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. However, the communication bottleneck among workers when running distributed SGD presents a non-trivial challenge. State-of-the-art frameworks such as TensorFlow, CNTK, and MXNet are built in a centralized fashion, where workers exchange gradients either via a centralized parameter server or the MPI AllReduce operation. 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 and (2) quantized communication. 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. 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.

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. This happens because existing quantization techniques are mostly designed for centralized SGD, where workers communicate via exchanging gradients. 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. 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. Previous work...
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). 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 \( \theta \) (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 \( (m_2)_j \mod \theta (\forall j \in [d]) \). Note that given \( \|m_1 - m_2\|_\infty \leq \theta \):

\[
(m_2)_j \mod \theta - (m_1)_j \mod \theta = ((m_2)_j - (m_2)_j) \mod \theta = (m_2)_j - (m_2)_j
\]

so \( w_1 \) can obtain the \( j \)-th coordinate of \( m_2 - m_1 \) by locally computing \( (m_2)_j \mod \theta - (m_1)_j \mod \theta \) with \( (m_2)_j \mod \theta \) received from \( w_2 \). Since \( (m_2)_j \mod \theta \) 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 modular arithmetic for communication quantization 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 where \( \delta \) with one bit and shows great empirical success on speech recognition (Seide et al., 2014). Other work (Koloskova et al., 2019; Tang et al., 2018c; 2019).

In comparison, Moniqua is extra-memory-free. DeepSqueeze, which applies an error-compensation method (Wu et al., 2018b) to decentralized quantizes communication via estimating model difference. Furthermore, Tang et al. (2019) proposes QCD, which has bounded error \( \delta \), 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.

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 \( Q_\delta \) that has bounded error

\[
\|Q_\delta(x) - x\|_\infty \leq \delta, \quad \forall x \in [-1,1]^d
\]

where \( \delta \) is some constant. In general, a smaller \( \delta \) denotes more fine-grained quantization requiring more bits. For example, a biased linear quantizer can achieve \( \|Q_\delta(x) - x\|_\infty \leq \delta \) by rounding \( x \) to the nearest number in the set \( \{2\delta n \mid n \in \mathbb{Z}\} \); this will require about \( \delta^{-1} \) quantization points to cover the interval \([-1,1]\), so such a linear quantizer can satisfy \( \|Q_\delta(x) - x\|_\infty \leq \delta \) using only \( \lceil \log_2 (\frac{1}{\delta} + 1) \rceil \) bits (Li et al., 2017; Gupta et al., 2015). Note that \( \|Q_\delta(x) - x\|_\infty \leq \delta \) can be satisfied (for appropriate values of \( \delta \)) 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 \( D_i \) and collaborates to optimize

\[
\min_{x \in \mathbb{R}^d} \ell(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\xi \sim D_i} f_i(x; \xi)
\]

where \( \xi \) is data sample from \( D_i \). In each iteration of D-PSGD, worker \( i \) computes a local gradient sample using \( D_i \). 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 \( \tilde{g}_{k,i} \) denote local model and sampled gradient on worker \( i \) at \( k \)-th iteration, respectively. Let \( \alpha \) denote the step size. The update rule of D-PSGD can be expressed as:

\[
x_{k+1,i} = \sum_{j=1}^{n} x_{k,j} W_{ji} - \alpha \tilde{g}_{k,i} = x_{k,i} - \sum_{j=1}^{n} (x_{k,i} - x_{k,j}) W_{ji} - \alpha \tilde{g}_{k,i}
\]

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 $Q_δ$ can be represented by

$$x_{k+1,i} = x_{k,i}W_{i,i} + \sum_{j \neq i} Q_δ(x_{k,j})W_{j,i} - αg_{k,i}$$

(4)

Based on Equation 4, we obtain the following theorem.

**Theorem 1** For some constant $δ$, suppose that we use an unbiased linear quantizer $Q$ with representable points $\{δn \mid n \in \mathbb{Z} \}$ to learn on the quadratic objective function $f(x) = (x - δ/2)^2 (x - δ/2)/2$ 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 $\mathbb{E}[\|∇f(x_{k,i})\|^2 ≥ \frac{φ^2}{8(1+σ^2)}]$. 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 section, 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.

\[ \|∇f_i(x) - ∇f_i(y)\| ≤ L\|x - y\|, \forall x, y \in \mathbb{R}^d \]

(A2) Spectral gap. The communication matrix $W$ is a symmetric doubly stochastic matrix and $\max(|λ_2(W)|, |λ_n(W)|) = ρ < 1$, where $λ_i(W)$ denotes the $i$th eigenvalue of $W$.

(A3) Bounded variance. There exist non-negative $σ$ and $ς$ such that

$$\mathbb{E}_{ξ_i \sim D_i} \left[ \|∇f_i(x; ξ_i) - ∇f_i(x)\|^2 \right] ≤ σ^2, \quad \mathbb{E}_{ξ_i \sim \{1, \ldots, n\}} \|∇f_i(x) - ∇f(x)\|^2 ≤ ς^2$$

where $∇f_i(x; ξ_i)$ 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 $\|g_{k,i}\|_∞ ≤ G_∞$, 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 $Q_δ$, neighbor list $N_i$

1: for $k = 0, 1, 2, \ldots, K - 1$ do
2: Compute a local stochastic gradient $q_{k,i}$ with data sample $ξ_{k,i}$ and current weight $x_{k,i}$
3: Compute modulo-ed model: $q_{k,i} \leftarrow θ \cdot Q_δ \left( \frac{x_{k,i} - ξ_{k,i}}{θ} \mod 1 \right)$ (element-wise division and mod)
4: Average with neighboring workers: $x_{k+1,i} \leftarrow x_{k,i} + \sum_{j \in N_i} (q_{k,j} - q_{k,i})W_{j,i}$
5: Update the local weight with local gradient: $x_{k+1,i} \leftarrow x_{k+1,i} - αg_{k,i}$
6: end for

**Output:** Averaged model $x_K = \frac{1}{n} Σ_{i=1}^n x_{K,i}$
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 \(\theta\) could increase since local models may move further apart due to quantization error. However, with appropriately chosen \(\delta\), we can still obtain a bound \(\theta\) 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

\[
\theta = \frac{2\log(16n)\alpha G_\infty}{1 - \rho}, \quad \delta = \frac{1 - \rho}{4\log(16n)}, \quad \alpha = \frac{1}{\zeta^{2/3}K^{1/3} + \sqrt{nK} + 2L},
\]

then the output of Algorithm 1 converges at the asymptotic rate

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

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

**Consistent with D-PSGD.** Note that D-PSGD converges at the asymptotic rate of \(O(\sigma/\sqrt{nK} + \zeta^2/K^2 + n/K)\), 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 \(l_\infty\)-norm to bound sample variance and gradient magnitude, respectively. Note that, when \(d\) gets larger, the variance \(\sigma^2\) 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 (Q3). If we use nearest rounding (Gupta et al., 2015) as Q3 in Theorem 2, it suffices to use at each step a number of bits \(B\) for each parameter sent, where

\[
B = \left\lfloor \log_2 \left( \frac{1}{\zeta} + 1 \right) \right\rfloor = \left\lfloor \log_2 \left( \frac{4\log(16n)}{1 - \rho} + 1 \right) \right\rfloor
\]

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 \(\mathbb{E}_{x_1, \ldots, x_n} \left\| \nabla f(x) - \nabla f(x) \right\|^2\) is no longer bounded by \(\zeta^2\) as assumed in D-PSGD (Assumption (A3)). This update rule presented can be explicitly expressed in two steps:

\[
X_{k+\frac{1}{2}} = X_k - X_{k-1} - \alpha G_k + \alpha \tilde{G}_{k-1}
\]

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

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

**Theorem 3** If we run \(D^2\) with Moniqua in a setting where

\[
\theta = (6D_1n + 8)\alpha G_\infty, \quad \delta = \frac{1}{6nD_2}, \quad \alpha = \frac{1}{\sigma \sqrt{K/n + 2L}},
\]

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:

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

---

1 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 Asynchronous 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:

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

where $W_k$ describes the communication behaviour between the $k$th and $(k+1)$th gradient update, and $\tau_k$ 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 $t_{\text{mix}}$, 

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

We call this constant $t_{\text{mix}}$ 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

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

the output has the following asymptotic convergence rate:

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

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 MPIICH as the communication backend. All the instances are running Ubuntu 16.04, and latency and bandwidth on the underlying network are configured using the `top` 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 stochastic rounding with quantization step $\delta$.

5Details in the supplementary material.

6Since 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.
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(c) to 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 allow these algorithms obtain acceleration compared to 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.

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 model on CIFAR10. We compare Moniqua with the following baselines: 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 $\theta = 0.01$ for stochastic rounding across all algorithms that use quantization. To prevent overflow, we use 16-bit integer $\text{torch.int16}$ 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(c) to 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 allow these algorithms obtain acceleration compared to 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,

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

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.
with 3-bit communication Moniqua achieves 85% training accuracy while other baselines are below 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., 2018) 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: $Q_\delta(x) = \delta \lfloor \frac{x}{\delta} + u \rfloor$ (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 $\theta$ in Figure 4(a) (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 $\theta$ 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 5.07, 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


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 G to H, with corollaries contained in the corresponding sections. Detailed algorithm statements for applying Moniqua on D² and AD-PSGD can be found in Section G Algorithm 2 and Section H Algorithm 5, 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 Q be δ = 1. For any z quantized using Q, let z_f = z - ⌊z⌋, the variance of quantization error can be expressed as

$$\mathbb{E}(\|Q(z) - z\|^2) = (1 - z_f)^2 z_f^2 + z_f(1 - z_f)^2 \equiv z_f(1 - z_f)$$  \hspace{1cm} (5)

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

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

We now show for ∀x, y ∈ R^d

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

With out the loss of generality, let x - [x] ≤ y - [y]. Let x_f = x - [x] and y_f = y - [y], then

\begin{align*}
  [x + u] &= [x] & \text{and} & \quad [y + u] &= [y], \text{with probability} & \quad [y] - y \\
  [x + u] &= [x] & \text{and} & \quad [y + u] &= [y], \text{with probability} & \quad x - [x] \\
  [x + u] &= [x] & \text{and} & \quad [y + u] &= [y], \text{with probability} & \quad ([x] - x) - ([y] - y)
\end{align*}

Then we have

\begin{align*}
\mathbb{E}(\|Q(x) - x\|) - (Q(y) - y)) &= \mathbb{E}\left(\left(\delta \left[\frac{x}{\delta} + u\right] - x\right) - \left(\delta \left[\frac{y}{\delta} + u\right] - y\right)\right)^2 \\
&= ([y] - y)(([x] - x) - ([y] - y))^2 + (x - [x])([x] - x) - ([y] - y))^2 \\
&+ (([x] - x) - ([y] - y)(([x] - x) - ([y] - y))^2 \\
&= (1 - y_f)(y_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}(Q(y - x) - (y - x))^2
\end{align*}

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

$$\Delta = y - x$$

$$r = Q(\Delta) - \Delta$$

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

$$r_h = Q(\Delta_h) - \Delta_h$$
\[ = \delta \begin{cases} \frac{-\Delta_h}{\delta} + \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor + 1, & p_t \leq \frac{\Delta_h}{\delta} - \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor \\ \frac{-\Delta_h}{\delta} + \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor, & \text{otherwise} \end{cases} \]
\[ = \delta \begin{cases} -q + 1, & p_t \leq q \\ -q, & \text{otherwise} \end{cases} \]

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

Based on that, we have
\[ \mathbb{E} [r_h^2] \leq \delta^2 (q + 1)^2 + (q - 1)^2 (1 - q) \]
\[ = \delta^2 q (1 - q) \]
\[ \leq \delta^2 \min\{q, 1 - q\} \]

Since \( \min\{q, 1 - q\} \leq \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor \), we have
\[ \mathbb{E} [r_h^2] \leq \delta^2 \left\lfloor \frac{\Delta_h}{\delta} \right\rfloor \leq \delta |\Delta_h| \]

Summing over the index \( h \) yields,
\[ \mathbb{E} \|r\|_2^2 \leq \delta \mathbb{E} \|\Delta\|_1 \leq \sqrt{d} \delta \mathbb{E} \|\Delta\|_2 \]

Pushing back \( x \) and \( r \), we have
\[ \mathbb{E} \|Q(y - x) - (y - x)\|^2 \leq \sqrt{d} \delta \mathbb{E} \|y - x\| = \sqrt{d} \delta \mathbb{E} \|x - y\| \]

Putting it back we have
\[ \mathbb{E} \|(Q(x) - x) - (Q(y) - y)\|^2 \leq \sqrt{d} \delta \mathbb{E} \|x - y\| \]

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 \( BW \) 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):

\[ \#\text{bits} = \frac{BW}{\#\text{neighbors} \cdot \#\text{workers} \cdot \#\text{params of model}} \]

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

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)

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

where \( \delta \) 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.

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.

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

Table 2: Bandwidth consumption under different $\theta$ and $\delta$ when applying linear quantizer in Moniqua

<table>
<thead>
<tr>
<th>MODEL</th>
<th>MOD PARAM $\theta$</th>
<th>QUANT STEP</th>
<th>BYTES/EPOCH</th>
<th>AVG BITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16</td>
<td>None</td>
<td>None</td>
<td>45594MB</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.01</td>
<td>5206MB</td>
<td>3.65</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.01</td>
<td>5192MB</td>
<td>3.64</td>
</tr>
<tr>
<td>ResNet110</td>
<td>None</td>
<td>None</td>
<td>3430MB</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>0.01</td>
<td>609MB</td>
<td>5.67</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>0.01</td>
<td>608MB</td>
<td>5.67</td>
</tr>
</tbody>
</table>

2. Randomized Gossip (Non-linear)

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

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 preserve 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 $\delta$ and $\theta$ in the experiment of aggressive quantization. And we plot the results in Figure 5(a) and Figure 5(b)
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.

![Figure 5: More statistics from Experiment of Aggressive Quantization under 3-bit communication](image)

![Figure 6: More statistics from Experiment of Aggressive Quantization under 2-bit communication](image)

C.5 DECREASING STEP SIZE AND CONSENSUS ERROR

In this subsection, we provide more experimental results with decreasing step size. We also provide and discuss results on consensus error in this experiment. We run the experiments in the following setting:

**Models, Datasets and Hyperparameters.** We launch 8 workers connected using a ring network. We train ResNet110 and ResNet18 on CIFAR10. The hyperparameters of Moniqua are: ResNet110 (Initial step size \( = 0.05 \), \( \theta = 3.0 \), batch size = 128, weight decay = \( 3e^{-4} \), and momentum = 0.9) and ResNet18 (Initial step size = 0.1, \( \theta = 2.5 \), batch size = 128, weight decay = \( 1e^{-4} \), and momentum = 0.9). Step size is decreased (times a 0.1 factor) every 30 epochs. To be consistent with the original paper, we use the stochastic rounding to quantize each number.

**Results of Decreasing Step Size** We plot the results of test accuracy in Figure 7. We can see from Figure 7(a) that Moniqua requires at least 6 bits to achieve the comparable (accuracy drop < 0.3%) test accuracy as the baseline (D-PSGD with 32 bits). Once the number of bits decreases to 5, there is a gap between Moniqua and D-PSGD. On the other hand, other baselines including DeepSqueeze, ChocoSGD and DCD/ECD-PSGD are not able to achieve comparable test accuracy with 6 bits. Similarly, we can see from Figure 7(b) that when training ResNet18 with 4 bits, Moniqua is able to achieve comparable test accuracy after 120 epochs, while other baselines suffer a certain accuracy gap.

We also plot the test accuracy of different algorithms under different bit-level communication in Figure 8 (ResNet110). We can see that compared to the baselines, Moniqua is generally robust to low
Figure 7: Test Accuracy and Consensus Error of Moniqua under decreasing step size.

Figure 8: Test Accuracy of different algorithms on training ResNet110 on CIFAR10.

bits-budget. (Some of the dots are missing for some algorithms, that means they do not converge under the corresponding bits-budget.)

**Results of Consensus Error.** To better measure the behaviour of workers reaching consensus, we define the consensus error at iteration $k$: $C_k$ as follows (Notations are the same as in the original paper):

$$C_k = \frac{1}{nk} \sum_{i=0}^{k-1} \sum_{i=1}^{n} \left( \frac{1}{n} \sum_{j=1}^{n} x_{t,j} - x_{t,i} \right)^2$$  \hspace{1cm} (6)

Note that $C_k$ is essentially the running average of distance among workers and the averaged model. Trivially, a decreasing $C_k$ indicates the workers are reaching consensus. We measure the consensus error in three aspects. We first provide the results of the original paper in Figure 9(a) where constant step size is adopted. We can see that even with extremely small number of bits as used in “Aggressive Quantization”, all the workers are able to reach consensus. We further plot consensus error under the setting where decreasing step size is adopted as defined in this section. We can see in Figure 9(b) and Figure 9(c) that all the workers are able to reach consensus.
D Why Naive Quantization Fails in D-PSGD (Proof to Theorem 11)

The update rule of naive quantization on D-PSGD is

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

where \( \alpha_k \) is allowed to vary with any policy. Let

\[
X_k = [x_{k,1}, \ldots, x_{k,n}] \in \mathbb{R}^{d \times n} \\
\Omega_k = \left[ \sum_{j \neq 1} W_{j1} (Q(x_{k,j}) - x_{k,1}), \ldots, \sum_{j \neq n} W_{jn} (Q(x_{k,j}) - x_{k,n}) \right] \in \mathbb{R}^{d \times n} \\
\tilde{G}_k = [\tilde{g}_{k,1}, \ldots, \tilde{g}_{k,n}] \in \mathbb{R}^{d \times n}
\]

by rewriting the update rule, we obtain

\[ X_{k+1} = X_k + \Omega_k - \alpha_k \tilde{G}_k \]

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

\[ Y_{k+1} e_i = Y_k e_i + \Omega_k e_i - \alpha_k Y_k e_i + \alpha_k (\tilde{G}_k - G_k) e_i \]

where \( (\tilde{G}_k - G_k) \) 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 \( \forall k \geq K \),

\[
E \| Y_{k+1} e_i \|^2 \leq E \| Y_k e_i \|^2 < \frac{\delta^2 \phi^2}{8(1 + \phi^2)}
\]  

(7)

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

\[
\frac{\delta^2}{4} \leq E \| Q(x_{k,i}) - x^* \|^2 \leq 2E \| Q(x_{k,i}) - x_{k,i} \|^2 + 2E \| x_{k,i} - x^* \|^2
\]

As a result

\[
E \| Q(x_{k,i}) - x_{k,i} \|^2 \geq \frac{\delta^2}{8} - \frac{\phi^2 \delta^2}{8(1 + \phi^2)} = \frac{\delta^2}{8(1 + \phi^2)}
\]

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

\[
E \| \Omega_k e_i \|^2
\]
with the structure of a Markov Chain. As a result, we obtain the following lemma:

\[
\sum_{j \neq i} W_{ji} (Q(x_{k,j}) - x_{k,i})^2 \leq \sum_{j \in N_i} W_{ji} \mathbb{E} \| (Q(x_{k,j}) - x_{k,i})^2 \| + \sum_{m \neq n \neq i} \mathbb{E} \| (Q(x_{k,m}) - x_{k,i}) W_{mi}, (Q(x_{k,n}) - x_{k,i}) W_{ni} \|
\]

\[
\geq \phi^2 \sum_{j \in N_i} \mathbb{E} \| (Q(x_{k,j}) - x_{k,i})^2 \| + \sum_{m \neq n \neq i} \mathbb{E} \| (Q(x_{k,m}) - x_{k,i}) W_{mi}, (Q(x_{k,n}) - x_{k,i}) W_{mi} \|
\]

\[
\geq \frac{\phi^2 \delta^2}{8(1 + \phi^2)}
\]

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

\[
\mathbb{E} \| Y_{k+1} e_i \|^2 = \mathbb{E} \| Y_k + \Omega_k - \alpha_k Y_k + \alpha_k (G_k - G_k) \|^2
\]

\[
= \mathbb{E} \| (1 - \alpha_k) Y_k e_i \|^2 + \mathbb{E} \| \Omega_k e_i \|^2 + \mathbb{E} \| \alpha_k (G_k - G_k) e_i \|^2
\]

\[
\geq \mathbb{E} \| \Omega_k e_i \|^2 + \frac{\phi^2 \delta^2}{8(1 + \phi^2)}
\]

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

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

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 \( \frac{1}{n} \) as its the stationary distribution \( (W^T \frac{1}{n} = \frac{1}{n}) \). 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.\(^{[3]}\)

### 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**

\[
\left\| W^t \left( I - \frac{1}{n} \frac{1}{n}^T \right) \right\|_1 \leq 2 \cdot 2^{- \left\lfloor \frac{t}{t_{mix}} \right\rfloor}
\]

**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,

\[
x = [2, -1]^T
\]

\(^{[3]}\)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](https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf)
\[ u = [2, 0]^T \]
\[ v = [0, 1]^T \]

And we have
\[
\| W^t \left( I - \frac{1_n 1_n^T}{n} \right) x \|_1
\]
\[
= \| W^t \left( I - \frac{1_n 1_n^T}{n} \right) (u - v) \|_1
\]
\[
\leq \| W^t \left( I - \frac{1_n 1_n^T}{n} \right) u \|_1 + \| W^t \left( I - \frac{1_n 1_n^T}{n} \right) v \|_1
\]
\[
= 1_n^T u \| W^t u \frac{1_n}{n} - \frac{1_n}{n} \|_1 + 1_n^T v \| W^t v \frac{1_n}{n} - \frac{1_n}{n} \|_1
\]
\[
\leq 2(1_n^T u + 1_n^T v) d(t)
\]
\[
\leq 2d(t) \| x \|_1
\]

Considering the definition of L1-norm, we have
\[
\| W^t \left( I - \frac{1_n 1_n^T}{n} \right) \|_1 = \max \frac{\| W^t \left( I - \frac{1_n 1_n^T}{n} \right) x \|_1}{\| x \|_1} \leq 2d(t)
\]

According to a well-known results on the theory of Markov Chains,\(^{10}\) \(d(t_{\text{mix}}) \leq 2^{-t} \) holds for any non-negative integer \( t \), so we have
\[
\| W^t \left( I - \frac{1_n 1_n^T}{n} \right) \|_1 \leq 2d(t) \leq 2d \left( \frac{t}{t_{\text{mix}}} \cdot t_{\text{mix}} \right) \leq 2d \left( \left\lfloor \frac{t}{t_{\text{mix}}} \right\rfloor \right) \leq 2 \cdot 2^{-\frac{1}{t_{\text{mix}}}}
\]

That completes the proof.

Additionally, based on standard results in the theory of reversible Markov Chains, we also have\(^{11}\)
\[
t_{\text{mix}} \leq \log \left( \frac{1/2}{1 - \rho} \right) \frac{1}{1 - \rho} \leq \log(4n) \frac{1}{1 - \rho}.
\]

### 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 \( t_{\text{mix}} \) (here \( t_{\text{mix}} \) is the same as \( t_{\text{mix}} \) in the paper) and for any \( k \) and any non-negative vector \( \mu \in \mathbb{R}^d \) such that \( 1_n^T \mu = 1 \), it must hold that
\[
\left\| \left( \prod_{i=1}^{t_{\text{mix}}} W_{k+i} \right) \mu - \frac{1_n}{n} \right\|_1 \leq \frac{1}{2}.
\]

We call this constant \( t_{\text{mix}} \) 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 \( t_{\text{mix}} \) such that
\[
\left\| \prod_{q=a}^{b} W_q \left( I - \frac{1_n 1_n^T}{n} \right) \right\|_1 \leq 2 \cdot 2^{-\frac{1+1}{1_{\text{mix}}}}
\]

\(^{10}\) 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.
As a result, from the submultiplicativity of the matrix induced norm, we obtain

\[ \left\| \prod_{q=1}^{t \text{mix}} W_{q+k} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) x \right\|_1 = \left\| \prod_{q=1}^{t \text{mix}} W_{q+k} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) (u - v) \right\|_1 \leq \left\| \prod_{q=1}^{t \text{mix}} W_{q+k} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) u \right\|_1 + \left\| \prod_{q=1}^{t \text{mix}} W_{q+k} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) v \right\|_1 \]

\[ = \mathbb{1}_n^\top u \left\| \prod_{q=1}^{t \text{mix}} W_{q+k} \frac{u - \mathbb{1}_n}{n} \right\|_1 + \mathbb{1}_n^\top v \left\| \prod_{q=1}^{t \text{mix}} W_{q+k} \frac{v - \mathbb{1}_n}{n} \right\|_1 \leq \frac{1}{2} (\mathbb{1}_n^\top u + \mathbb{1}_n^\top v) \leq \frac{1}{2} \|x\|_1 \]

Considering the definition of the induced \( \ell_1 \) operator norm, we have

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

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

\[ \left\| \prod_{q=a}^{b} W_q \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) \right\|_1 \leq \prod_{q=1}^{t \text{mix}} W_{a-1+q} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) \prod_{q=1}^{t \text{mix}} W_{a+q} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) \prod_{q=1}^{t \text{mix}} W_{a+q} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) \leq 2^{\frac{b-a+1}{t \text{mix}}} \left\| \prod_{q=1}^{t_r} W_{a+q} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) \right\|_1 \]

where \( t_r = (b-a+1) \mod t \text{mix}. \) Note that

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

Putting it back we obtain

\[ \left\| \prod_{q=a}^{b} W_{a+q} \left( I - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^\top \right) \right\|_1 \leq 2 \cdot 2^{\frac{b-a+1}{t \text{mix}}} \]

That completes the proof.

Note that in the analysis of Moniqua on AD-PSGD (Section F), 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 \( \theta \cdot Q_\delta \left( \frac{x}{\theta} \mod 1 \right) \) instead of \( Q_\delta \left( x \mod \theta \right) \) for theoretical analysis and how it covers both linear and non-linear quantizers. Note that typically, a linear quantizer has:

\[ \| Q_\delta(x) - x \|_\infty \leq \delta, \quad \forall x \in \mathbb{R}^d \]
while a non-linear quantizer has
\[ \|Q_\delta(x) - x\|_\infty \leq \delta \|x\|_\infty, \quad \forall x \in \mathbb{R}^d \]
so that for linear quantizer, with a given \( x \leq \theta \):
\[ \left\| \theta \cdot Q_\delta \left( \frac{x}{\theta} \text{ mod } 1 \right) - x \right\|_\infty = \theta \left\| Q_\delta \left( \frac{x}{\theta} \text{ mod } 1 \right) - \left( \frac{x}{\theta} \text{ mod } 1 \right) \right\|_\infty \leq \theta \delta = \theta \delta \]
And for non-linear quantizer, with a given \( x \leq \theta \):
\[ \left\| \theta \cdot Q_\delta \left( \frac{x}{\theta} \text{ mod } 1 \right) - x \right\|_\infty = \theta \left\| Q_\delta \left( \frac{x}{\theta} \text{ mod } 1 \right) - \left( \frac{x}{\theta} \text{ mod } 1 \right) \right\|_\infty \leq \theta \delta \cdot 1 = \theta \delta \]
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
\[
\begin{align*}
X_k &= [x_{k,1}, \cdots, x_{k,n}], & Q_k &= [q_{k,1}, \cdots, q_{k,n}], \\
\tilde{G}_k &= [\tilde{g}_{k,1}, \cdots, \tilde{g}_{k,n}], & G_k &= [g_{k,1}, \cdots, g_{k,n}], \\
\overline{X} &= \frac{1}{n} X, & \forall X \in \mathbb{R}^{d \times n}, & \Omega_k &= (Q_k - X_k)(W - I)
\end{align*}
\]
where \( g_{k,i} \) denotes gradient computed via the whole dataset \( D_t \) and \( x_{k,i} \)

From a local view, the update rule of Algorithm 4 on worker \( i \) at iteration \( k \) can be written as
\[ x_{k+1,i} = x_{k,i} + \sum_{j \in N_i} (q_{k,j} - q_{k,i}) W_{ji} - \alpha \tilde{g}_{k,i} \]
which is equivalent to
\[ x_{k+1,i} = x_{k,i} + \sum_{j=1}^{n} (x_{k,j} - x_{k,i}) W_{ji} - \alpha \tilde{g}_{k,i} + \sum_{j=1}^{n} ((q_{k,j} - x_{k,j}) - (q_{k,i} - x_{k,i})) W_{ji} \] (8)

From a global view, the update rule can be written as
\[ X_{k+1} = X_k + Q_k(W - I) - \alpha \tilde{G}_k = X_k W - \alpha \tilde{G}_k + (Q_k - X_k)(W - I) \] (9)

From Lemma 2 we have
\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|_p^2 \leq \frac{4(f(0) - f^*)}{\alpha K} + \frac{2 \alpha L}{n} \sigma^2 + \frac{8 \alpha^2 L^2 (\sigma^2 + 3 \varsigma^2)}{(1 - \rho)^2} + \frac{8 L^2}{n K (1 - \rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \left\| \Omega_k \right\|_F^2 \]

Note that
\[ \sum_{k=0}^{K-1} \mathbb{E} \left\| \Omega_k \right\|_F^2 = \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{j=1}^{n} ((q_{k,j} - x_{k,j}) - (q_{k,i} - x_{k,i})) W_{ji} \right\|^2 \]
\[ \leq \frac{1}{2} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \delta^2 \theta^2 d \leq \alpha^2 G_\infty^2 d n K \]

The last step holds because \( \delta \theta = \frac{1}{2} \alpha G_\infty \). Pushing it back we obtain
\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|_p^2 \leq \frac{4(f(0) - f^*)}{\alpha K} + \frac{2 \alpha L}{n} \sigma^2 + \frac{8 \alpha^2 L^2 (\sigma^2 + 3 \varsigma^2)}{(1 - \rho)^2} + \frac{8 \alpha^2 G_\infty^2 d L^2}{(1 - \rho)^2} \]
By setting \( \alpha = \frac{1}{\zeta^2 K + \sigma \sqrt{\frac{F}{\theta}} + 2L} \), we have

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 \leq \frac{8(f(0) - f^*)L}{K} + \frac{4\sigma(f(0) - f^* + L/2)}{\sqrt{nK}} + \frac{4\zeta^2(f(0) - f^*)}{K^2} \\
+ \frac{8L^2\sigma^2n}{(1 - \rho)^2(\sigma^2 K + 4nL^2)} + \frac{24L^2\zeta^2}{(1 - \rho)^2K^2} + \frac{8L^2G^2_\infty}{nK} + n
\]

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

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

Proof

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

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

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

Proof We use mathematical induction to prove this:

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

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

\[
\|X_{k+1}(e_i - e_j)\|_\infty = \| \left( X_k W - \alpha \tilde{G}_k + \Omega_k \right) (e_i - e_j) \|_\infty \\
X_0 = \sum_{t=0}^{k} \left( -\alpha \tilde{G}_t + \Omega_t \right) W^{k-t}(e_i - e_j) \|_\infty
\]
As a result, we have

\[
\begin{align*}
\|X_{k+1}(e_i - e_j)\|_\infty & \leq \sum_{t=0}^{K-1} \frac{1}{K} \mathbb{E} \left\| \nabla f(X_k) \right\|_2 \\
& \leq \frac{4(f(0) - f^*)}{\alpha K} + \frac{2\alpha L^2 n \sigma^2}{\alpha} + \frac{8\alpha^2 L^2 (\sigma^2 + \varsigma^2)}{(1 - \rho)^2} + \frac{8L^2}{nK(1 - \rho)} \sum_{k=1}^{K-1} \mathbb{E} \left\| \Omega_k \right\|^2_F 
\end{align*}
\]

Combining I and II, we complete the proof.

**Lemma 5** The output of Algorithm [7] has the following bound:

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|_2^2 \leq \frac{4(f(0) - f^*)}{\alpha K} + \frac{2\alpha L^2 n \sigma^2}{\alpha} + \frac{8\alpha^2 L^2 (\sigma^2 + \varsigma^2)}{(1 - \rho)^2} + \frac{8L^2}{nK(1 - \rho)} \sum_{k=1}^{K-1} \mathbb{E} \left\| \Omega_k \right\|^2_F
\]

For any \( t \geq 0 \), on one hand

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

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

\[
\|W^t(e_i - e_j)\|_1 \leq \mathbb{I}^T_n W^t e_i + \mathbb{I}^T_n W^t e_i = \mathbb{I}^T_n e_i + \mathbb{I}^T_n e_j = 2
\]

So

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

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

\[
\sum_{t=0}^{\infty} \|W^t(e_i - e_j)\|_1 = \sum_{t=0}^{T_0-1} \|W^t(e_i - e_j)\|_1 + \sum_{t=T_0}^{\infty} \|W^t(e_i - e_j)\|_1 \\
\leq T_0 - 1 + \sum_{t=T_0}^{\infty} 2\sqrt{n} \rho^t + T_0 \\
\leq 2 \left\lceil \frac{-\log(\sqrt{n})}{\log(\rho)} \right\rceil + \sum_{t=0}^{\infty} \left\lceil \frac{-\log(\sqrt{n})}{\log(\rho)} \right\rceil \rho^t \\
\leq 2 \frac{\log(\sqrt{n})}{1 - \rho} + 2 + \frac{2}{1 - \rho} \\
\leq \frac{\log(16n)}{1 - \rho}
\]

As a result, we have

\[
\|X_{k+1}(e_i - e_j)\|_\infty \leq (\alpha G_\infty + 2\delta \beta) \frac{\log(16n)}{1 - \rho}
\]

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

\[
\|X_{k+1}(e_i - e_j)\|_\infty \leq (\alpha G_\infty + 2\delta \beta) \frac{\log(16n)}{1 - \rho} \leq \frac{2\log(16n)}{1 - \rho} - \alpha G_\infty = \theta
\]
Proof: From Lemma 5 we have

\[
\frac{1 - \alpha L}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X_k)\|^2 + \frac{1}{R} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \\
\leq \frac{2\alpha_0 (f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \mathbb{E} \|\nabla f(X_k)\|^2 + L^2 \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \|x_k - x_{k,i}\|^2
\]

Lemma 5 \[2\alpha_0 (f(0) - f^*) \leq \frac{\alpha L}{n} \mathbb{E} \|\nabla f(X_k)\|^2 + \frac{2\alpha_0^2 L^2}{M_1 (1 - \rho)^2} \left( \sigma^2 + 3\rho^2 + \frac{3}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \right) + \frac{2L^2}{M_1 nK (1 - \rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\]

where

\[M_1 = 1 - \frac{6\alpha_0^2 L^2}{(1 - \rho)^2}\]

Rearrange the terms, we get

\[
\frac{1 - \alpha L}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X_k)\|^2 + \left( 1 - \frac{6\alpha_0^2 L^2}{M_1 (1 - \rho)^2} \right) \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \\
\leq \frac{2\alpha_0 (f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{2\alpha_0^2 L^2}{M_1 (1 - \rho)^2} \left( \sigma^2 + 3\rho^2 + \frac{3}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \right) + \frac{2L^2}{M_1 nK (1 - \rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\]

Let

\[M_2 = 1 - \frac{6\alpha_0^2 L^2}{M_1 (1 - \rho)^2}\]

we get

\[
\frac{1 - \alpha L}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X_k)\|^2 + \frac{M_2}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \\
\leq \frac{2\alpha_0 (f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{2\alpha_0^2 L^2}{M_1 (1 - \rho)^2} \left( \sigma^2 + 3\rho^2 + \frac{3}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \right) + \frac{2L^2}{M_1 nK (1 - \rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\]

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

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X_k)\|^2 \leq \frac{4\alpha_0 (f(0) - f^*)}{\alpha K} + \frac{2\alpha L}{n} \sigma^2 + \frac{8\alpha_0^2 L^2}{(1 - \rho)^2} \left( \sigma^2 + 3\rho^2 + \frac{3}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \right) \\
+ \frac{8L^2}{nK (1 - \rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\]

and that completes the proof.

Lemma 6 Let \(M_1 = 1 - \frac{6\alpha_0^2 L^2}{(1 - \rho)^2} > 0\), we have

\[
\frac{L^2}{nK} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \|x_k - x_{k,i}\|^2 \leq \frac{2\alpha_0^2 L^2}{M_1 (1 - \rho)^2} \left( \sigma^2 + 3\rho^2 + \frac{3}{K} \sum_{k=0}^{K-1} \mathbb{E} \|f(X_k) - f^*\|^2 \right) \\
+ \frac{2L^2}{M_1 nK (1 - \rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\]

Proof

\[
\sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \|x_k - x_{k,i}\|^2
\]
\[= \sum_{k=1}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| X_k \left( \frac{I_n}{n} - e_i \right) \right\|^2 \]

\[= \sum_{k=1}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| (X_{k-1}W - \alpha \bar{G}_{k-1} + \Omega_{k-1}) \left( \frac{I_n}{n} - e_i \right) \right\|^2 \]

\[= \sum_{k=1}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{t=0}^{k-1} (-\alpha \bar{G}_t + \Omega_t) \left( \frac{I_n}{n} - W^{k-t-1}e_i \right) \right\|^2 \]

\[\leq 2\alpha^2 \sum_{k=1}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{t=0}^{k-1} \frac{1}{n} - W^{k-t-1}e_i \right\|^2 \]

\[= 2\alpha^2 \sum_{k=1}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left( F \sum_{t=0}^{k-1} \| \bar{G}_t \|_F \right)^2 + 2 \sum_{k=1}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left( \| \Omega_t \|_F \right)^2 \]

**Lemma 7**

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

\[\frac{L^2}{nK} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| \bar{X}_k - x_{k,i} \right\|^2 \leq \frac{2\alpha^2 L^2}{M_1(1-\rho)^2} \left( \| \bar{G}_k \|_F \right)^2 + \frac{2}{(1-\rho)^2} \sum_{k=1}^{K-1} \mathbb{E} \left\| \Omega_k \right\|^2 \]

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

\[\mathbb{E} \| \bar{G}_k \|_F^2 = \sum_{i=1}^{n} \mathbb{E} \| \bar{g}_{k,i} \|^2 \]
Next, we derive the upper bound of $\mathbb{E} \| \tilde{g}_{k,i} \|^2$

$$\mathbb{E} \| \tilde{g}_{k,i} \|^2 = \mathbb{E} \| \tilde{g}_{k,i} - g_{k,i} + g_{k,i} \|^2 = \mathbb{E} \| \tilde{g}_{k,i} - g_{k,i} \|^2 + \mathbb{E} \| g_{k,i} \|^2 + 2\mathbb{E} \langle \tilde{g}_{k,i} - g_{k,i}, g_{k,i} \rangle = \mathbb{E} \| \tilde{g}_{k,i} - g_{k,i} \|^2 + \mathbb{E} \| g_{k,i} \|^2$$

$$\leq \sigma^2 + 3\mathbb{E} \| g_{k,i} - \nabla f_i(\bar{X}_k) \|^2 + 3\mathbb{E} \| \nabla f_i(\bar{X}_k) \|^2 + 3\mathbb{E} \| \nabla f(\bar{X}_k) \|^2$$

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

$$\sum_{k=0}^{K-1} \mathbb{E} \| G_k \|^2_F = \sum_{k=0}^{K-1} \mathbb{E} \| \tilde{g}_{k,i} \|^2 \leq \sum_{k=0}^{K-1} \sum_{i=1}^n \mathbb{E} \| \tilde{g}_{k,i} \|^2 \leq \sum_{k=0}^{K-1} \sum_{i=1}^n \sigma^2 + 3L^2 \sum_{k=0}^{K-1} \sum_{i=1}^n \mathbb{E} \| \bar{X}_k - x_{k,i} \|^2 + 3\sum_{k=0}^{K-1} \sum_{i=1}^n \mathbb{E} \| \nabla f(\bar{X}_k) \|^2$$

That completes the proof

**Lemma 8**

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

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

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

And by Taylor Expansion, we have

$$\mathbb{E} f(\bar{X}_{k+1}) = \mathbb{E} f\left( \frac{(X_k W - \alpha \bar{G}_k + \Omega_k) \mathbb{1}_n}{n} \right) = \mathbb{E} f\left( \bar{X}_k - \alpha \bar{G}_k \right) \leq \mathbb{E} f(\bar{X}_k) - \alpha \mathbb{E} (\nabla f(\bar{X}_k), \bar{G}_k) + \frac{\alpha^2 L}{2} \mathbb{E} \| \bar{G}_k \|^2$$

And for the last term, we have

$$\mathbb{E} \| \bar{G}_k \|^2 = \mathbb{E} \left\| \sum_{i=1}^n \frac{\tilde{g}_{k,i}}{n} \right\|^2 \leq \mathbb{E} \left\| \sum_{i=1}^n \frac{\tilde{g}_{k,i} - \frac{1}{n} \sum_{i=1}^n g_{k,i}}{n} + \frac{1}{n} \sum_{i=1}^n g_{k,i} \right\|^2 \leq \mathbb{E} \left\| \sum_{i=1}^n \frac{\tilde{g}_{k,i} - \frac{1}{n} \sum_{i=1}^n g_{k,i}}{n} \right\|^2 + \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n g_{k,i} \right\|^2$$
That completes the proof. Putting it back, we obtain

$$
\begin{align*}
Ef(X_{k+1}) & \leq Ef(X_k) - \alpha Ef(\nabla f(X_k), G_k) + \alpha^2 L \sigma^2 + \alpha^2 L/2 \sum_{i=1}^{n} \|g_{k,i}\|^2 \\
& = Ef(X_k) - \alpha - \alpha^2 L \|\nabla f(X_k)\|^2 - \alpha/2 \sum_{i=1}^{n} \|\nabla f(X_k)\|^2 + \alpha^2 L/2 \|\nabla f(X_k) - \nabla f(x_{k,i})\|^2 \\
& = L^2 n \sum_{i=1}^{n} \|\nabla f(X_k) - \nabla f(x_{k,i})\|^2
\end{align*}
$$

where the last step comes from $2(a, b) = \|a\|^2 + \|b\|^2 = \|a - b\|^2$. And

$$
\begin{align*}
Ef(X_k) - Ef(X_{k+1}) & \leq \alpha - \alpha^2 L \|\nabla f(X_k)\|^2 - \alpha/2 \sum_{i=1}^{n} \|\nabla f(X_k)\|^2 + \alpha^2 L/2 \|\nabla f(X_k) - \nabla f(x_{k,i})\|^2 \\
& \leq \frac{\alpha}{\alpha K} \sum_{k=0}^{K-1} \|\nabla f(X_k)\|^2 + \frac{1}{\alpha K} \sum_{k=0}^{K-1} \|\nabla f(X_k)\|^2 \\
& \leq \frac{2(f(0) - f^*)}{\alpha^2 L} + \frac{L^2}{\alpha n} \sum_{i=1}^{n} \|\nabla f(x_{k,i})\|^2 \\
& \leq \frac{1 - \alpha L}{K} \sum_{k=0}^{K-1} \|\nabla f(X_k)\|^2 + \frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(X_k)\|^2 \\
& \leq \frac{1}{1 - \rho} \sum_{s=1}^{t} \rho^{t-s} b_s \\
& \leq \frac{1}{1 - \rho} \sum_{s=1}^{t} b_s
\end{align*}
$$

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

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

$$
D_k = \sum_{t=1}^{k} a_t^2 \leq \frac{1}{(1 - \rho)^2} \sum_{s=1}^{k} b_s^2
$$
That completes the proof.

Algorithm 2

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

Lemma 10 For any \( X_t \in \mathbb{R}^{d \times n} \), we have
\[ \left\| \sum_{t=0}^{k-1} X_t \left( \frac{I_n \mathbb{1}^\top}{n} - W^{k-t-1} \right) \right\|_F^2 \leq \left( \sum_{t=0}^{k-1} \rho^{k-t-1} \| X_t \|_F \right)^2 \]

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

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 \( \alpha \), the discrepancy bound \( \theta \), 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 \( x_{k,i} \): \( \tilde{g}_{k,i} \)
4: if \( k = 0 \) then
5: Update local weight: \( x_{k + \frac{1}{2}, i} \leftarrow x_{k,i} - \alpha \tilde{g}_{k,i} \)
6: else
7: Update local weight: \( x_{k + \frac{1}{2}, i} \leftarrow 2x_{k,i} - x_{k-1,i} - \alpha \tilde{g}_{k,i} + \alpha \tilde{g}_{k-1,i} \)
8: end if
9: Compute modulo-ed model: \( q_{k,i} \leftarrow \theta \cdot Q_i \left( \frac{x_{k+\frac{1}{2}, i}}{\theta} \mod 1 \right) \) (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: \( \bar{x}_K = \frac{1}{n} \sum_{i=1}^{n} x_{K,i} \)
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**: 
   \[
   \mathbb{E}_{\xi_i \sim D_i} \left\| \nabla f_i(x_i; \xi_i) - \nabla f_i(x) \right\|^2 \leq \sigma^2, \forall i
   \]
   where $\nabla f_i(x_i; \xi_i)$ denotes gradient sample on worker $i$ computed via data sample $\xi_i$.
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 $\|g_{k,i}\|_\infty \leq G_\infty$ for some constant $G_\infty$.

G.3 Proof to Theorem 3

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

\[
\begin{align*}
\bar{x}_{k+\frac{1}{2},i} &= 2x_{k,i} - x_{k-1,i} - \alpha\tilde{g}_{k,i} + \alpha\tilde{g}_{k-1,i} \\
x_{k+1,i} &= \sum_{j=1}^{n} x_{k+\frac{1}{2},j} W_{ji} + \sum_{j=1}^{n} \left( \frac{q_{k,j} - x_{k+\frac{1}{2},j}}{2} \right) W_{ji}
\end{align*}
\]

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

\[
\begin{align*}
X_{k+\frac{1}{2}} &= 2X_k - X_{k-1} - \alpha \tilde{G}_k + \alpha \tilde{G}_{k-1} \\
X_{k+1} &= \sum_{j=1}^{n} X_{k+\frac{1}{2},j} W_{ji} + \sum_{j=1}^{n} \left( \frac{q_{k,j} - X_{k+\frac{1}{2},j}}{2} \right) W_{ji}
\end{align*}
\]

Define

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

Since $W$ is symmetric, it can be diagonalized as $W = P \Lambda P^T$, where the $i$-th column of $P$ and $\Lambda$ are $W$’s $i$-th eigenvector and eigenvalue, respectively. And we obtain

\[
X_{k+1} = 2X_k P \Lambda P^T - X_{k-1} P \Lambda P^T - \alpha \tilde{G}_k P \Lambda P^T + \alpha \tilde{G}_{k-1} P \Lambda P^T + \Omega_k
\]

and

\[
X_{k+1} = 2X_k P \Lambda - X_{k-1} \Lambda - \alpha \tilde{G}_k \Lambda + \alpha \tilde{G}_{k-1} \Lambda + \Omega_k P
\]

Denote $Y_k = X_k P$, $H(X_k; \xi_k) = \tilde{G}_k P$, and denote $y_{k,i}$, $h_{k,i}$ and $r_{k,i}$ as the $i$-th column of $Y_k$, $H_k$ and $\Omega_k P$, 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 7 of Constants $C_1$, $C_2$, $C_3$ and $C_4$ are defined in the Lemma 7, Constants $D_1$ and $D_2$ are defined in Lemma 12, we get

\[
\begin{align*}
&\left( 1 - \frac{3C_1 \alpha^2 L^2}{C_4} \right) \mathbb{E} \left\| \nabla f(0) \right\|^2 + \left( 1 - \alpha L - \frac{3C_2 C_4}{C_3} \right) \mathbb{E} \left\| \tilde{G}_k \right\|^2 + \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 \\
&\leq \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{3C_1 \alpha^2 L^2 (\sigma^2 + \frac{\sigma_0^2}{C_4})}{C_4 K} + \frac{C_2 C_4}{C_3} \alpha^2 \sigma^2 L^2 + \frac{3C_2}{nC_4} \alpha^2 \sigma^2 L^4
\end{align*}
\]
\[ \frac{C_3 L^2}{C_4} \left( \frac{3D_1 n + 4}{3D_2 n} \right)^2 \alpha^2 G_\infty^2 d \]

Let \( \alpha = \frac{1}{\sqrt{K}^{n+2L}} \), we have

\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 \]

\[ \leq \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{3C_1 \alpha^2 L^2 (\sigma^2 + \zeta_0^2)}{C_4 K} + \frac{6C_2}{C_4} \frac{\alpha^2 \sigma^2 L^2}{n} + \frac{3C_2}{nC_4} \alpha^4 \sigma^2 L^4 \]

\[ \quad + \left( \frac{3D_1 n + 4}{3D_2 n} \right)^2 \frac{C_3 L^2}{C_4} G_\infty^2 \alpha^2 \]

\[ \leq \frac{4(f(0) - f^*) L}{K} + \frac{2\sigma (f(0) - f^* + L/2)}{\sqrt{nK}} + \frac{3C_1 L^2 (\sigma^2 + \zeta_0^2) n}{C_4 (\sigma^2 K^2 + 4nL^2 K)} + \frac{6C_2 L^2 \sigma^2 n}{C_4 (\sigma^2 K^2 + 4nL^2 K)} \]

\[ \quad + \frac{3C_2 n \sigma^2 L^2}{C_4 (\sigma^2 K^2 + 16n^2 L^4)} + \left( \frac{3D_1 n + 4}{3D_2 n} \right)^2 \frac{C_3 G_\infty^2 d L^2}{C_4 (\sigma^2 K^2 + 4nL^2 K)} \]

\[ \leq \frac{1}{K} \frac{\sigma}{\sqrt{nK}} \left( \frac{\sigma^2 + \zeta_0^2}{\sigma^2 K^2 + nK} \right) + \frac{\sigma^2}{\sigma^2 K^2 + nK} + \frac{\sigma^2}{\sigma^4 K^2 + n^2} + \frac{G_\infty^2 d n}{\sigma^2 K + n} \]

\[ \leq \frac{1}{K} \frac{\sigma}{\sqrt{nK}} + \frac{\sigma^2}{\sigma^2 K + n} + \frac{G_\infty^2 d n}{\sigma^2 K + n} \]

That completes the proof.

G.4 Lemma for \( D^2 \)

**Lemma 11** Define

\[ D_1 = \max \left\{ \frac{\lambda_n}{1 - |v_n|}, \frac{2 |\lambda_n|}{1 - |v_n|} \right\} \]

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

\[ v_n = \lambda_n - \sqrt{\lambda_n^2 - \lambda_n} \]

Let \( \delta = \frac{1}{6n D_2} \), and we have for \( \forall i, j \)

\[ \left\| x_{k+1} (e_i - e_j) \right\|_\infty \leq \theta = (6D_1 n + 8) \alpha G_\infty \]

**Proof** We use mathematical induction to prove this:

I. When \( k = 0 \),

\[ \left\| X_{0+\frac{1}{2}} (e_i - e_j) \right\|_\infty = \left\| -a \tilde{G}_0 (e_i - e_j) \right\|_\infty \leq \alpha \left\| \tilde{G}_0 \right\|_1 \leq \alpha \left\| e_i - e_j \right\|_1 \leq 2 \alpha G_\infty \leq (6D_1 n + 8) \alpha G_\infty \]

II. Suppose for \( k \geq 0 \), \( \forall t \leq k \), we have \( \left\| X_{t+\frac{1}{2}} (e_i - e_j) \right\|_\infty \leq (6D_1 n + 8) \alpha G_\infty \), then for \( \forall i, j \)

\[ \left\| x_{k+1} (e_i - e_j) \right\|_\infty \leq \left\| X_{k+1} \left( \frac{1}{n} - e_i \right) \right\|_\infty + \left\| X_{k+1} \left( \frac{1}{n} - e_j \right) \right\|_\infty \]

\[ = X_{k+1} P P^\top e_i - X_{k+1} P \]

\[ = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} P^\top e_i \]
We obtain

\[ r \]

For \( s,i \), we have

\[ \| r_{s,i} \|_\infty = \| \Omega_k Pe_i \|_\infty \leq \| \Omega_k \|_1,\infty \| Pe_i \|_1 \leq 2\sqrt{n} \delta \theta \]

when \( \lambda_i < 0 \), we have

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

For \( \beta_{s,i} \), we have

\[ \| \beta_{s,i} \|_\infty = \| -\alpha h_{s,i} + \alpha h_{k-1,i} \|_\infty \leq 2\alpha (\| h_{k,i} \|_\infty + \| h_{k-1,i} \|_\infty ) \]

\[ \leq 2\alpha (\| G_k \|_{1,\infty} \| Pe_i \|_1 + \| G_{k-1} \|_{1,\infty} \| Pe_i \|_1 ) \]

\[ \leq 2\alpha \sqrt{n} G_\infty \]

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

\[ \| y_{k+1,i} \|_\infty \leq \| y_{1,i} \|_\infty \| v_i \|_k + \lambda_i \sum_{s=1}^{k} \| \beta_{s,i} \|_\infty \| v_i \|_{k-s} + \sum_{s=1}^{k} \| r_{s,i} \|_\infty \| v_i \|_{k-s} \]
As a result, we get

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

\[
+ \sum_{s=1}^{k} r_{s, i} \lambda_i^{\frac{k-s}{2}} \sin((k + s - 1) \theta_i)
\]

By taking norm, we get

\[
\|y_{k+1, i}\|_{\infty} \sin \theta_i = \|y_{1, i}\|_{\infty} \lambda_i^{\frac{k}{2}} \sin((t + 1) \theta_i) + \lambda_i \sum_{s=1}^{k} \|\beta_{s, i}\|_{\infty} \lambda_i^{\frac{k-s}{2}} \|\sin((k + s - 1) \theta_i)\|
\]

\[
+ \sum_{s=1}^{k} \|r_{s, i}\|_{\infty} \lambda_i^{\frac{k-s}{2}} \|\sin((k + s - 1) \theta_i)\|
\]

\[
\leq \|y_{1, i}\|_{\infty} \lambda_i^{\frac{k}{2}} + 2\alpha \sqrt{nG_{\infty}} \lambda_2 \sum_{s=1}^{\infty} \lambda_i^{\frac{s}{2}} + 2\sqrt{n\delta \theta} \sum_{s=1}^{\infty} \lambda_i^{\frac{s}{2}}
\]

\[
\leq \alpha \sqrt{nG_{\infty}} \sqrt{\lambda_2} + 2\alpha \sqrt{nG_{\infty}} \lambda_2 + 2\sqrt{n\delta \theta}
\]

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

\[
\|y_{k+1, i}\|_{\infty} \leq \alpha \sqrt{nG_{\infty}} \sqrt{\frac{\lambda_2}{1 - \lambda_2}} + 2\alpha \sqrt{nG_{\infty}} \lambda_2 + 2\sqrt{n\delta \theta}
\]

So there exists \(D_1, D_2\)

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

\[
D_2 = \max \left\{ \frac{2}{1 - |\lambda_n|}, \frac{2}{\sqrt{1 - \lambda_2}} \right\}
\]

such that

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

Putting it back we have \(\forall i, j\)

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

As a result

\[
\left\| X_{k+1}(e_i - e_j) \right\|_{\infty}
\]

\[
= \left\| (2X_k - X_{k-1} - \alpha \tilde{G}_k + \alpha \tilde{G}_{k-1})(e_i - e_j) \right\|_{\infty}
\]

\[
\leq 2 \|X_k(e_i - e_j)\|_{\infty} + \|X_{k-1}(e_i - e_j)\|_{\infty} + \alpha \left\| \tilde{G}_k \right\|_{1, \infty} \|e_i - e_j\|_1 + \alpha \left\| \tilde{G}_{k-1} \right\|_{1, \infty} \|e_i - e_j\|_1
\]

32
We obtain $\delta$ since Lemma 12.

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

Combining I and II we complete the proof.

**Lemma 12**

$$(1 - 12C_2 \alpha^2 L^2) \sum_{i=1}^{n} \sum_{k=1}^{K} E \left\| X_k - x_{k,i} \right\|^2$$

$$\leq 3C_1 \alpha^2 n \sigma^2 + 3C_1 \alpha^2 n \sigma^2_0 + 3C_1 \alpha^2 nE \left\| \nabla f(0) \right\| + 6C_2 \alpha^2 n \sigma^2 K + 3C_2 \alpha^4 \sigma^2 L^2 K$$

$$+ 3C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} E \left\| C_k \right\|^2 + C_3 \sum_{k=1}^{K-1} E \left\| \Omega_k \right\|^2_F$$

**Proof**

$$\sum_{i=1}^{n} \left\| X_k - x_{k,i} \right\|^2 = \sum_{i=1}^{n} \left\| X_k \left( e_i - \frac{1}{n} \mathbf{1} \right) \right\|^2$$

$$= \left\| X_k \left( I - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right) \right\|^2_F$$

$$= \left\| X_k P P^T - X_k v_1 v_1^T \right\|^2_F$$

$$\leq \left\| X_k P \right\| \left\| \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \right\|_F$$

$$= \sum_{i=2}^{n} \left\| y_{k,i} \right\|^2$$

From the update rule, we obtain,

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

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

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

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

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

Since

$$= \left\| \begin{bmatrix} u_i^{n+1} - v_i^{n+1} \\ u_i - v_i \end{bmatrix} \right\| \leq \left\| v_i \right\|^n \left\| \begin{bmatrix} u_i \\ v_i \end{bmatrix} \right\| - \left\| v_i \right\|^n \leq \left\| v_i \right\|^n$$

We obtain

$$\left\| y_{k+1,i} \right\|^2 \leq 3 \left\| y_{i,i} \right\|^2 \left\| v_i \right\|^2 + 3 \lambda_i^2 \left( \sum_{s=1}^{k} \left\| \beta_{s,i} \left\| v_i \right|^{k-s} \right\) + 3 \left( \sum_{s=1}^{k} \left\| r_{s,i} \left| v_i \right|^{k-s} \right\)$$
Summing over from \( k = 0 \) to \( t = K - 1 \), we obtain

\[
\sum_{k=0}^{K-1} \| y_{k+1,i} \|^2 = \sum_{k=1}^{K} \| y_{k,i} \|^2
\]

\[
\leq 3 \| y_{1,i} \|^2 \sum_{k=0}^{K-1} |v_i|^{2k} + 3\lambda_i^2 \sum_{k=1}^{K-1} \left( \sum_{s=1}^{k} \| \beta_{s,i} \| |v_i|^{k-s} \right)^2 + 3 \sum_{s=1}^{K-1} \left( \sum_{m=s}^{K-1} \| r_{m,i} \| |v_i|^{k-m} \right)^2
\]

\[
\leq 3 \| y_{1,i} \|^2 + \frac{3\lambda_i^2}{1 - |v_i|^2} \sum_{k=1}^{K-1} \| \beta_{k,i} \|^2 + \frac{3}{1 - |v_i|^2} \sum_{k=1}^{K-1} \| r_{k,i} \|^2
\]

where \( v_n = \lambda_n - \sqrt{\lambda_n^2 - \lambda_i} \).

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

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

\[+ \sum_{s=1}^{k} r_{s,i} \lambda_i^{\frac{k-s}{2}} \sin[(k + 1 - s)\theta_i] \]

And we have

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

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

Summing from \( k = 0 \) to \( K - 1 \), we have

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

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

\[
\leq 3 \| y_{1,i} \|^2 + \frac{3\lambda_i^2}{1 - \lambda_i} \sum_{k=1}^{K-1} \| \beta_{k,i} \|^2 + \frac{3}{1 - \lambda_i} \sum_{k=1}^{K-1} \| r_{k,i} \|^2
\]

Since \( \sin^2 \theta_i = 1 - \lambda_i \), we have

\[
\sum_{k=1}^{K} \| y_{k,i} \|^2 \leq \frac{3 \| y_{1,i} \|^2}{(1 - \lambda_i)^2} + \frac{3\lambda_i^2}{1 - \lambda_i} \sum_{k=1}^{K-1} \| \beta_{k,i} \|^2 + \frac{3}{1 - \lambda_i} \sum_{k=1}^{K-1} \| r_{k,i} \|^2
\]

\[\leq \frac{3 \| y_{1,i} \|^2}{(1 - \lambda_2)^2} + \frac{3\lambda_i^2}{1 - \lambda_2} \sum_{k=1}^{K-1} \| \beta_{k,i} \|^2 + \frac{3}{1 - \lambda_2} \sum_{k=1}^{K-1} \| r_{k,i} \|^2
\]
So there exists $C_1, C_2, C_3$

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

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

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

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

By taking expectation we have

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

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

$$
\sum_{i=2}^{n} E \|\beta_{k,i}\|^2 = \alpha^2 \sum_{i=2}^{n} E \|h_{k,i} - h_{k-1,i}\|^2
$$

$$
= \alpha^2 \sum_{i=1}^{n} E \left\| G_k P e_i - G_{k-1} P e_i \right\|^2
$$

$$
\leq \alpha^2 \sum_{i=1}^{n} E \left\| G_k P e_i - \tilde{G}_{k-1} P e_i \right\|^2
$$

$$
\leq \alpha^2 \sum_{i=1}^{n} \left\| G_k P - \tilde{G}_{k-1} P \right\|_F^2
$$

**Lemma 14**

$$
\leq \alpha^2 \sum_{i=1}^{n} \left\| G_k e_i - \tilde{G}_{k-1} e_i \right\|^2
$$

$$
\leq 3\alpha^2 \sum_{i=1}^{n} \left\| G_k e_i - G_{k-1} e_i \right\|^2 + 3\alpha^2 \sum_{i=1}^{n} \left\| G_{k-1} e_i - G_{k-1} e_i \right\|^2
$$

$$
+ 3\alpha^2 \sum_{i=1}^{n} \left\| G_k e_i - G_{k-1} e_i \right\|^2
$$

$$
\leq 6\alpha^2 n \sigma^2 + 3\alpha^2 \sum_{i=1}^{n} \left\| G_k e_i - G_{k-1} e_i \right\|^2
$$

$$
\leq 6\alpha^2 n \sigma^2 + 3\alpha^2 L^2 \sum_{i=1}^{n} \left\| x_{k,i} - x_{k-1,i} \right\|^2
$$

$$
\leq 6\alpha^2 n \sigma^2 + 3\alpha^2 L^2 \sum_{i=1}^{n} \left\| Y_k P^T e_i - Y_{k-1} P^T e_i \right\|^2
$$

$$
\leq 6\alpha^2 n \sigma^2 + 3\alpha^2 L^2 \left\| Y_k P^T - Y_{k-1} P^T \right\|_F^2
$$

**Lemma 14**

$$
\leq 6\alpha^2 n \sigma^2 + 3\alpha^2 L^2 \left\| Y_k - Y_{k-1} \right\|_F^2
$$
≤ 6α^2 nσ^2 + 3α^2 L^2 \sum_{i=1}^{n} E \| y_{k,i} - y_{k-1,i} \|^2

And Putting it back, we have

\[ \sum_{i=2}^{n} \sum_{k=1}^{K} E \| y_{k,i} \|^2 \leq C_1 E \| Y_1 \|^2_F + C_2 \sum_{i=2}^{n} \sum_{k=1}^{K-1} E \| \beta_{k,i} \|^2 + C_3 \sum_{k=1}^{K-1} \sum_{i=2}^{n} E \| r_{k,i} \|^2 \]

\[ \leq C_1 E \| Y_1 \|^2_F + C_2 \sum_{k=1}^{K-1} \left( 6α^2 nσ^2 + 3α^2 L^2 \sum_{i=1}^{n} E \| y_{k,i} - y_{k-1,i} \|^2 \right) + C_3 \sum_{k=1}^{K-1} \sum_{i=2}^{n} E \| r_{k,i} \|^2 \]

Lemma \[ \leq C_1 E \| Y_1 \|^2_F + 6C_2 α^2 nσ^2 K + 3C_2 α^4 L^2 \sum_{i=1}^{K-1} \sum_{k=1}^{n} E \| y_{k,i} - y_{k-1,i} \|^2 + C_3 \sum_{k=1}^{K-1} E \| \Omega_k \|^2_F \]

Since

\[ E \| y_{k,1} - y_{k-1,1} \|^2 = E \| X_k Pe_1 - X_{k-1} Pe_1 \|^2 = E \| X_k v_1 - X_{k-1} v_1 \|^2 \]

\[ = E \left\| X_k \frac{1}{\sqrt{n}} I_n - X_{k-1} \frac{1}{\sqrt{n}} I_n \right\|^2 = nE \left\| X_k - X_{k-1} \right\|^2 = nα^2 E \| G_k \|^2 \]

\[ \leq nα^2 E \left\| \overline{C}_k - \overline{C}_k \right\|^2 + nα^2 E \| \overline{C}_k \|^2 \leq nα^2 \frac{σ^2}{n} + nα^2 E \| \overline{C}_k \|^2 \]

\[ = α^2 σ^2 + nα^2 E \| \overline{C}_k \|^2 \]

Putting it back, and we obtain

\[ \sum_{i=2}^{n} \sum_{k=1}^{K} E \| y_{k,i} \|^2 \leq C_1 E \| Y_1 \|^2_F + 6C_2 α^2 nσ^2 K + 3C_2 α^4 L^2 K + 3C_2 α^4 nL^2 \sum_{k=1}^{K-1} E \| \overline{G}_k \|^2 \]

\[ + 3C_2 α^2 L^2 \sum_{k=1}^{K-1} \sum_{i=2}^{n} E \| y_{k,i} - y_{k-1,i} \|^2 + C_3 \sum_{k=1}^{K-1} E \| \Omega_k \|^2_F \]

\[ \leq C_1 E \| Y_1 \|^2_F + 6C_2 α^2 nσ^2 K + 3C_2 α^4 L^2 K + 3C_2 α^4 nL^2 \sum_{k=1}^{K-1} E \| \overline{C}_k \|^2 \]

\[ + 6C_2 α^2 L^2 \sum_{k=1}^{K-1} \sum_{i=2}^{n} E \left( \| y_{k,i} \|^2 + \| y_{k-1,i} \|^2 \right) + C_3 \sum_{k=1}^{K-1} E \| \Omega_k \|^2_F \]

\[ \leq C_1 E \| Y_1 \|^2_F + 6C_2 α^2 nσ^2 K + 3C_2 α^4 L^2 K + 3C_2 α^4 nL^2 \sum_{k=1}^{K-1} E \| \overline{C}_k \|^2 \]

\[ + 12C_2 α^2 L^2 \sum_{k=1}^{K-1} \sum_{i=2}^{n} E \| y_{k,i} \|^2 + C_3 \sum_{k=1}^{K-1} E \| \Omega_k \|^2_F \]

Rearrange the terms, we get

\[ (1 - 12C_2 α^2 L^2) \sum_{i=2}^{n} \sum_{k=1}^{K} E \| y_{k,i} \|^2 \]
\[
\begin{align*}
&\leq C_1 \mathbb{E} \|Y_1\|^2_F + 6C_2 \alpha^2 n \sigma^2 K + 3C_2 \alpha^4 \sigma^2 L^2 K + 3C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E} \|G_k\|^2 + C_3 \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F \\
&\leq C_1 \mathbb{E} \|X_1\|^2_F + 6C_2 \alpha^2 n \sigma^2 K + 3C_2 \alpha^4 \sigma^2 L^2 K + 3C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E} \|G_k\|^2 + C_3 \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\end{align*}
\]

Considering
\[
\mathbb{E} \|X_1\|^2_F = \alpha^2 \mathbb{E} \left\| \frac{1}{\alpha} G_0 \right\|^2_F
\]
\[
= \alpha^2 \sum_{i=1}^{n} \mathbb{E} \left\| \frac{1}{\alpha} (G_{0,i} - G_{0,i}) + \nabla f(0) \right\|^2
\]
\[
\leq 3\alpha^2 \sum_{i=1}^{n} \mathbb{E} \left\| \frac{1}{\alpha} (G_{0,i} - G_{0,i}) \right\|^2 + 3\alpha^2 \sum_{i=1}^{n} \mathbb{E} \|G_{0,i} - \nabla f(0)\|^2 + 3\alpha^2 \sum_{i=1}^{n} \mathbb{E} \|\nabla f(0)\|^2
\]
\[
\leq 3\alpha^2 n \sigma^2 + 3\alpha^2 n \sigma_0^2 + 3\alpha^2 n \mathbb{E} \|\nabla f(0)\|
\]

We finally get
\[
(1 - 12C_2 \alpha^2 L^2) \sum_{i=2}^{n} \sum_{k=1}^{K} \mathbb{E} \|y_{k,i}\|^2
\]
\[
= (1 - 12C_2 \alpha^2 L^2) \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E} \|X_k - x_{k,i}\|^2
\]
\[
\leq 3C_1 \alpha^2 n \sigma^2 + 3C_1 \alpha^2 n \sigma_0^2 + 3C_1 \alpha^2 n \mathbb{E} \|\nabla f(0)\| + 6C_2 \alpha^2 n \sigma^2 K + 3C_2 \alpha^4 \sigma^2 L^2 K
\]
\[
+ 3C_2 \alpha^4 n L^2 \sum_{k=1}^{K-1} \mathbb{E} \|G_k\|^2 + C_3 \sum_{k=1}^{K-1} \mathbb{E} \|\Omega_k\|^2_F
\]

That completes the proof.

**Lemma 13** Given \( \rho \in (-\frac{1}{2}, 0) \cup (0, 1) \), for any two sequence \( \{a_t\}_{t=1}^{\infty}, \{b_t\}_{t=1}^{\infty} \) and \( \{c_t\}_{t=1}^{\infty} \) that satisfy
\[
\begin{align*}
& a_0 = b_0 = 0, \\
& a_{t+1} = \rho (2a_t - a_{t-1}) + b_t - b_{t-1} + c_t, \forall t \geq 1
\end{align*}
\]
we have
\[
\begin{align*}
& a_{t+1} = a_1 \left( \frac{u^{t+1} - v^{t+1}}{u - v} \right) + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) \left( \frac{u^{t-s+1} - v^{t-s+1}}{u - v} \right), \forall t \geq 0
\end{align*}
\]
where
\[
\begin{align*}
& u = \rho + \sqrt{\rho^2 - \rho}, \quad v = \rho - \sqrt{\rho^2 - \rho}
\end{align*}
\]
Moreover, if \( 0 < \rho < 1 \), we have
\[
\begin{align*}
& a_{t+1} = a_1 \rho^{\frac{1}{2}} \frac{\sin((t + 1)\theta)}{\sin \theta} + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) \rho^{\frac{1-t}{2}} \frac{\sin((t - s + 1)\theta)}{\sin \theta}
\end{align*}
\]
where
\[
\begin{align*}
& \theta = \arccos(\sqrt{\rho})
\end{align*}
\]
**Proof** when \( t \geq 1 \), we have
\[
\begin{align*}
& a_{t+1} = 2 \rho a_t - \rho a_{t-1} + b_t - b_{t-1} + c_t
\end{align*}
\]
since,
\[
\begin{align*}
& u = \rho + \sqrt{\rho^2 - \rho}, \quad v = \rho - \sqrt{\rho^2 - \rho}
\end{align*}
\]
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
\[
a_{t+1} - u a_t = (a_t - u a_{t-1}) v + b_t - b_{t-1} + c_t \\
= (a_{t-1} - u a_{t-2}) v^2 + (b_{t-1} - b_{t-2} + c_{t-1}) v + b_t - b_{t-1} + c_t \\
= (a_1 - u a_0) v^t + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{t-s} \\
= a_1 v^t + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{t-s}
\]

Dividing both sides by \( u^{t+1} \), we have
\[
\frac{a_{t+1}}{u^{t+1}} = \frac{a_t}{u^t} + u^{-(t+1)} \left( a_1 v^t + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{t-s} \right) \\
= \frac{a_{t-1}}{u^{t-1}} + u^{-t} \left( a_1 v^{t-1} + \sum_{s=1}^{t-1} (b_s - b_{s-1} + c_s) v^{t-1-s} \right) \\
+ u^{-(t+1)} \left( a_1 v^t + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{t-s} \right) \\
= \frac{a_1}{u} + \sum_{k=1}^{t} u^{-k-1} \left( a_1 v^k + \sum_{s=1}^{k} (b_s - b_{s-1} + c_s) v^{k-s} \right)
\]

Multiplying both sides by \( u^{t+1} \)
\[
a_{t+1} = a_1 u^t + \sum_{k=1}^{t} u^{-k} \left( a_1 v^k + \sum_{s=1}^{k} (b_s - b_{s-1} + c_s) v^{k-s} \right) \\
= a_1 u^t \left( 1 + \sum_{k=1}^{t} \left( \frac{v}{u} \right)^k \right) + u^t \sum_{k=1}^{t} \sum_{s=1}^{k} (b_s - b_{s-1} + c_s) v^{k-s} \left( \frac{v}{u} \right)^k \\
= a_1 u^t \sum_{k=0}^{t} \left( \frac{v}{u} \right)^k + u^t \sum_{s=1}^{t} \sum_{k=s}^{t} (b_s - b_{s-1} + c_s) v^{k-s} \left( \frac{v}{u} \right)^k \\
= a_1 u^t \left( \frac{1 - (\frac{v}{u})^{t+1}}{1 - \frac{v}{u}} \right) + u^t \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) v^{k-s} \left( \frac{v}{u} \right)^k \cdot \frac{1 - (\frac{v}{u})^{t-s-1}}{1 - \frac{v}{u}} \\
= a_1 \left( \frac{u^{t+1} - v^{t+1}}{u - v} \right) + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) \frac{u^{t-s+1} - v^{t-s+1}}{u - v}
\]

Note that when \( 0 < \rho < 1 \), both \( u \) and \( v \) are complex numbers, we have
\[
u = \sqrt{\rho} e^{i\theta}, \quad v = \sqrt{\rho} e^{-i\theta}
\]

where \( \theta = \arccos \sqrt{\rho} \). And under this context, we have
\[
a_{t+1} = a_1 \rho^t \frac{\sin((t+1)\theta)}{\sin \theta} + \sum_{s=1}^{t} (b_s - b_{s-1} + c_s) \rho^{t-s} \frac{\sin((t-s+1)\theta)}{\sin \theta}
\]

That completes the proof.

**Lemma 14** For any matrix \( X \in \mathbb{R}^{N \times n} \), we have
\[
\sum_{i=1}^{n} \|X v_i\|^2 \leq \sum_{i=1}^{n} \|X v_i\|^2 = \|X\|_F^2 \\
\sum_{i=1}^{n} \|X P^T e_i\|^2 = \|X P^T\|_F^2 = \|X\|_F^2
\]
As a result, we can reuse Lemma 8 from D-PSGD, thus we have
\[ \sum_{i=1}^{n} \|X_i v_i\|^2 \leq \sum_{i=1}^{n} \|X_i P\|^2 = Tr(X_i P P^T X_i^T) = Tr(X_i X_i^T) = \|X_i\|^2_F. \]
And similarly,
\[ \sum_{i=1}^{n} \|X P^T e_i\|^2 = \|X P^T\|^2_F = Tr(X_i P^T P X_i^T) = Tr(X_i X_i^T) = \|X_i\|^2_F. \]
That completes the proof.

**Lemma 15** If we run Algorithm 2 for K iterations the following inequality holds:
\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X_k)\|^2 \\
\leq \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{3C_1 \alpha^2 L^2 (\sigma^2 + \mathbb{E})}{C_4 K} + \frac{C_2}{C_4} \alpha^2 \sigma^2 L^2 + \frac{C_2}{nC_4} \alpha^4 \sigma^2 L^4 \\
+ \frac{C_4}{C_4} \left( \frac{3D_1 n + 4}{3D_2 n} \right) \alpha^2 G_{\infty} d
\]
where
\[
C_1 = \max \left\{ \frac{3}{1 - |v_n|^2}, \frac{3}{(1 - \lambda_2)^2} \right\} \\
C_2 = \max \left\{ \frac{3\lambda^2}{(1 - |v_n|^2)}, \frac{3\lambda^2}{(1 - \sqrt{\lambda_2})^2 (1 - \lambda_2)} \right\} \\
C_3 = \max \left\{ \frac{3}{(1 - |v_n|^2)}, \frac{3}{(1 - \sqrt{\lambda_2})^2 (1 - \lambda_2)} \right\} \\
C_4 = 1 - 12C_2 \alpha^2 L^2
\]
\[
Q_k e_i = \sum_{j=1}^{n} \left( q_{k,j} - x_{k+\frac{1}{2}, j} \right) (q_{k,i} - x_{k+\frac{1}{2}, i}) W_{ji}
\]
**Proof** Since
\[
\bar{X}_{k+1} = (2X_k - X_{k-1} - \alpha \bar{G}_k + \alpha \bar{G}_{k-1}) W \frac{\mathbb{I}_n}{n} + (Q_k - X_{k+\frac{1}{2}})(W - I) \frac{\mathbb{I}_n}{n}
\]
\[
= 2X_k - \bar{X}_{k-1} - \alpha \bar{G}_k + \alpha \bar{G}_{k-1}
\]
and we have
\[
\bar{X}_{k+1} - \bar{X}_k = \bar{X}_k - \bar{X}_{k-1} - \alpha \bar{G}_k + \alpha \bar{G}_{k-1}
\]
\[
= \bar{X}_1 - \bar{X}_0 - \alpha \sum_{i=1}^{k} (G_i - \bar{G}_{i-1})
\]
\[
= -\alpha \bar{G}_k
\]
As a result, we can reuse Lemma 8 from D-PSGD, thus we have
\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\bar{G}_k\|^2 + \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \|\nabla f(X_k)\|^2 \\
\leq \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{L^2}{nK} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \|\bar{X}_k - x_{k,i}\|^2
\]
From Lemma 12 we obtain

\[
\frac{1 - \alpha L}{K} \sum_{k=0}^{K-1} E \| G_k \|^2 + \frac{1}{K} \sum_{k=0}^{K-1} E \| \nabla f(X_k) \|^2 \\
\leq \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{3C_1 \alpha^2 L^2 (\sigma^2 + \varsigma^2_0 + E \| \nabla f(0) \|)}{C_4 K} + 6 \frac{C_2}{C_4} \alpha^2 \sigma^2 L^2 + 3 \frac{C_2}{n C_4} \alpha^4 \sigma^2 L^4 \\
+ \frac{3 C_2}{C_4} \alpha^4 L^4 \frac{1}{K} \sum_{k=1}^{K-1} E \| G_k \|^2 + \frac{C_3 L^2}{C_4 n K} \sum_{k=1}^{K-1} E \| \Omega_k \|^2_F
\]

Rearrange the terms, we get

\[
\left( 1 - \frac{3 C_1 \alpha^2 L^2}{C_4} \right) E \| \nabla f(0) \| + \left( 1 - \alpha L - 3 \frac{C_2}{C_4} \alpha^4 L^4 \right) \frac{1}{K} \sum_{k=1}^{K-1} E \| G_k \|^2 \\
+ \frac{1}{K} \sum_{k=0}^{K-1} E \| \nabla f(X_k) \|^2 \\
\leq \frac{2(f(0) - f^*)}{\alpha K} + \frac{\alpha L}{n} \sigma^2 + \frac{3C_1 \alpha^2 L^2 (\sigma^2 + \varsigma^2_0)}{C_4 K} + \frac{C_3 L^2}{C_4 n K} \sum_{k=1}^{K-1} E \| \Omega_k \|^2_F \\
+ \frac{C_3 L^2}{C_4} \left( \frac{3 D_1 n + 4}{3 D_2 n} \right)^2 \alpha^2 G_\infty^2 \delta^2 d
\]

That completes the proof.

Lemma 16

\[
\sum_{k=0}^{K-1} E \| \Omega_k \|^2_F \leq \left( \frac{3 D_1 n + 4}{3 D_2 n} \right)^2 \alpha^2 G_\infty^2 \delta^2 d n K
\]

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

\[
\sum_{k=0}^{K-1} E \| \Omega_k \|^2_F = \sum_{k=0}^{K-1} \sum_{i=0}^{n} E \left[ \sum_{j=1}^{n} \left( (q_{k,j} - x_{k,j+\frac{1}{2},j}) - (q_{k,i} - x_{k,i+\frac{1}{2},i}) \right) W_{ji} \right]^2 \\
\leq \frac{4}{3} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \delta^2 \theta^2 \delta d \leq \left( \frac{3 D_1 n + 4}{3 D_2 n} \right)^2 \alpha^2 G_\infty^2 \delta^2 d n K
\]

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 $\alpha$, the discrepancy bound $\theta$, number of iterations $K$, quantization function $Q$, 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 $\tau_k$: $\tilde{g}_k,i_k \leftarrow x_{k,i_k} + \frac{1}{2}g_{k,i_k} - \frac{1}{2}g_{k,i_k}$
4: Compute modulo-ed model: $q_{k,i_k} \leftarrow \theta \cdot Q_{\delta}(x_{k,i_k})$ (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: $x_{k+1,i_k} \leftarrow x_{k,i_k} + \alpha \tilde{g}_{k,i_k}$
6: Update the local weight with local gradient: $x_{k+1,i_k} \leftarrow x_{k,i_k} - \alpha \tilde{g}_{k,i_k}$
7: end for

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

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

$$\nabla F(X_k) = [p_1 g_{k,1}, \cdots, p_n g_{k,n}] \in \mathbb{R}^{d \times n}$$

$$\nabla \tilde{F}(X_k) = [p_1 \tilde{g}_{k,1}, \cdots, p_n \tilde{g}_{k,n}] \in \mathbb{R}^{d \times n}$$

$$\tilde{G}_k = [\cdots, \tilde{g}_{k,i_k}, \cdots]$$

$$G_k = [\cdots, g_{k,i_k}, \cdots]$$

$$\Lambda_{a}^{b} = \frac{1}{n} n^\top - \prod_{q=a}^{b} W_q$$

H.3 ASSUMPTION

We makes the following assumptions:

1. Lipschitzian Gradient: All the function $f_i$ have $L$-Lipschitzian gradients.

2. Communication Matrix\footnote{Please refer to Section E for more details} 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

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

3. Bounded Variance:

$$\mathbb{E}_{\xi_i \sim D_i} \left[ \left\| \nabla \tilde{f}_i(x_i; \xi_i) - \nabla f_i(x) \right\|_2^2 \right] \leq \sigma^2, \forall i$$
\[ \mathbb{E}_{i \sim \{1, \ldots, n\}} \| \nabla f_i(x) - \nabla f(x) \|^2 \leq \epsilon^2, \forall i \]

where \( \nabla f_i(x; \xi_i) \) denotes gradient sample on worker \( i \) computed via data sample \( \xi_i \).

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 \( \|g_{k,i}\|_\infty \leq G_\infty \) for some constant \( G_\infty \).

### H.4 Proof to Theorem 4

**Proof** We start from

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \| \nabla f(X_k) \|^2 + \left( 1 - \frac{2\alpha L}{n} \right) \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \| \nabla \bar{f}(X_{k-T}) \|^2
\]

**Lemma 17**

\[
\frac{2}{\alpha K} \frac{1}{n} \sum_{k=0}^{K-1} \mathbb{E} \| \nabla f(X_k) \|^2 \leq \frac{2n(f(0) - f^*)}{\alpha K} + \frac{\sigma^2 + 6\xi^2}{n} \mathbb{E} \| X_{k-T} \| \left( \frac{1}{n} - \epsilon_i \right)^2
\]

Rearrange the terms, we get

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \| \nabla f(X_k) \|^2 \leq \frac{2n(f(0) - f^*)}{\alpha K} + \frac{\sigma^2 + 6\xi^2}{n} \mathbb{E} \| X_{k-T} \| \left( \frac{1}{n} - \epsilon_i \right)^2 + \frac{128\alpha^2 \tau_{mix}^2 L^2}{A_1} \left( \sigma^2 + 6\xi^2 \right) + \frac{2pK}{n} \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E} \left\| g_{k-T,k,i} \right\|^2 + G_\infty^2 d
\]

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

\[
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \| \nabla f(X_k) \|^2 \leq \frac{1}{K} + \frac{\sqrt{\sigma^2 + 6\xi^2}}{\sqrt{K}} + \frac{p}{n} \sum_{k=0}^{K-1} \frac{\sigma^2 + 6\xi^2}{\sqrt{K}} \left( \frac{\sigma^2 + 6\xi^2}{\sigma^2 + 6\xi^2} \right) + \frac{n^2 \tau_{mix}^2 G_\infty^2 d}{A_1}
\]

**H.5 Lemma for Moniqua on AD-PSGD**

**Lemma 17**

\[
\mathbb{E} \left\| g_{k-T,k} \frac{1}{n} \right\|^2 \leq \frac{\sigma^2}{n^2} + \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E} \| g_{k-T,k,i} \|^2, \forall k \geq 0.
\]
Lemma 18

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

Proof
\[ \sum_{i=1}^{n} p_i E \left\| g_{k-\tau_k,i} \right\|^2 = \sum_{i=1}^{n} p_i E \left\| g_{k-\tau_k,i} - \sum_{i=1}^{n} p_i g_{k-\tau_k,i} + \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \]
\[ \leq 2 \sum_{i=1}^{n} p_i E \left\| g_{k-\tau_k,i} - \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 + 2 \sum_{i=1}^{n} p_i E \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \]
\[ \leq 2 \sum_{i=1}^{n} p_i E \left\| g_{k-\tau_k,i} - \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 + 2E \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \]

And
\[ \sum_{i=1}^{n} p_i E \left\| g_{k-\tau_k,i} - \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \]
\[ \leq 3 \sum_{i=1}^{n} p_i E \left\| g_{k-\tau_k,i} - \nabla f_i(X_{k-\tau_k}) \right\|^2 + 3 \sum_{i=1}^{n} p_i E \left\| \nabla f_i(X_{k-\tau_k}) - \sum_{j=1}^{n} p_j \nabla f_j(X_{k-\tau_k}) \right\|^2 \]
\[ + 3 \sum_{i=1}^{n} p_i E \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} - \sum_{j=1}^{n} p_j \nabla f_j(X_{k-\tau_k}) \right\|^2 \]
\[ \leq 3L^2 \sum_{i=1}^{n} p_i E \left\| x_{k-\tau_k,i} - X_{k-\tau_k} \right\|^2 + 3 \sum_{i=1}^{n} p_i E \left\| \nabla f_i(X_{k-\tau_k}) - \sum_{j=1}^{n} p_j \nabla f_j(X_{k-\tau_k}) \right\|^2 \]
\[ + 3E \sum_{i=1}^{n} p_i E \left\| p_i g_{k-\tau_k,i} - \sum_{j=1}^{n} p_j \nabla f_j(X_{k-\tau_k}) \right\|^2 \]
\[ \leq 3L^2 \sum_{i=1}^{n} p_i E \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 + 3 \sum_{i=1}^{n} p_i E \left\| \nabla f_i(X_{k-\tau_k}) - \nabla f(X_{k-\tau_k}) \right\|^2 \]
\[ + 3 \sum_{j=1}^{n} p_j E \left\| g_{k-\tau_k,j} - \nabla f_j(X_{k-\tau_k}) \right\|^2 \]
\[ \leq 6L^2 \sum_{i=1}^{n} p_i E \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 + 3\varsigma^2 \]

That completes the proof.
Lemma 19 Let $A_1 = 1 - 192\rho^2 \tau_{\text{mix}}^2 L^2$,
\[
\sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i |X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) |^2 \leq \frac{32\rho^2 \tau_{\text{mix}}^2}{A_1} \left( (\sigma^2 + 6\rho^2)E + 2p \sum_{k=0}^{K-1} E \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k, i} \right\|^2 + G^2 \sqrt{dK} \right)
\]

Proof
\[
\sum_{i=1}^{n} p_i |X_k \left( \frac{1}{n} - e_i \right) |^2 = \sum_{i=1}^{n} p_i \left\| X_{k-1} W_{k-1} - \alpha \tilde{G}_{k-1-\tau_{k-1}} + \Omega_{k-1} \right\| \left( \frac{1}{n} - e_i \right) |^2
\]
\[
X_0 \leq \sum_{i=1}^{n} p_i \left\| \sum_{t=0}^{k-1} \left( -\alpha \tilde{G}_{t-\tau_t} + \Omega_t \right) \Lambda_{t+1}^{k-1} e_i \right\|^2 
\]
\[
\leq 2 \sum_{i=1}^{n} p_i \left\| \sum_{t=0}^{k-1} \alpha \tilde{G}_{t-\tau_t} \Lambda_{t+1}^{k-1} e_i \right\|^2 + 2 \sum_{i=1}^{n} p_i \left\| \sum_{t=0}^{k-1} \Omega_t \Lambda_{t+1}^{k-1} e_i \right\|^2 
\]

Now for the first term, we have
\[
2 \sum_{i=1}^{n} p_i \left\| \sum_{t=0}^{k-1} \alpha \tilde{G}_{t-\tau_t} \Lambda_{t+1}^{k-1} e_i \right\|^2 \leq 2\rho^2 E \left\| \sum_{t=0}^{k-1} \tilde{G}_{t-\tau_t} \Lambda_{t+1}^{k-1} \right\|_F^2 
\]
\[
\leq 2\rho^2 E \left( \sum_{t=0}^{k-1} \left\| \tilde{G}_{t-\tau_t} \right\|_F \left\| \Lambda_{t+1}^{k-1} \right\|_F \right)^2 
\]
\[
\leq 2\rho^2 E \left( \sum_{t=0}^{k-1} \left\| \tilde{G}_{t-\tau_t} \right\|_F \left\| \Lambda_{t+1}^{k-1} \right\|_F \right)^2 
\]
\[
\leq 8\rho^2 E \left( \sum_{t=0}^{k-1} \left\| \tilde{G}_{t-\tau_t} \right\|_F 2^{-\left( t - \tau_t - 1 \right)_{\text{min}}} \right)^2 
\]

Now we replace $k$ with $k - \tau_k$, that is
\[
\sum_{i=1}^{n} p_i |X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) |^2 \leq 8\rho^2 E \left( \sum_{t=0}^{k-\tau_k-1} \left\| \tilde{G}_{t-\tau_t} \right\|_F 2^{-\left( t - \tau_t - 1 \right)_{\text{min}}} \right)^2 
\]

Summing from $k = 0$ to $K - 1$ on both sides, we obtain
\[
\sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i |X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) |^2 \leq 8\rho^2 \sum_{k=0}^{K-1} \left( \sum_{t=0}^{k-\tau_k-1} \left\| \tilde{G}_{t-\tau_t} \right\|_F 2^{-\left( t - \tau_t - 1 \right)_{\text{min}}} \right)^2 
\]
\[
+ 2 \sum_{i=1}^{n} p_i \sum_{k=0}^{K-1} \left\| \sum_{t=0}^{k-\tau_k-1} \Omega_t \Lambda_{t+1}^{k-\tau_k-1} e_i \right\|^2 
\]
\[
\leq 8\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left( \left( \sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F^2 \right)^2 \right) \\
+ 2 \sum_{i=1}^{n} p_k \sum_{k=0}^{K-1} \mathbb{E} \left( \sum_{t=0}^{k-\tau_k-1} \left\| \Omega_t \right\|_{1,2}^2 \left\| \Lambda_t^{k-\tau_k-1} \right\|_2^2 \left\| e_t \right\|_1^2 \right)^2 \\
\leq 8\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left( \left( \sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F^2 \right)^2 \right) \\
+ 8 \sum_{i=1}^{n} p_k \sum_{k=0}^{K-1} \mathbb{E} \left( \sum_{t=0}^{k-\tau_k-1} \left\| \Omega_t \right\|_{1,2}^2 \left\| \widetilde{G}_{t-\tau_t} \right\|_F^2 \left\| e_t \right\|_1 \right)^2 \\
\leq 8\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left( \left( \sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F^2 \right)^2 \right) \\
+ 32t_{mix}^2 \sum_{i=1}^{n} p_k \sum_{k=0}^{K-1} \mathbb{E} \left\| \Omega_k \right\|_{1,2}^2 \\
\leq 8\alpha^2 \sum_{k=0}^{K-1} \mathbb{E} \left( \left( \sum_{t=0}^{k-\tau_k-1} \left\| \widetilde{G}_{t-\tau_t} \right\|_F^2 \right)^2 \right) + 128\delta^2 \theta^2 d_{mix}^2 K \\
\leq 32\alpha^2 t_{mix}^2 \sum_{k=0}^{K-1} \mathbb{E} \left\| \widetilde{G}_{k-\tau_k} \right\|_F^2 + 128\delta^2 \theta^2 d_{mix}^2 K \\
\]

Note that for the first term, we have

\[
\sum_{k=0}^{K-1} \mathbb{E} \left\| \widetilde{G}_{k-\tau_k} \right\|_F^2 = \sum_{k=0}^{K-1} \mathbb{E} \left\| g_{k-\tau_k,i_k} \right\|^2 \\
= \sum_{k=0}^{K-1} \mathbb{E} \left\| g_{k-\tau_k,i_k} - g_{k-\tau_k,j_k} \right\|^2 + \sum_{k=0}^{K-1} \mathbb{E} \left\| g_{k-\tau_k,j_k} \right\|^2 \\
\leq \sigma^2 K + \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_k \mathbb{E} \left\| g_{t-\tau_t,i} \right\|^2 \\
\leq (\sigma^2 + 6\kappa^2) K + 12L^2 \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_k \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{I_n}{n} - e_i \right) \right\|^2 + 2 \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_k g_{k-\tau_k,i} \right\|^2 \\
\]

Putting these two terms back, we obtain

\[
\sum_{k=0}^{K-1} \sum_{i=1}^{n} p_k \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{I_n}{n} - e_i \right) \right\|^2 \\
\leq 32\alpha^2 t_{mix}^2 \left( (\sigma^2 + 6\kappa^2) K + 12L^2 \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_k \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{I_n}{n} - e_i \right) \right\|^2 + 2 \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_k g_{k-\tau_k,i} \right\|^2 \right) + 128\delta^2 \theta^2 d_{mix}^2 K \\
\]

Rearrange the terms, we obtain

\[
(1 - 192\alpha^2 t_{mix}^2 L^2) \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_k \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{I_n}{n} - e_i \right) \right\|^2 \\
\leq 32\alpha^2 t_{mix}^2 \left( (\sigma^2 + 6\kappa^2) K + 2 \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_k g_{k-\tau_k,i} \right\|^2 \right) + 128\delta^2 \theta^2 t_{mix}^2 K \\
\]
Lemma 20

Under review as a conference paper at ICLR 2020

\[ \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 \leq \frac{32\alpha^2}{A_1} \left( (\sigma^2 + 2\epsilon^2) p K + 2p \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 + G^2 d K \right) \]

Let \( A_1 = 1 - 192 p \alpha^2 \frac{L^2}{\tau K} \), we obtain

\[ \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 \leq \frac{32\alpha^2}{A_1} \left( (\sigma^2 + 2\epsilon^2) p K + 2p \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 + G^2 d K \right) \]

Lemma 20

\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 + \left( 1 - \frac{2\alpha L}{n} \right) \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla \overline{F}(X_{k-\tau_k}) \right\|^2 \leq \frac{2n(f(0) - f^*)}{\alpha K} + \frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \frac{1}{n} \right\|^2 \]

\[ + \left( 2L^2 + \frac{12\alpha L^3}{n} \right) \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 + \frac{(\sigma^2 + 2\epsilon^2)\alpha L}{n} \]

Proof

We start from \( f(X_{k+1}) \) Since

\[ X_{k+1} = X_k W_k \frac{1}{n} + (Q_k - X_k)(W_k - I) \frac{1}{n} - \alpha \overline{G}_{k-\tau_k} = \overline{X}_k - \alpha \overline{G}_{k-\tau_k} \]

Then from Taylor Expansion, we have

\[ \mathbb{E} f(\overline{X}_{k+1}) = \mathbb{E} f(X_k) \]

\[ = \mathbb{E} f(X_k) - \alpha \mathbb{E} \left\| \nabla f(X_k), G_{k-\tau_k} \right\| + \alpha^2 \frac{L}{2} \mathbb{E} \left\| G_{k-\tau_k} \right\|^2 \]

\[ \leq \mathbb{E} f(X_k) - \alpha \mathbb{E} \left\| \nabla f(X_k), \overline{G}_{k-\tau_k} \right\| - \alpha \mathbb{E} \left\| \nabla f(X_k), G_{k-\tau_k} - \overline{G}_{k-\tau_k} \right\| + \frac{\alpha^2 L}{2} \mathbb{E} \left\| \overline{G}_{k-\tau_k} \right\|^2 \]

\[ \leq \mathbb{E} f(X_k) - \alpha \mathbb{E} \left\| \nabla f(X_k), \nabla \overline{F}(X_{k-\tau_k}) \right\| + \frac{\alpha^2 L}{2} \mathbb{E} \left\| \nabla f(X_k), \nabla \overline{F}(X_{k-\tau_k}) \right\| \]

\[ \leq \mathbb{E} f(X_k) - \frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(X_k), \nabla \overline{F}(X_{k-\tau_k}) \right\|^2 + \frac{\alpha^2 L}{2n^2} \mathbb{E} \left\| g_{k-\tau_k,i} \right\|^2 + \frac{\alpha^2 L \sigma^2}{4n^2} + \frac{\alpha^2 L}{2n^2} \sum_{i=1}^{n} p_i \mathbb{E} \left\| g_{k-\tau_k,i} \right\|^2 \]

Rearrange these terms, we can get

\[ \frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 + \frac{\alpha}{2n} \mathbb{E} \left\| \nabla \overline{F}(X_{k-\tau_k}) \right\|^2 \]

\[ \leq \mathbb{E} f(X_k) - \mathbb{E} f(X_{k+1}) + \frac{\alpha}{2n} \mathbb{E} \left\| \nabla f(X_k) - \nabla \overline{F}(X_{k-\tau_k}) \right\|^2 \]
Putting it back, we have

\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 + \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla F(X_{k-\tau_k}) \right\|^2 \]

\[ \leq \frac{2n(f(0) - f^*)}{\alpha K} + \frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \mathbb{1}_n \right\|^2 \]

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

Using Lemma 12, we can simplify this further.

\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 + \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla F(X_{k-\tau_k}) \right\|^2 \]

\[ \leq \frac{2n(f(0) - f^*)}{\alpha K} + \frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \mathbb{1}_n \right\|^2 \]

\[ + \frac{2L^2}{K} \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 + \frac{\alpha L \sigma^2}{n} + \frac{\alpha L}{nK} \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| g_{k-\tau_k, i} \right\|^2 \]
As a result

\[ \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{k=0}^{K-1} p_i g_{k-\tau_k} \right\|^2 \]

Note that

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

Moving it to the left side, we finally get

\[ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla f(X_k) \right\|^2 + \left( 1 - \frac{2\alpha L}{n} \right) \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \nabla F(X_{k-\tau_k}) \right\|^2 \]

\[ \leq \frac{2n(f(0) - f^*)}{\alpha K} + \frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \mathbb{I}_n \right\|^2 \]

\[ + \left( 2L^2 + \frac{12\alpha L^3}{n} \right) \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 + \frac{(\sigma^2 + 6\sigma^2)\alpha L}{n} \]

That completes the proof.

**Lemma 21** For all k ≥ 0, we have

\[ \frac{2L^2}{K} \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \mathbb{I}_n \right\|^2 \]

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

\[ + \frac{4\alpha^2 T^2 L^2}{n^2K} \sum_{k=0}^{K-1} \mathbb{E} \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \]

**Proof** From Lemma 20 we know the fact

\[ X_{k+1} = X_k W_k \frac{1}{n} + (Q_k - X_k)(W_k - I) \frac{1}{n} - \alpha \overline{G}_{k-\tau_k} = X_k - \alpha \overline{G}_{k-\tau_k} \]

As a result

\[ \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \mathbb{I}_n \right\|^2 \]

\[ = \sum_{k=0}^{K-1} \mathbb{E} \left\| \tau_k \sum_{i=1}^{\tau_k} \alpha \overline{G}_{k-t} \mathbb{I}_n \right\|^2 \]

\[ \leq \alpha^2 \tau_k \sum_{t=1}^{\tau_k} \mathbb{E} \left\| \overline{G}_{k-t} \mathbb{I}_n \right\|^2 \]

\[ \leq \alpha^2 \tau_k \sum_{t=1}^{\tau_k} \left( \frac{\sigma^2}{n^2} + \frac{1}{n^2} \sum_{i=1}^{n} p_i \mathbb{E} \left\| g_{k-t,i} \right\|^2 \right) \]

\[ \leq \frac{\alpha^2 T^2 \sigma^2 K}{n^2} + \frac{\alpha^2 T^2 \sigma^2 K}{n^2} \sum_{k=0}^{K-1} \sum_{t=1}^{\tau_k} \sum_{i=1}^{n} p_i \mathbb{E} \left\| g_{k-t,i} \right\|^2 \]

\[ \leq \frac{\alpha^2 T^2 \sigma^2 K}{n^2} + \frac{\alpha^2 T^2 \sigma^2 K}{n^2} \sum_{k=0}^{K-1} \sum_{t=1}^{\tau_k} \left( 12L^2 \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-t} \left( \frac{1}{n} - e_i \right) \right\|^2 + 6\sigma^2 + 2E \left\| \sum_{i=1}^{n} p_i g_{k-t,i} \right\|^2 \right) \]

\[ \leq \frac{\alpha^2 T^2 \sigma^2 K}{n^2} + \frac{\alpha^2 T^2 \sigma^2 K}{n^2} \sum_{k=0}^{K-1} \left( 12L^2 \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2 + 6\sigma^2 + 2E \left\| \sum_{i=1}^{n} p_i g_{k-\tau_k,i} \right\|^2 \right) \]
And we get
\[
\frac{2L^2K}{n^2} \sum_{k=0}^{K-1} \mathbb{E} \left\| (X_k - X_{k-\tau_k}) \frac{1}{n} \right\|^2 \leq \frac{2\alpha^2 T^2(\sigma^2 + 6\xi^2)K}{n^2} + \frac{24L^2\alpha^2 T^2}{n^2K} \sum_{k=0}^{K-1} n \sum_{i=1}^{n} p_i \mathbb{E} \left\| X_{k-\tau_k} \left( \frac{1}{n} - e_i \right) \right\|^2
\]
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
\[
a_t = \sum_{s=1}^{t-\tau_k} \rho^{ \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho} b_s
\]
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
\[
S_k = \sum_{t=1}^{k} a_t = \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_k} \rho^{ \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho} b_s \leq \sum_{t=1}^{k} \sum_{s=1}^{t} \rho^{ \max \left( \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho \right) } b_s = \sum_{t=1}^{k} \sum_{s=1}^{t} \rho^{ \max \left( \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho \right) } b_s
\]
\[
D_k = \sum_{t=1}^{k} a_t^2 = \sum_{t=1}^{k} \sum_{s=1}^{t-\tau_k} \rho^{ \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho} b_s \sum_{r=1}^{t-\tau_k} \rho^{ \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho} b_r = \sum_{t=1}^{k} \sum_{s=1}^{t} \sum_{r=1}^{t} \rho^{ \lfloor \frac{t-\tau_k-\rho}{T} \rfloor - \rho} b_s b_r
\]
Lemma 23 for \( \forall i, j \) and \( \forall k \geq 0 \), we have
\[
\| X_k(e_i - e_j) \|_{\infty} \leq \theta = 16t_{\text{mix}} \alpha G_{\infty}
\]
**Proof** Similar to Section 6 and Section 7, we use mathematical induction to prove this.

I. First, for \( k = 0 \), we have
\[
\| X_0(e_i - e_j) \|_\infty = 0 \leq \theta = 16t_{\text{mix}}\alpha G_\infty
\]

II. Suppose for \( k \geq 0 \), we have \( \| X_t(e_i - e_j) \|_\infty \leq \theta, \forall t \leq k \), then we have
\[
\| X_{k+1}(e_i - e_j) \|_\infty \\
\leq \| X_{k+1}\left(\frac{I}{n} - e_i\right) \|_\infty + \| X_{k+1}\left(\frac{I}{n} - e_j\right) \|_\infty \\
\leq \| X_{k+1}\left(I - \frac{I}{n}\right) \|_{1,\infty} \| e_i \|_1 + \| X_{k+1}\left(I - \frac{I}{n}\right) \|_{1,\infty} \| e_j \|_1 \\
= 2\| X_{k+1}\left(I - \frac{I}{n}\right) \|_{1,\infty} \\
\leq 2\left\| X_kW_k - \alpha \tilde{G}_{k_\tau_k} + \Omega_k\right\|_{1,\infty} \\
= 2\left\| \sum_{t=0}^{k} \left(-\alpha \tilde{G}_{t_\tau_t} + \Omega_t\right) \left( \prod_{q=t+1}^{k} W_q - \frac{I}{n}\right) \right\|_{1,\infty} \\
\leq 2\sum_{t=0}^{k} \left\| -\alpha \tilde{G}_{t_\tau_t} + \Omega_t\right\|_{1,\infty} \left\| \prod_{q=t+1}^{k} W_q - \frac{I}{n}\right\|_{1,\infty} \\
\leq 2\sum_{t=0}^{k} \left\| -\alpha \tilde{G}_{t_\tau_t} + \Omega_t\right\|_{1,\infty} \left\| \prod_{q=t+1}^{k} W_q - \frac{I}{n}\right\|_{1,\infty} \\
\leq 4(\alpha G_\infty + 2\delta \theta) \sum_{t=0}^{k} 2^{-\left\lfloor (k-t)/t_{\text{mix}} \right\rfloor} \\
\leq 4(\alpha G_\infty + 2\delta \theta) \sum_{t=0}^{t_{\text{mix}}-1} \sum_{r=0}^{\infty} 2^{-r} \\
\leq 8(\alpha G_\infty + 2\delta \theta)t_{\text{mix}}
\]

Put in \( \delta = \frac{1}{2t_{\text{mix}}} \), we obtain
\[
\| X_{k+1}(e_i - e_j) \|_2 \leq 8(\alpha G_\infty + 2\delta \theta)t_{\text{mix}} = 8t_{\text{mix}}\alpha G_\infty + 8t_{\text{mix}}\alpha G_\infty = 16t_{\text{mix}}\alpha G_\infty
\]

Combining I and II and we complete the proof.