A Additional Analysis

Before presenting the main results, we introduce some necessary background on the delay compensated gradients.

A.1 Connection Between PC Steps

As discussed above, PC-ASGD relies upon the two steps to determine the updates for each agent at every time step, as displayed in Fig. 7. We first turn to the clipping step (line 7 of Algorithm 1) where all stale

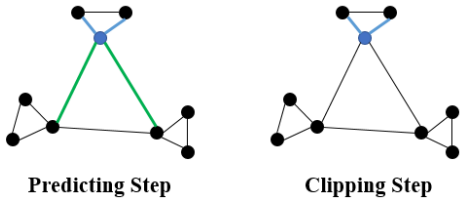

Figure 7: Predicting-Clipping Steps: in the predicting step, blue lines indicate no delay transmission; green lines represent delayed transmission that requires gradient prediction to reduce the stale effect; in the clipping step, the agent selectively drops the delayed information while only receiving information without delay.

information is dropped, which is equivalent to 'clipping' the original graph to become a smaller scale graph. Therefore, between the predicting step and the clipping step, we can observe two static graphs switching alternatively. This also suggests that element values of the mixing matrix $\tilde{W}$ in the clipping step are different from those in the predicting step. In the predicting step (line 6 of Algorithm 1), the agent still requires all the information from its neighbors while asking for gradient prediction from the unreliable neighbors. However, the update is determined by the combination of these two steps in Algorithm 1, which relies on the θ value to balance the tradeoff. For simplicity, we set the initialization of each agent 0.

We now turn to the practical variant of PC-ASGD in Algorithm 2 in the Appendix. The condition (line 9) adopted for PC-ASGD is based on the approximate cosine value of the angle between $g_i(x_t^i)$ and Δ_{pre} (or Δ_{clip}). When the angle between $g_i(x_t^i)$ and Δ_{pre} (or Δ_{clip}) is smaller, leading to a larger cosine value, the corresponding step should be chosen as it enables a larger descent amount along with the direction of $g_i(x_t^i)$. Hence, with a sequence of graphs and the properly set condition, these two alternating steps are connected to each other, allowing for convergence.

A.2 Delay compensated gradient

We detail how to arrive at Eq. 2. Specifically, given the outdated weights of agent k, $x_{t-\tau}^k$, due to the delay equal to τ, by induction, we can obtain for agent k

$$\begin{aligned} x_{t-\tau+1}^k =& x_{t-\tau}^k - \eta g_k(x_{t-\tau}^k) \\ =& x_{t-\tau}^k - \eta \sum_{r=0}^{0} [g_k(x_{t-\tau}^k) + \lambda g_k(x_{t-\tau}^k) \odot g_k(x_{t-\tau}^k) \odot (x_{t-\tau+r}^i - x_{t-\tau}^i)] \end{aligned} \tag{17}$$

$$\begin{aligned} x_{t-\tau+2}^k =& x_{t-\tau+1}^k - \eta g_k(x_{t-\tau+1}^k) = x_{t-\tau}^k - \eta g_k(x_{t-\tau}^k) - \eta g_k(x_{t-\tau+1}^k) \\ \approx& x_{t-\tau}^k - \eta \sum_{r=0}^{1} [g_k(x_{t-\tau}^k) + \lambda g_k(x_{t-\tau}^k) \odot g_k(x_{t-\tau}^k) \odot (x_{t-\tau+r}^i - x_{t-\tau}^i)] \end{aligned} \tag{18}$$

$$\cdots$$

$$x_t^k \approx x_{t-\tau}^k - \eta \sum_{r=0}^{\tau-1} [g_k(x_{t-\tau}^k) + \lambda g_k(x_{t-\tau}^k) \odot g_k(x_{t-\tau}^k) \odot (x_{t-\tau+r}^i - x_{t-\tau}^i)] \tag{19}$$

As we mentioned in the main contents, the term $(x^i_{t-\tau+r} - x^i_{t-\tau})$ is from agent i due to the outdated information of agent k, which intuitively illustrates that the compensation is driven by the agent i when agent k is in its neighborhood and deemed an unreliable one.

A.3 Compact Form of PC Steps

We next briefly discuss how to arrive at the compact form of the predicting and clipping steps for the analysis. For the convenience of analysis, we set the current time step as $t+\tau$ such that line 6 in Algorithm 1 shifts τ time steps ahead. Let us start with the predicting step and discuss its associated term $\sum_{j\in\mathcal{R}} w_{ij}x^j_{t+\tau} + \sum_{k\in\mathcal{R}^c} w_{ik}x^k_{t+\tau}$, where for the time being, it essentially holds that $x^k_{t+\tau} := x^k_t$. Note that $\mathcal{R}$ includes the agents i itself. Although unreliable neighbors are outdated, in the context, the update for agent i still requires such outdated information, which suggests that the whole graph applies. Additionally, the consensus is performed in parallel with the local computation, so this term boils down to a similar term in the existing consensus-based optimization algorithms in the literature. Thus, one can convert the current consensus term for weights to $\sum_p w_{ip}x^p_{t+\tau}, p \in V$. To show the evolution of predicting gradient over the past steps ranging from 0 to $\tau-1$, we use $g^{dc,r}_k(x^k_t)$ to represent.

Hence, the update law for the predicting step can be rewritten as:

$$x^i_{t+\tau+1} = \sum_p w_{ip}x^p_{t+\tau} - \eta(g_k(x^i_{t+\tau}) + \sum_{k\in\mathcal{R}^c} w_{ik}\sum_{r=0}^{\tau-1} g^{dc,r}_k(x^k_t)) \tag{20}$$

One may argue that for those outdated agent $k \in \mathcal{R}^c$, they have no information ahead of time t, which is τ time steps back from the current time. As the graph is undirected and connected, the time scale will not change the connections among agents. Also, for agent i, it receives always information from other agents, either the current or the outdated to update its weights. Thus, we have,

$$x^p_{t+\tau} = \begin{cases} x^j_{t+\tau} & p = j, j \in \mathcal{R} \\ x^k_t & p = k, k \in \mathcal{R}^c \end{cases} \tag{21}$$

Since the term $\sum_{k\in\mathcal{R}^c} w_{ik}\sum_{r=0}^{\tau-1} g^{dc,r}_k(x^k_t)$ applies to unreliable neighbors only, for the convenience of analysis, we expand it to the whole graph. It means that we establish an expanded graph to cover all of agents by setting some elements in the mixing matrix $\underline{W'} \in \mathbb{R}^{N\times N}$ equal to 0, but keeping the same connections as in $\underline{W}$. Then Eq. 20 can be modified as

$$x^i_{t+\tau+1} = \sum_p w_{ip}x^p_{t+\tau} - \eta(g_k(x^i_{t+\tau}) + \sum_q w'_{iq}\sum_{r=0}^{\tau-1} g^{dc,r}_k(x^q_t)) \tag{22}$$

where

$$w'_{iq} = \begin{cases} w_{ik} & if \;\; q = k, k \in \mathcal{R}^c \\ 0 & if \;\; q \in \mathcal{R} \end{cases} \tag{23}$$

Thus, we know via the above setting that $\underline{W'}$ is at least a row stochastic matrix. We rewrite the update law into a compact form such that

$$\mathbf{x}_{t+\tau+1} = W\mathbf{x}_{t+\tau} - \eta(\mathbf{g}(\mathbf{x}_{t+\tau}) + \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)). \tag{24}$$

where $W = \underline{W} \otimes I_{d\times d}$ and $W' = \underline{W'} \otimes I_{d\times d}$. Similarly, we rewrite the clipping steps in a vector form as follows:

$$\mathbf{x}_{t+\tau+1} = \tilde{W}\mathbf{x}_{t+\tau} - \eta\mathbf{g}(\mathbf{x}_{t+\tau}) \tag{25}$$

where $\tilde{W} = \underline{\tilde{W}} \otimes I_{d\times d}$. We are now ready to give the generalized step

$$\mathbf{x}_{t+\tau+1} = \mathcal{W}_{t+\tau}\mathbf{x}_{t+\tau} - \eta(\mathbf{g}(\mathbf{x}_{t+\tau}) + \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)), \tag{26}$$

where $\mathcal{W}_{t+\tau}$ is denoted as $\theta_{t+\tau}W + (1-\theta_{t+\tau})\tilde{W}$ throughout the rest of the analysis. Though the original graphs corresponding to the predicting and clipping steps are static, the equivalent graph $\mathcal{W}_{t+\tau}$ has become time-varying due to the time-varying θ value.

A.4 Approximate Hessian Matrix

Based on the update law, we know that the key part of PC-ASGD is the delay compensated gradients using Taylor expansion and Hessian approximation. Therefore, the Taylor expansion of the stochastic gradient $\mathbf{g}(\mathbf{x}_{t+\tau})$ at $\mathbf{x}_t$ can be written as follows:

$$\mathbf{g}(\mathbf{x}_{t+\tau}) = \mathbf{g}(\mathbf{x}_t) + \nabla\mathbf{g}(\mathbf{x}_t)(\mathbf{x}_{t+\tau} - \mathbf{x}_t) + O((\mathbf{x}_{t+\tau} - \mathbf{x}_t)^2)I, \tag{27}$$

where $\nabla\mathbf{g}$ denotes the matrix with the element $\nabla g_{ij} = \frac{\partial F}{\partial x^i \partial x^j}$ for all $i, j \in V$.

In most asynchronous SGD works, they used the zero-order item in Taylor expansion as its approximation to $\mathbf{g}(\mathbf{x}_{t+\tau})$ by ignoring the higher order term. Following from Zheng et al. (2017), we have

$$\mathbf{g}(\mathbf{x}_{t+\tau}) \approx \mathbf{g}(\mathbf{x}_t) + \nabla\mathbf{g}(\mathbf{x}_t)(\mathbf{x}_{t+\tau} - \mathbf{x}_t), \tag{28}$$

Directly adopting the above equation would be difficult in practice since $\nabla\mathbf{g}(\mathbf{x}_t)$ is generically computationally intractable when the model is very large, such as deep neural networks. To make the delay compensated gradients in PC-ASGD technically feasible, we apply approximation techniques for the Hessian matrix. We first use $O(\mathbf{x}_t)$ to denote the outer product matrix of the gradient at $\mathbf{x}_t$, i.e.,

$$O(\mathbf{x}_t) = (\frac{\partial}{\partial \mathbf{x}_t}F(\mathbf{x}_t))(\frac{\partial}{\partial \mathbf{x}_t}F(\mathbf{x}_t))^T \tag{29}$$

When the objective functions take the form of the cross-entropy loss or negative log-likelihood, the outer product of the gradient is an asymptotically unbiased estimation of the Hessian, according to the two equivalent methods to calculate the Fisher information matrix Friedman et al. (2001). That is,

$$\epsilon_t = \mathbb{E}[\|O(\mathbf{x}_t) - H(\mathbf{x}_t)\|] \to 0, \quad t \to 0 \tag{30}$$

where $H(\mathbf{x}_t)$ is the Hessian matrix of F at point $\mathbf{x}_t$.

The above equivalence relies on assumptions that the underlying distribution equals the model distribution with parameter $\mathbf{x}^*$ and that the training model $\mathbf{x}_t$ asymptotically converges to the (globally or locally) optimal model $\mathbf{x}^*$. According to the universal approximation theorem for DNN and some recent results on the optimality of the local optimal, such assumptions are technically reasonable. As the above equivalence was only developed by the negative log-likelihood form, that may not be applicable when we use PC-ASGD for the mean square error form, such as some time-series predictions with LSTM networks. Therefore, we introduce one assumption on the top of such an equivalence as follows,

$$\mathbb{E}[\|O(\mathbf{x}_t) - H(\mathbf{x}_t)\|] \leq \epsilon \quad \exists\epsilon > 0 \tag{31}$$

which primarily eliminates the computational complexity when directly calculating $H(\mathbf{x}_t)$. Another concern would be the large variance probably caused by $O(\mathbf{x}_t)$, though it is an unbiased estimation of $H(\mathbf{x}_t)$. Similar to Zheng et al. (2017), we introduce a new approximator $\lambda O(\mathbf{x}_t) \triangleq \lambda(\frac{\partial}{\partial \mathbf{x}_t}F(\mathbf{x}_t))(\frac{\partial}{\partial \mathbf{x}_t}F(\mathbf{x}_t))^T$. The authors in Zheng et al. (2017) have proved that $\lambda O(\mathbf{x}_t)$ is able to lead to smaller variance during training. Thus we refer interested readers to Zheng et al. (2017) for more details.

To reduce the storage of the approximator $\lambda O(\mathbf{x}_t)$, one widely-used diagonalization trick is adopted Becker & Lecun (1989). Hence, in the update law for PC-ASGD, we can see in the delay compensated gradient involving $\lambda g(\mathbf{x}_t) \odot \lambda g(\mathbf{x}_t)$. By denoting the diagonalized approximator as $Diag(\lambda O(\mathbf{x}_t))$, the following relationship is obtained:

$$Diag(\lambda O(\mathbf{x}_t)) = \lambda g(\mathbf{x}_t) \odot \lambda g(\mathbf{x}_t) \tag{32}$$

However, for analysis, when we apply diagonalization to $H(\mathbf{x}_t)$, it could cause diagonalization error such that we assume that the error is upper bounded by a constant $\epsilon_D > 0$, i.e.,

$$\|Diag(H(\mathbf{x}_t)) - H(\mathbf{x}_t)\| \leq \epsilon_D \tag{33}$$

B Additional Proof

For completeness, when presenting proof, we re-present statements for all lemmas and theorems.

Lemma 3: The iterates generated by PC-ASGD satisfy $\forall t \geq 0$, and $\tau \geq 2$:

$$\mathbf{x}_{t+\tau} = \prod_{v=0}^{t+\tau-1} \mathcal{W}_{t+\tau-1-v}\mathbf{x}_0 - \eta \sum_{s=0}^{t+\tau-1} \prod_{v=s+1}^{t+\tau-1} \mathcal{W}_{t+\tau+s-v}\mathbf{g}(\mathbf{x}_s) - \eta \sum_{s=t}^{t+\tau-1} \prod_{v=s+1}^{t+\tau-1} \theta_{s+1}\mathcal{W}_{t+\tau+s-v} \sum_{r=0}^{\tau-2} W'\mathbf{g}(\mathbf{x}_{s+1}). \tag{34}$$

Proof. Based on the vector form of the update law, we obtain

$$\mathbf{x}_{t+\tau} = \mathcal{W}_{t+\tau-1}\mathbf{x}_{t+\tau-1} - \eta(\mathbf{g}(\mathbf{x}_{t+\tau-1}) + \theta_{t+\tau-1} \sum_{r=0}^{\tau-2} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)) \tag{35}$$

With the above equation, it can be observed that $\mathbf{x}_{t+\tau}$ is a function with respect to $\mathbf{x}_t$, which contains all of agents. This suggests that by $\mathbf{x}_t$, there were no delay compensated gradients, while after $\mathbf{x}_{t+1}$, the unreliable neighbors need the delay compensated gradients due to delay. Hence, applying the above equation from 0 to $t + \tau - 1$ yields the desired result. □

Bounded (stochastic) gradient assumption: As $\mathbb{E}[\|\mathbf{g}(\mathbf{x})\|^2] \leq G^2$ and $\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla F(\mathbf{x})$, one can get that $\|\nabla F(\mathbf{x})\| = \|\mathbb{E}[\mathbf{g}(\mathbf{x})]\| \leq \mathbb{E}[\|\mathbf{g}(\mathbf{x})\|] = \sqrt{(\mathbb{E}[\|\mathbf{g}(\mathbf{x})\|])^2} \leq \sqrt{\mathbb{E}[\|\mathbf{g}(\mathbf{x})\|^2]} = G$.

Lemma 1: Let Assumptions 2 and 3 hold. Assume that the delay compensated gradients are uniformly bounded, i.e., there exists a scalar $B > 0$, such that

$$\|\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq B, \quad \forall t \geq 0 \ \textit{and} \ 0 \leq r \leq \tau - 1, \tag{36}$$

Then for all $i \in V$ and $t \geq 0$, $\exists \eta > 0$, we have

$$\mathbb{E}[\|x_t^i - y_t\|] \leq \eta \frac{G + (\tau - 1)B\theta_m}{1 - \delta_2}, \tag{37}$$

where $\theta_m = \max\{\theta_{s+1}\}_{s=t}^{t+\tau-1}$, $\delta_2 = \max\{\theta_s e_2 + (1 - \theta_s)\tilde{e}_2\}_{s=0}^{t+\tau-1} < 1$, where $e_2 := e_2(W) < 1$ and $\tilde{e}_2 := e_2(\tilde{W}) < 1$.

Proof. Since

$$\begin{aligned} \|x_{t+\tau}^i - y_{t+\tau}\| &\leq \|\mathbf{x}_{t+\tau} - y_{t+\tau}\mathbf{1}\| = \|\mathbf{x}_{t+\tau} - \frac{1}{N}\mathbf{1}^T\mathbf{x}_{t+\tau}\mathbf{1}\| \\ &= \|\mathbf{x}_{t+\tau} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\mathbf{x}_{t+\tau}\| = \|(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T)\mathbf{x}_{t+\tau}\|, \end{aligned} \tag{38}$$

where $\mathbf{1}$ is the column vector with entries all being 1. According to Assumption 2, we have $\frac{1}{N}\mathbf{1}\mathbf{1}^T\mathcal{W} = \frac{1}{N}\mathbf{1}\mathbf{1}^T$. Hence, by induction, setting $\mathbf{x}_0 = 0$, and Lemma 3, the following relationship can be obtained

$$\begin{aligned}
&\|\mathbf{x}_{t+\tau} - y_{t+\tau}\mathbf{1}\| \\
=&\eta\| \sum_{s=0}^{t+\tau-1} (\prod_{v=s+1}^{t+\tau-1} \mathcal{W}_{t+\tau+s-v} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)\mathbf{g}(\mathbf{x}_s) + \sum_{s=t}^{t+\tau-1} (\prod_{v=s+1}^{t+\tau-1} \mathcal{W}_{t+\tau+s-v} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)\theta_{s+1}\sum_{r=0}^{\tau-2} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \\
\leq&\eta \sum_{s=0}^{t+\tau-1} \| \prod_{v=s+1}^{t+\tau-1} \mathcal{W}_{t+\tau+s-v} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|\|\mathbf{g}(\mathbf{x}_s)\| + \eta \sum_{s=t}^{t+\tau-1} \| \prod_{v=s+1}^{t+\tau-1} \mathcal{W}_{t+\tau+s-v} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\|\|\theta_{s+1}\sum_{r=0}^{\tau-2} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \\
\leq&\eta G \sum_{s=0}^{t+\tau-1} \delta_2^{t+\tau-1-s} + \eta \sum_{s=t}^{t+\tau-1} \delta_2^{t+\tau-1-s}\theta_{s+1}(\tau-1)B \\
\leq&\eta G\frac{1}{1-\delta_2} + \eta(\tau-1)B\theta_m\frac{\delta_2^t - \delta_2^{t+\tau-1}}{1-\delta_2} \\
\leq&\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2}.
\end{aligned} \tag{39}$$

The second inequality follows from the Triangle inequality and Cauthy-Schwartz inequality and the third inequality follows from Assumption 2 and that the matrix $\frac{1}{N}\mathbf{1}\mathbf{1}^T$ is the projection of $\mathcal{W}$ onto the eigenspace associated with the eigenvalue equal to 1. The last inequality follows from the property of geometric sequence. The proof is completed by replacing $t+\tau$ with t on the left hand side. □

To prove the main results, we present several auxiliary lemmas first. We define

$$\begin{aligned}
\mathcal{G}^h(\mathbf{x}_t) &= \sum_{r=0}^{\tau-1} \mathbf{g}(\mathbf{x}_{t+r}) + H(\mathbf{x}_t)(\mathbf{v}_{t+r} - \mathbf{x}_t) \\
\nabla\mathcal{F}^h(\mathbf{x}_t) &= \sum_{r=0}^{\tau-1} \nabla F(\mathbf{x}_{t+r}) + \mathbb{E}[H(\mathbf{x}_t)(\mathbf{v}_{t+r} - \mathbf{x}_t)]
\end{aligned} \tag{40}$$

which are the incrementally delay compensated gradient and its expectation, respectively. It can be observed that $\mathcal{G}^h(\mathbf{x}_t)$ is the unbiased estimator of $\nabla\mathcal{F}^h(\mathbf{x}_t)$. It should be noted that $H(\mathbf{x}_t) = \nabla\mathbf{g}(\mathbf{x}_t)$. Let $\mathbf{v}_{t+\tau} = \mathcal{W}_{t+\tau}\mathbf{x}_{t+\tau}$. We next present a lemma to upper bound $\|\nabla F(\mathbf{v}_{t+r}) - \nabla\mathcal{F}^{h,r}(\mathbf{x}_t)\|$, where $\nabla\mathcal{F}^{h,r}(\mathbf{x}_t) = \nabla F(\mathbf{x}_{t+r}) + \mathbb{E}[H(\mathbf{x}_t)(\mathbf{v}_{t+r} - \mathbf{x}_t)]$.

Lemma 4: Let Assumptions 1,2 and 3 hold. Assume that $\nabla F(\mathbf{x}_t)$ is ξ_m-smooth. For $t \geq 0$, the iterates generated by PC-ASGD satisfy the following relationship, when $r \geq 1$

$$\|\nabla F(\mathbf{v}_{t+r}) - \nabla\mathcal{F}^{h,r}(\mathbf{x}_t)\| \leq \frac{\xi_m}{2}\eta^2(\frac{2G + (r-1)B\theta_m}{1-\delta_2})^2; \tag{41}$$

when $r = 0$, we have

$$\|\nabla F(\mathbf{v}_t) - \nabla F(\mathbf{x}_t)\| \leq 2\gamma_m\frac{\eta(G + (\tau-1)B\theta_m)}{1-\delta_2}. \tag{42}$$

Proof. By the smoothness condition for $\nabla F(\mathbf{x})$, we have

$$\|\nabla F(\mathbf{v}_{t+r}) - \nabla\mathcal{F}^{h,r}(\mathbf{x}_t)\| \leq \frac{\xi_m}{2}\|\mathbf{v}_{t+r} - \mathbf{x}_t\|^2 \leq \frac{\xi_m}{2}\|\mathbf{x}_{t+r} - \mathbf{x}_t\|^2 \tag{43}$$

Let $\mathbf{\Delta}_{t+r} = \mathbf{x}_{t+r} - \mathbf{x}_t$. Thus, based on Lemma 1, we have

$$\mathbf{x}_{t+r} = \prod_{v=t}^{t+r-1} \mathcal{W}_{t+r-1-v}\mathbf{x}_t - \eta \sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1} \mathcal{W}_{t+r+s-v}\mathbf{g}(\mathbf{x}_s) - \eta \sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1} \mathcal{W}_{t+s+r-v}\sum_{z=0}^{r-2}\theta_{s+1}W'\mathbf{g}^{dc,z}(\mathbf{x}_{s+1-r}) \tag{44}$$

Hence, we can obtain

$$\|\mathbf{\Delta}_{t+r}\|^2 = \|(\prod_{v=t}^{t+r-1}\mathcal{W}_{t+r-1-v}-I)\mathbf{x}_t-\eta\sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+r+s-v}\mathbf{g}(\mathbf{x}_s)-\eta\sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+s+r-v}\sum_{z=0}^{r-2}\theta_{s+1}W'\mathbf{g}^{dc,z}(\mathbf{x}_{s+1-r})\|^2 \tag{45}$$

Due to $\mathbf{x}_0 = 0$ and no delay compensated gradients before time step t, we can obtain

$$\begin{aligned}
&\|\mathbf{\Delta}_{t+r}\|^2\\
=&\|-\eta\sum_{s=0}^{t+r-1}\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+r+s-v}\mathbf{g}(\mathbf{x}_s)-\eta\sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+s+r-v}\sum_{z=0}^{r-2}\theta_{s+1}W'\mathbf{g}^{dc,z}(\mathbf{x}_{s+1-r})+\eta\sum_{s=0}^{t}\prod_{v=s}^{t}\mathcal{W}_{t+s-v}\mathbf{g}(\mathbf{x}_s)\|^2\\
\leq&\eta^2(\|\sum_{s=0}^{t+r-1}\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+r+s-v}\mathbf{g}(\mathbf{x}_s)\|+\|\sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+s+r-v}\sum_{z=0}^{r-2}\theta_{s+1}W'\mathbf{g}^{dc,z}(\mathbf{x}_{s+1-r})\|+\|\sum_{s=0}^{t}\prod_{v=s}^{t}\mathcal{W}_{t+s-v}\mathbf{g}(\mathbf{x}_s)\|)^2\\
\leq&\eta^2(\sum_{s=0}^{t+r-1}\|\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+r+s-v}\mathbf{g}(\mathbf{x}_s)\|+\sum_{s=t}^{t+r-1}\|\prod_{v=s+1}^{t+r-1}\mathcal{W}_{t+s+r-v}\sum_{z=0}^{r-2}\theta_{s+1}W'\mathbf{g}^{dc,z}(\mathbf{x}_{s+1-r})\|+\sum_{s=0}^{t}\|\prod_{v=s}^{t}\mathcal{W}_{t+s-v}\mathbf{g}(\mathbf{x}_s)\|)^2\\
\leq&\eta^2(\sum_{s=0}^{t+r-1}\prod_{v=s+1}^{t+r-1}\|\mathcal{W}_{t+r+s-v}\|\|\mathbf{g}(\mathbf{x}_s)\|+\sum_{s=t}^{t+r-1}\prod_{v=s+1}^{t+r-1}\|\mathcal{W}_{t+s+r-v}\|\|\sum_{z=0}^{r-2}\theta_{s+1}W'\mathbf{g}^{dc,z}(\mathbf{x}_{s+1-r})\|\\
&+\sum_{s=0}^{t}\prod_{v=s}^{t}\|\mathcal{W}_{t+s-v}\|\|\mathbf{g}(\mathbf{x}_s)\|)^2\\
\leq&\eta^2(\frac{2G}{1-\delta_2}+\frac{1}{1-\delta_2}B(r-1)\theta_m)^2\\
\leq&\eta^2(\frac{2G+\theta_m(r-1)B}{1-\delta_2})^2
\end{aligned} \tag{46}$$

The first inequality follows from the Triangle inequality. The second inequality follows from the Jensen inequality. The third inequality follows from the Cauthy-Schwartz inequality and the submultiplicative matrix norm applied to stochastic matrices. The fourth inequality follows from the Assumption 2 and bounded gradient. We have observed that this holds when $r \geq 1$. While $r = 0$ enables $\|\nabla F(\mathbf{v}_{t+r})-\mathcal{F}^{h,r}(\mathbf{x}_t)\|$ to degenerate to $\|\nabla F(\mathbf{v}_t) - \nabla F(\mathbf{x}_t)\|$ based on the definition of $\mathcal{F}^h(\mathbf{x}_t)$. Using the smoothness condition of $F(\mathbf{x})$, we can immediately obtain

$$\|\nabla F(\mathbf{v}_t) - \nabla F(\mathbf{x}_t)\| \leq 2\gamma_m\eta\frac{G+(\tau-1)B\theta_m}{1-\delta_2}. \tag{47}$$

The proof is completed. □

Lemma 5: Let Assumptions 1, 2 and 3 hold. Assume that the delay compensated gradients are uniformly bounded, i.e., there exists a scalar $B > 0$ such that

$$\|\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq B, \quad \forall t \geq 0 \ \textit{and} \ 0 \leq r \leq \tau-1, \tag{48}$$

Then for the iterates generated by PC-ASGD, $\exists\eta > 0$, they satisfy

$$\begin{aligned}
&\|\mathbb{E}[\mathcal{G}^h(\mathbf{x}_t)]-\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|\\
&\leq\sum_{r=1}^{\tau-1}(\gamma_m+\epsilon_D+\epsilon+(1-\lambda)G^2)\eta\frac{2G+(r-1)B\theta_m}{1-\delta_2}+\tau\sigma
\end{aligned} \tag{49}$$

Proof. Based on the definition of $\mathbb{E}\mathcal{G}^h(\mathbf{x}_t)$, we have

$$\begin{aligned}
&\|\mathbb{E}[\mathcal{G}^h(\mathbf{x}_t)] - \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| = \|\mathbb{E}[\sum_{r=0}^{\tau-1}\mathbf{g}(\mathbf{x}_{t+r}) + \sum_{r=0}^{\tau-1} H(\mathbf{x}_t)(\mathbf{x}_{t+r}-\mathbf{x}_t)] - \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \\
=&\|\mathbb{E}[\mathcal{G}^{h,r=0}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r=0}(\mathbf{x}_t) + \mathbb{E}[\mathcal{G}^{h,r=1}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r=1}(\mathbf{x}_t) + \cdots + \mathbb{E}[\mathcal{G}^{h,r=\tau-1}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r=\tau-1}(\mathbf{x}_t)\| \\
\leq&\|\mathbb{E}[\mathcal{G}^{h,r=0}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r=0}(\mathbf{x}_t)\| + \|\mathbb{E}[\mathcal{G}^{h,r=1}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r=1}(\mathbf{x}_t)\| + \cdots + \|\mathbb{E}[\mathcal{G}^{h,r=\tau-1}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r=\tau-1}(\mathbf{x}_t)\|
\end{aligned} \tag{50}$$

The last inequality follows from the Triangle inequality. Now let us discuss $\|\mathbb{E}\mathcal{G}^{h,r}(\mathbf{x}_t) - W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|$. The following analysis is for cases where $r \geq 1$. We give a brief analysis for the case in which $r = 0$ subsequently.

$$\begin{aligned}
&\|\mathbb{E}[\mathcal{G}^h(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \\
=&\|\mathbb{E}[\mathbf{g}(\mathbf{x}_{t+r}) + H(\mathbf{x}_t)(\mathbf{x}_{t+r}-\mathbf{x}_t)]W'[\mathbf{g}(\mathbf{x}_t) + \lambda\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)\odot(\mathbf{x}_{t+r}-\mathbf{x}_t)]\| \\
=&\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t) + [H(\mathbf{x}_t) - \lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)](\mathbf{x}_{t+r}-\mathbf{x}_t)\| \\
\leq&\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t)\| + \|[H(\mathbf{x}_t) - \lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)](\mathbf{x}_{t+r}-\mathbf{x}_t)\| \\
\leq&\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t)\| + \|[H(\mathbf{x}_t) - \lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) \\
&- Diag(H(\mathbf{x}_t)) + Diag(H(\mathbf{x}_t))](\mathbf{x}_{t+r}-\mathbf{x}_t)\| \\
\leq&\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t)\| + \|\mathbf{x}_{t+r}-\mathbf{x}_t\|\|(\lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)) + (\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) \\
&- Diag(H(\mathbf{x}_t))) + (Diag(H(\mathbf{x}_t)) - H(\mathbf{x}_t))\| \\
\leq&\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t)\| + \|\mathbf{x}_{t+r}-\mathbf{x}_t\|(\|\lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)\| + \|\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t) \\
&- Diag(H(\mathbf{x}_t))\| + \|Diag(H(\mathbf{x}_t)) - H(\mathbf{x}_t)\|)
\end{aligned}$$

The third inequality follows from Cauthy-Schwarz inequality while the last one follows from the Triangle inequality. It should be noted that when we combine $H(\mathbf{x}_t)(\mathbf{x}_{t+r}-\mathbf{x}_t)$ and $\lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)\odot(\mathbf{x}_{t+r}-\mathbf{x}_t)$, we follow the update law. Since in a rigorously mathematical sense, $\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)$ should be $\mathbf{g}(\mathbf{x}_t)\mathbf{g}(\mathbf{x}_t)^T$. However, for reducing the computational complexity when implementing the algorithm, as discussed above, we have made the approximation and diagonalization trick. Hence, we assume that $H(\mathbf{x}_t) - \lambda W'\mathbf{g}(\mathbf{x}_t)\odot\mathbf{g}(\mathbf{x}_t)$ can hold for simplicity and convenience.

Then we discuss $\mathbb{E}[\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t)\|]$.

$$\begin{aligned}
&\mathbb{E}[\|\nabla F(\mathbf{x}_{t+r}) - W'\mathbf{g}(\mathbf{x}_t)\|] \leq \mathbb{E}[\|\nabla F(\mathbf{x}_{t+r}) - \mathbf{g}(\mathbf{x}_t)\|] \\
=&\mathbb{E}[\|\nabla F(\mathbf{x}_{t+r}) - \nabla F(\mathbf{x}_t) + \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\|] \\
\leq&\mathbb{E}[\|\nabla F(\mathbf{x}_{t+r}) - \nabla F(\mathbf{x}_t)\|] + \mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\|] \\
\leq&\gamma_m\|\mathbf{x}_{t+r}-\mathbf{x}_t\| + \sqrt{(\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\|])^2} \\
\leq&\gamma_m\eta\frac{2G + (r-1)B\theta_m}{1-\delta_2} + \sqrt{\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\|]^2} \\
\leq&\gamma_m\eta\frac{2G + (r-1)B\theta_m}{1-\delta_2} + \sigma
\end{aligned} \tag{51}$$

Hence, we have

$$\begin{aligned}
\|\mathbb{E}[\mathcal{G}^h(\mathbf{x}_t)] - \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq&\gamma_m\eta\frac{2G + (r-1)B\theta_m}{1-\delta_2} + [(1-\lambda)G^2 + \epsilon_D + \epsilon]\eta\frac{2G + (r-1)B\theta_m}{1-\delta_2} + \sigma \\
=&(\gamma_m + \epsilon_D + \epsilon + (1-\lambda)G^2)\eta\frac{2G + (r-1)B\theta_m}{1-\delta_2} + \sigma
\end{aligned} \tag{52}$$

The above relationship is obtained for cases where $r \geq 1$. There still is $r = 0$ left. For $r = 0$,

$$\|\nabla F(\mathbf{x}_t) - W'\mathbf{g}(\mathbf{x}_t)\| \leq \sigma \tag{53}$$

Thus, combining each upper bound for $\|\mathbb{E}[\mathcal{G}^{h,r}(\mathbf{x}_t)] - W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|$, we can obtain

$$\|\mathbb{E}[\mathcal{G}^{h}(\mathbf{x}_t)] - \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq \sum_{r=1}^{\tau-1}(\gamma_m + \epsilon_D + \epsilon + (1-\lambda)G^2)\eta\frac{2G + (r-1)B\theta_m}{1-\delta_2} + \tau\sigma, \tag{54}$$

which completes the proof. □

Lemma 6: Let Assumptions 1, 2 and 3 hold. Assume that the delay compensated gradients are uniformly bounded, i.e., there exists a scalar $B > 0$ such that

$$\|\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq B, \quad \forall t \geq 0 \ \textit{and} \ 0 \leq r \leq \tau - 1, \tag{55}$$

Then for the iterates generated by PC-ASGD, $\exists \eta > 0$, they satisfy

$$F(\mathbf{x}_{t+\tau}) \geq F(\mathbf{v}_{t+\tau}) - 2G\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2} \tag{56}$$

Proof. Due to the convexity, we have

$$\begin{aligned}
F(\mathbf{x}_{t+\tau}) &\geq F(\mathbf{v}_{t+\tau}) + \nabla F(\mathbf{v}_{t+\tau})(\mathbf{x}_{t+\tau} - \mathbf{v}_{t+\tau}) \\
&\geq F(\mathbf{v}_{t+\tau}) - \|\nabla F(\mathbf{v}_{t+\tau})\|\|\mathbf{v}_{t+\tau} - \mathbf{x}_{t+\tau}\| \\
&\geq F(\mathbf{v}_{t+\tau}) - G\|\mathbf{v}_{t+\tau} - \mathbf{x}_{t+\tau}\| \\
&\geq F(\mathbf{v}_{t+\tau}) - G\|\mathbf{v}_{t+\tau} - y_{t+\tau}\mathbf{1} + y_{t+\tau}\mathbf{1} - \mathbf{x}_{t+\tau}\| \\
&\geq F(\mathbf{v}_{t+\tau}) - G(\|\mathbf{v}_{t+\tau} - y_{t+\tau}\mathbf{1}\| + \|y_{t+\tau}\mathbf{1} - \mathbf{x}_{t+\tau}\|) \\
&\geq F(\mathbf{v}_{t+\tau}) - 2G\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2}
\end{aligned} \tag{57}$$

The second inequality follows from the Cauthy-Schwarz inequality. The proof is completed. □

Theorem 1: Let Assumptions 1,2 and 3 hold. Assume that the delay compensated gradients are uniformly bounded, i.e., there exists a scalar $B > 0$ such that

$$\|\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq B, \quad \forall t \geq 0 \ \textit{and} \ 0 \leq r \leq \tau - 1, \tag{58}$$

and that $\nabla F(\mathbf{x}_t)$ is ξ_m-smooth for all $t \geq 0$. Then for the iterates generated by PC-ASGD, when $0 < \eta \leq \frac{1}{2\mu\tau}$ and the objective satisfies the PL condition, they satisfy

$$\mathbb{E}[F(\mathbf{x}_t) - F^*] \leq (1 - 2\mu\eta\tau)^{t-1}(F(\mathbf{x}_1) - F^* - \frac{Q}{2\mu\eta\tau}) + \frac{Q}{2\mu\eta\tau}, \tag{59}$$

$$\begin{aligned}
Q = {} & 2(1-2\mu\eta\tau)G\eta C_1 + \frac{\eta^3\xi_m G}{2}\sum_{r=1}^{\tau-1} C_r + 2\eta^2 G\gamma_m C_1 \\
& + G\eta\tau\sigma + \eta^2 G(\gamma_m + \epsilon_D + \epsilon + (1-\lambda)G^2)\sum_{r=1}^{\tau-1} C_r + \eta G^2 + \eta^2\gamma_m G\tau C_2
\end{aligned} \tag{60}$$

and,

$$\begin{aligned}
C_1 &= \frac{G + (\tau-1)B\theta_m}{1-\delta_2} \\
C_r &= \frac{2G + (r-1)B\theta_m}{1-\delta_2} \\
C_2 &= \frac{2G + (\tau-1)B\theta_m}{1-\delta_2},
\end{aligned} \tag{61}$$

$\epsilon_D > 0$ and $\epsilon > 0$ are upper bounds for the approximation errors of the Hessian matrix.

Proof. According to the smoothness condition of $F(\mathbf{x})$. We have

$$\mathbb{E}[F(\mathbf{x}_{t+\tau+1}) - F(\mathbf{x}^*)] \leq \mathbb{E}[F(\mathbf{v}_{t+\tau}) - F(\mathbf{x}^*)] + \mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), (\mathbf{x}_{t+\tau+1} - \mathbf{v}_{t+\tau})\rangle] + \frac{\gamma_m}{2}\mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{v}_{t+\tau}\|^2] \tag{62}$$

Based on the update law, we can obtain

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{x}_{t+\tau+1}) - F(\mathbf{x}^*)] \\
\leq &\mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*] - \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \mathbf{g}(\mathbf{x}_{t+\tau})\rangle] - \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle] \\
&+ \frac{\gamma_m\eta^2}{2}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau}) + \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] \\
\leq &\mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*] - \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \mathbf{g}(\mathbf{x}_{t+\tau})\rangle] - \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \tau\nabla F(\mathbf{v}_{t+\tau})\rangle] \\
&+ \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \tau\nabla F(\mathbf{v}_{t+\tau}) - \sum_{r=0}^{\tau-1} \nabla F(\mathbf{v}_{t+r})\rangle] + \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \sum_{r=0}^{\tau-1} \nabla F(\mathbf{v}_{t+r}) - \mathcal{F}^h(\mathbf{x}_t)\rangle] \\
&+ \eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \mathbb{E}[\mathcal{G}^h] - \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle] + \frac{\gamma_m\eta^2}{2}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau}) + \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2]
\end{aligned} \tag{63}$$

We next investigate each term on the right hand side. Based on Lemma 6, we can obtain

$$F(\mathbf{x}_{t+\tau}) \geq F(\mathbf{v}_{t+\tau}) - 2G\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2} \tag{64}$$

such that

$$F(\mathbf{x}_{t+\tau}) - F^* \geq F(\mathbf{v}_{t+\tau}) - F^* - 2G\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2} \tag{65}$$

For the term $-\eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), g(\mathbf{x}_{t+\tau})\rangle]$, we can quickly get that is is bounded above by ηG^2 due to the Cauthy-Schwarz inequality. Then for term $-\eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \tau\nabla F(\mathbf{v}_{t+\tau})\rangle]$, one can get the following relationship due to the PL condition.

$$-\eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \tau\nabla F(\mathbf{v}_{t+\tau})\rangle] \leq -2\eta\tau\mu(F(\mathbf{v}_{t+\tau}) - F^*) \tag{66}$$

Combining $F(\mathbf{v}_{t+\tau}) - F^*$, we have

$$\begin{aligned}
&(1 - 2\eta\tau\mu)(F(\mathbf{v}_{t+\tau}) - F^*) \\
&\leq (1 - 2\eta\tau\mu)[(F(\mathbf{x}_{t+\tau}) - F^*) + 2G\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2}]
\end{aligned} \tag{67}$$

Based on Lemma 4, we have known that

$$\|\nabla F(\mathbf{v}_{t+r}) - \nabla\mathcal{F}^{h,r}(\mathbf{x}_t)\| \leq \frac{\xi_m}{2}\eta^2[\frac{2G + (r-1)B\theta_m}{1-\delta_2}]^2; \tag{68}$$

for $r \geq 1$, while for $r = 0$, it can be obtained that

$$\|\nabla F(\mathbf{v}_t) - \nabla F(\mathbf{x}_t)\| \leq 2\gamma_m\eta\frac{G + (\tau-1)B\theta_m}{1-\delta_2}. \tag{69}$$

Since

$$\begin{aligned}
\eta\mathbb{E}[\langle \nabla F(\mathbf{v}_{t+\tau}), \sum_{r=0}^{\tau-1} \nabla F(\mathbf{v}_{t+r}) - \mathcal{F}^h(\mathbf{x}_t)\rangle] &\leq \eta\mathbb{E}[\|\nabla F(\mathbf{v}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} \nabla F(\mathbf{v}_{t+r}) - \mathcal{F}^h(\mathbf{x}_t)\|] \\
&\leq \mathbb{E}[\|\nabla F(\mathbf{v}_{t+\tau})\|\sum_{r=0}^{\tau-1} \|\nabla F(\mathbf{v}_{t+r}) - \mathcal{F}^h(\mathbf{x}_t)\|]
\end{aligned} \tag{70}$$

The first inequality follows from Cauthy-Schwarz inequality and the second one follows from Triangle inequality. Hence, we can have

$$\eta\mathbb{E}[\langle\nabla F(\mathbf{v}_{t+\tau}),\sum_{r=0}^{\tau-1}\nabla F(\mathbf{v}_{t+r})-\mathcal{F}^h(\mathbf{x}_t)\rangle]\leq\frac{\eta^3\xi_m G}{2(1-\delta_2)}\sum_{r=1}^{\tau-1}[2G+B(r-1)\theta_m]+2\eta^2G\gamma_m\frac{G+(\tau-1)B\theta_m}{1-\delta_2} \tag{71}$$

According to Lemma 4, the following relationship can be obtained,

$$\mathbb{E}[\langle\nabla F(\mathbf{v}_{t+\tau}),\mathbb{E}[\mathcal{G}^h(\mathbf{x}_t)]-\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle]\leq\frac{\eta^2G}{1-\delta_2}(\gamma_m+\epsilon_D+\epsilon+(1-\lambda)G^2)\sum_{r=1}^{\tau-1}[2G+(r-1)B\theta_m]+G\eta\tau\sigma \tag{72}$$

The last term is $\eta\mathbb{E}[\langle\nabla F(\mathbf{v}_{t+\tau}),\tau\nabla F(\mathbf{v}_{t+\tau})-\sum_{r=0}^{\tau-1}\nabla F(\mathbf{v}_{t+r})\rangle]$, which can be rewritten such that

$$\begin{aligned}&\eta\mathbb{E}[\langle\nabla F(\mathbf{v}_{t+\tau}),\tau\nabla F(\mathbf{v}_{t+\tau})-\sum_{r=0}^{\tau-1}\nabla F(\mathbf{v}_{t+r})\rangle]\\ \leq&\eta\mathbb{E}[\|\nabla F(\mathbf{v}_{t+\tau})\|\|\nabla F(\mathbf{v}_{t+\tau})-\nabla F(\mathbf{v}_t)+\cdots+\nabla F(\mathbf{v}_{t+\tau})-\nabla F(\mathbf{v}_{t+\tau-1})\|]\\ \leq&\eta\mathbb{E}[\|\nabla F(\mathbf{v}_{t+\tau})\|\|\nabla F(\mathbf{v}_{t+\tau})-\nabla F(\mathbf{v}_t)\|+\cdots+\|\nabla F(\mathbf{v}_{t+\tau})-\nabla F(\mathbf{x}_{t+\tau-1})\|]\end{aligned} \tag{73}$$

Using the smoothness condition, we then can bound the term by deriving the following relationship with Lemma 1 and Lemma 3,

$$\eta\mathbb{E}[\langle\nabla F(\mathbf{v}_{t+\tau}),\tau\nabla F(\mathbf{v}_{t+\tau})-\sum_{r=0}^{\tau-1}\nabla F(\mathbf{v}_{t+r})\rangle]\leq\eta^2\gamma_mG\tau\frac{2G+(\tau-1)B\theta_m}{1-\delta_2} \tag{74}$$

We combine the upper bounds of each term on the right hand side to produce the following relationship.

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}_{t+\tau+1})-F(\mathbf{x}^*)]\leq&(1-2\eta\mu\tau)(F(\mathbf{x}_{t+\tau})-F^*)+2(1-2\eta\mu\tau)G\eta\frac{G+(\tau-1)B\theta_m}{1-\delta_2}\\ &+\frac{\eta^3\xi_mG}{2(1-\delta_2)}\sum_{r=1}^{\tau-1}[2G+(r-1)B\theta_m]+2\eta^2G\gamma_m\frac{G+(\tau-1)B\theta_m}{1-\delta_2}+G\eta\tau\sigma+\eta G^2\\ &+\frac{\eta^2G}{1-\delta_2}(\gamma_m+\epsilon_D+\epsilon+(1-\lambda)G^2)\sum_{r=1}^{\tau-1}[2G+(r-1)B\theta_m]+\eta^2\gamma_mG\tau\frac{2G+(\tau-1)B\theta_m}{1-\delta_2}.\end{aligned} \tag{75}$$

We now know that

$$\mathbb{E}[F(\mathbf{x}_{t+1})-F^*]\leq(1-2\eta\tau\mu)\mathbb{E}[F(\mathbf{x}_t)-F^*]+Q, \tag{76}$$

subtracting the constant $\frac{Q}{2\mu\tau\eta}$ from both sides, one obtains

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}_{t+1})-F^*]-\frac{Q}{2\mu\tau\eta}&\leq(1-2\eta\mu\tau)\mathbb{E}[F(\mathbf{x}_t)-F^*]+Q-\frac{Q}{2\mu\tau\eta}\\ &=(1-2\eta\mu\tau)(\mathbb{E}[F(\mathbf{x}_t)-F^*]-\frac{Q}{2\mu\tau\eta})\end{aligned} \tag{77}$$

Observe that the above inequality is a contraction inequality since $0<2\eta\mu\tau\leq1$ due to $0<\eta\leq\frac{1}{2\mu\tau}$. The result thus follows by applying the inequality repeatedly through iteration $t\in\mathbb{N}$. □

Another scenario that could be of interest is the strongly convex objective. As Theorem 1 has shown with a properly set constant step size, PC-ASGD is able to converge to the neighborhood of the optimal solution

with a linear rate. This also applies to the strongly convex objective in which the strong convexity implies the PL condition, while the constants are subject to changes. We now proceed to give the proof for the generally convex case.

Theorem 2: Let Assumptions 1, 2 and 3 hold. Assume that the delay compensated gradients are uniformly bounded, i.e., there exists a scalar $B > 0$ such that for all $T \geq 1$

$$\|\mathbf{g}^{dc,r}(\mathbf{x}_t)\| \leq B, \quad \forall t \geq 0 \ \textit{and} \ 0 \leq r \leq \tau - 1, \tag{78}$$

and there exists $C > 0$,

$$\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|] \leq C, \tag{79}$$

where $\mathbf{x}^* \in \text{argmin} F(\mathbf{x})$. Then for the iterations generated by PC-ASGD, there exists $0 < \eta < \frac{1}{20\gamma_m}$, such that

$$\mathbb{E}[F(\bar{\mathbf{x}}_T) - F^*] \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{T\eta} + \frac{A}{\eta}, \tag{80}$$

where $A = 10\eta^2\sigma_*^2 + 10\eta^2\sigma^2 + 20\eta^4 G^2 C_1^2 + 5\eta^2\theta_m^2\tau^2 B^2 + 2\eta C\theta_m \tau B + 2G\eta^2 C_1(2C+1), C_1 = \frac{G+(\tau-1)B\theta_m}{1-\delta_2}, \sigma_*^2 := \mathbb{E}\|\mathbf{g}(\mathbf{x}^*) - \nabla F(\mathbf{x}^*)\|^2, \bar{\mathbf{x}}_T := \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$.

Proof. According to the compact update law, we have

$$\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2 = \|\mathcal{W}_{t+\tau}\mathbf{x}_{t+\tau} - \eta(\mathbf{g}(\mathbf{x}_{t+\tau}) + \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)) - \mathbf{x}^*\|^2. \tag{81}$$

As $\mathbf{v}_{t+\tau} = \mathcal{W}_{t+\tau}\mathbf{x}_{t+\tau}$, we can obtain

$$\begin{aligned}\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2 = \|\mathbf{v}_{t+\tau} - \mathbf{x}^*\|^2 - 2\eta\langle \mathbf{v}_{t+\tau} - \mathbf{x}^*, \mathbf{g}(\mathbf{x}_{t+\tau}) + \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle \\ + \eta^2\|\mathbf{g}(\mathbf{x}_{t+\tau}) + \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2.\end{aligned} \tag{82}$$

For convenience, we define that $\Gamma_{t+\tau} = \mathbf{g}(\mathbf{x}_{t+\tau}) + \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)$. Hence, the above equation can be rewritten as

$$\begin{aligned}\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2 = \|\mathbf{v}_{t+\tau} - \mathbf{x}^*\|^2 + \eta^2\|\Gamma_{t+\tau}\|^2 \\ + 2\eta\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \mathbf{g}(\mathbf{v}_{t+\tau})\rangle + 2\eta\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \mathbf{g}(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{v}_{t+\tau})\rangle \\ + 2\eta\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle.\end{aligned} \tag{83}$$

Taking expectation on both sides leads to the following relationship:

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] + \eta^2\mathbb{E}[\|\Gamma_{t+\tau}\|^2] \\ + 2\eta\mathbb{E}[\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \nabla F(\mathbf{v}_{t+\tau})\rangle] + 2\eta\mathbb{E}[\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\rangle] \\ + 2\eta\mathbb{E}[\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \theta_{t+\tau}\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle].\end{aligned} \tag{84}$$

The inequality holds due to the basic property for the projection Sundhar Ram et al. (2010). For the last two terms on the right hand side of the above inequality, we can leverage Cauchy-Schwartz inequality to obtain the upper bounds. For $2\eta\mathbb{E}[\langle \mathbf{x}^* - \mathbf{v}_{t+\tau}, \nabla F(\mathbf{v}_{t+\tau})\rangle]$, we will use Lemma 2 to reformulate. We next investigate $\eta^2\mathbb{E}[\|\Gamma_{t+\tau}\|^2]$. Before that, we introduce a theoretical fact for the generally convex smooth functions.

Variance transfer: gradient noise (Lemma 4.20) in Garrigos & Gower (2023). If F is smooth and convex, then for all $\mathbf{x}$ we have that

$$\mathbb{E}[\|\mathbf{g}(\mathbf{x})\|^2] \leq 4\gamma_m(F(\mathbf{x}) - F^*) + 2\sigma_*^2, \tag{85}$$

where $\mathbf{g}(\mathbf{x})$ is the stochastic gradient, σ_*^2 is the variance of stochastic gradient at $\mathbf{x}^*$. Rewrite $\|\Gamma_{t+\tau}\|^2 = \|\mathbf{g}(\mathbf{x}_{t+\tau}) + \nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{x}_{t+\tau}) + \mathbf{g}(\mathbf{v}_{t+\tau}) - \mathbf{g}(\mathbf{v}_{t+\tau}) + \nabla F(\mathbf{v}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau}) + \theta_{t+\tau} \sum_{r=0}^{\tau-1} W' \mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2$. We then have the following relationship:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2] \leq\ & \mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] + 5\eta^2 \mathbb{E}[\|\mathbf{g}(\mathbf{v}_{t+\tau})\|^2] + 5\eta^2 \mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{x}_{t+\tau})\|^2] \\
& + 5\eta^2 \mathbb{E}[\|\mathbf{g}(\mathbf{v}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|^2] + 5\eta^2 \mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|^2] \\
& + 5\eta^2 \mathbb{E}[\|\theta_{t+\tau} \sum_{r=0}^{\tau-1} W' \mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] + 2\eta \mathbb{E}[F^* - F(\mathbf{v}_{t+\tau})] \\
& + 2\eta \mathbb{E}[\|\mathbf{x}^* - \mathbf{v}_{t+\tau}\| \|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|] + 2\eta \mathbb{E}[\|\mathbf{x}^* - \mathbf{v}_{t+\tau}\| \|\theta_{t+\tau} \sum_{r=0}^{\tau-1} W' \mathbf{g}^{dc,r}(\mathbf{x}_t)\|].
\end{aligned} \tag{86}$$

The last inequality holds due to the basic inequality $\|\sum_{i=1}^N \mathbf{a}_i\|^2 \leq N \sum_{i=1}^N \|\mathbf{a}_i\|^2$, the convexity property, and Cauchy-Schwartz inequality. By substituting Eq. 85 into Eq. 86, the following relationship can be obtained

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2] \leq\ & \mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] + 20\eta^2 \gamma_m \mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*] + 10\eta^2 \sigma_*^2 \\
& + 5\eta^2 \mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{x}_{t+\tau})\|^2] \\
& + 5\eta^2 \mathbb{E}[\|\mathbf{g}(\mathbf{v}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|^2] + 5\eta^2 \mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|^2] \\
& + 5\eta^2 \mathbb{E}[\|\theta_{t+\tau} \sum_{r=0}^{\tau-1} W' \mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] + 2\eta \mathbb{E}[F^* - F(\mathbf{v}_{t+\tau})] \\
& + 2\eta \mathbb{E}[\|\mathbf{x}^* - \mathbf{v}_{t+\tau}\| \|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|] + 2\eta \mathbb{E}[\|\mathbf{x}^* - \mathbf{v}_{t+\tau}\| \|\theta_{t+\tau} \sum_{r=0}^{\tau-1} W' \mathbf{g}^{dc,r}(\mathbf{x}_t)\|] \\
\leq\ & \mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] + 2\eta(10\gamma_m \eta - 1) \mathbb{E}[F(\mathbf{x}_{t+\tau}) - F^*] + 10\eta^2 \sigma_*^2 \\
& + 10\eta^2 \sigma^2 + 20\eta^4 G^2 C_1^2 + 5\eta^2 \theta_m^2 \tau^2 B^2 + 2\eta C(2G\eta C_1 + \theta_m \tau B).
\end{aligned} \tag{87}$$

The second inequality follows from Assumption 3, Eq. 47 and bounds for the predicted gradients. With mathematical manipulation, the above inequality can be written as

$$\begin{aligned}
2\eta(1 - 10\gamma_m \eta) \mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*] \leq\ & \mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] - \mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2] \\
& + 10\eta^2 \sigma_*^2 + 10\eta^2 \sigma^2 + 20\eta^4 G^2 C_1^2 + 5\eta^2 \theta_m^2 \tau^2 B^2 + 2\eta C(2G\eta C_1 + \theta_m \tau B)
\end{aligned} \tag{88}$$

Due to $\eta \leq \frac{1}{20\gamma_m}$, $1 - 10\gamma_m \eta \geq \frac{1}{2}$ such that $\eta \mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*] \leq 2\eta(1 - 10\gamma_m \eta) \mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*]$. Dividing both sides of Eq. 88 by η yields the following

$$\begin{aligned}
\mathbb{E}[F(\mathbf{v}_{t+\tau}) - F^*] \leq\ & \frac{1}{\eta}(\mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] - \mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2]) \\
& + \frac{1}{\eta}(10\eta^2 \sigma_*^2 + 10\eta^2 \sigma^2 + 20\eta^4 G^2 C_1^2 + 5\eta^2 \theta_m^2 \tau^2 B^2 + 2\eta C(2G\eta C_1 + \theta_m \tau B)).
\end{aligned} \tag{89}$$

Similar to Lemma 6, we can obtain that $F(\mathbf{v}_{t+\tau}) \geq F(\mathbf{x}_{t+\tau}) - 2G\eta \frac{G + (\tau-1)B\theta_m}{1-\delta_2}$. Then it is immediately obtained that $F(\mathbf{v}_{t+\tau}) - F^* \geq F(\mathbf{x}_{t+\tau}) - F^* - 2G\eta \frac{G + (\tau-1)B\theta_m}{1-\delta_2}$. With this, the following relationship can be obtained

$$\mathbb{E}[F(\mathbf{x}_{t+\tau}) - F^*] \leq \frac{1}{\eta}(\mathbb{E}[\|\mathbf{x}_{t+\tau} - \mathbf{x}^*\|^2] - \mathbb{E}[\|\mathbf{x}_{t+\tau+1} - \mathbf{x}^*\|^2]) + \frac{A}{\eta}, \tag{90}$$

where $A = 10\eta^2 \sigma_*^2 + 10\eta^2 \sigma^2 + 20\eta^4 G^2 C_1^2 + 5\eta^2 \theta_m^2 \tau^2 B^2 + 2\eta C \theta_m \tau B + 2G\eta^2 C_1(2C + 1)$. Recursively summing over t from 1 to T and replacing $t + \tau$ with t grants us the following relationship

$$\sum_{t=1}^{T} \mathbb{E}[F(\mathbf{x}_t) - F^*] \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{\eta} + \frac{AT}{\eta}. \tag{91}$$

Dividing both sides by T in the last relationship attains the following

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\mathbf{x}_t)-F^*]\leq\frac{\|\mathbf{x}_1-\mathbf{x}^*\|^2}{T\eta}+\frac{A}{\eta}. \tag{92}$$

Using that F is convex with Jensen inequality gives the desirable result. □

In the sequel, we provide the details for the smooth nonconvex functions.

Theorem 3: Let Assumptions 1,2 and 3 hold. Assume that the delay compensated gradients are uniformly bounded, i.e., there exists a scalar $B>0$ such that

$$\|\mathbf{g}^{dc,r}(\mathbf{x}_t)\|\leq B,\quad \forall t\geq 0\ \ and\ \ 0\leq r\leq\tau-1, \tag{93}$$

and that

$$\mathbb{E}[\|\mathbf{g}^{dc}(\mathbf{x}_t)\|^2]\leq M. \tag{94}$$

Then for the iterates generated by PC-ASGD, there exists $0<\eta<\frac{1}{\gamma_m}$, such that for all $T\geq 1$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2]\leq\frac{2(F(\mathbf{x}_1)-F^*)}{T\eta}+\frac{R}{\eta}, \tag{95}$$

where

$$R=2G\eta^2C_1+\frac{\tau\eta^2\gamma_m M}{2}+\frac{\eta\sigma^2}{2}+\eta\sigma\tau B+2\eta^2\gamma_m(\tau B+G)C_1.$$

Proof. According to the smoothness condition of $F(\mathbf{x})$, we have

$$\begin{aligned}
&F(\mathbf{x}_{t+\tau+1})-F(\mathbf{v}_{t+\tau})\\
\leq&\langle\nabla F(\mathbf{v}_{t+\tau}),\mathbf{x}_{t+\tau+1}-\mathbf{v}_{t+\tau}\rangle+\frac{\gamma_m}{2}+\|\mathbf{x}_{t+\tau+1}-\mathbf{v}_{t+\tau}\|^2\\
=&\langle\nabla F(\mathbf{v}_{t+\tau}),-\eta(\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau}))\rangle+\frac{\eta^2\gamma_m}{2}\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}+\mathbf{g}(\mathbf{x}_{t+\tau})\|^2\\
=&\langle\nabla F(\mathbf{v}_{t+\tau})-\nabla F(\mathbf{x}_{t+\tau})+\nabla F(\mathbf{x}_{t+\tau}),\eta(\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau}))\rangle+\frac{\eta^2\gamma_m}{2}\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}+\mathbf{g}(\mathbf{x}_{t+\tau})\|^2\\
=&-\eta\langle\nabla F(\mathbf{x}_{t+\tau}),\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau})\rangle+\eta\langle(\nabla F(\mathbf{v}_{t+\tau})-\nabla F(\mathbf{x}_{t+\tau}),\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau}))\rangle\\
&+\frac{\eta^2\gamma_m}{2}\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}+\mathbf{g}(\mathbf{x}_{t+\tau})\|^2\\
=&-\frac{\eta}{2}[\|\nabla F(\mathbf{x}_{t+\tau})\|^2+\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau})\|^2-\|\nabla F(\mathbf{x}_{t+\tau})-(\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau}))\|^2]\\
&+\eta\langle\nabla F(\mathbf{x}_{t+\tau})-\nabla F(\mathbf{v}_{t+\tau}),\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau})\rangle+\frac{\eta^2\gamma_m}{2}\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}+\mathbf{g}(\mathbf{x}_{t+\tau})\|^2\\
=&-\frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau})\|^2-\frac{\eta}{2}\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau})\|^2+\frac{\eta}{2}(\|\nabla F(\mathbf{x}_{t+\tau})-\mathbf{g}(\mathbf{x}_{t+\tau})\|^2+\|\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2\\
&-2\langle\nabla F(\mathbf{x}_{t+\tau})-\mathbf{g}(\mathbf{x}_{t+\tau}),\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle)+\eta\langle\nabla F(\mathbf{x}_{t+\tau})-\nabla F(\mathbf{v}_{t+\tau}),\sum_{r=0}^{\tau-1}W'\mathbf{g}^{dc,r}(\mathbf{x}_t)+\mathbf{g}(\mathbf{x}_{t+\tau})\rangle
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\eta^2\gamma_m}{2}\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r} + \mathbf{g}(\mathbf{x}_{t+\tau})\|^2 \\
=&-\frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau})\|^2 - (\frac{\eta}{2} - \frac{\eta^2\gamma_m}{2})\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_{t+\tau})\|^2 + \frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|^2 + \frac{\eta}{2}\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2 \\
&- \eta\langle\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau}), \sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle + \eta\langle\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau}), \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_{t+\tau})\rangle \\
=&-\frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau})\|^2 + (\frac{\eta^2\gamma_m}{2} - \frac{\eta}{2})\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2 + (\frac{\eta^2\gamma_m}{2} - \frac{\eta}{2})\|\mathbf{g}(\mathbf{x}_{t+\tau})\|^2 \\
&+ (\frac{\eta^2\gamma_m}{2} - \frac{\eta}{2})\langle\mathbf{g}(\mathbf{x}_{t+\tau}), \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle + \frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|^2 + \frac{\eta}{2}\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2 \\
&- \eta\langle\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau}), \sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\rangle + \eta\langle\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau}), \sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_{t+\tau})\rangle \\
\leq&-\frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau})\|^2 + (\frac{\eta^2\gamma_m}{2} - \frac{\eta}{2})\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2 + (\frac{\eta^2\gamma_m}{2} - \frac{\eta}{2})\|\mathbf{g}(\mathbf{x}_{t+\tau})\|^2 \\
&+ (\frac{\eta^2\gamma_m}{2} - \frac{\eta}{2})\|\mathbf{g}(\mathbf{x}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| + \frac{\eta}{2}\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|^2 + \frac{\eta}{2}\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2 \\
&+ \eta\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\| + \eta\|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_{t+\tau})\|.
\end{aligned}
$$

The first inequality follows from the smooth property of the objective. The last inequality follows from Cauthy-Schwarz inequality. The left hand side of the above inequality can be rewritten:

$$F(\mathbf{x}_{t+\tau+1}) - F(\mathbf{x}_{t+\tau}) + F(\mathbf{x}_{t+\tau}) - F(\mathbf{v}_{t+\tau})$$

Taking expectations for both sides, with the last inequality, we have

$$
\begin{aligned}
&\mathbb{E}[F(\mathbf{x}_{t+\tau+1}) - F(\mathbf{x}_{t+\tau})] \\
\leq&\mathbb{E}[F(\mathbf{v}_{t+\tau}) - F(\mathbf{x}_{t+\tau})] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau})\|^2] + \frac{\eta^2\gamma_m - \eta}{2}\mathbb{E}[\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] + \frac{\eta^2\gamma_m - \eta}{2}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau})\|^2] \\
&+ \frac{\eta^2\gamma_m - \eta}{2}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|] + \frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|^2] + \frac{\eta}{2}\mathbb{E}[\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] \\
&+ \eta\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|] + \eta\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_{t+\tau})\|] \\
\leq&G\mathbb{E}[\|\mathbf{v}_{t+\tau} - \mathbf{x}_{t+\tau}\|] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau})\|^2] + \frac{\eta^2\gamma_m - \eta}{2}\tau\sum_{r=0}^{\tau-1}\mathbb{E}[\|W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] + \frac{\eta^2\gamma_m - \eta}{2}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau})\|^2] \\
&+ \frac{\eta^2\gamma_m - \eta}{2}\mathbb{E}[\|\mathbf{g}(\mathbf{x}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|] + \frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|^2] + \frac{\eta}{2}\mathbb{E}[\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|^2] \\
&+ \eta\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \mathbf{g}(\mathbf{x}_{t+\tau})\|\|\sum_{r=1}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t)\|] + \eta\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau}) - \nabla F(\mathbf{v}_{t+\tau})\|\|\sum_{r=0}^{\tau-1} W'\mathbf{g}^{dc,r}(\mathbf{x}_t) + \mathbf{g}(\mathbf{x}_{t+\tau})\|] \\
\leq&-\frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{x}_{t+\tau})\|^2] + \frac{\tau^2\eta^2\gamma_m M}{2} + \frac{\eta\sigma^2}{2} + \eta\sigma\tau B + 2\eta^2\gamma_m(\tau B + G + \frac{G}{\eta\gamma_m})\frac{G + (\tau-1)B\theta_m}{1-\delta_2}
\end{aligned}
\tag{96}
$$

The last inequality follows from the smoothness condition of $F(\mathbf{x})$ and the bounded gradient, respectively, as well as $\eta < \frac{1}{\gamma_m}$. Hence, by replacing $t+\tau$ with t, one can obtain

$$\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] + R \tag{97}$$

where R indicates the constant term on the right hand side of the inequality. As we assume that $F(\mathbf{x})$ is bounded from below, applying the last inequality from 1 to T, one can get

$$F^* - F(\mathbf{x}_1) \leq \mathbb{E}[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_1) \leq -\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] + TR \tag{98}$$

which results in

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] \leq \frac{2[(F(\mathbf{x}_1) - F^*) + TR]}{\eta} \tag{99}$$

Dividing both sides by T, the desirable results are obtained. □

C Detailed Settings of Deep Learning Models

Model Settings For the PreResNet110 (*model 1*), DenseNet (*model 2*), ResNet20 (*model 3*) and EfficientNet (*model 4*), models' architectures are shown in He et al. (2016b), Huang et al. (2017), He et al. (2016a) and Tan & Le (2019) respectively. The batch size is selected as 128. After hyperparameter searching in $(0.1, 0.01, 0.001)$, the learning rate is set as 0.01 for the first 160 epochs and changed to 0.001. The decays are applied in epochs $(80, 120, 160, 200)$. The approximation coefficient λ is set as 1. $\lambda = 0.001$ is first tried as suggested by DC-ASGD Zheng et al. (2017) and the results show that the predicting step doesn't affect the training process. By considering the upper bound of 1, a set of values $(0.001, 0.1, 1)$ are tried, and $\lambda = 1$ is applied according to the performance.

Hardware environment. Our experiments are implemented and evaluated at GTX-1080 ti with Intel Xenon 2.55GHz processor with 32GB RAM.

Table 5: Performance comparison in TinyImageNet and Time Series dataset

Model & dataset	Pre110 TinyImageNet	DesNet TinyImageNet	EfficientNet TinyImageNet	LSTM Wind Turbine Data
PC-ASGD (Ours)	$\mathbf{58.0 \pm 1.4}$	$\mathbf{61.4 \pm 0.7}$	$\mathbf{74.8 \pm 0.9}$	$\mathbf{71.2 \pm 0.5}$
D-ASGD Lian et al. (2017)	52.1 ± 0.3	57.5 ± 0.2	70.4 ± 0.5	66.2 ± 0.1
DC-s3gd Rigazzi (2019)	55.1 ± 0.8	58.5 ± 1.4	73.2 ± 1.2	61.4 ± 1.1
D-ASGD with IS Du et al. (2020)	53.2 ± 0.9	58.1 ± 1.2	73.4 ± 0.7	69.2 ± 0.2
Adaptive Braking Venigalla et al. (2020)	55.2 ± 1.2	60.2 ± 1.1	67.3 ± 1.5	66.5 ± 1.2

Table 6: Performance evaluation of ResNet20 on CIFAR-10

20 agents							
Model & dataset	PC-ASGD		P-ASGD		C-ASGD		Baseline
	acc. (%)	o.p. (%)	acc. (%)	o.p. (%)	acc.(%)	o.p. (%)	acc. (%)
ResNet 20, CIFAR-10	$\mathbf{84.9 \pm 0.6}$	$\mathbf{2.4 \pm 0.7}$	82.9 ± 0.7	0.4 ± 0.8	83.8 ± 0.8	1.3 ± 0.9	82.5 ± 0.1

acc.–accuracy, o.p.–outperformed comparing to baseline.

D More Empirical Results with different datasets

We also adopt our numerical studies on TinyImageNet Le & Yang (2015) and Wind turbine data set Liu et al. (2014). For TinyImageNet, we adopt PreResNet110 He et al. (2016b), DenseNet Huang et al. (2017), and EfficientNet Tan & Le (2019). For the wind turbine data set, we use LSTM[3] in Lei et al. (2019) to classify the fault in the wind turbine.

Results in Tab. 5 shows the effectiveness of our proposed methods in different models, datasets, and even different tasks (time series classification). It further demonstrates the generality of our proposed framework.

We also supplement the experiment with ResNet20 on CIFAR-10 to ablate the functions of the P-step and C-step in Tab. 6. The quantitative results are consistent with Tab. 2, showing the benefits of our PC steps design.

[3]Actually, we use SGD-based optimizer for better analysis instead of Adam in Lei et al. (2019), hence we do not achieve the best results in Lei et al. (2019). But our framework shows the best performances among other framework handling delay.