Optimal Complexity in Decentralized Training (Supplementary Materials)

A. Experimental Details

A.1. Hyperparameter Tuning

In the experiment of training LeNet on CIFAR10, we tune the step size using grid search inside the following range: {5e-3, 1e-3, 5e-4, 2.5e-4, 1e-4, 5e-5}. Note that this range is in general smaller than the one chosen in (Zhang & You, 2019c), since here we are working with unshuffled data, and we found original range in baselines causes algorithms to diverge easily. Following (Tang et al., 2018b), we let each run warm up for 10 epochs with step size 1e-5. For DeTAG, we further tune the accelerated gossip parameter η within $\{0, 0.1, 0.2, 0.4\}$ and phase length R within $\{1, 2, 3\}$. We fix the momentum term to be 0.9 and weight decay to be 1e-4.

In the experiment of training Resnet20 on CIFAR100, we tune the step size using grid search inside the following range: $\{0.5, 0.1, 0.05, 0.01, 0.005\}$. For DeTAG, we further tune the accelerated gossip parameter η within $\{0, 0.1, 0.2, 0.4\}$ and phase length R within $\{1, 2, 3\}$. We fix the momentum term to be 0.9 and weight decay to be 5e-4.

The hyperparameters adopted for each runs are shown in Table 3 and Table 4.

A.2. Techniques of Running DeTAG

We can see in the main loop of DeTAG, several gradient queries are made at the same point. This essentially is equivalent to a large mini-batch size. In practice, however, we can modify this to use local-steps and get better empirical results (Lin et al., 2018). Another technique is to use warm-up epochs when data is decentralized. We observe it ensures a smooth convergence in practice. Last but not least, since at first the noise in the algorithms is generally large, we can use a dynamic phase length to obtain better results. That is, we start from phase length 1 for the first few epochs, and let DeTAG follow the special case of DSGT. Then we can gradually increase the phase length following given policies. The intuition is that as algorithm converges, we would need less noise from communication, and thus a longer phase length can benefit.

Table 3. (Initial) Step size α used for each experiments.

Experiment	Setting	Algorithm			
		D-PSGD	D^2	DSGT	DeTAG
LeNet/CIFAR10	100% Shuffled	5e-3	5e-3	5e-3	5e-3
	50% Shuffled	5e-5	2.5e-4	2.5e-4	5e-4
	25% Shuffled	5e-5	1e-4	2.5e-4	5e-4
	0% Shuffled	5e-5	1e-4	2.5e-4	5e-4
Resnet20/CIFAR100	$\kappa = 1$	0.5	0.5	0.5	0.5
	$\kappa = 0.1$	0.5	0.5	0.5	0.5
	$\kappa = 0.05$	0.5	0.5	0.5	0.5
	$\kappa = 0.01$	0.5	0.5	0.5	0.5

Table 4. DeTAG-specific hyperparameters used for each experiments.

Experiment	Setting	Accelerate Factor η	Phase Length R
LeNet/CIFAR10	100% Shuffled	0	1
	50% Shuffled	0.2	2
	25% Shuffled	0.2	2
	0% Shuffled	0.2	2
Resnet20/CIFAR100	$\kappa = 1$	0	1
	$\kappa = 0.1$	0.2	2
	$\kappa = 0.05$	0.2	2
	$\kappa = 0.01$	0.4	2

B. Technical Proof

B.1. Proof to Theorem 1

Proof. To prove this theorem, it suffices for us to provide two examples, each has a (set of) loss function $f \in \mathcal{F}_{\Delta,L}$, a set of underlying oracles $O \in \mathcal{O}_{\sigma^2}$, a graph $G \in \mathcal{G}_{n,D}$, such that $\inf_{A\in\mathcal{A}_B} T_\epsilon(A, f, O, G)$ is lower bounded by $\Omega\left(\frac{\Delta L\sigma^2}{nB\epsilon^4}\right)$ and $\Omega\left(\frac{\Delta LD}{\epsilon^2}\right)$ iterations on these two examples, respectively. Then we will obtain the final bound as $\max\left\{\Omega\left(\frac{\Delta L\sigma^2}{nB\epsilon^4}\right), \Omega\left(\frac{\Delta LD}{\epsilon^2}\right)\right\}$, i.e., $\Omega\left(\frac{\Delta L\sigma^2}{nB\epsilon^4} + \frac{\Delta LD}{\epsilon^2}\right)$ as desired. For simplicity, we denote $\boldsymbol{z}^{(i)}$ as the i-th coordinate of vector $\boldsymbol{z} \in \mathbb{R}^d$.

For each setting, our constructions contain three main steps.

(1) The first step is to follow the construction of a zero chain function model (Carmon et al., 2017; 2019). Following (Arjevani et al., 2019) and define

$$\text{prog}(\boldsymbol{z}) = \max\{i \geq 0 | \boldsymbol{z}^{(i)} \neq 0\}, \forall \boldsymbol{z} \in \mathbb{R}^d. \tag{18}$$

A zero chain function f has the following property:

$$\text{prog}(\nabla f(\boldsymbol{x})) \leq \text{prog}(\boldsymbol{x}) + 1, \tag{19}$$

that means, for a model start from $\boldsymbol{x} = \mathbf{0}$, a single gradient evaluation can only make at most one more coordinate to be non-zero. The name of "chain" comes from the fact that the adjacent coordinates are linked like a chain and only if the previous coordinate becomes non-zero that the current coordinate can become non-zero via a gradient update. Consider a model with d dimension, if we show that $\|\nabla f(\boldsymbol{x})\| \geq \epsilon$ for any $\boldsymbol{x} \in \mathbb{R}^d$ with $\boldsymbol{x}^{(d)} = 0$, we will obtain d as a lower bound on the gradient calls to obtain the ϵ-stationary point. We refer such sequential lower bound as T_0.

(2) Step two is to construct a graph $G \in \mathcal{G}_{n,D}$ and a set of oracle $O \in \mathcal{O}_{\sigma^2}$. To do this, our basic idea is to follow (Arjevani et al., 2019) and introduce randomness on the $\text{prog}(\boldsymbol{x})$, and thus the whole chain only make progress with probability p. As will be shown later, this requires $\Omega(T_0/p)$ iterations in total.

(3) The third and last step is to rescale the function and distribution so as to make it belong to the function and oracle classes we consider. In other words, this step is to guarantee the result is shown in terms of Δ, L, σ, n and D.

We start from a smooth and (potentially) non-convex zero chain function $\hat{f}$ (Carmon et al., 2019) as defined below:

$$\hat{f}(\boldsymbol{x}) = -\Psi(1)\Phi(\boldsymbol{x}^{(1)}) + \sum_{i=1}^{T-1}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)}\Phi(\boldsymbol{x}^{(i+1)})], \tag{20}$$

where for $\forall z \in \mathbb{R}$

$$\Psi(z) = \begin{cases} 0 & z \leq 1/2 \\ \exp\left(1 - \frac{1}{(2z-1)^2}\right) & z > 1/2 \end{cases}, \qquad \Phi(z) = \sqrt{e}\int_{-\infty}^{z} e^{\frac{1}{2}t^2} dt. \tag{21}$$

This function, as shown in previous works (Carmon et al., 2019; Arjevani et al., 2019), is a zero-chain function and thus is generally "hard" to optimize: it costs at least T gradient evaluations to find a stationary point. We summarize some properties of Equation (20) as the following (Proof can be found in Lemma 2 in (Arjevani et al., 2019)):

1. $\hat{f}(\boldsymbol{x}) - \inf_x \hat{f}(\boldsymbol{x}) \leq \Delta_0 T, \forall \boldsymbol{x} \in \mathbb{R}^d$, where $\Delta_0 = 12$.
2. $\hat{f}$ is l_1-smooth, where $l_1 = 152$.
3. $\forall \boldsymbol{x} \in \mathbb{R}^T$, $\|\nabla \hat{f}(\boldsymbol{x})\|_\infty \leq G_\infty$, where $G_\infty = 23$.
4. $\forall \boldsymbol{x} \in \mathbb{R}^T$, if $\text{prog}(\boldsymbol{x}) < T$, then $\|\hat{f}(\boldsymbol{x})\|_\infty \geq 1$.

(Setting 1) Next we discuss the first setting with lower bound $\Omega\left(\frac{\Delta L \sigma^2}{nB\epsilon^4}\right)$. (Setting 1, Step 1) The loss functions are defined as

$$\hat{f}_i(\boldsymbol{x}) = \hat{f}(\boldsymbol{x}), \tag{22}$$

note that $1/n \sum_{i=1}^n \hat{f}_i = \hat{f}$. It can be seen from Property 2 that all the $\hat{f}_i$ are l_1-smooth. (Setting 1, Step 2) For this setting we consider complete graph. We construct the oracle on worker i as the following:

$$[\hat{g}_i(\boldsymbol{x})]_j = \nabla_j \hat{f}_i(\boldsymbol{x}) \cdot \left(1 + \mathbb{1}\{j > \text{prog}(\boldsymbol{x})\}\left(\frac{z}{p} - 1\right)\right), \tag{23}$$

where $z \sim \text{Bernoulli}(p)$. It can be seen that

$$\mathbb{E}[\hat{g}_i(\boldsymbol{x})] = \nabla \hat{f}_i(\boldsymbol{x}), \tag{24}$$

and from Property 3 we know

$$\mathbb{E}\|\hat{g}_i(\boldsymbol{x}) - \nabla \hat{f}_i(\boldsymbol{x})\|^2 = |\nabla_{\text{prog}(\boldsymbol{x})+1} \hat{f}(\boldsymbol{x})|^2 \mathbb{E}\left(\frac{z}{p} - 1\right)^2 \leq \frac{\|\nabla \hat{f}_i(\boldsymbol{x})\|_\infty^2 (1-p)}{p} \leq \frac{\|\nabla \hat{f}(\boldsymbol{x})\|_\infty^2 (1-p)}{p} \leq \frac{G_\infty^2 (1-p)}{p}.$$

(Setting 1, Step 3) Finally we rescale each function as $f_i = L\lambda^2 / l_1 \hat{f}_i(\boldsymbol{x}/\lambda)$ where λ is a parameter subject to change. For L: note that all f_i are $\frac{L}{l_1} \cdot l_1 = L$-smooth. For the Δ,

$$f - f^* = \frac{L\lambda^2}{l_1}(\hat{f} - \hat{f}^*) = \frac{L\lambda^2 \Delta_0 T}{l_1} \leq \Delta. \tag{25}$$

For the oracle, to be consistent with f_i, we rescale it as $g_i(\boldsymbol{x}) = L\lambda / l_1 \hat{g}_i(\boldsymbol{x}/\lambda)$, and we have

$$\mathbb{E}\|g_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x})\|^2 \leq \frac{L^2\lambda^2}{l_1^2} \mathbb{E}\left\|g_i\left(\frac{\boldsymbol{x}}{\lambda}\right) - \nabla f_i\left(\frac{\boldsymbol{x}}{\lambda}\right)\right\|^2 \leq \frac{L^2\lambda^2 G_\infty^2 (1-p)}{l_1^2 p} \leq \sigma^2. \tag{26}$$

We assign $\lambda = 2l_1\epsilon/L$, then Equation (25) and (26) are fulfilled with

$$T = \left\lfloor \frac{\Delta}{\Delta_0 l_1 (2\epsilon)^2} \right\rfloor,$$
$$p = \min\{(2G_\infty \epsilon)^2 / \sigma^2, 1\}.$$

Take $\delta = 1/2$ in Lemma 2, we have for probability at least 1/2, $\|\nabla f(\hat{\boldsymbol{x}}^{(t)})\| \geq \epsilon$ for all $t \leq \frac{T + \log(\delta)}{\min\{nBp, 1\}(e-1)}$. Use Property 4, for any $\boldsymbol{x} \in \mathbb{R}^T$ such that $\text{prog}(\boldsymbol{x}) < T$ it holds that $\|\nabla f(\boldsymbol{x})\| \geq 2\epsilon$, therefore,

$$\mathbb{E}\|\nabla f(\hat{\boldsymbol{x}}_T)\| > \epsilon. \tag{27}$$

Then with small ϵ it follows that

$$T_\epsilon(A, f, O, G) \geq \frac{T-1}{nBp(e-1)} \geq \Omega\left(\frac{\Delta L \sigma^2}{nB\epsilon^4}\right), \tag{28}$$

and that completes the proof for setting 1.

(Setting 2) We proceed to the prove second bound $\Omega\left(\frac{\Delta L D}{\epsilon^2}\right)$.

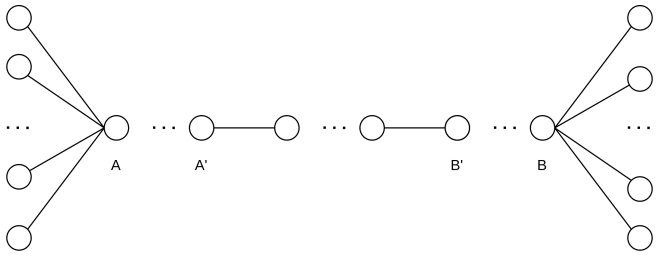

Figure 4. Illustration graph for setting 2 to in the proof of Theorem 1.

(Setting 2 Step 1 & Step 2) We assign all the workers with index from 1 to n, we first define two indices set

$$\begin{aligned} I_0 &= \{1, \cdots, |I_0|\}, \\ I_1 &= \{n, n-1, \cdots, n-|I_1|+1\}. \end{aligned} \tag{29}$$

where $|\cdot|$ denotes a cardinality of a set. Consider the construction of G in Figure 4:

If $D \geq n - 2\lceil n/3 \rceil + 2$, then it implies the number of nodes between A and B is larger than $\lceil n/3 \rceil$. In this case, denote A' and B' as a sub linear graph where its number of nodes is exactly $\lceil n/3 \rceil$. Let all the nodes on the left of A' be in I_0 and all the nodes on the right of B' be I_1 We define all the local functions on such graph as following:

$$\hat{f}_i(\boldsymbol{x}) = \begin{cases} -\frac{2n}{n-\lceil n/3 \rceil}\Psi(1)\Phi(\boldsymbol{x}^{(1)}) + \sum_{i=2k, k\in\{1,2,\cdots\}, i<T} \frac{2n}{n-\lceil n/3 \rceil}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)})\Phi(\boldsymbol{x}^{(i+1)})] & i \in I_0, \\ \sum_{i=2k-1, k\in\{1,2,\cdots\}, i<T} \frac{2n}{n-\lceil n/3 \rceil}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)})\Phi(\boldsymbol{x}^{(i+1)})] & i \in I_1, \\ 0 & i \notin I_0, I_1. \end{cases} \tag{30}$$

If $D < n - 2\lceil n/3 \rceil + 2$, the distance between node A and node B is $D-2$ and the sub linear graph whose end points are A and B contains $D-1$ nodes. We let the number of nodes on the left of A be $\lceil \frac{n-D+1}{2} \rceil$, we denote the set of indices of all such nodes as I_0; and then we let the number of nodes on the right of B be $\lfloor \frac{n-D+1}{2} \rfloor$, we denote the set of indices of all such nodes as I_I. Since $D < n - 2\lceil n/3 \rceil + 2$, this implies $|I_0|, |I_1| > n/3$. We define all the local functions on such graph as following:

$$\hat{f}_i(\boldsymbol{x}) = \begin{cases} -\frac{n}{|I_0|}\Psi(1)\Phi(\boldsymbol{x}^{(1)}) + \sum_{i=2k, k\in\{1,2,\cdots\}, i<T} \frac{n}{|I_0|}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)})\Phi(\boldsymbol{x}^{(i+1)})] & i \in I_0, \\ \sum_{i=2k-1, k\in\{1,2,\cdots\}, i<T} \frac{n}{|I_1|}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)})\Phi(\boldsymbol{x}^{(i+1)})] & i \in I_1, \\ 0 & i \notin I_0, I_1. \end{cases} \tag{31}$$

In both cases discussed based on D, we can see that $\hat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \hat{f}_i(\boldsymbol{x})$, and we are splitting hard zero-chain function into two main different part: the even components of the chain and the odd components of the chain. It is easy to see that for the zero chain function to make progress, it takes at least $\lceil n/3 \rceil$, i.e., $\Omega(D)$ number of iterations in the first case (since here $D = \tilde{\gamma} n$ for some $\tilde{\gamma} > 1/3$) and D number of iterations in the seconds case. Then the total number of iterations is lower bounded by $\Omega(TD)$.

For the oracle, we let oracle on worker i as

$$[\hat{g}_i(\boldsymbol{x})]_j = \nabla_j \hat{f}_i(\boldsymbol{x}). \tag{32}$$

(Setting 2, Step 3) The last step is to rescale the parameters. Compared to setting 1, we know here all the $\hat{f}_i$ are $3l_1$-smooth, as before we let

$$f_i(\boldsymbol{x}) = \frac{L\lambda^2}{3l_1}\hat{f}_i\left(\frac{\boldsymbol{x}}{\lambda}\right), \qquad \lambda = \frac{6l_1\epsilon}{L}. \tag{33}$$

For the Δ bound we have

$$L\lambda^2\Delta_0 T/3l_1 \leq \Delta \tag{34}$$

to fulfill this it suffices to set

$$T = \left\lfloor \frac{\Delta L}{\Delta_0 l_1 (12\epsilon)^2} \right\rfloor. \tag{35}$$

It also can be seen that f is L-smooth. So in this setting,

$$T_\epsilon(A, f, O, G) \geq \Omega(TD) = \Omega\left(\frac{\Delta L D}{\epsilon^2}\right). \tag{36}$$

Combining Setting 1 and 2 we complete the proof. □

Lemma 2. *In setting 1 in the proof of Theorem 1, with probability at least* $1-\delta$, $\|\nabla f(\boldsymbol{x}_t)\| \geq \epsilon$ *for all* $t \leq \frac{T+\log(\delta)}{\min\{nBp,1\}(e-1)}$.

Proof. Define a filtration at iteration t as the sigma field of all the previous events happened before iteration t. Let $i_j^{(t)} = \text{prog}(\boldsymbol{x}_{t,j}), \forall j \in [n]$ and $i^{(t)} = \max_j i_j^{(t)}$. And we denote $\mathcal{E}^{(t,m,j)}$ as the event of `the` $i_m^{(t)}+1$`-th coordinate of output of` j`-th query on worker` m `at iteration` t `is non-zero`. Based on the independent sampling, these events are independent. Thus we know:

$$\mathbb{P}[i^{(t+1)} - i^{(t)} = 1 | \mathcal{U}^{(t)}] = \mathbb{P}\left[\bigcup_{\substack{i\in[n]\\ j\leq B}} \mathcal{E}^{(t,i,j)} | \mathcal{U}^{(t)}\right] \leq \sum_{i\in[n], j\leq B} \mathbb{P}\left[\mathcal{E}^{(t,i,j)} | \mathcal{U}^{(t)}\right] \leq \min\{nBp, 1\}. \tag{37}$$

Let $q^{(t)} = i^{(t+1)} - i^{(t)}$, with Chernoff bound, we obtain

$$\mathbb{P}[i^{(t)} \geq T] = \mathbb{P}[e^{\sum_{j=0}^{t-1} q^{(j)}} \geq e^T] \leq e^{-T}\mathbb{E}[e^{\sum_{j=0}^{t-1} q^{(j)}}]. \tag{38}$$

For the expectation term we know that

$$\mathbb{E}[e^{\sum_{j=0}^{t-1} q^{(j)}}] = \mathbb{E}\left[\prod_{j=0}^{t-1} \mathbb{E}\left[e^{q^{(j)}} | \mathcal{U}^{(j)}\right]\right] \leq (1 - \min\{nBp,1\} + \min\{nBp,1\}e)^t \leq e^{\min\{nBp,1\}t(e-1)}. \tag{39}$$

Thus we know

$$\mathbb{P}[i^{(t)} \geq T] \leq e^{(e-1)\min\{nBp,1\}t - T} \leq \delta, \tag{40}$$

for every $t \leq \frac{T+\log(\delta)}{\min\{nBp,1\}(e-1)}$. □

B.2. Proof to Corollary

Proof. The proof of the Corollary consists of two parts: In the first part, we first prove given any n, how to construct the graph and the gossip matrix with $\lambda = 0, \cos(\pi/n)$. Then we proceed to discuss how the other $\lambda \in (0, \cos(\pi/n))$ can be achieved.

We start from the first part of proving the lower bound. We propose two settings of construction. The first setting is the same as the one shown in the proof of Theorem 1 as the complete graph. For setting 2, consider the two special cases of dumbbell graph in Figure 4: if the graph is fully connected, then we can use the average consensus matrix $\boldsymbol{W}$ that fulfills $\lambda = 0$, and this can be seen as the special case where $D = 1$; on the other hand, if the graph is a linear graph, we can first prove the lower bound using the diameter as follows:

(Linear graph, Step 1) We first let $|I_0| = |I_1| = \lceil n/3 \rceil$ in the proof of Theomrem 1, meaning I_0 denotes the first $\lceil n/3 \rceil$ workers and I_1 denotes the last $\lceil n/3 \rceil$ workers. We define all the local functions as following:

$$\hat{f}_i(\boldsymbol{x}) = \begin{cases} -\frac{n}{\lceil n/3 \rceil}\Psi(1)\Phi(\boldsymbol{x}^{(1)}) + \sum_{i=2k, k\in\{1,2,\cdots\}, i<T} \frac{n}{\lceil n/3 \rceil}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)})\Phi(\boldsymbol{x}^{(i+1)})] & i \in I_0, \\ \sum_{i=2k-1, k\in\{1,2,\cdots\}, i<T} \frac{n}{\lceil n/3 \rceil}[\Psi(-\boldsymbol{x}^{(i)})\Phi(-\boldsymbol{x}^{(i+1)}) - \Psi(\boldsymbol{x}^{(i)})\Phi(\boldsymbol{x}^{(i+1)})] & i \in I_1, \\ 0 & i \notin I_0, I_1. \end{cases} \tag{41}$$

we can see that $\hat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_i(\boldsymbol{x})$. (Linear graph, Step 2) We consider linear graph in this setting and from one end to the other, the worker's index is 1 to n, without the loss of generality. It is easy to see that for the zero chain function to make progress, it takes at least $n - 2\lceil n/3\rceil + 1$ number of iterations. Note that in linear graph $n - 1 = D$, the total number of iterations is at least

$$\Omega\left(TD\right). \tag{42}$$

For the oracle, we let oracle on worker i as

$$[\hat{g}_i(\boldsymbol{x})]_j = \nabla_j \hat{f}_i(\boldsymbol{x}) \tag{43}$$

(Linear graph, Step 3) The last step is to rescale the parameters. Compared to setting 1, we know here all the $\hat{f}_i$ are $3l_1$-smooth, as before we let

$$f_i(\boldsymbol{x}) = \frac{L\lambda^2}{3l_1}\hat{f}_i\left(\frac{\boldsymbol{x}}{\lambda}\right), \qquad \lambda = \frac{6l_1\epsilon}{L}. \tag{44}$$

For the Δ bound we have

$$L\lambda^2\Delta_0 T/3l_1 \leq \Delta, \tag{45}$$

to fulfill this it suffices to set

$$T = \left\lfloor \frac{\Delta L}{\Delta_0 l_1 (12\epsilon)^2} \right\rfloor. \tag{46}$$

It also can be seen that f is L-smooth. So in this setting,

$$T_\epsilon(A, f, O, G) \geq \Omega(TD) \geq \Omega\left(\frac{\Delta L D}{\epsilon^2}\right). \tag{47}$$

Given the bound, we use two additional results on linear graph as (Berthier et al., 2020): the random walk matrix $\boldsymbol{W}_{rw}$ on linear graph with λ fulfilling

$$\frac{1}{\sqrt{1-\lambda}} = O(D). \tag{48}$$

Then we can rewrite the lower bound in the form of λ as shown in Corollary 1.

So far we've proved the two boundary cases, now the dumbbell graph can be seen as the intermediates between the two boundary cases. Take any dumbbell graph as shown in Figure 4 and performs a random walk, we can easily see the second largest eigenvalue for that random walk is some $0 < \lambda' < \lambda = \cos(\pi/n)$, where $\cos(\pi/n)$ is the second largest eigenvalue for the linear graph for $n \geq 2$. Denote all these λ' associated with different dumbbell graph as $\{\tilde{\lambda}_1, \cdots, \tilde{\lambda}_K\}$, then for any $0 \leq \tilde{\lambda}_i < \lambda < \tilde{\lambda}_{i+1} \leq \cos(\pi/n)$, a corresponding matrix $\boldsymbol{W}$ with second largest eigenvalue λ can always be achieved by performing linear combination of the matrices with second largest eigenvalue $\tilde{\lambda}_i$ and $\tilde{\lambda}_{i+1}$.

Finally, using the conclusion of $\lambda = \cos(\pi/n)$ for $n \in \{2, 3, \cdots,\}$ on linear graph we complete the proof.

□

B.3. Proof to Theorem 2

Proof. As (partially) discussed in the paper, DeFacto is statistically equivalent to centralized SGD. Specifically, it conduct $K = T/2R$ gradient steps where each step contains a mini-batch of R at the point of $\boldsymbol{x}_{k,i}, \forall i \in [n]$. Take the well-known convergence rate for centralized SGD:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\hat{\boldsymbol{x}})\|^2 \leq O\left(\frac{\Delta L\sigma}{\sqrt{nBT}} + \frac{\Delta L}{T}\right). \tag{49}$$

The convergence rate of DeFacto can be expressed as:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\hat{\boldsymbol{x}})\|^2 \leq O\left(\frac{\Delta L\sigma/\sqrt{R}}{\sqrt{nBK}} + \frac{\Delta L}{K}\right) = O\left(\frac{\Delta L\sigma}{\sqrt{nBT}} + \frac{\Delta LR}{T}\right) = O\left(\frac{\Delta L\sigma}{\sqrt{nBT}} + \frac{\Delta LD}{T}\right), \tag{50}$$

then we obtain for DeFacto, when $T = O(\Delta L\sigma^2(nB\epsilon^4)^{-1} + \Delta LD\epsilon^{-2})$,

$$\min_{t=0,1,\cdots,T-1} \mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}\right)\right\| \leq \sqrt{\min_{t=0,1,\cdots,T-1} \mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}\right)\right\|^2} \leq \sqrt{O\left(\frac{\Delta L\sigma}{\sqrt{nBT}} + \frac{\Delta LD}{T}\right)} \leq \epsilon, \tag{51}$$

that completes the proof. □

B.4. Proof to Theorem 3

Proof. In this proof, we adopt an updated version of notation: we denote at the beginning of phase k, the three quantities of interests are $\boldsymbol{X}_k$, $\boldsymbol{Y}_k$ and $\tilde{\boldsymbol{G}}_k$, and the update rule becomes:

$$\boldsymbol{Y}_{k+1} = \mathcal{M}(\boldsymbol{Y}_k + \tilde{\boldsymbol{G}}_k - \tilde{\boldsymbol{G}}_{k-1}), \tag{52}$$
$$\boldsymbol{X}_{k+1} = \mathcal{M}(\boldsymbol{X}_k - \alpha\boldsymbol{Y}_k), \tag{53}$$

with

$$\tilde{\boldsymbol{G}}_{k+1} = \left[\nabla\tilde{f}_1(\boldsymbol{x}_{k,1}), \cdots, \nabla\tilde{f}_n(\boldsymbol{x}_{k,n})\right] \in \mathbb{R}^{d\times n}, \tag{54}$$
$$\boldsymbol{G}_{k+1} = [\nabla f_1(\boldsymbol{x}_{k,1}), \cdots, \nabla f_n(\boldsymbol{x}_{k,n})] \in \mathbb{R}^{d\times n}, \tag{55}$$
$$\boldsymbol{X}_k = [\boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n}] \in \mathbb{R}^{d\times n}, \tag{56}$$
$$\boldsymbol{Y}_k = \left[\boldsymbol{y}_{k,1}, \cdots, \boldsymbol{y}_{k,n}\right] \in \mathbb{R}^{d\times n}, \tag{57}$$

where $\nabla\tilde{f}_i$ denotes the stochastic gradient oracle on worker i, and ∇f_i denotes the full gradient oracle on worker i. We use $\overline{\boldsymbol{X}}$ denote $\boldsymbol{X}\frac{\mathbf{1}}{n}$ for any matrix $\boldsymbol{X}$ with appropriate shape. We use $\lambda_i(\boldsymbol{W})$ to denote the i-th general largest eigenvalue of matrix $\boldsymbol{W}$. Under such notation, λ in the main paper is equavilent to $\lambda_2(\boldsymbol{W})$. We use $\mathcal{M}(\cdot)$ to denote the R-step accelerated gossip which has the following property (Liu & Morse, 2011):

$$\|\mathcal{M}(\boldsymbol{X}) - \overline{\boldsymbol{X}}\| \leq \rho\|\boldsymbol{X} - \overline{\boldsymbol{X}}\|; \quad \mathcal{M}(\boldsymbol{X})\frac{\mathbf{1}}{n} = \boldsymbol{X}\frac{\mathbf{1}}{n}, \tag{58}$$

where $\rho = \left(1 - \sqrt{1-\lambda_2(\boldsymbol{W})}\right)^R$. The proof to the statement of Equation (58) can be found in (Ye et al., 2020).

For the stochastic oracle, based on the oracle class assumption, we have

$$\mathbb{E}\|\nabla\tilde{f}_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x})\|^2 \leq \sigma^2, \tag{59}$$

and we denote $\tilde{\sigma}^2 = \frac{\sigma^2}{BR}$ as the variance of mini-batch of R.

First, from the update rule of DeTAG,

$$\overline{\boldsymbol{Y}}_k = \mathcal{M}(\boldsymbol{Y}_{k-1} + \tilde{\boldsymbol{G}}_{k-1} - \tilde{\boldsymbol{G}}_{k-2})\frac{\mathbf{1}}{n} = \overline{\boldsymbol{Y}}_{k-1} + \overline{\tilde{\boldsymbol{G}}}_{k-1} - \overline{\tilde{\boldsymbol{G}}}_{k-2} = \overline{\boldsymbol{Y}}_{-1} + \sum_{j=-1}^{k-1}(\overline{\tilde{\boldsymbol{G}}}_j - \overline{\tilde{\boldsymbol{G}}}_{j-1}) = \overline{\tilde{\boldsymbol{G}}}_{k-1} \tag{60}$$

and

$$\overline{\boldsymbol{X}}_{k+1} = \mathcal{M}(\boldsymbol{X}_k - \alpha\boldsymbol{Y}_k)\frac{\mathbf{1}}{n} = \overline{\boldsymbol{X}}_k - \alpha\overline{\boldsymbol{Y}}_k. \tag{61}$$

By Taylor Theorem, we obtain

$$\mathbb{E}f\left(\overline{\boldsymbol{X}}_{k+1}\right) = \mathbb{E}f\left(\overline{\boldsymbol{X}}_k - \alpha\overline{\boldsymbol{Y}}_k\right) \tag{62}$$
$$\leq \mathbb{E}f\left(\overline{\boldsymbol{X}}_k\right) - \alpha\mathbb{E}\left\langle\nabla f\left(\overline{\boldsymbol{X}}_k\right), \overline{\boldsymbol{Y}}_k\right\rangle + \frac{\alpha^2 L}{2}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2 \tag{63}$$
$$\overset{(60)}{=} \mathbb{E}f\left(\overline{\boldsymbol{X}}_k\right) - \alpha\mathbb{E}\left\langle\nabla f\left(\overline{\boldsymbol{X}}_k\right), \overline{\boldsymbol{G}}_{k-1}\right\rangle + \frac{\alpha^2 L}{2}\mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}}_{k-1}\right\|^2. \tag{64}$$

For the last term, we have

$$\mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}}_{k-1}\right\|^2 =\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}\right\|^2+\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}-\overline{\tilde{\boldsymbol{G}}}_{k-1}\right\|^2+2\mathbb{E}\left\langle\overline{\boldsymbol{G}}_{k-1},\overline{\boldsymbol{G}}_{k-1}-\overline{\tilde{\boldsymbol{G}}}_{k-1}\right\rangle \tag{65}$$

$$=\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}\right\|^2+\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}-\overline{\tilde{\boldsymbol{G}}}_{k-1}\right\|^2 \tag{66}$$

$$=\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}\right\|^2+\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left\|\boldsymbol{G}_{k-1}\boldsymbol{e}_i-\tilde{\boldsymbol{G}}_{k-1}\boldsymbol{e}_i\right\|^2 \tag{67}$$

$$\leq\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}\right\|^2+\frac{\tilde{\sigma}^2}{n}, \tag{68}$$

where in the second step, we use the fact that the sampling noise is independent of the gradient itself. Putting it back we obtain

$$\mathbb{E}f\left(\overline{\boldsymbol{X}}_{k+1}\right)\leq\mathbb{E}f\left(\overline{\boldsymbol{X}}_k\right)-\alpha\mathbb{E}\left\langle\nabla f\left(\overline{\boldsymbol{X}}_k\right),\overline{\boldsymbol{G}}_{k-1}\right\rangle+\frac{\alpha^2L}{2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}\right\|^2+\frac{\alpha^2\tilde{\sigma}^2L}{2n} \tag{69}$$

$$=\mathbb{E}f\left(\overline{\boldsymbol{X}}_k\right)-\frac{\alpha}{2}\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2-\frac{\alpha-\alpha^2L}{2}\left\|\overline{\boldsymbol{G}}_{k-1}\right\|^2+\frac{\alpha^2\tilde{\sigma}^2L}{2n}+\frac{\alpha}{2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}-\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2, \tag{70}$$

where the last step we use $2\langle a,b\rangle=\|a\|^2+\|b\|^2-\|a-b\|^2$. Expand the last term, we obtain

$$\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}-\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \tag{71}$$

$$\leq2\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k-1}-\overline{\boldsymbol{G}}_{k+1}\right\|^2+2\mathbb{E}\left\|\overline{\boldsymbol{G}}_{k+1}-\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \tag{72}$$

$$=2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_{k,i})-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_{k-2,i})\right\|^2+2\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\boldsymbol{x}_{k,i})-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\overline{\boldsymbol{X}}_k)\right\|^2 \tag{73}$$

$$\leq\frac{2}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(\boldsymbol{x}_{k,i})-\nabla f_i(\boldsymbol{x}_{k-2,i})\right\|^2+\frac{2}{n}\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(\boldsymbol{x}_{k,i})-\nabla f_i(\overline{\boldsymbol{X}}_k)\right\|^2 \tag{74}$$

$$\leq\frac{2L^2}{n}\mathbb{E}\left\|\boldsymbol{X}_k-\boldsymbol{X}_{k-2}\right\|_F^2+\frac{2L^2}{n}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2. \tag{75}$$

Denote $f(\mathbf{0})-f^*\leq\Delta$, we obtain

$$\sum_{k=0}^{K-1}\alpha(1-\alpha L)\left\|\overline{\boldsymbol{G}}_k\right\|^2+\sum_{k=0}^{K-1}\alpha\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \tag{76}$$

$$\leq2\Delta+\frac{\alpha^2\tilde{\sigma}^2LK}{n}+\frac{2\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{2\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\boldsymbol{X}_{k-2}\right\|_F^2 \tag{77}$$

$$\leq2\Delta+\frac{\alpha^2\tilde{\sigma}^2LK}{n}+\frac{16\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{6\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top-\overline{\boldsymbol{X}}_{k-2}\mathbf{1}_n^\top\right\|_F^2, \tag{78}$$

where in the last step we use

$$\frac{2\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\boldsymbol{X}_{k-2}\right\|_F^2 \tag{79}$$

$$\leq\frac{6\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{6\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_{k-2}-\overline{\boldsymbol{X}}_{k-2}\mathbf{1}_n^\top\right\|_F^2+\frac{6\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top-\overline{\boldsymbol{X}}_{k-2}\mathbf{1}_n^\top\right\|_F^2. \tag{80}$$

In addition, for the last term we have

$$\frac{6\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top-\overline{\boldsymbol{X}}_{k-2}\mathbf{1}_n^\top\right\|_F^2=\frac{6\alpha L^2n}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{X}}_k-\overline{\boldsymbol{X}}_{k-2}\right\|^2 \tag{81}$$

$$\overset{(61)}{=}\frac{24\alpha^3L^2n}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}}_k\right\|^2 \tag{82}$$

$$\overset{(65)}{\leq}24\alpha^3L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{24\alpha^3\tilde{\sigma}^2L^2}{n}. \tag{83}$$

Push it back we have

$$\sum_{k=0}^{K-1}\alpha(1-\alpha L-24\alpha^2L^2)\left\|\overline{\boldsymbol{G}}_k\right\|^2+\sum_{k=0}^{K-1}\alpha\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \tag{84}$$

$$\leq 2\Delta+\frac{\alpha^2\tilde{\sigma}^2LK}{n}+\frac{16\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{24\alpha^3\tilde{\sigma}^2L^2}{n}. \tag{85}$$

The rest of the proof is to bound $\frac{16\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2$.

We start from

$$\left\|\boldsymbol{X}_{k+1}-\overline{\boldsymbol{X}}_{k+1}\mathbf{1}_n^\top\right\|_F^2 \tag{86}$$

$$\overset{(61)}{=}\left\|\mathcal{M}(\boldsymbol{X}_k-\alpha\boldsymbol{Y}_k)-(\overline{\boldsymbol{X}}_k-\alpha\overline{\boldsymbol{Y}}_k)\mathbf{1}_n^\top\right\|_F^2 \tag{87}$$

$$=\left\|\mathcal{M}(\boldsymbol{X}_k)-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2-2\alpha\left\langle\mathcal{M}(\boldsymbol{X}_k)-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top,\mathcal{M}(\boldsymbol{Y}_k)-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\rangle+\alpha^2\left\|\mathcal{M}(\boldsymbol{Y}_k)-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2 \tag{88}$$

$$\overset{(58)}{\leq}\rho^2\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{\rho^2(1-\rho^2)}{1+\rho^2}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{\rho^2(1+\rho^2)\alpha^2}{1-\rho^2}\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2 \tag{89}$$

$$+\alpha^2\rho^2\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2 \tag{90}$$

$$=\frac{2\rho^2}{(1+\rho^2)}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{2\rho^2\alpha^2}{1-\rho^2}\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2, \tag{91}$$

where in the third step we use

$$-2\langle\boldsymbol{a},\boldsymbol{b}\rangle\leq\frac{1-\rho^2}{1+\rho^2}\|\boldsymbol{a}\|^2+\frac{1+\rho^2}{1-\rho^2}\|\boldsymbol{b}\|^2. \tag{92}$$

Similarly, for $\boldsymbol{Y}_{k+1}$, we obtain

$$\mathbb{E}\left\|\boldsymbol{Y}_{k+1}-\overline{\boldsymbol{Y}}_{k+1}\mathbf{1}_n^\top\right\|_F^2 \tag{93}$$

$$\overset{(60)}{=}\mathbb{E}\left\|\mathcal{M}(\boldsymbol{Y}_k+\tilde{\boldsymbol{G}}_k-\tilde{\boldsymbol{G}}_{k-1})-(\overline{\boldsymbol{Y}}_k+\overline{\tilde{\boldsymbol{G}}}_k-\overline{\tilde{\boldsymbol{G}}}_{k-1})\mathbf{1}_n^\top\right\|_F^2 \tag{94}$$

$$=\mathbb{E}\left\|\mathcal{M}(\boldsymbol{Y}_k)-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2+\mathbb{E}\left\|\mathcal{M}(\tilde{\boldsymbol{G}}_k-\tilde{\boldsymbol{G}}_{k-1})-(\overline{\tilde{\boldsymbol{G}}}_k-\overline{\tilde{\boldsymbol{G}}}_{k-1})\mathbf{1}_n^\top\right\|_F^2 \tag{95}$$

$$+2\mathbb{E}\left\langle\mathcal{M}(\boldsymbol{Y}_k)-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top,\mathcal{M}(\tilde{\boldsymbol{G}}_k-\tilde{\boldsymbol{G}}_{k-1})-(\overline{\tilde{\boldsymbol{G}}}_k-\overline{\tilde{\boldsymbol{G}}}_{k-1})\mathbf{1}_n^\top\right\rangle \tag{96}$$

$$\overset{(92)(58)}{\leq}\rho^2\mathbb{E}\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2+\rho^2\mathbb{E}\left\|\boldsymbol{G}_k-\boldsymbol{G}_{k-1}-(\overline{\boldsymbol{G}}_k-\overline{\boldsymbol{G}}_{k-1})\mathbf{1}_n^\top\right\|_F^2 \tag{97}$$

$$+\frac{(1-\rho^2)\rho^2}{1+\rho^2}\mathbb{E}\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{(1+\rho^2)\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{G}_k-\boldsymbol{G}_{k-1}-(\overline{\boldsymbol{G}}_k-\overline{\boldsymbol{G}}_{k-1})\mathbf{1}_n^\top\right\|_F^2 \tag{98}$$

$$+2\rho^2\mathbb{E}\|\boldsymbol{G}_k-\tilde{\boldsymbol{G}}_k\|_F^2+2\rho^2\mathbb{E}\|\boldsymbol{G}_{k-1}-\tilde{\boldsymbol{G}}_{k-1}\|_F^2+2\rho^2\mathbb{E}\|\overline{\boldsymbol{G}_k}\mathbf{1}_n^\top-\overline{\tilde{\boldsymbol{G}}}_k\mathbf{1}_n^\top\|_F^2+2\rho^2\mathbb{E}\|\overline{\boldsymbol{G}_{k-1}}\mathbf{1}_n^\top-\overline{\tilde{\boldsymbol{G}}}_{k-1}\mathbf{1}_n^\top\|_F^2 \tag{99}$$

$$\leq\frac{2\rho^2}{1+\rho^2}\mathbb{E}\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{G}_{k+2}-\boldsymbol{G}_{k+1}\right\|_F^2 \tag{100}$$

$$+\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{G}_{k+2}-\boldsymbol{G}_{k+1}-\boldsymbol{G}_k+\boldsymbol{G}_{k-1}\right\|_F^2+8n\rho^2\tilde{\sigma}^2, \tag{101}$$

where in the last step we use $\|I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\| \leq 1$ and $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|$.

For the second term, we have

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{G}_{k+2} - \boldsymbol{G}_{k+1}\right\|_F^2 && (102)\\
=&\sum_{i=1}^n \mathbb{E}\left\|\nabla f(\boldsymbol{x}_{k+1,i}) - \nabla f(\boldsymbol{x}_{k,i})\right\|^2 && (103)\\
\leq& L^2 \sum_{i=1}^n \mathbb{E}\left\|\boldsymbol{x}_{k+1,i} - \boldsymbol{x}_{k,i}\right\|^2 && (104)\\
=& L^2 \mathbb{E}\left\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\right\|_F^2 && (105)\\
\overset{(61)}{=}& L^2 \mathbb{E}\left\|\mathcal{M}(\boldsymbol{X}_k) - \boldsymbol{X}_k - \alpha\mathcal{M}(\boldsymbol{Y}_k)\right\|_F^2 && (106)\\
=& L^2 \mathbb{E}\left\|\mathcal{M}(\boldsymbol{X}_k - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top) - (\boldsymbol{X}_k - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top) - \alpha\mathcal{M}(\boldsymbol{Y}_k)\right\|_F^2 && (107)\\
\leq& 4L^2 \mathbb{E}\left\|\mathcal{M}(\boldsymbol{X}_k) - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 + 4L^2\mathbb{E}\left\|\boldsymbol{X}_k - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 + 4\alpha^2L^2\mathbb{E}\left\|\mathcal{M}(\boldsymbol{Y}_k) - \overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2 && (108)\\
&+ 4\alpha^2 nL^2\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2 && (109)\\
\leq& 4(1+\rho^2)L^2\mathbb{E}\left\|\boldsymbol{X}_k - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 + 4\alpha^2\rho^2L^2\mathbb{E}\left\|\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2 + 4\alpha^2nL^2\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2. && (110)
\end{aligned}$$

Putting it back we obtain

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{Y}_{k+1} - \overline{\boldsymbol{Y}}_{k+1}\mathbf{1}_n^\top\right\|_F^2 && (111)\\
\leq& \left(\frac{2\rho^2}{1+\rho^2} + \frac{16\alpha^2\rho^4L^2}{1-\rho^2}\right)\mathbb{E}\left\|\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2 + \frac{16\rho^2(1+\rho^2)L^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{X}_k - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 + \frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2 && (112)\\
&\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{G}_{k+2} - \boldsymbol{G}_{k+1} - \boldsymbol{G}_k + \boldsymbol{G}_{k-1}\right\|_F^2 + 8n\rho^2\tilde{\sigma}^2. && (113)
\end{aligned}$$

Combining Equation (91) and Equation (100), we have

$$\begin{aligned}
\begin{bmatrix}\mathbb{E}\left\|\boldsymbol{X}_{k+1} - \overline{\boldsymbol{X}}_{k+1}\mathbf{1}_n^\top\right\|_F^2\\ \mathbb{E}\left\|\boldsymbol{Y}_{k+1} - \overline{\boldsymbol{Y}}_{k+1}\mathbf{1}_n^\top\right\|_F^2\end{bmatrix} \preceq& \begin{bmatrix}\boldsymbol{P}_{11} & \boldsymbol{P}_{12}\\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22}\end{bmatrix}\begin{bmatrix}\mathbb{E}\left\|\boldsymbol{X}_k - \overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2\\ \mathbb{E}\left\|\boldsymbol{Y}_k - \overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2\end{bmatrix} && (114)\\
&+ \begin{bmatrix}0\\ \frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2 + \frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2 + 8n\rho^2\tilde{\sigma}^2\end{bmatrix}, && (115)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{P}_{11} =& \frac{2\rho^2}{(1+\rho^2)} && (116)\\
\boldsymbol{P}_{12} =& \frac{2\rho^2\alpha^2}{1-\rho^2} && (117)\\
\boldsymbol{P}_{21} =& \frac{16\rho^2(1+\rho^2)L^2}{1-\rho^2} && (118)\\
\boldsymbol{P}_{22} =& \frac{2\rho^2}{1+\rho^2} + \frac{16\alpha^2\rho^4L^2}{1-\rho^2} && (119)\\
\boldsymbol{U}_k =& \boldsymbol{G}_{k+2} - \boldsymbol{G}_{k+1} - \boldsymbol{G}_k + \boldsymbol{G}_{k-1}. && (120)
\end{aligned}$$

For simplicity, define

$$\boldsymbol{z}_k = \begin{bmatrix}\mathbb{E}\left\|\boldsymbol{X}_{k+1} - \overline{\boldsymbol{X}}_{k+1}\mathbf{1}_n^\top\right\|_F^2\\ \mathbb{E}\left\|\boldsymbol{Y}_{k+1} - \overline{\boldsymbol{Y}}_{k+1}\mathbf{1}_n^\top\right\|_F^2\end{bmatrix} \tag{121}$$

$$\boldsymbol{P}=\begin{bmatrix}\boldsymbol{P}_{11} & \boldsymbol{P}_{12}\\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22}\end{bmatrix} \tag{122}$$

$$\boldsymbol{u}_k=\begin{bmatrix}0\\ \frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2+8n\rho^2\tilde{\sigma}^2,\end{bmatrix} \tag{123}$$

then we can write this linear system as

$$\boldsymbol{z}_k\preceq\boldsymbol{P}\boldsymbol{z}_{k-1}+\boldsymbol{u}_{k-1}\preceq\boldsymbol{P}^k\boldsymbol{z}_0+\sum_{t=0}^{k-1}\boldsymbol{P}^{k-t}\boldsymbol{u}_t, \tag{124}$$

for simplicity.

Let $\lambda_1(\boldsymbol{P}),\lambda_2(\boldsymbol{P})$ denote the two eigenvalues of $\boldsymbol{P}$ (without the loss of generality, we denote $\lambda_1(\boldsymbol{P})<\lambda_2(\boldsymbol{P})$), define

$$\Psi=\sqrt{(\boldsymbol{P}_{11}-\boldsymbol{P}_{22})^2+4\boldsymbol{P}_{12}\boldsymbol{P}_{21}}, \tag{125}$$

then with eigendecomposition, we obtain

$$\lambda_1(\boldsymbol{P})=\frac{\boldsymbol{P}_{11}+\boldsymbol{P}_{22}-\Psi}{2} \tag{126}$$

$$\lambda_2(\boldsymbol{P})=\frac{\boldsymbol{P}_{11}+\boldsymbol{P}_{22}+\Psi}{2}=\frac{2\rho^2}{1+\rho^2}+\frac{8\alpha^2\rho^4L^2}{1-\rho^2}+\frac{16\alpha\rho^2L\sqrt{\alpha^2\rho^4L^2+(1+\rho^2)}}{1-\rho^2} \tag{127}$$

$$\boldsymbol{P}^k\preceq\begin{bmatrix}\frac{\lambda_1^k(\boldsymbol{P})+\lambda_2^k(\boldsymbol{P})}{2}+\frac{(\boldsymbol{P}_{11}-\boldsymbol{P}_{22})(\lambda_2^k(\boldsymbol{P})-\lambda_1^k(\boldsymbol{P}))}{2\Psi} & \frac{\boldsymbol{P}_{12}}{\Psi}(\lambda_2^k(\boldsymbol{P})-\lambda_1^k(\boldsymbol{P}))\\ \frac{\boldsymbol{P}_{21}}{\Psi}(\lambda_2^k(\boldsymbol{P})-\lambda_1^k(\boldsymbol{P})) & \frac{\lambda_1^k(\boldsymbol{P})+\lambda_2^k(\boldsymbol{P})}{2}+\frac{(\boldsymbol{P}_{11}-\boldsymbol{P}_{22})(\lambda_1^k(\boldsymbol{P})-\lambda_2^k(\boldsymbol{P}))}{2\Psi}\end{bmatrix}, \tag{128}$$

when the step size is small enough such that

$$\alpha L<\frac{(1-\rho)^2}{32}, \tag{129}$$

it can be verified that $\lambda_2(\boldsymbol{P})\leq\frac{\sqrt{\rho}+\rho}{1+\rho}$, and then we can compute the $\mathbb{E}\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\|^2$ and $\mathbb{E}\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\|^2$. We use $\boldsymbol{X}[1:]$ to denote the first row of matrix $\boldsymbol{X}$. First for $\boldsymbol{X}_k$, we obtain:

$$\boldsymbol{P}^k\boldsymbol{z}_0[1:]\leq\boldsymbol{P}_{12}k\lambda_2^{k-1}(\boldsymbol{P})\mathbb{E}\|\boldsymbol{Y}_0-\overline{\boldsymbol{Y}}_0\mathbf{1}_n^\top\|_F^2=\frac{2\rho^2\alpha^2k}{1-\rho^2}\lambda_2^{k-1}(\boldsymbol{P})\mathbb{E}\|\boldsymbol{Y}_0-\overline{\boldsymbol{Y}}_0\mathbf{1}_n^\top\|_F^2 \tag{130}$$

where we use the property that $\lambda_2^k(\boldsymbol{P})-\lambda_1^k(\boldsymbol{P})=(\lambda_2(\boldsymbol{P})-\lambda_1(\boldsymbol{P}))\sum_{l=0}^{k-1}\lambda_2(\boldsymbol{P})^l\lambda_1(\boldsymbol{P})^{k-1-l}=\Psi k\lambda_2^{k-1}(\boldsymbol{P})$ and, similarly

$$\boldsymbol{P}^{k-t}\boldsymbol{u}_t[1:] \tag{131}$$

$$\leq\frac{2\rho^2\alpha^2(k-t)}{1-\rho^2}\lambda_2^{k-t-1}(\boldsymbol{P})\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_t\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_t\right\|^2+8n\rho^2\tilde{\sigma}^2\right) \tag{132}$$

$$=\frac{2\rho^2\alpha^2(k-t)}{1-\rho^2}\lambda_2^{k-t-1}(\boldsymbol{P})\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_t\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\tilde{\boldsymbol{G}}}_{t-1}\right\|^2+8n\rho^2\tilde{\sigma}^2\right) \tag{133}$$

$$\overset{(65)}{\leq}\frac{2\rho^2\alpha^2(k-t)}{1-\rho^2}\lambda_2^{k-t-1}(\boldsymbol{P})\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_t\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_{t-1}\right\|^2+\frac{16\alpha^2\rho^2\tilde{\sigma}^2L^2}{1-\rho^2}+8n\rho^2\tilde{\sigma}^2\right), \tag{134}$$

then we obtain

$$\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 \tag{135}$$

$$\leq\frac{2\rho^2\alpha^2k}{1-\rho^2}\lambda_2^{k-1}(\boldsymbol{P})\mathbb{E}\|\boldsymbol{Y}_0-\overline{\boldsymbol{Y}}_0\mathbf{1}_n^\top\|_F^2 \tag{136}$$

$$+\sum_{t=0}^{k-1}\frac{2\rho^2\alpha^2(k-t)}{1-\rho^2}\lambda_2^{k-t-1}(\boldsymbol{P})\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_t\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_{t-1}\right\|^2+\frac{16\alpha^2\rho^2\tilde{\sigma}^2L^2}{1-\rho^2}+8n\rho^2\tilde{\sigma}^2\right). \tag{137}$$

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

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 \tag{138}$$

$$\leq\frac{2\rho^2\alpha^2}{(1-\rho^2)(1-\lambda_2(\boldsymbol{P}))^2}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{Y}_0-\overline{\boldsymbol{Y}}_0\mathbf{1}_n^\top\|_F^2 \tag{139}$$

$$+\frac{2\rho^2\alpha^2}{(1-\rho^2)(1-\lambda_2(\boldsymbol{P}))^2}\sum_{k=0}^{K-1}\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{16\alpha^2\rho^2\tilde{\sigma}^2L^2}{1-\rho^2}+8n\rho^2\tilde{\sigma}^2\right) \tag{140}$$

$$\leq\frac{2\rho^2\alpha^2(1+\rho)nK\varsigma_0^2}{(1-\rho)(1-\sqrt{\rho})^2}+\frac{32\rho^4\alpha^4nL^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{8\rho^4\alpha^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{U}_k\|_F^2 \tag{141}$$

$$+\frac{32\rho^4\alpha^4\tilde{\sigma}^2L^2K}{(1-\rho)^2(1-\sqrt{\rho})^2}+\frac{8\rho^4\alpha^2n\tilde{\sigma}^2(1+\rho)K}{(1-\rho)(1-\sqrt{\rho})^2}] \tag{142}$$

$$\leq\frac{2\rho^2\alpha^2(1+\rho)nK\varsigma_0^2}{(1-\rho)(1-\sqrt{\rho})^2}+\frac{32\rho^4\alpha^4nL^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{8\rho^4\alpha^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{U}_k\|_F^2 \tag{143}$$

$$+\frac{16\rho^4\alpha^2n\tilde{\sigma}^2(1+\rho)K}{(1-\rho)(1-\sqrt{\rho})^2}, \tag{144}$$

where in the second step we used $\frac{1}{1-\lambda_2(\boldsymbol{P})}<\frac{1+\rho}{1-\sqrt{\rho}}$ since $\lambda_2(\boldsymbol{P})\leq\frac{\sqrt{\rho}+\rho}{1+\rho}$. And the third step holds due to Equation (129). We proceed to analyze the case in $\boldsymbol{Y}_k$: we first have

$$[\boldsymbol{P}^k]_{22}=\frac{\lambda_1^k(\boldsymbol{P})+\lambda_2^k(\boldsymbol{P})}{2}+\frac{(\boldsymbol{P}_{11}-\boldsymbol{P}_{22})(\lambda_1^k(\boldsymbol{P})-\lambda_2^k(\boldsymbol{P}))}{2\Psi} \tag{145}$$

$$\leq\lambda_2^k(\boldsymbol{P})+\frac{8\alpha^2\rho^4L^2k\lambda_2^{k-1}(\boldsymbol{P})}{1-\rho^2}, \tag{146}$$

then we can have

$$\boldsymbol{P}^k\boldsymbol{z}_0[2:]\leq\left(\lambda_2^k(\boldsymbol{P})+\frac{8\alpha^2\rho^4L^2k\lambda_2^{k-1}(\boldsymbol{P})}{1-\rho^2}\right)\mathbb{E}\|\boldsymbol{Y}_0-\overline{\boldsymbol{Y}}_0\mathbf{1}_n^\top\|_F^2, \tag{147}$$

and

$$\boldsymbol{P}^{k-t}\boldsymbol{u}_t[2:] \tag{148}$$

$$\leq\left(\lambda_2^{k-t}(\boldsymbol{P})+\frac{8\alpha^2\rho^4L^2(k-t)\lambda_2^{k-t-1}(\boldsymbol{P})}{1-\rho^2}\right)\cdot\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_t\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_t\right\|^2+8n\rho^2\tilde{\sigma}^2\right) \tag{149}$$

$$\leq\left(\lambda_2^{k-t}(\boldsymbol{P})+\frac{8\alpha^2\rho^4L^2(k-t)\lambda_2^{k-t-1}(\boldsymbol{P})}{1-\rho^2}\right)\cdot\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_t\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_{t-1}\right\|^2+\frac{16\alpha^2\rho^2\tilde{\sigma}^2L^2}{1-\rho^2}+8n\rho^2\tilde{\sigma}^2\right). \tag{150}$$

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

$$\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\|^2 \tag{151}$$

$$\leq\frac{(1+\rho)nK\varsigma_0^2}{1-\sqrt{\rho}}+\frac{8\alpha^2\rho^4(1+\rho)L^2nK\varsigma_0^2}{(1-\rho)(1-\sqrt{\rho})^2} \tag{152}$$

$$+\left(\frac{1+\rho}{1-\sqrt{\rho}}+\frac{8\alpha^2\rho^4(1+\rho)L^2}{(1-\rho)(1-\sqrt{\rho})^2}\right)\sum_{k=0}^{K-1}\left(\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2+\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{16\alpha^2\rho^2\tilde{\sigma}^2L^2}{1-\rho^2}+8n\rho^2\tilde{\sigma}^2\right). \tag{153}$$

We next solve $\|\boldsymbol{U}_k\|_F$, from the definition of $\boldsymbol{U}_k$ we obtain that

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2=\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{G}_{k+2}-\boldsymbol{G}_{k+1}-\boldsymbol{G}_k+\boldsymbol{G}_{k-1}\right\|_F^2 \tag{154}$$

$$\leq 2\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{G}_{k+2}-\boldsymbol{G}_{k+1}\right\|_F^2+2\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{G}_k-\boldsymbol{G}_{k-1}\right\|_F^2 \tag{155}$$

$$\leq 4\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{G}_{k+2}-\boldsymbol{G}_{k+1}\right\|_F^2 \tag{156}$$

$$=4\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\boldsymbol{x}_{k+1,i})-\nabla f(\boldsymbol{x}_{k,i})\right\|^2 \tag{157}$$

$$\leq 4L^2\sum_{k=0}^{K-1}\sum_{i=1}^{n}\mathbb{E}\left\|\boldsymbol{x}_{k+1,i}-\boldsymbol{x}_{k,i}\right\|^2 \tag{158}$$

$$=4L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_{k+1}-\boldsymbol{X}_k\right\|_F^2. \tag{159}$$

Fit in the derivation from Equation (110) we obtain

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2 \tag{160}$$

$$\leq 4L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_{k+1}-\boldsymbol{X}_k\right\|_F^2 \tag{161}$$

$$\leq 16(1+\rho^2)L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+16\alpha^2\rho^2L^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{Y}_k-\overline{\boldsymbol{Y}}_k\mathbf{1}_n^\top\right\|_F^2+16\alpha^2nL^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2 \tag{162}$$

$$\leq \frac{32\rho^2\alpha^2(1+\rho)^2nK\varsigma_0^2L^2}{(1-\rho)(1-\sqrt{\rho})^2}+\frac{512\rho^4(1+\rho)\alpha^4nL^4}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{256\rho^4(1+\rho)\alpha^2L^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{U}_k\|_F^2 \tag{163}$$

$$+\frac{256\rho^4\alpha^2n\tilde{\sigma}^2(1+\rho)^2KL^2}{(1-\rho)(1-\sqrt{\rho})^2} \tag{164}$$

$$+\frac{16\alpha^2\rho^2(1+\rho)nK\varsigma_0^2L^2}{1-\sqrt{\rho}}+\frac{128\alpha^4\rho^6(1+\rho)L^4nK\varsigma_0^2}{(1-\rho)(1-\sqrt{\rho})^2} \tag{165}$$

$$+\left(\frac{16\alpha^2\rho^2(1+\rho)L^2}{1-\sqrt{\rho}}+\frac{128\alpha^4\rho^6(1+\rho)L^4}{(1-\rho)(1-\sqrt{\rho})^2}\right)\sum_{k=0}^{K-1}\frac{4\rho^2}{1-\rho^2}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2 \tag{166}$$

$$+\left(\frac{16\alpha^2\rho^2(1+\rho)L^2}{1-\sqrt{\rho}}+\frac{128\alpha^4\rho^6(1+\rho)L^4}{(1-\rho)(1-\sqrt{\rho})^2}\right)\sum_{k=0}^{K-1}\left(\frac{16\alpha^2\rho^2nL^2}{1-\rho^2}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{16\alpha^2\rho^2\tilde{\sigma}^2L^2}{1-\rho^2}+8n\rho^2\tilde{\sigma}^2\right) \tag{167}$$

$$+16\alpha^2nL^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{Y}}_k\right\|^2 \tag{168}$$

$$\leq \frac{64\rho^2\alpha^2(1+\rho)^2nK\varsigma_0^2L^2}{(1-\rho)(1-\sqrt{\rho})^2}+\frac{512\rho^4(1+\rho)\alpha^2L^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\|\boldsymbol{U}_k\|_F^2+\frac{512\rho^4\alpha^2n\tilde{\sigma}^2(1+\rho)^2KL^2}{(1-\rho)(1-\sqrt{\rho})^2} \tag{169}$$

$$+32\alpha^2nL^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2, \tag{170}$$

where in the third step we use the derivation from Equation (138) and (151), in the fourth step we repeatedly use Equa-

tion (129) and Equation (65), solve it we obtain

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{U}_k\right\|_F^2 \leq \frac{128\rho^2\alpha^2(1+\rho)^2nK\varsigma_0^2L^2}{(1-\rho)(1-\sqrt{\rho})^2}+\frac{1024\rho^4\alpha^2n\tilde{\sigma}^2(1+\rho)^2KL^2}{(1-\rho)(1-\sqrt{\rho})^2}+64\alpha^2nL^2\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2, \tag{171}$$

where again we use Equation (129), combine it with Equation (138) we obtain

$$\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2 \tag{172}$$

$$\leq \frac{4\rho^2\alpha^2(1+\rho)nK\varsigma_0^2}{(1-\rho)(1-\sqrt{\rho})^2}+\frac{544\rho^4\alpha^4nL^2}{(1-\rho)^2(1-\sqrt{\rho})^2}\sum_{k=0}^{K-1}\mathbb{E}\left\|\overline{\boldsymbol{G}}_k\right\|^2+\frac{32\rho^4\alpha^2n\tilde{\sigma}^2(1+\rho)K}{(1-\rho)(1-\sqrt{\rho})^2}, \tag{173}$$

where we use Equation (129).

Recall from Equation (84) that

$$\sum_{k=0}^{K-1}\alpha(1-\alpha L-24\alpha^2L^2)\left\|\overline{\boldsymbol{G}}_k\right\|^2+\sum_{k=0}^{K-1}\alpha\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \tag{174}$$

$$\leq 2\Delta+\frac{\alpha^2\tilde{\sigma}^2LK}{n}+\frac{16\alpha L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left\|\boldsymbol{X}_k-\overline{\boldsymbol{X}}_k\mathbf{1}_n^\top\right\|_F^2+\frac{24\alpha^3\tilde{\sigma}^2L^2}{n}. \tag{175}$$

Combine Equation (129) and (173), we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \leq O\left(\frac{\Delta}{\alpha K}+\frac{\alpha\tilde{\sigma}^2L}{n}+\frac{\rho^2\alpha^2L^2\varsigma_0^2}{(1-\rho)^3}+\frac{\rho^4\alpha^2\tilde{\sigma}^2L^2}{(1-\rho)^3}+\frac{\alpha^2\tilde{\sigma}^2L^2}{nK}\right), \tag{176}$$

where we omit the numerical constants. Set

$$\alpha=\frac{1}{\tilde{\sigma}\sqrt{KL/n\Delta}+\frac{\rho^{\frac{2}{3}}L^{\frac{2}{3}}\varsigma_0^{\frac{2}{3}}K^{\frac{1}{3}}}{\Delta^{\frac{1}{3}}(1-\rho)}+\frac{32L}{(1-\rho)^2}}, \tag{177}$$

we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \leq O\left(\frac{\sqrt{\Delta L}\tilde{\sigma}}{\sqrt{nK}}+\frac{(\rho\Delta L\varsigma_0)^{\frac{2}{3}}}{(1-\rho)K^{\frac{2}{3}}}+\frac{\rho^2n\Delta L}{(1-\rho)^3K}+\frac{\Delta L}{(1-\rho)^2K}\right). \tag{178}$$

Fit in $T=KR$ and $\tilde{\sigma}^2=\sigma^2/BR$, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \leq O\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}+\frac{(\rho\Delta L\varsigma_0R)^{\frac{2}{3}}}{(1-\rho)T^{\frac{2}{3}}}+\frac{\rho^2nR\Delta L}{(1-\rho)^3T}+\frac{R\Delta L}{(1-\rho)^2T}\right), \tag{179}$$

set

$$R=\frac{1}{\sqrt{1-\lambda_2(\boldsymbol{W})}}\max\left(\frac{1}{2}\log(n),\frac{1}{2}\log\left(\frac{\varsigma_0^2T}{\Delta L}\right)\right),$$

we first have $\rho\leq 1/\sqrt{2}$ since

$$R\geq\frac{\log(n)}{2\sqrt{1-\lambda_2(\boldsymbol{W})}}\geq\frac{-\log(n)}{2\log(1-\sqrt{1-\lambda_2(\boldsymbol{W})})}\Rightarrow\left(1-\sqrt{1-\lambda_2(\boldsymbol{W})}\right)^R\leq\frac{1}{\sqrt{n}}\leq\frac{1}{\sqrt{2}}, \tag{180}$$

this implies

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \leq O\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}+\frac{(\rho\Delta L\varsigma_0R)^{\frac{2}{3}}}{T^{\frac{2}{3}}}+\frac{\rho^2nR\Delta L}{T}+\frac{R\Delta L}{T}\right) \tag{181}$$

$$\leq O\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}+\frac{(\rho\varsigma_0\sqrt{T}R/\sqrt{\Delta L})^{\frac{2}{3}}\Delta L}{T}+\frac{\rho^2 nR\Delta L}{T}+\frac{R\Delta L}{T}\right), \tag{182}$$

with the assignment of R, $\rho^2 n<1$ and $\rho\varsigma_0\sqrt{T}/\sqrt{\Delta L}<1$, so since it also holds that $R\geq 1$ (and so $R^{2/3}\leq R$),

$$\min_t \|\nabla f\left(\overline{\boldsymbol{X}}_t\right)\|^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla f\left(\overline{\boldsymbol{X}}_k\right)\right\|^2 \tag{183}$$

$$\leq O\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}+\frac{(R)^{\frac{2}{3}}\Delta L}{T}+\frac{R\Delta L}{T}+\frac{R\Delta L}{T}\right) \tag{184}$$

$$\leq O\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}+\frac{R\Delta L}{T}\right) \tag{185}$$

$$= O\left(\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}+\frac{\Delta L}{T\sqrt{1-\lambda_2(\boldsymbol{W})}}\cdot\max\left(\log(n),\log\left(\frac{\varsigma_0^2 T}{\Delta L}\right)\right)\right), \tag{186}$$

when

$$T\leq O\left(\frac{\Delta L\sigma^2}{nB\epsilon^4}\right), \tag{187}$$

we have

$$\frac{\sqrt{\Delta L}\sigma}{\sqrt{nBT}}\leq O(\epsilon^2). \tag{188}$$

On the other hand, when

$$T\leq O\left(\max\left(\frac{\log(n)\Delta L}{\epsilon^2\sqrt{1-\lambda_2(\boldsymbol{W})}},\frac{\Delta L}{\epsilon^2\sqrt{1-\lambda_2(\boldsymbol{W})}}\log\left(\frac{\varsigma_0^2}{\epsilon^2\Delta L}\right)\right)\right), \tag{189}$$

we have

$$\frac{\Delta L}{T\sqrt{1-\lambda_2(\boldsymbol{W})}}\cdot\max\left(\log(n),\log\left(\frac{\varsigma_0^2 T}{\Delta L}\right)\right)\leq O(\epsilon^2), \tag{190}$$

to see this, note that

$$\frac{\Delta L}{T\sqrt{1-\lambda_2(\boldsymbol{W})}}\log\left(\frac{\varsigma_0^2 T}{\Delta L}\right)=\epsilon^2\frac{\log\left(\frac{\varsigma_0^2}{\epsilon^2\Delta L}\log\left(\frac{\varsigma_0^2}{\epsilon^2\Delta L}\right)\right)}{\log\left(\frac{\varsigma_0^2}{\epsilon^2\Delta L}\right)}\leq O(\epsilon^2). \tag{191}$$

Finally, we can obtain the upper bound

$$T\leq O\left(\frac{\Delta L\sigma^2}{nB\epsilon^4}+\max\left(\frac{\log(n)\Delta L}{\epsilon^2\sqrt{1-\lambda_2(\boldsymbol{W})}},\frac{\Delta L}{\epsilon^2\sqrt{1-\lambda_2(\boldsymbol{W})}}\log\left(\frac{\varsigma_0^2}{\epsilon^2\Delta L}\right)\right)\right) \tag{192}$$

$$= O\left(\frac{\Delta L\sigma^2}{nB\epsilon^4}+\frac{\Delta L}{\epsilon^2\sqrt{1-\lambda_2(\boldsymbol{W})}}\log\left(n+\frac{\varsigma_0 n}{\epsilon\sqrt{\Delta L}}\right)\right), \tag{193}$$

as desired. □

C. Details to footnotes

C.1. Asynchronous Algorithm (Footnote 2)

In the full paper, we focus on the synchronous algorithms, i.e., we assume the existence of a synchronization process among workers between two adjacent iterations. We now extend our formulation to asynchronous algorithms. Since workers now

update and communicate asynchronously, we define any gradient update that took place on a randomly chosen worker as one iteration. This randomness depends on system implementation, stochastic events, etc. This is a commonly adopted definition in the analysis of (decentralized) asynchronous algorithms (Lian et al., 2017b). To obtain a lower bound in such case, consider the two settings as shown in the proof of Theorem 1. In setting 1, it can be easily verified that the lower bound for sample complexity is

$$\Omega\left(\frac{\Delta L\sigma^2}{B\epsilon^4}\right). \tag{194}$$

This holds because in the extreme case, only one worker is making contributions to the optimization. And since we have not made any assumption on how workers are sampled to conduct the next iteration, this is a valid bound for arbitrary distribution. On the other hand, considering setting 2, the lower bound is still $\Omega(T_0 D)$ where $T_0 = \Omega(\Delta L\epsilon^{-2})$ is the lower bound in the sequential case, since the systems need at least $\Omega(D)$ iterations for the workers in I_0 and I_2 to contact. The lower bound for communication complexity is then

$$\Omega\left(\frac{\Delta LD}{\epsilon^2}\right). \tag{195}$$

Combining them together, we can get the final lower bound as:

$$\Omega\left(\frac{\Delta L\sigma^2}{B\epsilon^4} + \frac{\Delta LD}{\epsilon^2}\right). \tag{196}$$

Note that this bound holds with probability 1. It is possible to propose finer-grained assumption on how workers are chosen (e.g. uniformly random) and use concentration inequalities (e.g. Hoeffding's inequality) to get tighter bounds, we leave this as future work.

C.2. Relax zero-respecting assumption (Footnote 3)

To relax the zero-respecting assumption, we can use the technique proposed by (Carmon et al., 2019) (See their proofs to Proposition 1 and 2). The basic idea is that to adversarially construct the loss function and rotate the non-zero coordinates in t-th iterations, such that when the algorithm operates on the rotated function, the first t iterations match with that of the old function. However, the new rotated function is still zero-respecting to the algorithm after t-th iteration so is generally hard to optimize. The details can be found in (Carmon et al., 2019).

C.3. Specific algorithm for Average Consensus (Footnote 8)

Many algorithms have been proposed on solving the Average Consensus problem, readers can find details in many previous works on graph theory such as (Georgopoulos, 2011; Hendrickx et al., 2014; Ko, 2010). A straightforward algorithm is the Minimum Spanning Tree, that is, we first generate a spanning tree of the graph, and then the workers send and receive message using propagation on the tree. Specifically, starting from the leaves, all the children nodes of the tree send its accumulated value to the parents and the root compute the averaged value after gathering the information from the graph. And then reversely, the parent nodes send the value back to the child nodes and eventually all the nodes will get the averaged value. This algorithm is also known as the `GATHER-PROPAGATE` algorithm as discussed in (Ko, 2010), section 3. We include the detailed pseudo-code[11] in Algorithm 4.

[11] This code is proposed by Ko (2010), we do not intend to take credit for this.

Algorithm 4 GATHER-PROPAGATE (Spanning Tree) for a single coordinate

Require: communication graph G, a single coordinate on workers (all the coordinates follow the same instructions) to be communicated $\boldsymbol{X} \in \mathbb{R}^n$.

$\boldsymbol{d} \leftarrow$ vector of 1's indexed by $V(G)$ (vertices set of graph G).
$\mathcal{I} \leftarrow$ a spnning tree of G with root r arbitrarily picked.
for $v \in V(\mathcal{I})$ **do**
  $l_v \leftarrow \bar{D}(r, v)$ (the distance between r and v)
end for
for $\alpha = \max_v l_v, \cdots, 1$ **do**
  for v with $l_v = \alpha$ **do**
    v gives all its value onto its parents u:

$$\begin{bmatrix} \boldsymbol{X}_u \\ \boldsymbol{X}_v \end{bmatrix} \leftarrow \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_u \\ \boldsymbol{X}_v \end{bmatrix}$$

    $\boldsymbol{d}_u \leftarrow \boldsymbol{d}_u + \boldsymbol{d}_v$
  end for
end for
for $\alpha = 0, \cdots, \max_v l_v - 1$ **do**
  for u with $l_u = \alpha$ **do**
    $\{v_1, \cdots, v_\beta\} \leftarrow$ set of children of u
    re-distribute the results:

$$\begin{bmatrix} \boldsymbol{X}_u \\ \boldsymbol{X}_{v_1} \\ \vdots \\ \boldsymbol{X}_{v_\beta} \end{bmatrix} \leftarrow \frac{1}{\boldsymbol{d}_u} \begin{bmatrix} \boldsymbol{d}_u - \boldsymbol{d}_u - \cdots - \boldsymbol{d}_{v_\beta} \\ \boldsymbol{d}_{v_1} \\ \vdots \\ \boldsymbol{d}_{v_\beta} \end{bmatrix} \boldsymbol{X}_u$$

  end for
end for
return $X \frac{\mathbf{1}\mathbf{1}^\top}{n}$