nabla F(\tilde{w}^s)\|^2, \end{aligned}$$

where the inequality is the result of applying Lemma 1. Further, taking expectation on both sides w.r.t. worker p, we get

$$\begin{aligned} \mathbb{E}_p\left[\mathbb{E}_i\|g_i^*\|^2\right] = & 2L\mathbb{E}_p\left(F_p(\tilde{w}^s) - F_p(w^*) - \langle \nabla F_p(w^*), \tilde{w}^s - w^*\rangle\right) \\ & - 2\mathbb{E}_p\langle \nabla F(\tilde{w}^s), \nabla F_p(\tilde{w}^s) - \nabla F_p(w^*)\rangle + \|\nabla F(\tilde{w}^s)\|^2, \\ = & 2L\left(F(\tilde{w}^s) - F(w^*)\right) - 2\|\nabla F(\tilde{w}^s)\|^2 + \|\nabla F(\tilde{w}^s)\|^2, \\ \leq & 2L(F(\tilde{w}^s) - F(w^*)) \end{aligned}$$

Lemma 3 *(Feyzmahdavian, Aytekin, and Johansson 2014) Let V^t be a sequence satisfying*

$$V^{t+1} \leq pV^t + q\max_{t-\tau_t \leq k \leq t} V^k + E,$$

where p, q and E are positive constants. If $p + q < 1$ and $0 \leq \tau_t \leq \tau$, then

$$V^t \leq \rho^t V^0 + \epsilon,$$

where $\rho = (p+q)^{\frac{1}{1+\tau}}$ and $\epsilon = \frac{E}{1-p-q}$.

To make our proof self-contained, here we also include the proof for the lemma.

Proof 3 *We use induction to show that (3) holds for all $t \geq 0$. It is easy to verify that it holds for $t = 0$. Let's assume that it holds for $t > 0$, thus*

$$V^t \leq \rho^t V^0 + \epsilon, \tag{1}$$

$$V^k \leq \rho^k V^0 + \epsilon, k = t - \tau_t, \ldots, t. \tag{2}$$

Using (1) and (2), we have

$$\begin{aligned}
V^{t+1} \leq\ & pV^t + q \max_{t-\tau_t \leq k \leq t} V^k + E \\
\leq\ & p\rho^t V^0 + p\epsilon + q \max_{t-\tau_t \leq k \leq t} \rho^k V^0 + q\epsilon + E \\
\leq\ & p\rho^t V^0 + p\epsilon + q\rho^{t-\tau} V^0 + q\epsilon + E \\
=\ & (p + q\rho^{-\tau})\rho^t V^0 + \epsilon \\
\leq\ & (p + q)\rho^{-\tau}\rho^t V^0 + \epsilon \\
=\ & \rho^{t+1} V^0 + \epsilon
\end{aligned}$$

Proof of the main theorem

Here we turn to prove the main theorem. Let $\beta = \frac{\eta}{\theta}$. We denote $f_{p,i}$ the loss function associated with sample i owned by worker p, $g_{p,i}^t = \nabla f_{p,i}(w^t) - \nabla f_{p,i}(\tilde{w}) + \nabla F(\tilde{w})$. Let w^* denote the optimal solution. At a specific stage s, let w^0 be the initial value at the beginning of the stage. Recall that, the update rule of our algorithm can be written as

$$\begin{aligned}
\hat{w}^{t-\tau_t} =& w^{t-\tau_t} - \beta g_{p,i}^{t-\tau_t} \\
w^{t+1} =& (1-\theta)w^t + \theta \hat{w}^{t-\tau_t},
\end{aligned}$$

thus,

$$\begin{aligned}
F(w^{t+1}) - F(w^*) =\ & F((1-\theta)w^t + \theta \hat{w}^{t-\tau_t}) - F(w^*) \\
\leq\ & (1-\theta)[F(w^t) - F(w^*)] + \theta[F(\hat{w}^{t-\tau_t}) - F(w^*)],
\end{aligned} \tag{3}$$

where the inequality is the result from invoking the Jensen's inequality. We next bound the last term.

$$\begin{aligned}
F(\hat{w}^{t-\tau_t}) \leq\ & F(w^{t-\tau_t}) + \langle \nabla F(w^{t-\tau_t}), \hat{w}^{t-\tau_t} - w^{t-\tau_t} \rangle + \frac{L}{2}\|\hat{w}^{t-\tau_t} - w^{t-\tau_t}\|^2 \\
\leq\ & F(w^{t-\tau_t}) - \beta \langle \nabla F(w^{t-\tau_t}), g_{p,i}^{t-\tau_t} \rangle + \frac{\beta^2 L}{2}\|g_{p,i}^{t-\tau_t}\|^2 \\
\leq\ & F(w^{t-\tau_t}) - \beta \langle \nabla F(w^{t-\tau_t}), g_{p,i}^{t-\tau_t} \rangle + \beta^2 L \|g_{p,i}^{t-\tau_t} - g_{p,i}^*\|^2 + \beta^2 L \|g_{p,i}^*\|^2 \\
=\ & F(w^{t-\tau_t}) - \beta \langle \nabla F(w^{t-\tau_t}), g_{p,i}^{t-\tau_t} \rangle + \beta^2 L \|\nabla f_{p,i}(w^{t-\tau_t}) - \nabla f_{p,i}(w^*)\|^2 + \beta^2 L \|g_{p,i}^*\|^2.
\end{aligned} \tag{4}$$

Subtracting $F(w^*)$ and taking expectation on both sides w.r.t. i and then p, we reach

$$\begin{aligned}
\mathbb{E}\left[F(\hat{w}^{t-\tau_t}) - F(w^*)\right] \leq\ & F(w^{t-\tau_t}) - F(w^*) \\
& - \beta \|\nabla F(w^{t-\tau_t})\|^2 + \beta^2 L \mathbb{E}\left[\|\nabla f_{p,i}(w^{t-\tau_t}) - \nabla f_{p,i}(w^*)\|^2\right] + \beta^2 L \mathbb{E}\|g_{p,i}^*\|^2.
\end{aligned} \tag{5}$$

To the fourth term on the R.H.S, we apply Lemma 1, thus

$$\mathbb{E}\left[\|\nabla f_i(w^{t-\tau_t}) - \nabla f_i(w^*)\|^2\right] \leq\ 2L\left(F(w^{t-\tau_t}) - F(w^*)\right) \tag{6}$$

Next we bound the third term on the R.H.S. Using the strongly convexity of F, we know that

$$\begin{aligned}
F(w^*) \geq\ & F(w^{t-\tau_t}) + \langle \nabla F(w^{t-\tau_t}), w^* - w^{t-\tau_t} \rangle + \frac{\mu}{2}\|w^* - w^{t-\tau_t}\|^2 \\
\geq\ & F(w^{t-\tau_t}) + \langle \nabla F(w^{t-\tau_t}), \bar{x} - w^{t-\tau_t} \rangle + \frac{\mu}{2}\|\bar{x} - w^{t-\tau_t}\|^2,
\end{aligned} \tag{7}$$

where $\bar{x} = w^{t-\tau_t} - \frac{1}{\mu}\nabla F(w^{t-\tau_t})$ is the optimal solution of the quadratic function $Q(w) = \langle \nabla F(w^{t-\tau_t}), w - w^{t-\tau_t} \rangle + \frac{\mu}{2}\|w - w^{t-\tau_t}\|^2$, thus,

$$2\mu(F(w^{t-\tau_t}) - F(w^*)) \leq\ \|\nabla F(w^{t-\tau_t})\|^2. \tag{8}$$

Therefore, by combining (5), (6) and (8), as well as invoking lemma 2, we get

$$\begin{aligned}\mathbb{E}\left[F(\hat{w}^{t-\tau_t}) - F(w^*)\right] \leq\ & \left(1-2\mu\beta(1-\frac{\beta L^2}{\mu})\right)\mathbb{E}\left[F(w^{t-\tau_t}) - F(w^*)\right] \\ & + 2\beta^2L^2[F(\tilde{w}^s) - F(w^*)].\end{aligned} \tag{9}$$

Suppose $\beta \in (0, \frac{\mu}{L^2})$, it is easy to check that $\left(1-2\mu\beta(1-\frac{\beta L^2}{\mu})\right) \in (0,1)$.

Taking expection on both sizes of (3), and substituting (9) into (3), we get

$$\begin{aligned}\mathbb{E}[F(w^{t+1}) - F(w^*)] \leq\ & (1-\theta)\mathbb{E}[F(w^t) - F(w^*)] \\ & + \theta\left(1-2\mu\beta(1-\frac{\beta L^2}{\mu})\right)\mathbb{E}\left[F(w^{t-\tau_t}) - F(w^*)\right] \\ & + 2\theta\beta^2L^2[F(\tilde{w}^s) - F(w^*)],\end{aligned} \tag{10}$$

Let $p = 1-\theta$, $q = \theta\left(1-2\mu\beta(1-\frac{\beta L^2}{\mu})\right)$, and $E = 2\theta\beta^2L^2[F(\tilde{w}^s) - F(w^*)]$. By defining the sequence $V^t = \mathbb{E}[F(w^t) - F(w^*)]$, using the fact that $\beta = \frac{\eta}{\theta}$, and invoking Lemma 3, then running m iterations for each stage, we have

$$\mathbb{E}[F(w^m) - F(w^*)] \leq \rho^m[F(w^0) - F(w^*)] + \epsilon \tag{11}$$

where $\rho = \left(1-2\eta(\mu-\frac{\eta L^2}{\theta})\right)^{\frac{1}{1+\tau}}$ and $\epsilon = \frac{\eta L^2}{\theta\mu-\eta L^2}[F(\tilde{w}^s) - F(w^*)]$.

Let $\tilde{w}^{s+1} = w^m$ and $w^0 = \tilde{w}^s$. Suppose $\eta \in (0, \frac{\theta\mu}{2L^2})$, we know that $\frac{\eta L^2}{\theta\mu-\eta L^2} \in (0,1)$, therefore

$$\begin{aligned}\mathbb{E}[F(\tilde{w}^{s+1}) - F(w^*)] \leq\ & \rho^m[F(\tilde{w}^s) - F(w^*)] + \frac{\eta L^2}{\mu-\eta L^2}[F(\tilde{w}^s) - F(w^*)] \\ =\ & (\rho^m + \frac{\eta L^2}{\theta\mu-\eta L^2})[F(\tilde{w}^s) - F(w^*)] \\ \leq\ & \left(\rho^m + \frac{\eta L^2}{\theta\mu-\eta L^2}\right)^s[F(\tilde{w}^0) - F(w^*)] \\ =\ & \left(\left(1-2\eta(\mu-\frac{\eta L^2}{\theta})\right)^{\frac{m}{1+\tau}} + \frac{\eta L^2}{\theta\mu-\eta L^2}\right)^{s+1}[F(\tilde{w}^0) - F(w^*)]\end{aligned} \tag{12}$$

Proof of Lemma 4

Lemma 4 *On a specific stage s, let $\hat{\nabla} f_{p,i}(w) = \nabla f_i(w) - \nabla f_i(\tilde{w}^s) + \nabla F(\tilde{w}^s), i \in \mathcal{D}_p$. By taking expectation first w.r.t. on the samples of each worker and then on all the workers, we have*

$$\mathbb{E}\|\hat{\nabla} f_{p,i}(w)\| \leq 4L[F(w) - F(w^*) + F(\tilde{w}^S) - F(w^*)].$$

Proof 4

$$\mathbb{E}\|\hat{\nabla} f_{p,i}(w)\|^2 \leq 2\mathbb{E}\|\nabla f_i(w) - \nabla f_i(w^*)\|^2 + 2\mathbb{E}\|\nabla f_i(\tilde{w}^s) - \nabla f_i(w^*) - \nabla F(\tilde{w}^s)\|^2$$

Invoking Lemma 1 and 2, we reach the conclusion.

References

Feyzmahdavian, H. R.; Aytekin, A.; and Johansson, M. 2014. A delayed proximal gradient method with linear convergence rate. In *Proceedings of the International Workshop on Machine Learning for Signal Processing*, 1–6.

Johnson, R., and Zhang, T. 2013. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 315–323.