A. Appendix

A.1. Proof of Lemma 1

This section presents the detailed proof for Lemma 1. To begin with, we provide some technical auxiliary lemmas and the associated proof. We start with bounding the ensemble average of local optimal gradients.

The core update law for *CGA* is:

Lemma 2. *Let all assumptions hold. Let g^i be the unbiased estimate of $\nabla f_i(\mathbf{x}^i)$ at the point $\mathbf{x}^i$ such that $\mathbb{E}[\mathbf{g}^i] = \nabla f_i(\mathbf{x}^i)$, for all $i \in [N] := \{1, 2, ..., N\}$. Thus the following relationship holds*

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}^i\right\|^2\right] \leq \frac{2\sigma^2}{N} + 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^i)\right\|^2\right] + 2\epsilon^2. \tag{15}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}^i\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}^i - \mathbf{g}^i + \mathbf{g}^i)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}^i - \mathbf{g}^i) + \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}^i\right\|^2\right] \\
&\overset{a}{\leq} 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}^i - \mathbf{g}^i)\right\|^2 + \left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}^i\right\|^2\right] \overset{b}{\leq} 2\frac{1}{N^2}\mathbb{E}\left[N\sum_{i=1}^{N}\left\|\tilde{\mathbf{g}}^i - \mathbf{g}^i\right\|^2\right] + 2\left(\frac{\sigma^2}{N} + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^i)\right\|^2\right]\right) \\
&\leq \frac{2}{N}\mathbb{E}\left[\sum_{i=1}^{N}\left\|\tilde{\mathbf{g}}^i - \mathbf{g}^i\right\|^2\right] + 2\frac{\sigma^2}{N} + 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^i)\right\|^2\right] = \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}^i - \mathbf{g}^i\right\|^2\right] + 2\frac{\sigma^2}{N} + 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^i)\right\|^2\right] \\
&\overset{c}{\leq} 2\epsilon^2 + \frac{2\sigma^2}{N} + 2\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}^i)\right\|^2\right]
\end{aligned} \tag{16}$$

(a) refers to the fact that the inequality $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$. (b) holds as $\|\sum_{i=1}^{N}\mathbf{a}_i\|^2 \leq N\sum_{i=1}^{N}\|\mathbf{a}_i\|^2$. The second term in the second inequality is the conclusion of Lemma 1 in (Yu et al., 2019) (c) follows from Assumption 3. □

Multiplying the update law by $\frac{1}{N}\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1}$ is the column vector with entries being 1, we obtain:

$$\begin{aligned}
\bar{\mathbf{v}}_k &= \beta\bar{\mathbf{v}}_{k-1} - \alpha\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}^i_{k-1} \\
\bar{\mathbf{x}}_k &= \bar{\mathbf{x}}_{k-1} + \bar{\mathbf{v}}_k
\end{aligned} \tag{17}$$

We define an auxiliary sequence such that

$$\bar{\mathbf{z}}_k := \frac{1}{1-\beta}\bar{\mathbf{x}}_k - \frac{\beta}{1-\beta}\bar{\mathbf{x}}_{k-1} \tag{18}$$

Where $k > 0$. If $k = 0$ then $\bar{\mathbf{z}}_k = \bar{\mathbf{x}}_k$. For the rest of the analysis, the initial value will be directly set to 0.

Lemma 3. *Define the sequence $\{\bar{\mathbf{z}}_k\}_{k\geq 0}$ as in Eq. 18. Based on CGA, we have the following relationship*

$$\bar{\mathbf{z}}_{k+1} - \bar{\mathbf{z}}_k = -\frac{\alpha}{1-\beta}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}^i_k. \tag{19}$$

Proof. Using mathematical induction we have:

$$
\begin{aligned}
&k = 0: \\
&\bar{\mathbf{z}}_{k+1} - \bar{\mathbf{z}}_k = \bar{\mathbf{z}}_1 - \bar{\mathbf{z}}_0 = \frac{1}{1-\beta}\bar{\mathbf{x}}_1 - \frac{\beta}{1-\beta}\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_0 = \frac{1}{1-\beta}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0) = \frac{1}{1-\beta}(\bar{\mathbf{v}}_1) = \frac{-\alpha}{N(1-\beta)}\sum_{i=1}^{N}\tilde{\mathbf{g}}_0^i \\
&k \geq 1: \\
&\bar{\mathbf{z}}_{k+1} - \bar{\mathbf{z}}_k = \frac{1}{1-\beta}\bar{\mathbf{x}}_{k+1} - \frac{\beta}{1-\beta}\bar{\mathbf{x}}_k - \frac{1}{1-\beta}\bar{\mathbf{x}}_k + \frac{\beta}{1-\beta}\bar{\mathbf{x}}_{k-1} = \\
&\frac{1}{1-\beta}((\bar{\mathbf{x}}_{k+1} - \bar{\mathbf{x}}_k) - (\beta(\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{k-1}))) = \frac{1}{1-\beta}\underbrace{(\bar{\mathbf{v}}_{k+1} - \beta(\bar{\mathbf{v}}_k))}_{-\alpha\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i} = \frac{-\alpha}{N(1-\beta)}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i
\end{aligned}
\tag{20}
$$

□

Lemma 4. *Define respectively the sequence $\{\bar{\mathbf{x}}_k\}_{k\geq 0}$ as in Eq. 17 and the sequence $\{\bar{\mathbf{z}}_k\}_{k\geq 0}$ as in Eq. 18. For all $K \geq 1$, CGA ensures the following relationship*

$$
\sum_{k=0}^{K-1}\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\|^2 \leq \frac{\alpha^2\beta^2}{(1-\beta)^4}\sum_{k=0}^{K-1}\left\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\right\|^2. \tag{21}
$$

Proof. As $\bar{v}_0 = 0$, we can apply 17 recursively to achieve an update rule for $\bar{v}_k$. Therefor, we have :

$$
\bar{\mathbf{v}}_k = -\alpha\sum_{\tau=0}^{k-1}\beta^{k-1-\tau}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_\tau^i\right] \quad \forall k \geq 1 \tag{22}
$$

Also, based on Eq. 18 we have:

$$
\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k = \frac{\beta}{1-\beta}[\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_{k-1}] = \frac{\beta}{1-\beta}\bar{\mathbf{v}}_k \tag{23}
$$

Based on Equations 22 and 23 we have:

$$
\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k = \frac{-\alpha\beta}{1-\beta}\sum_{\tau=0}^{k-1}\beta^{k-1-\tau}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_\tau^i\right] \quad \forall k \geq 1 \tag{24}
$$

We define $s_k = \sum_{\tau=0}^{k-1}\beta^{k-1-\tau} = \frac{1-\beta^k}{1-\beta} \quad \forall k \geq 1$. We have:

$$
\begin{aligned}
&||\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k||^2 = \frac{\alpha^2\beta^2}{(1-\beta)^2}s_k^2\left\|\sum_{\tau=0}^{k-1}\frac{\beta^{k-1-\tau}}{s_k}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_\tau^i\right]\right\|^2 \overset{Jensen Inequality}{\leq} \\
&\frac{\alpha^2\beta^2}{(1-\beta)^2}s_k^2\sum_{\tau=0}^{k-1}\frac{\beta^{k-1-\tau}}{s_k}\left\|\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_\tau^i\right]\right\|^2 = \frac{\alpha^2\beta^2(1-\beta^k)}{(1-\beta)^3}\sum_{\tau=0}^{k-1}\beta^{k-1-\tau}\left\|\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_\tau^i\right]\right\|^2 \leq \\
&\frac{\alpha^2\beta^2}{(1-\beta)^3}\sum_{\tau=0}^{k-1}\beta^{k-1-\tau}\left\|\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_\tau^i\right]\right\|^2
\end{aligned}
\tag{25}
$$

Setting $K \geq 1$, As $\bar{\mathbf{z}}_0 - \bar{\mathbf{x}}_0 = 0$, by summing Eq. 25 over $k \in \{1, 2, \ldots, K-1\}$:

$$\begin{aligned}
&\sum_{k=0}^{K-1} ||\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k||^2 \leq \frac{\alpha^2\beta^2}{(1-\beta)^3} \sum_{k=1}^{K-1}\sum_{\tau=0}^{k-1} \beta^{k-1-\tau} \left\| \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_\tau^i\right]\right\|^2 \\
&= \frac{\alpha^2\beta^2}{(1-\beta)^3} \sum_{\tau=0}^{K-2} \left(\left\| \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_\tau^i\right]\right\|^2 \sum_{l=\tau+1}^{K-1} \beta^{l-1-\tau} \right) \overset{a}{\leq} \\
&\frac{\alpha^2\beta^2}{(1-\beta)^4} \sum_{\tau=0}^{K-2} \left\| \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_\tau^i\right]\right\|^2 \leq \frac{\alpha^2\beta^2}{(1-\beta)^4} \sum_{\tau=0}^{K-1} \left\| \left[\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_\tau^i\right]\right\|^2
\end{aligned} \tag{26}$$

Here (a) refers to $\sum_{l=\tau+1}^{K-1} \beta^{l-1-\tau} = \frac{1-\beta^{K-1-\tau}}{1-\beta} \leq \frac{1}{1-\beta}$. □

Before proceeding to prove Lemma 1, we introduce some key notations and facts that serve to characterize the lemma.

We define the following notations:

$$\begin{aligned}
\tilde{\mathbf{G}}_k &\triangleq [\tilde{\mathbf{g}}_k^1, \tilde{\mathbf{g}}_k^2, ..., \tilde{\mathbf{g}}_k^N] \\
\mathbf{V}_k &\triangleq [\mathbf{v}_k^1, \mathbf{v}_k^2, ..., \mathbf{v}_k^N] \\
\mathbf{X}_k &\triangleq [\mathbf{x}_k^1, \mathbf{x}_k^2, ..., \mathbf{x}_k^N] \\
\mathbf{G}_k &\triangleq [\mathbf{g}_k^1, \mathbf{g}_k^2, ..., \mathbf{g}_k^N] \\
\mathbf{H}_k &\triangleq [\nabla f_1(\mathbf{x}_k^1), \nabla f_2(\mathbf{x}_k^2), ..., \nabla f_N(\mathbf{x}_k^N)]
\end{aligned} \tag{27}$$

We can observe that the above matrices are all with dimension $d \times N$ such that any matrix $\mathbf{A}$ satisfies $\|\mathbf{A}\|_{\mathfrak{F}}^2 = \sum_{i=1}^{N} \|\mathbf{a}_i\|^2$, where $\mathbf{a}_i$ is the i-th column of the matrix $\mathbf{A}$. Thus, we can obtain that:

$$\|\mathbf{X}_k(\mathbf{I} - \mathbf{Q})\|_{\mathfrak{F}}^2 = \sum_{i=1}^{N} \|\mathbf{x}_k^i - \bar{\mathbf{x}}_k\|^2. \tag{28}$$

Fact 1. *Define* $\mathbf{Q} = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$. *For each doubly stochastic matrix* $\mathbf{\Pi}$*, the following properties can be obtained*

- $\mathbf{Q}\mathbf{\Pi} = \mathbf{\Pi}\mathbf{Q}$*;*
- $(\mathbf{I} - \mathbf{Q})\mathbf{\Pi} = \mathbf{\Pi}(\mathbf{I} - \mathbf{Q})$*;*
- *For any integer* $k \geq 1$, $\|(\mathbf{I} - \mathbf{Q})\mathbf{\Pi}\|_{\mathfrak{S}} \leq (\sqrt{\rho})^k$*, where* $\|\cdot\|_{\mathfrak{S}}$ *is the spectrum norm of a matrix.*

Fact 2. *Let* $\mathbf{A}_i, i \in \{1, 2, ..., N\}$ *be* N *arbitrary real square matrices. It follows that*

$$\|\sum_{i=1}^{N} \mathbf{A}_i\|_{\mathfrak{F}}^2 \leq \sum_{i=1}^{N}\sum_{j=1}^{N} \|\mathbf{A}_i\|_{\mathfrak{F}} \|\mathbf{A}_j\|_{\mathfrak{F}}. \tag{29}$$

The properties shown in Facts 1 and 2 have been well established and in this context, we skip the proof here. We are now ready to prove Lemma 1.

Proof. Since $\mathbf{X}_k = \mathbf{X}_{k-1}\mathbf{\Pi} + \mathbf{V}_k$ we have:

$$\mathbf{X}_k(\mathbf{I} - \mathbf{Q}) = \mathbf{X}_{k-1}(\mathbf{I} - \mathbf{Q})\mathbf{\Pi} + \mathbf{V}_k(\mathbf{I} - \mathbf{Q}) \tag{30}$$

Applying the above equation k times we have:

$$\mathbf{X}_k(\mathbf{I}-\mathbf{Q}) = \mathbf{X}_0(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^k + \sum_{\tau=1}^{k}\mathbf{V}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-\tau} \overset{\mathbf{X}_0=0}{=} \sum_{\tau=1}^{k}\mathbf{V}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-\tau} \tag{31}$$

As $\bar{\mathbf{V}}_k = \beta\bar{\mathbf{V}}_{k-1} - \alpha\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{G}}_{k-1}^i \overset{\mathbf{V}_0=0}{=} -\alpha\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{G}}_{k-1}^i$, we can get:

$$\begin{aligned}\mathbf{X}_k(\mathbf{I}-\mathbf{Q}) &= -\alpha\sum_{\tau=1}^{k}\sum_{l=0}^{\tau-1}\tilde{\mathbf{G}}_l\beta^{\tau-1-l}(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-\tau} = -\alpha\sum_{\tau=1}^{k}\sum_{l=0}^{\tau-1}\tilde{\mathbf{G}}_l\beta^{\tau-1-l}\mathbf{\Pi}^{k-\tau-l}(\mathbf{I}-\mathbf{Q})\\ &- \alpha\sum_{n=1}^{k-1}\tilde{\mathbf{G}}_n[\sum_{l=n+1}^{k}\beta^{l-1-n}\mathbf{\Pi}^{k-1-n}(\mathbf{I}-\mathbf{Q}) = -\alpha\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}\tilde{\mathbf{G}}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}.\end{aligned} \tag{32}$$

Therefore, for $k \geq 1$, we have:

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{X}_k(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right] = \alpha^2\mathbb{E}\left[\left\|\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}\tilde{\mathbf{G}}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right]\\ &\overset{a}{\leq} \underbrace{2\alpha^2\mathbb{E}\left[\left\|\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}(\tilde{\mathbf{G}}_\tau-\mathbf{G}_\tau)(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right]}_{I} + \underbrace{2\alpha^2\mathbb{E}\left[\left\|\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}\mathbf{G}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right]}_{II}\end{aligned} \tag{33}$$

(a) follows from the inequality $\|\mathbf{A}+\mathbf{B}\|_{\mathfrak{F}}^2 \leq 2\|\mathbf{A}\|_{\mathfrak{F}}^2 + 2\|\mathbf{B}\|_{\mathfrak{F}}^2$.

We develop upper bounds of term **I**:

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}(\tilde{\mathbf{G}}_\tau-\mathbf{G}_\tau)(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right] \overset{a}{\leq} \sum_{\tau=0}^{k-1}\mathbb{E}\left[\left\|\frac{1-\beta^{k-\tau}}{1-\beta}(\tilde{\mathbf{G}}_\tau-\mathbf{G}_\tau)(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right]\\ &\overset{b}{\leq} \frac{1}{(1-\beta)^2}\sum_{\tau=0}^{k-1}\rho^{k-1-\tau}\mathbb{E}\left[\left\|\tilde{\mathbf{G}}_\tau-\mathbf{G}_\tau\right\|_{\mathfrak{F}}^2\right] \overset{c}{\leq} \frac{1}{(1-\beta)^2}\sum_{\tau=0}^{k-1}\rho^{k-1-\tau}N\epsilon^2 \overset{d}{\leq} \frac{N\epsilon^2}{(1-\beta)^2(1-\rho)}\end{aligned} \tag{34}$$

(a) follows from Jensen inequality. (b) follows from the inequality $|\frac{1-\beta^{k-\tau}}{1-\beta}| \leq \frac{1}{1-\beta}$. (c) follows from Assumption 3 and Frobenius norm. (d) follows from Assumption 4.

We then proceed to find the upper bound for term **II**.

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}\mathbf{G}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right] \overset{a}{\leq} \sum_{\tau=0}^{k-1}\sum_{\tau'=0}^{k-1}\mathbb{E}\left[\left\|\frac{1-\beta^{k-\tau}}{1-\beta}\mathbf{G}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}\right.\\ &\left.\left\|\frac{1-\beta^{k-\tau}}{1-\beta}\mathbf{G}_{\tau'}(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau'}\right\|_{\mathfrak{F}}\right] \leq \frac{1}{(1-\beta)^2}\sum_{\tau=0}^{k-1}\sum_{\tau'=0}^{k-1}\rho^{(k-1-\frac{\tau+\tau'}{2})}\mathbb{E}\left[\|\mathbf{G}_\tau\|_{\mathfrak{F}}\|\mathbf{G}_{\tau'}\|_{\mathfrak{F}}\right] \overset{b}{\leq}\\ &\frac{1}{(1-\beta)^2}\sum_{\tau=0}^{k-1}\sum_{\tau'=0}^{k-1}\rho^{(k-1-\frac{\tau+\tau'}{2})}\left(\frac{1}{2}\mathbb{E}[\|\mathbf{G}_\tau\|_{\mathfrak{F}}^2] + \frac{1}{2}\mathbb{E}[\|\mathbf{G}_{\tau'}\|_{\mathfrak{F}}^2]\right) = \frac{1}{(1-\beta)^2}\sum_{\tau=0}^{k-1}\sum_{\tau'=0}^{k-1}\rho^{(k-1-\frac{\tau+\tau'}{2})}\mathbb{E}[\|\mathbf{G}_\tau\|_{\mathfrak{F}}^2]\\ &\overset{c}{\leq} \frac{1}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{\tau=0}^{k-1}\rho^{(\frac{k-1-\tau}{2})}\mathbb{E}[\|\mathbf{G}_\tau\|_{\mathfrak{F}}^2]\end{aligned} \tag{35}$$

(a) follows from Fact 2. (b) follows from the inequality $xy \leq \frac{1}{2}(x^2+y^2)$ for any two real numbers x, y. (c) is derived using $\sum_{\tau_1=0}^{k-1} \rho^{k-1-\frac{\tau_1+\tau}{2}} \leq \frac{\rho^{\frac{k-1-\tau}{2}}}{1-\sqrt{\rho}}$.

We then proceed with finding the bounds for $\mathbb{E}[\|\mathbf{G}_\tau\|_{\mathfrak{F}}^2]$:

$$\begin{aligned}\mathbb{E}[\|\mathbf{G}_\tau\|_{\mathfrak{F}}^2] &= \mathbb{E}[\|\mathbf{G}_\tau - \mathbf{H}_\tau + \mathbf{H}_\tau - \mathbf{H}_\tau\mathbf{Q} + \mathbf{H}_\tau\mathbf{Q}\|_{\mathfrak{F}}^2] \\ &\leq 3\mathbb{E}[\|\mathbf{G}_\tau - \mathbf{H}_\tau\|_{\mathfrak{F}}^2] + 3\mathbb{E}[\|\mathbf{H}_\tau(I-\mathbf{Q})\|^2\mathfrak{F}] + 3\mathbb{E}[\|\mathbf{H}_\tau\mathbf{Q}\|_{\mathfrak{F}}^2] \overset{a}{\leq} 3N\sigma^2 + 3N\delta^2 + 3\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2]\end{aligned} \tag{36}$$

(a) holds because $\mathbb{E}[\|\mathbf{H}_\tau\mathbf{Q}\|_{\mathfrak{F}}^2] \leq \mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2]$

Substituting (36) in (35):

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{\tau=0}^{k-1}\frac{1-\beta^{k-\tau}}{1-\beta}\mathbf{G}_\tau(\mathbf{I}-\mathbf{Q})\mathbf{\Pi}^{k-1-\tau}\right\|_{\mathfrak{F}}^2\right] \leq \frac{1}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{\tau=0}^{k-1}\rho^{(\frac{k-1-\tau}{2})}\left[3N\sigma^2 + 3N\delta^2 + 3\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2]\right] \\ &\leq \frac{3N(\sigma^2+\delta^2)}{(1-\beta)^2(1-\sqrt{\rho})^2} + \frac{3N}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{\tau=0}^{k-1}\rho^{(\frac{k-1-\tau}{2})}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2]\end{aligned} \tag{37}$$

substituting (37) and (34) into the main inequality (33):

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{X}_k(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right] \leq \frac{2\alpha^2 N\epsilon^2}{(1-\beta)^2(1-\rho)} + \frac{2\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\left(\frac{3N(\sigma^2)}{1-\sqrt{\rho}} + \frac{3N(\delta^2)}{1-\sqrt{\rho}} + \right. \\ &\left. 3N\sum_{\tau=0}^{k-1}\rho^{(\frac{k-1-\tau}{2})}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2]\right) = \frac{2\alpha^2}{(1-\beta)^2}\left(\frac{N\epsilon^2}{1-\rho} + \frac{3N\sigma^2}{(1-\sqrt{\rho})^2} + \frac{3N\delta^2}{(1-\sqrt{\rho})^2}\right) + \\ &\frac{6N\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{\tau=0}^{k-1}\rho^{(\frac{k-1-\tau}{2})}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2]\end{aligned} \tag{38}$$

Summing over $k \in \{1, \ldots, K-1\}$ and noting that $\mathbb{E}\left[\left\|\mathbf{X}_0(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right] = 0$:

$$\begin{aligned}&\sum_{k=1}^{K-1}\mathbb{E}\left[\left\|\mathbf{X}_k(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right] \leq CK + \frac{6N\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{k=1}^{K-1}\sum_{\tau=0}^{k-1}\rho^{(\frac{k-1-\tau}{2})}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_\tau^i)\|^2] \leq \\ &CK + \frac{6N\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{k=0}^{K-1}\frac{1-\rho^{(\frac{K-1-k}{2})}}{1-\sqrt{\rho}}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_k^i)\|^2] \leq \\ &CK + \frac{6N\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{k=0}^{K-1}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_k^i)\|^2]\end{aligned} \tag{39}$$

Where $C = \frac{2\alpha^2}{(1-\beta)^2}\left(\frac{N\epsilon^2}{1-\rho} + \frac{3N\sigma^2}{(1-\sqrt{\rho})^2} + \frac{3N\delta^2}{(1-\sqrt{\rho})^2}\right)$.

Dividing both sides by N:

$$\sum_{k=1}^{K-1}\frac{1}{N}\mathbb{E}\left[\left\|\mathbf{X}_k(\mathbf{I}-\mathbf{Q})\right\|_{\mathfrak{F}}^2\right]\leq \\ \frac{2\alpha^2}{(1-\beta)^2}\left(\frac{\epsilon^2}{1-\rho}+\frac{3\sigma^2}{(1-\sqrt{\rho})^2}+\frac{3\delta^2}{(1-\sqrt{\rho})^2}\right)K+\frac{6\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{k=0}^{K-1}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2] \tag{40}$$

We immediately have:

$$\sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\bar{\mathbf{x}}_k-\mathbf{x}_k^i\right\|^2\right]\leq \\ \frac{2\alpha^2}{(1-\beta)^2}\left(\frac{\epsilon^2}{1-\rho}+\frac{3\sigma^2}{(1-\sqrt{\rho})^2}+\frac{3\delta^2}{(1-\sqrt{\rho})^2}\right)K+\frac{6\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})}\sum_{k=0}^{K-1}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2] \tag{41}$$

□

A.2. Proof for Theorem 1

Proof. Using the smoothness properties for $\mathcal{F}$ we have:

$$\mathbb{E}[\mathcal{F}(\bar{\mathbf{z}}_{k+1})]\leq\mathbb{E}[\mathcal{F}(\bar{\mathbf{z}}_k)]+\mathbb{E}[\langle\nabla\mathcal{F}(\bar{\mathbf{z}}_k),\bar{\mathbf{z}}_{k+1}-\bar{\mathbf{z}}_k\rangle]+\frac{L}{2}\mathbb{E}[\|\bar{\mathbf{z}}_{k+1}-\bar{\mathbf{z}}_k\|^2] \tag{42}$$

Using Lemma 3 we have:

$$\mathbb{E}[\langle\nabla\mathcal{F}(\bar{\mathbf{z}}_k),\bar{\mathbf{z}}_{k+1}-\bar{\mathbf{z}}_k\rangle]=\frac{-\alpha}{1-\beta}\mathbb{E}[\langle\nabla\mathcal{F}(\bar{\mathbf{z}}_k),\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\rangle]= \\ \underbrace{\frac{-\alpha}{1-\beta}\mathbb{E}[\langle\nabla\mathcal{F}(\bar{\mathbf{z}}_k)-\nabla\mathcal{F}(\bar{\mathbf{x}}_k),\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i\rangle]}_{I}\underbrace{-\frac{\alpha}{1-\beta}\mathbb{E}[\langle\nabla\mathcal{F}(\bar{\mathbf{x}}_k),\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i\rangle]}_{II} \tag{43}$$

We proceed by analysing (I):

$$\frac{-\alpha}{1-\beta}\mathbb{E}[\langle\nabla\mathcal{F}(\bar{\mathbf{z}}_k)-\nabla\mathcal{F}(\bar{\mathbf{x}}_k),\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i\rangle]\leq \\ \frac{(1-\beta)}{2\beta L}\mathbb{E}[\|\nabla\mathcal{F}(\bar{\mathbf{z}}_k)-\nabla\mathcal{F}(\bar{\mathbf{x}}_k)\|^2]+\frac{\beta L\alpha^2}{2(1^{-\beta})^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\|^2]\leq \\ \frac{(1-\beta)L}{2\beta}\mathbb{E}[\|\bar{\mathbf{z}}_k-\bar{\mathbf{x}}_k\|^2]+\frac{\beta L\alpha^2}{2(1-\beta)^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\|^2] \tag{44}$$

For term (II) we have:

$$\langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_k^i \rangle = \langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \left(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i + \mathbf{g}_k^i\right) \rangle = \underbrace{\langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \left(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i\right) \rangle}_{\star} + \underbrace{\langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_k^i \rangle}_{\star\star} \tag{45}$$

We first analyse ($\star$):

$$\frac{-\alpha}{(1-\beta)}\mathbb{E}[\langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \left(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i\right) \rangle] \leq \frac{(1-\beta)\alpha^2}{2\beta L}\mathbb{E}[\|\nabla \mathcal{F}(\bar{\mathbf{x}}_k)\|^2] + \frac{\beta L}{2(1-\beta)^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i)\|^2] \tag{46}$$

This holds as $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2$ where $\mathbf{a} = \frac{-\alpha\sqrt{1-\beta}}{\beta L}\nabla \mathcal{F}(\bar{\mathbf{x}}_k)$ and $\mathbf{b} = -\frac{\sqrt{\beta L}}{(1-\beta)^{\frac{3}{2}}}\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i)$.

Analysing ($\star\star$):

$$\mathbb{E}\left[\langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{g}}_k^i \rangle\right] = \mathbb{E}\left[\langle \nabla \mathcal{F}(\bar{\mathbf{x}}_k), \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(\mathbf{x}_k^i) \rangle\right] \tag{47}$$

The above equality holds because $\bar{\mathbf{x}}_k$ and $\mathbf{x}_k^i$ are determined by $\zeta_{k-1} = [\zeta_0, \ldots, \zeta_{k-1}]$ which is independent of ζ_k, and $\mathbb{E}[\mathbf{g}_k^i|\zeta_{k-1}] = \mathbb{E}[\mathbf{g}_k^i] = \nabla f_i(\mathbf{x}_k^i)$. With the aid of the equity $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2}[\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a} - \mathbf{b}\|^2]$, we have :

$$\langle \nabla \mathcal{F}\left(\bar{\mathbf{x}}_k\right), \frac{1}{N}\sum_{i=1}^{N} \nabla f_i\left(\mathbf{x}_k^i\right) \rangle = \frac{1}{2}\left(\|\nabla F\left(\bar{\mathbf{x}}_k\right)\|^2 + \|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2 - \|\nabla \mathcal{F}(\bar{\mathbf{x}}_k) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2\right) \overset{a}{\geq} \frac{1}{2}\left(\|\nabla \mathcal{F}(\bar{\mathbf{x}}_k)\|^2 + \|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2 - L^2\frac{1}{N}\sum_{i=1}^{N}\|\bar{\mathbf{x}}_k - \mathbf{x}_k^i\|^2\right) \tag{48}$$

(a) follows because $\|\nabla \mathcal{F}(\bar{\mathbf{x}}_k) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2 = \|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{\mathbf{x}}_k) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\|\nabla f_i(\bar{\mathbf{x}}_k) - \nabla f_i(\mathbf{x}_k^i)\|^2 \leq \frac{1}{N}\sum_{i=1}^{N} L^2\|\bar{\mathbf{x}}_k - \mathbf{x}_k^i\|^2$.

Substituting (48) into (47) and (46), (47) into (45) and (44), (45) into (43):

$$\mathbb{E}[\langle \nabla \mathcal{F}(\bar{\mathbf{z}}_k), \bar{\mathbf{z}}_{k+1} - \bar{\mathbf{z}}_k \rangle] \leq \frac{(1-\beta)L}{2\beta}\mathbb{E}[\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\|^2] + \frac{\beta L\alpha^2}{2(1-\beta)^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i)\|^2] + \left(\frac{(1-\beta)\alpha^2}{2\beta L} - \frac{\alpha}{2(1-\beta)}\right)$$
$$\mathbb{E}[\|\nabla \mathcal{F}(\bar{\mathbf{x}}_k)\|^2] - \frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2] + \frac{\beta L}{2(1-\beta)^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i)\|^2] + \frac{\alpha L^2}{2(1-\beta)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\bar{\mathbf{x}}_k - \mathbf{x}_k^i\|^2] \tag{49}$$

Lemma 3 states that:

$$\mathbb{E}[\|\bar{\mathbf{z}}_{k+1} - \bar{\mathbf{z}}_k\|^2] = \frac{\alpha^2}{(1-\beta)^2}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\|^2]. \tag{50}$$

Substituting (49),(50) in (42):

$$\mathbb{E}[\mathcal{F}(\bar{\mathbf{z}}_{k+1})] \leq \mathbb{E}[\mathcal{F}(\bar{\mathbf{z}}_k)] + \frac{(1-\beta)L}{2\beta}\mathbb{E}[\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\|^2] + \frac{\beta L\alpha^2}{2(1-\beta)^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i)\|^2] + \left(\frac{(1-\beta)\alpha^2}{2\beta L} - \frac{\alpha}{2(1-\beta)}\right)$$
$$\mathbb{E}[\|\nabla\mathcal{F}(\bar{\mathbf{x}}_k)\|^2] - \frac{\alpha}{2(1-\beta)}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2] + \frac{\beta L}{2(1-\beta)^3}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i)\|^2] +$$
$$\frac{\alpha L^2}{2(1-\beta)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\|\bar{\mathbf{x}}_k - \mathbf{x}_k^i\|^2] + \frac{\alpha^2}{(1-\beta)^2}\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\|^2]. \tag{51}$$

Rearranging the terms and dividing by $C_1 = \frac{\alpha}{2(1-\beta)} - \frac{(1-\beta)\alpha^2}{2\beta L}$to find the bound for $\mathbb{E}[\|\nabla\mathcal{F}(\bar{\mathbf{x}}_k)\|^2]$:

$$\mathbb{E}[\|\nabla\mathcal{F}(\bar{\mathbf{x}}_k)\|^2] \leq \frac{1}{C_1}\left(\mathbb{E}[\mathcal{F}(\bar{\mathbf{z}}_k)] - \mathbb{E}[\mathcal{F}(\bar{\mathbf{z}}_{k+1})]\right) + C_2\,\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i)\|^2] + C_3\,\mathbb{E}[\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\|^2]$$
$$- C_6\,\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_k^i)\|^2] + C_4\,\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i)\|^2] + C_5\sum_{i=1}^{N}\mathbb{E}[\|\bar{\mathbf{x}}_k - \mathbf{x}_k^i\|^2] \tag{52}$$

Where $C_2 = \left(\frac{\beta L\alpha^2}{2(1-\beta)^3} + \frac{\alpha^2 L}{(1-\beta)^2}\right)/C_1$, $C_3 = \frac{(1-\beta)L}{2\beta}/C_1$, $C_4 = \frac{\beta L}{2(1-\beta)^3}/C_1$, $C_5 = \frac{\alpha L^2}{2(1-\beta)}/C_1$, $C_6 = \frac{\alpha}{2(1-\beta)}/C_1$.

Summing over $k \in \{0, 1, \ldots, K-1\}$:

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\mathcal{F}\left(\bar{\mathbf{x}}_k\right)\|^2\right] \leq \frac{1}{C_1}\left(\mathbb{E}\left[\mathcal{F}\left(\bar{\mathbf{z}}_0\right)\right] - \mathbb{E}\left[\mathcal{F}\left(\bar{\mathbf{z}}_k\right)\right]\right) - C_6\sum_{k=0}^{k-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_k^i\right)\right\|^2\right] + C_2\sum_{k=0}^{k-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{g}}_k^i\right\|^2\right]$$
$$+ C_3\sum_{k=0}^{k-1}\mathbb{E}\left[\left\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\right\|^2\right] + C_4\sum_{k=0}^{k-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\mathbf{g}}_k^i - \mathbf{g}_k^i\right)\right\|^2\right] + C_5\sum_{k=0}^{k-1}\frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\left\|\bar{\mathbf{x}}_k - \mathbf{x}_k^i\right\|^2\right] \tag{53}$$

Substituting Lemma 1, Lemma 2, and Lemma 4 and Assumption 3 into the above equation we have:

$$\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\mathcal{F}\left(\bar{\mathbf{x}}_k\right)\|^2\right] \leq \frac{1}{C_1}\left(\mathbb{E}\left[\mathcal{F}\left(\bar{\mathbf{z}}_0\right)\right] - \mathbb{E}\left[\mathcal{F}\left(\bar{\mathbf{z}}_k\right)\right]\right) - \left(C_6 - C_5\frac{6\alpha^2}{(1-\beta)(1-\sqrt{\rho})} - 2C_2 - 2C_3\frac{\alpha^2\beta^2}{(1-\beta)^4}\right)$$
$$\sum_{k=0}^{k-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(\mathbf{x}_k^i\right)\right\|^2\right] + \left(C_2 + C_3\frac{\alpha^2\beta}{(1-\beta)^4}\right)\left(\frac{2\sigma^2}{N} + 2\epsilon^2\right)K + C_4\epsilon^2 K + C_5\frac{2\alpha^2}{(1-\beta)^2}\left(\frac{\epsilon^2}{1-\rho} + \right. \tag{54}$$
$$\left.\frac{3\sigma^2}{(1-\sqrt{\rho})^2} + \frac{3\delta^2}{(1-\sqrt{\rho})^2}\right)K$$

Dividing both sides by K:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\nabla\mathcal{F}\left(\bar{\mathbf{x}}_k\right)\|^2\right] \leq \frac{1}{C_1 K}\left(\mathcal{F}\left(\bar{\mathbf{x}}_0\right) - \mathcal{F}^\star\right) + \left(C_2 + C_3\frac{\alpha^2\beta}{(1-\beta)^4}\right)\left(\frac{2\sigma^2}{N} + 2\epsilon^2\right) + C_4\epsilon^2 +$$
$$C_5\frac{2\alpha^2}{(1-\beta)^2}\left(\frac{\epsilon^2}{1-\rho} + \frac{3\sigma^2}{(1-\sqrt{\rho})^2} + \frac{3\delta^2}{(1-\sqrt{\rho})^2}\right)K \tag{55}$$

The above follows from the fact that $\bar{\mathbf{z}}_0 = \bar{\mathbf{x}}_0$ and $\left(C_6 - C_5\frac{6\alpha^2}{(1-\beta)(1-\sqrt{\rho})} - 2C_2 - 2C_3\frac{\alpha^2\beta^2}{(1-\beta)^4}\right) \geq 0$.

Therefor we have :

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{\mathbf{x}}_k\right)\right\|^2\right] \leq \frac{1}{C_1 K}\left(\mathcal{F}\left(\bar{\mathbf{x}}_0\right) - \mathcal{F}^*\right) + \left(2C_2 + C_3\frac{\alpha^2\beta}{(1-\beta)^4} + C_4 + C_5\frac{2\alpha^2}{(1-\beta)^2(1-\rho)}\right)\epsilon^2 + \left(\frac{2}{N}\left(C_2 + C_3\frac{\alpha^2\beta}{(1-\beta)^4}\right) + C_5\frac{6\alpha^2}{(1-\beta)^2(1-\sqrt{p})^2}\right)\sigma^2 + C_5\frac{6\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})^2}\delta^2 \tag{56}$$

□

A.3. Discussion on the Step Size

Recalling the conditions for the step size α in Theorem 1,

$$1 - \frac{6\alpha^2L^2}{(1-\beta)(1-\sqrt{\rho})^2} - \frac{4L\alpha}{(1-\beta)^2} \geq 0.$$

Solving the last inequality, combining the fact that $\alpha > 0$, we have then the specific form of α^*

$$\alpha^* = \frac{(1-\sqrt{\rho})\sqrt{16(1-\sqrt{\rho})^2 + 24(1-\beta)^3} - 4(1-\sqrt{\rho})^2}{12L(1-\beta)}.$$

Therefore, if the step size α is defined as

$$\alpha \leq \min\left\{\frac{\beta L}{(1-\beta)^2}, \frac{(1-\sqrt{\rho})\sqrt{16(1-\sqrt{\rho})^2 + 24(1-\beta)^3} - 4(1-\sqrt{\rho})^2}{12L(1-\beta)}\right\},$$

Eq. 56 naturally holds true.

A.4. Proof for Corollary 1

Proof. According to Eq. 56, on the right hand side, there are four terms with different coefficients with respect to the step size α. We separately investigate each term in the following. As $C_1 = \mathcal{O}(\frac{\sqrt{N}}{\sqrt{K}})$. Therefore,

$$\frac{\mathcal{F}(\bar{\mathbf{x}}_0) - \mathcal{F}^*}{C_1 K} = \mathcal{O}(\frac{1}{\sqrt{NK}}).$$

While for the second term, we have

$$C_2 = \mathcal{O}(\frac{\sqrt{N}}{\sqrt{K}}), C_3 = \mathcal{O}(\frac{\sqrt{K}}{\sqrt{N}}), C_4 = \mathcal{O}(\frac{\sqrt{K}}{\sqrt{N}}), C_5 = \mathcal{O}(1),$$

such that

$$2C_2\epsilon^2 = \mathcal{O}(\frac{\sqrt{N}}{K^{1.5}}), C_3\frac{\alpha^2\beta}{(1-\beta)^4}\epsilon^2 = \mathcal{O}(\frac{\sqrt{N}}{K^{1.5}}), C_4\epsilon^2 = \mathcal{O}(\frac{1}{\sqrt{NK}}), C_5\frac{2\alpha^2}{(1-\beta)^2(1-\rho)}\epsilon^2 = \mathcal{O}(\frac{N}{K^2}).$$

Similarly, we can obtain for the third term and the last term,

$$\frac{2}{N}\left(C_2 + C_3\frac{\alpha^2\beta}{(1-\beta)^4}\right)\sigma^2 = \mathcal{O}(\frac{1}{\sqrt{NK}}), C_5\frac{6\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})^2}\sigma^2 = \mathcal{O}(\frac{N}{K}),$$

and

$$C_5\frac{6\alpha^2}{(1-\beta)^2(1-\sqrt{\rho})^2}\delta^2 = \mathcal{O}(\frac{N}{K}).$$

Hence, By omitting the constant N in this context, there exists a constant $C > 0$ such that the overall convergence rate is as follows:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\mathcal{F}\left(\bar{\mathbf{x}}_k\right)\right\|^2\right] \leq C\left(\frac{1}{\sqrt{NK}} + \frac{1}{K} + \frac{1}{K^{1.5}} + \frac{1}{K^2}\right), \tag{57}$$

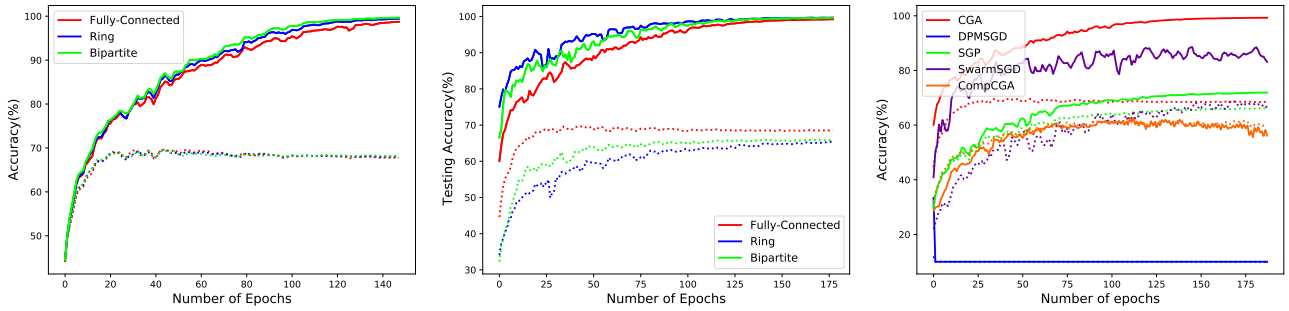
which suggests when N is fixed and K is sufficiently large, *CGA* enables the convergence rate of $\mathcal{O}(\frac{1}{\sqrt{NK}})$. □

A.5. Additional CIFAR-10 Results

In this section, we provide more experimental results for CIFAR10 dataset trained using a CNN architecture and more complex VGG11 model architecture:

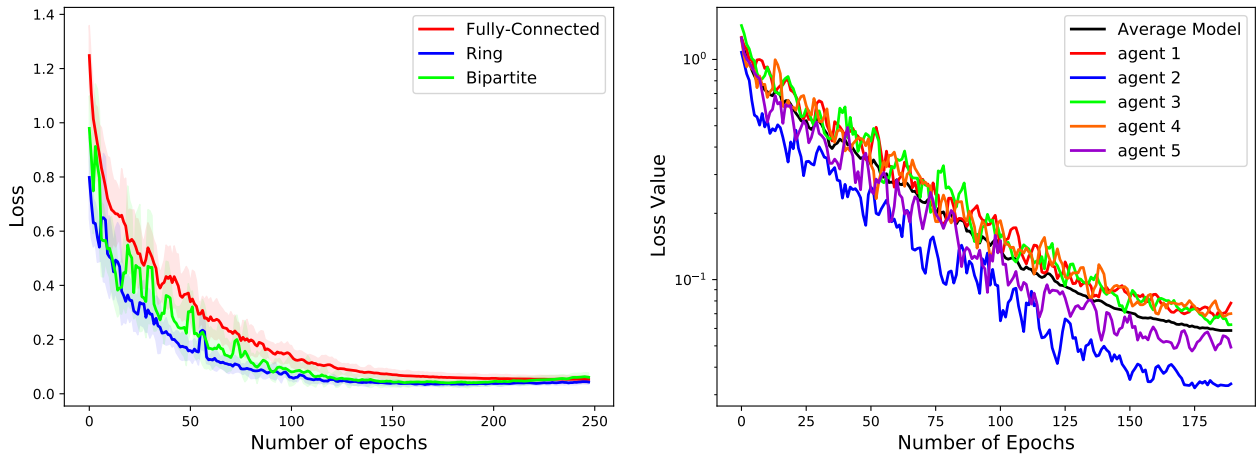
Additional CIFAR10 results trained using CNN :

We start by providing the corresponding accuracy plots for Figure 2 in the main paper:



(a) (b) (c)

Figure 5. Average training and validation accuracy for (a) CGA method on IID (b) CGA method on non-IID data distributions (c) different methods on non-IID data distributions for training 5 agents using CNN model architecture



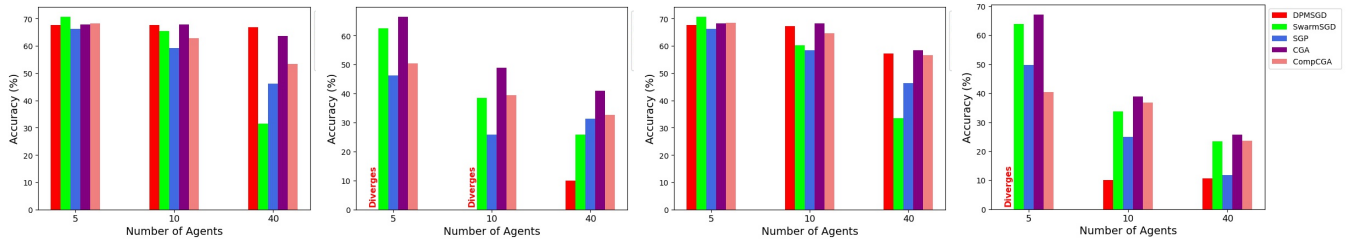
(a) (b)

Figure 6. Average training loss for (a) different topologies trained using CGA algorithm (b) individual agents along with the average model during training using CGA algorithm (log scale)

Based on Figure 5(a), (b) *CGA* achieves a high accuracy for different graph topologies when learning from both IID and non-IID data distributions. However other methods i.e. DPMSGD suffer from maintaining the high accuracy when learning from non-IID data distributions. The adverse effect of non-IIDness in the data can be more elaborated upon by looking at Figure 7. Comparing (a) with (b) and (c) with (d) we can see that although the migration from IID to non-IID affects all the methods, *CGA* suffers less than other methods for different combinations of graph topology and graph type. The same observation can be made by looking at Figure 8 which shows the accuracy obtained for different methods *w.r.t* the graph type.

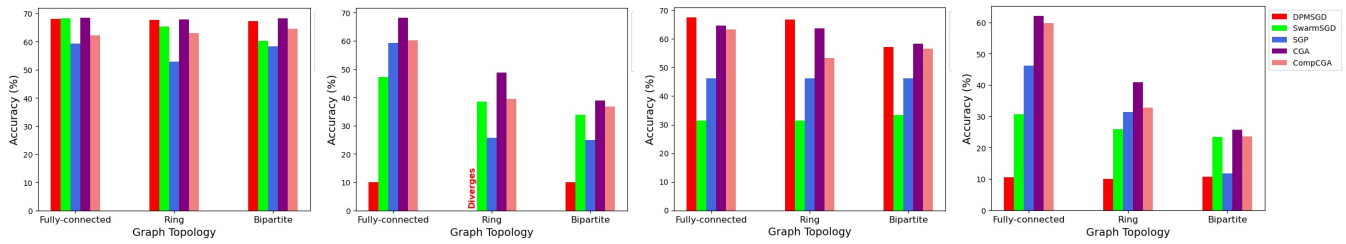
While Figure 2(a) harps on the phenomenon of faster convergence with sparser graph topology which is an observation that have been made by earlier research works in Federated Learning (McMahan et al., 2017) by reducing the client fraction which makes the mixing matrix sparser and decentralized learning (Jiang et al., 2017). However, as Figure 6(a) shows, by training for more epochs, all converge to similar loss values. Figure 6 shows that the loss value associated with the

consensus model is very close to the loss values corresponding to all other agents which means the projected gradient using QP is capturing the correct direction.



(a) (b) (c) (d)

Figure 7. Average testing accuracy for different methods w.r.t the number of learning agents learning from (a) IID data distributions for Ring graph topology (b) non-IID data distributions for Ring graph topology (c) IID data distributions for Bipartite graph topology (d) non-IID data distributions for Bipartite graph topology



(a) (b) (c) (d)

Figure 8. Average testing accuracy for different methods w.r.t the graph topology learning from (a) IID data distributions learning from 10 agents (b) non-IID data distributions learning from 10 agents(c) IID data distributions learning from 40 agents (d) non-IID data distributions learning from 40 agents

CIFAR10 with VGG11:

We now extend our experimental analysis by using a more complex model architecture (e.g. VGG11) for CIFAR10 dataset. Tables 4 and 5 summarize the performance of *CGA* compared to other methods. Similar to CNN model architecture, *CGA* can maintain the performance when migrating from IID to non-IID data distributions. However, we observe that as VGG11 model is much more complex than CNN, all the methods suffer from an increase in the number of learning agents and complexity of graph topology.

A.6. MNIST Results

Same as what we did for CIFAR-10, we are comparing different methods performance on MNIST dataset. The results are summarized in Tables 6 and 7. Although the accuracies are generally high when learning from MNIST dataset, and most of the methods work in most of the settings, we can see that although *CGA* can maintain the model performance while learning from non-IID data, DPMSGD, SGP and SwarmSGD suffer from non-IIDness in the data specially when the number of agents and the graph topology combinations become more complex.

Table 4. Model Accuracy Comparison for training CIFAR10 using VGG11 with IID data distribution

Model	Fully-connected	Ring	Bipartite
DPMSGD	67.8% (5)	61.9% (5)	61.0% (5)
	60.8% (10)	60.5% (10)	60.7% (10)
	59.8% (40)	**60.1%** (40)	**60.1%** (40)
SGP	72.5% (5)	72.0% (5)	71.1% (5)
	70.3% (10)	42.8% (10)	**70.2%** (10)
	41.1% (40)	41.6% (40)	41.5% (40)
SwarmSGD	75.8% (5)	73.1% (5)	78.3% (5)
	71.5% (10)	**71.4%** (10)	70.1% (10)
	21.8% (40)	20.6% (40)	20.3% (40)
CGA (ours)	**81.1%** (5)	**81.8%** (5)	**81.5%** (5)
	68.8% (10)	68.3% (10)	68.2% (10)
	21.9% (40)	18.5% (40)	20.3% (40)

Table 5. Model Accuracy Comparison for training CIFAR10 with non-IID data distribution using VGG11

Model	Fully-connected	Ring	Bipartite
DPMSGD	Diverges (5)	Diverges (5)	Diverges (5)
	Diverges (10)	Diverges (10)	10% (10)
	12% (40)	Diverges (40)	10.7% (40)
SGP	20.4% (5)	20.8% (5)	20.3% (5)
	10.1% (10)	10.0% (10)	Diverges (10)
	Diverges (40)	10.0% (40)	10.1% (40)
SwarmSGD	19.4% (5)	19.9% (5)	20.2% (5)
	10.0% (10)	Diverges (10)	Diverges (10)
	9.9% (40)	10.2% (40)	10% (40)
CGA (ours)	**74.6%** (5)	**75.8%** (5)	**77.5%** (5)
	69.8% (10)	**38.9%** (10)	**18.7%** (10)
	12.8% (40)	**20.5%** (40)	**23.6%** (40)

Table 6. Model Accuracy Comparison for training MNIST using CNN with IID data distribution

Model	Fully-connected	Ring	Bipartite
DPSGD	**98.8%** (5)	**98.8%** (5)	**98.8%** (5)
	98.6% (10)	**98.5%** (10)	98.5% (10)
	96.9% (40)	**96.8%** (40)	**96.8%** (40)
SGP	96.2% (5)	96.3% (5)	96.2% (5)
	93.2% (10)	93.2% (10)	93.2% (10)
	71.4% (40)	71.4% (40)	71.4% (40)
SwarmSGD	98.4% (5)	98.4% (5)	98.5% (5)
	96.1% (10)	96.1% (10)	96.0% (10)
	38.3% (40)	38.3% (40)	39.7% (40)
CGA (ours)	98.6 % (5)	98.7% (5)	98.7% (5)
	98.2% (10)	98.3% (10)	**98.6%** (10)
	94.7% (40)	95.5% (40)	**96.8%** (40)

Table 7. Model Accuracy Comparison for training MNIST with non-IID data distribution using CNN

Model	**Fully-connected**	**Ring**	**Bipartite**
DPSGD	98.3% (5)	98.2% (5)	98.2% (5)
	87.1% (10)	74.5% (10)	70.9% (10)
	85.3% (40)	72.5% (40)	34.3% (40)
SGP	95.9% (5)	96.0% (5)	95.9% (5)
	92.7% (10)	91.3% (10)	90.2% (10)
	71.2% (40)	74.6% (40)	62.2% (40)
SwarmSGD	98.2% (5)	98.1% (5)	98.2% (5)
	93.2% (10)	90.9% (10)	91.4% (10)
	24.8% (40)	33.5% (40)	18.3% (40)
CGA (ours)	**98.6%** (5)	**98.5%** (5)	**98.5%** (5)
	98.2% (10)	**96.2%** (10)	**96.2%** (10)
	94.1% (40)	**91.6%** (40)	**91.8%** (40)