Supplementary Material: Asynchronous Stochastic Gradient Descent with Delay Compensation

A. Theorem 3.1 and Its Proof

Theorem 3.1:

Assume the loss function is L_1-Lipschitz. If $\lambda \in [0,1]$ make the following inequality holds,

$$\sum_{k=1}^{K} \frac{1}{\sigma_k^3(x, \boldsymbol{w}_t)} \geq 2\left[C_{ij}\left(\sum_{k=1}^{K}\frac{1}{\sigma_k(x,\boldsymbol{w}_t)}\right)^2 + C_{ij}^{'} L_1^2 |\epsilon_t|\right], \quad (1)$$

where $C_{ij} = \frac{1}{1+\lambda}(\frac{u_i u_j \beta}{l_i l_j \sqrt{\alpha}})^2$, $C_{ij}^{'} = \frac{1}{(1+\lambda)\alpha(l_i l_j)^2}$, and the model converges to the optimal model, then the MSE of $\lambda G(\boldsymbol{w}_t)$ is smaller than the MSE of $G(\boldsymbol{w}_t)$ in approximating Hessian $H(\boldsymbol{w}_t)$.

Proof:

For simplicity, we abbreviate $\mathbb{E}_{(Y|x,w^*)}$ as $\mathbb{E}$, G_t as $G(\mathbf{w}_t)$ and H_t as $H(\mathbf{w}_t)$. First, we calculate the MSE of G_t, λG_t to approximate H_t for each element of G_t. We denote the element in the i-th row and j-th column of $G(w_t)$ as G_{ij}^t and $H(w_t)$ as $H_{ij}(t)$.

The MSE of G_{ij}^t:

$$\mathbb{E}(G_{ij}^t - \mathbb{E}H_{ij}^t)^2 = \mathbb{E}(G_{ij}^t - \mathbb{E}G_{ij}^t)^2 + (\mathbb{E}H_{ij}^t - \mathbb{E}G_{ij}^t)^2 = \mathbb{E}(G_{ij}^t)^2 - (\mathbb{E}G_{ij}^t)^2 + \epsilon_t^2 \quad (2)$$

The MSE of λg_{ij}:

$$\begin{aligned}\mathbb{E}(\lambda G_{ij}^t - \mathbb{E}H_{ij}^t)^2 &= \lambda^2 \mathbb{E}(G_{ij}^t - \mathbb{E}G_{ij}^t)^2 + (\mathbb{E}H_{ij}^t - \lambda\mathbb{E}G_{ij}^t)^2 \\ &= \lambda^2\mathbb{E}(G_{ij}^t)^2 - \lambda^2(\mathbb{E}G_{ij}^t)^2 + (1-\lambda)^2(\mathbb{E}G_{ij}^t)^2 + \epsilon_t^2 + 2(\lambda-1)\mathbb{E}G_{ij}^t\epsilon_t \end{aligned} \quad (3)$$

The condition for $\mathbb{E}(G_{ij}^t - \mathbb{E}H_{ij}^t)^2 \geq \mathbb{E}(\lambda G_{ij}^t - \mathbb{E}H_{ij}^t)^2$ is

$$(1-\lambda^2)(\mathbb{E}(G_{ij}^t)^2 - (\mathbb{E}G_{ij}^t)^2) \geq 2(1-\lambda)(\mathbb{E}G_{ij}^t)^2 + 2(\lambda-1)\mathbb{E}G_{ij}^t\epsilon_t \quad (4)$$

Inequality (4) is equivalent to

$$(1+\lambda)\mathbb{E}(G_{ij}^t)^2 \geq 2[(\mathbb{E}G_{ij}^t)^2 - \mathbb{E}G_{ij}^t\epsilon_t] \quad (5)$$

Next we calculate $\mathbb{E}(G_{ij}^t)^2$, and $(\mathbb{E}G_{ij}^t)^2$ which appear in Eqn.(5). For simplicity, we denote $\sigma_k(x, \mathbf{w}_t)$ as σ_k, and $I_{[Y=k]}$

as z_k. Then we can get:

$$\mathbb{E}(g_{ij})^2 = \mathbb{E}_{(Y|x,\mathbf{w}_t)}\left(\frac{\partial}{\partial w_i}\log P(Y|x,\mathbf{w}_t)\right)^2\left(\frac{\partial}{\partial w_j}\log P(Y|x,\mathbf{w}_t)\right)^2 \tag{6}$$

$$\geq \mathbb{E}_{(Y|x,\mathbf{w}^*)}\left(\sum_{k=1}^{K}\left(-\frac{z_k}{\sigma_k}\right)\right)^4 (l_i l_j)^2$$

$$= \alpha\,(l_i l_j)^2\left(\sum_{k=1}^{K}\frac{1}{\sigma_k^3(x,\mathbf{w}_t)}\right) \tag{7}$$

$$(\mathbb{E}h_{ij})^2 = \left(\mathbb{E}_{(Y|x,\mathbf{w}^*)}\sum_{k=1}^{K}\frac{\partial\sigma_k}{\partial w_i}\left(-\frac{z_k}{\sigma_k}\right)\cdot\sum_{k=1}^{K}\frac{\partial\sigma_k}{\partial w_j}\left(-\frac{z_k}{\sigma_k}\right)\right)^2$$

$$\leq \beta^2\,(u_i u_j)^2\left(\sum_{k=1}^{K}\frac{1}{\sigma_k(x,\mathbf{w}_t)}\right)^2. \tag{8}$$

By substituting Ineq.(7) and Ineq.(8) into Ineq.(5), a sufficient condition for Ineq.(5) to be satisfied is $\sum_{k=1}^{K}\frac{1}{\sigma_k^3(x,\mathbf{w}_t)} \geq 2\left[C_{ij}\left(\sum_{k=1}^{K}\frac{1}{\sigma_k(x,\mathbf{w}_t)}\right)^2 + C_{ij}^{'}L_1^2|\epsilon_t|\right]$ because $G_{ij}^t \leq L_1^2$. $\square$

B. Corollary 3.2 and Its Proof

Corollary 3.2: *A sufficient condition for inequality (1) is* $\lambda \in [0,1]$ *and* $\exists k_0 \in [K]$ *such that* $\sigma_{k_0} \in \left[1-\frac{K-1}{2(C_{ij}K^2+C_{ij}^{'}L_1^2\epsilon_t)}, 1\right]$.

Proof:
Denote $\Delta = \frac{K-1}{2C_{ij}K^2}$ and $F(\sigma_1,...,\sigma_K) = \sum_{k=1}^{K}\frac{1}{\sigma_k^3(x,\mathbf{w}_t)} - 2C_{ij}\left(\sum_{k=1}^{K}\frac{1}{\sigma_k(x,\mathbf{w}_t)}\right)^2 - 2C_{ij}^{'}L_1^2|\epsilon_t|$. If $\exists k_1 \in [K]$ such that $\sigma_{k_1} \in [1-\Delta, 1]$, we have for $k \neq k_1$ $\sigma_k \in [0,\Delta]$. Therefore

$$F(\sigma_1,...,\sigma_K) \geq \frac{1}{(\sigma_{k_1})^3} + \frac{K-1}{\Delta^3} - 2C_{ij}\left(\frac{1}{\sigma_{k_1}} + \frac{K-1}{\Delta}\right)^2 - 2C_{ij}^{'}L_1^2|\epsilon_t| \tag{9}$$

$$\geq \frac{K-1}{\Delta^3} - 2C_{ij}\left(\left(\frac{K-1}{\Delta}\right)^2 + \frac{1}{\sigma_{k_1}^2} + \frac{2(K-1)}{\sigma_{k_1}\Delta}\right) - 2C_{ij}^{'}L_1^2|\epsilon_t| \tag{10}$$

$$\geq \frac{K-1}{\Delta^3} - 2C_{ij}\left(\frac{(K-1)^2}{\Delta^2} + \frac{2K-1}{\sigma_{k_1}\Delta}\right) - 2C_{ij}^{'}L_1^2|\epsilon_t| \tag{11}$$

$$= \frac{1}{\Delta}\left(\frac{K-1}{\Delta^2} - 2C_{ij}\left(\frac{(K-1)^2}{\Delta} + \frac{2K-1}{\sigma_{k_1}}\right)\right) - 2C_{ij}^{'}L_1^2|\epsilon_t| \tag{12}$$

$$\geq \frac{1}{\Delta}\left(\frac{K-1}{\Delta^2} - 2C_{ij}\left(\frac{(K-1)^2+2K-1}{\Delta}\right)\right) - 2C_{ij}^{'}L_1^2|\epsilon_t| \tag{13}$$

$$\geq \frac{1}{\Delta^2}\left(\frac{K-1}{\Delta} - 2C_{ij}K^2 - 2C_{ij}^{'}L_1^2|\epsilon_t|\right) \tag{14}$$

$$= 0 \tag{15}$$

where Ineq.(11) and (13) is established since $\sigma_{k_1} > \Delta$; and Eqn.(15) is established by putting $\Delta = \frac{K-1}{2(C_{ij}K^2+C_{ij}^{'}L_1^2|\epsilon_t|)}$ in Eqn.(14). $\square$

C. Uniform upper bound of MSE

Lemma C.1 *Assume the loss function is L_1-Lipschitz, and the diagonalization error of Hessian is upper bounded by ϵ_D, i.e., $||Diag(H(\boldsymbol{w}_t)) - H(\boldsymbol{w}_t)|| \leq \epsilon_D$,* [1] *then we have, for $\forall t$,*

$$mse^t(Diag(\lambda G)) \leq 4\lambda^2 V_1 + 4(1-\lambda)^2 L_1^4 + 4\epsilon_t^2 + 4\epsilon_D, \tag{16}$$

where V_1 is the upper bound of the variance of $G(\boldsymbol{w}_t)$.

Proof:

$$\begin{aligned}
&mse^t(Diag(\lambda G)) && (17)\\
\leq &\mathbb{E}\|Diag(\lambda G(w_t)) - H(w_t)\|^2 && (18)\\
\leq &4\mathbb{E}\|Diag(\lambda G(w_t)) - \mathbb{E}(Diag(\lambda G(w_t)))\|^2 + 4\|\mathbb{E}(Diag(\lambda G(w_t))) - \mathbb{E}(Diag(G(w_t)))\|^2 && (19)\\
&+ 4\|\mathbb{E}(Diag(G(w_t))) - \mathbb{E}(Diag(H(w_t)))\|^2 + 4\|\mathbb{E}(Diag(H(w_t))) - \mathbb{E}H(w_t)\|^2 && (20)\\
\leq &4\lambda^2 V_1 + 4(1-\lambda)^2 L_1^4 + 4\epsilon_t^2 + 4\epsilon_D && (21)
\end{aligned}$$

D. Convergence Rate for DC-ASGD: Convex Case

DC-ASGD is a general method to compensate delay in ASGD. We first show the convergence rate for convex loss function. If the loss function $f(w)$ is convex about w, we can add a regularization term $\frac{\rho}{2}\|w\|^2$ to make the objective function $F(w) + \frac{\rho}{2}\|w\|^2$ strongly convex. Thus, we assume that the objective function is μ-strongly convex.

Theorem 4.1: (Strongly Convex) *If $f(w)$ is L_2-smooth and μ-strongly convex about w, $\nabla f(w)$ is L_3-smooth about w and the expectation of the $\|\cdot\|_2^2$ norm of the delay compensated gradient is upper bounded by a constant G. By setting the learning rate $\eta_t = \frac{1}{\mu t}$, DC-ASGD has convergence rate as*

$$\mathbb{E}F(w_t) - F(w^*) \leq \frac{2L_2^2 G^2}{t\mu^4}(1 + 4\tau C_\lambda) + \frac{2G^2 L_2^2 \theta\sqrt{\tau}}{\mu^4 t\sqrt{t}} + \frac{L^3 L_2^3 \tau^2 G^3}{\mu^6 t^2},$$

where $\theta = \frac{2HKLG}{\mu}\sqrt{\frac{L_2}{\mu}(1 + \frac{\tau G L_3}{\mu L_2})}$ and $C_\lambda = (1-\lambda)L_1^2 + \epsilon_D$, and the expectation is taking with respect to the random sampling of DC-ASGD and $\mathbb{E}_{(y|x,w^*)}$.

Proof:

We denote $g^{dc}(w_t) = g(w_t) + \lambda g(w_t) \odot g(w_t) \odot (w_{t+\tau} - w_t)$, $g^h(w_t) = g(w_t) + H_{i_t}(w_t)(w_{t+\tau} - w_t)$ and $\nabla F^h(w_t) = \nabla F(w_t) + \mathbb{E}_{i_t} H_{i_t}(w_t)(w_{t+\tau} - w_t)$. Obviously, we have $\mathbb{E}g^h(w_t) = \nabla F^h(w_t)$. By the smoothness condition, we have

$$\begin{aligned}
&\mathbb{E}F(w_{t+\tau+1}) - F(w^*) && (22)\\
\leq\ & F(w_{t+\tau}) - F(w^*) - \langle \nabla F(w_{t+\tau}), w_{t+\tau+1} - w_{t+\tau}\rangle + \frac{L_2}{2}\|w_{t+\tau+1} - w_{t+\tau}\|^2 && (23)\\
\leq\ & F(w_{t+\tau}) - F(w^*) - \eta_{t+\tau}\langle \nabla F(w_{t+\tau}), g^{dc}(w_t)\rangle + \frac{L_2 \eta_{t+\tau}^2 G^2}{2} && (24)\\
=\ & F(w_{t+\tau}) - F(w^*) - \eta_{t+\tau}\langle \nabla F(w_{t+\tau}), \nabla F(w_{t+\tau})\rangle + \eta_{t+\tau}\langle \nabla F(w_{t+\tau}), \nabla F(w_{t+\tau}) - \nabla F^h(w_t)\rangle && (25)\\
& + \eta_{t+\tau}\langle \nabla F(w_{t+\tau}), \mathbb{E}g^h(w_t) - g^{dc}(w_t)\rangle + \frac{L_2 \eta_{t+\tau}^2 G^2}{2} && (26)
\end{aligned}$$

Since $f(w)$ is L_2-smooth and μ strongly convex, we have

$$-\langle \nabla F(w_{t+\tau}), \nabla F(w_{t+\tau})\rangle \leq -\mu^2\|w_{t+\tau} - w^*\|^2 \leq -\frac{2\mu^2}{L_2}(F(w_{t+\tau}) - F(w^*)). \tag{27}$$

[1](LeCun, 1987) demonstrated that the diagonal approximation to Hessian for neural networks is an efficient method with no much drop on accuracy

For the term $\eta_{t+\tau}\langle\nabla F(w_{t+\tau}),\nabla F(w_{t+\tau})-\nabla F^h(w_t)\rangle$, we have

$$\begin{aligned}&\eta_{t+\tau}\langle\nabla F(w_{t+\tau}),\nabla F(w_{t+\tau})-\nabla F^h(w_t)\rangle &(28)\\ \leq\ &\eta_{t+\tau}\|\nabla F(w_{t+\tau})\|\|\nabla F(w_{t+\tau})-\nabla F^h(w_t)\| &(29)\\ \leq\ &\eta_{t+\tau}G\|\nabla F(w_{t+\tau})-\nabla F^h(w_t)\| &(30)\end{aligned}$$

By the smoothness condition for $\nabla F(w)$, we have

$$\|\nabla F(w_{t+\tau})-\nabla F^h(w_t)\|\leq\frac{L_3}{2}\|w_{t+\tau}-w_t\|^2\leq\frac{L_3\tau G^2}{2}\sum_{j=0}^{\tau-1}\eta_{t+j}^2 \tag{31}$$

Let $\eta_t=\frac{L_2}{\mu^2 t}$, we can get $\sum_{j=1}^{\tau}\eta_{t+j}^2\leq\frac{L_2^2}{\mu^4}\cdot\frac{\tau}{t(t+\tau)}\leq\frac{2L_2^2\tau}{\mu^4(t+\tau)^2}$.

For the term $\eta_{t+\tau}\langle\nabla F(w_{t+\tau}),\mathbb{E}g^h(w_t)-g^{dc}(w_t)\rangle$, we have

$$\langle\nabla F(w_{t+\tau}),\mathbb{E}(g^h(w_t)-g^{dc}(w_t))\rangle \tag{32}$$

$$\leq\|\nabla F(w_{t+\tau})\|\|\mathbb{E}(\lambda g(w_t)\odot g(w_t)-H(w_t))(w_{t+\tau}-w_t)\| \tag{33}$$

$$\leq G^2\tau\sum_{j=0}^{\tau-1}\eta_{t+j}(\|\mathbb{E}(\lambda g(w_t)\odot g(w_t)-g(w_t)\odot g(w_t)\|+\|g(w_t)\odot g(w_t)-Diag(H(w_t))\|+\|Diag(H(w_t))-H(w_t)\|) \tag{34}$$

$$\leq\frac{2G^2L_2\tau}{(t+\tau)\mu^2}(C_\lambda+\epsilon_t), \tag{35}$$

where $C_\lambda=(1-\lambda)L_1^2+\epsilon_D$.

Using Lemma F.1, $\epsilon_t\leq\theta\sqrt{\frac{1}{t}}\leq\theta\sqrt{\frac{\tau}{t+\tau}}$. Putting inequality 27 and 31 in inequality 26, we have

$$\mathbb{E}F(w_{t+\tau+1})-F(w^*)\leq\left(1-\frac{2}{t+\tau}\right)(\mathbb{E}F(w_t)-F(w^*))+\frac{L_3L_2^3\tau^2G^3}{\mu^6(t+\tau)^3} \tag{36}$$

$$+\frac{2G^2L_2^2\tau}{\mu^4(t+\tau)^2}\left(C_\lambda+\theta\sqrt{\frac{\tau}{t+\tau}}\right)+\frac{L_2^2G^2}{2(t+\tau)^2\mu^4} \tag{37}$$

We can get

$$\mathbb{E}F(w_t)-F(w^*)\leq\frac{2L_2^2G^2}{t\mu^4}(1+4\tau C_\lambda)+\frac{2G^2L_2^2\theta\sqrt{\tau}}{\mu^4t\sqrt{t}}+\frac{L^3L_2^3\tau^2G^3}{\mu^6t^2}. \tag{38}$$

by induction. □

Discussion:

(1). Following the above proof steps and using $\|\nabla F(w_{t+\tau})-\nabla F(w_t)\|\leq L_2\|w_{t+\tau}-w_t\|$, we can get the convergence rate of ASGD is

$$\mathbb{E}F(w_t)-F(w^*)\leq\frac{2L_2^2G^2}{t\mu^4}(1+4\tau L_2). \tag{39}$$

Compared the convergence rate of DC-ASGD with ASGD, the extra term $\frac{2G^2L_2^2\theta\sqrt{\tau}}{\mu^4t\sqrt{t}}+\frac{L^3L_2^3\tau^2G^3}{\mu^6t^2}$ converge to zero faster than $\frac{2L_2^2G^2}{t\mu^4}(1+4\tau C_\lambda)$ in terms of the order of t. Thus, when t is large, the extra term has smaller value. We assume that t is large and the term can be neglected. Then the condition for DC-ASGD outperforming ASGD is $L_2>C_\lambda$.

E. Convergence Rate for DC-ASGD: Nonconvex Case

Theorem 5.1: (Nonconvex Case) *Assume that Assumptions 1-4 hold. Set the learning rate*

$$\eta_t=\sqrt{\frac{2(F(w_1)-F(w^*))}{bTV^2L_2}}, \tag{40}$$

where b is the mini-batch size, and V is the upper bound of the variance of the delay-compensated gradient. If $T \geq \max\{\mathcal{O}(1/r^4), 2D_0bL_2/V^2\}$ and delay τ is upper-bounded as below,

$$\tau \leq \min\left\{\frac{L_2V}{C_\lambda}\sqrt{\frac{L_2T}{2D_0b}}, \frac{V}{C_\lambda}\sqrt{\frac{L_2T}{2D_0b}}, \frac{TV}{\tilde{C}}\sqrt{\frac{L_2}{bD_0}}, \frac{VL_2T}{4\tilde{C}}\sqrt{\frac{TL_2}{2D_0b}}\right\}. \tag{41}$$

then DC-ASGD has the following ergodic convergence rate,

$$\min_{t=\{1,\cdots,T\}} \mathbb{E}(\|\nabla F(\boldsymbol{w}_t)\|^2) \leq V\sqrt{\frac{2D_0L_2}{bT}}, \tag{42}$$

where the expectation is taken with respect to the random sampling in SGD and the data distribution $P(Y|x, \boldsymbol{w}^*)$.

Proof:

We denote $g_m(w_t) + \lambda g_m(w_t) \odot g_m(w_t) \odot (w_{t+\tau} - w_t)$ as $g_m^{dc}(w_t)$ where $m \in \{1, \cdots, b\}$ is the index of instances in the minibatch. From the proof the Theorem 1 in ASGD (Lian et al., 2015), we can get

$$\begin{aligned}
& \mathbb{E}F(w_{t+\tau+1}) - F(w_{t+\tau}) && (43)\\
\leq\ & \langle\nabla F(w_{t+\tau}), w_{t+\tau} - w_t\rangle + \frac{L_2}{2}\|w_{t+\tau+1} - w_{t+\tau}\|^2 && (44)\\
\leq\ & -\eta_{t+\tau}\langle\nabla F(w_{t+\tau}), \sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\rangle + \frac{\eta_{t+\tau}^2L_2}{2}\mathbb{E}\left(\left\|\sum_{m=1}^{b} g_m^{dc}(w_t)\right\|^2\right) && (45)\\
\leq\ & -\frac{b\eta_{t+\tau}}{2}\left(\|\nabla F(w_{t+\tau})\|^2 + \left\|\sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\right\|^2 - \left\|\nabla F(w_{t+\tau}) - \sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\right\|^2\right) \\
& + \frac{\eta_{t+\tau}^2L_2}{2}\mathbb{E}\left(\left\|\sum_{m=1}^{b} g_m^{dc}(w_t)\right\|^2\right) && (46)
\end{aligned}$$

For the term $T_1 = \left\|\nabla F(w_{t+\tau}) - \sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\right\|^2$, by using the smooth condition of g, we have

$$\begin{aligned}
T_1 &= \left\|\nabla F(w_{t+\tau}) - \sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\right\|^2 && (47)\\
&\leq \left\|\nabla F(w_{t+\tau}) - \nabla F^h(w_t) + \nabla F^h(w_t) - \sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\right\|^2 && (48)\\
&\leq 2\left\|\frac{L_3}{2}\|w_{t+\tau} - w_t\|^2\right\|^2 + 2\left\|\nabla F^h(w_t) - \sum_{m=1}^{b}\mathbb{E}g_m^{dc}(w_t)\right\|^2 && (49)\\
&\leq (L_3^2\pi^2/2 + 2(((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2))\|w_{t+\tau} - w_t\|^2 && (50)
\end{aligned}$$

Thus by following the proof of ASGD, we have

$$\mathbb{E}(T_1) \leq 4(L_3^2\pi^2/4 + ((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)\left(b\tau\eta_{t+\tau}^2V^2 + \tau^2\eta_{t+\tau}^2\left\|b\mathbb{E}g_m^{dc}(w_t)\right\|^2\right). \tag{51}$$

For the term $T_2 = \mathbb{E}\left(\left\|\sum_{m=1}^{b} g_m^{dc}(w_t)\right\|^2\right)$, it has

$$\mathbb{E}(T_2) \leq bV^2 + \left\|b\mathbb{E}g_m^{dc}(w_t)\right\|^2. \tag{52}$$

By putting Ineq.(51) and Ineq.(52) in Ineq.(46), we can get

$$\begin{aligned} & \mathbb{E}(F(w_{t+\tau+1}) - F(w_{t+\tau}) && (53)\\ \leq \quad & -\frac{b\eta_{t+\tau}}{2}\mathbb{E}\|\nabla F(w_{t+\tau})\|^2 + \left(\frac{\eta_{t+\tau}^2 L_2}{2} - \frac{\eta_{t+\tau}}{2b}\right)\mathbb{E}\left(\left\|b\mathbb{E}g_m^{dc}(w_t)\right\|^2\right) \\ & + \left(\frac{\eta_{t+\tau}^2 bL_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)b^2\tau\eta_{t+\tau}^3\right)V^2 && (54)\\ & + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)b\tau^2\eta_{t+\tau}^3\mathbb{E}\left(\left\|b\mathbb{E}g_m^{dc}(w_t)\right\|^2\right) && (55) \end{aligned}$$

Summarizing the Ineq.(55) from $t = 1$ to $t + \tau = T$, we have

$$\begin{aligned} & \mathbb{E}F(w_{T+1}) - F(w_1) && (56)\\ \leq & -\frac{b}{2}\sum_{t=1}^{T}\eta_t\mathbb{E}\|\nabla F(w_t)\|^2 + \sum_{t=1}^{T}\left(\frac{\eta_{t+\tau}^2 bL_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)b^2\tau\eta_{t+\tau}^3\right)V^2 && (57)\\ & + \sum_{t=1}^{T}\left(\frac{\eta_t^2 L_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)b\tau^2\eta_t^3 - \frac{\eta_t}{2b}\right)\mathbb{E}\left\|b\mathbb{E}g_m^{dc}(w_{\max\{t-\tau,1\}})\right\|^2. && (58) \end{aligned}$$

By Lemma F.1 and under our assumptions, we have when $t > T_0$, w_t will goes into a strongly convex neighbourhood of some local optimal w_{loc}. Thus, $\epsilon_t \leq \epsilon_{nc} + \theta\sqrt{1/(t-T_0)}$, when $t > T_0$ and $\epsilon_t < \max_{s\in 1,\cdots,T_0}\epsilon_s$ when $t < T_0$.

Let $\eta_t = \sqrt{\frac{2(F(w_1)-F(w^*)}{bTV^2L_2}}$. It follows that

$$\begin{aligned} & \sum_{t=1}^{T}\frac{\eta_t L_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)b\tau^2\eta_t^2 && (59)\\ \leq & \sum_{t=1}^{T}\left\{\frac{\eta_t L_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + 2\epsilon_{nc}^2)b\tau^2\eta_t^2\right\} + 2b\tau^2\eta_t^2(4T_0\max_{s\in 1,\cdots,T_0}(\epsilon_s)^2 + 4\theta^2\log(T-T_0)) && (60) \end{aligned}$$

We ignore the $\log(T - T_0)$ term and regards $\tilde{C}^2 = 4T_0\max_{s\in 1,\cdots,T_0}(\epsilon_s)^2 + 4\theta^2\log(T-T_0)$ as a constant, which yields

$$\begin{aligned} & \sum_{t=1}^{T}\frac{\eta_t L_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + \epsilon_t^2)b\tau^2\eta_t^2 && (61)\\ \leq & \sum_{t=1}^{T}\left\{\frac{\eta_t L_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + 2\epsilon_{nc}^2)b\tau^2\eta_t^2\right\} + 2\tau^2\eta_t^2 b\tilde{C}^2 && (62) \end{aligned}$$

η_t should be set to make

$$\sum_{t=1}^{T}\left(\frac{\eta_t^2 L_2}{2} + (L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + 2\epsilon_{nc}^2)b\tau^2\eta_t^3 + \frac{2\tau^2\eta_t^3 b\tilde{C}^2}{T} - \frac{\eta_t}{2b}\right) \leq 0. \quad (63)$$

Then we can get

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w_t)\|^2 && (64)\\ \leq & \frac{2(F(w_1) - F(w^*) + Tb(\eta_t^2 L_2 + 2(L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + 2\epsilon_{nc}^2)b\tau\eta_t^3)V^2 + \frac{\eta_t^3\tilde{C}^2 4b\tau}{T}V^2}{bT\eta_t} && (65)\\ \leq & \frac{2(F(w_1) - F(w^*)}{bT\eta_t} + (\eta_t L_2 + 2(L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + 2\epsilon_{nc}^2)b\tau\eta_t^2)V^2 + \frac{\eta_t^2\tilde{C}^2 4b\tau V^2}{T} && (66)\\ & && (67) \end{aligned}$$

We set η_t to make

$$(2(L_3^2\pi^2/2 + 2((1-\lambda)L_1^2 + \epsilon_D)^2 + 2\epsilon_{nc}^2)b\tau\eta_t^2) + \frac{\eta_t^2\tilde{C}^2 4b\tau}{T} \le \eta_t L_2 \tag{68}$$

Thus let $\eta_t = \sqrt{\frac{2(F(w_1)-F(w^*)}{bTV^2L_2}}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w_t)\|^2 \le V\sqrt{\frac{2D_0L_2}{bT}}. \tag{69}$$

And we can get the condition for T by putting η in ineq.63 and ineq.68, we can get that

$$\tau \le \min\left\{\frac{L_2V}{C_\lambda}\sqrt{\frac{L_2T}{2D_0b}}, \frac{V}{C_\lambda}\sqrt{\frac{L_2T}{2D_0b}}, \frac{TV}{\tilde{C}}\sqrt{\frac{L_2}{bD_0}}, \frac{VL_2T}{4\tilde{C}}\sqrt{\frac{TL_2}{2D_0b}}\right\}. \tag{70}$$

F. Decreasing rate of the approximation error ϵ_t

Since ϵ_t is contained the proof of the convergence rate for DC-ASGD , in this section we will introduce a lemma which describes the approximation error ϵ_t the for both convex and nonconvex cases.

Lemma F.1 *Assume that the true label y is generated according to the distribution $\mathbb{P}(Y=k|x,w^*) = \sigma_k(x,w^*)$ and $f(x,y,\boldsymbol{w}) = -\sum_{k=1}^{K}(I_{[y=k]}\log\sigma_k(x;\boldsymbol{w}))$. If we assume that the loss function is μ-strongly convex about w. We denote $\boldsymbol{w}_t$ is the output of DC-ASGD by using the outerproduct approximation of Hessian, we have*

$$\epsilon_t = \left|\mathbb{E}_{(x,y|\boldsymbol{w}^*)}\frac{\partial^2}{\partial \boldsymbol{w}^2}f(x,y,\boldsymbol{w}_t) - \mathbb{E}_{(x,y|\boldsymbol{w}^*)}\left(\frac{\partial}{\partial \boldsymbol{w}}f(x,y,\boldsymbol{w}_t)\right)\otimes\left(\frac{\partial}{\partial \boldsymbol{w}}f(x,y,\boldsymbol{w}_t)\right)\right| \le \theta\sqrt{\frac{1}{t}},$$

where $\theta = \frac{2HKLVL_2}{\mu^2}\sqrt{\frac{1}{\mu}(1+\frac{L_2+\lambda L_1^2}{L_2}\tau)}$.

If we assume that the loss function is μ-strongly convex in a neighborhood of each local optimal $d(\boldsymbol{w}_{loc}, r)$, $\left|\frac{\partial^2\mathbb{P}(Y=k|x,\boldsymbol{w})}{\partial^2\boldsymbol{w}} \times \frac{1}{P(Y=k|x,w)}\right| \le H$, $\forall k,x,w$, each $\sigma_k(\boldsymbol{w})$ is L-Lipschitz continuous about $\boldsymbol{w}$. We denote $\boldsymbol{w}_t$ is the output of DC-ASGD by using the outerproduct approximation of Hessian, we have

$$\epsilon_t = \left|\mathbb{E}_{(x,y|\boldsymbol{w}^*)}\frac{\partial^2}{\partial \boldsymbol{w}^2}f(x,y,\boldsymbol{w}_t) - \mathbb{E}_{(x,y|\boldsymbol{w}^*)}\left(\frac{\partial}{\partial \boldsymbol{w}}f(x,y,\boldsymbol{w}_t)\right)\otimes\left(\frac{\partial}{\partial \boldsymbol{w}}f(x,y,\boldsymbol{w}_t)\right)\right| \le \theta\sqrt{\frac{1}{t-T_0}} + \epsilon_{nc}.$$

where $t > T_0 \ge \mathcal{O}(\frac{1}{r^8})$.

Proof:

$$\begin{aligned}
\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\partial^2}{\partial \mathbf{w}^2}f(x,Y,\mathbf{w}_t) &= -\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\partial^2}{\partial \mathbf{w}^2}\left(\sum_{k=1}^{K}(I_{[y=k]}\log\sigma_k(x;\mathbf{w}_t))\right)\\
&= -\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\partial^2}{\partial \mathbf{w}^2}\log\left(\prod_{k=1}^{K}\sigma_k(x,\mathbf{w}_t)^{I_{[y=k]}}\right)\\
&= -\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\partial^2}{\partial \mathbf{w}^2}\log\mathbb{P}(y|x,\mathbf{w}_t)\\
&= -\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)} + \mathbb{E}_{(y|x,\mathbf{w}^*)}\left(\frac{\frac{\partial}{\partial\omega}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)}\right)^2\\
&= -\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)} + \mathbb{E}_{(y|x,\mathbf{w}^*)}\left(\frac{\partial}{\partial\omega}\log\mathbb{P}(y|x,\mathbf{w}_t)\right)^2.\\
&= -\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)} + \mathbb{E}_{(y|x,\mathbf{w}^*)}\left(\frac{\partial}{\partial\omega}f(x,Y,\mathbf{w}_t)\right)^2.
\end{aligned} \tag{71}$$

Since $\mathbb{E}_{(y|x,\mathbf{w}_t)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)}=0$ by the two equivalent methods to calculating fisher information matrix (Friedman et al., 2001), we have

$$\begin{aligned}\left|\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)}\right| &= \left|\mathbb{E}_{(y|x,\mathbf{w}^*)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)} - \mathbb{E}_{(y|x,\mathbf{w}_t)}\frac{\frac{\partial^2}{\partial\omega^2}\mathbb{P}(y|x,\mathbf{w}_t)}{\mathbb{P}(y|x,\mathbf{w}_t)}\right| \\ &= \left|\sum_{k=1}^{K}\frac{\partial^2}{\partial\omega^2}\mathbb{P}(Y=k|X=x,\mathbf{w}_t)\times\frac{\mathbb{P}(Y=k|x,\mathbf{w}^*)-\mathbb{P}(Y=k|x,\mathbf{w}_t)}{\mathbb{P}(Y=k|x,\mathbf{w}_t)}\right| \quad (72)\\ &\le H\cdot\sum_{k=1}^{K}|\mathbb{P}(Y=k|x,\mathbf{w}^*)-\mathbb{P}(Y=k|x,\mathbf{w}_t)| \\ &\le HKL\|\mathbf{w}_t-\mathbf{w}_{loc}\| + HK\max_{k=1,\cdots,K}|\mathbb{P}(Y=k|x,\mathbf{w}_{loc})-\mathbb{P}(Y=k|x,\mathbf{w}^*)| \quad (73)\\ &\le HKL\|\mathbf{w}_t-\mathbf{w}_{loc}\|+\epsilon_{nc}. \quad (74)\end{aligned}$$

For strongly convex objective functions, $\epsilon_{nc}=0$ and $w_{loc}=w^*$. The only thing we need is to prove the convergence of DC-ASGD without using the information of ϵ_t like before. By the smoothness condition, we have

$$\begin{aligned} & \mathbb{E}F(w_{t+\tau+1})-F(w^*) \quad (75)\\ \le\ & F(w_{t+\tau})-F(w^*)-\eta_{t+\tau}\langle\nabla F(w_{t+\tau}),\mathbb{E}g^{dc}(w_t)\rangle+\frac{L_2\eta_{t+\tau}^2V^2}{2} \quad (76)\\ =\ & F(w_{t+\tau})-F(w^*)-\eta_{t+\tau}\langle\nabla F(w_{t+\tau}),\nabla F(w_{t+\tau})\rangle \quad (77)\\ & +\eta_{t+\tau}\langle\nabla F(w_{t+\tau}),\nabla F(w_{t+\tau})-\mathbb{E}g^{dc}(w_t)\rangle+\frac{L_2\eta_{t+\tau}^2V^2}{2} \quad (78)\\ \le\ & (1-\frac{2\eta_{t+\tau}\mu^2}{L_2})(F(w_{t+\tau})-F(w^*))+\eta_{t+\tau}\|\nabla F(w_{t+\tau})\|\|\nabla F(w_{t+\tau})-\mathbb{E}g^{dc}(w_t)\|+\frac{L_2\eta_{t+\tau}^2V^2}{2} \quad (79)\\ \le\ & (1-\frac{2\eta_{t+\tau}\mu^2}{L_2})(F(w_{t+\tau})-F(w^*))+\eta_{t+\tau}V\cdot(L_2+\lambda L_1^2)\|w_{t+\tau}-w_t\|+\frac{L_2\eta_{t+\tau}^2V^2}{2} \quad (80)\\ \le\ & (1-\frac{2\eta_{t+\tau}\mu^2}{L_2})(F(w_{t+\tau})-F(w^*))+\eta_{t+\tau}V\cdot(L_2+\lambda L_1^2)\|\sum_{j=1}^{\tau}\eta_{t+\tau-j}g^{dc}(w_t)\|+\frac{L_2\eta_{t+\tau}^2V^2}{2} \quad (81)\end{aligned}$$

Taking expectation to the above inequality, we can get

$$\begin{aligned}\mathbb{E}F(w_{t+\tau+1})-F(w^*) &\le (1-\frac{2\eta_{t+\tau}\mu^2}{L_2})(\mathbb{E}F(w_{t+\tau})-F(w^*))+\frac{\eta_{t+\tau}^2(L_2+\lambda L_1^2)V^2\tau}{2}+\frac{L_2\eta_{t+\tau}^2V^2}{2} \quad (82)\\ &\le (1-\frac{2\eta_{t+\tau}\mu^2}{L_2})(\mathbb{E}F(w_{t+\tau})-F(w^*))+\frac{\eta_{t+\tau}^2V^2L_2}{2}(1+\frac{L_2+\lambda L_1^2}{L_2}\tau). \quad (83)\end{aligned}$$

Let $\eta_t=\frac{L_2}{\mu^2 t}$, we have

$$\mathbb{E}F(w_{t+1})-F(w^*)\le\left(1-\frac{2}{t}\right)(\mathbb{E}F(w_t)-F(w^*))+\frac{V^2L_2^2}{2\mu^4t^2}\left(1+\frac{L_2+\lambda L_1^2}{L_2}\tau\right). \quad (84)$$

We can get

$$\mathbb{E}F(w_t)-F(w^*)\le\frac{2L_2^2V^2}{t\mu^4}\left(1+\frac{L_2+\lambda L_1^2}{L_2}\tau\right). \quad (85)$$

by induction. Then we can get

$$\|w_t-w^*\|^2\le\frac{4L_2^2V^2}{t\mu^5}\left(1+\frac{L_2+\lambda L_1^2}{L_2}\tau\right). \quad (86)$$

By putting Ineq.86 into Ineq.73, we can get the result in the theorem.

For nonconvex case, if $\mathbf{w}_t\in\mathcal{B}(\mathbf{w}_{loc},r)$, we have $\mathbb{E}(\mathbf{w}_t-\mathbf{w}_{loc})\le\frac{1}{\mu}\mathbb{E}\nabla F(\mathbf{w}_t)$ under the assumptions. Next we will prove that, for nonconvex loss function $f(x,y,\mathbf{w}_t)$, DC-ASGD has ergodic convergence rate. $\min_{t=1,\cdots,T}\mathbb{E}\|\frac{\partial}{\partial\mathbf{w}_t}F(x,y,\mathbf{w}_t)\|^2=\mathcal{O}(1/\sqrt{T})$, where the expectation is taking with respect to the stochastic sampling.

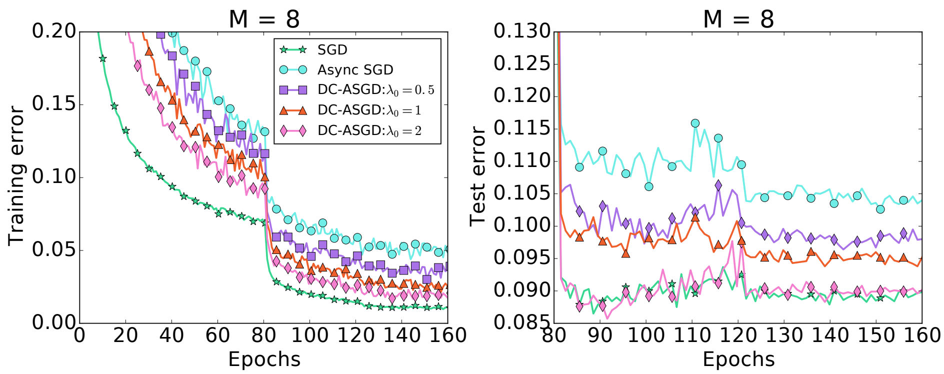

Figure 1. Error rates of the global model with Different λ_0 w.r.t. number of effective passes on CIFAR-10

Compared with the proof of ASGD (Lian et al., 2015), DC-ASGD with Hessian approximation has

$$\begin{aligned} T_1 &= \|\nabla F(w_{t+\tau}) - \mathbb{E}g^{dc}(w_t)\|^2 && (87)\\ &= \|\nabla F(w_{t+\tau}) - \nabla F(w_t) - \lambda \mathbb{E}g(w_t)\odot g(w_t)\cdot(w_{t+\tau}-w_t)\|^2 && (88)\\ &\le 2\|\nabla F(w_{t+\tau}) - \nabla F(w_t)\|^2 + 2\|\lambda \mathbb{E}g(w_t)\odot g(w_t)\cdot(w_{t+\tau}-w_t)\|^2 && (89)\\ &\le 2(L_2^2+\lambda^2 L_1^4)\|w_{t+\tau}-w_t\|^2, && (90)\end{aligned}$$

since L_1 is the upper bound of $\nabla f(w)$ and L_2 is the smooth coefficient of $f(w)$. Suppose that $\eta = \sqrt{\frac{2D_0}{bTV^2L_2}}$ and τ is upper bounded as Theorem 5.1,

$$\min_{t=1,\cdots,T} \mathbb{E}\|\nabla F(w_t)\|^2 \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w_t)\|^2 \le \mathcal{O}(\frac{1}{T^{1/2}}). \quad (91)$$

Referring to a recent work of Lee $et.al$ (Lee et al., 2016), GD with a random initialization and sufficiently small constant step size converges to a local minimizer almost surely under the assumptions in Theorem 1.2. Thus, the assumption that $F(w)$ is μ-strongly convex in the r-neighborhood of arbitrary local minimum w_{loc} is easily to be satistied with probability one. By the L_1-Lipschitz assumption, we have $P(Y=k|x,w_t) - P(Y=k|x,w_{loc}) \le L_1\|w_t - w_{loc}\|$. By the L_2-smooth assumption, we have $L_2\|w_t - w_{loc}\|^2 \ge \langle \nabla F(w_t), w_t - w_{loc}\rangle$. Thus for $w_t \in \mathcal{B}(w_{loc}, r)$, we have $\|\nabla F(w_t)\| \le L_2\|w_t - w_{loc}\| \le L_2 r$. By the continuously twice differential assumption, we can assume that $\|\nabla F(w_t)\| \le L_2\|w_t - w_{loc}\| \le L_2 r$ for $w_t \in \mathcal{B}(w_{loc}, r)$ and $\|\nabla F(w_t)\| \le L_2\|w_t - w_{loc}\| > L_2 r$ for $w_t \notin \mathcal{B}(w_{loc}, r)$ without loss of generality [2]. Therefore $\min_{t=1,\cdots,T}\mathbb{E}\|\nabla F(w_t)\|^2 \le L_2^2 r^2$ is a sufficient condition for $\mathbb{E}\|w_T - w_{loc}\| \le r$.

$$\min_{t=1,\cdots,T_0} \mathbb{E}\|\nabla F(w_t)\|^2 \le \mathcal{O}(\frac{1}{T_0^{1/2}}) \le r^2. \quad (92)$$

We have $T_0 \ge \mathcal{O}\left(\frac{1}{r^4}\right)$.

Thus we have finished the proof for nonconvex case.

G. Experimental Results on the Influence of λ

In this section, we show how the parameter λ affect our DC-ASGD algorithm. We compare the performance of respectively sequential SGD, ASGD and DC-ASGD-a with different value of initial λ_0 [3]. The results are given in Figure 1. This experiment reflects to the discussion in Section 5, too large value of this parameter ($\lambda_0 > 2$ in this setting) will introduce large variance and lead to a wrong gradient direction, meanwhile too small will make the compensation influence nearly disappear. As λ decreasing, DC-ASGD will gradually degrade to ASGD. A proper λ will lead to significant better accuracy.

[2] We can choose r small enough to make it satisfied.

[3] We also compare different λ_0 for DC-ASGD-c and the results are very similar to DC-ASGD-a.

H. Large Mini-batch Synchronous SGD with Delay-Compensated Gradient

In this section, we discuss how delay-compensated gradient can be used in synchronous SGD. The effective mini-batch size in SSGD is usually enlarged M times comparing with sequential SGD. A learning rate scaling trick is commonly used to overcome the influence of large mini-batch size in SSGD (Goyal et al., 2017): when the mini-batch size is multiplied by M, multiply the learning rate by M. For sequential mini-batch SGD with learning rate η we have:

$$\mathbf{w}_{t+M} = \mathbf{w}_t - \eta \sum_{j=0}^{M-1} g(\mathbf{w}_{t+j}, z_{t+j}), \tag{93}$$

where z_{t+j} is the $t + j$-th minibatch.

On the other hand, taking one step with M times large mini-batch size and learning rate $\hat{\eta} = M\eta$ in synchronous SGD yields:

$$\hat{\mathbf{w}}_{t+1} = \mathbf{w}_t - \hat{\eta} \frac{1}{M} \sum_{j=0}^{M-1} g(\mathbf{w}_t, z_t^j), \tag{94}$$

where z_t^j is the t-th minibatch on local machine j.

Assume that $z_{t+j} = z_t^j$. The assumption $g(\mathbf{w}_{t+j}, z_{t+j}) \approx g(\mathbf{w}_t, z_t^j)$ was made in synchronous SGD(Goyal et al., 2017). However, it often may not hold.

If we denote $\tilde{\mathbf{w}}_{t+1}^j = \mathbf{w}_t - \hat{\eta} \frac{1}{M} \sum_{i<j} g(\mathbf{w}_t, z_t^i)$, we can unfold the summation in Eq.94 to

$$\tilde{\mathbf{w}}_{t+1}^{j+1} = \tilde{\mathbf{w}}_{t+1}^j - \hat{\eta} \frac{1}{M} g(\mathbf{w}_t, z_t^j), j < M, \tag{95}$$

then we have $\hat{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_{t+1}^M$. We propose to use Eq.(5) in the main paper to compensate this assumption and apply delay-compensated gradient to update Eq.95 with:

$$g(\mathbf{w}_{t+j}, z_{t+j}) \approx \tilde{g}(\tilde{\mathbf{w}}_{t+1}^j, z_t^j) := g(\mathbf{w}_t, z_t^j) + \lambda g(\mathbf{w}_t, z_t^j) \odot g(\mathbf{w}_t, z_t^j) \odot (\tilde{\mathbf{w}}_{t+1}^j - \mathbf{w}_t)\big), \tag{96}$$

$$\tilde{\mathbf{w}}_{t+1}^{j+1} = \tilde{\mathbf{w}}_{t+1}^j - \hat{\eta} \frac{1}{M} \tilde{g}(\tilde{\mathbf{w}}_{t+1}^j, z_t^j), j < M. \tag{97}$$

Please note that we redefine the previous $\tilde{\mathbf{w}}_{t+1}^{j+1}$ in Eq.97. For $j > 1$, we need to design an order to make $\tilde{\mathbf{w}}_{t+1}^j \approx \mathbf{w}_{t+j}$. Choosing $\tilde{\mathbf{w}}_{t+1}^j$ according to the increasing order of $\|\tilde{\mathbf{w}}_{t+1}^j - \mathbf{w}_t\|^2$ can be used since the smaller distance with $\mathbf{w}_t$ will induce more accurate approximation by using Taylor expansion.

References

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Goyal, Priya, Dollar, Piotr, Girshick, Ross, Noordhuis, Pieter, Wesolowski, Lukasz, Kyrola, Aapo, Tulloch, Andrew, Jia, Yangqing, and He, Kaiming. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

LeCun, Yann. *Modèles connexionnistes de lapprentissage*. PhD thesis, These de Doctorat, Universite Paris 6, 1987.

Lee, Jason D, Simchowitz, Max, Jordan, Michael I, and Recht, Benjamin. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

Lian, Xiangru, Huang, Yijun, Li, Yuncheng, and Liu, Ji. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 2737–2745, 2015.