MONIQUA: MODULO QUANTIZED COMMUNICATION IN DECENTRALIZED SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Decentralized stochastic gradient descent (SGD), where parallel workers are connected to form a graph and communicate adjacently, has shown promising results both theoretically and empirically. In this paper we propose Moniqua, a technique that allows decentralized SGD to use quantized communication. We prove in theory that Moniqua communicates a provably bounded number of bits per iteration, while converging at the same asymptotic rate as the original algorithm does with full-precision communication. Moniqua improves upon prior works in that it (1) requires no additional memory, (2) applies to non-convex objectives, and (3) supports biased or linear quantizers. We demonstrate empirically that Moniqua converges faster with respect to wall clock time than other quantized decentralized algorithms. We also show that Moniqua is robust to very low bit-budgets, allowing less than 4-bits-per-parameter communication without affecting convergence when training VGG16 on CIFAR10.

1 INTRODUCTION

Stochastic gradient descent (SGD), as a widely adopted optimization algorithm for machine learning, has shown promising performance when running at large scale (Zhang, 2004; Bottou, 2010; Dean et al., 2012; Goyal et al., 2017). However, the communication bottleneck among workers[1] when running distributed SGD presents a non-trivial challenge (Alistarh, 2018). State-of-the-art frameworks such as TensorFlow (Abadi et al., 2016), CNTK (Seide and Agarwal, 2016) and MXNet (Chen et al., 2015) are built in a centralized fashion, where workers exchange gradients either via a centralized parameter server (Li et al., 2014a;b) or the MPI AllReduce operation (Gropp et al., 1999). Such a design, however, puts heavy pressure on the central server and strict requirements on the underlying network. In other words, when the underlying network is poorly constructed, i.e. high latency or low bandwidth, it can easily cause degradation of training performance due to communication congestion in the central server or stragglers (slow workers) in the system.

There are two general approaches to deal with these problems: (1) decentralized training (Lian et al., 2017a;b; Tang et al., 2018a; Hendrikx et al., 2018) and (2) quantized communication[2] (Zhang et al., 2017; Alistarh et al., 2017; Wen et al., 2017). In decentralized training, all the workers are connected to form a graph and each worker communicates only with adjacent workers by averaging model parameters. This balances load and is robust to scenarios where workers can only be partially connected or the communication latency is high. On the other hand, quantized communication reduces the amount of data exchanged among workers, which leads to faster convergence with respect to wall clock time (Alistarh et al., 2017; Seide et al., 2014; Doan et al., 2018; Zhang et al., 2017; Wang et al., 2018). This is especially useful when the communication bandwidth is restricted.

At this point, a natural question is: *Can we apply quantized communication to decentralized training, and thus benefit from both of them?* Unfortunately, directly combining them together negatively affects the convergence rate (Tang et al., 2018b). This happens because existing quantization techniques are mostly designed for centralized SGD, where workers communicate via exchanging gradients (Alistarh et al., 2017; Seide et al., 2014; Wangni et al., 2018). Gradients are robust to quantization since they get smaller in magnitude near local optimum and in some sense carry less information, causing quantization error to approach zero (De Sa et al., 2018). In contrast, decentralized workers are communicating model parameters, which do not necessarily approach zero, and so quantization error does not diminish unless precision is explicitly increased (Tang et al., 2018c). Previous work

[1] A worker could refer to any computing unit that is capable of computing, communicating and has local memory such as CPU, GPU, or even a single thread, etc.

[2] These approaches include low-precision, sparsification, and compression techniques more generally.

solved this problem by adding an error tracker to compensate quantization errors (Tang et al., 2019) or adding replicas of neighboring models and focusing on quantizing model difference which does approach zero (Koloskova et al., 2019; Tang et al., 2018b). However, these methods suffer from trade-offs and limitations in that: (1) the extra replicas or error tracking incurs substantial memory overhead that is proportional to model size (more details in Section 2); and (2) these methods are statistically restricted, in the sense that they are either limited to convex problems (Koloskova et al., 2019) or require unbiased or non-linear quantizers (Koloskova et al., 2019; Tang et al., 2018b; 2019).

To address these problems, in this paper we propose Moniqua, an extra-memory-free (details in Section 2) method for decentralized training to use quantized communication. Moniqua supports both biased and linear quantizers, as well as non-convex objectives.

Intuition behind Moniqua. In a communication step of decentralized training, a worker w_1 updates its model parameter m_1 by averaging with a neighboring worker w_2's model parameter m_2: $m_1 \leftarrow \frac{1}{2}(m_1 + m_2)$. Note that $\frac{1}{2}(m_1 + m_2) = m_1 + \frac{1}{2}(m_2 - m_1)$, so averaging is equivalent to letting w_1 obtain $m_2 - m_1$ (same logic for w_2). Since m_1 and m_2 will approach the same local optimum as the algorithm converges, we can expect the higher-order bits of m_1 and m_2 to get close. Then we can save communication by having w_2 not communicate those higher-order bits to w_1. More explicitly, if we know that $\|m_1 - m_2\|_\infty \leq \theta$ for some known parameter θ (later we will show it can be derived in theory), then instead of sending the entire model m_2 which might cause overhead, w_2 can just send its j-th coordinate $(m_2)_j$ as $(m_2)_j \bmod \theta$ ($\forall j \in [d]$). Note that given $\|m_1 - m_2\|_\infty \leq \theta$:

$$(m_2)_j \bmod \theta - (m_1)_j \bmod \theta = ((m_2)_j - (m_2)_j) \bmod \theta = (m_2)_j - (m_2)_j$$

so w_1 can obtain the j-th coordinate of $m_2 - m_1$ by locally computing $(m_2)_j \bmod \theta - (m_1)_j \bmod \theta$ with $(m_2)_j \bmod \theta$ received from w_2. Since $(m_2)_j \bmod \theta$ is generally a smaller number than $(m_2)_j$, w_2 can send fewer bits with the same level of absolute error.

In this paper, we make the following contributions.

- We show by example that directly quantizing communication in decentralized training, even with an unbiased quantizer, can fail to converge asymptotically. (Section 3)
- We propose **Moniqua**, a general algorithm that uses **mo**dular arithmetic for commu**ni**cation **qua**ntization in decentralized training. We prove applying Moniqua achieves the same asymptotic convergence rate as the baseline full-precision algorithm (D-PSGD) while requiring at most $O(\log \log n)$ number of bits per parameter communicated, where n is the number of parallel workers. (Section 4)
- We apply Moniqua to decentralized algorithms with variance reduction and asynchronous communication (D^2 and AD-PSGD) and prove Moniqua enjoys the same asymptotic rate as with full-precision communication when applied to these cases. (Section 5)
- We empirically evaluate Moniqua and show it outperforms all the related algorithms given an identical quantizer. We also show Moniqua is scalable and robust to very low bit-budgets, and we introduce techniques we found empirically useful to run Moniqua even more efficiently. (Section 6)

2 Related Work

Decentralized Stochastic Gradient Descent (SGD) Decentralized algorithms (Mokhtari and Ribeiro, 2015; Sirb and Ye, 2016; Lan et al., 2017; Wu et al., 2018a) have been widely studied with consideration of communication efficiency, privacy and scalability. In the domain of large-scale machine learning, D-PSGD was the first Decentralized SGD algorithm that enjoys the same asymptotic convergence rate $O(1/\sqrt{Kn})$ (where K is the number of total iterations and n is the number of workers) as centralized algorithms (Lian et al., 2017a). After D-PSGD came D^2, which improves D-PSGD and is applicable to the case where workers are not sampling from identical data sources (Tang et al., 2018a). Another extension was AD-PSGD, which lets workers communicate *asynchronously* and has a convergence rate of $O(1/\sqrt{K})$ (Lian et al., 2017b). In this paper we prove that Moniqua is applicable to all of these three algorithms. Other relevant work includes: He et al. (2018), which investigates decentralized learning on linear models; Nazari et al. (2019), which introduces decentralized algorithms with online learning; Zhang and You (2019), which analyzes the case when workers cannot mutually communicate; and Assran et al. (2018), which investigates Decentralized SGD specifically for deep learning.

Quantized Communication in Centralized SGD Prior research on quantized communication is often focused on centralized algorithms, such as randomized quantization (Doan et al., 2018; Suresh et al., 2017; Zhang et al., 2017) and randomized sparsification (Wangni et al., 2018; Stich et al., 2018; Wang et al., 2018; Alistarh et al., 2018). Many examples of prior work focus on studying quantization in the communication of deep learning tasks specifically (Han et al., 2015; Wen et al., 2017; Grubic et al., 2018). Alistarh et al. (2017) proposes QSGD, which uses an encoding-efficient scheme, and discusses its communication complexity. Another method, 1bitSGD, quantizes exchanged gradients with one bit and shows great empirical success on speech recognition (Seide et al., 2014). Other work discusses the convergence rate under sparsified or quantized communication (Jiang and Agrawal, 2018; Stich et al., 2018). Acharya et al. (2019) theoretically analyzes sublinear communication for distributed training.

Quantized Communication in Decentralized SGD Quantized communication for decentralized algorithms is a rising topic in the optimization community. Previous work has proposed decentralized algorithms with quantized communication for strongly convex objectives (Reisizadeh et al., 2018; Koloskova et al., 2019). Following that, Tang et al. (2018b) proposes DCD/ECD-PSGD, which quantizes communication via estimating model difference. Furthermore, Tang et al. (2019) proposes DeepSqueeze, which applies an error-compensation method (Wu et al., 2018b) to decentralized setting. From a systems perspective, Koloskova et al. (2019) and Tang et al. (2018b) require $O(d \cdot l)$ and Tang et al. (2019) requires $O(d)$ extra memory compared to D-PSGD to implement quantized communication, where d denotes the dimension of the model and l denotes the number of connections in the network. In comparison, Moniqua is extra-memory-free.

3 Setting and Notation

In this section, we introduce our notation and the general assumptions we will make about the quantizers for our results to hold. Then we describe D-PSGD (Lian et al., 2017a), the basic algorithm for Decentralized SGD, and we show how naive quantization can fail in decentralized training.

Quantizers. Throughout this paper, we assume that we use a quantizer $\mathcal{Q}_\delta$ that has bounded error

$$\|\mathcal{Q}_\delta(x) - x\|_\infty \leq \delta, \quad \forall x \in [-1,1]^d \tag{1}$$

where δ is some constant. In general, a smaller δ denotes more fine-grained quantization requiring more bits. For example, a biased linear quantizer can achieve (1) by rounding x to the nearest number in the set $\{2\delta n \mid n \in \mathbb{Z}\}$; this will require about δ^{-1} quantization points to cover the interval $[-1,1]$, so such a linear quantizer can satisfy (1) using only $\left\lceil \log_2\left(\frac{1}{\delta}+1\right)\right\rceil$ bits (Li et al., 2017; Gupta et al., 2015). Note that (1) can be satisfied (for appropriate values of δ) by both linear (Gupta et al., 2015; De Sa et al., 2017) and non-linear (Stich, 2018; Alistarh et al., 2017) quantizers, and thus it is more general than assumptions used in previous works where only non-linear quantizers are considered (Koloskova et al., 2019; Tang et al., 2018c; 2019).

Decentralized parallel SGD (D-PSGD). D-PSGD (Lian et al., 2017a) is the first and most basic Decentralized SGD algorithm. In D-PSGD, n workers are connected to form a graph. Each worker i stores a copy of model $x \in \mathbb{R}^d$ and a local dataset $\mathcal{D}_i$ and collaborates to optimize

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_{\xi\sim\mathcal{D}_i} f_i(x;\xi)}_{f_i(x)} \tag{2}$$

where ξ is data sample from $\mathcal{D}_i$. In each iteration of D-PSGD, worker i computes a local gradient sample using $\mathcal{D}_i$. Then it *averages* its model parameters with its neighbors according to a symmetric and doubly stochastic matrix W, where W_{ij} denotes the ratio worker j averages from worker i. Formally: Let $x_{k,i}$ and $\widetilde{g}_{k,i}$ denote local model and sampled gradient on worker i at k-th iteration, respectively. Let α denote the step size. The update rule of D-PSGD can be expressed as:

$$x_{k+1,i} = \sum_{j=1}^{n} x_{k,j}W_{ji} - \alpha\widetilde{g}_{k,i} = x_{k,i} \underbrace{-\sum_{j=1}^{n}(x_{k,i}-x_{k,j})W_{ji}}_{\text{communicate to reduce difference}} \underbrace{-\alpha\widetilde{g}_{k,i}}_{\text{gradient step}} \tag{3}$$

From (3) we can see the update of a single local model contains two parts: communication to reduce model difference and a gradient step. Lian et al. (2017a) shows that all local models in D-PSGD are able to reach the same stationary point.

Failure with direct quantization. Here, we illustrate why directly quantizing communication in decentralized training —naively quantizing the exchanged data—can fail to converge asymptotically even on a simple problem. This naive approach with quantizer $\mathcal{Q}_\delta$ can be represented by

$$x_{k+1,i} = x_{k,i}W_{ii} + \sum\nolimits_{j\neq i} \mathcal{Q}_\delta(x_{k,j})W_{ji} - \alpha\widetilde{g}_{k,i} \tag{4}$$

Based on Equation 4, we obtain the following theorem.

Theorem 1 *For some constant δ, suppose that we use an unbiased linear quantizer $\mathcal{Q}$ with representable points $\{\delta n \mid n \in \mathbb{Z}\}$ to learn on the quadratic objective function $f(x) = (x-\delta/2)^\top(x-\delta/2)/2$ with the direct quantization approach (4). Let ϕ denote the smallest value of a non-zero entry in W. Regardless of what step size we adopt, it will always hold for all iterations k and local model indices i that $\mathbb{E}\left\|\nabla f(x_{k,i})\right\|^2 \geq \frac{\phi^2\delta^2}{8(1+\phi^2)}$. That is, the local iterates will fail to asymptotically converge to a region of small gradient magnitude in expectation.*

4 MONIQUA

Theorem 1 shows that when directly quantizing communication in decentralized SGD, even with an unbiased quantizer, any local model can fail to converge on a simple quadratic objective. In this seciton, we propose a technique, Moniqua, that solves this problem. Moniqua works under the following common assumptions for analyzing decentralized optimization algorithms (Lian et al., 2017a; Tang et al., 2018b; Koloskova et al., 2019).

- **(A1) Lipschitzian gradient.** All the functions f_i have L-Lipschitzian gradients.
$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x-y\|, \forall x, y \in \mathbb{R}^d$$
- **(A2) Spectral gap.** The communication matrix W is a symmetric doubly stochastic matrix and $\max\{|\lambda_2(W)|, |\lambda_n(W)|\} = \rho < 1$, where $\lambda_i(W)$ denotes the ith eigenvalue of W.
- **(A3) Bounded variance.** There exist non-negative σ and ς such that
$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left\|\nabla\widetilde{f}_i(x;\xi_i) - \nabla f_i(x)\right\|^2 \leq \sigma^2, \qquad \mathbb{E}_{i\sim\{1,\cdots,n\}}\left\|\nabla f_i(x) - \nabla f(x)\right\|^2 \leq \varsigma^2$$
where $\nabla\widetilde{f}_i(x;\xi_i)$ denotes gradient sample on worker i computed via data sample ξ_i.
- **(A4) Initialization.** All the local models are initialized by the same weight: $x_{0,i} = x_0$, for all i and without loss of generality $x_0 = 0$.
- **(A5) Bounded gradient magnitude.** The norm of a sampled gradient is bounded by $\|\widetilde{g}_{k,i}\|_\infty \leq G_\infty$, for all i and k with some constant G_∞.

In Section 1, we described how a modulo operation can be used to avoid sending redundant bits if a bound θ on model difference is known. Here we outline how we can obtain such a bound. We do so by leveraging the following insight: in decentralized training, all the workers initialize local models at same point and average with each other periodically. The only difference among their models is caused by the sampled gradients (updated with the step size), and this difference is reduced each time they communicate. Since we have an upper bound on the magnitude of the gradients (A5) as well as a bound characterizing how quickly the communication process converges (A2), we can combine these to get an a priori bound θ on how much the models can differ. We can then pass this bound θ as a parameter to the algorithm, which can proceed to modulo-quantize the communication via the process described in Section 1. We formalize this approach as Moniqua (Algorithm 1).

Algorithm 1 Pseudo-code of Moniqua on worker i

Input: initial point $x_{0,i} = x_0$, step size α, the priori bound θ, communication matrix W, number of iterations K, quantizer $\mathcal{Q}_\delta$, neighbor list $\mathcal{N}_i$

1: **for** $k = 0, 1, 2, \cdots, K-1$ **do**
2: Compute a local stochastic gradient $\widetilde{g}_{k,i}$ with data sample $\xi_{k,i}$ and current weight $x_{k,i}$
3: Compute modulo-ed model: $q_{k,i} \leftarrow \theta \cdot \mathcal{Q}_\delta\left(\frac{x_{k,i}}{\theta} \bmod 1\right)$ (element-wise division and mod)
4: Average with neighboring workers: $x_{k+\frac{1}{2},i} \leftarrow x_{k,i} + \sum_{j\in\mathcal{N}_i}(q_{k,j} - q_{k,i})W_{ji}$
5: Update the local weight with local gradient: $x_{k+1,i} \leftarrow x_{k+\frac{1}{2},i} - \alpha\widetilde{g}_{k,i}$
6: **end for**

Output: Averaged model $\overline{X}_K = \frac{1}{n}\sum_{i=1}^n x_{K,i}$

In line 3 we rescale each coordinate so that the number to be quantized falls in the region of $[-1, 1]$, which is required for (1) to apply. Note that with quantization, the priori bound θ could increase since local models may move further apart due to quantization error. However, with appropriately chosen δ, we can still obtain a bound θ and apply modulo-quantized communication that allows Moniqua to converge. We present these parameter choices in Theorem 2, along with the resulting convergence rate for Moniqua.

Theorem 2 *If we run Algorithm 1 in a setting where*

$$\theta = \frac{2\log(16n)\alpha G_\infty}{1-\rho}, \qquad \delta = \frac{1-\rho}{4\log(16n)}, \quad \textit{and} \qquad \alpha = \frac{1}{\varsigma^{2/3}K^{1/3} + \sigma\sqrt{K/n} + 2L},$$

then the output of Algorithm 1 converges at the asymptotic rate

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \lesssim \frac{1}{K} + \frac{\sigma}{\sqrt{nK}} + \frac{\varsigma^{\frac{2}{3}}}{K^{\frac{2}{3}}} + \frac{\sigma^2 n}{\sigma^2 K + n} + \frac{G_\infty^2 dn}{\sigma^2 K + n}.$$

where ρ, $f(0) - f^*$ and L are omitted as constants.

Consistent with D-PSGD. Note that D-PSGD converges at the asymptotic rate of $O(\sigma/\sqrt{nK} + \varsigma^{\frac{2}{3}}/K^{\frac{2}{3}} + n/K)$, and thus Moniqua has the same asymptotic rate as D-PSGD (Lian et al., 2017a). In other words, the asymptotic convergence rate is not negatively impacted by the quantization.

Robust to large d. In Assumptions (A3) and (A5), we use l_2-norm and l_∞-norm to bound sample variance and gradient magnitude, respectively. Note that, when d gets larger, the variance σ^2 will also grow proportionally. So, the last term will tend to remain n/K asymptotically with large d.

How many bits does Moniqua need? The specific number of bits required by Moniqua depends on the underlying quantizer ($\mathcal{Q}_\delta$). If we use nearest rounding (Gupta et al., 2015) as $\mathcal{Q}_\delta$ in Theorem 2, it suffices to use at each step a number of bits $\mathcal{B}$ for each parameter sent, where

$$\mathcal{B} = \left\lceil \log_2\left(\tfrac{1}{\delta} + 1\right)\right\rceil = \left\lceil \log_2\left(\tfrac{4\log_2(16n)}{1-\rho} + 1\right)\right\rceil$$

Note that this bound is independent of model dimension d. When the system scales up, the number of required bits grows at a rate of $O(\log\log n)$.

5 Scalable Moniqua

Previous work has extended D-PSGD to D^2 (Tang et al., 2018a) (to make Decentralized SGD applicable to workers sampling from different data sources) and AD-PSGD (Lian et al., 2017b) (an asynchronous version of D-PSGD). In this section, we theoretically prove Moniqua is applicable to both of these algorithms.

Moniqua with Decentralized Data Decentralized data refers to the case where all the local datasets $\mathcal{D}_i$ are not identically distributed (Tang et al., 2018a). More explicitly, the outer variance $\mathbb{E}_{i\sim\{1,\cdots,n\}}\left\|\nabla f_i(x) - \nabla f(x)\right\|^2$ is no longer bounded by ς^2 as assumed in D-PSGD (Assumption (A3)). This update rule presented can be explicitly expressed in two steps[3]:

$$X_{k+\frac{1}{2}} = 2X_k - X_{k-1} - \alpha\widetilde{G}_k + \alpha\widetilde{G}_{k-1}$$
$$X_{k+1} = X_{k+\frac{1}{2}}W + (Q_k - X_{k+\frac{1}{2}})(W - I)$$

where X_k, $\widetilde{G}_k$ and Q_k are matrix in the shape of $\mathbb{R}^{d\times n}$, where their i-th column are $x_{k,i}$, $\widetilde{g}_{k,i}$ and $q_{k,i}$ respectively. And X_{-1} and $\widetilde{G}_{-1}$ are $0^{d\times n}$ by convention. Based on this, we obtain the following convergence theorem.

Theorem 3 *If we run D^2 with Monqiua in a setting where*

$$\theta = (6D_1 n + 8)\alpha G_\infty, \qquad \delta = \frac{1}{6nD_2}, \quad \textit{and} \qquad \alpha = \frac{1}{\sigma\sqrt{K/n} + 2L},$$

where D_1 and D_2 are two constants that only depend on the eigenvalues of W (definition can be found in supplementary material), the output has the following asymptotic convergence rate:

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \lesssim \frac{1}{K} + \frac{\sigma}{\sqrt{nK}} + \frac{\sigma^2 n}{\sigma^2 K + n} + \frac{G_\infty^2 dn}{\sigma^2 K + n}.$$

[3]Detailed pseudo-code in the supplementary material.

Note that D^2 (Tang et al., 2018a) with full-precision communication has the asymptotic convergence rate of $O\left(1/K + \sigma/\sqrt{nK} + n/K\right)$, Moniqua on D^2 has the same asymptotic rate.

Moniqua with Asychronous Communication. Both D-PSGD and D^2 are synchronous algorithms as they require global synchronization at the end of each iteration, which can become a bottleneck when such synchronization is not cheap. Another algorithm, AD-PSGD, avoids this overhead by letting workers communicate asynchronously (Lian et al., 2017b). In the analysis of AD-PSGD, an iteration represents a *single* gradient update on *one* randomly-chosen worker, rather than a synchronous bulk update of all the workers. This single-worker-update analysis models the asynchronous nature of the algorithm. We apply Moniqua to AD-PSGD and obtain the following update rule[4]:

$$X_{k+1} = X_k W_k + (Q_k - X_k)(W_k - I) - \alpha \widetilde{G}_{k-\tau_k}$$

where W_k describes the communication behaviour between the kth and $(k+1)$th gradient update, and τ_k denotes the delay (measured as a number of iterations) between when the gradient is computed and updated to the model. Note that unlike D-PSGD, here W_k can be different at each update step and usually each individually has $\rho = 1$, 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},

$$\forall \mu \in \mathbb{R}^n,\ \forall k \in \mathbb{N},\ \text{if } \mu_i \geq 0 \text{ and } \mathbb{1}^\top \mu = 1,\ \text{it must hold that } \left\| \left(\prod_{i=1}^{t_{\text{mix}}} W_{k+i} \right) \mu - \tfrac{\mathbb{1}}{n} \right\|_1 \leq \tfrac{1}{2}.$$

We call this constant t_{mix} because it is effectively the *mixing time* of the time-inhomogeneous Markov chain with transition probability matrix W_k at time k (Levin and Peres, 2017). Note that this condition is more general than those used in previous work on AD-PSGD because it does not require that the W_k are sampled independently or in an unbiased manner. Using this, we obtain the following convergence theorem.

Theorem 4 *If we run AD-PSGD with Moniqua in a setting where*

$$\theta = 16 t_{\text{mix}} \alpha G_\infty, \qquad \delta = \frac{1}{32 t_{\text{mix}}}, \quad \textit{and} \qquad \alpha = \frac{n}{2L + \sqrt{K(\sigma^2 + 6\varsigma^2)}},$$

the output has the following asymptotic convergence rate:

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

Note that AD-PSGD (Lian et al., 2017b) with full-precision communication has the asymptotic convergence rate of $O\left(1/K + \sqrt{\sigma^2 + 6\varsigma^2}/\sqrt{K} + n^2/K\right)$, Moniqua converges at the same rate.

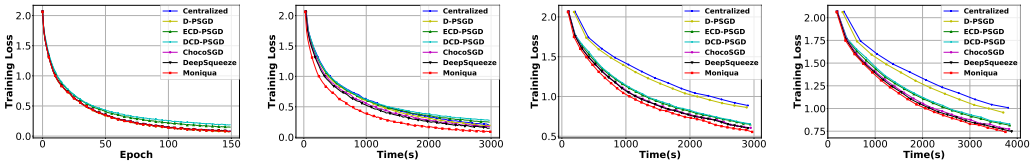
6 Experiments

In this section, we evaluate Moniqua empirically. First, we compare Moniqua and other quantized decentralized training algorithms' convergence under different network configurations. Second, we evaluate Moniqua's scalability on D^2 and AD-PSGD. Third, we introduce two additional techniques to run Moniqua more efficiently and empirically investigate the limits of Moniqua.

Configuration. All the models and training scripts in this section are implemented in PyTorch and run on Google Cloud Platform. We launch an instance as one worker, each configured with a 2-core CPU with 4 GB memory and an NVIDIA Tesla P100 GPU. We use MPICH as the communication backend. All the instances are running Ubuntu 16.04, and latency and bandwidth on the underlying network are configured using the `tc` command in Linux. In all the experiments, we use the following hyperparameters by default: batch size $= 128$, weight decay $= 1e-4$, and momentum $= 0.9$, which are default values adopted in previous works (Lian et al., 2017b; Grubic et al., 2018). We tune the step size from set $\{0.5, 0.1, 0.05, 0.01\}$ for each algorithm. Throughout our experiments, we adopt the commonly used (Gupta et al., 2015; Li et al., 2017) stochastic rounding[5] with quantization step δ.

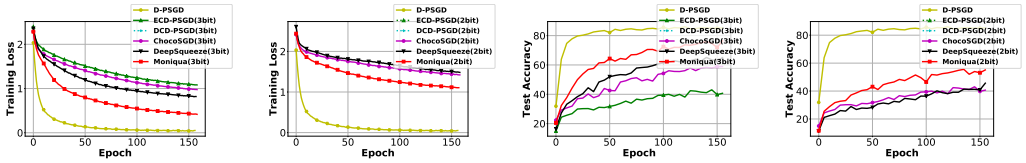
[4] Details in the supplementary material.

[5] Since several baselines are not applicable to biased quantizers, for fair comparison we consistently use stochastic rounding (unbiased). More experiments using different quantizers including biased and non-linear quantizers on Moniqua can be found in supplementary material.



(a) Training Loss vs Epoch (b) Bandwidth=100Mbps, Latency=0.15ms (c) Bandwidth=20Mbps, Latency=0.15ms (d) Bandwidth=100Mbps, Latency=10ms

Figure 1: Performance of different algorithms under different network configurations



(a) 3-Bit Training Loss (b) 2-Bit Training Loss (c) 3-Bit Test Accuracy (d) 2-Bit Test Accuracy

Figure 2: Performance of Moniqua and other quantization algorithms under extreme bit-budget.

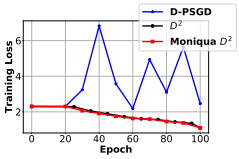
Wall-clock Time Evaluation. We start by evaluating the performance of Moniqua and other baseline algorithms under different network configurations. We launch 8 workers connected in a ring topology and train a ResNet110 (He et al., 2016) model on CIFAR10 (Krizhevsky et al., 2014). We compare Moniqua with the following baselines:[6] Centralized (implemented as a standard AllReduce operation), D-PSGD (Lian et al., 2017a) with full-precision communication, DCD/ECD-PSGD (Tang et al., 2018b), ChocoSGD (Koloskova et al., 2019) and DeepSqueeze (Tang et al., 2019). We set $\delta = 0.01$ for stochastic rounding across all algorithms that use quantization. To prevent overflow, we use 16-bit integers[7] `torch.int16` as the floored output on the sender side. For Moniqua, we set $\theta = 3.0$.

We plot our results in Figure 1. As can be seen in Figure 1(a), with respect to epochs, All the algorithms have similar convergence curve while DCD/ECD-PSGD have slightly slower convergence curves. We can see from Figures 1(b) and 1(c) that when the network bandwidth decreases, the curves begin to separate. AllReduce and full-precision D-PSGD suffer the most, since they require a large volume of high-precision exchanged data. And from Figure 1(b) to Figure 1(d), when the network latency increases, we observe similar behavior. On the other hand, from Figure 1(b) to Figure 1(c) and Figure 1(d), curves of all the quantized baselines (DCD/ECD-PSGD, ChocoSGD and DeepSqueeze) are getting closer to Moniqua. This is because, as shown in Figure 1(b), the extra updating of the replicas in DCD/ECD-PSGD and ChocoSGD as well as the error tracking in DeepSqueeze counteract the benefits from accelerated communication. However, when network bandwidth decreases or latency increases, communication becomes the bottleneck and makes these algorithms diverge from centralized SGD and D-PSGD. Delay between Moniqua and quantized baselines does not vary with the network since that only depends on the their extra local computation (error tracking and replica update). We observe that compared to Moniqua, DCD/ECD-PSGD is approximately 13 seconds slower while ChocoSGD and DeepSqueeze being 10 and 8 seconds slower repectively. From Figure 1 we can see that Moniqua outperforms all these other algorithms.

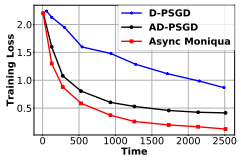
Aggressive Quantization Now we investigate how Moniqua and baselines behave under aggressive quantization. We enforce two strict bit-budget: 2bit and 3bit (per parameter). We plot the results in Figure 2. We can see that DCD-PSGD fails to converge in both cases and ECD-PSGD fails to converge with 2bit. This is consistent with results in previous work (Tang et al., 2018c; 2019). On the other hand, Moniqua converges faster than any other baselines. We observe at the end of 150 epoch, with 3-bit communication Moniqua achieves 85% training accuracy while other baselines are below

[6]Other algorithms are not applicable to non-convex DNN problems, so we are not comparing them here.

[7]Since we are measuring the system performance, the specific number of bits is not the focus here. In later section we will discuss statistical performance with small number of bits.

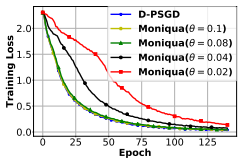


(a) Training Loss (Decentralized Data)

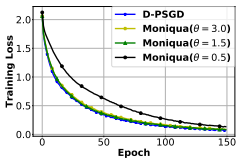


(b) Training Loss (Asynchronous Communication)

Figure 3: Performance of applying Moniqua on D^2 and AD-PSGD



(a) Training Loss (VGG16)



(b) Training Loss (ResNet110)

Figure 4: Performance of Moniqua on VGG16 and ResNet110 under different θ

70% (full precision achieves 97%). Compared to the theoretical results in Section 4, we show that Moniqua is much more robust to low-bits budget in practice.

Scalability of Moniqua. We evaluate how Moniqua can be applied to D^2 (Tang et al., 2018a) and AD-PSGD (Lian et al., 2017b). First, we demonstrate how applying Moniqua to D^2 can handle decentralized data. We launch 10 workers, collaborating to train a VGG16 (Simonyan and Zisserman, 2014) model on CIFAR10. Similar to the setting of D^2 (Tang et al., 2018a), we let each worker have exclusive access to 1 labels (of the 10 labels total in CIFAR10). In this way, the data variance among workers is maximized. We plot the results in Figure 3(a). We observe that applying Moniqua on D^2 does not affect the convergence rate while D-PSGD can no longer converge because of the outer variance. Here we omit the wall clock time comparison since the communication volume is the same in comparison of Moniqua and Centralized algorithm in Figure 1.

Next, we evaluate Moniqua on AD-PSGD. We launch 6 workers organized in a ring topology, collaborating to train a ResNet110 model on CIFAR10. We set the network bandwidth to be 20Mbps and latency to be 0.15ms. We plot the results in Figure 3(b). We can see that both AD-PSGD and asynchronous Moniqua outperform D-PSGD. Besides, Moniqua outperforms AD-PSGD in that communication is reduced, which is aligned with the intuition and theory.

Efficient Moniqua. There are two techniques we have observed to improve the performance of Moniqua when using stochastic rounding: $\mathcal{Q}_\delta(x) = \delta \lfloor \frac{x}{\delta} + u \rfloor$ (where u is uniformly sampled from $[0, 1]$), $\forall x \in \mathbb{R}^d$. The **first** is to use *shared randomness*, in which the same random seed is used for stochastic rounding on all the workers. That is, if two workers are exchanging tensors x and y respectively, then the floored tensors $\lfloor \frac{x}{\delta} + u \rfloor$ and $\lfloor \frac{y}{\delta} + u \rfloor$ they send use the *same* randomly sampled value u. This provably reduces the error due to quantization (more details are in the supplementary material). The **second** technique is to use a standard entropy compressor like `bzip` to further compress the communicated tensors. This can help further reduce the number of bits because the modulo operation in Moniqua can introduce some redundancy in the higher-order bits, which a traditional compression algorithm can easily remove.

To evaluate these methods, we train both ResNet110 and VGG16 on CIFAR10 using 8 ring-connected workers. We plot the training loss under different θ in Figure 4 (with $\delta = 0.01$ for stochastic rounding). Note that for VGG16, it can tolerate small $\theta = 0.08$ while still preserving the convergence rate. On the other hand, for ResNet110, it begins to diverge when θ decreases to 0.5. This is because VGG16 has more fully connected layers than ResNet110, and these layers are less sensitive to quantization, as claimed in (Grubic et al., 2018). We observed that the fewest number of bits per number needed to communicate by Moniqua for VGG16 and ResNet110 to guarantee convergence (accuracy loss $< 0.3\%$, criterion adopted by (Grubic et al., 2018)) are 3.64 and 5.67, respectively (details in the supplementary material).

7 Conclusions

In this paper we propose Moniqua, a simple unified method of quantizing the communication in decentralized training algorithms. Theoretically, Moniqua supports biased quantizer and non-convex problems, while enjoying the same asymptotic convergence rate as full-precision-communication algorithms without incurring storage or computation overhead. Empirically, we observe Moniqua converges faster than other related algorithms with respect to wall clock time. Additionally, Moniqua is robust to very low bits-budget.

REFERENCES

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116. ACM, 2004.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.

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

Dan Alistarh. A brief tutorial on distributed and concurrent machine learning. In *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing*, pages 487–488. ACM, 2018.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

Frank Seide and Amit Agarwal. Cntk: Microsoft's open-source deep-learning toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2135–2135. ACM, 2016.

Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *OSDI*, volume 14, pages 583–598, 2014a.

Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pages 19–27, 2014b.

William Gropp, Rajeev Thakur, and Ewing Lusk. *Using MPI-2: Advanced features of the message passing interface*. MIT press, 1999.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017a.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. *arXiv preprint arXiv:1710.06952*, 2017b.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D2: Decentralized training over decentralized data. *arXiv preprint arXiv:1803.07068*, 2018a.

Hadrien Hendrikx, Laurent Massoulié, and Francis Bach. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives. *arXiv preprint arXiv:1810.02660*, 2018.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. Zipml: Training linear models with end-to-end low precision, and a little bit of deep learning. In *International Conference on Machine Learning*, pages 4035–4043, 2017.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems*, pages 1509–1519, 2017.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. On the convergence of distributed subgradient methods under quantization. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 567–574. IEEE, 2018.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, pages 9850–9861, 2018.

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *Advances in Neural Information Processing Systems*, pages 7663–7673, 2018b.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1306–1316, 2018.

Christopher De Sa, Megan Leszczynski, Jian Zhang, Alana Marzoev, Christopher R Aberger, Kunle Olukotun, and Christopher Ré. High-accuracy low-precision training. *arXiv preprint arXiv:1803.03383*, 2018.

Hanlin Tang, Chen Yu, Cedric Renggli, Simon Kassing, Ankit Singla, Dan Alistarh, Ji Liu, and Ce Zhang. Distributed learning over unreliable networks. *arXiv preprint arXiv:1810.07766*, 2018c.

Hanlin Tang, Xiangru Lian, Shuang Qiu, Lei Yuan, Ce Zhang, Tong Zhang, and Ji Liu. Deepsqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. *arXiv preprint arXiv:1907.07346*, 2019.

Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. *arXiv preprint arXiv:1902.00340*, 2019.

Aryan Mokhtari and Alejandro Ribeiro. Decentralized double stochastic averaging gradient. In *Signals, Systems and Computers, 2015 49th Asilomar Conference on*, pages 406–410. IEEE, 2015.

Benjamin Sirb and Xiaojing Ye. Consensus optimization with delayed and stochastic gradients on decentralized networks. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 76–85. IEEE, 2016.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *arXiv preprint arXiv:1701.03961*, 2017.

Tianyu Wu, Kun Yuan, Qing Ling, Wotao Yin, and Ali H Sayed. Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):293–307, 2018a.

Lie He, An Bian, and Martin Jaggi. Cola: Decentralized linear learning. In *Advances in Neural Information Processing Systems*, pages 4541–4551, 2018.

Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *arXiv preprint arXiv:1901.09109*, 2019.

Jiaqi Zhang and Keyou You. Asynchronous decentralized optimization in directed networks. *arXiv preprint arXiv:1901.08215*, 2019.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Michael Rabbat. Stochastic gradient push for distributed deep learning. *arXiv preprint arXiv:1811.10792*, 2018.

Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3329–3337. JMLR. org, 2017.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4452–4463, 2018.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5973–5983, 2018.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

D Grubic, L Tam, Dan Alistarh, and Ce Zhang. Synchronous multi-gpu deep learning with low-precision communication: An experimental study. *Proceedings of the EDBT 2018*, 2018.

Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Advances in Neural Information Processing Systems*, pages 2525–2536, 2018.

Jayadev Acharya, Christopher De Sa, Dylan J Foster, and Karthik Sridharan. Distributed learning with sublinear communication. *arXiv preprint arXiv:1902.11259*, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, S. Hamed Hassani, and Ramtin Pedarsani. Quantized decentralized consensus optimization. *CoRR*, abs/1806.11536, 2018. URL `http://arxiv.org/abs/1806.11536`.

Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized sgd and its applications to large-scale distributed optimization. *arXiv preprint arXiv:1806.08054*, 2018b.

Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems*, pages 5811–5821, 2017.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning*, pages 1737–1746, 2015.

Christopher De Sa, Matthew Feldman, Christopher Ré, and Kunle Olukotun. Understanding and optimizing asynchronous low-precision stochastic gradient descent. In *ACM SIGARCH Computer Architecture News*, volume 45, pages 561–574. ACM, 2017.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Supplementary Material

A Overview

This supplementary material contains proofs of all the theoretical results and extra experimental results of Moniqua. It is organized as follows: In Section B, we provably explain why using shared randomness in communication with stochastic rounding can improve performance (theoretical explanation for technique 1 in Experiment *Efficient Moniqua*). Then we demonstrate more experimental results in Section C. In Section D, we illustrate why naively 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, 3 and 4 from Section F to H, with corollaries contained in the corresponding sections. Detailed algorithm statements for applying Moniqua on D^2 and AD-PSGD can be found in Section G Algorithm 2 and Section H Algorithm 3, respectively.

B Shared Randomness (Experiment of *Efficient Moniqua*)

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 $\mathcal{Q}$ be $\delta = 1$. For any z quantized using $\mathcal{Q}$, let $z_f = z - \lfloor z \rfloor$, the variance of quantization error can be expressed as

$$\mathbb{E}\left\|\mathcal{Q}(z) - z\right\|^2 = (1 - z_f)(-z_f)^2 + z_f(1 - z_f)^2 = z_f(1 - z_f) \tag{5}$$

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

$$\mathbb{E}\left\|(q_{k,j} - x_{k,j}) - (q_{k,i} - x_{k,i})\right\|^2$$

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

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

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

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

Then we have

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

The last equality holds due to equation 5. Next, let

$$\begin{aligned}
\Delta &= y - x \\
r &= \mathcal{Q}(\Delta) - \Delta
\end{aligned}$$

And let r_h denote h-th entry of r, let Δ_h denote h-th entry of Δ. We obtain

$$r_h = \mathcal{Q}(\Delta_h) - \Delta_h$$

$$
\begin{aligned}
=&\delta \begin{cases} -\frac{\Delta_h}{\delta} + \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor + 1, & p_t \leq \frac{\Delta_h}{\delta} - \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor \\ -\frac{\Delta_h}{\delta} + \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor, & \text{otherwise} \end{cases} \\
=&\delta \begin{cases} -q+1, & p_t \leq q \\ -q, & \text{otherwise} \end{cases}
\end{aligned}
$$

where

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

Based on that, we have

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

Since $\min\{q, 1-q\} \leq \left|\frac{x_h}{\delta}\right|$, we have

$$
\mathbb{E}\left[r_h^2\right] \leq \delta^2 \left|\frac{\Delta_h}{\delta}\right| \leq \delta \left|\Delta_h\right|
$$

Summing over the index h yields,

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

Pushing back x and r, we have

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

Putting it back we have

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

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.

C More Experimental Results

C.1 Compute Number of Bits

In Experiment of *Efficient Moniqua*, we calculate the number of bits in the following way: First, we calculate the total number of bits each worker send out, sum them up and divided by number of epochs, and we get the average bandwidth consumption $\overline{BW}$ of the whole system in each epoch. Then we compute the number of bits required for each number in the following way (note that every worker has 2 neighbors in a ring topology):

$$
\#bits = \frac{\overline{BW}}{\text{\#neighbors} \cdot \text{\#workers} \cdot \text{\#params of model}}
$$

In our experiments, #neighbors=2, #workers=8. For VGG16, #params of model= 15,245,130 while for ResNet110, #params of model=1,146,842. We formalize the results in Table C.1 [8].

C.2 Various Quantizers

In this section, we will verify Moniqua is applicable to other quantizers aside from linear quantizer as shown in the paper. We test it on two more quantizers:

1. **Nearest Rounding (Biased)**

$$
\mathcal{Q}(x) = \delta \left\lfloor \frac{x}{\delta} + 0.5 \right\rfloor
$$

where δ is the quantization step as defined in the linear quantizer. In this experiment, we set $\delta = 0.01$, the same value as we used in the paper with stochastic rounding.

[8] Note that we only put results that's 'close to the limit of Moniqua' here

Table 1: Wall Clock Time consumption (Seconds)/Epoch in average under different network in Experiment of Evaluation of Moniqua.

	100mbps/0.15ms	20mbps/0.15ms	100mbps/10ms	Extra Memory
Centralized	38.92	206.14	343.28	N/A
D-PSGD	36.25	189.48	310.98	N/A
DCD-PSGD	32.99	105.40	202.42	20.4 MB
ECD-PSGD	31.96	105.26	202.04	20.4 MB
ChocoSGD	32.03	105.18	201.18	20.4 MB
DeepSqueeze	30.01	103.67	193.92	13.6 MB
Moniqua	22.42	95.08	184.86	0 B

Table 2: Bandwidth consumption under different θ and δ when applying linear quantizer in Moniqua

MODEL	MOD PARAM θ	QUANT STEP	BYTES/EPOCH	AVG BITS
VGG16	NONE	NONE	45594MB	32
	1.0	0.01	5206MB	3.65
	0.08	0.01	5192MB	3.64
RESNET110	NONE	NONE	3430MB	32
	2.0	0.01	609MB	5.67
	1.3	0.01	608MB	5.67

2. **Randomized Gossip (Non-linear)**

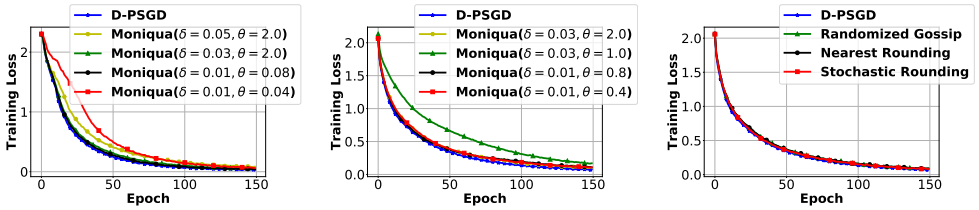
$$\mathcal{Q}(x) = \begin{cases} x, & \text{with probability} \quad p \\ 0, & \text{with probability} \quad 1-p \end{cases}$$

In this experiment, we set $p = 0.7$.

We train ResNet110 on CIFAR10, and plot the results in Figure 5(c). We can see that the training curves of using three quantizers are all aligned with D-PSGD with full-precision communication. Note that in the paper we show that previous work cannot perserve the aligned curve even with stochastic rounding (unbiased), thus we are not comparing them here.

C.3 MORE RESULTS ON DIFFERENT HYPERPARAMETERS

In this experiment, we plot more result of training ResNet110 and VGG16 on CIFAR10 under different δ and θ in the experiment of aggressive quantization. And we plot the results in Figure 5(a) and Figure 5(b).



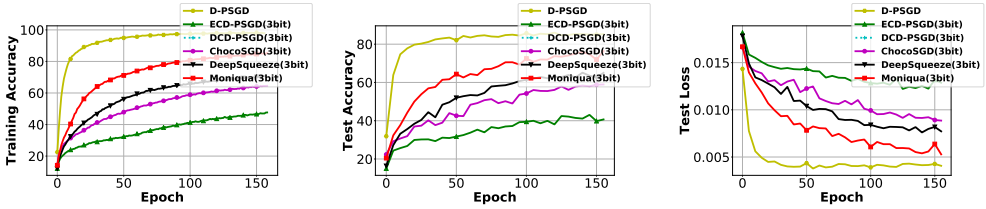
(a) Performance of Moniqua on VGG16 under different θ and δ

(b) Performance of Moniqua on ResNet110 under different θ and δ

(c) Performance of algorithms under different quantizer

C.4 PERFORMANCE ON THE TESTSET UNDER AGGRESSIVE QUANTIZATION

We report the results in experiment of "Aggressive Quantization" and report the test error and test accuracy in the Figure 5 and Figure 6.

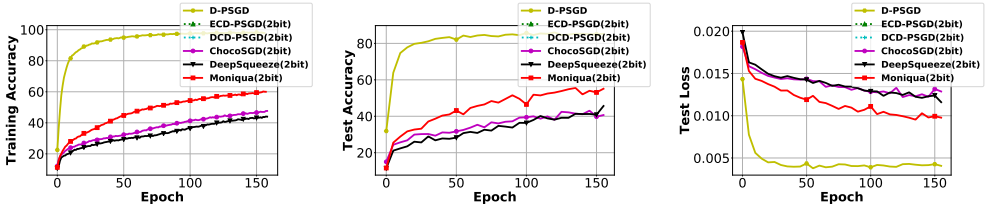


(d) Training Accuracy under different algorithms with 3-bit communication

(e) Test Accuracy under different algorithms with 3-bit communication

(f) Test Loss under different algorithms with 3-bit communication

Figure 5: More statistics from Experiment of Aggressive Quantization under 3-bit communication



(a) Training Accuracy under different algorithms with 2-bit communication

(b) Test Accuracy under different algorithms with 2-bit communication

(c) Test Loss under different algorithms with 2-bit communication

Figure 6: More statistics from Experiment of Aggressive Quantization under 2-bit communication

D WHY NAIVE QUANTIZATION FAILS IN D-PSGD (PROOF TO THEOREM 1)

The update rule of naive quantization on D-PSGD is

$$x_{k+1,i} = x_{k,i}W_{ii} + \sum_{j=1,j\neq i}^{n} \mathcal{Q}(x_{k,j})W_{ji} - \alpha_k \widetilde{g}_{k,i} = x_{k,i} + \sum_{j=1,j\neq i}^{n} (\mathcal{Q}(x_{k,j}) - x_{k,i})W_{ji} - \alpha_k \widetilde{g}_{k,i}$$

where α_k is allowed to vary with any policy. Let

$$X_k = [x_{k,1}, \cdots, x_{k,n}] \in \mathbb{R}^{d\times n}$$

$$\Omega_k = \left[\sum_{j\neq 1} W_{j1}\left(\mathcal{Q}(x_{k,j}) - x_{k,1}\right), \cdots, \sum_{j\neq n} W_{jn}\left(\mathcal{Q}(x_{k,j}) - x_{k,n}\right)\right] \in \mathbb{R}^{d\times n}$$

$$\widetilde{G}_k = [\widetilde{g}_{k,1}, \cdots, \widetilde{g}_{k,n}] \in \mathbb{R}^{d\times n}$$

by rewritting the update rule, we obtain

$$X_{k+1} = X_k + \Omega_k - \alpha_k \widetilde{G}_k$$

Let $Y_k = X_k - x^* \mathbb{1}_n^\top$, and considering the fact that $\nabla f(x) = x - \delta/2 = x - x^*$, we can rewrite the update rule as

$$Y_{k+1}e_i = Y_k e_i + \Omega_k e_i - \alpha_k Y_k e_i + \alpha_k \left(\widetilde{G}_k - G_k\right) e_i$$

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

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

$$\mathbb{E}\left\|Y_{k+1}e_i\right\|^2 \leq \mathbb{E}\left\|Y_k e_i\right\|^2 < \frac{\phi^2\delta^2}{8(1+\phi^2)} \tag{6}$$

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

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

As a result

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left\|Y_{k+1}e_i\right\|^2 =&\mathbb{E}\left\|\left(Y_k + \Omega_k - \alpha_k Y_k + \alpha_k\left(\widetilde{G}_k - G_k\right)\right)e_i\right\|^2 \\
\overset{(*)}{=}&\mathbb{E}\left\|(1-\alpha_k)Y_k e_i\right\|^2 + \mathbb{E}\left\|\Omega_k e_i\right\|^2 + \mathbb{E}\left\|\alpha_k\left(\widetilde{G}_k - G_k\right)e_i\right\|^2 \\
\geq&\mathbb{E}\left\|\Omega_k e_i\right\|^2 \\
\geq&\frac{\phi^2\delta^2}{8(1+\phi^2)}
\end{aligned}$$

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 $\frac{\phi^2\delta^2}{8(1+\phi^2)} \leq \mathbb{E}\left\|x_{k+1} - x^*\right\|^2 < \frac{\phi^2\delta^2}{8(1+\phi^2)}$. That being said, for $\forall k, i$

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

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 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 $\frac{\mathbb{1}_n}{n}$ as its the stationary distribution $\left(W\frac{\mathbb{1}_n}{n} = \frac{\mathbb{1}_n}{n}\right)$. 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.[9]

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 1

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

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

$$x = [2, -1]^\top$$
$$u = [2, 0]^\top$$
$$v = [0, 1]^\top$$

And we have

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

Considering the definition of L1-norm, we have

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

According to a well-known results on the theory of Markov Chains,[10] $d(lt_{\text{mix}}) \leq 2^{-l}$ holds for any non-negative integer l, so we have

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

That completes the proof.

Additionally, based on standard results in the theory of reversible Markov Chains, we also have[11]

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

[9] 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`

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

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

E.2 AD-PSGD

Note that unlike D-PSGD, here 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 $\mu \in \mathbb{R}^d$ such that $\mathbb{1}_n^\top \mu = 1$, it must hold that

$$\left\| \left(\prod_{i=1}^{t_{\text{mix}}} W_{k+i} \right) \mu - \frac{\mathbb{1}_n}{n} \right\|_1 \leq \frac{1}{2}.$$

We call this constant t_{mix} because it is effectively the *mixing time* of the time-inhomogeneous Markov chain with transition probability matrix 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 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 2 *For any $k \geq 0$ and for any $b \geq a \geq 0$, there exists t_{mix} such that*

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

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

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

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

$$\left\| \prod_{q=1}^{t_{\text{mix}}} W_{q+k} \left(I - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n} \right) \right\|_1 = \max_x \frac{\left\| \prod_{q=1}^{t_{\text{mix}}} W_{q+k} \left(I - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n} \right) x \right\|_1}{\|x\|_1} \leq \frac{1}{2}$$

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

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

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

$$\left\| \prod_{q=1}^{t_r} W_q \left(I - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n} \right) \right\|_1 \leq 1 - \frac{1}{n} + (n-1)\frac{1}{n} = 2 - \frac{2}{n} \leq 2$$

Putting it back we obtain

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

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)

Consistent with linear and non-linear quantizer Here we briefly explain why using $\theta \cdot \mathcal{Q}_\delta\left(\frac{x}{\theta} \bmod 1\right)$ instead of $\mathcal{Q}_\delta\left(x \bmod \theta\right)$ for theoretical analysis and how it covers both linear and non-linear quantizers. Note that typically, a linear quantizer has:

$$\|\mathcal{Q}_\delta(x)-x\|_\infty \leq \delta, \quad \forall x\in\mathbb{R}^d$$

while a non-linear quantizer has

$$\|\mathcal{Q}_\delta(x)-x\|_\infty \leq \delta\|x\|_\infty, \quad \forall x\in\mathbb{R}^d$$

so that for linear quantizer, with a given $x \leq \theta$:

$$\left\|\theta\cdot\mathcal{Q}_\delta\left(\frac{x}{\theta} \bmod 1\right)-x\right\|_\infty = \theta\left\|\mathcal{Q}_\delta\left(\frac{x}{\theta} \bmod 1\right)-\left(\frac{x}{\theta} \bmod 1\right)\right\|_\infty \leq \theta\delta = \theta\delta$$

And for non-linear quantizer, with a given $x \leq \theta$:

$$\left\|\theta\cdot\mathcal{Q}_\delta\left(\frac{x}{\theta} \bmod 1\right)-x\right\|_\infty = \theta\left\|\mathcal{Q}_\delta\left(\frac{x}{\theta} \bmod 1\right)-\left(\frac{x}{\theta} \bmod 1\right)\right\|_\infty \leq \theta\delta\cdot 1 = \theta\delta$$

As a result, we can use the same bound $\theta\delta$ for quantizers with both of the properties, which we will show in the rest of the proof.

F.1 PROOF TO THEOREM 2

Proof *For convenience, we define the following notation*

$$\begin{aligned} X_k &= [x_{k,1},\cdots,x_{k,n}], \qquad Q_k = [q_{k,1},\cdots,q_{k,n}] \\ \widetilde{G}_k &= [\widetilde{g}_{k,1},\cdots,\widetilde{g}_{k,n}], \qquad G_k = [g_{k,1},\cdots,g_{k,n}] \\ \overline{X} &= X\frac{\mathbb{1}_n}{n}, \forall X\in\mathbb{R}^{d\times n}, \qquad \Omega_k = (Q_k-X_k)(W-I) \end{aligned}$$

where $g_{k,i}$ denotes gradient computed via the whole dataset $\mathcal{D}_i$ and $x_{k,i}$

From a local view, the update rule of Algorithm 1 on worker i at iteration k can be written as

$$x_{k+1,i} \leftarrow x_{k,i} + \sum\nolimits_{j\in\mathcal{N}_i}\left(q_{k,j}-q_{k,i}\right)W_{ji} - \alpha\widetilde{g}_{k,i}$$

which is equivalent to

$$x_{k+1,i} = x_{k,i} + \sum_{j=1}^{n}\left(x_{k,j}-x_{k,i}\right)W_{ji} - \alpha\widetilde{g}_{k,i} + \sum_{j=1}^{n}\left(\left(q_{k,j}-x_{k,j}\right)-\left(q_{k,i}-x_{k,i}\right)\right)W_{ji} \tag{7}$$

From a global view, the update rule can be written as

$$X_{k+1} = X_k + Q_k(W-I) - \alpha\widetilde{G}_k = X_kW - \alpha\widetilde{G}_k + (Q_k-X_k)(W-I) \tag{8}$$

From Lemma 5 we have

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

$$+\frac{8L^2}{nK(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2$$

Note that

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2=\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{j=1}^{n}\left(\left(q_{k,j}-x_{k,j}\right)-\left(q_{k,i}-x_{k,i}\right)\right)W_{ji}\right\|^2$$
$$\overset{\text{Lemma 3}}{\leq}4\sum_{k=0}^{K-1}\sum_{i=1}^{n}\delta^2\theta^2 d\leq\alpha^2G_\infty^2dnK$$

The last step holds because $\delta\theta=\frac{1}{2}\alpha G_\infty$. Pushing it back we obtain

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

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

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

That completes the proof of Theorem 2.

F.2 Lemma for Moniqua on D-PSGD

Lemma 3 *If $\|x_{t,i}-x_{t,j}\|_\infty\leq\theta$, $\forall i,j$ holds at iteration t, then*

$$\left\|\sum_{j=1}^{n}\left(\left(q_{t,j}-x_{t,j}\right)-\left(q_{t,i}-x_{t,i}\right)\right)W_{ji}\right\|_\infty\leq 2\delta\theta$$

Proof

$$\begin{aligned}&\left\|\sum_{j=1}^{n}\left(\left(q_{t,j}-x_{t,j}\right)-\left(q_{t,i}-x_{t,i}\right)\right)W_{ji}\right\|_\infty\\\leq&\sum_{j=1}^{n}W_{ji}\left\|\left(q_{t,j}-x_{t,j}\right)-\left(q_{t,i}-x_{t,i}\right)\right\|_\infty\\=&\sum_{j=1}^{n}W_{ji}\left\|\theta\mathcal{Q}\left(\frac{x_{t,j}}{\theta}\bmod 1\right)-\theta\mathcal{Q}\left(\frac{x_{t,i}}{\theta}\bmod 1\right)-\left(x_{t,j}-x_{t,i}\right)\right\|_\infty\\=&\sum_{j=1}^{n}W_{ji}\left\|\theta\mathcal{Q}\left(\frac{x_{t,j}}{\theta}\bmod 1\right)-\theta\mathcal{Q}\left(\frac{x_{t,j}}{\theta}\bmod 1\right)-\theta\left(\frac{x_{t,j}-x_{t,i}}{\theta}\bmod 1\right)\right\|_\infty\\=&\sum_{j=1}^{n}W_{ji}\left\|\theta\mathcal{Q}\left(\frac{x_{t,j}}{\theta}\bmod 1\right)-\theta\left(\frac{x_{t,j}}{\theta}\bmod 1\right)-\left(\theta\mathcal{Q}\left(\frac{x_{t,i}}{\theta}\bmod 1\right)-\theta\left(\frac{x_{t,i}}{\theta}\bmod 1\right)\right)\right\|_\infty\\\leq&\sum_{j=1}^{n}W_{ji}\left\|\theta\mathcal{Q}\left(\frac{x_{t,j}}{\theta}\bmod 1\right)-\theta\left(\frac{x_{t,j}}{\theta}\bmod 1\right)\right\|_\infty\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{n} W_{ji} \left\| \theta \mathcal{Q}\left(\frac{x_{t,i}}{\theta} \bmod 1\right) - \theta\left(\frac{x_{t,i}}{\theta} \bmod 1\right) \right\|_\infty \\
\leq & 2\delta\theta
\end{aligned}
$$

Lemma 4 *In any iteration $k \geq 0$, and for any two worker i and j, we have:*

$$
\left\| X_k(e_i - e_j) \right\|_\infty \leq \theta = \frac{2\log(16n)}{1-\rho} \alpha G_\infty
$$

Proof *We use mathematical induction to prove this:*

I. When $k = 0$, $\left\| X_0(e_i - e_j) \right\|_\infty = 0 \leq \theta, \forall i, j$

II. Suppose for $\left\| X_k(e_i - e_j) \right\|_\infty \leq \theta, k \geq 0, \forall i, j$, we have

$$
\begin{aligned}
\left\| X_{k+1}(e_i - e_j) \right\|_\infty &= \left\| \left(X_k W - \alpha \widetilde{G}_k + \Omega_k \right)(e_i - e_j) \right\|_\infty \\
&\overset{X_0 = 0}{=} \left\| \sum_{t=0}^{k} \left(-\alpha \widetilde{G}_t + \Omega_t \right) W^{k-t}(e_i - e_j) \right\|_\infty \\
&\leq \sum_{t=0}^{k} \left\| \left(-\alpha \widetilde{G}_t + \Omega_t \right) W^{k-t}(e_i - e_j) \right\|_\infty \\
&\leq \sum_{t=0}^{k} \left\| -\alpha \widetilde{G}_t + \Omega_t \right\|_{1,\infty} \left\| W^{k-t}(e_i - e_j) \right\|_1 \\
&\leq \sum_{t=0}^{k} \left(\alpha \left\| \widetilde{G}_t \right\|_{1,\infty} + \left\| \Omega_t \right\|_{1,\infty} \right) \left\| W^{k-t}(e_i - e_j) \right\|_1 \\
&\overset{\text{induction hypothesis}}{\leq} (\alpha G_\infty + 2\delta\theta) \sum_{t=0}^{k} \left\| W^{k-t}(e_i - e_j) \right\|_1 \\
&\leq (\alpha G_\infty + 2\delta\theta) \sum_{t=0}^{\infty} \left\| W^{t}(e_i - e_j) \right\|_1
\end{aligned}
$$

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

$$
\left\| W^t(e_i - e_j) \right\|_1 \leq \sqrt{n} \left\| W^t(e_i - e_j) \right\|_2 \leq \sqrt{n} \left\| W^t e_i - \frac{\mathbb{1}_n}{n} \right\| + \sqrt{n} \left\| W^t e_j - \frac{\mathbb{1}_n}{n} \right\| \leq 2\sqrt{n}\rho^t
$$

where the last step holds due to the diagonalizability of W. On the other hand,

$$
\left\| W^t(e_i - e_j) \right\|_1 \leq \mathbb{1}_n^\top W^t e_i + \mathbb{1}_n^\top W^t e_i = \mathbb{1}_n^\top e_i + \mathbb{1}_n^\top e_j = 2
$$

So

$$
\left\| W^t(e_i - e_j) \right\|_1 \leq \min\{2\sqrt{n}\rho^t, 2\}
$$

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

$$
\begin{aligned}
\sum_{t=0}^{\infty} \left\| W^t(e_i - e_j) \right\|_1 &= \sum_{t=0}^{T_0 - 1} \left\| W^t(e_i - e_j) \right\|_1 + \sum_{t=T_0}^{\infty} \left\| W^t(e_i - e_j) \right\|_1 \\
&\leq \sum_{t=0}^{T_0 - 1} 2 + \sum_{t=0}^{\infty} 2\sqrt{n}\rho^{t+T_0} \\
&\leq 2\left\lceil \frac{-\log(\sqrt{n})}{\log(\rho)} \right\rceil + \sum_{t=0}^{\infty} 2\left(\sqrt{n}\rho^{T_0}\right)\rho^t \\
&\leq \frac{2\log(\sqrt{n})}{1-\rho} + 2 + \frac{2}{1-\rho}
\end{aligned}
$$

$$\leq \frac{\log(16n)}{1-\rho}$$

As a result, we have

$$\left\|X_{k+1}(e_i - e_j)\right\|_\infty \leq (\alpha G_\infty + 2\delta\theta)\,\frac{\log(16n)}{1-\rho}$$

Since $\delta = \frac{1-\rho}{4\log(16n)}$*, we have*

$$\left\|X_{k+1}(e_i - e_j)\right\|_\infty \leq (\alpha G_\infty + 2\delta\theta)\frac{\log(16n)}{1-\rho} \leq \frac{2\log(16n)}{1-\rho}\alpha G_\infty = \theta$$

Combining I and II, we complete the proof.

Lemma 5 *The output of Algorithm 1 has the following bound:*

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

Proof *From Lemma 8, we have*

$$\begin{aligned}
&\frac{1-\alpha L}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2 + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \\
\leq &\frac{2(f(0)-f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{L^2}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k - x_{k,i}\right\|^2 \\
\overset{Lemma\ 6}{\leq} &\frac{2(f(0)-f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{2\alpha^2L^2}{M_1(1-\rho)^2}\left(\sigma^2+3\varsigma^2 + \frac{3}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2\right) \\
&+\frac{2L^2}{M_1nK(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}$$

where

$$M_1 = 1 - \frac{6\alpha^2L^2}{(1-\rho)^2}$$

Rearrange the terms, we get

$$\begin{aligned}
&\frac{1-\alpha L}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2 + \left(1 - \frac{6\alpha^2L^2}{M_1(1-\rho)^2}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \\
\leq &\frac{2(f(0)-f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{2\alpha^2L^2\left(\sigma^2+3\varsigma^2\right)}{M_1(1-\rho)^2} + \frac{2L^2}{M_1nK(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}$$

Let

$$M_2 = 1 - \frac{6\alpha^2L^2}{M_1(1-\rho)^2}$$

we get

$$\begin{aligned}
&\frac{1-\alpha L}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2 + \frac{M_2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \\
\leq &\frac{2(f(0)-f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{2\alpha^2L^2\left(\sigma^2+3\varsigma^2\right)}{M_1(1-\rho)^2} + \frac{2L^2}{M_1nK(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}$$

Let $M_1, M_2 \geq \frac{1}{2}$ and rearrange the terms, we have

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

and that completes the proof

Lemma 6 *Let $M_1 = 1 - \frac{6\alpha^2L^2}{(1-\rho)^2} > 0$, we have*

$$\frac{L^2}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k - x_{k,i}\right\|^2 \leq \frac{2\alpha^2L^2}{M_1(1-\rho)^2}\left(\sigma^2+3\varsigma^2+\frac{3}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2\right) + \frac{2L^2}{M_1nK(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2$$

Proof

$$\begin{aligned}
&\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k - x_{k,i}\right\|^2 \\
=&\sum_{k=1}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|X_k\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2 \\
=&\sum_{k=1}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\left(X_{k-1}W-\alpha\widetilde{G}_{k-1}+\Omega_{k-1}\right)\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2 \\
\overset{x_{0,i}=0}{=}&\sum_{k=1}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{t=0}^{k-1}\left(-\alpha\widetilde{G}_t+\Omega_t\right)\left(\frac{\mathbb{1}_n}{n}-W^{k-t-1}e_i\right)\right\|^2 \\
\leq&2\alpha^2\sum_{k=1}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{t=0}^{k-1}\widetilde{G}_t\left(\frac{\mathbb{1}_n}{n}-W^{k-t-1}e_i\right)\right\|^2 \\
&+2\sum_{k=1}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{t=0}^{k-1}\Omega_t\left(\frac{\mathbb{1}_n}{n}-W^{k-t-1}e_i\right)\right\|^2 \\
=&2\alpha^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\sum_{t=0}^{k-1}\widetilde{G}_t\left(\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n}-W^{k-t-1}\right)\right\|_F^2 + 2\sum_{k=1}^{K-1}\mathbb{E}\left\|\sum_{t=0}^{k-1}\Omega_t\left(\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n}-W^{k-t-1}\right)\right\|_F^2 \\
\overset{Lemma\ 10}{\leq}&2\alpha^2\sum_{k=1}^{K-1}\mathbb{E}\left(\sum_{t=0}^{k-1}\rho^{k-t-1}\left\|\widetilde{G}_t\right\|_F\right)^2 + 2\sum_{k=1}^{K-1}\mathbb{E}\left(\sum_{t=0}^{k-1}\rho^{k-t-1}\left\|\Omega_t\right\|_F\right)^2 \\
\overset{Lemma\ 9}{\leq}&\frac{2\alpha^2}{(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\widetilde{G}_k\right\|_F^2 + \frac{2}{(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2 \\
\overset{Lemma\ 7}{\leq}&\frac{2\alpha^2}{(1-\rho)^2}\left(n\sigma^2K+3L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k-x_{k,i}\right\|^2+3n\varsigma^2K+3n\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2\right) \\
&+\frac{2}{(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}$$

Rearrange the terms, we have

$$\left(1-\frac{6\alpha^2L^2}{(1-\rho)^2}\right)\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k-x_{k,i}\right\|^2$$

$$\leq \frac{2\alpha^2}{(1-\rho)^2}\left(n\sigma^2 K + 3n\varsigma^2 K + 3n\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2\right) + \frac{2}{(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2$$

Let $M_1 = 1 - \frac{6\alpha^2 L^2}{(1-\rho)^2} > 0$*, we have*

$$\frac{L^2}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k - x_{k,i}\right\|^2 \leq \frac{2\alpha^2 L^2}{M_1(1-\rho)^2}\left(\sigma^2 + 3\varsigma^2 + \frac{3}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2\right) + \frac{2L^2}{M_1 nK(1-\rho)^2}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2$$

Lemma 7

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\widetilde{G}_k\right\|_F^2 \leq n\sigma^2 K + 3L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\overline{X}_k - x_{k,i}\right\|^2 + 3n\varsigma^2 K + 3n\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2$$

Proof *From the property of Frobenius norm, we have*

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

Next, we derive the upper bound of $\mathbb{E}\left\|\widetilde{g}_{k,i}\right\|^2$

$$\begin{aligned}
&\mathbb{E}\left\|\widetilde{g}_{k,i}\right\|^2 \\
=&\mathbb{E}\left\|\widetilde{g}_{k,i} - g_{k,i} + g_{k,i}\right\|^2 \\
=&\mathbb{E}\left\|\widetilde{g}_{k,i} - g_{k,i}\right\|^2 + \mathbb{E}\left\|g_{k,i}\right\|^2 + 2\mathbb{E}\left\langle\widetilde{g}_{k,i} - g_{k,i}, g_{k,i}\right\rangle \\
=&\mathbb{E}\left\|\widetilde{g}_{k,i} - g_{k,i}\right\|^2 + \mathbb{E}\left\|g_{k,i}\right\|^2 \\
\leq&\sigma^2 + 3\mathbb{E}\left\|g_{k,i} - \nabla f_i(\overline{X}_k)\right\|^2 + 3\mathbb{E}\left\|\nabla f_i(\overline{X}_k) - \nabla f(\overline{X}_k)\right\|^2 + 3\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \\
\leq&\sigma^2 + 3L^2\mathbb{E}\left\|\overline{X}_k - x_{k,i}\right\|^2 + 3\varsigma^2 + 3\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2
\end{aligned}$$

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

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

That completes the proof

Lemma 8

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

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

$$\overline{X}_{k+1} = (X_k W - \alpha\widetilde{G}_k + \Omega_k)\frac{\mathbb{1}_n}{n} = \overline{X}_k - \alpha\overline{\widetilde{G}}_k + (Q_k - X_k)(W - I)\frac{\mathbb{1}_n}{n} = \overline{X}_k - \alpha\overline{\widetilde{G}}_k$$

And by Taylor Expansion, we have

$$\begin{aligned}\mathbb{E}f(\overline{X}_{k+1}) &= \mathbb{E}f\left(\frac{(X_kW-\alpha\widetilde{G}_k+\Omega_k)\mathbb{1}_n}{n}\right)\\ &= \mathbb{E}f\left(\overline{X}_k-\alpha\overline{\widetilde{G}}_k\right)\\ &\leq \mathbb{E}f(\overline{X}_k)-\alpha\mathbb{E}\langle\nabla f(\overline{X}_k),\overline{\widetilde{G}}_k\rangle+\frac{\alpha^2L}{2}\mathbb{E}\left\|\overline{\widetilde{G}}_k\right\|^2\end{aligned}$$

And for the last term, we have

$$\begin{aligned}\mathbb{E}\left\|\overline{\widetilde{G}}_k\right\|^2 &= \mathbb{E}\left\|\frac{\sum_{i=1}^n\widetilde{g}_{k,i}}{n}\right\|^2\\ &= \mathbb{E}\left\|\frac{\sum_{i=1}^n\widetilde{g}_{k,i}-\sum_{i=1}^n g_{k,i}}{n}+\frac{\sum_{i=1}^n g_{k,i}}{n}\right\|^2\\ &= \mathbb{E}\left\|\frac{\sum_{i=1}^n\widetilde{g}_{k,i}-\sum_{i=1}^n g_{k,i}}{n}\right\|^2+\mathbb{E}\left\|\frac{\sum_{i=1}^n g_{k,i}}{n}\right\|^2\\ &+\mathbb{E}\left\langle\frac{\sum_{i=1}^n\widetilde{g}_{k,i}-\sum_{i=1}^n g_{k,i}}{n},\frac{\sum_{i=1}^n g_{k,i}}{n}\right\rangle\\ &= \mathbb{E}\left\|\frac{\sum_{i=1}^n\widetilde{g}_{k,i}-\sum_{i=1}^n g_{k,i}}{n}\right\|^2+\mathbb{E}\left\|\frac{\sum_{i=1}^n g_{k,i}}{n}\right\|^2\\ &\overset{Assumption\ (A3)}{=} \frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\widetilde{g}_{k,i}-g_{k,i}\right\|^2+\mathbb{E}\left\|\frac{\sum_{i=1}^n g_{k,i}}{n}\right\|^2\\ &\overset{Assumption\ (A3)}{\leq} \frac{\sigma^2}{n}+\mathbb{E}\left\|\frac{\sum_{i=1}^n g_{k,i}}{n}\right\|^2\end{aligned}$$

Putting it back, we obtain

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

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

$$\begin{aligned}\mathbb{E}\left\|\nabla f(\overline{X}_k)-\overline{G}_k\right\|^2 &\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i\left(\frac{\sum_{i'=1}^n x_{k,i'}}{n}\right)-\nabla f_i(x_{k,i})\right\|^2\\ &\overset{Assumption\ (A1)}{\leq} \frac{L^2}{n}\sum_{i=1}^n\mathbb{E}\left\|\frac{\sum_{i'=1}^n x_{k,i'}}{n}-x_{k,i}\right\|^2\\ &= \frac{L^2}{n}\sum_{i=1}^n\mathbb{E}\left\|\overline{X}_k-x_{k,i}\right\|^2\end{aligned}$$

putting it back, we have

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

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

$$\frac{1-\alpha L}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2$$

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

That completes the proof.

Lemma 9 *Given two non-negative sequences $\{a_t\}_{t=1}^{\infty}$ and $\{b_t\}_{t=1}^{\infty}$ that satisfying*

$$a_t = \sum_{s=1}^{t}\rho^{t-s}b_s$$

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

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

Proof

$$S_k = \sum_{t=1}^{k} a_t = \sum_{t=1}^{k}\sum_{s=1}^{t}\rho^{t-s}b_s = \sum_{s=1}^{k}\sum_{t=s}^{k}\rho^{t-s}b_s = \sum_{s=1}^{k}\sum_{t=0}^{k-s}\rho^{t}b_s \leq \frac{1}{1-\rho}\sum_{s=1}^{k} b_s$$
$$D_k = \sum_{t=1}^{k} a_t = \sum_{t=1}^{k}\sum_{s=1}^{t}\rho^{t-s}b_s\sum_{r=1}^{t}\rho^{t-r}b_r = \sum_{s=1}^{k}\sum_{t=s}^{k}\rho^{t-s}b_s = \sum_{t=1}^{k}\sum_{s=1}^{t}\sum_{r=1}^{t}\rho^{2t-s-r}b_s b_r$$
$$\leq \sum_{t=1}^{k}\sum_{s=1}^{t}\sum_{r=1}^{t}\rho^{2t-s-r}\frac{b_s^2+b_r^2}{2} = \sum_{t=1}^{k}\sum_{s=1}^{t}\sum_{r=1}^{t}\rho^{2t-s-r}b_s^2$$
$$\leq \frac{1}{1-\rho}\sum_{t=1}^{k}\sum_{s=1}^{t}\rho^{t-s}b_s^2 \leq \frac{1}{(1-\rho)^2}\sum_{s=1}^{k} b_s^2$$

Lemma 10 *For any $X_t \in \mathbb{R}^{d\times n}$, we have*

$$\left\|\sum_{t=0}^{k-1} X_t\left(\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n} - W^{k-t-1}\right)\right\|_F^2 \leq \left(\sum_{t=0}^{k-1}\rho^{k-t-1}\left\|X_t\right\|_F\right)^2$$

Proof

$$\left\|\sum_{t=0}^{k-1} X_t\left(\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n} - W^{k-t-1}\right)\right\|_F^2 = \left(\left\|\sum_{t=0}^{k-1} X_t\left(\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n} - W^{k-t-1}\right)\right\|_F\right)^2$$
$$\leq \left(\sum_{t=0}^{k-1}\left\|X_t\left(\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n} - W^{k-t-1}\right)\right\|_F\right)^2$$
$$\leq \left(\sum_{t=0}^{k-1}\left\|X_t\right\|_F\left\|\frac{\mathbb{1}_n\mathbb{1}_n^\top}{n} - W^{k-t-1}\right\|\right)^2$$
$$\leq \left(\sum_{t=0}^{k-1}\rho^{k-t-1}\left\|X_t\right\|_F\right)^2$$

That completes the proof.

G MONIQUA ON D^2 (PROOF TO THEOREM 3)

G.1 ALGORITHM

Algorithm 2 Moniqua with Variance Reduction on worker i

Input: initial point $x_{0,i} = x_0$, step size α, the discrepency bound θ, communication matrix W, number of iterations K, neighbor list of worker i: $\mathcal{N}_i$

1: **for** $k = 0, 1, 2, \cdots, K-1$ **do**
2: Randomly sample data $\xi_{k,i}$ from local memory
3: Compute a local stochastic gradient based on $\xi_{k,i}$ and current weight $x_{k,i}$: $\widetilde{g}_{k,i}$
4: **if** $k = 0$ **then**
5: Update local weight: $x_{k+\frac{1}{2},i} \leftarrow x_{k,i} - \alpha\widetilde{g}_{k,i}$
6: **else**
7: Update local weight: $x_{k+\frac{1}{2},i} \leftarrow 2x_{k,i} - x_{k-1,i} - \alpha\widetilde{g}_{k,i} + \alpha\widetilde{g}_{k-1,i}$
8: **end if**
9: Compute modulo-ed model: $q_{k,i} \leftarrow \theta \cdot \mathcal{Q}_\delta\left(\frac{x_{k+\frac{1}{2},i}}{\theta} \bmod 1\right)$ (element-wise division and mod)
10: Average with neighboring workers:: $x_{k+1,i} \leftarrow x_{k+\frac{1}{2},i} + \sum_{j\in\mathcal{N}_i}(q_{k,j} - q_{k,i})W_{ji}$
11: **end for**

Output: $\overline{X}_K = \frac{1}{n}\sum_{i=1}^n x_{K,i}$

G.2 ASSUMPTIONS

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 W is a symmetric doubly stochastic matrix. Let the eigenvalues of $W \in \mathbb{R}^{n\times n}$ be $\lambda_1 \geq \cdots \geq \lambda_n$. We assume $\lambda_2 < 1, \lambda_n > -\frac{1}{3}$.
3. **Bounded Variance**:
$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left\|\nabla\widetilde{f}_i(x_i;\xi_i) - \nabla f_i(x)\right\|^2 \leq \sigma^2, \forall i$$
where $\nabla\widetilde{f}_i(x;\xi_i)$ denotes gradient sample on worker i computed via data sample ξ_i.
4. **Initialization**: All the models are initialized by the same parameters: $x_{0,i} = x_0, \forall i$ and with out the loss of generality $x_0 = 0$.
5. **Gradient magnitude**: The norm of a sampled gradient is bounded by $\|\widetilde{g}_{k,i}\|_\infty \leq G_\infty$ for some constant G_∞.

G.3 PROOF TO THEOREM 3

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

$$x_{k+\frac{1}{2},i} = 2x_{k,i} - x_{k-1,i} - \alpha\widetilde{g}_{k,i} + \alpha\widetilde{g}_{k-1,i}$$
$$x_{k+1,i} = \sum_{j=1}^n x_{k+\frac{1}{2},j}W_{ji} + \sum_{j=1}^n\left((q_{k,j} - x_{k+\frac{1}{2},j}) - (q_{k,i} - x_{k+\frac{1}{2},i})\right)W_{ji}$$

From a global view, the update rule can be written as

$$X_{k+\frac{1}{2}} = 2X_k - X_{k-1} - \alpha\widetilde{G}_k + \alpha\widetilde{G}_{k-1}$$
$$X_{k+1} = X_{k+\frac{1}{2}}W + (Q_k - X_{k+\frac{1}{2}})(W - I)$$

Define

$$\Omega_k = (Q_k - X_{k+\frac{1}{2}})(W - I)$$

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

$$X_{k+1} = 2X_kP\Lambda P^\top - X_{k-1}P\Lambda P^\top - \alpha\widetilde{G}_kP\Lambda P^\top + \alpha\widetilde{G}_{k-1}P\Lambda P^\top + \Omega_k$$

and

$$X_{k+1}P = 2X_kP\Lambda - X_{k-1}P\Lambda - \alpha\widetilde{G}_kP\Lambda + \alpha\widetilde{G}_{k-1}P\Lambda + \Omega_kP$$

Denote $Y_k = X_kP$, $H(X_k;\xi_k) = \widetilde{G}_kP$, and denote $y_{k,i}$, $h_{k,i}$ and $r_{k,i}$ as the i-th column of Y_k, H_k and Ω_kP, respectively. Then we have

$$y_{k+1,i} = \lambda_i(2y_{k,i} - y_{k-1,i} - \alpha h_{k,i} + \alpha h_{k-1,i}) + r_{k,i}$$

From Lemma 15 (Constants C_1, C_2, C_3 andn C_4 are defined in the Lemma 11. Constants D_1 and D_2 are defined in Lemma 15) we get

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

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

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

That completes the proof.

G.4 LEMMA FOR D^2

Lemma 11 *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 $\delta = \frac{1}{6nD_2}$, and we have for $\forall i, j$

$$\left\|x_{k+\frac{1}{2}}(e_i - e_j)\right\|_\infty \leq \theta = (6D_1n+8)\alpha G_\infty$$

Proof *We use mathematical induction to prove this:*

I. When $k = 0$*,*

$$\left\|X_{0+\frac{1}{2}}(e_i - e_j)\right\|_\infty = \left\|-\alpha\widetilde{G}_0(e_i - e_j)\right\|_\infty \leq \alpha\left\|\widetilde{G}_0\right\|_{1,\infty}\|e_i - e_j\|_1 \leq 2\alpha G_\infty \leq (6D_1n + 8)\alpha G_\infty$$

II. Suppose for $k \geq 0$*,* $\forall t \leq k$*, we have* $\left\|X_{t+\frac{1}{2}}(e_i - e_j)\right\| \leq (6D_1n + 8)\alpha G_\infty$*, then for* $\forall i, j$

$$\begin{aligned}
&\|X_{k+1}(e_i - e_j)\|_\infty \\
\leq&\left\|X_{k+1}\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|_\infty + \left\|X_{k+1}\left(\frac{\mathbb{1}_n}{n} - e_j\right)\right\|_\infty \\
=&\left\|X_{k+1}PP^\top e_i - X_{k+1}P\begin{bmatrix}1&0&0&\dots&0\\0&0&0&\dots&0\\0&0&0&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&0\end{bmatrix}P^\top e_i\right\|_\infty \\
&+\left\|X_{k+1}PP^\top e_j - X_{k+1}P\begin{bmatrix}1&0&0&\dots&0\\0&0&0&\dots&0\\0&0&0&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&0\end{bmatrix}P^\top e_j\right\|_\infty \\
\leq&\left\|X_{k+1}P\begin{bmatrix}0&0&0&\dots&0\\0&1&0&\dots&0\\0&0&1&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&1\end{bmatrix}\right\|_{1,\infty}\|P^\top e_i\|_1 + \left\|X_{k+1}P\begin{bmatrix}0&0&0&\dots&0\\0&1&0&\dots&0\\0&0&1&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&1\end{bmatrix}\right\|_{1,\infty}\|P^\top e_j\|_1 \\
\leq&2\sqrt{n}\left\|X_{k+1}P\begin{bmatrix}0&0&0&\dots&0\\0&1&0&\dots&0\\0&0&1&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&1\end{bmatrix}\right\|_{1,\infty}
\end{aligned}$$

From the update rule, we have

$$y_{k+1,i} = \lambda_i(2y_{k,i} - y_{k-1,i} - \alpha h_{k,i} + \alpha h_{k-1,i}) + r_{k,i} = \lambda_i(2y_{k,i} - y_{k-1,i}) + \lambda_i\beta_{k,i} + r_{k,i}$$

where$\beta_{k,i} = -\alpha h_{k,i} + \alpha h_{k-1,i}$*, for all* y_i *with* $-\frac{1}{3} < \lambda_i < 0$*, from Lemma 13 we have*

$$y_{k+1,i} = y_{1,i}\left(\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i}\right) + \sum_{s=1}^{k}(\lambda_i\beta_{s,i} + r_{s,i})\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}$$

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

$$\begin{aligned}
\|y_{k+1,i}\|_\infty \leq \|y_{1,i}\|_\infty\left|\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i}\right| + |\lambda_i|\sum_{s=1}^{k}\|\beta_{s,i}\|_\infty\left|\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}\right| \\
+\sum_{s=1}^{k}\|r_{s,i}\|_\infty\left|\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}\right|
\end{aligned}$$

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

$$\|y_{k+1,i}\|_\infty \leq \|y_{1,i}\|_\infty |v_i|^k + |\lambda_i| \sum_{s=1}^{k} \|\beta_{s,i}\|_\infty |v_i|^{k-s} + \sum_{s=1}^{k} \|r_{s,i}\|_\infty |v_i|^{k-s}$$

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

$$\begin{aligned}\|\beta_{s,i}\|_\infty =& \|-\alpha h_{k,i} + \alpha h_{k-1,i}\|_\infty \leq 2\alpha(\|h_{k,i}\|_\infty + \|h_{k-1,i}\|_\infty) \\ \leq& 2\alpha(\|G_k\|_{1,\infty}\|Pe_i\|_1 + \|G_{k-1}\|_{1,\infty}\|Pe_i\|_1) \\ \leq& 2\alpha\sqrt{n}G_\infty\end{aligned}$$

For $r_{s,i}$, we have

$$\|r_{k,i}\|_\infty = \|\Omega_k Pe_i\|_\infty \leq \|\Omega_k\|_{1,\infty} \|Pe_i\|_1 \leq 2\sqrt{n}\delta\theta$$

when $\lambda_i < 0$, we have

$$\begin{aligned}\|y_{k+1,i}\|_\infty \leq& \|y_{1,i}\|_\infty |v_i|^k + |\lambda_i| \sum_{s=1}^{k} \|\beta_{s,i}\|_\infty |v_i|^{k-s} + \sum_{s=1}^{k} \|r_{s,i}\|_\infty |v_i|^{k-s} \\ \leq& \|y_{1,i}\|_\infty |v_n|^k + |\lambda_n| \sum_{s=1}^{k} \|\beta_{s,i}\|_\infty |v_n|^{k-s} + \sum_{s=1}^{k} \|r_{s,i}\|_\infty |v_n|^{k-s} \\ \leq& \alpha\sqrt{n}G_\infty |v_n|^k + 2\alpha\sqrt{n}G_\infty |\lambda_n| \sum_{s=1}^{\infty} |v_n|^{k-s} + 2\sqrt{n}\delta\theta \sum_{s=1}^{\infty} |v_n|^{k-s} \\ \leq& \alpha\sqrt{n}G_\infty |v_n| + \frac{2\alpha\sqrt{n}G_\infty |\lambda_n|}{1-|v_n|} + \frac{2\sqrt{n}\delta\theta}{1-|v_n|}\end{aligned}$$

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

On the other hand, when $0 \leq \lambda_i < 1$, from Lemma 13 we have

$$\begin{aligned}y_{k+1,i} \sin\theta_i =& y_{1,i}\lambda_i^{\frac{k}{2}} \sin[(t+1)\theta_i] + \lambda_i \sum_{s=1}^{k} \beta_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\theta_i] \\ &+ \sum_{s=1}^{k} r_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\theta_i]\end{aligned}$$

By taking norm, we get

$$\begin{aligned}\|y_{k+1,i}\|_\infty |\sin\theta_i| =& \|y_{1,i}\|_\infty \lambda_i^{\frac{k}{2}} |\sin[(t+1)\theta_i]| + \lambda_i \sum_{s=1}^{k} \|\beta_{s,i}\|_\infty |\lambda_i^{\frac{k-s}{2}}| |\sin[(k+1-s)\theta_i]| \\ &+ \sum_{s=1}^{k} \|r_{s,i}\|_\infty |\lambda_i^{\frac{k-s}{2}}| |\sin[(k+1-s)\theta_i]| \\ \leq& \|y_{1,i}\|_\infty \lambda_2^{\frac{k}{2}} + 2\alpha\sqrt{n}G_\infty\lambda_2 \sum_{s=1}^{\infty} \lambda_2^{\frac{s}{2}} + 2\sqrt{n}\delta\theta \sum_{s=1}^{\infty} \lambda_2^{\frac{s}{2}} \\ \leq& \alpha\sqrt{n}G_\infty\sqrt{\lambda_2} + \frac{2\alpha\sqrt{n}G_\infty\lambda_2 + 2\sqrt{n}\delta\theta}{\sqrt{1-\lambda_2}}\end{aligned}$$

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

$$\|y_{k+1,i}\| \leq \alpha\sqrt{n}G_\infty \sqrt{\frac{\lambda_2}{1-\lambda_2}} + \frac{2\alpha\sqrt{n}G_\infty\lambda_2 + 2\sqrt{n}\delta\theta}{1-\lambda_2}$$

So there exists D_1, D_2

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

$$D_2 = \max\left\{\frac{2}{1-|v_n|}, \frac{2}{\sqrt{1-\lambda_2}}\right\}$$

such that

$$\|y_{k+1,i}\|_\infty \leq D_1\alpha\sqrt{n}G_\infty + D_2\sqrt{n}\delta\theta$$

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

$$\|X_{k+1}(e_i - e_j)\|_\infty \leq D_1\alpha n G_\infty + D_2 n\delta\theta$$

As a result

$$\begin{aligned}
&\left\|X_{k+\frac{1}{2}}(e_i - e_j)\right\|_\infty \\
=&\left\|(2X_k - X_{k-1} - \alpha\widetilde{G}_k + \alpha\widetilde{G}_{k-1})(e_i - e_j)\right\|_\infty \\
\leq& 2\|X_k(e_i - e_j)\|_\infty + \|X_{k-1}(e_i - e_j)\|_\infty + \alpha\left\|\widetilde{G}_k\right\|_{1,\infty}\|e_i - e_j\|_1 + \alpha\left\|\widetilde{G}_{k-1}\right\|_{1,\infty}\|e_i - e_j\|_1 \\
\leq& 3(D_1\alpha n G_\infty + D_2 n\delta\theta) + 4\alpha G_\infty \\
\leq& (6D_1 n + 8)\alpha G_\infty
\end{aligned}$$

The last step is because $\delta = \frac{1}{6nD_2}$

Combining I and II we complete the proof.

Lemma 12

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

Proof

$$\begin{aligned}
\sum_{i=1}^n\left\|\overline{X}_k - x_{k,i}\right\|^2 &= \sum_{i=1}^n\left\|X_k\left(e_i - \frac{\mathbb{1}_n}{n}\right)\right\|^2 \\
&= \left\|X_k\left(I - \frac{\mathbb{1}_n\mathbb{1}_n^\top}{n}\right)\right\|_F^2 \\
&= \left\|X_kPP^\top - X_kv_1v_1^\top\right\|_F^2 \\
&\overset{\textit{Lemma 14}}{=} \left\|X_kP\begin{bmatrix}0&0&0&\dots&0\\0&1&0&\dots&0\\0&0&1&\dots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\0&0&0&\dots&1\end{bmatrix}\right\|_F^2 \\
&= \sum_{i=2}^n\|y_{k,i}\|^2
\end{aligned}$$

From the update rule, we obtain,

$$y_{k+1,i} = \lambda_i(2y_{k,i} - y_{k-1,i} - \alpha h_{k,i} + \alpha h_{k-1,i}) + r_{k,i} = \lambda_i(2y_{k,i} - y_{k-1,i}) + \lambda_i\beta_{k,i} + r_{k,i}$$

where$\beta_{k,i} = -\alpha h_{k,i} + \alpha h_{k-1,i}$*, for all* y_i *with* $-\frac{1}{3} < \lambda_i < 0$*, from Lemma 13 we have*

$$y_{k+1,i} = y_{1,i}\left(\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i}\right) + \sum_{s=1}^k(\lambda_i\beta_{s,i} + r_{k,i})\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}$$

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

$$\|y_{k+1,i}\|^2 \leq 3\|y_{1,i}\|^2 \left(\frac{u_i^{k+1} - v_i^{k+1}}{u_i - v_i}\right)^2 + 3\lambda_i^2 \left(\sum_{s=1}^{k} \|\beta_{s,i}\| \left|\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}\right|\right)^2 + 3\left(\sum_{s=1}^{k} \|r_{s,i}\| \left|\frac{u_i^{k-s+1} - v_i^{k-s+1}}{u_i - v_i}\right|\right)^2$$

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

$$\|y_{k+1,i}\|^2 \leq 3\|y_{1,i}\|^2 |v_i|^{2t} + 3\lambda_i^2 \left(\sum_{s=1}^{k} \|\beta_{s,i}\| |v_i|^{k-s}\right)^2 + 3\left(\sum_{s=1}^{k} \|r_{s,i}\| |v_i|^{k-s}\right)^2$$

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

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

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

On the other hand, when $0 \leq \lambda_i < 1$*, from Lemma 13 we have*

$$\begin{aligned}
y_{k+1,i} \sin\theta_i =& y_{1,i}\lambda_i^{\frac{k}{2}} \sin[(t+1)\theta_i] + \lambda_i \sum_{s=1}^{k} \beta_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\theta_i] \\
&+ \sum_{s=1}^{k} r_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\theta_i]
\end{aligned}$$

And we have

$$\begin{aligned}
\|y_{k+1,i}\|^2 \sin^2\theta_i \leq& 3\|y_{1,i}\|^2 \lambda_i^k \sin^2[(t+1)\theta_i] + 3\lambda_i^2 \left(\sum_{s=1}^{k} \beta_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\theta_i]\right)^2 \\
&+ 3\left(\sum_{s=1}^{k} r_{s,i}\lambda_i^{\frac{k-s}{2}} \sin[(k+1-s)\theta_i]\right)^2 \\
\leq& 3\|y_{1,i}\|^2 \lambda_i^k + 3\lambda_i^2 \left(\sum_{s=1}^{k} \beta_{s,i}\lambda_i^{\frac{k-s}{2}}\right)^2 + 3\left(\sum_{s=1}^{k} r_{s,i}\lambda_i^{\frac{k-s}{2}}\right)^2
\end{aligned}$$

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

$$\sum_{k=0}^{K-1} \|y_{k+1,i}\|^2 \sin^2\theta_i = \sum_{k=1}^{K} \|y_{k,i}\|^2 \sin^2\theta_i$$

$$
\begin{aligned}
\leq & 3\left\|y_{1,i}\right\|^2 \sum_{k=0}^{K-1} \lambda_i^t + 3\lambda_i^2 \sum_{k=1}^{K-1} \left(\sum_{s=1}^{k} \left\|\beta_{s,i}\right\| \lambda_i^{\frac{t-s}{2}} \right)^2 + 3 \sum_{k=1}^{K-1} \left(\sum_{s=1}^{k} r_{s,i} \lambda_i^{\frac{k-s}{2}} \right)^2 \\
\leq & \frac{3\left\|y_{1,i}\right\|^2}{1-\lambda_i} + \frac{3\lambda_i^2}{(1-\sqrt{\lambda_i})^2} \sum_{k=1}^{K-1} \left\|\beta_{k,i}\right\|^2 + \frac{3}{(1-\sqrt{\lambda_i})^2} \sum_{k=1}^{K-1} \left\|r_{k,i}\right\|^2
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{k=1}^{K} \left\|y_{k,i}\right\|^2 \leq & \frac{3\left\|y_{1,i}\right\|^2}{(1-\lambda_i)^2} + \frac{3\lambda_i^2}{(1-\sqrt{\lambda_i})^2(1-\lambda_i)} \sum_{k=1}^{K-1} \left\|\beta_{k,i}\right\|^2 + \frac{3}{(1-\sqrt{\lambda_i})^2(1-\lambda_i)} \sum_{k=1}^{K-1} \left\|r_{k,i}\right\|^2 \\
\leq & \frac{3\left\|y_{1,i}\right\|^2}{(1-\lambda_2)^2} + \frac{3\lambda_2^2}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)} \sum_{k=1}^{K-1} \left\|\beta_{k,i}\right\|^2 + \frac{3}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)} \sum_{k=1}^{K-1} \left\|r_{k,i}\right\|^2
\end{aligned}
$$

So there exists C_1, C_2, C_3

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

$$
\sum_{k=1}^{K} \left\|y_{k,i}\right\|^2 \leq C_1 \left\|y_{1,i}\right\|^2 + C_2 \sum_{k=1}^{K-1} \left\|\beta_{k,i}\right\|^2 + C_3 \sum_{k=1}^{K-1} \left\|r_{k,i}\right\|^2
$$

By taking expectation we have

$$
\sum_{k=1}^{K} \mathbb{E}\left\|y_{k,i}\right\|^2 \leq C_1 \mathbb{E}\left\|y_{1,i}\right\|^2 + C_2 \sum_{k=1}^{K-1} \mathbb{E}\left\|\beta_{k,i}\right\|^2 + C_3 \sum_{k=1}^{K-1} \mathbb{E}\left\|r_{k,i}\right\|^2
$$

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

$$
\begin{aligned}
& \sum_{i=2}^{n} \mathbb{E}\left\|\beta_{k,i}\right\|^2 \\
= & \alpha^2 \sum_{i=2}^{n} \mathbb{E}\left\|h_{k,i} - h_{k-1,i}\right\|^2 \\
= & \alpha^2 \sum_{i=2}^{n} \mathbb{E}\left\| \widetilde{G}_k P e_i - \widetilde{G}_{k-1} P e_i \right\|^2 \\
\leq & \alpha^2 \sum_{i=1}^{n} \mathbb{E}\left\| \widetilde{G}_k P e_i - \widetilde{G}_{k-1} P e_i \right\|^2 \\
\leq & \alpha^2 \mathbb{E}\left\| \widetilde{G}_k P - \widetilde{G}_{k-1} P \right\|_F^2 \\
\overset{\text{Lemma 14}}{\leq} & \alpha^2 \mathbb{E}\left\| \widetilde{G}_k - \widetilde{G}_{k-1} \right\|_F^2 \\
= & \alpha^2 \sum_{i=1}^{n} \mathbb{E}\left\| \widetilde{G}_k e_i - \widetilde{G}_{k-1} e_i \right\|^2 \\
\leq & 3\alpha^2 \sum_{i=1}^{n} \mathbb{E}\left\| \widetilde{G}_k e_i - G_k e_i \right\|^2 + 3\alpha^2 \sum_{i=1}^{n} \mathbb{E}\left\| \widetilde{G}_{k-1} e_i - G_{k-1} e_i \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|G_k e_i - G_{k-1}e_i\right\|^2 \\
\leq& 6\alpha^2 n\sigma^2 + 3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|G_k e_i - G_{k-1}e_i\right\|^2 \\
\leq& 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2\sum_{i=1}^{n}\mathbb{E}\left\|x_{k,i} - x_{k-1,i}\right\|^2 \\
\leq& 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2\sum_{i=1}^{n}\mathbb{E}\left\|Y_k P^\top e_i - Y_{k-1}P^\top e_i\right\|^2 \\
\leq& 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2\mathbb{E}\left\|Y_k P^\top - Y_{k-1}P^\top\right\|_F^2 \\
\overset{\text{Lemma 14}}{\leq}& 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2\mathbb{E}\left\|Y_k - Y_{k-1}\right\|_F^2 \\
\leq& 6\alpha^2 n\sigma^2 + 3\alpha^2 L^2\sum_{i=1}^{n}\mathbb{E}\left\|y_{k,i} - y_{k-1,i}\right\|^2
\end{aligned}
$$

And Putting it back, we have

$$
\begin{aligned}
&\sum_{i=2}^{n}\sum_{k=1}^{K}\mathbb{E}\left\|y_{k,i}\right\|^2 \\
\leq& C_1\mathbb{E}\left\|Y_1\right\|_F^2 + C_2\sum_{i=2}^{n}\sum_{k=1}^{K-1}\mathbb{E}\left\|\beta_{k,i}\right\|^2 + C_3\sum_{k=1}^{K-1}\sum_{i=2}^{n}\mathbb{E}\left\|r_{k,i}\right\|^2 \\
\leq& C_1\mathbb{E}\left\|Y_1\right\|_F^2 + C_2\sum_{k=1}^{K-1}\left(6\alpha^2 n\sigma^2 + 3\alpha^2 L^2\sum_{i=1}^{n}\mathbb{E}\left\|y_{k,i} - y_{k-1,i}\right\|^2\right) + C_3\sum_{k=1}^{K-1}\sum_{i=2}^{n}\mathbb{E}\left\|r_{k,i}\right\|^2 \\
\overset{\text{Lemma 14}}{\leq}& C_1\mathbb{E}\left\|Y_1\right\|_F^2 + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^2 L^2\sum_{k=1}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|y_{k,i} - y_{k-1,i}\right\|^2 + C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}
$$

Since

$$
\begin{aligned}
&\mathbb{E}\left\|y_{k,1} - y_{k-1,1}\right\|^2 = \mathbb{E}\left\|X_k P e_1 - X_{k-1}P e_1\right\|^2 = \mathbb{E}\left\|X_k v_1 - X_{k-1}v_1\right\|^2 \\
=& \mathbb{E}\left\|X_k\frac{1}{\sqrt{n}}\mathbb{1}_n - X_{k-1}\frac{1}{\sqrt{n}}\mathbb{1}_n\right\|^2 = n\mathbb{E}\left\|\overline{X}_k - \overline{X}_{k-1}\right\|^2 = n\alpha^2\mathbb{E}\left\|\overline{\widetilde{G}_k}\right\|^2 \\
\leq& n\alpha^2\mathbb{E}\left\|\overline{\widetilde{G}_k} - \overline{G}_k\right\|^2 + n\alpha^2\mathbb{E}\left\|\overline{G}_k\right\|^2 \leq n\alpha^2\frac{\sigma^2}{n} + n\alpha^2\mathbb{E}\left\|\overline{G}_k\right\|^2 \\
=& \alpha^2\sigma^2 + n\alpha^2\mathbb{E}\left\|\overline{G}_k\right\|^2
\end{aligned}
$$

Putting it back, and we obtain

$$
\begin{aligned}
&\sum_{i=2}^{n}\sum_{k=1}^{K}\mathbb{E}\left\|y_{k,i}\right\|^2 \\
\leq& C_1\mathbb{E}\left\|Y_1\right\|_F^2 + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^4\sigma^2 L^2 K + 3C_2\alpha^4 n L^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2 \\
&+3C_2\alpha^2 L^2\sum_{k=1}^{K-1}\sum_{i=2}^{n}\mathbb{E}\left\|y_{k,i} - y_{k-1,i}\right\|^2 + C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2 \\
\leq& C_1\mathbb{E}\left\|Y_1\right\|_F^2 + 6C_2\alpha^2 n\sigma^2 K + 3C_2\alpha^4\sigma^2 L^2 K + 3C_2\alpha^4 n L^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+6C_2\alpha^2L^2\sum_{k=1}^{K-1}\sum_{i=2}^{n}\mathbb{E}\left(\|y_{k,i}\|^2+\|y_{k-1,i}\|^2\right)+C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2\\
\leq&C_1\mathbb{E}\left\|Y_1\right\|_F^2+6C_2\alpha^2n\sigma^2K+3C_2\alpha^4\sigma^2L^2K+3C_2\alpha^4nL^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2\\
&+12C_2\alpha^2L^2\sum_{k=1}^{K-1}\sum_{i=2}^{n}\mathbb{E}\left\|y_{k,i}\right\|^2+C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}
$$

Rearrange the terms, we get

$$
\begin{aligned}
&(1-12C_2\alpha^2L^2)\sum_{i=2}^{n}\sum_{k=1}^{K}\mathbb{E}\left\|y_{k,i}\right\|^2\\
\leq&C_1\mathbb{E}\left\|Y_1\right\|_F^2+6C_2\alpha^2n\sigma^2K+3C_2\alpha^4\sigma^2L^2K+3C_2\alpha^4nL^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2+C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2\\
\leq&C_1\mathbb{E}\left\|X_1\right\|_F^2+6C_2\alpha^2n\sigma^2K+3C_2\alpha^4\sigma^2L^2K+3C_2\alpha^4nL^2\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2+C_3\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2
\end{aligned}
$$

Considering

$$
\begin{aligned}
\mathbb{E}\left\|X_1\right\|_F^2&=\alpha^2\mathbb{E}\left\|\widetilde{G}_0\right\|_F^2\\
&=\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\widetilde{G}_{0,i}-G_{0,i}+G_{0,i}-\nabla f(0)+\nabla f(0)\right\|^2\\
&\leq 3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\widetilde{G}_{0,i}-G_{0,i}\right\|^2+3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|G_{0,i}-\nabla f(0)\right\|^2+3\alpha^2\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(0)\right\|^2\\
&\leq 3\alpha^2n\sigma^2+3\alpha^2n\varsigma_0^2+3\alpha^2n\mathbb{E}\left\|\nabla f(0)\right\|
\end{aligned}
$$

We finally get

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

That completes the proof.

Lemma 13 *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*

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

we have

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

where

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

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

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

where

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

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

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

Recursively we have

$$\begin{aligned}
a_{t+1} - u a_t &= (a_t - u a_{t-1}) v + b_t - b_{t-1} + c_t \\
&= (a_{t-1} - u a_{t-2}) v^2 + (b_{t-1} - b_{t-2} + c_{t-1}) v + b_t - b_{t-1} + c_t \\
&= (a_1 - u a_0) v^t + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{t-s} \\
&= a_1 v^t + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{t-s}
\end{aligned}$$

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

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

Multiplying both sides by u^{t+1}

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

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

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

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

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

That completes the proof.

Lemma 14 *For any matrix $X \in \mathbb{R}^{N\times n}$, we have*

$$\sum_{i=2}^{n}\|Xv_i\|^2 \leq \sum_{i=1}^{n}\|Xv_i\|^2 = \|X\|_F^2$$
$$\sum_{i=1}^{n}\left\|XP^\top e_i\right\|^2 = \left\|XP^\top\right\|_F^2 = \|X\|_F^2$$

Proof

$$\sum_{i=2}^{n}\|X_tv_i\|^2 \leq \sum_{i=1}^{n}\|X_tv_i\|^2 = \|X_tP\|_F^2 = Tr(X_tPP^\top X_t^\top) = Tr(X_tX_t^\top) = \|X_t\|_F^2$$

And similarly,

$$\sum_{i=1}^{n}\left\|XP^\top e_i\right\|^2 = \left\|XP^\top\right\|_F^2 = Tr(X_tP^\top PX_t^\top) = Tr(X_tX_t^\top) = \|X_t\|_F^2$$

That completes the proof.

Lemma 15 *If we run Algorithm 2 for K iterations the following inequality holds:*

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

where

$$\begin{aligned}
C_1 &= \max\left\{\frac{3}{1-|v_n|^2}, \frac{3}{(1-\lambda_2)^2}\right\} \\
C_2 &= \max\left\{\frac{3\lambda_n^2}{(1-|v_n|)^2}, \frac{3\lambda_2^2}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\} \\
C_3 &= \max\left\{\frac{3}{(1-|v_n|)^2}, \frac{3}{(1-\sqrt{\lambda_2})^2(1-\lambda_2)}\right\} \\
C_4 &= 1 - 12C_2\alpha^2L^2 \\
\Omega_ke_i &= \sum_{j=1}^{n}\left((q_{k,j} - x_{k+\frac{1}{2},j}) - (q_{k,i} - x_{k+\frac{1}{2},i})\right)W_{ji}
\end{aligned}$$

Proof *Since*

$$\overline{X}_{k+1} = (2X_k - X_{k-1} - \alpha\widetilde{G}_k + \alpha\widetilde{G}_{k-1})W\frac{\mathbb{1}_n}{n} + (Q_k - X_{k+\frac{1}{2}})(W-I)\frac{\mathbb{1}_n}{n}$$

$$= 2\overline{X}_k - \overline{X}_{k-1} - \alpha\overline{\widetilde{G}}_k + \alpha\overline{\widetilde{G}}_{k-1}$$

and we have

$$\begin{aligned}\overline{X}_{k+1} - \overline{X}_k &= \overline{X}_k - \overline{X}_{k-1} - \alpha\overline{\widetilde{G}}_k + \alpha\overline{\widetilde{G}}_{k-1} \\ &= \overline{X}_1 - \overline{X}_0 - \alpha\sum_{t=1}^{k}(\overline{\widetilde{G}}_t - \overline{\widetilde{G}}_{t-1}) \\ &= -\alpha\overline{\widetilde{G}}_k\end{aligned}$$

As a result, we can reuse Lemma 8 from D-PSGD, thus we have

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

From Lemma 12 we obatin

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

Rearrange the terms, we get

$$\begin{aligned}&\left(1-\frac{3C_1\alpha^2L^2}{C_4}\right)\mathbb{E}\left\|\nabla f(0)\right\| + \left(1-\alpha L - 3\frac{C_2}{C_4}\alpha^4L^4\right)\frac{1}{K}\sum_{k=1}^{K-1}\mathbb{E}\left\|\overline{G}_k\right\|^2 \\ &+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2 \\ \leq&\frac{2(f(0)-f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2)}{C_4K} + 6\frac{C_2}{C_4}\alpha^2\sigma^2L^2 + 3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4 \\ &+\frac{C_3L^2}{C_4nK}\sum_{k=1}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2 \\ \overset{\text{Lemma 16}}{\leq}&\frac{2(f(0)-f^*)}{\alpha K} + \frac{\alpha L}{n}\sigma^2 + \frac{3C_1\alpha^2L^2(\sigma^2+\varsigma_0^2)}{C_4K} + 6\frac{C_2}{C_4}\alpha^2\sigma^2L^2 + 3\frac{C_2}{nC_4}\alpha^4\sigma^2L^4 \\ &+\frac{C_3L^2}{C_4}\left(\frac{3D_1n+4}{3D_2n}\right)^2\alpha^2G_\infty^2d\end{aligned}$$

That completes the proof.

Lemma 16

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2 \leq \left(\frac{3D_1n+4}{3D_2n}\right)^2\alpha^2G_\infty^2dnK$$

Proof *Similar to the case in D-PSGD, we have*

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left\|\Omega_k\right\|_F^2 &= \sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\sum_{j=1}^{n}\left((q_{k,j}-x_{k+\frac{1}{2},j})-(q_{k,i}-x_{k+\frac{1}{2},i})\right)W_{ji}\right\|^2 \\ &\overset{\text{Lemma 3}}{\leq} 4\sum_{k=0}^{K-1}\sum_{i=1}^{n}\delta^2\theta^2d \leq \left(\frac{3D_1n+4}{3D_2n}\right)^2\alpha^2G_\infty^2dnK\end{aligned}$$

That completes the proof.

H MONIQUA ON AD-PSGD (PROOF TO THEOREM 4)

H.1 ALGORITHM

Algorithm 3 Moniqua with Asynchronous Communication

Input: initial point $x_{0,i} = x_0$, step size α, the discrepency bound θ, number of iterations K, quantization function $\mathcal{Q}$, 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 W_k.
3: Compute a local stochastic gradient with model delayed by τ_k: $\widetilde{g}_{k-\tau_k,i_k}$
4: Compute modulo-ed model: $q_{k,i_k} \leftarrow \theta \cdot \mathcal{Q}_\delta\left(\frac{x_{k,i_k}}{\theta} \bmod 1\right)$ (element-wise division and mod)
5: Randomly select one of the neighbors j_k and average local weights with remote weights while subtracting the biased term: $x_{k+\frac{1}{2},i_k} \leftarrow x_{k,i_k} + \frac{1}{2}q_{k,j_k} - \frac{1}{2}q_{k,i_k}$
6: Update the local weight with local gradient: $x_{k+1,i_k} \leftarrow x_{k,i_k} - \alpha\widetilde{g}_{k-\tau_k,i_k}$
7: **end for**

Output: $\overline{X}_K = \frac{1}{n}\sum_{i=1}^n x_{K,i}$

H.2 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

$$f(x) = \sum_{i=1}^n p_i f_i(x)$$

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

$$\begin{aligned}
\nabla F(X_k) &= n\left[p_1 g_{k,1}, \cdots, p_n g_{k,n}\right] \in \mathbb{R}^{d\times n} \\
\nabla \widetilde{F}(X_k) &= n\left[p_1 \widetilde{g}_{k,1}, \cdots, p_n \widetilde{g}_{k,n}\right] \in \mathbb{R}^{d\times n} \\
\widetilde{G}_k &= \left[\cdots, \widetilde{g}_{k,i_k}, \cdots\right] \\
G_k &= \left[\cdots, g_{k,i_k}, \cdots\right] \\
\Lambda_a^b &= \frac{\mathbb{1}_n\mathbb{1}_n^\top}{n} - \prod_{q=a}^b W_q
\end{aligned}$$

H.3 ASSUMPTION

We makes the following assumptions:

1. **Lipschitzian Gradient**: All the function f_i have L-Lipschitzian gradients.
2. **Communication Matrix** [12]: The communication matrix W_k is doubly stochastic for any $k \geq 0$ and for any $b \geq a \geq 0$, there exists t_{mix} such that

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

3. **Bounded Variance**:

$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left\|\nabla\widetilde{f}_i(x_i;\xi_i) - \nabla f_i(x)\right\|^2 \leq \sigma^2, \forall i$$

[12] Please refer to Section E for more details

$$\mathbb{E}_{i\sim\{1,\cdots,n\}}\left\|\nabla f_i(x)-\nabla f(x)\right\|^2\leq\varsigma^2,\forall i$$

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

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 $\left\|\widetilde{g}_{k,i}\right\|_\infty\leq G_\infty$ for some constant G_∞.

H.4 Proof to Theorem 4

Proof *We start from*

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2+\left(1-\frac{2\alpha L}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla\overline{F}(X_{k-\tau_k})\right\|^2\\
\overset{\text{Lemma 20}}{\leq}&\frac{2n(f(0)-f^*)}{\alpha K}+\frac{(\sigma^2+6\varsigma^2)\alpha L}{n}+\left(2L^2+\frac{12\alpha L^3}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2\\
&+\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(X_k-X_{k-\tau_k})\mathbb{1}_n}{n}\right\|^2\\
\overset{\text{Lemma 21}}{\leq}&\frac{2n(f(0)-f^*)}{\alpha K}+\frac{(\sigma^2+6\varsigma^2)\alpha L}{n}+\frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2}+\frac{4\alpha^2T^2L^2}{n^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_ig_{k-\tau_k,i}\right\|^2\\
&+\left(2L^2+\frac{12\alpha L^3}{n}+\frac{24L^4\alpha^2T^2}{n^2}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2\\
\overset{\text{Lemma 19}}{\leq}&\frac{2n(f(0)-f^*)}{\alpha K}+\frac{(\sigma^2+6\varsigma^2)\alpha L}{n}+\frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2}+\frac{4\alpha^2T^2L^2}{n^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_ig_{k-\tau_k,i}\right\|^2\\
&+\frac{128\alpha^2t_{\text{mix}}^2L^2}{A_1}\left((\sigma^2+6\varsigma^2)p+\frac{2p}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_ig_{k-\tau_k,i}\right\|^2+G_\infty^2d\right)
\end{aligned}$$

where $A_1=1-192p\alpha^2t_{\text{mix}}^2L^2$ as defined in Lemma 19.

Rearrange the terms, we get

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

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

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

H.5 Lemma for Moniqua on AD-PSGD

Lemma 17

$$\mathbb{E}\left\|\widetilde{G}_{k-\tau_k}\frac{\mathbb{1}_n}{n}\right\|^2\leq\frac{\sigma^2}{n^2}+\frac{1}{n^2}\sum_{i=1}^{n}p_i\mathbb{E}\left\|g_{k-\tau_k,i}\right\|^2,\forall k\geq0.$$

Proof

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

Lemma 18

$$\sum_{i=1}^n p_i\mathbb{E}\left\|g_{k-\tau_k,i}\right\|^2 \leq 12L^2\sum_{i=1}^n p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2+6\varsigma^2+2\mathbb{E}\left\|\sum_{i=1}^n p_i g_{k-\tau_k,i}\right\|^2, \forall k\geq 0.$$

Proof

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

And

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

That completes the proof.

Lemma 19 *Let* $A_1 = 1 - 192p\alpha^2 t_{\text{mix}}^2 L^2$,

$$\begin{aligned}
&\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\| X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|^2 \\
\leq &\frac{32\alpha^2 t_{\text{mix}}^2}{A_1}\left((\sigma^2 + 6\varsigma^2)pK + 2p\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n} p_i g_{k-\tau_k,i}\right\|^2 + G_\infty^2 dK\right)
\end{aligned}$$

Proof

$$\begin{aligned}
&\sum_{i=1}^{n} p_i \mathbb{E}\left\| X_k\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|^2 \\
=&\sum_{i=1}^{n} p_i \mathbb{E}\left\| \left(X_{k-1}W_{k-1} - \alpha\widetilde{G}_{k-1-\tau_{k-1}} + \Omega_{k-1}\right)\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|^2 \\
\overset{X_0=0}{=}&\sum_{i=1}^{n} p_i \mathbb{E}\left\| \sum_{t=0}^{k-1}\left(-\alpha\widetilde{G}_{t-\tau_t} + \Omega_t\right)\Lambda_{t+1}^{k-1} e_i\right\|^2 \\
\leq &2\sum_{i=1}^{n} p_i \mathbb{E}\left\| \sum_{t=0}^{k-1}\alpha\widetilde{G}_{t-\tau_t}\Lambda_{t+1}^{k-1} e_i\right\|^2 + 2\sum_{i=1}^{n} p_i \mathbb{E}\left\| \sum_{t=0}^{k-1}\Omega_t\Lambda_{t+1}^{k-1} e_i\right\|^2
\end{aligned}$$

Now for the first term, we have

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

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

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

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

$$\begin{aligned}
&\sum_{k=0}^{K-1}\sum_{i=1}^{n} p_i \mathbb{E}\left\| X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|^2 \\
\leq &8p\alpha^2\sum_{k=0}^{K-1}\mathbb{E}\left(\sum_{t=0}^{k-\tau_k-1}\left\|\widetilde{G}_{t-\tau_t}\right\|_F 2^{-\left\lfloor\frac{k-\tau_k-t-1}{t_{\text{mix}}}\right\rfloor}\right)^2 \\
&+ 2\sum_{i=1}^{n} p_i \sum_{k=0}^{K-1}\mathbb{E}\left\| \sum_{t=0}^{k-\tau_k-1}\Omega_t\Lambda_{t+1}^{k-\tau_k-1} e_i\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left(\sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F 2^{-\left\lfloor \frac{k-\tau_k-t-1}{t_{\text{mix}}} \right\rfloor} \right)^2 \\
&\quad + 2\sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \mathbb{E} \left(\sum_{t=0}^{k-\tau_k-1} \left\| \Omega_t \right\|_{1,2} \left\| \Lambda_{t+1}^{k-\tau_k-1} \right\|_1 \left\| e_i \right\|_1 \right)^2 \\
&\leq 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left(\sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F 2^{-\left\lfloor \frac{k-\tau_k-t-1}{t_{\text{mix}}} \right\rfloor} \right)^2 \\
&\quad + 8\sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \mathbb{E} \left(\sum_{t=0}^{k-\tau_k-1} \left\| \Omega_t \right\|_{1,2} 2^{-\left\lfloor \frac{k-\tau_k-t-1}{t_{\text{mix}}} \right\rfloor} \right)^2 \\
&\overset{\text{Lemma 22}}{\leq} 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left(\sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F 2^{-\left\lfloor \frac{k-\tau_k-t-1}{t_{\text{mix}}} \right\rfloor} \right)^2 + 32t_{\text{mix}}^2 \sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \mathbb{E} \left\| \Omega_k \right\|_{1,2}^2 \\
&\leq 8p\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left(\sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F 2^{-\left\lfloor \frac{k-\tau_k-t-1}{t_{\text{mix}}} \right\rfloor} \right)^2 + 128\delta^2\theta^2 d t_{\text{mix}}^2 K \\
&\overset{\text{Lemma 22}}{\leq} 32p\alpha^2 t_{\text{mix}}^2 \sum_{k=0}^{K-1} \mathbb{E} \left\| \widetilde{G}_{k-\tau_k} \right\|_F^2 + 128\delta^2\theta^2 d t_{\text{mix}}^2 K
\end{aligned}
$$

Note that for the first term, we have

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

Putting these two terms back, we obtain

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left(\frac{\mathbb{1}_n}{n} - e_i \right) \right\|^2 \\
&\leq 32p\alpha^2 t_{\text{mix}}^2 \left((\sigma^2 + 6\varsigma^2)K + 12L^2 \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left(\frac{\mathbb{1}_n}{n} - e_i \right) \right\|^2 + 2\sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \right) \\
&+ 128\delta^2\theta^2 d t_{\text{mix}}^2 K
\end{aligned}
$$

Rearrange the terms, we obtain

$$
\begin{aligned}
&\left(1 - 192p\alpha^2 t_{\text{mix}}^2 L^2\right) \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left(\frac{\mathbb{1}_n}{n} - e_i \right) \right\|^2 \\
&\leq 32p\alpha^2 t_{\text{mix}}^2 \left((\sigma^2 + 6\varsigma^2)K + 2\sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \right) + 128\delta^2\theta^2 t_{\text{mix}}^2 K
\end{aligned}
$$

$$\overset{\text{Lemma 23}}{\leq} 32\alpha^2 t_{\text{mix}}^2 \left((\sigma^2 + 6\varsigma^2)pK + 2p \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 + G_\infty^2 dK \right)$$

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

$$\begin{aligned}
&\sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left(\frac{\mathbb{1}_n}{n} - e_i \right) \right\|^2 \\
\leq &\frac{32\alpha^2 t_{\text{mix}}^2}{A_1} \left((\sigma^2 + 6\varsigma^2)pK + 2p \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 + G_\infty^2 dK \right)
\end{aligned}$$

Lemma 20

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

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

$$\overline{X}_{k+1} = X_k W_k \frac{\mathbb{1}_n}{n} + (Q_k - X_k)(W_k - I)\frac{\mathbb{1}_n}{n} - \alpha \overline{\widetilde{G}}_{k-\tau_k} = \overline{X}_k - \alpha \overline{\widetilde{G}}_{k-\tau_k}$$

Then from Taylor Expansion, we have

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

Rearrange these terms, we can get

$$\begin{aligned}
&\frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(\overline{X}_k) \right\|^2 + \frac{\alpha}{2n} \mathbb{E} \left\| \nabla \overline{F}(X_{k-\tau_k}) \right\|^2 \\
&\leq \mathbb{E} f(\overline{X}_k) - \mathbb{E} f(\overline{X}_{k+1}) + \frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(\overline{X}_k) - \nabla \overline{F}(X_{k-\tau_k}) \right\|^2
\end{aligned}$$

$$+\frac{\alpha^2 L\sigma^2}{2n^2}+\frac{\alpha^2 L}{2n^2}\sum_{i=1}^{n}p_i\mathbb{E}\|g_{k-\tau_k,i}\|^2$$

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

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla\overline{F}(X_{k-\tau_k})\right\|^2\\&\leq\frac{2n(f(0)-f^*)}{\alpha K}+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)-\nabla\overline{F}(X_{k-\tau_k})\right\|^2+\frac{\alpha L\sigma^2}{n}+\frac{\alpha L}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\|g_{k-\tau_k,i}\|^2\end{aligned}$$

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

$$\begin{aligned}&\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)-\nabla\overline{F}(X_{k-\tau_k})\right\|^2\\\leq&2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)-\nabla f(\overline{X}_{k-\tau_k})\right\|^2+2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_{k-\tau_k})-\nabla\overline{F}(X_{k-\tau_k})\right\|^2\\=&2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)-\nabla f(\overline{X}_{k-\tau_k})\right\|^2+2\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i\left(\nabla f_i(\overline{X}_{k-\tau_k})-g_{k-\tau_k,i}\right)\right\|^2\\\leq&2\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)-\nabla f(\overline{X}_{k-\tau_k})\right\|^2+2\sum_{k=0}^{K-1}\mathbb{E}\sum_{i=1}^{n}p_i\left\|\nabla f_i(\overline{X}_{k-\tau_k})-g_{k-\tau_k,i}\right\|^2\\\leq&2L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(X_k-X_{k-\tau_k})\mathbb{1}_n}{n}\right\|^2+2L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2\end{aligned}$$

Putting it back, we have

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f(\overline{X}_k)\right\|^2+\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla\overline{F}(X_{k-\tau_k})\right\|^2\\\leq&\frac{2n(f(0)-f^*)}{\alpha K}+\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(X_k-X_{k-\tau_k})\mathbb{1}_n}{n}\right\|^2\\&+\frac{2L^2}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2+\frac{\alpha L\sigma^2}{n}+\frac{\alpha L}{nK}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\|g_{k-\tau_k,i}\|^2\\\overset{Lemma\ 18}{\leq}&\frac{2n(f(0)-f^*)}{\alpha K}+\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(X_k-X_{k-\tau_k})\mathbb{1}_n}{n}\right\|^2\\&+\frac{2L^2}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2+\frac{\alpha L\sigma^2}{n}\\&+\frac{\alpha L}{nK}\sum_{k=0}^{K-1}\left(12L^2\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2+6\varsigma^2+2\mathbb{E}\left\|\sum_{i=1}^{n}p_i g_{k-\tau_k,i}\right\|^2\right)\\=&\frac{2n(f(0)-f^*)}{\alpha K}+\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{(X_k-X_{k-\tau_k})\mathbb{1}_n}{n}\right\|^2\\&+\left(2L^2+\frac{12\alpha L^3}{n}\right)\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2\end{aligned}$$

$$+\frac{(\sigma^2+6\varsigma^2)\alpha L}{n}+\frac{2\alpha L}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_i g_{k-\tau_k,i}\right\|^2$$

Note that

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

Moving it to the left side, we finally get

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

That completes the proof.

Lemma 21 *For all $k\geq 0$, we have*

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

Proof *From Lemma 20, we know the fact*

$$\overline{X}_{k+1}=X_kW_k\frac{\mathbb{1}_n}{n}+(Q_k-X_k)(W_k-I)\frac{\mathbb{1}_n}{n}-\alpha\overline{\widetilde{G}}_{k-\tau_k}=\overline{X}_k-\alpha\overline{\widetilde{G}}_{k-\tau_k}$$

As a result

$$\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}\left\|(X_k-X_{k-\tau_k})\frac{\mathbb{1}_n}{n}\right\|^2\\
=&\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{t=1}^{\tau_k}\alpha\widetilde{G}_{k-t}\frac{\mathbb{1}_n}{n}\right\|^2\\
\leq&\alpha^2\sum_{k=0}^{K-1}\tau_k\sum_{t=1}^{\tau_k}\mathbb{E}\left\|\widetilde{G}_{k-t}\frac{\mathbb{1}_n}{n}\right\|^2\\
\leq&\alpha^2\sum_{k=0}^{K-1}\tau_k\sum_{t=1}^{\tau_k}\left(\frac{\sigma^2}{n^2}+\frac{1}{n^2}\sum_{i=1}^{n}p_i\mathbb{E}\left\|g_{k-t,i}\right\|^2\right)\\
\leq&\frac{\alpha^2T^2\sigma^2K}{n^2}+\frac{\alpha^2T}{n^2}\sum_{k=0}^{K-1}\sum_{t=1}^{\tau_k}\sum_{i=1}^{n}p_i\mathbb{E}\left\|g_{k-t,i}\right\|^2\\
\leq&\frac{\alpha^2T^2\sigma^2K}{n^2}+\frac{\alpha^2T}{n^2}\sum_{k=0}^{K-1}\sum_{t=1}^{\tau_k}\left(12L^2\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-t}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2+6\varsigma^2+2\mathbb{E}\left\|\sum_{i=1}^{n}p_i g_{k-t,i}\right\|^2\right)\\
\leq&\frac{\alpha^2T^2\sigma^2K}{n^2}+\frac{\alpha^2T^2}{n^2}\sum_{k=0}^{K-1}\left(12L^2\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2+6\varsigma^2+2\mathbb{E}\left\|\sum_{i=1}^{n}p_i g_{k-\tau_k,i}\right\|^2\right)
\end{aligned}$$

$$=\frac{\alpha^2T^2(\sigma^2+6\varsigma^2)K}{n^2}+\frac{12L^2\alpha^2T^2}{n^2}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2$$
$$+\frac{2\alpha^2T^2}{n^2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_ig_{k-\tau_k,i}\right\|^2$$

And we get

$$\frac{2L^2}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|(X_k-X_{k-\tau_k})\frac{\mathbb{1}_n}{n}\right\|^2$$
$$\leq\frac{2\alpha^2T^2(\sigma^2+6\varsigma^2)L^2}{n^2}+\frac{24L^4\alpha^2T^2}{n^2K}\sum_{k=0}^{K-1}\sum_{i=1}^{n}p_i\mathbb{E}\left\|X_{k-\tau_k}\left(\frac{\mathbb{1}_n}{n}-e_i\right)\right\|^2$$
$$+\frac{4\alpha^2T^2L^2}{n^2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\sum_{i=1}^{n}p_ig_{k-\tau_k,i}\right\|^2$$

That completes the proof.

Lemma 22 *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*

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

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

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

Proof

$$S_k=\sum_{t=1}^{k}a_t=\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor}b_s\leq\sum_{t=1}^{k}\sum_{s=1}^{t}\rho^{\max(\lfloor\frac{t-\tau_t-s}{T}\rfloor,0)}b_s=\sum_{s=1}^{k}\sum_{t=s}^{k}\rho^{\max(\lfloor\frac{t-\tau_t-s}{T}\rfloor,0)}b_s$$
$$=\sum_{s=1}^{k}\sum_{t=0}^{k-\tau_k-s}\rho^{\lfloor\frac{t}{T}\rfloor}b_s+\sum_{s=1}^{k}\sum_{t=1}^{\tau_k}\rho^0b_s\leq\sum_{s=1}^{k}\left(\sum_{t=0}^{T-1}\sum_{m=0}^{\infty}\rho^m\right)b_s+\tau_k\sum_{s=1}^{k}b_s\leq\left(T+\frac{T}{1-\rho}\right)\sum_{s=1}^{k}b_s$$
$$D_k=\sum_{t=1}^{k}a_t^2=\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor}b_s\sum_{r=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-r}{T}\rfloor}b_r=\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}\sum_{r=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor+\lfloor\frac{t-\tau_t-r}{T}\rfloor}b_sb_r$$
$$\leq\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}\sum_{r=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor+\lfloor\frac{t-\tau_t-r}{T}\rfloor}\frac{b_s^2+b_r^2}{2}=\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}\sum_{r=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor+\lfloor\frac{t-\tau_t-r}{T}\rfloor}b_s^2$$
$$\leq\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}b_s^2\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor}\sum_{r=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-r}{T}\rfloor}\leq\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}b_s^2\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor}\sum_{r=0}^{T-1}\sum_{m=0}^{\infty}\rho^m$$
$$cs6\leq\frac{T}{1-\rho}\sum_{t=1}^{k}\sum_{s=1}^{t-\tau_t}\rho^{\lfloor\frac{t-\tau_t-s}{T}\rfloor}b_s^2\overset{UsingS_k}{\leq}\frac{(2-\rho)T^2}{(1-\rho)^2}\sum_{s=1}^{k}b_s^2$$

Lemma 23 *for $\forall i,j$ and $\forall k\geq0$, we have*

$$\|X_k(e_i-e_j)\|_{\infty}\leq\theta=16t_{\text{mix}}\alpha G_{\infty}$$

Proof *Similar to Section F and Section G, we use mathmatical induction to prove this.*

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

$$\|X_k(e_i - e_j)\|_\infty = 0 \leq \theta = 16 t_{\text{mix}} \alpha G_\infty$$

II. Suppose for $k \geq 0$, we have $\|X_t(e_i - e_j)\|_\infty \leq \theta$, $\forall t \leq k$, then we have

$$\begin{aligned}
&\|X_{k+1}(e_i - e_j)\|_\infty \\
\leq &\left\| X_{k+1}\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|_\infty + \left\| X_{k+1}\left(\frac{\mathbb{1}_n}{n} - e_j\right)\right\|_\infty \\
\leq &\left\| X_{k+1}\left(I - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n}\right)\right\|_{1,\infty} \|e_i\|_1 + \left\| X_{k+1}\left(I - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n}\right)\right\|_{1,\infty} \|e_j\|_1 \\
= &2\left\| X_{k+1}\left(I - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n}\right)\right\|_{1,\infty} \\
\leq &2\left\| \left(X_k W_k - \alpha \widetilde{G}_{k-\tau_k} + \Omega_k\right)\left(\frac{\mathbb{1}_n}{n} - e_i\right)\right\|_{1,\infty} \\
= &2\left\| \sum_{t=0}^{k} \left(-\alpha \widetilde{G}_{t-\tau_t} + \Omega_t\right)\left(\prod_{q=t+1}^{k} W_q - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n}\right)\right\|_{1,\infty} \\
\leq &2\sum_{t=0}^{k} \left\| \left(-\alpha \widetilde{G}_{t-\tau_t} + \Omega_t\right)\left(\prod_{q=t+1}^{k} W_q - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n}\right)\right\|_{1,\infty} \\
\leq &2\sum_{t=0}^{k} \left\| -\alpha \widetilde{G}_{t-\tau_t} + \Omega_t\right\|_{1,\infty} \left\| \prod_{q=t+1}^{k} W_q - \frac{\mathbb{1}_n \mathbb{1}_n^\top}{n}\right\|_1 \\
\leq &4(\alpha G_\infty + 2\delta\theta) \sum_{t=0}^{k} 2^{-\lfloor (k-t)/t_{\text{mix}} \rfloor} \\
\leq &4(\alpha G_\infty + 2\delta\theta) \sum_{t=0}^{t_{\text{mix}}-1} \sum_{r=0}^{\infty} 2^{-r} \\
\leq &8(\alpha G_\infty + 2\delta\theta) t_{\text{mix}}
\end{aligned}$$

Put in $\delta = \frac{1}{32 t_{\text{mix}}}$, we obtain

$$\|X_{k+1}(e_i - e_j)\|_2 \leq 8(\alpha G_\infty + 2\delta\theta) t_{\text{mix}} = 8 t_{\text{mix}} \alpha G_\infty + 8 t_{\text{mix}} \alpha G_\infty = 16 t_{\text{mix}} \alpha G_\infty$$

Combining I and II and we complete the proof.