DELAY-TOLERANT EFFICIENT DISTRIBUTED TRAINING WITH LOCAL SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

The communication overheads have been the notorious bottleneck for scaling up the distributed training of deep neural networks using more workers. Meanwhile, delay-sensitive applications widely exist in the real world. Thus, it is crucial to develop a *delay-tolerant* AND *communication-efficient* training method for the distributed learning scenario with both communication delay and bandwidth sensitive constraints. However, no existing research work comprehensively solves this training problem under both constraints. To address this important problem, we propose a novel training framework OLCO_3 with staleness and compression compensation for the common compression schemes and also the more complex vector quantization schemes. The proposed method is compatible with local computation to further reduce the communication rounds via periodic averaging. To validate the delay and compression tolerance of our new method, the comprehensive theoretical analysis of OLCO_3 is conducted to show the same sub-linear convergence rate as stochastic gradient descent (SGD). Extensive experiments on deep learning tasks verify the effectiveness of our method and superiority over its counterparts.

1 INTRODUCTION

Synchronous distributed training is currently the workhorse algorithm to accelerate the large-scale deep learning tasks by splitting a very large batch-size of data into much smaller data batches and assigning them to a large number of workers (*i.e.*, GPUs in the context of deep learning) (Goyal et al., 2017; You et al., 2017; Huo et al., 2020). However, the communication overheads cause significant inefficiency for large deep neural networks (DNNs) when scaling up the training with a large number of workers. In general, the communication bottleneck comes in two forms: 1) *high communication delay* due to the unstable network latency; and 2) *large communication budget* with a limited network bandwidth, which grows with the size of DNNs. In real-world applications, the workers are often faced with more unstable links than the data center environment, *e.g.* the workers can be geo-distributed and placed under different networks (Ethernet, cellular networks, Wi-Fi, etc.) in federated learning (Konečnỳ et al., 2016).

Existing works to address the communication inefficiency of synchronous SGD can be roughly classified into three categories: 1) *pipelining* (Pipe-SGD (Li et al., 2018)); 2) *gradient compression* (Strom, 2015; Aji & Heafield, 2017; Stich et al., 2018; Alistarh et al., 2018; Wangni et al., 2018; Yu et al., 2018; Vogels et al., 2019); and 3) *periodic averaging* (also known as Local SGD) (Stich, 2019; Lin et al., 2018a). In pipelining, current computation is overlapped with current communication to remove the communication barrier by using stale information that has been outdated for s *rounds of communication*. Therefore, they can run simultaneously to achieve *delay tolerance*. Gradient compression reduces the size of data transferred in each SGD iteration with a compressor $\mathcal{C}(\cdot)$. Representative methods include scalar quantization (Alistarh et al., 2017; Wen et al., 2017; Bernstein et al., 2018), gradient sparsification (Strom, 2015; Aji & Heafield, 2017; Stich et al., 2018; Alistarh et al., 2018; Wangni et al., 2018), and vector quantization (Yu et al., 2018; Vogels et al., 2019). Periodic averaging reduces the communication frequency by synchronizing the workers every p (larger than 1) SGD iterations. It is also effective for federated learning (McMahan et al., 2017). In summary, pipelining addresses the high communication delay issue, while gradient compression and periodic averaging address the large communication budget issue. However, all existing methods **fail to address both**.

Table 1: Comparison of communication-efficient methods for distributed DNN training. Period $p \in \mathbb{N}_+$ is the communication interval for periodic averaging and staleness $s \in \mathbb{N}+$ is the number of communication rounds that the information used in computation has been outdated for. Delay tolerance $\mathcal{T} = sp$.

Methods	**Overlap Comput. with Comm.**	**Comm. Compression**	p **Iterations per Comm. Round**	**Staleness** s **Supported**
Gradient Compression	$\times$	$\surd$	$= 1$	$= 0$
Periodic Averaging (Local SGD)	$\times$	$\times$	≥ 1	$= 0$
Pipelining (Pipe-SGD)	$\surd$	$\times$	$= 1$	≥ 1
CoCoD-SGD	$\surd$	$\times$	≥ 1	$= 1$
OverlapLocalSGD	$\surd$	$\times$	≥ 1	$= 1$
OLCO_3 (Ours)	$\surd$	$\surd$	≥ 1	≥ 1

In this paper, we propose a novel framework Overlap Local Computation with Compressed Communication (*i.e.*, OLCO_3) to make distributed training both *delay-tolerant* AND *communication efficient* by improving the combination of the three communication-efficient techniques. In Table 1, we compare OLCO_3 with the aforementioned techniques and succeeding delay-tolerant methods CoCoD-SGD (Shen et al., 2019) and OverlapLocalSGD (Wang et al., 2020). In the conext of periodic averaging, we still define staleness s as *the number of communication rounds* that the information used in computation has been outdated for. CoCoD-SGD and OverlapLocalSGD combines pipelining and periodic averaging by using stale information from last communication round ($s = 1, p \geq 1$), while our OLCO_3 supports all features and larger staleness ($s \geq 1, p \geq 1$). Assume that the computation time of one SGD iteration is T_{comput}, then one method can pipeline the communication and the computation when the communication delay time is less than $sp \cdot T_{comput}$. We consequently define the delay tolerance $\mathcal{T} = sp$. We summarize our contributions as follows.

- We propose OLCO_3, which achieves extreme communication efficiency by addressing both high communication delay and large communication budget issues.
- We design novel staleness compensation and compression compensation techniques to make OLCO_3 robust. Convergence analysis shows that OLCO_3 achieves the same rate as SGD.
- Extensive experimental results of deep learning tasks show that OLCO_3 significantly outperforms existing delay-tolerant methods in communication efficiency and model performance.

2 Backgrounds & Related Works

SGD and Pipelining. Without loss of generality, suppose each worker $k \in [K]$ stores the same number of training data such that the global loss function $f(\cdot) = \frac{1}{K}\sum_{k=1}^{K} f_k(\cdot)$, where $f_k(\cdot)$ is the local loss function at worker k. The training is to $\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. At iteration t, vanilla synchronous SGD updates the model $\mathbf{x}_t \in \mathbb{R}^d$ with learning rate η_t via $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta_t}{K}\sum_{k=1}^{K} \nabla F_k(\mathbf{x}_t; \xi_t^{(k)})$, where $\xi_t^{(k)}$ is the stochastic sampling variable and $\nabla F_k(\mathbf{x}_t; \xi_t^{(k)})$ is the corresponding stochastic gradient at worker k. Throughout this paper, we assume that the stochastic gradient is an unbiased estimator: $\mathbb{E}_{\xi_t^{(k)}} \nabla F_k(\mathbf{x}_t; \xi_t^{(k)}) = \nabla f_k(\mathbf{x}_t)$.

Pipe-SGD (Li et al., 2018) pipelines and parallelizes the communication and computation of SGD. At iteration t, worker k computes stochastic gradient $\nabla F_k(\mathbf{x}_t; \xi_t^{(k)})$ with current model and communicates to get the averaged stochastic gradient $\frac{1}{K}\sum_{k=1}^{K} \nabla F_k(\mathbf{x}_t; \xi_t^{(k)})$. Instead of waiting the communication to finish, it concurrently updates the current model with stale averaged stochastic gradient via $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\eta_t}{K}\sum_{k=1}^{K} \nabla F_k(\mathbf{x}_{t-s}; \xi_{t-s}^{(k)})$. Pipelining is significantly different from delay-intolerant asynchronous SGD (Ho et al., 2013; Lian et al., 2015) which computes stochastic gradient using stale model. But Pipe-SGD is sensitive to high communication delay (large s).

Pipelining with Periodic Averaging. CoCoD-SGD (Shen et al., 2019) parallelizes the local update and averaging of the local model to reduce the communication rounds by concurrently conducting

$$\mathbf{x}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)} \quad \text{and} \quad \mathbf{x}_{t+p}^{(k)} = \mathbf{x}_t^{(k)} - \sum_{\tau=t}^{t+p-1}\eta_\tau \nabla F_k(\mathbf{x}_\tau^{(k)};\xi_\tau^{(k)})\,. \tag{1}$$

When both the above equations are finished, $\mathbf{x}_{t+p}^{(k)} \leftarrow \mathbf{x}_t + \mathbf{x}_{t+p}^{(k)} - \mathbf{x}_t^{(k)}$ and $t \leftarrow t+p$. Therefore the current update relies on the local model $\mathbf{x}_t^{(k)}$ other than the averaged model $\mathbf{x}_t$ which requires communication. Here we add a superscript (k) to $\mathbf{x}_t$ to denote the local model at worker k because it is not consistent across workers anymore compared with Pipe-SGD. The performance of CoCoD-SGD is empirically sensitive to p.

OverlapLocalSGD (Wang et al., 2020) extends CoCoD-SGD by heuristically pulling $\mathbf{x}_{t+p}^{(k)}$ back a little to the $\mathbf{x}_t$ after the completion of Eq. (1) to reduce the inconsistency in the local models across workers: $\mathbf{x}_{t+p}^{(k)} \leftarrow (1-\alpha)\mathbf{x}_{t+p}^{(k)} + \alpha\mathbf{x}_t$ where α is a hyperparameter, and $t \leftarrow t+p$. For the momentum variant, OverlapLocalSGD maintains a slow momentum buffer for $\mathbf{x}_t$ following SlowMo (Wang et al., 2019). However, these methods suffer from the large communication budget.

Gradient Compression. We denote the compressor for compression as $\mathcal{C}(\cdot)$. Specifically, Scalar quantization is to round a 32-bit floating-point gradient component to a low-precision value of only several bits. Throughout this paper, we refer to scaled SignSGD as SignSGD (Bernstein et al., 2018; Karimireddy et al., 2019) which compresses a vector $\mathbf{v} \in \mathbb{R}^d$ to $\mathcal{C}(\mathbf{v}) = \frac{\|\mathbf{v}\|_1}{d}\,\mathrm{sign}(\mathbf{v})$. Gradient sparsification only selects large gradient components for communication. Vector quantization is to transform the gradient from a high-dimensional space to a low-dimensional space. With the help of local error feedback (Seide et al., 2014; Lin et al., 2018b; Wu et al., 2018; Karimireddy et al., 2019; Zheng et al., 2019) which adds the current compression error $(\mathbf{v} - \mathcal{C}(\mathbf{v}))$ to the next iteration, gradient compression can achieve comparable performance with the full-precision training. Local error feedback also works for both *one-way compression* (compress the communication from worker to sever) (Karimireddy et al., 2019) and *two-way compression* (compress the communication between worker and server) (Zheng et al., 2019).

Difficulties. Achieving communication compression with pipelining and periodic averaging requires careful design due to 1) the staleness introduced by pipelining; and 2) for state-of-the-art vector quantization method PowerSGD (Vogels et al., 2019), the compression $\mathcal{C}(\cdot)$ requires an additional round of communication which is unfavorable in high communication delay scenario.

3 THE PROPOSED FRAMEWORK: OLCO$_3$

In this section, we propose the delay-tolerant efficient training framework OLCO$_3$ with two versions (OLCO$_3$-TC and OLCO$_3$-VQ) dedicated for different compression methods in Algorithm 1 with novel compression and staleness compensations techniques. Some differences of OLCO$_3$-VQ compared with OLCO$_3$-TC are marked with red color. For simplicity, we use the phrase "line x" to represent the x-th line in Algorithm 1.

3.1 OLCO$_3$-TC FOR TWO-WAY COMPRESSION

Motivation. OLCO$_3$-TC is summarized in the green part of Algorithm 1 for efficient master-slave distributed training. Naively pipelining local computation with compressed communication will certainly break the momentum SGD update rule of the averaged model $\mathbf{x}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)}$ and lead to non-convergence. Instead, we consider the *auxiliary variable* $\tilde{\mathbf{x}}_t := \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_t^{(k)} - \mathbf{e}_t$ where $\mathbf{e}_t^{(k)}$ is the local compression error at worker k and $\mathbf{e}_t$ is the compression error at server. If $\tilde{\mathbf{x}}_t$ approximately follows the momentum SGD update rule, then the real trained model $\mathbf{x}_t$ will gradually approach $\tilde{\mathbf{x}}_t$ as training converges with the gradient and local error $\mathbf{e}_t^{(k)} \to \mathbf{0}$.

Pipelining. In non-communication iteration, we perform the local update following Local SGD (line 4). To pipeline the communication and computation in communication iteration which happens once per p iteration, we compress the local update $\Delta_{t+1}^{(k)}$ (line 7) for efficient communication but concurrently try to update the model with a *stale compressed global update* $\mathcal{C}(\Delta_{t+1-sp})$ (line 13) that has been outdated for s communication rounds (*i.e.*, staleness is s). We may maintain or reset

Algorithm 1 Overlap Local Computation with Compressed Communication (OLCO$_3$) on worker $k \in [K]$. Green part: OLCO$_3$-TC; Yellow part: OLCO$_3$-VQ.

1: **Input:** period $p \geq 1$, staleness $s \geq 0$, number of iterations T, number of workers K, learning rate $\{\eta_t\}_{t=0}^{T-1}$, and compression scheme $\mathcal{C}(\cdot)$.
2: **Initialize:** Local model $\mathbf{x}_0^{(k)} = \mathbf{x}_0$, local error $\mathbf{e}_0^{(k)} = \mathbf{0}$, server error $\mathbf{e}_0 = \mathbf{0}$, local momentum buffer $\mathbf{m}_0^{(k)} = \mathbf{0}$, and momentum constant $0 < \mu < 1$. Variables with negative subscripts are $\mathbf{0}$.
3: **for** $t = 0, 1, \cdots, T-1$ **do**
4: $\quad \mathbf{m}_{t+1}^{(k)} = \mu \mathbf{m}_t^{(k)} + \nabla f_k(\mathbf{x}_t^{(k)}; \xi_t^{(k)})$, $\mathbf{x}_{t+1}^{(k)} = \mathbf{x}_t^{(k)} - \eta_t \mathbf{m}_{t+1}^{(k)}$ *// Momentum Local SGD.*
5: $\quad$ **if** $(t+1) \mod p = 0$ **then**
6: $\quad\quad$ Maintain or reset the momentum buffer.
7: $\quad\quad \Delta_{t+1}^{(k)} = \mathbf{x}_{t+1-p}^{(k)} - \mathbf{x}_{t+1}^{(k)} + \mathbf{e}_t^{(k)}$ *// Compression compensation.*
8: $\quad\quad \mathbf{e}_{t+1}^{(k)} = \mathbf{e}_{t+2}^{(k)} = \cdots = \mathbf{e}_{t+p}^{(k)} = \Delta_{t+1}^{(k)} - \mathcal{C}(\Delta_{t+1}^{(k)})$ *// Compression.*
9: $\quad\quad$ Invoke the communication thread *in parallel* which does:
10: $\quad\quad\quad$ (1) Send $\mathcal{C}(\Delta_{t+1}^{(k)})$ to and receive $\mathcal{C}(\Delta_{t+1})$ from the server node.
11: $\quad\quad\quad$ (2) Server: $\Delta_{t+1} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{C}(\Delta_{t+1}^{(k)}) + \mathbf{e}_t$; $\mathbf{e}_{t+1} = \mathbf{e}_{t+2} = \cdots = \mathbf{e}_{t+p} = \Delta_{t+1} - \mathcal{C}(\Delta_{t+1})$.
12: $\quad\quad$ Block until $\mathcal{C}(\Delta_{t+1-sp})$ is ready.
13: $\quad\quad \mathbf{x}_{t+1} = \mathbf{x}_{t+1-p} - \mathcal{C}(\Delta_{t+1-sp})$
14: $\quad\quad \mathbf{x}_{t+1}^{(k)} \leftarrow \mathbf{x}_{t+1} - \sum_{i=0}^{s-1} \mathcal{C}(\Delta_{t+1-ip}^{(k)})$ *// Staleness compensation.*
15: $\quad\quad \Delta_{t+1}^{(k)} = \mathbf{x}_{t+1-p}^{(k)} - \mathbf{x}_{t+1}^{(k)} + \mathbf{e}_{t-sp}^{(k)}$ *// Compression compensation.*
16: $\quad\quad$ Invoke the communication thread *in parallel* which does:
17: $\quad\quad\quad$ (1) $\mathbf{e}_{t+1}^{(k)} = \mathbf{e}_{t+2}^{(k)} = \cdots = \mathbf{e}_{t+p}^{(k)} = \Delta_{t+1}^{(k)} - \mathcal{C}(\Delta_{t+1}^{(k)})$ *// Compression.*
18: $\quad\quad\quad$ (2) Average $\frac{1}{K}\sum_{k=1}^{K} \mathcal{C}(\Delta_{t+1}^{(k)})$ by ring all-reduce or master-slave communication.
19: $\quad\quad$ Block until $\frac{1}{K}\sum_{k=1}^{K} \mathcal{C}(\Delta_{t+1-sp}^{(k)})$ and $\mathbf{e}_{t+1-sp}^{(k)}$ is ready.
20: $\quad\quad \mathbf{x}_{t+1} = \mathbf{x}_{t+1-p} - \frac{1}{K}\sum_{k=1}^{K} \mathcal{C}(\Delta_{t+1-sp}^{(k)})$
21: $\quad\quad \mathbf{x}_{t+1}^{(k)} \leftarrow \mathbf{x}_{t+1} - \sum_{i=0}^{s-1} \Delta_{t+1-ip}^{(k)}$ *// Staleness compensation.*
22: $\quad$ **end if**
23: **end for**
24: **Output:** averaged model $\mathbf{x}_T = \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_T^{(k)}$

the momentum buffer to zero every p iteration (line 6). If the delay tolerance $\mathcal{T} = sp$ is larger than the communication delay, then the blocking in line 12 becomes a no-op and there will be no synchronization barrier. The server also compresses the sum of compressed local update from all workers (line 10) and send it back (line 11), making OLCO$_3$-TC an efficient two-way compression method.

Compensation. To make the update of $\tilde{\mathbf{x}}_t$ in OLCO$_3$-TC approximately follows SGD, we propose to 1) compensate the staleness with all *compressed local update* of staleness $\in [0, s-1]$ which requires no communication (line 14), such that less stale local update will also affect the local model; and 2) maintain a local error (line 8) and add it to the next local update (line 7) to compensate the compression loss. With the compensation techniques applied in OLCO$_3$-TC, Lemma 1 shows that the update of the auxiliary variable $\tilde{\mathbf{x}}_t$ is *exactly the same* as momentum SGD with the averaged momentum $\frac{1}{K}\sum_{k=1}^{K} \mathbf{m}_t^{(k)}$. Note that there is a "gradient mismatch" problem as the local momentum $\mathbf{m}_t^{(k)}$ is computed with stochastic gradient at local model $\mathbf{x}_t^{(k)}$ updates the auxiliary variable $\tilde{\mathbf{x}}_t$ (Karimireddy et al., 2019; Xu et al., 2020). But theoretical results will show that it does not affect the convergence rate.

Lemma 1. *For OLCO$_3$-TC, let* $\tilde{\boldsymbol{x}}_t := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{e}_t^{(k)} - \boldsymbol{e}_{t-sp}$*, then we have* $\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^{K} \boldsymbol{m}_t^{(k)}$.

We also consider OLCO$_3$ for one-way compression (*i.e.*, OLCO$_3$-OC), a special case of OLCO$_3$-TC by setting the compressor at the server side to identity function. The server error $\mathbf{e}_t$ will also become $\mathbf{0}$. Its auxiliary variable also satisfies Lemma 2. We can see that the delay tolerance of OLCO$_3$ satisfies $\mathcal{T} = sp(s \geq 1, p \geq 1)$. Its complexity of memory overheads is $\mathcal{O}(sd)$ due to storing information of staleness $\in [0, s-1]$. For common compression schemes $\mathcal{C}(\cdot)$ such as SignSGD, the complexity of its computation overheads is usually $\mathcal{O}(d)$ in OLCO$_3$-OC and OLCO$_3$-TC.

Lemma 2. *For OLCO$_3$-OC, let* $\tilde{\boldsymbol{x}}_t := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{e}_t^{(k)}$*, then we have* $\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^{K} \boldsymbol{m}_t^{(k)}$.

3.2 OLCO$_3$-VQ for Commutative Vector Quantization

OLCO$_3$-TC and OLCO$_3$-OC only achieve common compression in the master-slave communication paradigm. In OLCO$_3$-VQ (the yellow part of Algorithm 1), we also consider state-of-the-art commutative vector quantization scheme PowerSGD (Vogels et al., 2019) for both master-slave and ring all-reduce communication paradigms. However, it is essentially more difficult than common compressors as the compression using PowerSGD requires communication, which suffers from high communication delay. Specifically, PowerSGD first arouses the communication across all workers to compute a transformation matrix, which then is used to transform the local update to a low-dimension compressed form. It features commutability of compression ($\mathcal{C}(\mathbf{v}_1) + \mathcal{C}(\mathbf{v}_2) = \mathcal{C}(\mathbf{v}_1 + \mathbf{v}_2)$, Yu et al. (2018)) such that we only have to compress once both in master-slave and ring all-reduce, but adds to the difficulty of pipelining.

Pipelining with Communication Dependent Compressor. To address this issue, we observe that the staleness compensation in the main thread of OLCO$_3$-TC and OLCO$_3$-OC relies on *immediate* compressed results to acquire the up-to-date local error. When the compression scheme itself needs communication as PowerSGD does, we propose a novel compression compensation technique with a stale local error (line 13) to remove the dependencies on the up-to-date local error in OLCO$_3$-VQ. Consequently, we can move the whole compression and averaging process to the communication thread (lines 17 and 18). For the staleness compensation of OLCO$_3$-VQ, we store all *uncompressed local update* of staleness $\in [0, s-1]$. Following this design, Lemma 3 in OLCO$_3$-VQ where the auxiliary variable associated with the *stale local error* also exists for convergence guarantee.

Lemma 3. *For OLCO$_3$-VQ, let* $\tilde{\boldsymbol{x}}_t := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{e}_{t-sp}^{(k)}$*, then we have* $\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^{K} \boldsymbol{m}_t^{(k)}$.

4 Theoretical Results

In this section, we conduct convergence analysis of OLCO$_3$-VQ and OLCO$_3$-TC for both SGD and momentum SGD maintaining momentum with common assumptions below. As OLCO$_3$-OC is a special case of OLCO$_3$-TC, we only analyze OLCO$_3$-TC and OLCO$_3$-VQ. The detailed proof of Theorem 1, 2, 3, and 4 can be found in Appendix Section D, E, F, and G respectively. The detailed proof of key Lemma 1, 2, and 3 can be found in Appendix Section C.

Assumption 1. *(L-Lipschitz Smoothness) Both the local ($f_k(\cdot)$) and global ($f(\cdot) = \frac{1}{K}\sum_{k=1}^{K} f_k(\cdot)$) loss functions are L-smooth, i.e.,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d\,, \tag{2}$$

$$\|\nabla f_k(\boldsymbol{x}) - \nabla f_k(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \forall k \in [K], \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d\,. \tag{3}$$

Assumption 2. *(Local Bounded Variance) The local stochastic gradient $\nabla F_k(\boldsymbol{x};\xi)$ has a bounded variance, i.e., $\mathbb{E}_{\xi\sim\mathcal{D}_k}\|\nabla F_k(\boldsymbol{x};\xi) - \nabla f_k(\boldsymbol{x})\|_2^2 \leq \sigma^2, \forall k \in [K], \forall \boldsymbol{x} \in \mathbb{R}^d$. Note that $\mathbb{E}_{\xi\sim\mathcal{D}_k}\nabla F_k(\boldsymbol{x};\xi) = \nabla f_k(\boldsymbol{x})$.*

Assumption 3. *(Bounded Variance across Workers) The L_2 norm of the difference of the local and global full gradient is bounded, i.e., $\|\nabla f_k(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|_2^2 \leq \kappa^2, \forall k \in [K], \forall \boldsymbol{x} \in \mathbb{R}^d$. $\kappa = 0$ leads to i.i.d. data distributions across workers.*

Assumption 4. *(Bounded Full Gradient) The second moment of the global full gradient is bounded, i.e., $\|\nabla f(\boldsymbol{x})\|_2^2 \leq G^2, \forall \boldsymbol{x} \in \mathbb{R}^d$.*

Assumption 5. *(Karimireddy et al., 2019) The compression function $\mathcal{C}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a δ-approximate compressor for $0 < \delta \leq 1$ if for all $\boldsymbol{v} \in \mathbb{R}^d$, $\|\mathcal{C}(\boldsymbol{v}) - \boldsymbol{v}\|_2^2 \leq (1-\delta)\|\boldsymbol{v}\|_2^2$.*

4.1 SGD

Theorem 1. *For OLCO$_3$-VQ with vanilla SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \min\{\frac{1}{6L(s+1)p}, \frac{1}{9L}\}$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(f(\boldsymbol{x}_0)-f_*)}{\eta T} + \frac{9\eta L\sigma^2}{K} + 12\eta^2L^2(s+1)p\sigma^2[1+ \frac{14(1-\delta)}{\delta^2}(s+1)p] + 36\eta^2L^2(s+1)^2p^2\kappa^2(1+\frac{5(1-\delta)}{\delta^2}) + \frac{168(1-\delta)}{\delta^2}\eta^2L^2(s+1)^2p^2G^2\,. \tag{4}$$

If we further choose the learning rate $\eta = \mathcal{O}(K^{\frac{1}{2}}T^{-\frac{1}{2}})$ and the communication interval $p = \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}}(s+1)^{-1})$, then the convergence rate will be $\mathcal{O}(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

Theorem 2. *For OLCO$_3$-TC with vanilla SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \min\{\frac{1}{6L(s+1)p}, \frac{1}{9L}\}$ and let $h(\delta) = \frac{1-\delta}{\delta^2}(1+\frac{4(2-\delta)}{\delta^2})$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(f(\boldsymbol{x}_0)-f_*)}{\eta T} + \frac{9\eta L\sigma^2}{K} + 12\eta^2p\sigma^2(s+1+80h(\delta)p) + 12\eta^2p^2\kappa^2(3(s+1)^2+80h(\delta)) + 960\eta^2p^2G^2h(\delta)\,. \tag{5}$$

If we further choose the learning rate $\eta = \mathcal{O}(K^{\frac{1}{2}}T^{-\frac{1}{2}})$ and the communication interval $p = \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}}(s+1)^{-1})$, then the convergence rate will be $\mathcal{O}(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$. For i.i.d. data distributions across workers ($\kappa = 0$), if we choose the learning rate $\eta = \mathcal{O}(K^{\frac{1}{2}}T^{-\frac{1}{2}})$ and the communication interval $p = \min\{\mathcal{O}(K^{-\frac{3}{2}}T^{\frac{1}{2}}(s+1)^{-1}), \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}})\}$ ($p = \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}})$ for a enough large T) instead, then the convergence rate will still be $\mathcal{O}(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

4.2 MOMENTUM SGD

Theorem 3. *For OLCO$_3$-VQ with Momentum SGD and under Assumption 1, 2, 3, 4, 5, if the learning rate $\eta \leq \min\{\frac{1-\mu}{\sqrt{72}L(s+1)p}, \frac{1-\mu}{9L}\}$ and let $g(\mu,\delta,s,p) = \frac{15}{(1-\mu)^2} + \frac{60(1-\delta)(s+1)^2p^2}{\delta^2}$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(1-\mu)[f(\boldsymbol{z}_0)-f(\boldsymbol{z}_T)]}{\eta} + \frac{9L\eta\sigma^2}{(1-\mu)K} + \frac{4\eta^2L^2}{(1-\mu)^2}[(4(s+1)p+g(\mu,\delta,s,p))\sigma^2 + (12(s+1)^2p^2+g(\mu,\delta,s,p))\kappa^2 + g(\mu,\delta,s,p)G^2]\,. \tag{6}$$

The same convergence rate and communication complexity are achieved as Theorem 1.

Theorem 4. *For OLCO$_3$-TC with Momentum SGD and under Assumption 1, 2, 3, 4, 5, if the learning rate $\eta \leq \min\{\frac{1-\mu}{\sqrt{72}L(s+1)p}, \frac{1-\mu}{9L}\}$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(1-\mu)[f(\boldsymbol{z}_0)-f(\boldsymbol{z}_T)]}{\eta} + \frac{9L\eta\sigma^2}{(1-\mu)K} + \frac{6\eta^2L^2}{(1-\mu)^2}[\sigma^2(\frac{9}{(1-\mu)^2}+2(s+1)p+168h(\delta)p^2) + \kappa^2(\frac{9}{(1-\mu)^2}+6(s+1)^2p^2+168h(\delta)p^2) + G^2(\frac{9}{(1-\mu)^2}+168h(\delta)p^2)]\,, \tag{7}$$

where $h(\delta) = \frac{1-\delta}{\delta^2}(1+\frac{4(2-\delta)}{\delta^2})$. The same convergence rate and communication complexity are achieved as Theorem 2.

Therefore, OLCO$_3$-TC has a better complexity of communication interval p than OLCO$_3$-VQ in *i.i.d.* settings. But they are similar in *non-i.i.d.* settings.

5 EXPERIMENTS

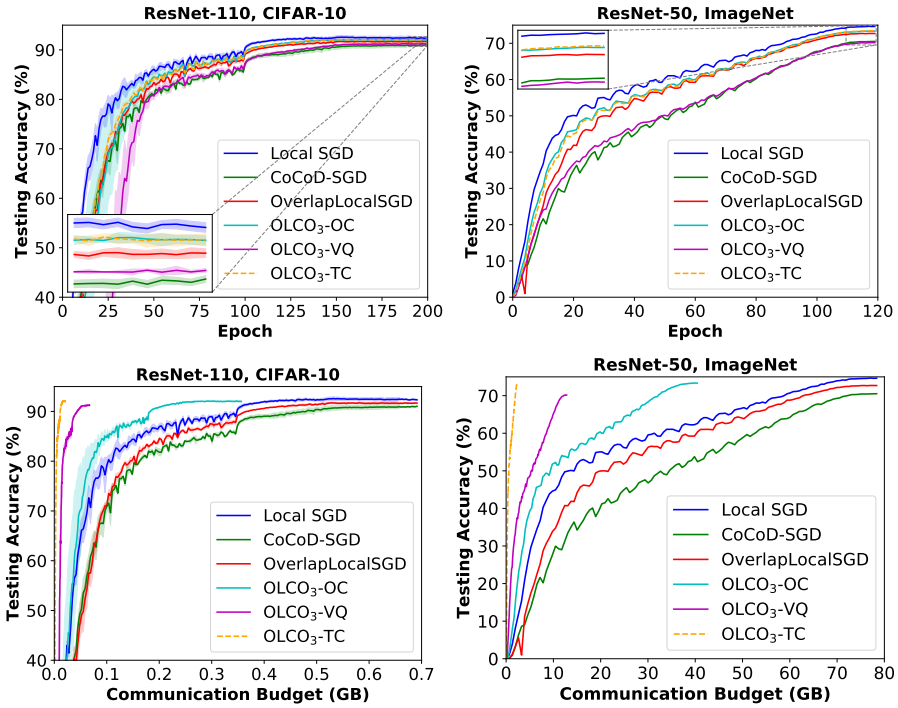

Figure 1: Training curves using $\mathcal{T} = 56$ and $s = 1$ for delay-tolerant methods except that $\mathcal{T} = 0$ and $p = 56$ for Local SGD. Best view in color. Numerical values can be found in Appendix Section A.3.

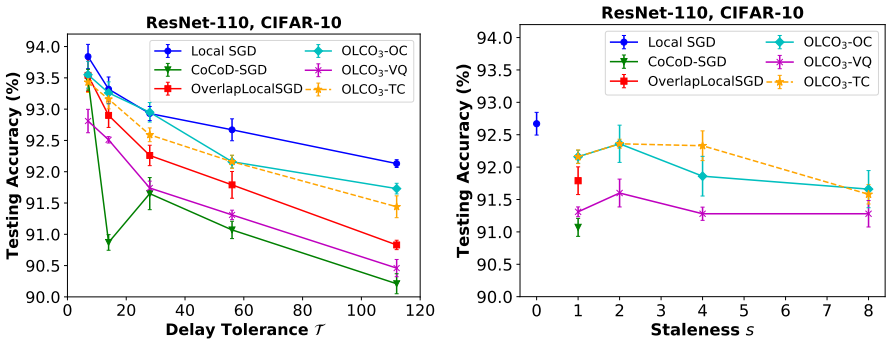

Figure 2: **Left**: Vary delay tolerance $\mathcal{T} = sp$ with fixed staleness $s = 1$. For Local SGD, $\mathcal{T} = 0$ but we use the same period p as delay-tolerant methods. **Right**: Vary staleness s with fixed delay tolerance $\mathcal{T} = 56$ for OLCO$_3$. More settings of s and p can be found in Appendix Section A.4.

We compare the following methods: 1) Local SGD (**baseline, NO delay tolerance** with $\mathcal{T} = 0$); 2) Pipe-SGD; 3) CoCoD-SGD; 4) OverlapLocalSGD, the hyperparamters of which follow Wang et al. (2020); 5) OLCO$_3$-OC with SignSGD compression; 6) OLCO$_3$-VQ with PowerSGD compressoin; 7) OLCO$_3$-TC with SignSGD compression. We maintain the momentum buffer in default unless specified otherwise. We do not report the results of Pipe-SGD as it does not converge for large delay tolerance $\mathcal{T}$. We train the ResNet-110 (He et al., 2016) model with 8 workers on CIFAR-10 (Krizhevsky et al., 2009) image classification task and report the mean and standard deviation metrics over 3 runs in both *i.i.d.* and *non-i.i.d.* settings. We also train the ResNet-50 model with 16 workers on ImageNet (Russakovsky et al., 2015) image classification tasks. Detailed description can be found in Appendix Section A.1.

Table 2: *Non-i.i.d.* testing accuracy of ResNet-110 on CIFAR-10 using $\mathcal{T} = 56$ for delay-tolerant methods except that $\mathcal{T} = 0$ and $p = 56$ for Local SGD. Training curves can be found in Appendix Section A.2.

Method	Delay Tolerance $\mathcal{T}$	Compression Ratio	ResNet110
Local SGD	0	1	89.93 ± 0.08
Pipe-SGD	56	1	Diverges
CoCoD-SGD	56	1	87.73 ± 0.25
OverlapLocalSGD	56	1	88.97 ± 0.19
$OLCO_3$-OC	56	1/2 + 1/32	$\mathbf{89.75 \pm 0.21}$
$OLCO_3$-VQ	56	0.0973	89.31 ± 0.08
$OLCO_3$-TC	56	**1/32**	$\mathbf{89.73 \pm 0.25}$

Lower Communication Budget with Delay Tolerance. The training curves of ResNet-110 on CIFAR-10 and ResNet-50 on ImageNet are shown in Figure 1. $s = 1$ because CoCoD-SGD and OverlapLocalSGD does not support $s \geq 2$. In comparison with other delay-tolerant methods, our proposed $OLCO_3$ achieves a significantly lower communication budget due to compressed communication. $OLCO_3$ features extreme efficiency with compressed communication, delay tolerance, and low communication frequency due to periodic averaging.

Better Model Performance. The upper two figures in Figure 1 shows that $OLCO_3$-OC and $OLCO_3$-TC outperforms other delay-tolerant methods and is comparable to the baseline Local SGD in terms of the model performance, with $OLCO_3$-VQ similar to CoCoD-SGD but inferior to OverlapLocalSGD. $OLCO_3$-OC and $OLCO_3$-TC perform similarly. However, in *non-i.i.d.* experiments illustrated in Table 2, all $OLCO_3$-OC, $OLCO_3$-VQ, $OLCO_3$-TC are superior to other delay-tolerant methods regarding testing accuracy. This confirms the theoretical results in Theorem 1, 2, 3, and 4 that $OLCO_3$-TC can deal with a larger p than $OLCO_3$-VQ in *i.i.d.* settings, but they are similar in *non-i.i.d.* settings. $OLCO_3$ also performs very close to Local SGD. In average regarding the testing accuracy, $OLCO_3$-OC and $OLCO_3$-TC improves CoCoD-SGD and OverlapLocalSGD by 2.0% and 0.8% respectively, while $OLCO_3$-VQ improves CoCoD-SGD and OverlapLocalSGD by 1.6% and 0.4% respectively. This also confirms the effectiveness of staleness and compression compensation in $OLCO_3$.

Varying Delay Tolerance. We vary the delay tolerance $\mathcal{T}$ with fixed staleness $s = 1$ in the left of Figure 2 to show that the same conclusion can be reached in various settings of the delay tolerance. $OLCO_3$-OC and $OLCO_3$-TC always outperform other delay-tolerant methods with comparable performance to corresponding Local SGD. $OLCO_3$-VQ always outperforms CoCoD-SGD with a much smaller communication budget.

Varying Staleness. We also vary the staleness s of our $OLCO_3$ in the right of Figure 2 with fixed delay tolerance $\mathcal{T}$. Local SGD only supports $s = 0$ with no delay tolerance and CoCoD-SGD and OverlapLocalSGD only support $s = 1$, so there is only one result for them in the figure. With the increasing of staleness s, the deterioration of model performance training by $OLCO_3$ is very small, especially for $OLCO_3$-VQ. It confirms the effectiveness of the proposed staleness compensation. Typically, we see a better model performance by choosing $s = 2$ because an appropriate staleness may introduce noise for better generalization. In comparison, we cannot tune staleness s in CoCoD-SGD and OverlapLocalSGD.

6 Conclusion

In this work, we propose $OLCO_3$ to achieve extreme communication efficiency with high delay tolerance and low communication budget via novel staleness and compression compensation. The theoretical result shows that it converges as fast as vanilla SGD. Experiments confirm the analysis and show a significant improvement regarding communication budget and model performance over other existing delay-tolerant methods.

REFERENCES

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

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

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

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/bernstein18a.html`.

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

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

Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B Gibbons, Garth A Gibson, Greg Ganger, and Eric P Xing. More effective distributed ml via a stale synchronous parallel parameter server. In *Advances in neural information processing systems*, pp. 1223–1231, 2013.

Zhouyuan Huo, Bin Gu, and Heng Huang. Large batch training does not need warmup. *arXiv preprint arXiv:2002.01576*, 2020.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261, 2019.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Youjie Li, Mingchao Yu, Songze Li, Salman Avestimehr, Nam Sung Kim, and Alexander Schwing. Pipe-sgd: A decentralized pipelined sgd framework for distributed deep net training. In *Advances in Neural Information Processing Systems*, pp. 8045–8056, 2018.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 2737–2745, 2015.

Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018a.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018b. URL `https://openreview.net/forum?id=SkhQHMW0W`.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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

Shuheng Shen, Linli Xu, Jingchang Liu, Xianfeng Liang, and Yifei Cheng. Faster distributed deep net training: Computation and communication decoupled stochastic gradient descent. *arXiv preprint arXiv:1906.12043*, 2019.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1g2JnRcFX`.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pp. 4447–4458, 2018.

Nikko Strom. Scalable distributed dnn training using commodity gpu cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 14259–14268, 2019.

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

Jianyu Wang, Hao Liang, and Gauri Joshi. Overlap local-sgd: An algorithmic approach to hide communication delays in distributed sgd. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8871–8875. IEEE, 2020.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 1299–1309, 2018.

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

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

An Xu, Zhouyuan Huo, and Heng Huang. Training faster with compressed gradient. *arXiv preprint arXiv:2008.05823*, 2020.

Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 6, 2017.

Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander Schwing, Murali Annavaram, and Salman Avestimehr. Gradiveq: Vector quantization for bandwidth-efficient gradient aggregation in distributed cnn training. In *Advances in Neural Information Processing Systems*, pp. 5123–5133, 2018.

Shuai Zheng, Ziyue Huang, and James Kwok. Communication-efficient distributed blockwise momentum sgd with error-feedback. In *Advances in Neural Information Processing Systems*, pp. 11446–11456, 2019.

A ADDITIONAL EXPERIMENTAL RESULTS

A.1 EXPERIMENTAL SETTING

All experiments are implemented with PyTorch (Paszke et al., 2019) and run on a cluster of Nvidia Tesla P40 GPUs. Each node is connected by 40Gbps Ethernet and equipped with 4 GPUs.

CIFAR. We train the ResNet-110 (He et al., 2016) model with 8 workers on CIFAR-10 (Krizhevsky et al., 2009) image classification task. We report the mean and standard deviation metrics over 3 runs. The base learning rate is 0.4 and the total batch size is 512. The momentum constant is 0.9 and the weight decay is 1×10^{-4}. For momentum SGD the model is trained for 200 epochs with a learning rate decay of 0.1 at epoch 100 and 150. We linearly warm up the learning rate from 0.05 to 0.4 in the beginning 5 epochs. For OLCO$_3$ with staleness $s \in \{2, 4, 8\}$, we set the base learning rate to 0.2 due to increased staleness. The rank of PowerSGD is 4. Random cropping, random flipping, and standardization are applied as data augmentation techniques. We also train ResNet-56 to explore more combinations of s and p in Section A.4 with the same other settings.

ImageNet. We train the ResNet-50 model with 16 workers on ImageNet (Russakovsky et al., 2015) image classification tasks. The model is trained for 120 epochs with a cosine learning rate scheduling (Loshchilov & Hutter, 2016). The base learning rate is 0.4 and the total batch size is 2048. The momentum constant is 0.9 and the weight decay is 1×10^{-4}. We linearly warm up the learning rate from 0.025 to 0.4 in the beginning 5 epochs. The rank of PowerSGD is 50. Random cropping, random flipping, and standardization are applied as data augmentation techniques.

***Non-i.i.d.* Settings.** Similar to Wang et al. (2020), we randomly choose fraction α of the whole data, sort the data by the class, and evenly assign them to all workers in order. For the rest fraction $1 - \alpha$ of the whole data, we randomly and evenly distribute them to all workers (Figure 3). When $0 < \alpha \leq 1$ is large, the data distribution across workers are *non-i.i.d* and highly skewed. When $\alpha = 0$, it becomes *i.i.d.* data distribution across workers. In our *non-i.i.d.* experiments, we choose $\alpha = 0.8$.

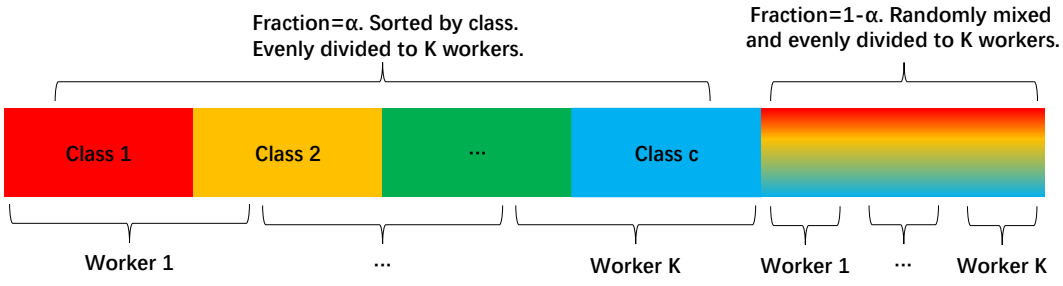

Figure 3: *Non-i.i.d.* data partition across workers. Best view in color.

A.2 TRAINING CURVE

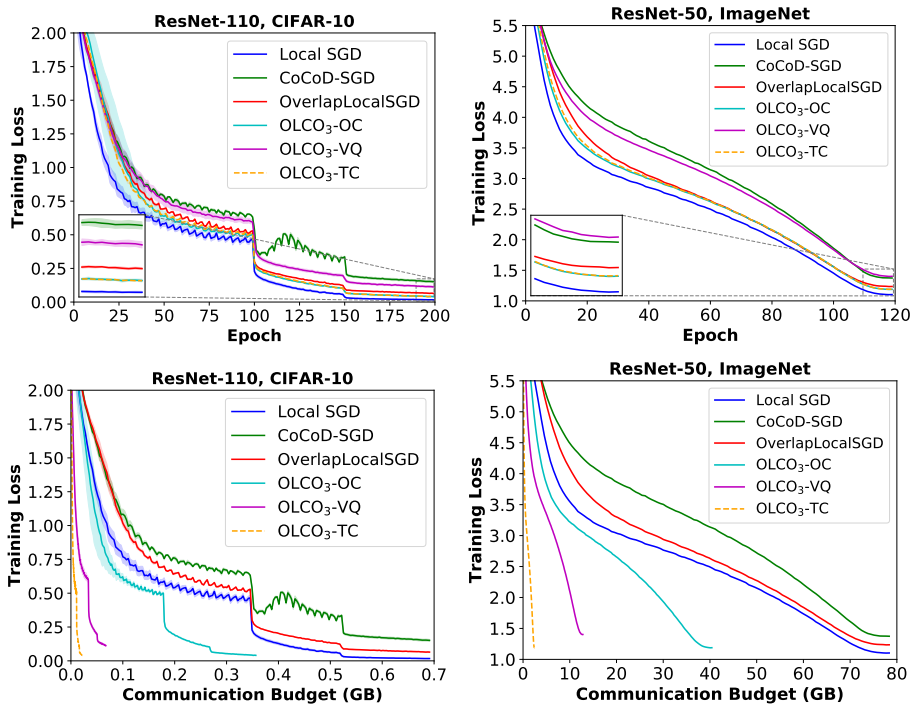

Figure 4: Training curves using $\mathcal{T} = 56$ for delay-tolerant methods except that $\mathcal{T} = 0$ and $p = 56$ for Local SGD. Best view in color.

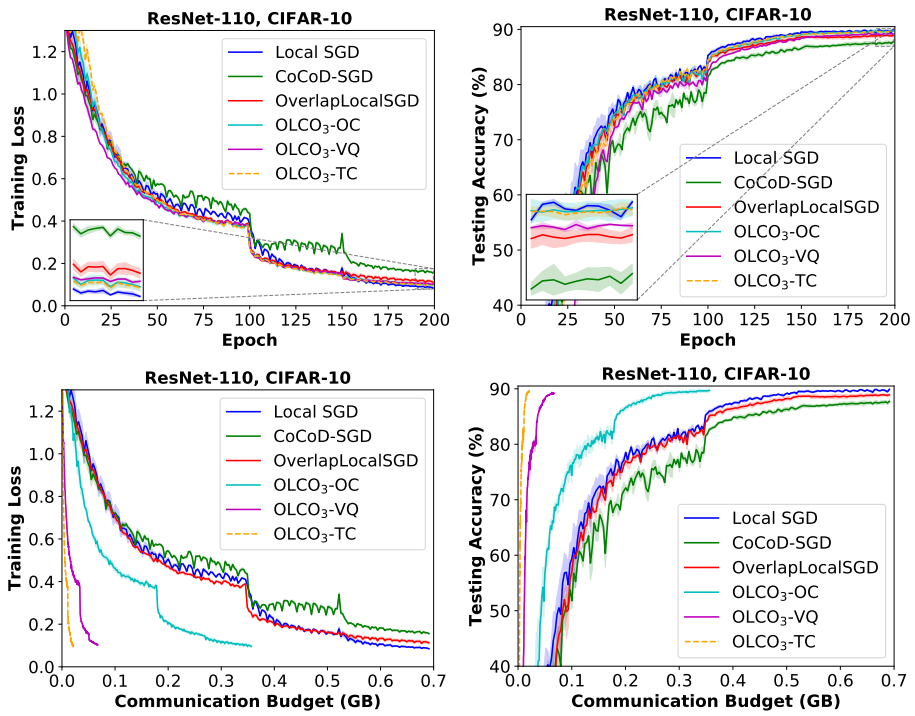

Figure 5: *Non-i.i.d.* training curves using $\mathcal{T} = 56$ for delay-tolerant methods except that $\mathcal{T} = 0$ and $p = 56$ for Local SGD. Best view in color.

A.3 TESTING ACCURACY

Table 3: Testing accuracy of ResNet-110 on CIFAR-10 and ResNet-50 on ImageNet using $\mathcal{T} = 56$ for delay-tolerant methods except that $\mathcal{T} = 0$ and $p = 56$ for Local SGD. CR stands for compression ratio. For each method, the first row denotes maintaining momentum and the second row denotes resetting momentum.

Method	Delay Tolerance $\mathcal{T}$	CR	ResNet110	CR	ResNet-50
Local SGD	0	1	92.67 ± 0.18 92.68 ± 0.21	1	74.68 74.76
Pipe-SGD	56	1	Diverges	1	Diverges
CoCoD-SGD	56	1	91.07 ± 0.14 90.84 ± 0.06	1	70.52 70.27
OverlapLocalSGD	56	1	91.78 ± 0.21 91.45 ± 0.14	1	72.71 72.87
OLCO_3-OC	56	1/2 + 1/32	$\mathbf{92.16 \pm 0.10}$ $\mathbf{92.48 \pm 0.12}$	1/2 + 1/32	**73.36** **73.40**
OLCO_3-VQ	56	0.0973	91.31 ± 0.08 91.05 ± 0.21	0.1633	70.38 70.32
OLCO_3-TC	56	**1/32**	$\mathbf{92.15 \pm 0.11}$ $\mathbf{92.31 \pm 0.10}$	**1/32**	**73.49** **73.47**

A.4 HYPERPARAMETERS s & p

runninggggggggg

B ASSUMPTIONS

Assumption 1. *(L-Lipschitz Smoothness) Both the local ($f_k(\cdot)$) and global ($f(\cdot) = \frac{1}{K}\sum_{k=1}^{K} f_k(\cdot)$) loss functions are L-smooth, i.e.,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d\,, \tag{8}$$

$$\|\nabla f_k(\boldsymbol{x}) - \nabla f_k(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \forall k \in [K], \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d\,. \tag{9}$$

Assumption 2. *(Local Bounded Variance) The local stochastic gradient $\nabla F_k(\boldsymbol{x};\xi)$ has a bounded variance, i.e.,*

$$\mathbb{E}_{\xi\sim\mathcal{D}_k}\|\nabla F_k(\boldsymbol{x};\xi) - \nabla f_k(\boldsymbol{x})\|_2^2 \leq \sigma^2, \forall k \in [K], \forall \boldsymbol{x} \in \mathbb{R}^d\,. \tag{10}$$

Note that $\mathbb{E}_{\xi\sim\mathcal{D}_k}\nabla F_k(\boldsymbol{x};\xi) = \nabla f_k(\boldsymbol{x})$.

Assumption 3. *(Bounded Variance across Workers) The L_2 norm of the difference of the local and global full gradient is bounded, i.e.,*

$$\|\nabla f_k(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|_2^2 \leq \kappa^2, \forall k \in [K], \forall \boldsymbol{x} \in \mathbb{R}^d\,, \tag{11}$$

where $\kappa = 0$ leads to i.i.d. data distributions across workers.

Assumption 4. *(Bounded Full Gradient) The second moment of the global full gradient is bounded, i.e.,*

$$\|\nabla f(\boldsymbol{x})\|_2^2 \leq G^2, \forall \boldsymbol{x} \in \mathbb{R}^d\,. \tag{12}$$

Assumption 5. *(δ-approximate compressor) The compression function $\mathcal{C}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a δ-approximate compressor for $0 < \delta \leq 1$ if for all $\boldsymbol{v} \in \mathbb{R}^d$,*

$$\|\mathcal{C}(\boldsymbol{v}) - \boldsymbol{v}\|_2^2 \leq (1-\delta)\|\boldsymbol{v}\|_2^2\,. \tag{13}$$

C BASIC LEMMAS

Lemma 1. *For OLCO$_3$-TC, let* $\tilde{\boldsymbol{x}}_t \coloneqq \frac{1}{K}\sum_{k=1}^K \boldsymbol{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^K \boldsymbol{e}_t^{(k)} - \boldsymbol{e}_{t-sp}$*, then we have*

$$\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^K \boldsymbol{m}_t^{(k)} . \tag{14}$$

Proof. For $t = np$ where n is some integer,

$$\begin{aligned}
\tilde{\mathbf{x}}_{np} &= \frac{1}{K}\sum_{k=1}^K \mathbf{x}_{np}^{(k)} - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_{np}^{(k)} - \mathbf{e}_{(n-s)p} = \mathbf{x}_{np} - \frac{1}{K}\sum_{k=1}^K\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_{np}^{(k)} - \mathbf{e}_{(n-s)p} \\
&= \mathbf{x}_{(n-1)p} - \mathcal{C}(\Delta_{(n-s)p}) - \frac{1}{K}\sum_{k=1}^K\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_{np}^{(k)} - \mathbf{e}_{(n-s)p} \\
&= \frac{1}{K}\sum_{k=1}^K(\mathbf{x}_{(n-1)p}^{(k)} + \sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-1-i)p}^{(k)})) - \mathcal{C}(\Delta_{(n-s)p}) - \frac{1}{K}\sum_{k=1}^K\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_{np}^{(k)} - \mathbf{e}_{(n-s)p} \\
&= \frac{1}{K}\sum_{k=1}^K \mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^K \Delta_{np}^{(k)} + \frac{1}{K}\sum_{k=1}^K \mathcal{C}(\Delta_{(n-s)p}^{(k)}) - \mathcal{C}(\Delta_{(n-s)p}) - \mathbf{e}_{(n-s)p} \\
&= \frac{1}{K}\sum_{k=1}^K \mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^K \Delta_{np}^{(k)} - \mathbf{e}_{(n-s)p-1} \\
&= \frac{1}{K}\sum_{k=1}^K \mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^K\sum_{\tau=(n-1)p}^{np-1}\eta_\tau \mathbf{m}_{\tau+1}^{(k)} - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_{np-1}^{(k)} - \mathbf{e}_{(n-s)p-1} = \tilde{\mathbf{x}}_{np-1} - \frac{1}{K}\sum_{k=1}^K \eta_{np-1}\mathbf{m}_{np}^{(k)} .
\end{aligned} \tag{15}$$

For $t \neq np$,

$$\tilde{\mathbf{x}}_t = \frac{1}{K}\sum_{k=1}^K \mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_t^{(k)} - \mathbf{e}_{t-sp} = \frac{1}{K}\sum_{k=1}^K \mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^K \mathbf{e}_{t-1}^{(k)} - \mathbf{e}_{t-sp-1} = \tilde{\mathbf{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^K \mathbf{m}_t^{(k)} . \tag{16}$$

□

Lemma 2. *For OLCO$_3$-OC, let* $\tilde{\boldsymbol{x}}_t \coloneqq \frac{1}{K}\sum_{k=1}^K \boldsymbol{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^K \boldsymbol{e}_t^{(k)}$*, then we have*

$$\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^K \boldsymbol{m}_t^{(k)} . \tag{17}$$

Proof. For $t = np$ where n is some integer,

$$
\begin{aligned}
\tilde{\mathbf{x}}_{np} &= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{np}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np}^{(k)} = \mathbf{x}_{np} - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np}^{(k)} \\
&= \mathbf{x}_{(n-1)p} - \frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{(n-s)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}_{(n-1)p}^{(k)} + \sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-1-i)p}^{(k)})) - \frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{(n-s)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(n-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{np}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\Delta_{np}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=(n-1)p}^{np-1}\eta_\tau \mathbf{m}_{\tau+1}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np-1}^{(k)} \\
&= \tilde{\mathbf{x}}_{np-1} - \frac{1}{K}\sum_{k=1}^{K}\eta_{np-1}\mathbf{m}_{np}^{(k)} .
\end{aligned}
\tag{18}
$$

For $t \neq np$,

$$
\tilde{\mathbf{x}}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_t^{(k)} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-1}^{(k)} = \tilde{\mathbf{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^{K}\mathbf{m}_t^{(k)} . \tag{19}
$$

□

Lemma 3. *For OLCO$_3$-VQ, let $\tilde{\boldsymbol{x}}_t \coloneqq \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{e}_{t-sp}^{(k)}$, then we have*

$$
\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^{K}\boldsymbol{m}_t^{(k)} . \tag{20}
$$

Proof. For $t = np$ where n is some integer,

$$\begin{aligned}
\tilde{\mathbf{x}}_{np} &= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{np}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{(n-s)p}^{(k)} = \mathbf{x}_{np} - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\Delta_{(n-i)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{(n-s)p}^{(k)} \\
&= \mathbf{x}_{(n-1)p} - \frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{(n-s)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\Delta_{(n-i)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{(n-s)p}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}_{(n-1)p}^{(k)} + \sum_{i=0}^{s-1}\Delta_{(n-1-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{(n-s)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\Delta_{(n-i)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{(n-s)p}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}_{(n-1)p}^{(k)} + \sum_{i=0}^{s-1}\Delta_{(n-1-i)p}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\Delta_{(n-s)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\Delta_{(n-i)p}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\Delta_{np}^{(k)} \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_{(n-1)p}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\sum_{\tau=(n-1)p}^{np-1}\eta_\tau \mathbf{m}_{\tau+1}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{np-1-sp}^{(k)} \\
&= \tilde{\mathbf{x}}_{np-1} - \frac{1}{K}\sum_{k=1}^{K}\eta_{np-1}\mathbf{m}_{np}^{(k)} .
\end{aligned} \tag{21}$$

For $t \neq np$,

$$\tilde{\mathbf{x}}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-1-sp}^{(k)} = \tilde{\mathbf{x}}_{t-1} - \frac{\eta_{t-1}}{K}\sum_{k=1}^{K}\mathbf{m}_t^{(k)} . \tag{22}$$

□

D PROOF OF THEOREM 1

Lemma 4. *For OLCO$_3$-VQ with vanilla SGD and under Assumption 2, 3, 4, and 5, the local error satisfies*

$$\mathbb{E}\|\boldsymbol{e}_t^{(k)}\|_2^2 \leq \frac{12(1-\delta)}{\delta^2}p^2\eta^2(\sigma^2 + \kappa^2 + G^2) . \tag{23}$$

Proof. First we have

$$\begin{aligned}
\mathbb{E}\|\nabla F(\mathbf{x}_t^{(k)};\xi_t^{(k)})\|_2^2 &\leq 3\mathbb{E}\|\nabla F(\mathbf{x}_t^{(k)};\xi_t^{(k)}) - \nabla f_k(\mathbf{x}_t^{(k)})\|^2 + 3\mathbb{E}\|\nabla f_k(\mathbf{x}_t^{(k)}) - \nabla f(\mathbf{x}_t^{(k)})\|_2^2 + 3\mathbb{E}\|\nabla f_k(\mathbf{x}_t^{(k)})\|_2^2 \\
&\leq 3\sigma^2 + 3\kappa^2 + 3G^2 .
\end{aligned} \tag{24}$$

Let $S_t = \lfloor \frac{t}{p} \rfloor$,

$$\begin{aligned}
\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 = \mathbb{E}\|\mathbf{e}_{S_tp}^{(k)}\|_2^2 &= \mathbb{E}\|\mathcal{C}(\Delta_{S_tp}^{(k)}) - \Delta_{S_tp}^{(k)}\|_2^2 \leq (1-\delta)\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 \\
&= (1-\delta)\mathbb{E}\|\sum_{t'=(S_t-1)p}^{S_tp-1}\eta\nabla F(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)}) + \mathbf{e}_{(S_t-s-1)p}^{(k)}\|_2^2 \\
&\leq (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-s-1)p}^{(k)}\|_2^2 + (1+\delta)(1+\frac{1}{\rho})\mathbb{E}\|\sum_{t'=(S_t-1)p}^{S_tp-1}\eta\nabla F(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2 \\
&\leq (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-s-1)p}^{(k)}\|_2^2 + 3(1+\delta)(1+\frac{1}{\rho})p^2\eta^2(\sigma^2+\kappa^2+G^2) .
\end{aligned} \tag{25}$$

Therefore,

$$\begin{aligned}\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 &\leq 3(1-\delta)(1+\frac{1}{\rho})p^2\eta^2(\sigma^2+\kappa^2+G^2)\sum_{i=0}^{\lfloor\frac{S_t}{s}\rfloor-1}[(1-\delta)(1+\rho)]^i \\ &\leq \frac{3(1-\delta)(1+\frac{1}{\rho})}{1-(1-\delta)(1+\rho)}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,.\end{aligned} \tag{26}$$

Let $\rho = \frac{\delta}{2(1-\delta)}$ such that $1+\frac{1}{\rho} = \frac{2-\delta}{\delta} \leq \frac{2}{\delta}$, then $\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 \leq \frac{12(1-\delta)}{\delta^2}p^2\eta^2(\sigma^2+\kappa^2+G^2)$. □

Lemma 5. *For OLCO$_3$-VQ with vanilla SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \frac{1}{6L(s+1)p}$, we have*

$$\begin{aligned}&\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t^{(k)}\|_2^2 \\ &\leq 3\eta^2(s+1)p\sigma^2(1+\frac{12(1-\delta)}{\delta^2}(s+1)p) + 9\eta^2(s+1)^2p^2\kappa^2(1+\frac{4(1-\delta)}{\delta^2}) + \frac{36(1-\delta)}{\delta^2}\eta^2(s+1)^2p^2G^2\,.\end{aligned} \tag{27}$$

Proof. Let $S_t = \lfloor\frac{t}{p}\rfloor$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}(-\sum_{i=0}^{s-1}\Delta_{(S_t-i)p}^{(k')} - \sum_{t'=S_tp}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})) \\ &\quad - (-\sum_{i=0}^{s-1}\Delta_{(S_t-i)p}^{(k)} - \sum_{t'=S_tp}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})) - \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')}\|_2^2 \\ &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\| - \frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')}) + \sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)}) \\ &\quad - \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')} - \frac{1}{K}\sum_{k'=1}^{K}\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k')} + \sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2 \\ &\leq \frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\| - \frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')}) + \sum_{t'=(S_t-s)p}^{t-1}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2 \\ &\quad + \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\| - \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')} - \frac{1}{K}\sum_{k'=1}^{K}\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k')} + \sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2\,.\end{aligned} \tag{28}$$

The first term is bounded by

$$
\begin{aligned}
&\frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})+\sum_{t'=(S_t-s)p}^{t-1}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2\\
&\leq \frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\sum_{t'=(S_t-s)p}^{t-1}\left(-\frac{1}{K}\sum_{k'=1}^{K}(\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})-\nabla f_{k'}(\mathbf{x}_{t'}^{(k')}))+(\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\nabla f_k(\mathbf{x}_{t'}^{(k)}))\right)\right\|_2^2\\
&\quad+\frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\sum_{t'=(S_t-s)p}^{t-1}(-\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_{t'}^{(k')})+\nabla f_k(\mathbf{x}_{t'}^{(k)}))\|_2^2\\
&=\frac{2\eta^2}{K}\sum_{k=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}(\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})-\nabla f_{k'}(\mathbf{x}_{t'}^{(k')}))+(\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\nabla f_k(\mathbf{x}_{t'}^{(k)}))\|_2^2\\
&\quad+\frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\sum_{t'=(S_t-s)p}^{t-1}(-\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_{t'}^{(k')})+\nabla f_k(\mathbf{x}_{t'}^{(k)}))\|_2^2\\
&\leq\frac{2\eta^2}{K}\sum_{k=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\mathbb{E}\|\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\nabla f_k(\mathbf{x}_{t'}^{(k)})\|_2^2\\
&\quad+\frac{2\eta^2}{K}\sum_{k=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}(t-(S_t-s)p)\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_{t'}^{(k')})+\nabla f_k(\mathbf{x}_{t'}^{(k)})\|_2^2\\
&\leq 2\eta^2(s+1)p\sigma^2+\frac{2\eta^2(s+1)p}{K}\sum_{t'=t-(s+1)p}^{t-1}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_{t'}^{(k')})+\nabla f_k(\mathbf{x}_{t'}^{(k)})\|_2^2\,,
\end{aligned}
\tag{29}
$$

where the third inequality follows $\frac{1}{K}\sum_{k=1}^{K}\|\frac{1}{K}\sum_{k'=1}^{K}a_{k'}-a_k\|_2^2=\frac{1}{K}\sum_{k=1}^{K}\|a_k\|_2^2-\|\frac{1}{K}\sum_{k=1}^{K}a_k\|_2^2\leq\frac{1}{K}\sum_{k=1}^{K}\|a_k\|_2^2$, and

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_t^{k'})+\nabla f_k(\mathbf{x}_t^{(k)})\|_2^2\\
&=\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f_k(\mathbf{x}_t^{(k)})-\nabla f_k(\tilde{\mathbf{x}}_t)\|_2^2+\|\nabla f_k(\tilde{\mathbf{x}}_t)-\nabla f(\tilde{\mathbf{x}}_t)\|_2^2+\|\nabla f(\tilde{\mathbf{x}}_t)-\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_t^{k'})\|_2^2]\\
&\leq\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}[L^2\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2+\kappa^2+\frac{1}{K}\sum_{k'=1}^{K}\|\nabla f_{k'}(\tilde{\mathbf{x}}_t)-\nabla f_{k'}(\mathbf{x}_t^{k'})\|_2^2]\\
&\leq\frac{6L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2+3\kappa^2\,.
\end{aligned}
\tag{30}
$$

The second term is bounded by

$$
\begin{aligned}
&\frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')}-\frac{1}{K}\sum_{k'=1}^{K}\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k')}+\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2\\
&\leq \frac{2(1+s)}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')}\|_2^2+\frac{2(1+\frac{1}{s})}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k')}+\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2\\
&\leq \frac{2(1+s)}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}^{(k)}\|_2^2+\frac{2(1+\frac{1}{s})}{K}\sum_{k=1}^{K}\mathbb{E}\|\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2\\
&\leq \frac{2(1+s)}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}^{(k)}\|_2^2+\frac{2(1+s)}{K}\sum_{k=1}^{K}\sum_{i=0}^{s-1}\mathbb{E}\|\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2\\
&= \frac{2(s+1)}{K}\sum_{k=1}^{K}\sum_{i=0}^{s}\mathbb{E}\|\mathbf{e}_{(S_t-i-s)p}^{(k)}\|_2^2\leq\frac{24(1-\delta)}{\delta^2}(s+1)^2p^2\eta^2(\sigma^2+\kappa^2+G^2)\,,
\end{aligned}
\tag{31}
$$

where the last inequality follows Lemma 4. Combine the bounds of the first term and the second term, we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\\
&\leq 2\eta^2(s+1)p\sigma^2+2\eta^2(s+1)p\sum_{t'=t-(s+1)p}^{t-1}(\frac{6L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2+3\kappa^2)\\
&\quad+\frac{24(1-\delta)}{\delta^2}(s+1)^2p^2\eta^2(\sigma^2+\kappa^2+G^2)\\
&\leq 2\eta^2(s+1)p\sigma^2(1+\frac{12(1-\delta)}{\delta^2}(s+1)p)+6\eta^2(s+1)^2p^2\kappa^2(1+\frac{4(1-\delta)}{\delta^2})+\frac{24(1-\delta)}{\delta^2}\eta^2(s+1)^2p^2G^2\\
&\quad+12\eta^2L^2(s+1)p\sum_{t'=t-(s+1)p}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\,.
\end{aligned}
\tag{32}
$$

Sum the above inequality from $t=0$ to $t=T-1$ and divide it by T,

$$
\begin{aligned}
&\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\\
&\leq 2\eta^2(s+1)p\sigma^2(1+\frac{12(1-\delta)}{\delta^2}(s+1)p)+6\eta^2(s+1)^2p^2\kappa^2(1+\frac{4(1-\delta)}{\delta^2})+\frac{24(1-\delta)}{\delta^2}\eta^2(s+1)^2p^2G^2\\
&\quad+12\eta^2L^2(s+1)^2p^2\cdot\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\,.
\end{aligned}
\tag{33}
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\\
&\leq\frac{2\eta^2(s+1)p\sigma^2(1+\frac{12(1-\delta)}{\delta^2}(s+1)p)+6\eta^2(s+1)^2p^2\kappa^2(1+\frac{4(1-\delta)}{\delta^2})+\frac{24(1-\delta)}{\delta^2}\eta^2(s+1)^2p^2G^2}{1-12\eta^2L^2(s+1)^2p^2}\,.
\end{aligned}
\tag{34}
$$

If we choose $\eta \leq \frac{1}{6L(s+1)p}$,

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2$$
$$\leq 3\eta^2(s+1)p\sigma^2(1+\frac{12(1-\delta)}{\delta^2}(s+1)p) + 9\eta^2(s+1)^2p^2\kappa^2(1+\frac{4(1-\delta)}{\delta^2}) + \frac{36(1-\delta)}{\delta^2}\eta^2(s+1)^2p^2G^2\,. \tag{35}$$

□

Theorem 1. *For OLCO$_3$-VQ with vanilla SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \min\{\frac{1}{6L(s+1)p}, \frac{1}{9L}\}$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(f(\boldsymbol{x}_0)-f_*)}{\eta T} + \frac{9\eta L\sigma^2}{K} + 12\eta^2L^2(s+1)p\sigma^2(1+\frac{14(1-\delta)}{\delta^2}(s+1)p)$$
$$+ 36\eta^2L^2(s+1)^2p^2\kappa^2(1+\frac{5(1-\delta)}{\delta^2}) + \frac{168(1-\delta)}{\delta^2}\eta^2L^2(s+1)^2p^2G^2\,. \tag{36}$$

If we further choose the learning rate $\eta = \mathcal{O}(K^{\frac{1}{2}}T^{-\frac{1}{2}})$ and the communication interval $p = \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}}(s+1)^{-1})$, then the convergence rate will be $\mathcal{O}(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

Proof. According to Assumption 1,

$$\mathbb{E}_t f(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t) \leq \mathbb{E}_t\langle\nabla f(\tilde{\mathbf{x}}_t), \tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_t\rangle + \frac{L}{2}\mathbb{E}_t\|\tilde{\mathbf{x}}_{t+1}-\tilde{\mathbf{x}}_t\|_2^2$$
$$= -\eta\langle\nabla f(\tilde{\mathbf{x}}_t), \frac{1}{K}\sum_{k=1}^{K}\nabla f_k(\mathbf{x}_t^{(k)})\rangle + \frac{L\eta^2}{2}\mathbb{E}_t\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_k(\mathbf{x}_t^{(k)};\xi_t^{(k)})\|_2^2\,. \tag{37}$$

For the first term,

$$-\langle\nabla f(\tilde{\mathbf{x}}_t), \frac{1}{K}\sum_{k=1}^{K}\nabla f_k(\mathbf{x}_t^{(k)})\rangle = -\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 - \langle\nabla f(\tilde{\mathbf{x}}_t), \frac{1}{K}\sum_{k=1}^{K}(\nabla f_k(\mathbf{x}_t^{(k)}) - \nabla f_k(\tilde{\mathbf{x}}_t))\rangle$$
$$\leq -\frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + \frac{1}{2}\|\frac{1}{K}\sum_{k=1}^{K}(\nabla f_k(\mathbf{x}_t^{(k)}) - \nabla f_k(\tilde{\mathbf{x}}_t))\|_2^2 \tag{38}$$
$$\leq -\frac{1}{2}\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + \frac{L^2}{2K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2\,,$$

where the first equality follows that $\nabla f(\tilde{\mathbf{x}}_t) = \frac{1}{K}\sum_{k=1}^{K}\nabla f_k(\tilde{\mathbf{x}}_t)$. For the second term,

$$\mathbb{E}_t\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_k(\mathbf{x}_t^{(k)};\xi_t^{(k)})\|_2^2$$
$$= \mathbb{E}_t\|\frac{1}{K}\sum_{k=1}^{K}(\nabla F_k(\mathbf{x}_t^{(k)};\xi_t^{(k)}) - \nabla f_k(\mathbf{x}_t^{(k)})) + \frac{1}{K}\sum_{k=1}^{K}(\nabla f_k(\mathbf{x}_t^{(k)}) - \nabla f_k(\tilde{\mathbf{x}}_t)) + \nabla f(\tilde{\mathbf{x}}_t)\|_2^2$$
$$\leq 3\mathbb{E}_t\|\frac{1}{K}\sum_{k=1}^{K}(\nabla F_k(\mathbf{x}_t^{(k)};\xi_t^{(k)}) - \nabla f_k(\mathbf{x}_t^{(k)}))\|_2^2 + 3\|\frac{1}{K}\sum_{k=1}^{K}(\nabla f_k(\mathbf{x}_t^{(k)}) - \nabla f_k(\tilde{\mathbf{x}}_t))\|_2^2 + 3\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2$$
$$\leq \frac{3\sigma^2}{K} + \frac{3L^2}{K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 + 3\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2\,. \tag{39}$$

Combine them and we have

$$\mathbb{E}_t(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t) \leq -\frac{\eta}{2}(1-3\eta L)\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + \frac{\eta L^2}{2}(1+3\eta L)\frac{1}{K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 + \frac{3\eta^2 L\sigma^2}{2K}\,. \tag{40}$$

If we choose $\eta \leq \frac{1}{9L}$,

$$\mathbb{E}_t(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t) \leq -\frac{\eta}{3}\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + \frac{2\eta L^2}{3}\frac{1}{K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 + \frac{3\eta^2 L\sigma^2}{2K}\,. \tag{41}$$

Then for the averaged parameters $\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)}$,

$$\begin{aligned}
\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 &\leq 2\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)}) - \nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 \\
&\leq 2L^2\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)}\|_2^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 \leq \frac{2L^2}{K}\sum_{k=1}^{K}\|\mathbf{e}_{t-sp}^{(k)}\|_2^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 \\
&\leq \frac{6[f(\tilde{\mathbf{x}}_t) - \mathbb{E}_t f(\tilde{\mathbf{x}}_{t+1})]}{\eta} + \frac{9\eta L\sigma^2}{K} + \frac{4L^2}{K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 + \frac{2L^2}{K}\sum_{k=1}^{K}\|\mathbf{e}_{t-sp}^{(k)}\|_2^2\,.
\end{aligned} \tag{42}$$

Take total expectation, sum from $t=0$ to $t=T-1$, and rearrange,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 \\
&\leq \frac{6[f(\tilde{\mathbf{x}}_0) - \mathbb{E}f(\tilde{\mathbf{x}_T})]}{\eta T} + \frac{9\eta L\sigma^2}{K} + \frac{2L^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{t-sp}^{(k)}\|_2^2 + \frac{4L^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 \\
&\leq \frac{6[f(\tilde{\mathbf{x}}_0) - \mathbb{E}f(\tilde{\mathbf{x}_T})]}{\eta T} + \frac{9\eta L\sigma^2}{K} + \frac{24(1-\delta)}{\delta^2}p^2\eta^2L^2(\sigma^2+\kappa^2+G^2) \\
&\quad + 12\eta^2L^2(s+1)p\sigma^2(1+\frac{12(1-\delta)}{\delta^2}(s+1)p) + 36\eta^2L^2(s+1)^2p^2\kappa^2(1+\frac{4(1-\delta)}{\delta^2}) \\
&\quad + \frac{144(1-\delta)}{\delta^2}\eta^2L^2(s+1)^2p^2G^2 \\
&\leq \frac{6[f(\tilde{\mathbf{x}}_0) - \mathbb{E}f(\tilde{\mathbf{x}_T})]}{\eta T} + \frac{9\eta L\sigma^2}{K} + 12\eta^2L^2(s+1)p\sigma^2(1+\frac{14(1-\delta)}{\delta^2}(s+1)p) \\
&\quad + 36\eta^2L^2(s+1)^2p^2\kappa^2(1+\frac{5(1-\delta)}{\delta^2}) + \frac{168(1-\delta)}{\delta^2}\eta^2L^2(s+1)^2p^2G^2\,,
\end{aligned} \tag{43}$$

where the second inequality follows Lemma 4 and 5. □

E PROOF OF THEOREM 2

Lemma 6. *For OLCO$_3$-TC with vanilla SGD and under Assumption 2, 3, 4, and 5, the local error satisfies*

$$\mathbb{E}\|\boldsymbol{e}_t^{(k)}\|_2^2 \leq \frac{12(1-\delta)}{\delta^2}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,. \tag{44}$$

Proof. Same as the proof of Lemma 4, except that $\mathbf{e}_{(S_t-s-1)p}^{(k)}$ is replaced with $\mathbf{e}_{(S_t-1)p}^{(k)}$. □

Lemma 7. *For OLCO$_3$-TC with vanilla SGD and under Assumption 2, 3, 4, and 5, the server error satisfies*

$$\mathbb{E}\|\boldsymbol{e}_t\|_2^2 \leq \frac{96(2-\delta)(1-\delta)}{\delta^4}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,. \tag{45}$$

Proof. Let $S_t = \lfloor \frac{t}{p} \rfloor$,

$$\begin{aligned}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{S_tp}^{(k)})\|_2^2 &\leq 2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{S_tp}^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\Delta_{S_tp}^{(k)}\|_2^2 + 2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\Delta_{S_tp}^{(k)}\|_2^2 \\ &\leq \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathcal{C}(\Delta_{S_tp}^{(k)}) - \Delta_{S_tp}^{(k)}\|_2^2 + \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 \\ &\leq \frac{2(2-\delta)}{K}\sum_{k=1}^{K}\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 \,.\end{aligned} \tag{46}$$

Following the proof of Lemma 4 we have $\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 \leq \frac{3(1+\frac{1}{\rho})}{1-(1-\delta)(1+\rho)}p^2\eta^2(\sigma^2+\kappa^2+G^2)$. Therefore,

$$\begin{aligned}&\mathbb{E}\|\mathbf{e}_t\|_2^2 = \mathbb{E}\|\mathbf{e}_{S_tp}\|_2^2 \leq (1-\delta)\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{S_tp}^{(k)}) + \mathbf{e}_{(S_t-1)p}\|^2 \\ &\leq (1-\delta)(1+\frac{1}{\rho})\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\Delta_{S_tp}^{(k)})\|_2^2 + (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-1)p}\|_2^2 \\ &\leq 2(2-\delta)(1-\delta)(1+\frac{1}{\rho})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 + (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-1)p}\|_2^2 \\ &\leq 2(2-\delta)(1-\delta)(1+\frac{1}{\rho})\frac{3(1+\frac{1}{\rho})}{1-(1-\delta)(1+\rho)}p^2\eta^2(\sigma^2+\kappa^2+G^2) + (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-1)p}\|_2^2 \\ &\leq \frac{6(2-\delta)(1-\delta)(1+\frac{1}{\rho})^2}{[1-(1-\delta)(1+\rho)]^2}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,.\end{aligned} \tag{47}$$

Let $\rho = \frac{\delta}{2(1-\delta)}$ such that $1+\frac{1}{\rho} = \frac{2-\delta}{\delta} \leq \frac{2}{\delta}$, then $\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 \leq \frac{96(2-\delta)(1-\delta)}{\delta^4}p^2\eta^2(\sigma^2+\kappa^2+G^2)$. □

Lemma 8. *For OLCO$_3$-TC with vanilla SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \frac{1}{6L(s+1)p}$ and let $h(\delta) = \frac{1-\delta}{\delta^2}(1+\frac{4(2-\delta)}{\delta^2})$, we have*

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t^{(k)}\|_2^2 \leq 3\eta^2p\sigma^2(s+1+72h(\delta)p) + 9\eta^2p^2\kappa^2((s+1)^2+24h(\delta)) + 216h(\delta)\eta^2p^2G^2\,. \tag{48}$$

Proof. Let $S_t = \lfloor \frac{t}{p} \rfloor$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}(-\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(S_t-i)p}^{(k')}) - \sum_{t'=S_tp}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})) \\ - (-\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(S_t-i)p}^{(k)}) - \sum_{t'=S_tp}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})) - \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_t^{(k')} - \mathbf{e}_{t-sp}\|_2^2\,,\end{aligned} \tag{49}$$

where

$$\begin{aligned}\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(S_t-i)p}^{(k)}) &= \sum_{i=0}^{s-1}[\Delta_{(S_t-i)p}^{(k)} - \mathbf{e}_{(S_t-i)p}^{(k)}] = \sum_{i=0}^{s-1}[\sum_{t'=(S_t-i-1)p}^{(S_t-i)p-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)}) + \mathbf{e}_{(S_t-i-1)p}^{(k)} - \mathbf{e}_{(S_t-i)p}^{(k)}] \\ &= \sum_{i=0}^{s-1}\sum_{t'=(S_t-i-1)p}^{(S_t-i)p-1}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)}) + \mathbf{e}_{(S_t-s)p}^{(k)} - \mathbf{e}_{S_tp}^{(k)}\,.\end{aligned} \tag{50}$$

Therefore

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})+\sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\\
&\quad-\frac{1}{K}\sum_{k'=1}^{K}(\mathbf{e}_{(S_t-s)p}^{(k')}-\mathbf{e}_{S_tp}^{(k')})+(\mathbf{e}_{(S_t-s)p}^{(k)}-\mathbf{e}_{S_tp}^{(k)})-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_t^{(k')}-\mathbf{e}_{t-sp}\|_2^2 \qquad (51)\\
&\le\frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})+\sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2\\
&\quad+\frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')}+\mathbf{e}_{(S_t-s)p}^{(k)}-\mathbf{e}_{S_tp}^{(k)}-\mathbf{e}_{(S_t-s)p}\|_2^2\,,
\end{aligned}
$$

where the first term can be bounded following Eq. (29,30). The second term satisfies

$$
\begin{aligned}
&\frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')}+\mathbf{e}_{(S_t-s)p}^{(k)}-\mathbf{e}_{S_tp}^{(k)}-\mathbf{e}_{(S_t-s)p}\|_2^2\\
&\le\frac{6}{K}\sum_{k=1}^{K}\mathbb{E}\|-\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')}+\mathbf{e}_{(S_t-s)p}^{(k)}\|_2^2+\frac{6}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{S_tp}^{(k)}\|_2^2+\frac{6}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}\|_2^2 \qquad (52)\\
&\le\frac{6}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}^{(k)}\|_2^2+\frac{6}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{S_tp}^{(k)}\|_2^2+\frac{6}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}\|_2^2\\
&\le\frac{1-\delta}{\delta^2}(1+\frac{4(2-\delta)}{\delta^2})\cdot 144p^2\eta^2(\sigma^2+\kappa^2+G^2)\,,
\end{aligned}
$$

where the last inequality follows Lemma 6 and 7. Let $h(\delta)=\frac{1-\delta}{\delta^2}(1+\frac{4(2-\delta)}{\delta^2})$. Combine the above two inequalities and we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\\
&\le 2\eta^2(s+1)p\sigma^2+2\eta^2(s+1)p\sum_{t'=t-(s+1)p}^{t-1}(\frac{6L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2+3\kappa^2)+144h(\delta)p^2\eta^2(\sigma^2+\kappa^2+G^2)\\
&\le 2\eta^2p\sigma^2(s+1+72h(\delta)p)+6\eta^2p^2\kappa^2((s+1)^2+24h(\delta))+144h(\delta)p^2\eta^2G^2\\
&\quad+12\eta^2L^2(s+1)p\sum_{t'=t-(s+1)p}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\,.
\end{aligned} \qquad (53)
$$

Following Eq. (33,34,35),

$$
\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K-1}\mathbb{E}\|\tilde{\mathbf{x}}_t-\mathbf{x}_t^{(k)}\|_2^2\le 3\eta^2p\sigma^2(s+1+72h(\delta)p)+9\eta^2p^2\kappa^2((s+1)^2+24h(\delta))+216h(\delta)\eta^2p^2G^2\,. \qquad (54)
$$

□

Theorem 2. *For OLCO$_3$-TC with vanilla SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \min\{\frac{1}{6L(s+1)p}, \frac{1}{9L}\}$ and let $h(\delta) = \frac{1-\delta}{\delta^2}(1 + \frac{4(2-\delta)}{\delta^2})$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(f(\boldsymbol{x}_0)-f_*)}{\eta T} + \frac{9\eta L\sigma^2}{K} + 12\eta^2 p\sigma^2(s+1+80h(\delta)p) + 12\eta^2p^2\kappa^2(3(s+1)^2+80h(\delta)) + 960\eta^2p^2G^2h(\delta)\,. \tag{55}$$

If we further choose the learning rate $\eta = \mathcal{O}(K^{\frac{1}{2}}T^{-\frac{1}{2}})$ and the communication interval $p = \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}}(s+1)^{-1})$, then the convergence rate will be $\mathcal{O}(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$. For i.i.d. data distributions across workers ($\kappa = 0$), if we choose the learning rate $\eta = \mathcal{O}(K^{\frac{1}{2}}T^{-\frac{1}{2}})$ and the communication interval $p = \min\{\mathcal{O}(K^{-\frac{3}{2}}T^{\frac{1}{2}}(s+1)^{-1}), \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}})\}$ ($p = \mathcal{O}(K^{-\frac{3}{4}}T^{\frac{1}{4}})$ for a enough large T) instead, then the convergence rate will still be $\mathcal{O}(K^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

Proof. Following the proof of Theorem 1, we have the same inequality as Eq. (41):

$$\mathbb{E}_t f(\tilde{\mathbf{x}}_{t+1}) - f(\tilde{\mathbf{x}}_t) \leq -\frac{\eta}{3}\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + \frac{2\eta L^2}{3}\frac{1}{K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 + \frac{3\eta^2 L\sigma^2}{2K}\,. \tag{56}$$

Then for the averaged parameters $\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)}$,

$$\begin{aligned}
&\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 \leq 2\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)}) - \nabla f(\tilde{\mathbf{x}}_t)\|_2^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2\\
&\leq 2L^2\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_t^{(k)} + \mathbf{e}_{t-sp}\|_2^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2\\
&\leq \frac{4L^2}{K}\sum_{k=1}^{K}\|\mathbf{e}_t^{(k)}\|_2^2 + \frac{4L^2}{K}\|\mathbf{e}_{t-sp}\|_2^2 + 2\|\nabla f(\tilde{\mathbf{x}}_t)\|_2^2\\
&\leq \frac{6[f(\tilde{\mathbf{x}}_t) - \mathbb{E}_t f(\tilde{\mathbf{x}}_{t+1})]}{\eta} + \frac{9\eta L\sigma^2}{K} + \frac{4L^2}{K}\sum_{k=1}^{K}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2 + \frac{4L^2}{K}\sum_{k=1}^{K}\|\mathbf{e}_t^{(k)}\|_2^2 + \frac{4L^2}{K}\|\mathbf{e}_{t-sp}\|_2^2\,.
\end{aligned} \tag{57}$$

Take total expectation, sum from $t = 0$ to $t = T-1$, and rearrange,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2\\
&\leq \frac{6[f(\tilde{\mathbf{x}}_0) - \mathbb{E}f(\tilde{\mathbf{x}_T})]}{\eta T} + \frac{9\eta L\sigma^2}{K} + \frac{4L^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{t-sp}^{(k)}\|_2^2 + \frac{4L^2}{KT}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{e}_{t-sp}\|_2^2\\
&\quad + \frac{4L^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t^{(k)}\|_2^2\\
&\leq \frac{6[f(\tilde{\mathbf{x}}_0) - \mathbb{E}f(\tilde{\mathbf{x}_T})]}{\eta T} + \frac{9\eta L\sigma^2}{K} + 12\eta^2p\sigma^2(s+1+72h(\delta)p + \frac{4(1-\delta)}{\delta^2}p + \frac{32(2-\delta)(1-\delta)}{\delta^4}p)\\
&\quad + 12\eta^2p^2\kappa^2(3(s+1)^2 + 72h(\delta) + \frac{4(1-\delta)}{\delta^2} + \frac{32(2-\delta)(1-\delta)}{\delta^4})\\
&\quad + 12\eta^2p^2G^2(72h(\delta) + \frac{4(1-\delta)}{\delta^2} + \frac{32(2-\delta)(1-\delta)}{\delta^4})\\
&\leq \frac{6[f(\tilde{\mathbf{x}}_0) - \mathbb{E}f(\tilde{\mathbf{x}_T})]}{\eta T} + \frac{9\eta L\sigma^2}{K}\\
&\quad + 12\eta^2p\sigma^2(s+1+80h(\delta)p) + 12\eta^2p^2\kappa^2(3(s+1)^2+80h(\delta)) + 960\eta^2p^2G^2h(\delta)\,,
\end{aligned} \tag{58}$$

where the second inequality follows Lemma 6, 7 and 8. □

F PROOF OF THEOREM 3

We first define two virtual variables $\mathbf{z}_t$ and $\mathbf{p}_t$ satisfying

$$\mathbf{p}_t = \begin{cases} \frac{\mu}{1-\mu}(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{t-1}), & t \geq 1 \\ \mathbf{0}, & t = 0 \end{cases} \tag{59}$$

and

$$\mathbf{z}_t = \tilde{\mathbf{x}}_t + \mathbf{p}_t \,. \tag{60}$$

Then the update rule of $\mathbf{z}_t$ satisfies

$$\begin{aligned} \mathbf{z}_{t+1} - \mathbf{z}_t &= (\tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t) + \frac{\mu}{1-\mu}(\tilde{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_t) - \frac{\mu}{1-\mu}(\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{t-1}) \\ &= -\frac{\eta}{K}\sum_{k=1}^{K}\mathbf{m}_{t+1}^{(k)} - \frac{\mu}{1-\mu}\frac{\eta}{K}\sum_{k=1}^{K}\mathbf{m}_{t+1}^{(k)} + \frac{\mu}{1-\mu}\frac{\eta}{K}\sum_{k=1}^{K}\mathbf{m}_{t}^{(k)} \\ &= -\frac{\eta}{(1-\mu)K}\sum_{k=1}^{K}(\mathbf{m}_{t+1}^{(k)} - \mu\mathbf{m}_{t}^{(k)}) \\ &= -\frac{\eta}{(1-\mu)K}\sum_{k=1}^{K}\nabla f_k(\mathbf{x}_t^{(k)};\xi_t^{(k)}) \,, \end{aligned} \tag{61}$$

which exists for OLCO$_3$-OC, OLCO$_3$-VQ and OLCO$_3$-TC.

Lemma 9. *For OLCO$_3$ with Momentum SGD, we have*

$$\mathbb{E}\|\boldsymbol{m}_t^{(k)}\|_2^2 \leq \frac{3(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2} \,. \tag{62}$$

Proof.

$$\begin{aligned} \mathbb{E}\|\mathbf{m}_t^{(k)}\|_2^2 &= \mathbb{E}\|\sum_{t'=0}^{t-1}\mu^{t-1-t'}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2 = (\sum_{t'=0}^{t-1}\mu^{t-1-t'})^2\mathbb{E}\|\sum_{t'=0}^{t-1}\frac{\mu^{t-1-t'}}{\sum_{t'=0}^{t-1}\mu^{t-1-t'}}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2 \\ &\leq (\sum_{t'=0}^{t-1}\mu^{t-1-t'})^2\mathbb{E}\|\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\|_2^2 \leq \frac{3(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2} \,. \end{aligned} \tag{63}$$

□

Lemma 10. *For OLCO$_3$ with Momentum SGD, we have*

$$\mathbb{E}\|\boldsymbol{p}_t\|^2 \leq \frac{3\mu^2\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^4} \,. \tag{64}$$

Proof.

$$\begin{aligned} \mathbb{E}\|\mathbf{p}_t\|^2 &= \frac{\mu^2}{(1-\mu)^2}\mathbb{E}\|\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{t-1}\|^2 = \frac{\mu^2\eta^2}{(1-\mu)^2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{m}_t^{(k)}\|^2 \leq \frac{\mu^2\eta^2}{(1-\mu)^2K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{m}_t^{(k)}\|_2^2 \\ &\leq \frac{3\mu^2\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^4} \,. \end{aligned} \tag{65}$$

□

Lemma 11. *For OLCO$_3$-VQ with Momentum SGD and under Assumption 2, 3, 4, and 5, the local error satisfies*

$$\mathbb{E}\|\boldsymbol{e}_t^{(k)}\|_2^2 \leq \frac{12(1-\delta)}{(1-\mu)^2\delta^2}p^2\eta^2(\sigma^2+\kappa^2+G^2) \,. \tag{66}$$

Proof. Let $S_t = \lfloor \frac{t}{p} \rfloor$,

$$\begin{aligned}
\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 &= \mathbb{E}\|\mathbf{e}_{S_t p}^{(k)}\|_2^2 = \mathbb{E}\|\mathcal{C}(\Delta_{S_t p}^{(k)}) - \Delta_{S_t p}^{(k)}\|_2^2 \leq (1-\delta)\mathbb{E}\|\Delta_{S_t p}^{(k)}\|_2^2 \\
&= (1-\delta)\mathbb{E}\|\sum_{t'=(S_t-1)p}^{S_T p - 1} \eta \mathbf{m}_{t'}^{(k)} + \mathbf{e}_{(S_t-s-1)p}^{(k)}\|_2^2 \\
&\leq (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-s-1)p}^{(k)}\|_2^2 + (1-\delta)(1+\frac{1}{\rho})\frac{3\eta^2 p^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2}\,.
\end{aligned} \tag{67}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 &\leq 3(1-\delta)(1+\frac{1}{\rho})\frac{p^2\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2}\sum_{i=0}^{\lfloor \frac{S_t}{s} \rfloor - 1}[(1-\delta)(1+\rho)]^i \\
&\leq \frac{3(1-\delta)(1+\frac{1}{\rho})}{1-(1-\delta)(1+\rho)}\frac{p^2\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2}\,.
\end{aligned} \tag{68}$$

Let $\rho = \frac{\delta}{2(1-\delta)}$ such that $1+\frac{1}{\rho} = \frac{2-\delta}{\delta} \leq \frac{2}{\delta}$, then $\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 \leq \frac{12(1-\delta)}{(1-\mu)^2\delta^2}p^2\eta^2(\sigma^2+\kappa^2+G^2)$. □

Lemma 12. *For OLCO$_3$-VQ with Momentum SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \frac{1-\mu}{\sqrt{72}L(s+1)p}$, we have*

$$\begin{aligned}
\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\boldsymbol{z}_t - \boldsymbol{x}_t^{(k)}\|_2^2 &\leq \frac{4\eta^2\sigma^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2} + (s+1)p + \frac{12(1-\delta)(s+1)^2p^2}{\delta^2}) \\
&+ \frac{12\eta^2\kappa^2}{(1-\mu)^2}(\frac{1}{(1-\mu)^2} + (s+1)^2p^2 + \frac{4(1-\delta)(s+1)^2p^2}{\delta^2}) + \frac{12\eta^2G^2}{(1-\mu)^2}(\frac{1}{(1-\mu)^2} + \frac{4(1-\delta)(s+1)^2p^2}{\delta^2})\,.
\end{aligned} \tag{69}$$

Proof. Let $S_t = \lfloor \frac{t}{p} \rfloor$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t - \mathbf{x}_t^{(k)}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}(-\sum_{i=0}^{s-1}\Delta_{(S_t-i)p}^{(k')} - (\mathbf{x}_{S_t p}^{(k')} - \mathbf{x}_t^{(k')})) \\
&\quad - (-\sum_{i=0}^{s-1}\Delta_{(S_t-i)p}^{(k)} - (\mathbf{x}_{S_t p}^{(k)} - \mathbf{x}_t^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')}\|_2^2 \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}(\eta\mathbf{m}_{(S_t-s)p}^{(k')}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1} + \sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}) \\
&\quad - (\eta\mathbf{m}_{(S_t-s)p}^{(k)}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1} + \sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_{k}(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}) \\
&\quad + \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')} + \frac{1}{K}\sum_{k'=1}^{K}\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k')} - \sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2 \\
&\leq \frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{m}_{(S_t-s)p}^{(k')}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1} - \mathbf{m}_{(S_t-s)p}^{(k)}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1}\|_2^2 \\
&\quad + \frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\sum_{\tau=0}^{t-1-t'}\mu^{\tau} - \sum_{t'=(S_t-s)p}^{t-1}\nabla F_{k}(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\|_2^2 \\
&\quad + \frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{t-sp}^{(k')} + \frac{1}{K}\sum_{k'=1}^{K}\sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k')} + \sum_{i=0}^{s-1}\mathbf{e}_{(S_t-i-s-1)p}^{(k)}\|_2^2\,.
\end{aligned} \tag{70}$$

The first term

$$\frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{m}_{(S_t-s)p}^{(k')}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1}-\mathbf{m}_{(S_t-s)p}^{(k)}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1}\|_2^2$$
$$\leq \frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{m}_{(S_t-s)p}^{(k)}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1}\|_2^2 \leq \frac{3\eta^2}{(1-\mu)^2K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{m}_{(S_t-s)p}^{(k)}\|_2^2 \leq \frac{9\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^4}\,, \tag{71}$$

where the last inequality follows Lemma 9. Following Eq. (29), the second term can be bounded by

$$\begin{aligned}
&\frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}-\sum_{t'=(S_t-s)p}^{t-1}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\|_2^2\\
&=\frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\sum_{t'=(S_t-s)p}^{t-1}[\frac{1}{K}\sum_{k'=1}^{K}(\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})-\nabla f_{k'}(\mathbf{x}_{t'}^{(k')}))-(\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\nabla f_k(\mathbf{x}_{t'}^{(k)}))]\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\|_2^2\\
&\quad+\frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\sum_{t'=(S_t-s)p}^{t-1}(\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_{t'}^{(k')})-\nabla f_k(\mathbf{x}_{t'}^{(k)}))\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\|_2^2\\
&\leq\frac{3\eta^2}{(1-\mu)^2K}\sum_{k=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\mathbb{E}\|\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})-\nabla f_k(\mathbf{x}_{t'}^{(k)})\|_2^2\\
&\quad+\frac{3\eta^2}{(1-\mu)^2K}(t-(S_t-s)p)\sum_{k=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\nabla f_{k'}(\mathbf{x}_{t'}^{(k')})-\nabla f_k(\mathbf{x}_{t'}^{(k)})\|_2^2\\
&\leq\frac{3\eta^2(s+1)p\sigma^2}{(1-\mu)^2}+\frac{3\eta^2(s+1)p}{(1-\mu)^2}\sum_{t'=t-(s+1)p}^{t-1}(\frac{6L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_{t'}-\mathbf{x}_{t'}^{(k)}\|_2^2+3\kappa^2)\,,
\end{aligned} \tag{72}$$

where the last inequality follows Eq. (30). Combine the bounds of the first and second term with Lemma 11 and Eq. (31),

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2\\
&\leq\frac{9\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^4}+\frac{3\eta^2(s+1)p\sigma^2}{(1-\mu)^2}+\frac{3\eta^2(s+1)p}{(1-\mu)^2}\sum_{t'=t-(s+1)p}^{t-1}(\frac{6L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_{t'}-\mathbf{x}_{t'}^{(k)}\|_2^2+3\kappa^2)\\
&\quad+\frac{3(s+1)}{K}\sum_{k=1}^{K}\sum_{i=0}^{s}\mathbb{E}\|\mathbf{e}_{(S_t-i-s)p}^{(k)}\|_2^2\\
&\leq\frac{9\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^4}+\frac{3\eta^2(s+1)p\sigma^2}{(1-\mu)^2}+\frac{9\eta^2(s+1)^2p^2\kappa^2}{(1-\mu)^2}+\frac{36(1-\delta)\eta^2(s+1)^2p^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2\delta^2}\\
&\quad+\frac{18\eta^2L^2(s+1)p}{(1-\mu)^2}\sum_{t'=t-(s+1)p}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_{t'}-\mathbf{x}_{t'}^{(k)}\|_2^2\,.
\end{aligned} \tag{73}$$

Sum the above inequality from $t=0$ to $t=T-1$ and divide it by T,

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2 \leq \frac{3\eta^2\sigma^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+(s+1)p+\frac{12(1-\delta)(s+1)^2p^2}{\delta^2})$$
$$+\frac{9\eta^2\kappa^2}{(1-\mu)^2}(\frac{1}{(1-\mu)^2}+(s+1)^2p^2+\frac{4(1-\delta)(s+1)^2p^2}{\delta^2})+\frac{9\eta^2G^2}{(1-\mu)^2}(\frac{1}{(1-\mu)^2}+\frac{4(1-\delta)(s+1)^2p^2}{\delta^2})$$
$$+\frac{18\eta^2L^2(s+1)^2p^2}{(1-\mu)^2}\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2\,. \tag{74}$$

If we choose $\eta \leq \frac{1-\mu}{\sqrt{72}L(s+1)p}$,

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2 \leq \frac{4\eta^2\sigma^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+(s+1)p+\frac{12(1-\delta)(s+1)^2p^2}{\delta^2})$$
$$+\frac{12\eta^2\kappa^2}{(1-\mu)^2}(\frac{1}{(1-\mu)^2}+(s+1)^2p^2+\frac{4(1-\delta)(s+1)^2p^2}{\delta^2})+\frac{12\eta^2G^2}{(1-\mu)^2}(\frac{1}{(1-\mu)^2}+\frac{4(1-\delta)(s+1)^2p^2}{\delta^2})\,. \tag{75}$$

□

Theorem 3. *For OLCO$_3$-VQ with Momentum SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \min\{\frac{1-\mu}{\sqrt{72}L(s+1)p}, \frac{1-\mu}{9L}\}$ and let $g(\mu,\delta,s,p)=\frac{15}{(1-\mu)^2}+\frac{60(1-\delta)(s+1)^2p^2}{\delta^2}$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq \frac{6(1-\mu)[f(\boldsymbol{z}_0)-f(\boldsymbol{z}_T)]}{\eta}+\frac{9L\eta\sigma^2}{(1-\mu)K}$$
$$+\frac{4\eta^2L^2}{(1-\mu)^2}[(4(s+1)p+g(\mu,\delta,s,p))\sigma^2+(12(s+1)^2p^2+g(\mu,\delta,s,p))\kappa^2+g(\mu,\delta,s,p)G^2]\,. \tag{76}$$

The same convergence rate and communication complexity are achieved as Theorem 1.

Proof. Following the proof of Theorem 1 and the update rule Eq. (61), we have a similar inequality as Eq. (41) by choosing $\eta \leq \frac{1-\mu}{9L}$:

$$\mathbb{E}_t f(\mathbf{z}_{t+1})-f(\mathbf{z}_t)$$
$$\leq \frac{\eta}{1-\mu}(-\frac{1}{2}\|\nabla f(\mathbf{z}_t)\|_2^2+\frac{L^2}{2K}\sum_{k=1}^{K}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2)+\frac{L\eta^2}{2(1-\mu)^2}(\frac{3\sigma^2}{K}+\frac{3L^2}{K}\sum_{k=1}^{K}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2+3\|\nabla f(\mathbf{z}_t)\|_2^2)$$
$$=-\frac{\eta}{2(1-\mu)}(1-\frac{3L\eta}{1-\mu})\|\nabla f(\mathbf{z}_t)\|_2^2+\frac{L^2\eta}{2(1-\mu)K}(1+\frac{3L\eta}{1-\mu})\sum_{k=1}^{K}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2+\frac{3L\eta^2\sigma^2}{2(1-\mu)^2K}$$
$$\leq -\frac{\eta}{3(1-\mu)}\|\nabla f(\mathbf{z}_t)\|_2^2+\frac{2\eta L^2}{3(1-\mu)K}\sum_{k=1}^{K}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2+\frac{3L\eta^2\sigma^2}{2(1-\mu)^2K}\,. \tag{77}$$

Then for the averaged parameters $\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)}$,

$$\begin{aligned}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 &\leq 2\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})-\nabla f(\mathbf{z}_t)\|_2^2+2\|\nabla f(\mathbf{z}_t)\|_2^2\\ &\leq 2L^2\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)}-\mathbf{p}_t\|_2^2+2\|\nabla f(\mathbf{z}_t)\|_2^2\\ &\leq 4L^2\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)}\|_2^2+4L^2\|\mathbf{p}_t\|_2^2+2\|\nabla f(\mathbf{z}_t)\|_2^2\,.\end{aligned} \tag{78}$$

Therefore

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 \leq \frac{6(1-\mu)[f(\mathbf{z}_0)-f(\mathbf{z}_T)]}{\eta} + \frac{9L\eta\sigma^2}{(1-\mu)K} + \frac{4L^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2$$
$$+\frac{4L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)}\|_2^2 + \frac{4L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{p}_t\|_2^2$$
$$\leq \frac{6(1-\mu)[f(\mathbf{z}_0)-f(\mathbf{z}_T)]}{\eta} + \frac{9L\eta\sigma^2}{(1-\mu)K}$$
$$+\frac{4\eta^2L^2}{(1-\mu)^2}[(4(s+1)p+g(\mu,\delta,s,p))\sigma^2+(12(s+1)^2p^2+g(\mu,\delta,s,p))\kappa^2+g(\mu,\delta,s,p)G^2]\,. \tag{79}$$

where the last inequality follows Lemma 10, 11 and 12 and $g(\mu,\delta,s,p)=\frac{15}{(1-\mu)^2}+\frac{60(1-\delta)(s+1)^2p^2}{\delta^2}$.

□

G PROOF OF THEOREM 4

Lemma 13. *For OLCO$_3$-TC with Momentum SGD and under Assumption 2, 3, 4, and 5, the local error satisfies*

$$\mathbb{E}\|\boldsymbol{e}_t^{(k)}\|_2^2 \leq \frac{12(1-\delta)}{(1-\mu)^2\delta^2}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,. \tag{80}$$

Proof. Same as the proof of Lemma 11, except that $\mathbf{e}_{(S_t-s-1)p}^{(k)}$ is replaced with $\mathbf{e}_{(S_t-1)p}^{(k)}$. □

Lemma 14. *For OLCO$_3$-TC with vanilla SGD and under Assumption 2, 3, 4, and 5, the server error satisfies*

$$\mathbb{E}\|\boldsymbol{e}_t\|_2^2 \leq \frac{96(2-\delta)(1-\delta)}{(1-\mu)^2\delta^4}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,. \tag{81}$$

Proof. Let $S_t = \lfloor \frac{t}{p} \rfloor$. Following the proof of Lemma 11 we have $\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 \leq \frac{3(1+\frac{1}{\rho})}{1-(1-\delta)(1+\rho)}\frac{p^2\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2}$. Therefore,

$$\mathbb{E}\|\mathbf{e}_t\|_2^2 \leq 2(2-\delta)(1-\delta)(1+\frac{1}{\rho})\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\Delta_{S_tp}^{(k)}\|_2^2 + (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-1)p}\|_2^2$$
$$\leq 2(2-\delta)(1-\delta)(1+\frac{1}{\rho})\frac{3(1+\frac{1}{\rho})}{1-(1-\delta)(1+\rho)}\frac{p^2\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^2} + (1-\delta)(1+\rho)\mathbb{E}\|\mathbf{e}_{(S_t-1)p}\|_2^2$$
$$\leq \frac{6(2-\delta)(1-\delta)(1+\frac{1}{\rho})^2}{[1-(1-\delta)(1+\rho)]^2(1-\mu)^2}p^2\eta^2(\sigma^2+\kappa^2+G^2)\,, \tag{82}$$

where the first inequality follows the proof of Lemma 7. Let $\rho=\frac{\delta}{2(1-\delta)}$ such that $1+\frac{1}{\rho}=\frac{2-\delta}{\delta}\leq\frac{2}{\delta}$, then $\mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 \leq \mathbb{E}\|\mathbf{e}_t^{(k)}\|_2^2 \leq \frac{96(2-\delta)(1-\delta)}{(1-\mu)^2\delta^4}p^2\eta^2(\sigma^2+\kappa^2+G^2)$. □

Lemma 15. *For OLCO$_3$-TC with Momentum SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \frac{1-\mu}{\sqrt{72}L(s+1)p}$, we have*

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\boldsymbol{z}_t-\boldsymbol{x}_t^{(k)}\|_2^2 \leq \frac{3\eta^2\sigma^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+(s+1)p+72h(\delta)p^2)$$
$$+\frac{3\eta^2\kappa^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+3(s+1)^2p^2+72h(\delta)p^2)+\frac{3\eta^2G^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+72h(\delta)p^2)\,, \tag{83}$$

where $h(\delta) = \frac{1-\delta}{\delta^2}(1 + \frac{4(2-\delta)}{\delta^2})$.

Proof. Let $S_t = \lfloor \frac{t}{p} \rfloor$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t - \mathbf{x}_t^{(k)}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}(-\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(S_t-i)p}^{(k')}) - (\mathbf{x}_{S_tp}^{(k')} - \mathbf{x}_t^{(k')}))\\ - (-\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(S_t-i)p}^{(k)}) - (\mathbf{x}_{S_tp}^{(k')} - \mathbf{x}_t^{(k)}) - \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_t^{(k)} - \mathbf{e}_{t-sp}\|_2^2\,,\end{aligned} \tag{84}$$

where

$$\begin{aligned}&\sum_{i=0}^{s-1}\mathcal{C}(\Delta_{(S_t-i)p}^{(k)}) + (\mathbf{x}_{S_tp}^{(k)} - \mathbf{x}_t^{(k)}) = \sum_{i=0}^{s-1}[\Delta_{(S_t-i)p}^{(k)} - \mathbf{e}_{(S_t-i)p}^{(k)}] + (\mathbf{x}_{S_tp}^{(k)} - \mathbf{x}_t^{(k)})\\ &= \sum_{i=0}^{s-1}[\mathbf{m}_{(S_t-i-1)p}^{(k)}\sum_{\tau=0}^{p-1}\mu^{\tau+1} + \sum_{t'=(S_t-i-1)p}^{(S_t-i)p-1}\eta\nabla F(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{(S_t-i)p-1-t'}\mu^{\tau} + \mathbf{e}_{(S_t-i-1)p}^{(k)} - \mathbf{e}_{(S_t-i)p}^{(k)}]\\ &\quad + \mathbf{m}_{S_tp}^{(k)}\sum_{\tau=0}^{t-1-S_tp}\mu^{\tau+1} + \sum_{t'=S_tp}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\\ &= \mathbf{m}_{(S_t-s)p}^{(k)}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1} + \sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau} + \mathbf{e}_{(S_t-s)p}^{(k)} - \mathbf{e}_{S_tp}^{(k)}\,.\end{aligned} \tag{85}$$

Therefore,

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t - \mathbf{x}_t^{(k)}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\eta\mathbf{m}_{(S_t-s)p}^{(k')}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1} - \eta\mathbf{m}_{(S_t-s)p}^{(k)}\sum_{\tau=0}^{t-1-(S_t-s)p}\mu^{\tau+1}\\ &\quad + \frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\sum_{\tau=0}^{t-1-t'}\mu^{\tau} - \sum_{t'=(S_t-s)p}^{t-1}\eta\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\\ &\quad + \frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')} - \mathbf{e}_{(S_t-s)p}^{(k)} + \mathbf{e}_{S_tp}^{(k)} + \mathbf{e}_{t-sp}\|_2^2\\ &= \frac{3\eta^2}{(1-\mu)^2K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{m}_{(S_t-s)p}^{(k)}\|_2^2 + \frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')} - \mathbf{e}_{(S_t-s)p}^{(k)} + \mathbf{e}_{S_tp}^{(k)} + \mathbf{e}_{t-sp}\|_2^2\\ &\quad + \frac{3\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\sum_{t'=(S_t-s)p}^{t-1}\nabla F_{k'}(\mathbf{x}_{t'}^{(k')};\xi_{t'}^{(k')})\sum_{\tau=0}^{t-1-t'}\mu^{\tau} - \sum_{t'=(S_t-s)p}^{t-1}\nabla F_k(\mathbf{x}_{t'}^{(k)};\xi_{t'}^{(k)})\sum_{\tau=0}^{t-1-t'}\mu^{\tau}\|_2^2\,,\end{aligned} \tag{86}$$

where the first term is bounded following Lemma 9 and the third term is bounded following Eq. (72). The second term

$$\begin{aligned}&\frac{3}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}\mathbf{e}_{(S_t-s)p}^{(k')} - \mathbf{e}_{(S_t-s)p}^{(k)} + \mathbf{e}_{S_tp}^{(k)} + \mathbf{e}_{t-sp}\|_2^2\\ &\le \frac{9}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}^{(k)}\|_2^2 + \frac{9}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{S_tp}^{(k)}\|_2^2 + \frac{9}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{e}_{(S_t-s)p}\|_2^2\\ &\le \frac{1-\delta}{(1-\mu)^2\delta^2}(1 + \frac{4(2-\delta)}{\delta^2})\cdot 216p^2\eta^2(\sigma^2 + \kappa^2 + G^2)\,.\end{aligned} \tag{87}$$

Combine these bounds,

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2 \leq \frac{9\eta^2(\sigma^2+\kappa^2+G^2)}{(1-\mu)^4}+\frac{1-\delta}{(1-\mu)^2\delta^2}(1+\frac{4(2-\delta)}{\delta^2})\cdot 216p^2\eta^2(\sigma^2+\kappa^2+G^2)$$
$$+\frac{3\eta^2(s+1)p\sigma^2}{(1-\mu)^2}+\frac{3\eta^2(s+1)p}{(1-\mu)^2}\sum_{t'=t-(s+1)p}^{t-1}(\frac{6L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_{t'}-\mathbf{x}_{t'}^{(k)}\|_2^2+3\kappa^2)$$
$$=\frac{18\eta^2L^2(s+1)p}{(1-\mu)^2}\sum_{t'=t-(s+1)p}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_{t'}-\mathbf{x}_{t'}^{(k)}\|_2^2 \tag{88}$$

Sum the above inequality from $t=0$ to $t=T-1$, divide it by T, and choose $\eta \leq \frac{1-\mu}{\sqrt{72}L(s+1)p}$,

$$\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2 \leq \frac{3\eta^2\sigma^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+(s+1)p+72h(\delta)p^2)$$
$$+\frac{3\eta^2\kappa^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+3(s+1)^2p^2+72h(\delta)p^2)+\frac{3\eta^2G^2}{(1-\mu)^2}(\frac{3}{(1-\mu)^2}+72h(\delta)p^2)\,. \tag{89}$$

□

Theorem 4. *For OLCO$_3$-TC with Momentum SGD and under Assumption 1, 2, 3, 4, and 5, if the learning rate $\eta \leq \min\{\frac{1-\mu}{\sqrt{72}L(s+1)p}, \frac{1-\mu}{9L}\}$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}_t^{(k)})\|_2^2 \leq$$
$$\leq \frac{6(1-\mu)[f(\boldsymbol{z}_0)-f(\boldsymbol{z}_T)]}{\eta}+\frac{9L\eta\sigma^2}{(1-\mu)K}+\frac{6\eta^2L^2}{(1-\mu)^2}[\sigma^2(\frac{9}{(1-\mu)^2}+2(s+1)p+168h(\delta)p^2)$$
$$+\kappa^2(\frac{9}{(1-\mu)^2}+6(s+1)^2p^2+168h(\delta)p^2)+G^2(\frac{9}{(1-\mu)^2}+168h(\delta)p^2)]\,, \tag{90}$$

where $h(\delta)=\frac{1-\delta}{\delta^2}(1+\frac{4(2-\delta)}{\delta^2})$. The same convergence rate and communication complexity are achieved as Theorem 2.

Proof. Following the proof of Theorem 3,

$$\mathbb{E}_t f(\mathbf{z}_{t+1})-f(\mathbf{z}_t) \leq -\frac{\eta}{3(1-\mu)}\|\nabla f(\mathbf{z}_t)\|_2^2+\frac{2\eta L^2}{3(1-\mu)K}\sum_{k=1}^{K}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2+\frac{3L\eta^2\sigma^2}{2(1-\mu)^2K}\,, \tag{91}$$

$$\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 \leq 2\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})-\nabla f(\mathbf{z}_t)\|_2^2+2\|\nabla f(\mathbf{z}_t)\|_2^2$$
$$\leq 2L^2\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_t^{(k)}+\mathbf{e}_{t-sp}-\mathbf{p}_t\|_2^2+2\|\nabla f(\mathbf{z}_t)\|_2^2 \tag{92}$$
$$\leq 6L^2\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)}\|_2^2+6L^2\|\mathbf{e}_{t-sp}\|_2^2+6L^2\|\mathbf{p}_t\|_2^2+2\|\nabla f(\mathbf{z}_t)\|_2^2\,.$$

Therefore

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_t^{(k)})\|_2^2 \leq \frac{6(1-\mu)[f(\mathbf{z}_0)-f(\mathbf{z}_T)]}{\eta} + \frac{9L\eta\sigma^2}{(1-\mu)K} + \frac{4L^2}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{z}_t-\mathbf{x}_t^{(k)}\|_2^2 \\
&\quad + \frac{6L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t-sp}^{(k)}\|_2^2 + \frac{6L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{e}_{t-sp}\|_2^2 + \frac{6L^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\mathbf{p}_t\|_2^2 \\
&\leq \frac{6(1-\mu)[f(\mathbf{z}_0)-f(\mathbf{z}_T)]}{\eta} + \frac{9L\eta\sigma^2}{(1-\mu)K} + \frac{6\eta^2L^2}{(1-\mu)^2}[\sigma^2(\frac{9}{(1-\mu)^2}+2(s+1)p+168h(\delta)p^2) \\
&\quad + \kappa^2(\frac{9}{(1-\mu)^2}+6(s+1)^2p^2+168h(\delta)p^2) + G^2(\frac{9}{(1-\mu)^2}+168h(\delta)p^2)]\,,
\end{aligned} \tag{93}$$

where the last inequality follows Lemma 10, 13, 14 and 15. □