Supplementary Material for Paper:
Moniqua: Modulo Quantized Communication in Decentralized SGD

A. Overview

This supplementary material contains proof to all the theoretical results. It is organized as follows: In Section B, we analyze how to work with Modulo and quantization, as proofs to Lemma 1 and Lemma 2 in the paper. In Section C, we provably explain why using shared randomness in communication with stochastic rounding can improve performance. In Section D, we illustrate why directly quantizing communication in D-PSGD fails to converge asymptotically, as a proof to Theorem 1. In Section E, we introduce some useful tools of modeling communication as a Markov Chain for the rest of the proof (part of the intuition is illustrated in the paper). We recommend to go through this before getting into Section F to H. Finally we will provide proof to Theorem 2 to 5 from Section F to H.

B. Modulo Operation with Quantization

Proof to Lemma 1.

Proof. Rewrite x and y as

$$\begin{aligned} x &= N_x a + r_x, -\frac{a}{2} \leq r_x < \frac{a}{2} \\ y &= N_y a + r_y, -\frac{a}{2} \leq r_y < \frac{a}{2} \end{aligned}$$

where N_x, $N_y \in \mathbb{Z}$ then,

$$\begin{aligned} \text{LHS} &= (r_x - r_y) \bmod a \\ \text{RHS} &= ((N_x - N_y)a + r_x - r_y) \bmod a = (r_x - r_y) \bmod a = \text{LHS} \end{aligned}$$

Thus we complete the proof. $\square$

Proof to Lemma 2.

Proof. We start from

$$B_\theta \mathcal{Q}_\delta \left(\frac{x}{B_\theta} \bmod 1 \right) - B_\theta \left(\frac{x}{B_\theta} \bmod 1 \right) + x = B_\theta \mathcal{Q}_\delta \left(\frac{x}{B_\theta} \bmod 1 \right) - B_\theta \left(\frac{x}{B_\theta} \bmod 1 \right) + x - y + y$$

If B_θ is sufficiently large such that $B_\theta \geq 2\theta + 2\delta B_\theta > 2|x-y| + 2\delta B_\theta$, we could put a "$\bmod B_\theta$" to the first four terms as follows:

$$\begin{aligned} & B_\theta \mathcal{Q}_\delta \left(\frac{x}{B_\theta} \bmod 1 \right) - B_\theta \left(\frac{x}{B_\theta} \bmod 1 \right) + x - y + y \\ =& \left(B_\theta \mathcal{Q}_\delta \left(\frac{x}{B_\theta} \bmod 1 \right) - B_\theta \left(\frac{x}{B_\theta} \bmod 1 \right) + x - y \right) \bmod B_\theta + y \\ \overset{\text{Lemma 1}}{=}& \left[\left(B_\theta \mathcal{Q}_\delta \left(\frac{x}{B_\theta} \bmod 1 \right) - B_\theta \left(\frac{x}{B_\theta} \bmod 1 \right) + x \right) \bmod B_\theta - y \bmod B_\theta \right] \bmod B_\theta + y \\ \overset{\text{Lemma 1}}{=}& \left\{ \left[B_\theta \mathcal{Q}_\delta \left(\frac{x}{B_\theta} \bmod 1 \right) \bmod B_\theta - \left(B_\theta \left(\frac{x}{B_\theta} \bmod 1 \right) - x \right) \bmod B_\theta \right] \bmod B_\theta - y \bmod B_\theta \right\} \bmod B_\theta + y \end{aligned}$$

Note that the term $\left(B_\theta\left(\frac{x}{B_\theta} \bmod 1\right)-x\right) \bmod B_\theta = 0$, then we can proceed as:

$$\begin{aligned}
&\left\{\left[B_\theta \mathcal{Q}_\delta\left(\frac{x}{B_\theta} \bmod 1\right) \bmod B_\theta - \left(B_\theta\left(\frac{x}{B_\theta} \bmod 1\right)-x\right) \bmod B_\theta\right] \bmod B_\theta - y \bmod B_\theta\right\} \bmod B_\theta + y \\
=&\left(B_\theta \mathcal{Q}_\delta\left(\frac{x}{B_\theta} \bmod 1\right) \bmod B_\theta - y \bmod B_\theta\right) \bmod B_\theta + y \\
=&\left(B_\theta \mathcal{Q}_\delta\left(\frac{x}{B_\theta} \bmod 1\right) - y\right) \bmod B_\theta + y
\end{aligned}$$

By moving x to the right side we obtain

$$\left|\left(B_\theta \mathcal{Q}_\delta\left(\frac{x}{B_\theta} \bmod 1\right) - y\right) \bmod B_\theta + y - x\right| = \left|B_\theta \mathcal{Q}_\delta\left(\frac{x}{B_\theta} \bmod 1\right) - B_\theta\left(\frac{x}{B_\theta} \bmod 1\right)\right| \le \delta B_\theta$$

That completes the proof. □

C. Shared Randomness

In this section, we provide a theoretical explanation why using shared randomness in the stochastic rounding is able to improve the performance. Without the loss of generality, in the following analysis, we let the quantization step associated with stochastic rounding quantizer $\mathcal{Q}_\delta$ be $\delta = 1$. For any $z \in \mathbb{R}$ quantized using $\mathcal{Q}_\delta$, let $z_f = z - \lfloor z \rfloor$, the variance of quantization error can be expressed as

$$\mathbb{E}|\mathcal{Q}_\delta(z) - z|^2 = (1-z_f)(-z_f)^2 + z_f(1-z_f)^2 = z_f(1-z_f) \tag{1}$$

Note that in Moniqua, the term asssociate with quantization error is

$$\mathbb{E}\left\|(\boldsymbol{q}_{k,j} - \boldsymbol{x}_{k,j}) - (\boldsymbol{q}_{k,i} - \boldsymbol{x}_{k,i})\right\|^2$$

We now show for $\forall x, y \in \mathbb{R}$

$$\mathbb{E}\left|(\mathcal{Q}_\delta(x) - x) - (\mathcal{Q}_\delta(y) - y)\right|^2 = \mathbb{E}\left|\mathcal{Q}_\delta(y-x) - (y-x)\right|^2$$

With out the loss of generality, let $x - \lfloor x \rfloor \le y - \lfloor y \rfloor$. Let $x_f = x - \lfloor x \rfloor$ and $y_f = y - \lfloor y \rfloor$, then

$$\begin{aligned}
&\lfloor x+u \rfloor = \lfloor x \rfloor \quad \text{and} \quad \lfloor y+u \rfloor = \lfloor y \rfloor, \text{with probability} \quad \lceil y \rceil - y \\
&\lfloor x+u \rfloor = \lceil x \rceil \quad \text{and} \quad \lfloor y+u \rfloor = \lceil y \rceil, \text{with probability} \quad x - \lfloor x \rfloor \\
&\lfloor x+u \rfloor = \lfloor x \rfloor \quad \text{and} \quad \lfloor y+u \rfloor = \lceil y \rceil, \text{with probability} \quad (\lceil x \rceil - x) - (\lceil y \rceil - y)
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left|(\mathcal{Q}_\delta(x) - x) - (\mathcal{Q}_\delta(y) - y)\right|^2 \\
=&\mathbb{E}\left|\left(\delta\left\lfloor\frac{x}{\delta}+u\right\rfloor - x\right) - \left(\delta\left\lfloor\frac{y}{\delta}+u\right\rfloor - y\right)\right|^2 \\
=&(\lceil y \rceil - y)((\lfloor x \rfloor - x) - (\lfloor y \rfloor - y))^2 + (x - \lfloor x \rfloor)((\lceil x \rceil - x) - (\lceil y \rceil - y))^2 \\
&+ ((\lceil x \rceil - x) - (\lceil y \rceil - y))((\lfloor x \rfloor - x) - (\lceil y \rceil - y))^2 \\
=&(1-y_f)(x_f - y_f)^2 + (x_f)(x_f - y_f) + (y_f - x_f)(y_f - x_f - 1)^2 \\
=&(1 - y_f + x_f)(y_f - x_f)^2 + (y_f - x_f)(y_f - x_f - 1)^2 \\
=&(1 - y_f + x_f)(y_f - x_f) \\
=&\mathbb{E}\left|\mathcal{Q}_\delta(y-x) - (y-x)\right|^2
\end{aligned}$$

The last equality holds due to equation 1. Next, for $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ let

$$\boldsymbol{\Delta} = \boldsymbol{y} - \boldsymbol{x}$$

$$\boldsymbol{r} = \mathcal{Q}_\delta(\boldsymbol{\Delta}) - \boldsymbol{\Delta}$$

And let $\boldsymbol{r}_h$ denote h-th entry of $\boldsymbol{r}$, let $\boldsymbol{\Delta}_h$ denote h-th entry of $\boldsymbol{\Delta}$. We obtain

$$\begin{aligned}
\boldsymbol{r}_h =& \mathcal{Q}_\delta(\boldsymbol{\Delta}_h) - \boldsymbol{\Delta}_h \\
=& \delta \begin{cases} -\frac{\boldsymbol{\Delta}_h}{\delta} + \left\lfloor \frac{\boldsymbol{\Delta}_h}{\delta} \right\rfloor + 1, & p_t \le \frac{\boldsymbol{\Delta}_h}{\delta} - \left\lfloor \frac{\boldsymbol{\Delta}_h}{\delta} \right\rfloor \\ -\frac{\boldsymbol{\Delta}_h}{\delta} + \left\lfloor \frac{\boldsymbol{\Delta}_h}{\delta} \right\rfloor, & \text{otherwise} \end{cases} \\
=& \delta \begin{cases} -q + 1, & p_t \le q \\ -q, & \text{otherwise} \end{cases}
\end{aligned}$$

where

$$q = \frac{\boldsymbol{\Delta}_h}{\delta} - \left\lfloor \frac{\boldsymbol{\Delta}_h}{\delta} \right\rfloor, q \in [0, 1]$$

Based on that, we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{r}_h^2\right] \le& \delta^2((-q+1)^2 q + (-q)^2(1-q)) \\
=& \delta^2 q(1-q) \\
\le& \delta^2 \min\{q, 1-q\}
\end{aligned}$$

Since $\min\{q, 1-q\} \le \left|\frac{\boldsymbol{x}_h}{\delta}\right|$, we have

$$\mathbb{E}\left[\boldsymbol{r}_h^2\right] \le \delta^2 \left|\frac{\boldsymbol{\Delta}_h}{\delta}\right| \le \delta \left|\boldsymbol{\Delta}_h\right|$$

Summing over the index h yields,

$$\mathbb{E}\left\|\boldsymbol{r}\right\|_2^2 \le \delta \mathbb{E}\left\|\boldsymbol{\Delta}\right\|_1 \le \sqrt{d}\delta \mathbb{E}\left\|\boldsymbol{\Delta}\right\|_2$$

Pushing back $\boldsymbol{x}$ and $\boldsymbol{r}$, we have

$$\mathbb{E}\left\|\mathcal{Q}_\delta(\boldsymbol{y}-\boldsymbol{x}) - (\boldsymbol{y}-\boldsymbol{x})\right\|^2 \le \sqrt{d}\delta \mathbb{E}\left\|\boldsymbol{y}-\boldsymbol{x}\right\| = \sqrt{d}\delta \mathbb{E}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|$$

Putting it back we have

$$\mathbb{E}\left\|(\mathcal{Q}_\delta(\boldsymbol{x}) - \boldsymbol{x}) - (\mathcal{Q}_\delta(\boldsymbol{y}) - \boldsymbol{y})\right\|^2 \le \sqrt{d}\delta \mathbb{E}\left\|\boldsymbol{x}-\boldsymbol{y}\right\|$$

Now we can see that the error term is bounded by the distance of two quantized tensor, which, in decentralized training, refers to the distance between two models on adjacent workers. In such a way, the error bound can be reduced since the workers are getting close to each other.

D. Why Naive Quantization Fails in D-PSGD (Proof to Theorem 1)

The update rule of naive quantization on D-PSGD is

$$\boldsymbol{x}_{k+1,i} = \boldsymbol{x}_{k,i}\boldsymbol{W}_{ii} + \sum_{j=1, j\neq i}^{n} \mathcal{Q}_\delta(\boldsymbol{x}_{k,j})\boldsymbol{W}_{ji} - \alpha_k \tilde{\boldsymbol{g}}_{k,i} = \boldsymbol{x}_{k,i} + \sum_{j=1, j\neq i}^{n} (\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,i})\boldsymbol{W}_{ji} - \alpha_k \tilde{\boldsymbol{g}}_{k,i}$$

where α_k is allowed to vary with any policy. Let

$$\begin{aligned}
\boldsymbol{X}_k =& [\boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n}] \in \mathbb{R}^{d\times n} \\
\boldsymbol{\Omega}_k =& \left[\sum_{j\neq 1} \boldsymbol{W}_{j1}\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,1}\right), \cdots, \sum_{j\neq n} \boldsymbol{W}_{jn}\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,n}\right)\right] \in \mathbb{R}^{d\times n} \\
\tilde{\boldsymbol{G}}_k =& \left[\tilde{\boldsymbol{g}}_{k,1}, \cdots, \tilde{\boldsymbol{g}}_{k,n}\right] \in \mathbb{R}^{d\times n}
\end{aligned}$$

by rewritting the update rule, we obtain

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \boldsymbol{\Omega}_k - \alpha_k \tilde{\boldsymbol{G}}_k$$

Let $\boldsymbol{Y}_k = \boldsymbol{X}_k - \boldsymbol{x}^*\mathbf{1}^\top$, and considering the fact that $\nabla f(\boldsymbol{x}) = \boldsymbol{x} - \delta\mathbf{1}/2 = \boldsymbol{x} - \boldsymbol{x}^*$, we can rewrite the update rule as

$$\boldsymbol{Y}_{k+1}\boldsymbol{e}_i = \boldsymbol{Y}_k\boldsymbol{e}_i + \boldsymbol{\Omega}_k\boldsymbol{e}_i - \alpha_k\boldsymbol{Y}_k\boldsymbol{e}_i + \alpha_k\left(\tilde{\boldsymbol{G}}_k - \boldsymbol{G}_k\right)\boldsymbol{e}_i$$

where $\left(\tilde{\boldsymbol{G}}_k - \boldsymbol{G}_k\right)$ denotes variance in the gradient sampling.

Suppose that by using the update rule of naive quantization, worker i converges to $\boldsymbol{x}^*$. Then there must exist a K such that $\forall k \geq K$,

$$\mathbb{E}\left\|\boldsymbol{Y}_{k+1}\boldsymbol{e}_i\right\|^2 \leq \mathbb{E}\left\|\boldsymbol{Y}_k\boldsymbol{e}_i\right\|^2 < \frac{\phi^2\delta^2}{8(1+\phi^2)} \tag{2}$$

Next we show that this assumption lets us derive a contradiction. Firstly, considering the property of linear quantizer,

$$\frac{\delta^2}{4} \leq \mathbb{E}\left\|\mathcal{Q}_\delta(\boldsymbol{x}_{k,i}) - \boldsymbol{x}^*\right\|^2 \leq 2\mathbb{E}\left\|\mathcal{Q}_\delta(\boldsymbol{x}_{k,i}) - \boldsymbol{x}_{k,i}\right\|^2 + 2\mathbb{E}\left\|\boldsymbol{x}_{k,i} - \boldsymbol{x}^*\right\|^2$$

As a result

$$\mathbb{E}\left\|\mathcal{Q}_\delta(\boldsymbol{x}_{k,i}) - \boldsymbol{x}_{k,i}\right\|^2 \geq \frac{\delta^2}{8} - \frac{\phi^2\delta^2}{8(1+\phi^2)} = \frac{\delta^2}{8(1+\phi^2)}$$

Since $\mathcal{Q}_\delta$ is unbiased, that means $\mathbb{E}[\mathcal{Q}_\delta(\boldsymbol{x}) - \boldsymbol{x}] = 0$, then we have

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{\Omega}_k\boldsymbol{e}_i\right\|^2 \\
=&\mathbb{E}\left\|\sum_{j\neq i}\boldsymbol{W}_{ji}\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,i}\right)\right\|^2 \\
=&\sum_{j\in\mathcal{N}_i}\boldsymbol{W}_{ji}^2\mathbb{E}\left\|\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,i}\right)\right\|^2 + \sum_{m\neq n\neq i}\mathbb{E}\left\langle\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,m}) - \boldsymbol{x}_{k,i}\right)\boldsymbol{W}_{mi}, \left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,n}) - \boldsymbol{x}_{k,i}\right)\boldsymbol{W}_{ni}\right\rangle \\
\geq&\phi^2\sum_{j\in\mathcal{N}_i}\mathbb{E}\left\|\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,i}\right)\right\|^2 + \sum_{m\neq n\neq i}\mathbb{E}\left\langle\left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,m}) - \boldsymbol{x}_{k,i}\right)\boldsymbol{W}_{mi}, \left(\mathcal{Q}_\delta(\boldsymbol{x}_{k,n}) - \boldsymbol{x}_{k,i}\right)\boldsymbol{W}_{ni}\right\rangle \\
\overset{(*)}{=}&\phi^2\sum_{j\in\mathcal{N}_i}\mathbb{E}\left\|\mathcal{Q}_\delta(\boldsymbol{x}_{k,j}) - \boldsymbol{x}_{k,i}\right\|^2 \\
\geq&\frac{\phi^2\delta^2}{8(1+\phi^2)}
\end{aligned}$$

where step $(*)$ holds due to unbiased quantizer. Putting it back to the update rule, we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{Y}_{k+1}\boldsymbol{e}_i\right\|^2 \\
=&\mathbb{E}\left\|\left(\boldsymbol{Y}_k + \boldsymbol{\Omega}_k - \alpha_k\boldsymbol{Y}_k + \alpha_k\left(\tilde{\boldsymbol{G}}_k - \boldsymbol{G}_k\right)\right)\boldsymbol{e}_i\right\|^2 \\
\overset{(*)}{=}&\mathbb{E}\left\|(1-\alpha_k)\boldsymbol{Y}_k\boldsymbol{e}_i\right\|^2 + \mathbb{E}\left\|\boldsymbol{\Omega}_k\boldsymbol{e}_i\right\|^2 + \mathbb{E}\left\|\alpha_k\left(\tilde{\boldsymbol{G}}_k - \boldsymbol{G}_k\right)e_i\right\|^2 \\
\geq&\mathbb{E}\left\|\boldsymbol{\Omega}_k\boldsymbol{e}_i\right\|^2 \\
\geq&\frac{\phi^2\delta^2}{8(1+\phi^2)}
\end{aligned}$$

where cross terms in the $(*)$ step are all 0 due to the unbiased quantizer and unbiased sampling of the gradient. Her we obtain the contradictory that $\frac{\phi^2\delta^2}{8(1+\phi^2)} \leq \mathbb{E}\left\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\right\|^2 < \frac{\phi^2\delta^2}{8(1+\phi^2)}$. That being said, for $\forall k, i$

$$\mathbb{E}\left\|\boldsymbol{x}_{k,i} - \boldsymbol{x}^*\right\|^2 = \mathbb{E}\left\|\nabla f(\boldsymbol{x}_{k,i})\right\|^2 \geq \frac{\phi^2\delta^2}{8(1+\phi^2)}$$

Thus we complete the proof.

E. A Markov Chain Analysis on the Communication

To better understand how the parallel workers reach consensus over a communication matrix, in this section we use theory from the analysis of Markov Chains to obtain some useful lemmas for proof of Moniqua on D-PSGD and AD-PSGD.

Since the communication matrix $\boldsymbol{W}$ is doubly stochastic (each row and column sum to 1), it has the same structure as the transition matrix of a Markov Chain with $\frac{\mathbf{1}}{n}$ as its the stationary distribution $\left(\boldsymbol{W}\frac{\mathbf{1}}{n} = \frac{\mathbf{1}}{n}\right)$. Now let t_{mix} and $d(t)$ denote the mixing time and maximal distance between initial state and stationary distribution as defined in Markov Chain theory.[1]

E.1. D-PSGD

In D-PSGD, the communication matrix is fixed during the training. That makes it perfectly aligned with the structure of a Markov Chain. As a result, we obtain the following lemma:

Lemma E.1.

$$\left\|\boldsymbol{W}^t\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1 \leq 2\cdot 2^{-\left\lfloor \frac{t}{t_{\text{mix}}} \right\rfloor}$$

Proof. For $\forall \boldsymbol{x} \in \mathbb{R}^d$, let $\boldsymbol{u} \in \mathbb{R}^d$ be such a vector that every entry of $\boldsymbol{u}$ is the positive entry of $\boldsymbol{x}$ and 0 otherwise. Let $\boldsymbol{v} \in \mathbb{R}^d$ be such a vector that every entry of $\boldsymbol{v}$ is the absolute value of negative entry of $\boldsymbol{x}$ and 0 otherwise. The setting above means $\boldsymbol{x} = \boldsymbol{u} - \boldsymbol{v}$. For example,

$$\begin{aligned}\boldsymbol{x} &= [2, -1]^\top \\ \boldsymbol{u} &= [2, 0]^\top \\ \boldsymbol{v} &= [0, 1]^\top\end{aligned}$$

And we have

$$\begin{aligned}
&\left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{x}\right\|_1 \\
=&\left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)(\boldsymbol{u}-\boldsymbol{v})\right\|_1 \\
\leq&\left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{u}\right\|_1 + \left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{v}\right\|_1 \\
=&\mathbf{1}^\top\boldsymbol{u}\left\|\boldsymbol{W}^t\frac{\boldsymbol{u}}{\mathbf{1}^\top\boldsymbol{u}} - \frac{\mathbf{1}}{n}\right\|_1 + \mathbf{1}^\top\boldsymbol{v}\left\|\boldsymbol{W}^t\frac{\boldsymbol{v}}{\mathbf{1}^\top\boldsymbol{v}} - \frac{\mathbf{1}}{n}\right\|_1 \\
\leq&2(\mathbf{1}^\top\boldsymbol{u} + \mathbf{1}^\top\boldsymbol{v})d(t) \\
\leq&2d(t)\left\|\boldsymbol{x}\right\|_1
\end{aligned}$$

Considering the definition of L1-norm, we have

$$\left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1 = \max\frac{\left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{x}\right\|_1}{\|\boldsymbol{x}\|_1} \leq 2d(t)$$

According to a well-known results on the theory of Markov Chains,[2] $d(lt_{\text{mix}}) \leq 2^{-l}$ holds for any non-negative integer l, so we have

$$\left\|\boldsymbol{W}^t\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1 \leq 2d(t) \leq 2d\left(\frac{t}{t_{\text{mix}}}\cdot t_{\text{mix}}\right) \leq 2d\left(\left\lfloor\frac{t}{t_{\text{mix}}}\right\rfloor t_{\text{mix}}\right) \leq 2\cdot 2^{-\left\lfloor \frac{t}{t_{\text{mix}}} \right\rfloor}$$

That completes the proof. $\square$

[1] Here we are using notation from Chapter 4.5 of *Markov Chains and Mixing Times* (Levin 2009), available at `https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf`

[2] Again, see *Markov Chains and Mixing Times* for more details.

Additionally, based on standard results in the theory of reversible Markov Chains, we also have[3]

$$t_{\text{mix}} \leq \log\left(\frac{1}{\frac{1}{4}\cdot\frac{1}{n}}\right)\frac{1}{1-\rho} \leq \frac{\log(4n)}{1-\rho}.$$

E.2. AD-PSGD

Note that unlike D-PSGD, here $\boldsymbol{W}_k$ can be different at each update step and usually each individually have spectral radius $\rho = 1$, so we can't expect to get a bound in terms of a bound on the spectral gap as we did in Theorems 2 and 3. Instead, we require the following condition, which is inspired by the literature on Markov chain Monte Carlo methods: for some constant t_{mix} (here t_{mix} is the same as t_{mix} in the paper) and for any k and any non-negative vector $\boldsymbol{\mu} \in \mathbb{R}^d$ such that $\mathbf{1}^\top \boldsymbol{\mu} = 1$, it must hold that

$$\left\|\left(\prod_{i=1}^{t_{\text{mix}}} \boldsymbol{W}_{k+i}\right)\boldsymbol{\mu} - \frac{\mathbf{1}}{n}\right\|_1 \leq \frac{1}{2}.$$

We call this constant t_{mix} because it is effectively the *mixing time* of the time-inhomogeneous Markov chain with transition probability matrix $\boldsymbol{W}_k$ at time k. Note that this condition is more general than those used in previous work on AD-PSGD because it does not require that the $\boldsymbol{W}_k$ are sampled independently or in an unbiased manner. Based on the above analysis, we can prove the following lemma, which is analogous to the lemma used in the synchronous case.

Lemma E.2. *For any $k \geq 0$ and for any $b \geq a \geq 0$, there exists t_{mix} such that*

$$\left\|\prod_{q=a}^{b} \boldsymbol{W}_q\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1 \leq 2\cdot 2^{-\left\lfloor \frac{b-a+1}{t_{\text{mix}}}\right\rfloor}$$

Proof. Note that for any $\boldsymbol{x} \in \mathbb{R}^d$, and let $\boldsymbol{u}$ and $\boldsymbol{v}$ be two vectors having same definition as in Lemma E.1 with respect to $\boldsymbol{x}$, then we have for any k

$$\begin{aligned}
&\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{x}\right\|_1 \\
=&\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)(\boldsymbol{u}-\boldsymbol{v})\right\|_1 \\
\leq&\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{u}\right\|_1 + \left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{v}\right\|_1 \\
=&\mathbf{1}^\top\boldsymbol{u}\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\frac{\boldsymbol{u}}{\mathbf{1}^\top\boldsymbol{u}} - \frac{\mathbf{1}}{n}\right\|_1 + \mathbf{1}^\top\boldsymbol{v}\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\frac{\boldsymbol{v}}{\mathbf{1}^\top\boldsymbol{v}} - \frac{\mathbf{1}}{n}\right\|_1 \\
\leq&\frac{1}{2}(\mathbf{1}^\top\boldsymbol{u} + \mathbf{1}^\top\boldsymbol{v}) \\
\leq&\frac{1}{2}\left\|\boldsymbol{x}\right\|_1
\end{aligned}$$

Considering the definition of the induced ℓ_1 operator norm, we have

$$\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1 = \max_{\boldsymbol{x}} \frac{\left\|\prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{q+k}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\boldsymbol{x}\right\|_1}{\left\|\boldsymbol{x}\right\|_1} \leq \frac{1}{2}$$

As a result, from the submultiplicativity of the matrix induced norm, we obtain

$$\left\|\prod_{q=a}^{b} \boldsymbol{W}_q\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1$$

[3] Detailed analysis and proofs of this result can be found in chapter 12.2 of *Markov Chains and Mixing Times*.

$$
\begin{aligned}
&\leq \left\| \prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{a-1+q} \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right) \right\|_1 \cdots \left\| \prod_{q=1}^{t_{\text{mix}}} \boldsymbol{W}_{\cdots+q} \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right) \right\|_1 \cdot \left\| \prod_{q=1}^{t_r} \boldsymbol{W}_{\cdots+q} \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right) \right\|_1 \\
&\leq 2^{-\left\lfloor \frac{b-a+1}{t_{\text{mix}}} \right\rfloor} \left\| \prod_{q=1}^{t_r} \boldsymbol{W}_{\cdots+q} \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right) \right\|_1
\end{aligned}
$$

where $t_r = (b-a+1) \bmod t_{\text{mix}}$. Note that

$$
\left\| \prod_{q=1}^{t_r} \boldsymbol{W}_q \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right) \right\|_1 \leq 1 - \frac{1}{n} + (n-1)\frac{1}{n} = 2 - \frac{2}{n} \leq 2
$$

Putting it back we obtain

$$
\left\| \prod_{q=a}^{b} \boldsymbol{W}_{\cdots+q} \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right) \right\|_1 \leq 2 \cdot 2^{-\left\lfloor \frac{b-a+1}{t_{\text{mix}}} \right\rfloor}
$$

That completes the proof. □

Note that in the analysis of Moniqua on AD-PSGD (Section H), we will use this lemma as an assumption.

F. Moniqua on D-PSGD (Proof to Theorem 2 and 3)

F.1. Notations

For convenience, we adopt the following notation

$$
\begin{aligned}
\boldsymbol{X}_k &= \left[\boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n}\right], \qquad & \hat{\boldsymbol{X}}_k &= \left[\hat{\boldsymbol{x}}_{k,1}, \cdots, \hat{\boldsymbol{x}}_{k,n}\right] \\
\tilde{\boldsymbol{G}}_k &= \left[\tilde{\boldsymbol{g}}_{k,1}, \cdots, \tilde{\boldsymbol{g}}_{k,n}\right], \qquad & \boldsymbol{G}_k &= \left[\boldsymbol{g}_{k,1}, \cdots, \boldsymbol{g}_{k,n}\right] \\
\overline{\boldsymbol{X}} &= \boldsymbol{X}\frac{\mathbf{1}}{n}, \forall \boldsymbol{X} \in \mathbb{R}^{d\times n}, \qquad & \boldsymbol{\Omega}_k &= (\hat{\boldsymbol{X}}_k - \boldsymbol{X}_k)(\boldsymbol{W} - \boldsymbol{I})
\end{aligned}
$$

where $\boldsymbol{g}_{k,i}$ denotes gradient computed via the whole dataset $\mathcal{D}_i$ and $\boldsymbol{x}_{k,i}$

From a local view, the update rule on worker i at iteration k can be written as

$$
\boldsymbol{x}_{k+1,i} \leftarrow \boldsymbol{x}_{k,i} + \sum\nolimits_{j\in\mathcal{N}_i} \left(\hat{\boldsymbol{x}}_{k,j} - \hat{\boldsymbol{x}}_{k,i}\right) \boldsymbol{W}_{ji} - \alpha_k \tilde{\boldsymbol{g}}_{k,i}
$$

which is equivalent to

$$
\boldsymbol{x}_{k+1,i} = \sum_{j=1}^{n} \boldsymbol{x}_{k,j} \boldsymbol{W}_{ji} - \alpha_k \tilde{\boldsymbol{g}}_{k,i} + \sum_{j=1}^{n} \left((\hat{\boldsymbol{x}}_{k,j} - \boldsymbol{x}_{k,j}) - (\hat{\boldsymbol{x}}_{k,i} - \boldsymbol{x}_{k,i}) \right) \boldsymbol{W}_{ji} \tag{3}
$$

with a more compact notation, this can be expressed as:

$$
\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \hat{\boldsymbol{X}}_k(\boldsymbol{W} - \boldsymbol{I}) - \alpha_k \tilde{\boldsymbol{G}}_k = \boldsymbol{X}_k \boldsymbol{W} - \alpha_k \tilde{\boldsymbol{G}}_k + (\hat{\boldsymbol{X}}_k - \boldsymbol{X}_k)(\boldsymbol{W} - \boldsymbol{I}) \tag{4}
$$

F.2. Proof to Theorem 2.

Proof. From Lemma F.4 we have

$$
\begin{aligned}
\sum_{k=0}^{K-1} \alpha_k \mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \leq & 4(\mathbb{E}f(\mathbf{0}) - \mathbb{E}f^*) + \frac{2\sigma^2 L}{n} \sum_{k=0}^{K-1} \alpha_k^2 + \frac{8\sigma^2 L^2}{(1-\rho)^2} \sum_{k=0}^{K-1} \alpha_k^3 + \frac{24\varsigma^2 L^2}{(1-\rho)^2} \sum_{k=0}^{K-1} \alpha_k^3 \\
& + \frac{8L^2}{n(1-\rho)^2} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

Note that

$$\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2=\sum_{k=0}^{K-1}\alpha_k\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{k,j}-\boldsymbol{x}_{k,j})-(\hat{\boldsymbol{x}}_{k,i}-\boldsymbol{x}_{k,i})\right)\boldsymbol{W}_{ji}\right\|^2\overset{\text{Lemma } F.1, F.3}{\leq}4\sum_{k=0}^{K-1}\alpha_k\delta^2B_{\theta_k}^2nd$$

By using Lemma F.3 and by assigning $\delta=\frac{1-\eta\rho}{8C_\alpha^2\eta\log(16n)+2(1-\eta\rho)}$, we obtain

$$\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2\leq\frac{G_\infty^2dn}{C_\alpha^2}\sum_{k=0}^{K-1}\alpha_k^3$$

Pushing it back we obtain

$$\begin{aligned}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq&4(\mathbb{E}f(\mathbf{0})-\mathbb{E}f^*)+\frac{2\sigma^2L}{n}\sum_{k=0}^{K-1}\alpha_k^2+\frac{8\sigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{24\varsigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3\\&+\frac{8G_\infty^2dL^2}{(1-\rho)^2C_\alpha^2}\sum_{k=0}^{K-1}\alpha_k^3\end{aligned}$$

That completes the proof. □

F.3. Proof to Corollary 1.

Proof. When $\alpha_k=\alpha$, $C_\alpha=\eta=1$, and we have:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq\frac{4(f(\mathbf{0})-f^*)}{\alpha K}+\frac{2\alpha L}{n}\sigma^2+\frac{8\alpha^2L^2\left(\sigma^2+3\varsigma^2\right)}{(1-\rho)^2}+\frac{8\alpha^2G_\infty^2dL^2}{(1-\rho)^2}$$

By setting $\alpha=\frac{1}{\varsigma^{\frac{2}{3}}K^{\frac{1}{3}}+\sigma\sqrt{\frac{K}{n}}+2L}$, we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq&\frac{8(f(\mathbf{0})-f^*)L}{K}+\frac{4\sigma(f(\mathbf{0})-f^*+L/2)}{\sqrt{nK}}+\frac{4\varsigma^{\frac{2}{3}}(f(\mathbf{0})-f^*)}{K^{\frac{2}{3}}}\\&+\frac{8L^2\sigma^2n}{(1-\rho)^2(\sigma^2K+4nL^2)}+\frac{24L^2\varsigma^{\frac{2}{3}}}{(1-\rho)^2K^{\frac{2}{3}}}+\frac{8G_\infty^2dnL^2}{(1-\rho)^2(\sigma^2K+4nL^2)}\\\lesssim&\frac{1}{K}+\frac{\sigma}{\sqrt{nK}}+\frac{\varsigma^{\frac{2}{3}}}{K^{\frac{2}{3}}}+\frac{\sigma^2n}{\sigma^2K+n}+\frac{G_\infty^2dn}{\sigma^2K+n}\end{aligned}$$

That completes the proof of Corollary 1. □

F.4. Lemma for Moniqua on D-PSGD

Lemma F.1. *If $\|\boldsymbol{x}_{t,i}-\boldsymbol{x}_{t,j}\|_\infty<\theta_t$, $\forall i,j$ holds at iteration t, then*

$$\left\|\sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{t,j}-\boldsymbol{x}_{t,j})-(\hat{\boldsymbol{x}}_{t,i}-\boldsymbol{x}_{t,i})\right)\boldsymbol{W}_{ji}\right\|_\infty\leq\frac{4\delta}{1-2\delta}\theta_t$$

Proof. Let $B_{\theta_t}=\frac{2}{1-2\delta}\theta_t$, based on the algorithm, we obtain

$$\begin{aligned}\hat{\boldsymbol{x}}_{t,j}&=\left(B_{\theta_t}\mathcal{Q}_\delta\left(\frac{\boldsymbol{x}_{t,j}}{B_{\theta_t}}\bmod 1\right)-\boldsymbol{x}_{t,i}\right)\bmod B_{\theta_t}+\boldsymbol{x}_{t,i}\\\hat{\boldsymbol{x}}_{t,i}&\overset{\text{Lemma 2}}{=}B_{\theta_t}\mathcal{Q}_\delta\left(\frac{\boldsymbol{x}_{t,i}}{B_{\theta_t}}\bmod 1\right)-B_{\theta_t}\left(\frac{\boldsymbol{x}_{t,i}}{B_{\theta_t}}\bmod 1\right)+\boldsymbol{x}_{t,i}\end{aligned}$$

We start from

$$\begin{aligned}\left\|\sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{t,j}-\boldsymbol{x}_{t,j})-(\hat{\boldsymbol{x}}_{t,i}-\boldsymbol{x}_{t,i})\right)\boldsymbol{W}_{ji}\right\|_{\infty} &\leq \sum_{j=1}^{n}\boldsymbol{W}_{ji}\left\|(\hat{\boldsymbol{x}}_{t,j}-\boldsymbol{x}_{t,j})-(\hat{\boldsymbol{x}}_{t,i}-\boldsymbol{x}_{t,i})\right\|_{\infty}\\ &\leq \sum_{j=1}^{n}\boldsymbol{W}_{ji}\left\|\hat{\boldsymbol{x}}_{t,j}-\boldsymbol{x}_{t,j}\right\|_{\infty}+\sum_{j=1}^{n}\boldsymbol{W}_{ji}\left\|\hat{\boldsymbol{x}}_{t,i}-\boldsymbol{x}_{t,i}\right\|_{\infty}\end{aligned}$$

On the first hand, due to Lemma 2 we obtain

$$\left\|\hat{\boldsymbol{x}}_{t,j}-\boldsymbol{x}_{t,j}\right\|_{\infty}\leq \delta B_{\theta_t}$$

on the other hand,

$$\left\|\hat{\boldsymbol{x}}_{t,i}-\boldsymbol{x}_{t,i}\right\|_{\infty}=\left\|B_{\theta_t}\mathcal{Q}_{\delta}\left(\frac{\boldsymbol{x}_{t,i}}{B_{\theta_t}}\bmod 1\right)-B_{\theta_t}\left(\frac{\boldsymbol{x}_{t,i}}{B_{\theta_t}}\bmod 1\right)\right\|_{\infty}\leq \delta B_{\theta_t}$$

Putting it back, we obtain

$$\left\|\sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{t,j}-\boldsymbol{x}_{t,j})-(\hat{\boldsymbol{x}}_{t,i}-\boldsymbol{x}_{t,i})\right)\boldsymbol{W}_{ji}\right\|_{\infty}\leq 2\delta B_{\theta_t}=\frac{4\delta}{1-2\delta}\theta_t$$

which completes the proof. □

Lemma F.2. *For any $\boldsymbol{X}_t\in\mathbb{R}^{d\times n}$, we have*

$$\left\|\sum_{t=0}^{k-1}\boldsymbol{X}_t\left(\frac{\mathbf{1}\mathbf{1}^{\top}}{n}-\boldsymbol{W}^{k-t-1}\right)\right\|_F^2\leq\left(\sum_{t=0}^{k-1}\rho^{k-t-1}\left\|\boldsymbol{X}_t\right\|_F\right)^2$$

Proof.

$$\begin{aligned}\left\|\sum_{t=0}^{k-1}\boldsymbol{X}_t\left(\frac{\mathbf{1}\mathbf{1}^{\top}}{n}-\boldsymbol{W}^{k-t-1}\right)\right\|_F^2 &=\left(\left\|\sum_{t=0}^{k-1}\boldsymbol{X}_t\left(\frac{\mathbf{1}\mathbf{1}^{\top}}{n}-\boldsymbol{W}^{k-t-1}\right)\right\|_F\right)^2\\ &\leq\left(\sum_{t=0}^{k-1}\left\|\boldsymbol{X}_t\left(\frac{\mathbf{1}\mathbf{1}^{\top}}{n}-\boldsymbol{W}^{k-t-1}\right)\right\|_F\right)^2\\ &\leq\left(\sum_{t=0}^{k-1}\left\|\boldsymbol{X}_t\right\|_F\left\|\frac{\mathbf{1}\mathbf{1}^{\top}}{n}-\boldsymbol{W}^{k-t-1}\right\|\right)^2\\ &\leq\left(\sum_{t=0}^{k-1}\rho^{k-t-1}\left\|\boldsymbol{X}_t\right\|_F\right)^2\end{aligned}$$

That completes the proof. □

Lemma F.3. *In any iteration $k\geq 0$, and for any two worker i and j, when $\delta=\frac{1-\eta\rho}{8C_{\alpha}^2\eta\log(16n)+2(1-\eta\rho)}$ we have:*

$$\left\|\boldsymbol{X}_k(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|_{\infty}<\frac{2\alpha_k G_{\infty}C_{\alpha}\eta\log(16n)}{1-\eta\rho}=\theta_k$$

Proof. We use mathematical induction to prove this:

I. When $k=0$, $\left\|\boldsymbol{X}_0(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|_{\infty}=0<\theta_0,\forall i,j$

II. Suppose $\|\boldsymbol{X}_t(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < \theta_t, \forall t \le k, \forall i, j$, we obtain

$$\begin{aligned}
\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty &= \left\|\sum_{t=0}^{k}(-\alpha_t \boldsymbol{G}_t + \boldsymbol{\Omega}_t)\boldsymbol{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty \\
&\le \sum_{t=0}^{k} \|-\alpha_t \boldsymbol{G}_t\|_{1,\infty} \left\|\boldsymbol{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 + \sum_{t=0}^{k} \|\boldsymbol{\Omega}_t\|_{1,\infty} \left\|\boldsymbol{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \\
&\overset{\text{Lemma } F.1}{\le} \sum_{t=0}^{k} \alpha_t G_\infty \left\|\boldsymbol{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 + \frac{4\delta}{1-2\delta} \sum_{t=0}^{k} \theta_t \left\|\boldsymbol{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \\
&\le \alpha_{k+1} G_\infty \sum_{t=0}^{k} \frac{\alpha_{k-t}}{\alpha_{k+1}} \left\|\boldsymbol{W}^{t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 + \frac{4\delta\theta_k}{1-2\delta} \sum_{t=0}^{k} \frac{\theta_t}{\theta_k} \left\|\boldsymbol{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \\
&< \alpha_{k+1} G_\infty C_\alpha \eta \sum_{t=0}^{\infty} \eta^t \left\|\boldsymbol{W}^{t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 + \frac{4\delta C_\alpha \theta_k}{1-2\delta} \sum_{t=0}^{\infty} \eta^t \left\|\boldsymbol{W}^{t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1
\end{aligned}$$

For any $t \ge 0$, on one hand

$$\left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \le \sqrt{n} \left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_2 \le \sqrt{n} \left\|\boldsymbol{W}^t \boldsymbol{e}_i - \frac{\mathbf{1}}{n}\right\| + \sqrt{n} \left\|\boldsymbol{W}^t \boldsymbol{e}_j - \frac{\mathbf{1}}{n}\right\| \le 2\sqrt{n}\rho^t$$

where the last step holds due to the diagonalizability of $\boldsymbol{W}$. On the other hand,

$$\left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \le \mathbf{1}^\top \boldsymbol{W}^t \boldsymbol{e}_i + \mathbf{1}^\top \boldsymbol{W}^t \boldsymbol{e}_i = \mathbf{1}^\top \boldsymbol{e}_i + \mathbf{1}^\top \boldsymbol{e}_j = 2$$

As a result

$$\eta^t \left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \le \min\{2\sqrt{n}(\eta\rho)^t, 2\}$$

Let $T_0 = \left\lceil \frac{-\log(\sqrt{n})}{\log(\eta\rho)} \right\rceil$, so that $\sqrt{n}(\eta\rho)^{T_0} \le 1$, then we have

$$\begin{aligned}
\sum_{t=0}^{\infty} \eta^t \left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 &= \sum_{t=0}^{T_0-1} \eta^t \left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 + \sum_{t=T_0}^{\infty} \eta^t \left\|\boldsymbol{W}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \\
&\le \sum_{t=0}^{T_0-1} 2 + \sum_{t=0}^{\infty} 2\sqrt{n}(\eta\rho)^{t+T_0} \\
&\le 2\left\lceil \frac{-\log(\sqrt{n})}{\log(\eta\rho)} \right\rceil + \sum_{t=0}^{\infty} 2\left(\sqrt{n}(\eta\rho)^{T_0}\right)(\eta\rho)^t \\
&\le \frac{2\log(\sqrt{n})}{1-\eta\rho} + 2 + \frac{2}{1-\eta\rho} \\
&\le \frac{\log(16n)}{1-\eta\rho}
\end{aligned}$$

As a result, we have

$$\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < \frac{\alpha_{k+1} G_\infty C_\alpha \eta \log(16n)}{1-\eta\rho} + \frac{4\delta C_\alpha}{1-2\delta} \cdot \frac{\log(16n)}{1-\eta\rho}\theta_k$$

with $\delta = \frac{1-\eta\rho}{8C_\alpha^2 \eta \log(16n) + 2(1-\eta\rho)}$,

$$\begin{aligned}
\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty &< \frac{\alpha_{k+1} G_\infty C_\alpha \eta \log(16n)}{1-\eta\rho} + \frac{4\delta C_\alpha}{1-2\delta} \cdot \frac{\log(16n)}{1-\eta\rho} \cdot \frac{2\alpha_k G_\infty C_\alpha \eta \log(16n)}{1-\eta\rho} \\
&\le \frac{\alpha_{k+1} G_\infty C_\alpha \eta \log(16n)}{1-\eta\rho} + \frac{4\delta C_\alpha}{1-2\delta} \cdot \frac{\log(16n)}{1-\eta\rho} \cdot \frac{2\alpha_{k+1} C_\alpha \eta G_\infty C_\alpha \eta \log(16n)}{1-\eta\rho}
\end{aligned}$$

$$\leq \frac{2\alpha_{k+1} G_\infty C_\alpha \eta \log(16n)}{1-\eta\rho} = \theta_{k+1}$$

Combining I and II, we complete the proof. □

Lemma F.4. *The running average of the gradient norm has the following bound:*

$$\sum_{k=0}^{K-1} \alpha_k \mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \leq 4(\mathbb{E}f(\mathbf{0}) - \mathbb{E}f^*) + \frac{2\sigma^2 L}{n}\sum_{k=0}^{K-1}\alpha_k^2 + \frac{8\sigma^2 L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3 + \frac{24\varsigma^2 L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3 + \frac{8L^2}{n(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2$$

Proof. Let $\mathbf{1}$ denote a n-dimensional vector with all the entries be 1. And we have

$$\overline{\boldsymbol{X}}_{k+1} = (\boldsymbol{X}_k\boldsymbol{W} - \alpha_k\tilde{\boldsymbol{G}}_k + \boldsymbol{\Omega}_k)\frac{\mathbf{1}}{n} = \overline{\boldsymbol{X}}_k - \alpha_k\overline{\tilde{\boldsymbol{G}}}_k + (\hat{\boldsymbol{X}}_k - \boldsymbol{X}_k)(\boldsymbol{W}-\boldsymbol{I})\frac{\mathbf{1}}{n} = \overline{\boldsymbol{X}}_k - \alpha_k\overline{\tilde{\boldsymbol{G}}}_k$$

And by Taylor Expansion, we have

$$\begin{aligned}
\mathbb{E}f(\overline{\boldsymbol{X}}_{k+1}) &= \mathbb{E}f\left(\frac{(\boldsymbol{X}_k\boldsymbol{W} - \alpha_k\tilde{\boldsymbol{G}}_k + \boldsymbol{\Omega}_k)\mathbf{1}}{n}\right)\\
&= \mathbb{E}f\left(\overline{\boldsymbol{X}}_k - \alpha_k\overline{\tilde{\boldsymbol{G}}}_k\right)\\
&\leq \mathbb{E}f(\overline{\boldsymbol{X}}_k) - \alpha_k\mathbb{E}\langle\nabla f(\overline{\boldsymbol{X}}_k), \overline{\tilde{\boldsymbol{G}}}_k\rangle + \frac{\alpha_k^2 L}{2}\mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}}_k\right\|^2
\end{aligned}$$

And for the last term, we have

$$\begin{aligned}
\mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}}_k\right\|^2 &= \mathbb{E}\left\|\frac{\sum_{i=1}^n \tilde{\boldsymbol{g}}_{k,i}}{n}\right\|^2\\
&= \mathbb{E}\left\|\frac{\sum_{i=1}^n \tilde{\boldsymbol{g}}_{k,i} - \sum_{i=1}^n \boldsymbol{g}_{k,i}}{n} + \frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2\\
&= \mathbb{E}\left\|\frac{\sum_{i=1}^n \tilde{\boldsymbol{g}}_{k,i} - \sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2 + \mathbb{E}\left\|\frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2 + \mathbb{E}\left\langle\frac{\sum_{i=1}^n \tilde{\boldsymbol{g}}_{k,i} - \sum_{i=1}^n \boldsymbol{g}_{k,i}}{n} + \frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\rangle\\
&= \mathbb{E}\left\|\frac{\sum_{i=1}^n \tilde{\boldsymbol{g}}_{k,i} - \sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2 + \mathbb{E}\left\|\frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2\\
&\overset{\text{Assumption 3}}{=} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k,i} - \boldsymbol{g}_{k,i}\right\|^2 + \mathbb{E}\left\|\frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2\\
&\overset{\text{Assumption 3}}{\leq} \frac{\sigma^2}{n} + \mathbb{E}\left\|\frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2
\end{aligned}$$

Putting it back, we obtain

$$\begin{aligned}
\mathbb{E}f(\overline{\boldsymbol{X}}_{k+1}) &\leq \mathbb{E}f(\overline{\boldsymbol{X}}_k) - \alpha_k\mathbb{E}\langle\nabla f(\overline{\boldsymbol{X}}_k), \overline{\tilde{\boldsymbol{G}}}_k\rangle + \frac{\alpha_k^2 L}{2n}\sigma^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}\left\|\frac{\sum_{i=1}^n \boldsymbol{g}_{k,i}}{n}\right\|^2\\
&= \mathbb{E}f(\overline{\boldsymbol{X}}_k) - \frac{\alpha_k - \alpha_k^2 L}{2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 - \frac{\alpha_k}{2}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 + \frac{\alpha_k^2 L}{2n}\sigma^2 + \frac{\alpha_k}{2}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \overline{\boldsymbol{G}}_k\right\|^2
\end{aligned}$$

where the last step comes from $2\langle\boldsymbol{a},\boldsymbol{b}\rangle = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 = \|\boldsymbol{a}-\boldsymbol{b}\|^2$ And

$$\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \overline{\boldsymbol{G}}_k\right\|^2 \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\|\nabla f_i\left(\frac{\sum_{i'=1}^n \boldsymbol{x}_{k,i'}}{n}\right) - \nabla f_i(\boldsymbol{x}_{k,i})\right\|^2$$

$$\overset{\text{Assumption 1}}{\leq} \frac{L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\frac{\sum_{i'=1}^{n}\boldsymbol{x}_{k,i'}}{n}-\boldsymbol{x}_{k,i}\right\|^2$$
$$=\frac{L^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2$$

by Lipschitz assumption, we obtain

$$\frac{\alpha_k-\alpha_k^2L}{2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{\alpha_k}{2}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq\mathbb{E}f(\overline{\boldsymbol{X}}_k)-\mathbb{E}f(\overline{\boldsymbol{X}}_{k+1})+\frac{\alpha_k^2L}{2n}\sigma^2+\frac{\alpha_kL^2}{2n}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2$$

summing over from $k=0$ to $K-1$ on both sides, we have

$$\sum_{k=0}^{K-1}(\alpha_k-\alpha_k^2L)\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq 2(\mathbb{E}f(\overline{\boldsymbol{X}}_0)-\mathbb{E}f(\overline{\boldsymbol{X}}_K))+\frac{\sigma^2L}{n}\sum_{k=0}^{K-1}\alpha_k^2$$
$$+\frac{L^2}{n}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2$$

From Lemma F.5, we have

$$\sum_{k=0}^{K-1}(\alpha_k-\alpha_k^2L)\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2$$
$$\leq 2(\mathbb{E}f(\overline{\boldsymbol{X}}_0)-\mathbb{E}f(\overline{\boldsymbol{X}}_K))+\frac{\sigma^2L}{n}\sum_{k=0}^{K-1}\alpha_k^2+\frac{L^2}{n}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2$$
$$\leq 2(\mathbb{E}f(\overline{\boldsymbol{X}}_0)-\mathbb{E}f(\overline{\boldsymbol{X}}_K))+\frac{\sigma^2L}{n}\sum_{k=0}^{K-1}\alpha_k^2+\frac{4\sigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{12\varsigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{12L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2$$
$$+\frac{4L^2}{n(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2$$

Rearrange the terms, we have

$$\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq 4(\mathbb{E}f(\mathbf{0})-\mathbb{E}f^*)+\frac{2\sigma^2L}{n}\sum_{k=0}^{K-1}\alpha_k^2+\frac{8\sigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{24\varsigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3$$
$$+\frac{8L^2}{n(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2$$

and that completes the proof ☐

Lemma F.5.

$$\frac{L^2}{n}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2\leq\frac{4\sigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{12\varsigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{12L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2$$
$$+\frac{4L^2}{n(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2$$

Proof.

$$\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2$$

$$
\begin{aligned}
&= \sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k \mathbb{E}\left\|\boldsymbol{X}_k\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_i\right)\right\|^2 \\
&= \sum_{k=1}^{K-1}\sum_{i=1}^{n}\alpha_k \mathbb{E}\left\|\left(\boldsymbol{X}_{k-1}\boldsymbol{W}-\alpha\tilde{\boldsymbol{G}}_{k-1}+\boldsymbol{\Omega}_{k-1}\right)\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_i\right)\right\|^2 \\
&\overset{\boldsymbol{x}_{0,i}=\mathbf{0}}{=} \sum_{k=1}^{K-1}\sum_{i=1}^{n}\alpha_k \mathbb{E}\left\|\sum_{t=0}^{k-1}\left(-\alpha_t\tilde{\boldsymbol{G}}_t+\boldsymbol{\Omega}_t\right)\left(\frac{\mathbf{1}}{n}-\boldsymbol{W}^{k-t-1}\boldsymbol{e}_i\right)\right\|^2 \\
&\le 2\sum_{k=1}^{K-1}\alpha_k\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{t=0}^{k-1}\alpha_t\tilde{\boldsymbol{G}}_t\left(\frac{\mathbf{1}}{n}-\boldsymbol{W}^{k-t-1}\boldsymbol{e}_i\right)\right\|^2+2\sum_{k=1}^{K-1}\alpha_k\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{t=0}^{k-1}\boldsymbol{\Omega}_t\left(\frac{\mathbf{1}}{n}-\boldsymbol{W}^{k-t-1}\boldsymbol{e}_i\right)\right\|^2 \\
&= 2\sum_{k=1}^{K-1}\alpha_k\mathbb{E}\left\|\sum_{t=0}^{k-1}\alpha_t\tilde{\boldsymbol{G}}_t\left(\frac{\mathbf{1}\mathbf{1}^\top}{n}-\boldsymbol{W}^{k-t-1}\right)\right\|_F^2+2\sum_{k=1}^{K-1}\mathbb{E}\left\|\sum_{t=0}^{k-1}\boldsymbol{\Omega}_t\left(\frac{\mathbf{1}\mathbf{1}^\top}{n}-\boldsymbol{W}^{k-t-1}\right)\right\|_F^2 \\
&\overset{Lemma\ F.2}{\le} 2\sum_{k=1}^{K-1}\alpha_k\left(\sum_{t=0}^{k-1}\rho^{k-t-1}\alpha_t\mathbb{E}\left\|\tilde{\boldsymbol{G}}_t\right\|_F\right)^2+2\sum_{k=1}^{K-1}\alpha_k\left(\sum_{t=0}^{k-1}\rho^{k-t-1}\mathbb{E}\left\|\boldsymbol{\Omega}_t\right\|_F\right)^2 \\
&\overset{Lemma\ F.7}{\le} \frac{2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\right\|_F^2+\frac{2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2 \\
&\overset{Lemma\ F.6}{\le} \frac{2}{(1-\rho)^2}\left(n\sigma^2\sum_{k=0}^{K-1}\alpha_k^3+3L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k^3\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2+3n\varsigma^2\sum_{k=0}^{K-1}\alpha_k^3+3n\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\right) \\
&\quad+\frac{2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

Rearrange the terms, we have

$$
\begin{aligned}
\sum_{k=0}^{K-1}\alpha_k\left(1-\frac{6\alpha_k^2L^2}{(1-\rho)^2}\right)\sum_{i=1}^{n}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2 \le& \frac{2n\sigma^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{6n\varsigma^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{6n}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \\
&+\frac{2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

Let $1-\frac{6\alpha_k^2L^2}{(1-\rho)^2}\ge\frac{1}{2}$, we have

$$
\begin{aligned}
\frac{L^2}{n}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2 \le& \frac{4\sigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{12\varsigma^2L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3+\frac{12L^2}{(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \\
&+\frac{4L^2}{n(1-\rho)^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

That completes the proof. □

Lemma F.6.

$$
\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\right\|_F^2 \le n\sigma^2\sum_{k=0}^{K-1}\alpha_k^3+3L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k^3\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2+3n\varsigma^2\sum_{k=0}^{K-1}\alpha_k^3+3n\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2
$$

Proof. From the property of Frobenius norm, we have

$$
\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\right\|_F^2=\sum_{i=1}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k,i}\right\|^2
$$

Since

$$\begin{aligned}\mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k,i}\right\|^2 =&\mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k,i}-\boldsymbol{g}_{k,i}\right\|^2+\mathbb{E}\left\|\boldsymbol{g}_{k,i}\right\|^2\\ =&\sigma^2+3\mathbb{E}\left\|\nabla f_i(\boldsymbol{x}_{k,i})-\nabla f_i(\overline{\boldsymbol{X}}_k)\right\|^2+3\mathbb{E}\left\|\nabla f_i(\overline{\boldsymbol{X}}_k)-\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2+3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\\ \le&\sigma^2+3L^2\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2+3\varsigma^2+3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\end{aligned}$$

Summing from $k=0$ to $K-1$, we obtain

$$\begin{aligned}&\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\right\|_F^2\\ =&\sum_{k=0}^{K-1}\alpha_k^3\sum_{i=1}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k,i}\right\|^2\\ \le&\sum_{k=0}^{K-1}\alpha_k^3\sum_{i=1}^{n}\sigma^2+3L^2\sum_{k=0}^{K-1}\alpha_k^3\sum_{i=1}^{n}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2+3\sum_{k=0}^{K-1}\alpha_k^3\sum_{i=1}^{n}\varsigma^2+3\sum_{k=0}^{K-1}\alpha_k^3\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\\ =&n\sigma^2\sum_{k=0}^{K-1}\alpha_k^3+3L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\alpha_k^3\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2+3n\varsigma^2\sum_{k=0}^{K-1}\alpha_k^3+3n\sum_{k=0}^{K-1}\alpha_k^3\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\end{aligned}$$

That completes the proof. □

Lemma F.7. *Given $0\le\rho<1$ and T, a positive integer. Also given non-negative sequences $\{a_t\}_{t=1}^{\infty}$ and $\{b_t\}_{t=1}^{\infty}$ with $\{a_t\}_{t=1}^{\infty}$ being non-increasing, the following inequalities holds:*

$$\begin{aligned}\sum_{t=1}^{k}a_t\left(\sum_{s=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s\right)\le&\frac{T}{1-\rho}\sum_{s=1}^{k}a_sb_s\\ \sum_{t=1}^{k}a_t\left(\sum_{s=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s\right)^2\le&\frac{T^2}{(1-\rho)^2}\sum_{s=1}^{k}a_sb_s^2\end{aligned}$$

Proof. Firstly,

$$S_k=\sum_{t=1}^{k}a_t\left(\sum_{s=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s\right)=\sum_{s=1}^{k}\sum_{t=s}^{k}\alpha_t\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s\le\sum_{s=1}^{k}a_sb_s\sum_{t=0}^{T-1}\sum_{m=0}^{\infty}\rho^m\le\frac{T}{1-\rho}\sum_{s=1}^{k}a_sb_s$$

further we have

$$\begin{aligned}&\sum_{t=1}^{k}a_t\left(\sum_{s=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s\right)^2=\sum_{t=1}^{k}a_t\sum_{s=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s\sum_{r=1}^{t}\rho^{-\lfloor\frac{t-r}{T}\rfloor}b_r=\sum_{t=1}^{k}a_t\sum_{s=1}^{t}\sum_{r=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor+\lfloor\frac{t-r}{T}\rfloor}b_sb_r\\ \le&\sum_{t=1}^{k}a_t\sum_{s=1}^{t}\sum_{r=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor+\lfloor\frac{t-r}{T}\rfloor}\frac{b_s^2+b_r^2}{2}=\sum_{t=1}^{k}a_t\sum_{s=1}^{t}\sum_{r=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor+\lfloor\frac{t-r}{T}\rfloor}b_s^2\\ \le&\sum_{t=1}^{k}a_t\sum_{s=1}^{t}b_s^2\rho^{-\lfloor\frac{t-s}{T}\rfloor}\sum_{r=1}^{t}\rho^{-\lfloor\frac{t-r}{T}\rfloor}\le\sum_{t=1}^{k}a_t\sum_{s=1}^{t}b_s^2\rho^{-\lfloor\frac{t-s}{T}\rfloor}\sum_{r=0}^{T-1}\sum_{m=0}^{\infty}\rho^m\\ \le&\frac{T}{1-\rho}\sum_{t=1}^{k}a_t\sum_{s=1}^{t}\rho^{-\lfloor\frac{t-s}{T}\rfloor}b_s^2\overset{\text{Using }S_k}{\le}\frac{T^2}{(1-\rho)^2}\sum_{s=1}^{k}a_sb_s^2\end{aligned}$$

That completes the proof. □

F.5. Proof to Theorem 3.

Proof. Let $\overline{\rho}$ denote the spectral gap of matrix $\overline{\boldsymbol{W}}$, it is straightforward to know that $\overline{\rho} = \gamma\rho + (1-\gamma)$. we first use mathematical induction to prove at iteration $\forall k \le K$, for any worker i and j, with probability $(1-\epsilon)^k$

$$\|\boldsymbol{X}_k(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < \theta = \frac{2\alpha \log(16n) G_\infty}{\gamma(1-\rho)}$$

where $\gamma = \frac{2}{1-\rho+\frac{16\delta^2}{(1-2\delta)^2}\cdot\frac{32\log(4n)}{1-\rho}\log\left(\frac{1}{\epsilon}\right)}$.

I. When $k=0$, $\|\boldsymbol{X}_0(\boldsymbol{e}_i-\boldsymbol{e}_j)\|_\infty = 0 < \theta$

II. Suppose $\|\boldsymbol{X}_t(\boldsymbol{e}_i-\boldsymbol{e}_j)\|_\infty < \theta$ holds for $\forall t \le k$, then for $k+1$ we have

$$\begin{aligned}
\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty &= \left\|\left(\boldsymbol{X}_k\overline{\boldsymbol{W}} - \alpha\tilde{\boldsymbol{G}}_k + \gamma\boldsymbol{\Omega}_k\right)(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty \\
&\overset{\boldsymbol{X}_0=0}{=} \left\|\sum_{t=0}^{k}\left(-\alpha\tilde{\boldsymbol{G}}_t + \gamma\boldsymbol{\Omega}_t\right)\overline{W}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty \\
&\le \left\|\sum_{t=0}^{k}\alpha\tilde{\boldsymbol{G}}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty + \left\|\sum_{t=0}^{k}\gamma\boldsymbol{\Omega}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty
\end{aligned}$$

We bound these two terms seperately. First from Lemma F.3 we know that

$$\sum_{t=0}^{\infty}\left\|\overline{\boldsymbol{W}}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 < \frac{\log(16n)}{1-\overline{\rho}} = \frac{\log(16n)}{\gamma(1-\rho)} \tag{5}$$

then we have for the first term,

$$\begin{aligned}
\left\|\sum_{t=0}^{k}\alpha\tilde{\boldsymbol{G}}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty &\le \sum_{t=0}^{k}\left\|\alpha\tilde{\boldsymbol{G}}_t\right\|_{1,\infty}\left\|\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \\
&\le \alpha G_\infty\sum_{t=0}^{\infty}\left\|\overline{\boldsymbol{W}}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_1 \\
&< \frac{\alpha\log(16n)G_\infty}{\gamma(1-\rho)}
\end{aligned}$$

Next, we bound the second term. Suppose the infinity norm of the term $\sum_{t=0}^{k}\gamma\boldsymbol{\Omega}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)$ is taken at coordinate h, then we have

$$\begin{aligned}
\left\|\sum_{t=0}^{k}\gamma\boldsymbol{\Omega}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty &= \gamma\left|\boldsymbol{e}_h^\top\left(\sum_{t=0}^{k}\boldsymbol{\Omega}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)\right| \\
&= \gamma\left|\sum_{t=0}^{k}\boldsymbol{e}_h^\top\left(\boldsymbol{\Omega}_t\overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)\right|
\end{aligned}$$

Let

$$u_t = \sum_{m=0}^{t}\boldsymbol{e}_h^\top\left(\boldsymbol{\Omega}_{k-m}\overline{\boldsymbol{W}}^m(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)$$

from the induction hypothesis we know that $\{u_t\}_{t\le k}$ is a martingale sequence. Note that,

$$\begin{aligned}
|u_t - u_{t-1}| &= \left|\boldsymbol{e}_h^\top\left(\boldsymbol{\Omega}_{k-t}\overline{\boldsymbol{W}}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)\right| \\
&\le \left\|\boldsymbol{\Omega}_{k-t}\overline{\boldsymbol{W}}^t(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\overset{Equation~5}{\leq} \|\mathbf{\Omega}_{k-t}\|_{1,\infty} \min\{2\sqrt{n}\overline{\rho}^t, 2\} \\
&\leq 2\delta B_\theta \min\{2\sqrt{n}\overline{\rho}^t, 2\}
\end{aligned}
$$

where $B_\theta = \frac{2}{1-2\delta}\theta$, then by using Azuma's inequality we obtain

$$
\begin{aligned}
\mathbb{P}\left[\left|\sum_{t=0}^{k} \boldsymbol{e}_h^\top \left(\mathbf{\Omega}_t \overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)\right| > a\right] &\leq \exp\left(-\frac{a^2}{8\delta^2 B_\theta^2 \sum_{t=0}^{k} \min\{2\sqrt{n}\overline{\rho}^t, 2\}^2}\right) \\
&\leq \exp\left(-\frac{a^2}{32\delta^2 B_\theta^2 \sum_{t=0}^{\infty} \min\{n\overline{\rho}^{2t}, 1\}}\right)
\end{aligned}
$$

Here we use the induction hypothesis. Similar as before, Let $T_0 = \left\lceil \frac{-\log(n)}{2\log(\overline{\rho})} \right\rceil$, so that $n\overline{\rho}^{2T_0} \leq 1$, then we have

$$
\begin{aligned}
\sum_{t=0}^{\infty} \min\{n\overline{\rho}^{2t}, 1\} &= \sum_{t=0}^{T_0-1} \min\{n\overline{\rho}^{2t}, 1\} + \sum_{t=T_0}^{\infty} \min\{n\overline{\rho}^{2t}, 1\} \\
&< \sum_{t=0}^{T_0-1} 1 + \sum_{t=0}^{\infty} n\overline{\rho}^{2t+2T_0} \\
&\leq \left\lceil \frac{-\log(n)}{2\log(\overline{\rho})} \right\rceil + \sum_{t=0}^{\infty} \left(n\overline{\rho}^{2T_0}\right)\overline{\rho}^{2t} \\
&\leq \frac{\log(n)}{1-\overline{\rho}^2} + 1 + \frac{1}{1-\overline{\rho}^2} \\
&\leq \frac{\log(4n)}{1-\overline{\rho}^2} \\
&= \frac{\log(4n)}{\gamma(1-\rho)(2-\gamma(1-\rho))}
\end{aligned}
$$

Putting it back, we obtain

$$
\mathbb{P}\left[\left|\sum_{t=0}^{k} \boldsymbol{e}_h^\top \left(\mathbf{\Omega}_t \overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)\right| > a\right] \leq \exp\left(-\frac{a^2\gamma(1-\rho)(2-\gamma(1-\rho))}{32\delta^2 B_\theta^2 \log(4n)}\right)
$$

In other words, with probability $1-\epsilon$,

$$
\left\|\sum_{t=0}^{k} \gamma \mathbf{\Omega}_t \overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty = \gamma\left|\sum_{t=0}^{k} \boldsymbol{e}_h^\top \left(\mathbf{\Omega}_t \overline{\boldsymbol{W}}^{k-t}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right)\right| \leq \delta B_\theta \sqrt{\frac{32\log(4n)\gamma}{(1-\rho)(2-\gamma(1-\rho))}\log\left(\frac{1}{\epsilon}\right)}
$$

Combine them together, we obtain

$$
\begin{aligned}
\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty &< \frac{\alpha\log(16n)G_\infty}{\gamma(1-\rho)} + \delta B_\theta \sqrt{\frac{32\log(4n)\gamma}{(1-\rho)(2-\gamma(1-\rho))}\log\left(\frac{1}{\epsilon}\right)} \\
&< \frac{\alpha\log(16n)G_\infty}{\gamma(1-\rho)} + \frac{2\delta}{1-2\delta}\theta\sqrt{\frac{32\log(4n)\gamma}{(1-\rho)(2-\gamma(1-\rho))}\log\left(\frac{1}{\epsilon}\right)}
\end{aligned}
$$

Let $\gamma = \frac{2}{1-\rho+\frac{16\delta^2}{(1-2\delta)^2}\cdot\frac{32\log(4n)}{1-\rho}\log\left(\frac{1}{\epsilon}\right)}$

$$
\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < \frac{\alpha\log(16n)G_\infty}{\gamma(1-\rho)} + \frac{1}{2}\theta \leq \theta
$$

Combining I and II, we complete the proof.

We proceed to obtain the convergence rate. From Theorem 2 we have with $\alpha_k = \alpha$

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \leq \frac{4(\mathbb{E}f(\mathbf{0}) - \mathbb{E}f^*)}{\alpha K} + \frac{2\alpha\sigma^2 L}{n} + \frac{8\alpha^2\sigma^2L^2}{(1-\overline{\rho})^2} + \frac{24\alpha^2\varsigma^2L^2}{(1-\overline{\rho})^2} + \frac{8\alpha L^2}{n(1-\overline{\rho})^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\gamma\boldsymbol{\Omega}_k\right\|_F^2$$

Note that with probability $(1-\epsilon)^K$

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\gamma\boldsymbol{\Omega}_k\right\|_F^2 = \gamma^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{k,j} - \boldsymbol{x}_{k,j}) - (\hat{\boldsymbol{x}}_{k,i} - \boldsymbol{x}_{k,i})\right)\boldsymbol{W}_{ji}\right\|^2 \overset{\text{Lemma } F.1}{\leq} \frac{16\delta^2\gamma^2}{(1-2\delta)^2}\theta^2 dnK$$

Fit in $\theta = \frac{2\alpha\log(16n)G_\infty}{\gamma(1-\rho)}$, we obtain

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\gamma\boldsymbol{\Omega}_k\right\|_F^2 \leq \frac{64\alpha^2\delta^2\log^2(16n)G_\infty^2}{(1-2\delta)^2(1-\rho)^2}dnK$$

Let $\mathcal{E}$ denote the event that the bound θ holds for all $0 \leq t \leq T-1$, then,

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 =& \left[\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2|\mathcal{E}\right]\mathbb{P}(\mathcal{E}) + \left[\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2|\neg\mathcal{E}\right]\mathbb{P}(\neg\mathcal{E}) \\
\leq& \frac{4(f(\mathbf{0}) - f^*)}{\alpha K} + \frac{2\alpha L}{n}\sigma^2 + \frac{8\alpha^2L^2\left(\sigma^2 + 3\varsigma^2\right)}{(1-\overline{\rho})^2} + \frac{8L^2}{nK(1-\overline{\rho})^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\gamma\boldsymbol{\Omega}_k\right\|_F^2 \\
&+ G_\infty^2 d\left(1 - (1-\epsilon)^K\right) \\
\leq& \frac{4(f(\mathbf{0}) - f^*)}{\alpha K} + \frac{2\alpha L}{n}\sigma^2 + \frac{8\alpha^2L^2\left(\sigma^2 + 3\varsigma^2\right)}{\gamma^2(1-\rho)^2} + \frac{512\alpha^2\delta^2L^2\log^2(16n)G_\infty^2 d}{\gamma^2(1-\rho)^4(1-2\delta)^2} \\
&+ G_\infty^2 d\left(1 - (1-\epsilon)^K\right)
\end{aligned}$$

Assign $\epsilon = \frac{1}{K^2}$ and set $\alpha = \frac{1}{\varsigma^{\frac{2}{3}}K^{\frac{1}{3}} + \sigma\sqrt{\frac{K}{n}} + 2L}$, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \lesssim \frac{\sigma}{\sqrt{nK}} + \frac{1}{K} + \frac{\varsigma^{\frac{2}{3}}\delta^4\log^2(n)\log^2(K)}{K^{\frac{2}{3}}(1-2\delta)^4} + \frac{\sigma^2 n\delta^4\log^2(n)\log^2(K)}{(\sigma^2K+n)(1-2\delta)^4} + \frac{n\delta^6\log^4(n)\log^2(K)}{(\sigma^2K+n)(1-2\delta)^6}$$

That completes the proof ◻

G. Moniqua on D^2 (Proof to Theorem 4)

G.1. Setting

We first show the pseudo code in Algorithm 1.

D^2 makes the following assumptions (1-4), and we add the additional assumption (5):

1. **Lipschitzian Gradient**: All the function f_i have L-Lipschitzian gradients.
2. **Communication Matrix**: Communication matrix $\boldsymbol{W}$ is a symmetric doubly stochastic matrix. Let the eigenvalues of $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ be $\lambda_1 \geq \cdots \geq \lambda_n$. We assume $\lambda_2 < 1, \lambda_n > -\frac{1}{3}$.
3. **Bounded Variance**:
$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left\|\nabla\tilde{f}_i(\boldsymbol{x};\xi_i) - \nabla f_i(\boldsymbol{x})\right\|^2 \leq \sigma^2, \forall i$$
where $\nabla\tilde{f}_i(\boldsymbol{x};\xi_i)$ denotes gradient sample on worker i computed via data sample ξ_i.
4. **Initialization**: All the models are initialized by the same parameters: $\boldsymbol{x}_{0,i} = \boldsymbol{x}_0, \forall i$ and with out the loss of generality $\boldsymbol{x}_0 = 0$.
5. **Gradient magnitude**: The norm of a sampled gradient is bounded by $\left\|\tilde{\boldsymbol{g}}_{k,i}\right\|_\infty \leq G_\infty$ for some constant G_∞.

Algorithm 1 Moniqua with Variance Reduction on worker i

Require: initial point $\boldsymbol{x}_{0,i} = \boldsymbol{x}_0$, step size α, the discrepency bound B_θ, communication matrix $\boldsymbol{W}$, number of iterations K, neighbor list of worker i: $\mathcal{N}_i$, quantizer Q_δ

for $k = 0, 1, 2, \cdots, K-1$ **do**
  Randomly sample data $\xi_{k,i}$ from local memory
  Compute a local stochastic gradient based on $\xi_{k,i}$ and current weight $\boldsymbol{x}_{k,i}$: $\tilde{\boldsymbol{g}}_{k,i}$
  if $k = 0$ **then**
    Update local weight: $\boldsymbol{x}_{k+\frac{1}{2},i} \leftarrow \boldsymbol{x}_{k,i} - \alpha\tilde{\boldsymbol{g}}_{k,i}$
  else
    Update local weight: $\boldsymbol{x}_{k+\frac{1}{2},i} \leftarrow 2\boldsymbol{x}_{k,i} - \boldsymbol{x}_{k-1,i} - \alpha\tilde{\boldsymbol{g}}_{k,i} + \alpha\tilde{\boldsymbol{g}}_{k-1,i}$
  end if
  Send modulo-ed model to neighbors: $\boldsymbol{q}_{k+\frac{1}{2},i} \leftarrow \mathcal{Q}_\delta\left(\frac{\boldsymbol{x}_{k+\frac{1}{2},i}}{B_\theta} \bmod 1\right)$
  Compute local biased term $\hat{\boldsymbol{x}}_{k+\frac{1}{2},i}$ as:

$$\hat{\boldsymbol{x}}_{k+\frac{1}{2},i} = \boldsymbol{q}_{k+\frac{1}{2},i}B_\theta - \boldsymbol{x}_{k+\frac{1}{2},i} \bmod B_\theta + \boldsymbol{x}_{k+\frac{1}{2},i}$$

  Recover model received from worker j as:

$$\hat{\boldsymbol{x}}_{k+\frac{1}{2},j} = (\boldsymbol{q}_{k+\frac{1}{2},j}B_\theta - \boldsymbol{x}_{k+\frac{1}{2},j}) \bmod B_\theta + \boldsymbol{x}_{k+\frac{1}{2},i}$$

  Average with neighboring workers: $\boldsymbol{x}_{k+1,i} \leftarrow \boldsymbol{x}_{k+\frac{1}{2},i} + \sum_{j\in\mathcal{N}_i}(\hat{\boldsymbol{x}}_{k+\frac{1}{2},j} - \hat{\boldsymbol{x}}_{k+\frac{1}{2},i})\boldsymbol{W}_{ji}$
end for
return $\overline{\boldsymbol{X}}_K = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{K,i}$

G.2. Proof to Theorem 4

Proof. From a local view, define $\boldsymbol{x}_{-1} = \tilde{\boldsymbol{g}}_{-1} = 0$, the update rule of Moniqua on D^2 on worker i in iteration k can be written as

$$\boldsymbol{x}_{k+\frac{1}{2},i} = 2\boldsymbol{x}_{k,i} - \boldsymbol{x}_{k-1,i} - \alpha\tilde{\boldsymbol{g}}_{k,i} + \alpha\tilde{\boldsymbol{g}}_{k-1,i}$$
$$\boldsymbol{x}_{k+1,i} = \sum_{j=1}^{n}\boldsymbol{x}_{k+\frac{1}{2},j}\boldsymbol{W}_{ji} + \sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{k+\frac{1}{2},j} - \boldsymbol{x}_{k+\frac{1}{2},j}) - (\hat{\boldsymbol{x}}_{k+\frac{1}{2},i} - \boldsymbol{x}_{k+\frac{1}{2},i})\right)\boldsymbol{W}_{ji}$$

For a more compact expression,

$$\boldsymbol{X}_{k+\frac{1}{2}} = 2\boldsymbol{X}_k - \boldsymbol{X}_{k-1} - \alpha\tilde{\boldsymbol{G}}_k + \alpha\tilde{\boldsymbol{G}}_{k-1}$$
$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_{k+\frac{1}{2}}\boldsymbol{W} + (\hat{\boldsymbol{X}}_{k+\frac{1}{2}} - \boldsymbol{X}_{k+\frac{1}{2}})(\boldsymbol{W} - \boldsymbol{I})$$

Define

$$\boldsymbol{\Omega}_k = (\hat{\boldsymbol{X}}_{k+\frac{1}{2}} - \boldsymbol{X}_{k+\frac{1}{2}})(\boldsymbol{W} - \boldsymbol{I})$$

Since $\boldsymbol{W}$ is symmetric, it can be diagonalized as $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$, where the i-th column of $\boldsymbol{P}$ and $\boldsymbol{\Lambda}$ are $\boldsymbol{W}$'s i-th eigenvector and eigenvalue, respectively. And we obtain

$$\boldsymbol{X}_{k+1} = 2\boldsymbol{X}_k\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top - \boldsymbol{X}_{k-1}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top - \alpha\tilde{\boldsymbol{G}}_k\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top + \alpha\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top + \boldsymbol{\Omega}_k$$

and

$$\boldsymbol{X}_{k+1}\boldsymbol{P} = 2\boldsymbol{X}_k\boldsymbol{P}\boldsymbol{\Lambda} - \boldsymbol{X}_{k-1}\boldsymbol{P}\boldsymbol{\Lambda} - \alpha\tilde{\boldsymbol{G}}_k\boldsymbol{P}\boldsymbol{\Lambda} + \alpha\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{P}\boldsymbol{\Lambda} + \boldsymbol{\Omega}_k\boldsymbol{P}$$

Denote $\boldsymbol{Y}_k = \boldsymbol{X}_k\boldsymbol{P}$, $\boldsymbol{H}(\boldsymbol{X}_k;\xi_k) = \tilde{\boldsymbol{G}}_k\boldsymbol{P}$, and denote $\boldsymbol{y}_{k,i}$, $\boldsymbol{h}_{k,i}$ and $\boldsymbol{r}_{k,i}$ as the i-th column of $\boldsymbol{Y}_k$, $\boldsymbol{H}_k$ and $\boldsymbol{\Omega}_k\boldsymbol{P}$, respectively. Then we have

$$\boldsymbol{y}_{k+1,i} = \lambda_i(2\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i} - \alpha\boldsymbol{h}_{k,i} + \alpha\boldsymbol{h}_{k-1,i}) + \boldsymbol{r}_{k,i}$$

From Lemma G.5 (Constants C_1, C_2, C_3 andn C_4 are defined in the Lemma G.1. Constants D_1 and D_2 are defined in Lemma G.5) we get

$$\left(1-\frac{3C_1\alpha^2L^2}{C_4}\right)\mathbb{E}\left\|\nabla f(\mathbf{0})\right\|+\left(1-\alpha L-3\frac{C_2}{C_4}\alpha^4L^4\right)\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2$$
$$\le\frac{2(f(0)-f^*)}{\alpha K}+\frac{\alpha L}{n}\sigma^2+\frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2)}{C_4K}+6\frac{C_2}{C_4}\alpha^2\sigma^2L^2+3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4+\frac{C_3L^2}{C_4}\left(\frac{6D_1n+8}{6D_2n+1}\right)^2\alpha^2G_\infty^2d$$

Let $\alpha=\frac{1}{\sigma\sqrt{K/n}+2L}$, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\\
\le&\frac{2(f(\mathbf{0})-f^*)}{\alpha K}+\frac{\alpha L}{n}\sigma^2+\frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2)}{C_4K}+6\frac{C_2}{C_4}\alpha^2\sigma^2L^2+3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4+\left(\frac{6D_1n+8}{6D_2n+1}\right)^2\frac{C_3L^2}{C_4}G_\infty^2d\alpha^2\\
\le&\frac{4(f(\mathbf{0})-f^*)L}{K}+\frac{2\sigma(f(\mathbf{0})-f^*+L/2)}{\sqrt{nK}}+\frac{3C_1L^2(\sigma^2+\varsigma_0^2)n}{C_4(\sigma^2K^2+4nL^2K)}+\frac{6C_2L^2\sigma^2n}{C_4(\sigma^2K+4nL^2)}\\
&+\frac{3C_2n\sigma^2L^2}{C_4(\sigma^4K^2+16n^2L^4)}+\left(\frac{6D_1n+8}{6D_2n+1}\right)^2\frac{C_3G_\infty^2dL^2n}{C_4(\sigma^2K+4nL^2)}\\
\lesssim&\frac{1}{K}+\frac{\sigma}{\sqrt{nK}}+\frac{(\sigma^2+\varsigma_0^2)n}{\sigma^2K^2+nK}+\frac{\sigma^2n}{\sigma^2K+n}+\frac{\sigma^2n}{\sigma^4K^2+n^2}+\frac{G_\infty^2dn}{\sigma^2K+n}\\
\lesssim&\frac{1}{K}+\frac{\sigma}{\sqrt{nK}}+\frac{\sigma^2n}{\sigma^2K+n}+\frac{G_\infty^2dn}{\sigma^2K+n}
\end{aligned}$$

That completes the proof. □

G.3. Lemma for D^2

Lemma G.1. *Define*

$$\begin{aligned}
D_1&=\max\left\{|v_n|+\frac{2|\lambda_n|}{1-|v_n|},\sqrt{\frac{\lambda_2}{1-\lambda_2}}+\frac{2\lambda_2}{1-\lambda_2}\right\}\\
D_2&=\max\left\{\frac{2}{1-|v_n|},\frac{2}{\sqrt{1-\lambda_2}}\right\}\\
v_n&=\lambda_n-\sqrt{\lambda_n^2-\lambda_n}
\end{aligned}$$

Let $\delta=\frac{1}{12nD_2+2}$, and we have for $\forall i,j$

$$\left\|\boldsymbol{x}_{k+\frac{1}{2}}(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|_\infty<\theta=(6D_1n+8)\alpha G_\infty$$

Proof. We use mathematical induction to prove this:

I. When $k=0$,

$$\left\|\boldsymbol{X}_{0+\frac{1}{2}}(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|_\infty=\left\|-\alpha\tilde{\boldsymbol{G}}_0(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|_\infty\le\alpha\left\|\tilde{\boldsymbol{G}}_0\right\|_{1,\infty}\left\|\boldsymbol{e}_i-\boldsymbol{e}_j\right\|_1<2\alpha G_\infty\le(6D_1n+8)\alpha G_\infty$$

II. Suppose for $k\ge0$, $\forall t\le k$, we have $\left\|\boldsymbol{X}_{t+\frac{1}{2}}(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|<(6D_1n+8)\alpha G_\infty$, then for $\forall i,j$

$$\begin{aligned}
&\left\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i-\boldsymbol{e}_j)\right\|_\infty\\
\le&\left\|\boldsymbol{X}_{k+1}\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_i\right)\right\|_\infty+\left\|X_{k+1}\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_j\right)\right\|_\infty
\end{aligned}$$

$$= \left\| \boldsymbol{X}_{k+1}\boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{e}_i - \boldsymbol{X}_{k+1}\boldsymbol{P} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \boldsymbol{P}^\top \boldsymbol{e}_i \right\|_\infty + \left\| \boldsymbol{X}_{k+1}\boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{e}_j - \boldsymbol{X}_{k+1}\boldsymbol{P} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \boldsymbol{P}^\top \boldsymbol{e}_j \right\|_\infty$$

$$\leq \left\| \boldsymbol{X}_{k+1}\boldsymbol{P} \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\|_{1,\infty} \|\boldsymbol{P}^\top \boldsymbol{e}_i\|_1 + \left\| \boldsymbol{X}_{k+1}\boldsymbol{P} \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\|_{1,\infty} \|\boldsymbol{P}^\top \boldsymbol{e}_j\|_1$$

$$\leq 2\sqrt{n} \left\| \boldsymbol{X}_{k+1}\boldsymbol{P} \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\|_{1,\infty}$$

From the update rule, we have

$$\boldsymbol{y}_{k+1,i} = \lambda_i(2\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i} - \alpha \boldsymbol{h}_{k,i} + \alpha \boldsymbol{h}_{k-1,i}) + \boldsymbol{r}_{k,i} = \lambda_i(2\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i}) + \lambda_i \boldsymbol{\beta}_{k,i} + \boldsymbol{r}_{k,i}$$

where$\boldsymbol{\beta}_{k,i} = -\alpha \boldsymbol{h}_{k,i} + \alpha \boldsymbol{h}_{k-1,i}$, for all $\boldsymbol{y}_i$ with $-\frac{1}{3} < \lambda_i < 0$, from Lemma G.3 we have

$$\boldsymbol{y}_{k+1,i} = \boldsymbol{y}_{1,i} \left(\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i} \right) + \sum_{s=1}^{k} (\lambda_i \boldsymbol{\beta}_{s,i} + \boldsymbol{r}_{s,i}) \frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}$$

where $u_i = \lambda_i + \sqrt{\lambda_i^2 - \lambda_i}$ and $v_i = \lambda_i - \sqrt{\lambda_i^2 - \lambda_i}$, we obtain

$$\left\|\boldsymbol{y}_{k+1,i}\right\|_\infty \leq \left\|\boldsymbol{y}_{1,i}\right\|_\infty \left| \frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i} \right| + |\lambda_i| \sum_{s=1}^{k} \left\|\boldsymbol{\beta}_{s,i}\right\|_\infty \left| \frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i} \right| + \sum_{s=1}^{k} \|\boldsymbol{r}_{s,i}\|_\infty \left| \frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i} \right|$$

Since

$$\left| \frac{u_i^{n+1} - v_i^{n+1}}{u_i - v_i} \right| \leq |v_i|^n \left| \frac{u_i \left(\frac{u_i}{v_i}\right)^n - v_i}{u_i - v_i} \right| \leq |v_i|^n$$

We obtain

$$\left\|\boldsymbol{y}_{k+1,i}\right\|_\infty \leq \left\|\boldsymbol{y}_{1,i}\right\|_\infty |v_i|^k + |\lambda_i| \sum_{s=1}^{k} \left\|\boldsymbol{\beta}_{s,i}\right\|_\infty |v_i|^{k-s} + \sum_{s=1}^{k} \|\boldsymbol{r}_{s,i}\|_\infty |v_i|^{k-s}$$

For $\beta_{s,i}$, we have

$$\begin{aligned} \left\|\boldsymbol{\beta}_{s,i}\right\|_\infty &= \|-\alpha \boldsymbol{h}_{k,i} + \alpha \boldsymbol{h}_{k-1,i}\|_\infty \leq 2\alpha(\|\boldsymbol{h}_{k,i}\|_\infty + \|\boldsymbol{h}_{k-1,i}\|_\infty) \\ &\leq 2\alpha(\|\boldsymbol{G}_k\|_{1,\infty}\|\boldsymbol{P}\boldsymbol{e}_i\|_1 + \|\boldsymbol{G}_{k-1}\|_{1,\infty}\|\boldsymbol{P}\boldsymbol{e}_i\|_1) \\ &\leq 2\alpha\sqrt{n}G_\infty \end{aligned}$$

For $\boldsymbol{r}_{s,i}$, we have

$$\|\boldsymbol{r}_{k,i}\|_\infty = \|\boldsymbol{\Omega}_k \boldsymbol{P}\boldsymbol{e}_i\|_\infty \leq \|\boldsymbol{\Omega}_k\|_{1,\infty} \|\boldsymbol{P}\boldsymbol{e}_i\|_1 \leq 2\sqrt{n}\delta B_\theta$$

when $\lambda_i < 0$, we have

$$\left\|\boldsymbol{y}_{k+1,i}\right\|_\infty \leq \left\|\boldsymbol{y}_{1,i}\right\|_\infty |v_i|^k + |\lambda_i| \sum_{s=1}^{k} \left\|\boldsymbol{\beta}_{s,i}\right\|_\infty |v_i|^{k-s} + \sum_{s=1}^{k} \|\boldsymbol{r}_{s,i}\|_\infty |v_i|^{k-s}$$

$$
\begin{aligned}
&\leq \left\|\boldsymbol{y}_{1,i}\right\|_\infty |v_n|^k + |\lambda_n| \sum_{s=1}^{k} \left\|\boldsymbol{\beta}_{s,i}\right\|_\infty |v_n|^{k-s} + \sum_{s=1}^{k} \|\boldsymbol{r}_{s,i}\|_\infty |v_n|^{k-s} \\
&\leq \alpha\sqrt{n}G_\infty |v_n|^k + 2\alpha\sqrt{n}G_\infty|\lambda_n| \sum_{s=1}^{\infty} |v_n|^{k-s} + 2\sqrt{n}\delta B_\theta \sum_{s=1}^{\infty} |v_n|^{k-s} \\
&\leq \alpha\sqrt{n}G_\infty|v_n| + \frac{2\alpha\sqrt{n}G_\infty|\lambda_n|}{1-|v_n|} + \frac{2\sqrt{n}\delta B_\theta}{1-|v_n|}
\end{aligned}
$$

where $v_n = \lambda_n - \sqrt{\lambda_n^2 - \lambda_n}$.

On the other hand, when $0 \leq \lambda_i < 1$, from Lemma G.3 we have

$$
\boldsymbol{y}_{k+1,i} \sin\phi_i = \boldsymbol{y}_{1,i}\lambda_i^{\frac{k}{2}} \sin[(t+1)\phi_i] + \lambda_i \sum_{s=1}^{k} \boldsymbol{\beta}_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\phi_i] + \sum_{s=1}^{k} \boldsymbol{r}_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\phi_i]
$$

By taking norm, we get

$$
\begin{aligned}
\left\|\boldsymbol{y}_{k+1,i}\right\|_\infty |\sin\phi_i| = &\left\|\boldsymbol{y}_{1,i}\right\|_\infty \lambda_i^{\frac{k}{2}} |\sin[(t+1)\phi_i]| + \lambda_i \sum_{s=1}^{k} \left\|\boldsymbol{\beta}_{s,i}\right\|_\infty |\lambda_i^{\frac{k-s}{2}}| |\sin[(k+1-s)\phi_i]| \\
&+ \sum_{s=1}^{k} \|\boldsymbol{r}_{s,i}\|_\infty |\lambda_i^{\frac{k-s}{2}}| |\sin[(k+1-s)\phi_i]| \\
&< \left\|\boldsymbol{y}_{1,i}\right\|_\infty \lambda_2^{\frac{k}{2}} + 2\alpha\sqrt{n}G_\infty\lambda_2 \sum_{s=1}^{\infty} \lambda_2^{\frac{s}{2}} + 2\sqrt{n}\delta B_\theta \sum_{s=1}^{\infty} \lambda_2^{\frac{s}{2}} \\
&\leq \alpha\sqrt{n}G_\infty\sqrt{\lambda_2} + \frac{2\alpha\sqrt{n}G_\infty\lambda_2 + 2\sqrt{n}\delta B_\theta}{\sqrt{1-\lambda_2}}
\end{aligned}
$$

Since $|\sin\phi_i| \geq \sqrt{1-\lambda_2}$, putting it back, we get

$$
\left\|\boldsymbol{y}_{k+1,i}\right\| < \alpha\sqrt{n}G_\infty\sqrt{\frac{\lambda_2}{1-\lambda_2}} + \frac{2\alpha\sqrt{n}G_\infty\lambda_2 + 2\sqrt{n}\delta B_\theta}{1-\lambda_2}
$$

So there exists D_1, D_2

$$
\begin{aligned}
D_1 &= \max\left\{|v_n| + \frac{2|\lambda_n|}{1-|v_n|}, \sqrt{\frac{\lambda_2}{1-\lambda_2}} + \frac{2\lambda_2}{1-\lambda_2}\right\} \\
D_2 &= \max\left\{\frac{2}{1-|v_n|}, \frac{2}{\sqrt{1-\lambda_2}}\right\}
\end{aligned}
$$

such that

$$
\left\|\boldsymbol{y}_{k+1,i}\right\|_\infty < D_1\alpha\sqrt{n}G_\infty + D_2\sqrt{n}\delta B_\theta
$$

Putting it back we have $\forall i, j$

$$
\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < D_1\alpha n G_\infty + D_2 n\delta B_\theta
$$

As a result

$$
\begin{aligned}
&\left\|\boldsymbol{X}_{k+1+\frac{1}{2}}(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty \\
=&\left\|(2\boldsymbol{X}_{k+1} - \boldsymbol{X}_k - \alpha\tilde{\boldsymbol{G}}_{k+1} + \alpha\tilde{\boldsymbol{G}}_k)(\boldsymbol{e}_i - \boldsymbol{e}_j)\right\|_\infty \\
\leq& 2\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty + \|\boldsymbol{X}_k(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty + \alpha\left\|\tilde{\boldsymbol{G}}_{k+1}\right\|_{1,\infty}\|\boldsymbol{e}_i - \boldsymbol{e}_j\|_1 + \alpha\left\|\tilde{\boldsymbol{G}}_k\right\|_{1,\infty}\|\boldsymbol{e}_i - \boldsymbol{e}_j\|_1 \\
<& 3(D_1\alpha n G_\infty + D_2 n\delta B_\theta) + 4\alpha G_\infty
\end{aligned}
$$

$$\leq (6D_1 n + 8)\alpha G_\infty$$

The last step is because $\delta = \frac{1}{12nD_2+2}$

Combining I and II we complete the proof. □

Lemma G.2. *By defining*

$$C_1 = \max\left\{\frac{3}{1-|v_n|^2}, \frac{3}{(1-\lambda_2)^2}\right\}$$
$$C_2 = \max\left\{\frac{3\lambda_n^2}{(1-|v_n|)^2}, \frac{3\lambda_2^2}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\}$$
$$C_3 = \max\left\{\frac{3}{(1-|v_n|)^2}, \frac{3}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\}$$

we have

$$\begin{aligned}
&(1-12C_2\alpha^2L^2)\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k - \boldsymbol{x}_{k,i}\right\|^2 \\
\leq & 3C_1\alpha^2 n\sigma^2 + 3C_1\alpha^2 n\varsigma_0^2 + 3C_1\alpha^2 n\mathbb{E}\left\|\nabla f(\mathbf{0})\right\| + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^4\sigma^2L^2K \\
&+ 3C_2\alpha^4 nL^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 + C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}\left\|\overline{\boldsymbol{X}}_k - \boldsymbol{x}_{k,i}\right\|^2 = & \sum_{i=1}^{n}\left\|\boldsymbol{X}_k\left(\boldsymbol{e}_i - \frac{\mathbf{1}}{n}\right)\right\|^2 \\
= & \left\|\boldsymbol{X}_k\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_F^2 \\
= & \left\|\boldsymbol{X}_k\boldsymbol{P}\boldsymbol{P}^\top - \boldsymbol{X}_k\boldsymbol{v}_1\boldsymbol{v}_1^\top\right\|_F^2 \\
\overset{\text{Lemma G.4}}{=} & \left\|\boldsymbol{X}_k\boldsymbol{P}\begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}\right\|_F^2 \\
= & \sum_{i=2}^{n}\left\|\boldsymbol{y}_{k,i}\right\|^2
\end{aligned}$$

From the update rule, we obtain,

$$\boldsymbol{y}_{k+1,i} = \lambda_i(2\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i} - \alpha\boldsymbol{h}_{k,i} + \alpha\boldsymbol{h}_{k-1,i}) + \boldsymbol{r}_{k,i} = \lambda_i(2\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i}) + \lambda_i\boldsymbol{\beta}_{k,i} + \boldsymbol{r}_{k,i}$$

where$\boldsymbol{\beta}_{k,i} = -\alpha\boldsymbol{h}_{k,i} + \alpha\boldsymbol{h}_{k-1,i}$, for all $\boldsymbol{y}_i$ with $-\frac{1}{3} < \lambda_i < 0$, from Lemma G.3 we have

$$\boldsymbol{y}_{k+1,i} = \boldsymbol{y}_{1,i}\left(\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i}\right) + \sum_{s=1}^{k}(\lambda_i\boldsymbol{\beta}_{s,i} + \boldsymbol{r}_{k,i})\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}$$

where $u_i = \lambda_i + \sqrt{\lambda_i^2 - \lambda_i}$ and $v_i = \lambda_i - \sqrt{\lambda_i^2 - \lambda_i}$, we obtain

$$\left\|\boldsymbol{y}_{k+1,i}\right\|^2 \leq 3\left\|\boldsymbol{y}_{1,i}\right\|^2\left(\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i}\right)^2 + 3\lambda_i^2\left(\sum_{s=1}^{k}\left\|\boldsymbol{\beta}_{s,i}\right\|\left|\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}\right|\right)^2$$

$$+ 3\left(\sum_{s=1}^{k}\|\boldsymbol{r}_{s,i}\|\left|\frac{u_i^{k-s+1}-v_i^{k-s+1}}{u_i-v_i}\right|\right)^2$$

Since

$$\left|\frac{u_i^{n+1}-v_i^{n+1}}{u_i-v_i}\right| \leq |v_i|^n\left|\frac{u_i\left(\frac{u_i}{v_i}\right)^n-v_i}{u_i-v_i}\right| \leq |v_i|^n$$

We obtain

$$\left\|\boldsymbol{y}_{k+1,i}\right\|^2 \leq 3\left\|\boldsymbol{y}_{1,i}\right\|^2|v_i|^{2t} + 3\lambda_i^2\left(\sum_{s=1}^{k}\left\|\boldsymbol{\beta}_{s,i}\right\||v_i|^{k-s}\right)^2 + 3\left(\sum_{s=1}^{k}\|\boldsymbol{r}_{s,i}\||v_i|^{k-s}\right)^2$$

Summing over from $k=0$ to $t=K-1$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{K-1}\left\|\boldsymbol{y}_{k+1,i}\right\|^2 = \sum_{k=1}^{K}\left\|\boldsymbol{y}_{k,i}\right\|^2\\
\leq& 3\left\|\boldsymbol{y}_{1,i}\right\|^2\sum_{k=0}^{K-1}|v_i|^{2k} + 3\lambda_i^2\sum_{k=1}^{K-1}\left(\sum_{s=1}^{k}\left\|\boldsymbol{\beta}_{s,i}\right\||v_i|^{k-s}\right)^2 + 3\sum_{k=1}^{K-1}\left(\sum_{s=1}^{k}\|\boldsymbol{r}_{s,i}\||v_i|^{k-s}\right)^2\\
\leq& \frac{3\left\|\boldsymbol{y}_{1,i}\right\|^2}{1-|v_i|^2} + \frac{3\lambda_i^2}{(1-|v_i|)^2}\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + \frac{3}{(1-|v_i|)^2}\sum_{k=1}^{K-1}\|\boldsymbol{r}_{k,i}\|^2\\
\leq& \frac{3\left\|\boldsymbol{y}_{1,i}\right\|^2}{1-|v_n|^2} + \frac{3\lambda_n^2}{(1-|v_n|)^2}\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + \frac{3}{(1-|v_n|)^2}\sum_{k=1}^{K-1}\|\boldsymbol{r}_{k,i}\|^2
\end{aligned}$$

where $v_n = \lambda_n - \sqrt{\lambda_n^2-\lambda_n}$.

On the other hand, when $0 \leq \lambda_i < 1$, from Lemma G.3 we have

$$\boldsymbol{y}_{k+1,i}\sin\phi_i = \boldsymbol{y}_{1,i}\lambda_i^{\frac{k}{2}}\sin[(t+1)\phi_i] + \lambda_i\sum_{s=1}^{k}\boldsymbol{\beta}_{s,i}\lambda_i^{\frac{k-s}{2}}\sin[(k+1-s)\phi_i] + \sum_{s=1}^{k}\boldsymbol{r}_{s,i}\lambda_i^{\frac{k-s}{2}}\sin[(k+1-s)\phi_i]$$

And we have

$$\begin{aligned}
\left\|\boldsymbol{y}_{k+1,i}\right\|^2\sin^2\phi_i \leq& 3\left\|\boldsymbol{y}_{1,i}\right\|^2\lambda_i^k\sin^2[(t+1)\phi_i] + 3\lambda_i^2\left(\sum_{s=1}^{k}\|\boldsymbol{\beta}_{s,i}\|\lambda_i^{\frac{k-s}{2}}\sin[(k+1-s)\phi_i]\right)^2\\
&+ 3\left(\sum_{s=1}^{k}\|\boldsymbol{r}_{s,i}\|\lambda_i^{\frac{k-s}{2}}\sin[(k+1-s)\phi_i]\right)^2\\
\leq& 3\left\|\boldsymbol{y}_{1,i}\right\|^2\lambda_i^k + 3\lambda_i^2\left(\sum_{s=1}^{k}\|\boldsymbol{\beta}_{s,i}\|\lambda_i^{\frac{k-s}{2}}\right)^2 + 3\left(\sum_{s=1}^{k}\|\boldsymbol{r}_{s,i}\|\lambda_i^{\frac{k-s}{2}}\right)^2
\end{aligned}$$

Summing from $k=0$ to $K-1$, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1}\left\|\boldsymbol{y}_{k+1,i}\right\|^2\sin^2\phi_i = \sum_{k=1}^{K}\left\|\boldsymbol{y}_{k,i}\right\|^2\sin^2\phi_i\\
\leq& 3\left\|\boldsymbol{y}_{1,i}\right\|^2\sum_{k=0}^{K-1}\lambda_i^t + 3\lambda_i^2\sum_{k=1}^{K-1}\left(\sum_{s=1}^{k}\left\|\boldsymbol{\beta}_{s,i}\right\|\lambda_i^{\frac{t-s}{2}}\right)^2 + 3\sum_{k=1}^{K-1}\left(\sum_{s=1}^{k}\|\boldsymbol{r}_{s,i}\|\lambda_i^{\frac{k-s}{2}}\right)^2\\
\leq& \frac{3\left\|\boldsymbol{y}_{1,i}\right\|^2}{1-\lambda_i} + \frac{3\lambda_i^2}{(1-\sqrt{\lambda_i})^2}\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + \frac{3}{(1-\sqrt{\lambda_i})^2}\sum_{k=1}^{K-1}\|\boldsymbol{r}_{k,i}\|^2
\end{aligned}$$

Since $\sin^2\phi_i = 1-\lambda_i$, we have

$$\begin{aligned}\sum_{k=1}^{K}\left\|\boldsymbol{y}_{k,i}\right\|^2 &\le \frac{3\left\|\boldsymbol{y}_{1,i}\right\|^2}{(1-\lambda_i)^2} + \frac{3\lambda_i^2}{(1-\sqrt{\lambda_i})^2(1-\lambda_i)}\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + \frac{3}{(1-\sqrt{\lambda_i})^2(1-\lambda_i)}\sum_{k=1}^{K-1}\left\|\boldsymbol{r}_{k,i}\right\|^2\\ &\le \frac{3\left\|\boldsymbol{y}_{1,i}\right\|^2}{(1-\lambda_2)^2} + \frac{3\lambda_2^2}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + \frac{3}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\sum_{k=1}^{K-1}\left\|\boldsymbol{r}_{k,i}\right\|^2\end{aligned}$$

So there exists C_1, C_2, C_3

$$\begin{aligned}C_1 &= \max\left\{\frac{3}{1-|v_n|^2}, \frac{3}{(1-\lambda_2)^2}\right\}\\ C_2 &= \max\left\{\frac{3\lambda_n^2}{(1-|v_n|)^2}, \frac{3\lambda_2^2}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\}\\ C_3 &= \max\left\{\frac{3}{(1-|v_n|)^2}, \frac{3}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\}\end{aligned}$$

$$\sum_{k=1}^{K}\left\|\boldsymbol{y}_{k,i}\right\|^2 \le C_1\left\|\boldsymbol{y}_{1,i}\right\|^2 + C_2\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + C_3\sum_{k=1}^{K-1}\left\|\boldsymbol{r}_{k,i}\right\|^2$$

By taking expectation we have

$$\sum_{k=1}^{K}\mathbb{E}\left\|\boldsymbol{y}_{k,i}\right\|^2 \le C_1\mathbb{E}\left\|\boldsymbol{y}_{1,i}\right\|^2 + C_2\sum_{k=1}^{K-1}\mathbb{E}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\boldsymbol{r}_{k,i}\right\|^2$$

We next analyze $\beta_{k,i}$:

$$\begin{aligned}&\sum_{i=2}^{n}\mathbb{E}\left\|\boldsymbol{\beta}_{k,i}\right\|^2\\ =&\alpha^2\sum_{i=2}^{n}\mathbb{E}\left\|\boldsymbol{h}_{k,i}-\boldsymbol{h}_{k-1,i}\right\|^2\\ =&\alpha^2\sum_{i=2}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\boldsymbol{P}\boldsymbol{e}_i-\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{P}\boldsymbol{e}_i\right\|^2\\ \le&\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\boldsymbol{P}\boldsymbol{e}_i-\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{P}\boldsymbol{e}_i\right\|^2\\ \le&\alpha^2\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\boldsymbol{P}-\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{P}\right\|_F^2\\ \overset{\text{Lemma G.4}}{\le}&\alpha^2\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k-\tilde{\boldsymbol{G}}_{k-1}\right\|_F^2\\ =&\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\boldsymbol{e}_i-\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{e}_i\right\|^2\\ \le&3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_k\boldsymbol{e}_i-\boldsymbol{G}_k\boldsymbol{e}_i\right\|^2 + 3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{e}_i-\boldsymbol{G}_{k-1}\boldsymbol{e}_i\right\|^2\\ &+ 3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\boldsymbol{G}_k\boldsymbol{e}_i-\boldsymbol{G}_{k-1}\boldsymbol{e}_i\right\|^2\\ \le&6\alpha^2n\sigma^2 + 3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\boldsymbol{G}_k\boldsymbol{e}_i-\boldsymbol{G}_{k-1}\boldsymbol{e}_i\right\|^2\end{aligned}$$

$$
\begin{aligned}
&\leq 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2 \sum_{i=1}^{n} \mathbb{E}\left\|\boldsymbol{x}_{k,i} - \boldsymbol{x}_{k-1,i}\right\|^2 \\
&\leq 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2 \sum_{i=1}^{n} \mathbb{E}\left\|\boldsymbol{Y}_k \boldsymbol{P}^\top \boldsymbol{e}_i - \boldsymbol{Y}_{k-1} \boldsymbol{P}^\top \boldsymbol{e}_i\right\|^2 \\
&\leq 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2 \mathbb{E}\left\|\boldsymbol{Y}_k \boldsymbol{P}^\top - \boldsymbol{Y}_{k-1} \boldsymbol{P}^\top\right\|_F^2 \\
&\overset{\text{Lemma G.4}}{\leq} 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2 \mathbb{E}\left\|\boldsymbol{Y}_k - \boldsymbol{Y}_{k-1}\right\|_F^2 \\
&\leq 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2 \sum_{i=1}^{n} \mathbb{E}\left\|\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i}\right\|^2
\end{aligned}
$$

Putting it back, we have

$$
\begin{aligned}
&\sum_{i=2}^{n}\sum_{k=1}^{K} \mathbb{E}\left\|\boldsymbol{y}_{k,i}\right\|^2 \\
\leq& C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2 + C_2 \sum_{i=2}^{n}\sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\beta}_{k,i}\right\|^2 + C_3 \sum_{k=1}^{K-1}\sum_{i=2}^{n} \mathbb{E}\left\|\boldsymbol{r}_{k,i}\right\|^2 \\
\leq& C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2 + C_2 \sum_{k=1}^{K-1} \left(6\alpha^2 n\sigma^2 + 3\alpha^2 L^2 \sum_{i=1}^{n} \mathbb{E}\left\|\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i}\right\|^2\right) + C_3 \sum_{k=1}^{K-1}\sum_{i=2}^{n} \mathbb{E}\left\|\boldsymbol{r}_{k,i}\right\|^2 \\
\overset{\text{Lemma G.4}}{\leq}& C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2 + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^2 L^2 \sum_{k=1}^{K-1}\sum_{i=1}^{n} \mathbb{E}\left\|\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i}\right\|^2 + C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

Since

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{y}_{k,1} - \boldsymbol{y}_{k-1,1}\right\|^2 = \mathbb{E}\left\|\boldsymbol{X}_k \boldsymbol{P}\boldsymbol{e}_1 - \boldsymbol{X}_{k-1}\boldsymbol{P}\boldsymbol{e}_1\right\|^2 = \mathbb{E}\left\|\boldsymbol{X}_k \boldsymbol{v}_1 - \boldsymbol{X}_{k-1}\boldsymbol{v}_1\right\|^2 \\
=& \mathbb{E}\left\|\boldsymbol{X}_k \frac{1}{\sqrt{n}}\boldsymbol{1} - X_{k-1}\frac{1}{\sqrt{n}}\boldsymbol{1}\right\|^2 = n\mathbb{E}\left\|\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}}_{k-1}\right\|^2 = n\alpha^2 \mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}_k}\right\|^2 \\
\leq& n\alpha^2 \mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}_k} - \overline{\boldsymbol{G}}_k\right\|^2 + n\alpha^2 \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 \leq n\alpha^2 \frac{\sigma^2}{n} + n\alpha^2 \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 \\
=& \alpha^2\sigma^2 + n\alpha^2 \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2
\end{aligned}
$$

Putting it back, and we obtain

$$
\begin{aligned}
&\sum_{i=2}^{n}\sum_{k=1}^{K} \mathbb{E}\left\|\boldsymbol{y}_{k,i}\right\|^2 \\
\leq& C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2 + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^4\sigma^2 L^2 K + 3C_2\alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 \\
&+ 3C_2\alpha^2 L^2 \sum_{k=1}^{K-1}\sum_{i=2}^{n} \mathbb{E}\left\|\boldsymbol{y}_{k,i} - \boldsymbol{y}_{k-1,i}\right\|^2 + C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2 \\
\leq& C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2 + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^4\sigma^2 L^2 K + 3C_2\alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 \\
&+ 6C_2\alpha^2 L^2 \sum_{k=1}^{K-1}\sum_{i=2}^{n} \mathbb{E}\left(\left\|\boldsymbol{y}_{k,i}\right\|^2 + \left\|\boldsymbol{y}_{k-1,i}\right\|^2\right) + C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
\leq & C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2+6 C_2 \alpha^2 n \sigma^2 K+3 C_2 \alpha^4 \sigma^2 L^2 K+3 C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 \\
& +12 C_2 \alpha^2 L^2 \sum_{k=1}^{K-1} \sum_{i=2}^n \mathbb{E}\left\|\boldsymbol{y}_{k, i}\right\|^2+C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

Rearrange the terms, we get

$$
\begin{aligned}
& \left(1-12 C_2 \alpha^2 L^2\right) \sum_{i=2}^n \sum_{k=1}^K \mathbb{E}\left\|\boldsymbol{y}_{k, i}\right\|^2 \\
\leq & C_1 \mathbb{E}\left\|\boldsymbol{Y}_1\right\|_F^2+6 C_2 \alpha^2 n \sigma^2 K+3 C_2 \alpha^4 \sigma^2 L^2 K+3 C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2 \\
\leq & C_1 \mathbb{E}\left\|\boldsymbol{X}_1\right\|_F^2+6 C_2 \alpha^2 n \sigma^2 K+3 C_2 \alpha^4 \sigma^2 L^2 K+3 C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

Considering

$$
\begin{aligned}
\mathbb{E}\left\|\boldsymbol{X}_1\right\|_F^2 & =\alpha^2 \mathbb{E}\left\|\tilde{\boldsymbol{G}}_0\right\|_F^2 \\
& =\alpha^2 \sum_{i=1}^n \mathbb{E}\left\|\tilde{\boldsymbol{G}}_{0, i}-\boldsymbol{G}_{0, i}+\boldsymbol{G}_{0, i}-\nabla f(\mathbf{0})+\nabla f(\mathbf{0})\right\|^2 \\
& \leq 3 \alpha^2 \sum_{i=1}^n \mathbb{E}\left\|\tilde{\boldsymbol{G}}_{0, i}-\boldsymbol{G}_{0, i}\right\|^2+3 \alpha^2 \sum_{i=1}^n \mathbb{E}\left\|\boldsymbol{G}_{0, i}-\nabla f(\mathbf{0})\right\|^2+3 \alpha^2 \sum_{i=1}^n \mathbb{E}\|\nabla f(\mathbf{0})\|^2 \\
& \leq 3 \alpha^2 n \sigma^2+3 \alpha^2 n \varsigma_0^2+3 \alpha^2 n \mathbb{E}\|\nabla f(\mathbf{0})\|
\end{aligned}
$$

We finally get

$$
\begin{aligned}
& \left(1-12 C_2 \alpha^2 L^2\right) \sum_{i=2}^n \sum_{k=1}^K \mathbb{E}\left\|\boldsymbol{y}_{k, i}\right\|^2 \\
= & \left(1-12 C_2 \alpha^2 L^2\right) \sum_{i=1}^n \sum_{k=1}^K \mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k, i}\right\|^2 \\
\leq & 3 C_1 \alpha^2 n \sigma^2+3 C_1 \alpha^2 n \varsigma_0^2+3 C_1 \alpha^2 n \mathbb{E}\|\nabla f(\mathbf{0})\|+6 C_2 \alpha^2 n \sigma^2 K+3 C_2 \alpha^4 \sigma^2 L^2 K \\
& +3 C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2
\end{aligned}
$$

That completes the proof. □

Lemma G.3. *Given $\rho \in \left(-\frac{1}{3}, 0\right) \cup (0,1)$, for any two sequence $\{a_t\}_{t=1}^{\infty}$, $\{b_t\}_{t=1}^{\infty}$ and $\{c_t\}_{t=1}^{\infty}$ that satisfying*

$$
\begin{aligned}
& a_0=b_0=0, \\
& a_{t+1}=\rho\left(2 a_t-a_{t-1}\right)+b_t-b_{t-1}+c_t, \forall t \geq 1
\end{aligned}
$$

we have

$$
a_{t+1}=a_1\left(\frac{u^{t+1}-v^{t+1}}{u-v}\right)+\sum_{s=1}^t\left(b_s-b_{s-1}+c_s\right)\left(\frac{u^{t-s+1}-v^{t-s+1}}{u-v}\right), \forall t \geq 0
$$

where

$$
u=\rho+\sqrt{\rho^2-\rho}, v=\rho-\sqrt{\rho^2-\rho}
$$

Moreover, if $0 < \rho < 1$*, we have*

$$a_{t+1} = a_1\rho^{\frac{t}{2}}\frac{\sin[(t+1)\phi]}{\sin\phi} + \sum_{s=1}^{t}(b_s - b_{s-1} + c_s)\rho^{\frac{t-s}{2}}\frac{\sin[(t-s+1)\phi]}{\sin\phi}$$

where

$$\phi = \arccos\left(\sqrt{\rho}\right)$$

Proof. when $t \geq 1$, we have

$$a_{t+1} = 2\rho a_t - \rho a_{t-1} + b_t - b_{t-1} + c_t$$

since,

$$u = \rho + \sqrt{\rho^2 - \rho}, v = \rho - \sqrt{\rho^2 - \rho}$$

we obtain

$$a_{t+1} - ua_t = (a_t - ua_{t-1})v + b_t - b_{t-1} + c_t$$

Recursively we have

$$\begin{aligned}
a_{t+1} - ua_t &= (a_t - ua_{t-1})v + b_t - b_{t-1} + c_t \\
&= (a_{t-1} - ua_{t-2})v^2 + (b_{t-1} - b_{t-2} + c_{t-1})v + b_t - b_{t-1} + c_t \\
&= (a_1 - ua_0)v^t + \sum_{s=1}^{t}(b_s - b_{s-1} + c_s)v^{t-s} \\
&= a_1v^t + \sum_{s=1}^{t}(b_s - b_{s-1} + c_s)v^{t-s}
\end{aligned}$$

Dividing both sides by u^{t+1}, we have

$$\begin{aligned}
\frac{a_{t+1}}{u^{t+1}} &= \frac{a_t}{u^t} + u^{-(t+1)}\left(a_1v^t + \sum_{s=1}^{t}(b_s - b_{s-1} + c_s)v^{t-s}\right) \\
&= \frac{a_{t-1}}{u^{t-1}} + u^{-t}\left(a_1v^{t-1} + \sum_{s=1}^{t-1}(b_s - b_{s-1} + c_s)v^{t-1-s}\right) \\
&+ u^{-(t+1)}\left(a_1v^t + \sum_{s=1}^{t}(b_s - b_{s-1} + c_s)v^{t-s}\right) \\
&= \frac{a_1}{u} + \sum_{k=1}^{t}u^{-k-1}\left(a_1v^k + \sum_{s=1}^{k}(b_s - b_{s-1} + c_s)v^{k-s}\right)
\end{aligned}$$

Multiplying both sides by u^{t+1}

$$\begin{aligned}
a_{t+1} &= a_1u^t + \sum_{k=1}^{t}u^{t-k}\left(a_1v^k + \sum_{s=1}^{k}(b_s - b_{s-1} + c_s)v^{t-s}\right) \\
&= a_1u^t\left(1 + \sum_{k=1}^{t}\left(\frac{v}{u}\right)^k\right) + u^t\sum_{k=1}^{t}\sum_{s=1}^{k}(b_s - b_{s-1} + c_s)v^{-s}\left(\frac{v}{u}\right)^k \\
&= a_1u^t\sum_{k=0}^{t}\left(\frac{v}{u}\right)^k + u^t\sum_{s=1}^{t}\sum_{k=s}^{t}(b_s - b_{s-1} + c_s)v^{-s}\left(\frac{v}{u}\right)^k \\
&= a_1u^t\left(\frac{1 - \left(\frac{v}{u}\right)^{t+1}}{1 - \frac{v}{u}}\right) + u^t\sum_{s=1}^{t}(b_s - b_{s-1} + c_s)v^{-s}\left(\frac{v}{u}\right)^s\frac{1 - \left(\frac{v}{u}\right)^{t-s-1}}{1 - \frac{v}{u}} \\
&= a_1\left(\frac{u^{t+1} - v^{t+1}}{u - v}\right) + \sum_{s=1}^{t}(b_s - b_{s-1} + c_s)\frac{u^{t-s+1} - v^{t-s+1}}{u - v}
\end{aligned}$$

Note that when $0 < \rho < 1$, both u and v are complex numbers, we have

$$u = \sqrt{\rho} e^{i\phi}, v = \sqrt{\rho} e^{-i\phi}$$

where $\phi = \arccos \sqrt{\rho}$. And under this context, we have

$$a_{t+1} = a_1 \rho^{\frac{t}{2}} \frac{\sin[(t+1)\phi]}{\sin \phi} + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) \rho^{\frac{t-s}{2}} \frac{\sin[(t-s+1)\phi]}{\sin \phi}$$

That completes the proof. □

Lemma G.4. *For any matrix $\boldsymbol{X} \in \mathbb{R}^{N \times n}$, we have*

$$\sum_{i=2}^{n} \|\boldsymbol{X} \boldsymbol{v}_i\|^2 \leq \sum_{i=1}^{n} \|\boldsymbol{X} \boldsymbol{v}_i\|^2 = \|\boldsymbol{X}\|_F^2$$
$$\sum_{i=1}^{n} \left\|\boldsymbol{X} \boldsymbol{P}^\top \boldsymbol{e}_i\right\|^2 = \left\|\boldsymbol{X} \boldsymbol{P}^\top\right\|_F^2 = \|\boldsymbol{X}\|_F^2$$

Proof.

$$\sum_{i=2}^{n} \|\boldsymbol{X}_t \boldsymbol{v}_i\|^2 \leq \sum_{i=1}^{n} \|\boldsymbol{X}_t \boldsymbol{v}_i\|^2 = \|\boldsymbol{X}_t \boldsymbol{P}\|_F^2 = Tr(\boldsymbol{X}_t \boldsymbol{P} \boldsymbol{P}^\top \boldsymbol{X}_t^\top) = Tr(\boldsymbol{X}_t \boldsymbol{X}_t^\top) = \|\boldsymbol{X}_t\|_F^2$$

And similarly,

$$\sum_{i=1}^{n} \left\|\boldsymbol{X} \boldsymbol{P}^\top \boldsymbol{e}_i\right\|^2 = \left\|\boldsymbol{X} \boldsymbol{P}^\top\right\|_F^2 = Tr(\boldsymbol{X}_t \boldsymbol{P}^\top \boldsymbol{P} \boldsymbol{X}_t^\top) = Tr(\boldsymbol{X}_t \boldsymbol{X}_t^\top) = \|\boldsymbol{X}_t\|_F^2$$

That completes the proof. □

Lemma G.5. *If we run Algorithm 1 for K iterations the following inequality holds:*

$$\begin{aligned}
&\left(1 - \frac{3C_1\alpha^2 L^2}{C_4}\right) \mathbb{E}\|\nabla f(\mathbf{0})\| + \left(1 - \alpha L - 3\frac{C_2}{C_4}\alpha^4 L^4\right) \frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2 + \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \\
\leq & \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{3C_1\alpha^2 L^2(\sigma^2 + \varsigma_0^2)}{C_4 K} + 6\frac{C_2}{C_4}\alpha^2\sigma^2 L^2 + 3\frac{C_2}{nC_4}\alpha^4\sigma^2 L^4 \\
+ & \frac{C_3 L^2}{C_4}\left(\frac{6D_1 n + 8}{6D_2 n + 1}\right)^2 \alpha^2 G_\infty^2 d
\end{aligned}$$

where

$$\begin{aligned}
C_1 &= \max\left\{\frac{3}{1 - |v_n|^2}, \frac{3}{(1-\lambda_2)^2}\right\} \\
C_2 &= \max\left\{\frac{3\lambda_n^2}{(1 - |v_n|)^2}, \frac{3\lambda_2^2}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\} \\
C_3 &= \max\left\{\frac{3}{(1 - |v_n|)^2}, \frac{3}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\} \\
C_4 &= 1 - 12C_2\alpha^2 L^2
\end{aligned}$$

Proof. Since

$$\begin{aligned}
\overline{\boldsymbol{X}}_{k+1} &= (2\boldsymbol{X}_k - \boldsymbol{X}_{k-1} - \alpha\tilde{\boldsymbol{G}}_k + \alpha\tilde{\boldsymbol{G}}_{k-1})\boldsymbol{W}\frac{\mathbf{1}}{n} + (\hat{\boldsymbol{X}}_{k+\frac{1}{2}} - \boldsymbol{X}_{k+\frac{1}{2}})(\boldsymbol{W} - \boldsymbol{I})\frac{\mathbf{1}}{n} \\
&= 2\overline{\boldsymbol{X}}_k - \overline{\boldsymbol{X}}_{k-1} - \alpha\overline{\tilde{\boldsymbol{G}}}_k + \alpha\overline{\tilde{\boldsymbol{G}}}_{k-1}
\end{aligned}$$

and we have

$$\overline{\boldsymbol{X}}_{k+1}-\overline{\boldsymbol{X}}_k=\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}}_{k-1}-\alpha\overline{\tilde{\boldsymbol{G}}}_k+\alpha\overline{\tilde{\boldsymbol{G}}}_{k-1}$$
$$=\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_0-\alpha\sum_{t=1}^{k}(\overline{\tilde{\boldsymbol{G}}}_t-\overline{\tilde{\boldsymbol{G}}}_{t-1})$$
$$=-\alpha\overline{\tilde{\boldsymbol{G}}}_k$$

Note that the update of the averaged model is exactly the same as D-PSGD, thus we can reuse the result from D-PSGD for D^2 as follows:

$$\frac{1-\alpha L}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\leq\frac{2(f(0)-f^*)}{\alpha K}+\frac{\alpha L}{n}\sigma^2+\frac{L^2}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\boldsymbol{x}_{k,i}\right\|^2$$

From Lemma G.2 we obatin

$$\begin{aligned}&\frac{1-\alpha L}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\\ \leq&\frac{2(f(\mathbf{0})-f^*)}{\alpha K}+\frac{\alpha L}{n}\sigma^2+\frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2+\mathbb{E}\left\|\nabla f(\mathbf{0})\right\|)}{C_4K}+6\frac{C_2}{C_4}\alpha^2\sigma^2L^2+3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4\\ &+3\frac{C_2}{C_4}\alpha^4L^4\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{C_3L^2}{C_4nK}\sum_{k=1}^{K-1}\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2\end{aligned}$$

Rearrange the terms, we get

$$\begin{aligned}&\left(1-\frac{3C_1\alpha^2L^2}{C_4}\right)\mathbb{E}\left\|\nabla f(\mathbf{0})\right\|+\left(1-\alpha L-3\frac{C_2}{C_4}\alpha^4L^4\right)\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\\ \leq&\frac{2(f(0)-f^*)}{\alpha K}+\frac{\alpha L}{n}\sigma^2+\frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2)}{C_4K}+6\frac{C_2}{C_4}\alpha^2\sigma^2L^2+3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4+\frac{C_3L^2}{C_4nK}\sum_{k=1}^{K-1}\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2\end{aligned}$$

Similar to the case in D-PSGD, we have

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{\Omega}_k\right\|_F^2=&\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{j=1}^{n}\left((\hat{\boldsymbol{x}}_{k+\frac{1}{2},j}-\boldsymbol{x}_{k+\frac{1}{2},j})-(\hat{\boldsymbol{x}}_{k+\frac{1}{2},i}-\boldsymbol{x}_{k+\frac{1}{2},i})\right)\boldsymbol{W}_{ji}\right\|^2\\ \overset{\text{Lemma F.1}}{\leq}&4\sum_{k=0}^{K-1}\sum_{i=1}^{n}\delta^2B_\theta^2d\leq\left(\frac{6D_1n+8}{6D_2n+1}\right)^2\alpha^2G_\infty^2dnK\end{aligned}$$

Putting it back, we obtain

$$\begin{aligned}&\left(1-\frac{3C_1\alpha^2L^2}{C_4}\right)\mathbb{E}\left\|\nabla f(\mathbf{0})\right\|+\left(1-\alpha L-3\frac{C_2}{C_4}\alpha^4L^4\right)\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2\\ \leq&\frac{2(f(\mathbf{0})-f^*)}{\alpha K}+\frac{\alpha L}{n}\sigma^2+\frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2)}{C_4K}+6\frac{C_2}{C_4}\alpha^2\sigma^2L^2+3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4+\frac{C_3L^2}{C_4}\left(\frac{6D_1n+8}{6D_2n+1}\right)^2\alpha^2G_\infty^2d\end{aligned}$$

That completes the proof. □

H. Moniqua on AD-PSGD (Proof to Theorem 5)

H.1. Definition and Notation

In the original analysis of AD-PSGD, to better capture the nature of workers computing at different speed, the objective function is expressed as

$$f(\boldsymbol{x})=\sum_{i=1}^{n}p_if_i(\boldsymbol{x})$$

Algorithm 2 Moniqua with Asynchronous Communication

Require: initial point $\boldsymbol{x}_{0,i} = \boldsymbol{x}_0$, step size α, the discrepency bound B_θ, number of iterations K, quantization function $\mathcal{Q}_\delta$, initial random seed

1: **for** $k = 0, 1, 2, \cdots, K-1$ **do**

2: worker i_k is updating the gradient while during this iteration the global communication behaviour is written in the form of $\boldsymbol{W}_k$.

3: Compute a local stochastic gradient with model delayed by τ_k: $\tilde{\boldsymbol{g}}_{k-\tau_k,i_k}$

4: Send modulo-ed model to one randomly selected neighbor j_k: $\boldsymbol{q}_{k,i_k} \leftarrow \mathcal{Q}_\delta\left(\frac{\boldsymbol{x}_{k,i_k}}{B_\theta} \bmod 1\right)$

5: Compute local biased term $\hat{\boldsymbol{x}}_{k,i_k}$ as:

$$\hat{\boldsymbol{x}}_{k,i_k} = \boldsymbol{q}_{k,i_k} B_\theta - \boldsymbol{x}_{k,i_k} \bmod B_\theta + \boldsymbol{x}_{k,i_k}$$

6: Randomly select one neighbor j_k and recover its model as:

$$\hat{\boldsymbol{x}}_{k,j_k} = (\boldsymbol{q}_{k,j_k} B_\theta - \boldsymbol{x}_{k,i_k}) \bmod B_\theta + \boldsymbol{x}_{k,i}$$

7: Average with neighboring workers: $\boldsymbol{x}_{k,i_k} \leftarrow \boldsymbol{x}_{k,i_k} + \sum_{j \in \mathcal{N}_i} (\hat{\boldsymbol{x}}_{k,j_k} - \hat{\boldsymbol{x}}_{k,i_k}) \boldsymbol{W}_{ji}$

8: Update the local weight with local gradient: $\boldsymbol{x}_{k+1,i_k} \leftarrow \boldsymbol{x}_{k,i_k} - \alpha \tilde{\boldsymbol{g}}_{k-\tau_k,i_k}$

9: **end for**

10: **return** $\overline{\boldsymbol{X}}_K = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{K,i}$

where p_i is a parameter denoting the speed of i-th worker gradient updates. In the rest of the proof, we denote $p = \max_i\{p_i\}$

For simplicity, we also define the following terms

$$\begin{aligned}
\nabla F(\boldsymbol{X}_k) &= n\left[p_1 \boldsymbol{g}_{k,1}, \cdots, p_n \boldsymbol{g}_{k,n}\right] \in \mathbb{R}^{d\times n} \\
\nabla \widetilde{F}(\boldsymbol{X}_k) &= n\left[p_1 \tilde{\boldsymbol{g}}_{k,1}, \cdots, p_n \tilde{\boldsymbol{g}}_{k,n}\right] \in \mathbb{R}^{d\times n} \\
\tilde{\boldsymbol{G}}_k &= \left[\cdots, \tilde{\boldsymbol{g}}_{k,i_k}, \cdots\right] \\
\boldsymbol{G}_k &= \left[\cdots, \boldsymbol{g}_{k,i_k}, \cdots\right] \\
\boldsymbol{\Lambda}_a^b &= \frac{\mathbf{1}\mathbf{1}^\top}{n} - \prod_{q=a}^b \boldsymbol{W}_q
\end{aligned}$$

H.2. Setting

The pseudo code can be found in Algorithm 2. We makes the following assumptions:

1. **Lipschitzian Gradient**: All the function f_i have L-Lipschitzian gradients.
2. **Communication Matrix** [4]: The communication matrix $\boldsymbol{W}_k$ is doubly stochastic for any $k \geq 0$ and for any $b \geq a \geq 0$, there exists t_{mix} such that

$$\left\|\prod_{q=a}^b \boldsymbol{W}_q \left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right\|_1 \leq 2 \cdot 2^{-\left\lfloor \frac{b-a+1}{t_{\text{mix}}} \right\rfloor}$$

3. **Bounded Variance**:

$$\begin{aligned}
\mathbb{E}_{\xi_i \sim \mathcal{D}_i} \left\|\nabla \tilde{f}_i(\boldsymbol{x};\xi_i) - \nabla f_i(\boldsymbol{x})\right\|^2 &\leq \sigma^2, \forall i \\
\mathbb{E}_{i \sim \{1,\cdots,n\}} \left\|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2 &\leq \varsigma^2, \forall i
\end{aligned}$$

where $\nabla \tilde{f}_i(\boldsymbol{x};\xi_i)$ denotes gradient sample on worker i computed via data sample ξ_i.

[4]Please refer to Section E for more details

4. **Bounded Staleness**: There exists T such that $\tau_k \leq T, \forall k$
5. **Gradient magnitude**: The norm of a sampled gradient is bounded by $\left\|\tilde{\boldsymbol{g}}_{k,i}\right\|_\infty \leq G_\infty$ for some constant G_∞.

H.3. Proof to Theorem 5.

Proof. We start from

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 + \left(1-\frac{2\alpha L}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2\\
\overset{\text{Lemma H.4}}{\leq}&\frac{2n(f(\boldsymbol{0})-f^*)}{\alpha K} + \frac{(\sigma^2+6\varsigma^2)\alpha L}{n} + \left(2L^2+\frac{12\alpha L^3}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\boldsymbol{1}}{n}-\boldsymbol{e}_i\right)\right\|^2\\
&+\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(\boldsymbol{X}_k-\boldsymbol{X}_{k-\tau_k})\boldsymbol{1}}{n}\right\|^2\\
\overset{\text{Lemma H.5}}{\leq}&\frac{2n(f(\boldsymbol{0})-f^*)}{\alpha K} + \frac{(\sigma^2+6\varsigma^2)\alpha L}{n} + \frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2} + \frac{4\alpha^2T^2L^2}{n^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i\boldsymbol{g}_{k-\tau_k,i}\right\|^2\\
&+\left(2L^2+\frac{12\alpha L^3}{n}+\frac{24L^4\alpha^2T^2}{n^2}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\boldsymbol{1}}{n}-e_i\right)\right\|^2\\
\overset{\text{Lemma H.3}}{\leq}&\frac{2n(f(\boldsymbol{0})-f^*)}{\alpha K} + \frac{(\sigma^2+6\varsigma^2)\alpha L}{n} + \frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2} + \frac{4\alpha^2T^2L^2}{n^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i\boldsymbol{g}_{k-\tau_k,i}\right\|^2\\
&+\frac{128\alpha^2t_{\text{mix}}^2L^2}{A_1}\left((\sigma^2+6\varsigma^2)p+\frac{2p}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i\boldsymbol{g}_{k-\tau_k,i}\right\|^2+G_\infty^2 d\right)
\end{aligned}$$

where $A_1 = 1-192p\alpha^2t_{\text{mix}}^2L^2$ as defined in Lemma H.3.

Rearrange the terms, we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \leq& \frac{2n(f(\boldsymbol{0})-f^*)}{\alpha K} + \frac{(\sigma^2+6\varsigma^2)\alpha L}{n} + \frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2}\\
&+\frac{128p\alpha^2t_{\text{mix}}^2L^2}{A_1}(\sigma^2+6\varsigma^2) + \frac{128\alpha^2t_{\text{mix}}^2L^2}{A_1}G_\infty^2 d
\end{aligned}$$

By setting $\alpha = \frac{n}{2L+\sqrt{K(\sigma^2+6\varsigma^2)}}$

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 \lesssim& \frac{1}{K} + \frac{\sqrt{\sigma^2+6\varsigma^2}}{\sqrt{K}} + \frac{pt_{\text{mix}}^2(\sigma^2+6\varsigma^2)n^2}{(\sigma^2+6\varsigma^2)K+4L^2} + \frac{n^2t_{\text{mix}}^2G_\infty^2 d}{(\sigma^2+6\varsigma^2)K+4L^2}\\
\lesssim& \frac{1}{K} + \frac{\sqrt{\sigma^2+6\varsigma^2}}{\sqrt{K}} + \frac{(\sigma^2+6\varsigma^2)t_{\text{mix}}^2n^2}{(\sigma^2+6\varsigma^2)K+1} + \frac{n^2t_{\text{mix}}^2G_\infty^2 d}{(\sigma^2+6\varsigma^2)K+1}
\end{aligned}$$

□

H.4. Lemma for Moniqua on AD-PSGD

Lemma H.1.

$$\mathbb{E}\left\|\tilde{\boldsymbol{G}}_{k-\tau_k}\frac{\boldsymbol{1}}{n}\right\|^2 \leq \frac{\sigma^2}{n^2} + \frac{1}{n^2}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i}\right\|^2, \forall k\geq 0.$$

Proof.

$$\begin{aligned}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_{k-\tau_k}\frac{\mathbf{1}}{n}\right\|^2 &\leq \sum_{i=1}^{n} p_i \mathbb{E}\left\|\frac{\tilde{\boldsymbol{g}}_{k-\tau_k,i}}{n}\right\|^2 \\ &= \sum_{i=1}^{n} p_i \mathbb{E}\left\|\frac{\tilde{\boldsymbol{g}}_{k-\tau_k,i}-\boldsymbol{g}_{k-\tau_k,i}}{n}\right\|^2 + \sum_{i=1}^{n} p_i \mathbb{E}\left\|\frac{\boldsymbol{g}_{k-\tau_k,i}}{n}\right\|^2 \\ &\leq \frac{\sigma^2}{n^2} + \frac{1}{n^2}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i}\right\|^2\end{aligned}$$

□

Lemma H.2.

$$\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i}\right\|^2 \leq 12L^2 \sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_i\right)\right\|^2 + 6\varsigma^2 + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2, \forall k \geq 0.$$

Proof.

$$\begin{aligned}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i}\right\|^2 &= \sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i} - \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i} + \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2 \\ &\leq 2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i} - \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2 + 2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2 \\ &= 2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i} - \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2 + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2\end{aligned}$$

And

$$\begin{aligned}&\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i} - \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2 \\ \leq& 3\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i} - \nabla f_i(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 + 3\sum_{i=1}^{n} p_i \mathbb{E}\left\|\nabla f_i(\overline{\boldsymbol{X}}_{k-\tau_k}) - \sum_{j=1}^{n} p_j \nabla f_j(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 \\ &+ 3\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i} - \sum_{j=1}^{n} p_j \nabla f_j(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 \\ \leq& 3L^2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{x}_{k-\tau_k,i} - \overline{\boldsymbol{X}}_{k-\tau_k}\right\|^2 + 3\sum_{i=1}^{n} p_i \mathbb{E}\left\|\nabla f_i(\overline{\boldsymbol{X}}_{k-\tau_k}) - \sum_{j=1}^{n} p_j \nabla f_j(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 \\ &+ 3\mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i} - \sum_{j=1}^{n} p_j \nabla f_j(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 \\ \leq& 3L^2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - e_i\right)\right\|^2 + 3\sum_{i=1}^{n} p_i \mathbb{E}\left\|\nabla f_i(\overline{\boldsymbol{X}}_{k-\tau_k}) - \nabla f(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 \\ &+ 3\sum_{j=1}^{n} p_j \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,j} - \nabla f_j(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 \\ \leq& 6L^2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 + 3\varsigma^2\end{aligned}$$

That completes the proof. $\square$

Lemma H.3. *Let* $A_1 = 1 - 192p\alpha^2 t_{\text{mix}}^2 L^2$,

$$\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - e_i\right)\right\|^2 \leq \frac{32\alpha^2 t_{\text{mix}}^2}{A_1}\left((\sigma^2 + 6\varsigma^2)pK + 2p\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2 + G_\infty^2 dK\right)$$

Proof.

$$\begin{aligned}
&\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_k\left(\frac{\mathbf{1}}{n} - e_i\right)\right\|^2\\
=&\sum_{i=1}^{n} p_i \mathbb{E}\left\|\left(\boldsymbol{X}_{k-1}\boldsymbol{W}_{k-1} - \alpha\tilde{\boldsymbol{G}}_{k-1-\tau_{k-1}} + \boldsymbol{\Omega}_{k-1}\right)\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2\\
\overset{X_0=0}{=}&\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{t=0}^{k-1}\left(-\alpha\tilde{\boldsymbol{G}}_{t-\tau_t} + \boldsymbol{\Omega}_t\right)\boldsymbol{\Lambda}_{t+1}^{k-1}\boldsymbol{e}_i\right\|^2\\
\leq&2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{t=0}^{k-1}\alpha\tilde{\boldsymbol{G}}_{t-\tau_t}\boldsymbol{\Lambda}_{t+1}^{k-1}\boldsymbol{e}_i\right\|^2 + 2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{t=0}^{k-1}\boldsymbol{\Omega}_t\boldsymbol{\Lambda}_{t+1}^{k-1}\boldsymbol{e}_i\right\|^2
\end{aligned}$$

Now for the first term, we have

$$\begin{aligned}
2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{t=0}^{k-1}\alpha\tilde{\boldsymbol{G}}_{t-\tau_t}\boldsymbol{\Lambda}_{t+1}^{k-1}e_i\right\|^2 \leq&2p\alpha^2\mathbb{E}\left\|\sum_{t=0}^{k-1}\tilde{\boldsymbol{G}}_{t-\tau_t}\boldsymbol{\Lambda}_{t+1}^{k-1}\right\|_F^2\\
\leq&2p\alpha^2\mathbb{E}\left(\sum_{t=0}^{k-1}\left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F\left\|\boldsymbol{\Lambda}_{t+1}^{k-1}\right\|\right)^2\\
\leq&2p\alpha^2\mathbb{E}\left(\sum_{t=0}^{k-1}\left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F\left\|\boldsymbol{\Lambda}_{t+1}^{k-1}\right\|_1\right)^2\\
\leq&8p\alpha^2\mathbb{E}\left(\sum_{t=0}^{k-1}\left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2
\end{aligned}$$

Now we replace k with $k-\tau_k$, that is

$$\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 \leq 8p\alpha^2\mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1}\left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 + 2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\sum_{t=0}^{k-\tau_k-1}\boldsymbol{\Omega}_t\boldsymbol{\Lambda}_{t+1}^{k-\tau_k-1}\boldsymbol{e}_i\right\|^2$$

Summing from $k=0$ to $K-1$ on both sides, we obtain

$$\begin{aligned}
&\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2\\
\leq&8p\alpha^2\sum_{k=0}^{K-1}\mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1}\left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2\\
&+2\sum_{i=1}^{n} p_i\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{t=0}^{k-\tau_k-1}\boldsymbol{\Omega}_t\boldsymbol{\Lambda}_{t+1}^{k-\tau_k-1}e_i\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1} \left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 \\
&\quad + 2\sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1} \|\boldsymbol{\Omega}_t\|_{1,2} \left\|\boldsymbol{\Lambda}_{t+1}^{k-\tau_k-1}\right\|_1 \|\boldsymbol{e}_i\|_1\right)^2 \\
&\leq 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1} \left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 \\
&\quad + 8\sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1} \|\boldsymbol{\Omega}_t\|_{1,2} 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 \\
&\overset{\text{Lemma H.6}}{\leq} 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1} \left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 + 32t_{\text{mix}}^2 \sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \mathbb{E}\|\boldsymbol{\Omega}_k\|_{1,2}^2 \\
&\leq 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1} \left\|\tilde{\boldsymbol{G}}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 + 128\delta^2 B_\theta^2 d t_{\text{mix}}^2 K \\
&\overset{\text{Lemma H.6}}{\leq} 32p\alpha^2 t_{\text{mix}}^2 \sum_{k=0}^{K-1} \mathbb{E}\left\|\tilde{\boldsymbol{G}}_{k-\tau_k}\right\|_F^2 + 128\delta^2 B_\theta^2 d t_{\text{mix}}^2 K
\end{aligned}
$$

Note that for the first term, we have

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left\|\tilde{\boldsymbol{G}}_{k-\tau_k}\right\|_F^2 \\
=&\sum_{k=0}^{K-1} \mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k-\tau_k,i_k}\right\|^2 \\
=&\sum_{k=0}^{K-1} \mathbb{E}\left\|\tilde{\boldsymbol{g}}_{k-\tau_k,i_k} - \boldsymbol{g}_{k-\tau_k,i_k}\right\|^2 + \sum_{k=0}^{K-1} \mathbb{E}\left\|\boldsymbol{g}_{k-\tau_k,i_k}\right\|^2 \\
\leq&\sigma^2 K + \sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{g}_{t-\tau_t,i}\right\|^2 \\
\leq&(\sigma^2 + 6\varsigma^2)K + 12L^2 \sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 + 2\sum_{k=0}^{K-1} \mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2
\end{aligned}
$$

Putting these two terms back, we obtain

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 \\
\leq&32p\alpha^2 t_{\text{mix}}^2 \left((\sigma^2 + 6\varsigma^2)K + 12L^2 \sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 + 2\sum_{k=0}^{K-1} \mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2\right) \\
&+128\delta^2 B_\theta^2 d t_{\text{mix}}^2 K
\end{aligned}
$$

Rearrange the terms, we obtain

$$
\begin{aligned}
&\left(1 - 192p\alpha^2 t_{\text{mix}}^2 L^2\right) \sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 \\
\leq&32p\alpha^2 t_{\text{mix}}^2 \left((\sigma^2 + 6\varsigma^2)K + 2\sum_{k=0}^{K-1} \mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2\right) + 128\delta^2 B_\theta^2 t_{\text{mix}}^2 K
\end{aligned}
$$

$$\overset{\text{Lemma H.7}}{\leq} 32\alpha^2 t_{\text{mix}}^2 \left((\sigma^2 + 6\varsigma^2)pK + 2p \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^n p_i \boldsymbol{g}_{k-\tau_k,i} \right\|^2 + G_\infty^2 dK \right)$$

Let $A_1 = 1 - 192p\alpha^2 t_{\text{mix}}^2 L^2$, we obtain

$$\sum_{k=0}^{K-1} \sum_{i=1}^n p_i \mathbb{E} \left\| \boldsymbol{X}_{k-\tau_k} \left(\frac{\boldsymbol{1}}{n} - \boldsymbol{e}_i \right) \right\|^2 \leq \frac{32\alpha^2 t_{\text{mix}}^2}{A_1} \left((\sigma^2 + 6\varsigma^2)pK + 2p \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^n p_i \boldsymbol{g}_{k-\tau_k,i} \right\|^2 + G_\infty^2 dK \right)$$

□

Lemma H.4.

$$\begin{aligned}
&\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(\overline{\boldsymbol{X}}_k) \right\|^2 + \left(1 - \frac{2\alpha L}{n}\right) \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \right\|^2 \\
\leq &\frac{2n(f(\boldsymbol{0}) - f^*)}{\alpha K} + \frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \frac{(\boldsymbol{X}_k - \boldsymbol{X}_{k-\tau_k})\boldsymbol{1}}{n} \right\|^2 \\
&+ \left(2L^2 + \frac{12\alpha L^3}{n}\right) \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^n p_i \mathbb{E} \left\| \boldsymbol{X}_{k-\tau_k} \left(\frac{\boldsymbol{1}}{n} - e_i \right) \right\|^2 + \frac{(\sigma^2 + 6\varsigma^2)\alpha L}{n}
\end{aligned}$$

Proof. We start from $f(\overline{\boldsymbol{X}}_{k+1})$ Since

$$\overline{\boldsymbol{X}}_{k+1} = \boldsymbol{X}_k \boldsymbol{W}_k \frac{\boldsymbol{1}}{n} + (\hat{\boldsymbol{X}}_k - \boldsymbol{X}_k)(\boldsymbol{W}_k - \boldsymbol{I})\frac{\boldsymbol{1}}{n} - \alpha \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k} = \overline{\boldsymbol{X}}_k - \alpha \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k}$$

Then from Taylor Expansion, we have

$$\begin{aligned}
&\mathbb{E} f(\overline{\boldsymbol{X}}_{k+1}) \\
=&\mathbb{E} f\left(\overline{\boldsymbol{X}}_k - \alpha \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k} \right) \\
\leq&\mathbb{E} f(\overline{\boldsymbol{X}}_k) - \alpha \mathbb{E} \langle \nabla f(\overline{\boldsymbol{X}}_k), \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k} \rangle + \frac{\alpha^2 L}{2} \mathbb{E} \left\| \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k} \right\|^2 \\
=&\mathbb{E} f(\overline{\boldsymbol{X}}_k) - \alpha \mathbb{E} \langle \nabla f(\overline{\boldsymbol{X}}_k), \overline{\boldsymbol{G}}_{k-\tau_k} \rangle - \alpha \mathbb{E} \langle \nabla f(\overline{\boldsymbol{X}}_k), \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k} - \overline{\boldsymbol{G}}_{k-\tau_k} \rangle + \frac{\alpha^2 L}{2} \mathbb{E} \left\| \overline{\tilde{\boldsymbol{G}}}_{k-\tau_k} \right\|^2 \\
=&\mathbb{E} f(\overline{\boldsymbol{X}}_k) - \frac{\alpha}{n} \mathbb{E} \langle \nabla f(\overline{\boldsymbol{X}}_k), \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \rangle + \frac{\alpha^2 L}{2} \mathbb{E} \left\| \frac{\tilde{\boldsymbol{g}}_{k-\tau_k,i_k}}{n} \right\|^2 \\
\leq&\mathbb{E} f(\overline{\boldsymbol{X}}_k) - \frac{\alpha}{n} \mathbb{E} \langle \nabla f(\overline{\boldsymbol{X}}_k), \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \rangle \\
&+ \frac{\alpha^2 L}{2} \sum_{i=1}^n p_i \mathbb{E} \left\| \frac{\tilde{\boldsymbol{g}}_{k-\tau_k,i_k} - \boldsymbol{g}_{k-\tau_k,i_k}}{n} \right\|^2 + \frac{\alpha^2 L}{2} \sum_{i=1}^n p_i \mathbb{E} \left\| \frac{\boldsymbol{g}_{k-\tau_k,i}}{n} \right\|^2 \\
\leq&\mathbb{E} f(\overline{\boldsymbol{X}}_k) - \frac{\alpha}{n} \mathbb{E} \langle \nabla f(\overline{\boldsymbol{X}}_k), \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \rangle + \frac{\alpha^2 L \sigma^2}{2n^2} + \frac{\alpha^2 L}{2n^2} \sum_{i=1}^n p_i \mathbb{E} \|\boldsymbol{g}_{k-\tau_k,i}\|^2 \\
=&\mathbb{E} f(\overline{\boldsymbol{X}}_k) + \frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(\overline{\boldsymbol{X}}_k) - \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \right\|^2 - \frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(\overline{\boldsymbol{X}}_k) \right\|^2 - \frac{\alpha}{2n} \mathbb{E} \left\| \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \right\|^2 \\
+&\frac{\alpha^2 L \sigma^2}{2n^2} + \frac{\alpha^2 L}{2n^2} \sum_{i=1}^n p_i \mathbb{E} \|\boldsymbol{g}_{k-\tau_k,i}\|^2
\end{aligned}$$

Rearrange these terms, we can get

$$\frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(\overline{\boldsymbol{X}}_k) \right\|^2 + \frac{\alpha}{2n} \mathbb{E} \left\| \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k}) \right\|^2$$

$$\leq \mathbb{E} f(\overline{\boldsymbol{X}}_k) - \mathbb{E} f(\overline{\boldsymbol{X}}_{k+1}) + \frac{\alpha}{2n}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2$$
$$+ \frac{\alpha^2 L\sigma^2}{2n^2} + \frac{\alpha^2 L}{2n^2}\sum_{i=1}^{n} p_i \mathbb{E}\|\boldsymbol{g}_{k-\tau_k,i}\|^2$$

Summing over $k = 0$ to $K-1$ on both sides, we can get

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2$$
$$\leq \frac{2n(f(\mathbf{0}) - f^*)}{\alpha K} + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2 + \frac{\alpha L\sigma^2}{n} + \frac{\alpha L}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\|\boldsymbol{g}_{k-\tau_k,i}\|^2$$

For $\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2$, we have

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2$$
$$\leq 2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \nabla f(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 + 2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_{k-\tau_k}) - \nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2$$
$$= 2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_{\boldsymbol{k}}) - \nabla f(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 + 2\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n} p_i\left(\nabla f_i(\overline{\boldsymbol{X}}_{k-\tau_k}) - \boldsymbol{g}_{k-\tau_k,i}\right)\right\|^2$$
$$\leq 2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k) - \nabla f(\overline{\boldsymbol{X}}_{k-\tau_k})\right\|^2 + 2\sum_{k=0}^{K-1}\mathbb{E}\sum_{i=1}^{n} p_i\left\|\nabla f_i(\overline{\boldsymbol{X}}_{k-\tau_k}) - \boldsymbol{g}_{k-\tau_k,i}\right\|^2$$
$$\leq 2L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(\boldsymbol{X}_k - \boldsymbol{X}_{k-\tau_k})\mathbf{1}}{n}\right\|^2 + 2L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - e_i\right)\right\|^2$$

Putting it back, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla \overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2$$
$$\leq \frac{2n(f(\mathbf{0}) - f^*)}{\alpha K} + \frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(\boldsymbol{X}_k - \boldsymbol{X}_{k-\tau_k})\mathbf{1}}{n}\right\|^2$$
$$+ \frac{2L^2}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - e_i\right)\right\|^2 + \frac{\alpha L\sigma^2}{n} + \frac{\alpha L}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\|\boldsymbol{g}_{k-\tau_k,i}\|^2$$
$$\overset{Lemma\ H.2}{\leq} \frac{2n(f(\mathbf{0}) - f^*)}{\alpha K} + \frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(\boldsymbol{X}_k - \boldsymbol{X}_{k-\tau_k})\mathbf{1}}{n}\right\|^2$$
$$+ \frac{2L^2}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - e_i\right)\right\|^2 + \frac{\alpha L\sigma^2}{n}$$
$$+ \frac{\alpha L}{nK}\sum_{k=0}^{K-1}\left(12L^2\sum_{i=1}^{n} p_i \mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i\right)\right\|^2 + 6\varsigma^2 + 2\mathbb{E}\left\|\sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i}\right\|^2\right)$$
$$= \frac{2n(f(\mathbf{0}) - f^*)}{\alpha K} + \frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(\boldsymbol{X}_k - \boldsymbol{X}_{k-\tau_k})\mathbf{1}}{n}\right\|^2$$

$$
\begin{aligned}
&+\left(2L^2+\frac{12\alpha L^3}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_i\right)\right\|^2\\
&+\frac{(\sigma^2+6\varsigma^2)\alpha L}{n}+\frac{2\alpha L}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i\boldsymbol{g}_{k-\tau_k,i}\right\|^2
\end{aligned}
$$

Note that

$$
\mathbb{E}\left\|\sum_{i=1}^{n}p_i\boldsymbol{g}_{k-\tau_k,i}\right\|^2=\mathbb{E}\left\|\nabla\overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2
$$

Moving it to the left side, we finally get

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{\boldsymbol{X}}_k)\right\|^2+\left(1-\frac{2\alpha L}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla\overline{F}(\boldsymbol{X}_{k-\tau_k})\right\|^2\\
\leq&\frac{2n(f(\mathbf{0})-f^*)}{\alpha K}+\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(\boldsymbol{X}_k-\boldsymbol{X}_{k-\tau_k})\mathbf{1}}{n}\right\|^2\\
+&\left(2L^2+\frac{12\alpha L^3}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n}-\boldsymbol{e}_i\right)\right\|^2+\frac{(\sigma^2+6\varsigma^2)\alpha L}{n}
\end{aligned}
$$

That completes the proof. □

Lemma H.5. *For all $k\geq 0$, we have*

$$
\begin{aligned}
&\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|(\boldsymbol{X}_k-\boldsymbol{X}_{k-\tau_k})\frac{\mathbf{1}}{n}\right\|^2\\
\leq&\frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2}+\frac{24L^4\alpha^2T^2}{n^2K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{X}_{k-\tau_k}\left(\frac{\mathbf{1}}{n}-e_i\right)\right\|^2\\
&+\frac{4\alpha^2T^2L^2}{n^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i\boldsymbol{g}_{k-\tau_k,i}\right\|^2
\end{aligned}
$$

Proof. From Lemma H.4, we know the fact

$$
\overline{\boldsymbol{X}}_{k+1}=\boldsymbol{X}_k\boldsymbol{W}_k\frac{\mathbf{1}}{n}+(\hat{\boldsymbol{X}}_k-\boldsymbol{X}_k)(\boldsymbol{W}_k-\boldsymbol{I})\frac{\mathbf{1}}{n}-\alpha\overline{\tilde{\boldsymbol{G}}}_{k-\tau_k}=\overline{\boldsymbol{X}}_k-\alpha\overline{\tilde{\boldsymbol{G}}}_{k-\tau_k}
$$

As a result

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}\left\|(\boldsymbol{X}_k-\boldsymbol{X}_{k-\tau_k})\frac{\mathbf{1}}{n}\right\|^2\\
=&\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{t=1}^{\tau_k}\alpha\tilde{\boldsymbol{G}}_{k-t}\frac{\mathbf{1}}{n}\right\|^2\\
\leq&\alpha^2\sum_{k=0}^{K-1}\tau_k\sum_{t=1}^{\tau_k}\mathbb{E}\left\|\tilde{\boldsymbol{G}}_{k-t}\frac{\mathbf{1}}{n}\right\|^2\\
\leq&\alpha^2\sum_{k=0}^{K-1}\tau_k\sum_{t=1}^{\tau_k}\left(\frac{\sigma^2}{n^2}+\frac{1}{n^2}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{g}_{k-t,i}\right\|^2\right)\\
\leq&\frac{\alpha^2T^2\sigma^2K}{n^2}+\frac{\alpha^2T}{n^2}\sum_{k=0}^{K-1}\sum_{t=1}^{\tau_k}\sum_{i=1}^{n}p_i\mathbb{E}\left\|\boldsymbol{g}_{k-t,i}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{\alpha^2 T^2 \sigma^2 K}{n^2} + \frac{\alpha^2 T}{n^2} \sum_{k=0}^{K-1} \sum_{t=1}^{\tau_k} \left(12L^2 \sum_{i=1}^{n} p_i \mathbb{E} \left\| \boldsymbol{X}_{k-t} \left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i \right) \right\|^2 + 6\varsigma^2 + 2\mathbb{E} \left\| \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-t,i} \right\|^2 \right) \\
&\leq \frac{\alpha^2 T^2 \sigma^2 K}{n^2} + \frac{\alpha^2 T^2}{n^2} \sum_{k=0}^{K-1} \left(12L^2 \sum_{i=1}^{n} p_i \mathbb{E} \left\| \boldsymbol{X}_{k-\tau_k} \left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i \right) \right\|^2 + 6\varsigma^2 + 2\mathbb{E} \left\| \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i} \right\|^2 \right) \\
&= \frac{\alpha^2 T^2 (\sigma^2 + 6\varsigma^2) K}{n^2} + \frac{12L^2 \alpha^2 T^2}{n^2} \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| \boldsymbol{X}_{k-\tau_k} \left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i \right) \right\|^2 \\
&\quad + \frac{2\alpha^2 T^2}{n^2} \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i} \right\|^2
\end{aligned}
$$

And we get

$$
\begin{aligned}
&\frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| (\boldsymbol{X}_k - \boldsymbol{X}_{k-\tau_k}) \frac{\mathbf{1}}{n} \right\|^2 \\
\leq& \frac{2\alpha^2 T^2 (\sigma^2 + 6\varsigma^2) L^2}{n^2} + \frac{24 L^4 \alpha^2 T^2}{n^2 K} \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| \boldsymbol{X}_{k-\tau_k} \left(\frac{\mathbf{1}}{n} - \boldsymbol{e}_i \right) \right\|^2 \\
&+ \frac{4\alpha^2 T^2 L^2}{n^2 K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i \boldsymbol{g}_{k-\tau_k,i} \right\|^2
\end{aligned}
$$

That completes the proof. □

Lemma H.6. *Given non-negative sequences $\{a_t\}_{t=1}^{\infty}$, $\{b_t\}_{t=1}^{\infty}$ and $\{\tau_t\}_{t=1}^{\infty}$ and a positive number T that satisfying*

$$
a_t = \sum_{s=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor} b_s
$$

with $0 \leq \rho < 1$,we have

$$
\begin{aligned}
S_k &= \sum_{t=1}^{k} a_t \leq \frac{(2-\rho)T}{1-\rho} \sum_{s=1}^{k} b_s \\
D_k &= \sum_{t=1}^{k} a_t^2 \leq \frac{(2-\rho)T^2}{(1-\rho)^2} \sum_{s=1}^{k} b_s^2
\end{aligned}
$$

Proof.

$$
\begin{aligned}
S_k &= \sum_{t=1}^{k} a_t = \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor} b_s \leq \sum_{t=1}^{k} \sum_{s=1}^{t} \rho^{\max\left(\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor, 0\right)} b_s = \sum_{s=1}^{k} \sum_{t=s}^{k} \rho^{\max\left(\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor, 0\right)} b_s \\
&= \sum_{s=1}^{k} \sum_{t=0}^{k-\tau_k-s} \rho^{\left\lfloor \frac{t}{T} \right\rfloor} b_s + \sum_{s=1}^{k} \sum_{t=1}^{\tau_k} \rho^0 b_s \leq \sum_{s=1}^{k} \left(\sum_{t=0}^{T-1} \sum_{m=0}^{\infty} \rho^m \right) b_s + \tau_k \sum_{s=1}^{k} b_s \leq \left(T + \frac{T}{1-\rho} \right) \sum_{s=1}^{k} b_s \\
D_k &= \sum_{t=1}^{k} a_t^2 = \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor} b_s \sum_{r=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-r}{T} \right\rfloor} b_r = \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_t} \sum_{r=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor + \left\lfloor \frac{t-\tau_t-r}{T} \right\rfloor} b_s b_r \\
&\leq \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_t} \sum_{r=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor + \left\lfloor \frac{t-\tau_t-r}{T} \right\rfloor} \frac{b_s^2 + b_r^2}{2} = \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_t} \sum_{r=1}^{t-\tau_t} \rho^{\left\lfloor \frac{t-\tau_t-s}{T} \right\rfloor + \left\lfloor \frac{t-\tau_t-r}{T} \right\rfloor} b_s^2
\end{aligned}
$$

$$\leq \sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t} b_s^2 \rho^{\lfloor \frac{t-\tau_t-s}{T} \rfloor} \sum_{r=1}^{t-\tau_t} \rho^{\lfloor \frac{t-\tau_t-r}{T} \rfloor} \leq \sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t} b_s^2 \rho^{\lfloor \frac{t-\tau_t-s}{T} \rfloor} \sum_{r=0}^{T-1}\sum_{m=0}^{\infty} \rho^m$$

$$cs6 \leq \frac{T}{1-\rho}\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t} \rho^{\lfloor \frac{t-\tau_t-s}{T} \rfloor} b_s^2 \overset{\text{Using} S_k}{\leq} \frac{(2-\rho)T^2}{(1-\rho)^2}\sum_{s=1}^{k} b_s^2$$

□

Lemma H.7. *for* $\forall i, j$ *and* $\forall k \geq 0$*, we have*

$$\|\boldsymbol{X}_k(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < \theta = 16 t_{\text{mix}} \alpha G_\infty$$

Proof. We use mathmatical induction to prove this.

I. First, for $k = 0$, we have

$$\|\boldsymbol{X}_k(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty = 0 < \theta = 16 t_{\text{mix}} \alpha G_\infty$$

II. Suppose for $k \geq 0$, we have $\|\boldsymbol{X}_t(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty < \theta, \forall t \leq k$, then we have

$$\begin{aligned}
&\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_\infty \\
\leq &\left\|\boldsymbol{X}_{k+1}\left(\frac{\boldsymbol{1}}{n} - \boldsymbol{e}_i\right)\right\|_\infty + \left\|\boldsymbol{X}_{k+1}\left(\frac{\boldsymbol{1}}{n} - \boldsymbol{e}_j\right)\right\|_\infty \\
\leq &\left\|\boldsymbol{X}_{k+1}\left(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{n}\right)\right\|_{1,\infty} \|\boldsymbol{e}_i\|_1 + \left\|\boldsymbol{X}_{k+1}\left(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{n}\right)\right\|_{1,\infty} \|\boldsymbol{e}_j\|_1 \\
= &2\left\|\boldsymbol{X}_{k+1}\left(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{n}\right)\right\|_{1,\infty} \\
\leq &2\left\|\left(\boldsymbol{X}_k\boldsymbol{W}_k - \alpha\tilde{\boldsymbol{G}}_{k-\tau_k} + \boldsymbol{\Omega}_k\right)\left(\frac{\boldsymbol{1}}{n} - \boldsymbol{e}_i\right)\right\|_{1,\infty} \\
= &2\left\|\sum_{t=0}^{k}\left(-\alpha\tilde{\boldsymbol{G}}_{t-\tau_t} + \boldsymbol{\Omega}_t\right)\left(\prod_{q=t+1}^{k}\boldsymbol{W}_q - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{n}\right)\right\|_{1,\infty} \\
\leq &2\sum_{t=0}^{k}\left\|\left(-\alpha\tilde{\boldsymbol{G}}_{t-\tau_t} + \boldsymbol{\Omega}_t\right)\left(\prod_{q=t+1}^{k}\boldsymbol{W}_q - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{n}\right)\right\|_{1,\infty} \\
\leq &2\sum_{t=0}^{k}\left\|-\alpha\tilde{\boldsymbol{G}}_{t-\tau_t} + \boldsymbol{\Omega}_t\right\|_{1,\infty}\left\|\prod_{q=t+1}^{k}\boldsymbol{W}_q - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{n}\right\|_1 \\
\leq &4(\alpha G_\infty + 2\delta B_\theta)\sum_{t=0}^{k} 2^{-\lfloor (k-t)/t_{\text{mix}} \rfloor} \\
< &4(\alpha G_\infty + 2\delta B_\theta)\sum_{t=0}^{t_{\text{mix}}-1}\sum_{r=0}^{\infty} 2^{-r} \\
\leq &8(\alpha G_\infty + 2\delta B_\theta) t_{\text{mix}}
\end{aligned}$$

Put in $\delta = \frac{1}{64 t_{\text{mix}}+2}$, we obtain

$$\|\boldsymbol{X}_{k+1}(\boldsymbol{e}_i - \boldsymbol{e}_j)\|_2 < 8(\alpha G_\infty + 2\delta B_\theta) t_{\text{mix}} = 8 t_{\text{mix}} \alpha G_\infty + 8 t_{\text{mix}} \alpha G_\infty = 16 t_{\text{mix}} \alpha G_\infty$$

Combining I and II and we complete the proof. □