B-ary Tree Push-Pull Method is Provably Efficient for Distributed Learning on Heterogeneous Data

Runze You[1]   **Shi Pu**[1*]

[1]The Chinese University of Hong Kong, Shenzhen

runzeyou@link.cuhk.edu.cn, pushi@cuhk.edu.cn

Abstract

This paper considers the distributed learning problem where a group of agents cooperatively minimizes the summation of their local cost functions based on peer-to-peer communication. Particularly, we propose a highly efficient algorithm, termed "B-ary Tree Push-Pull" (BTPP), that employs two B-ary spanning trees for distributing the information related to the parameters and stochastic gradients across the network. The simple method is efficient in communication since each agent interacts with at most $(B + 1)$ neighbors per iteration. More importantly, BTPP achieves linear speedup for smooth nonconvex and strongly convex objective functions with only $\tilde{O}(n)$ and $\tilde{O}(1)$ transient iterations, respectively, significantly outperforming the state-of-the-art results to the best of our knowledge. Our code is available at https://github.com/ryou98/BTPP.

1   Introduction

In this paper, we consider a group of agents, labeled as $\mathcal{N} := \{1, 2, \ldots, n\}$, in which each agent i holds its own local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$ and communicates only within its direct neighborhood. We investigate how the agents collaborate to locate $x \in \mathbb{R}^p$ that minimizes the average of all the cost functions:

$$\min_{x \in \mathbb{R}^p} f(x) \left(= \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right), \tag{1}$$

where $f_i(x) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [F_i(x; \xi_i)]$. Here ξ_i denotes the local data of agent i that follows the local distribution $\mathcal{D}_i$. Data heterogeneity exists if $\{\mathcal{D}_i\}_{i=1}^{n}$ are not identical.

To solve problem (1), we assume each agent i queries a stochastic oracle ($\mathcal{SO}$) to obtain noisy gradient samples. Stochastic gradients appear in many areas including online distributed learning [23, 3], reinforcement learning [17, 15], generative modeling [5, 6], and parameter estimation [2, 27]. Assumption 1.1 ensures that the gradient estimator $g_i(x; \xi_i)$ remains unbiased with a bounded variance for any given **x**, while independent samples ξ_i are gathered continuously over time. In addition, the assumption is critical in simulation-based optimization as gradient estimation often encounters noise from multiple sources, such as modeling and discretization errors, or limitations due to finite sample sizes in Monte-Carlo methods [7].

Modern optimization and machine learning typically involve tremendous data samples and model parameters. The scale of these problems calls for efficient distributed algorithms across multiple computing nodes. Recently, distributed algorithms dealing with problem (1) have been studied extensively in the literature; see, e.g., [19, 14, 4, 34]. Traditional distributed learning approaches typically follow a centralized master-worker setup, where each worker node communicates with a (virtual) central server [12]. However, such a communication pattern incurs significant communication overheads and long latency, especially when the training requires a large number of computing nodes.

---

*Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Decentralized learning is an emerging paradigm to save communication costs, where the computing nodes are connected through a certain network topology (e.g., ring, grid, hypercube). Decentralized algorithms do not rely on central servers: the agents maintain the similarity among their copies of model parameters through peer-to-peer messages passing by communicating locally with immediate neighbors in the network. Such a setup allows each node to communicate with only a few peers and hence incurs much lower communication overhead [1]. Moreover, it offers strong promise for new applications, allowing agents to collaboratively train a model while respecting the data locality and privacy of each contributor.

Specifically, in decentralized stochastic gradient methods, the agents share their local stochastic gradient updates through gossip communication [32]. At every iteration, the local updates are sent to the neighbors of each agent who iteratively propagate the information through the network. Typically, the agents employ iterative gossip averaging of their neighbors' models with their own, where the averaging weights are chosen to ensure asymptotic uniform distribution of each update across the network. However, local averaging is less effective in "mixing" information which makes decentralized algorithms converge slower than their centralized counterparts. Generally speaking, the network topology determines both the number of per-iteration communications and the convergence rates of decentralized algorithms, leading to a trade-off. For example, a densely-connected graph enables decentralized methods to converge faster but results in less efficient communication since each node needs to communicate with more neighbors. By contrast, a sparsely-connected topology results in a slower convergence rate but also reduces the per-iteration communication cost [19, 21, 35]. In particular, for smooth and non-convex objective functions, it has been shown that decentralized stochastic gradient methods (with arbitrary topology) can achieve the same convergence rate as the centralized SGD method, but only after an initial period of iterations has passed [14, 34, 20]. The number of transient iterations (transient time) heavily depends on the network topology, and thus a practical decentralized stochastic gradient algorithm should aim to minimize the transient time while keeping the number of per-iteration communications small (e.g., over a a sparsely-connected topology). Such an observation has motivated several recent works, which consider network topologies with $\Theta(1)$ per-iteration communications (or degree) for each node; see, e.g., [34, 26].

This work considers an alternative mechanism to gossip averaging, called "B-ary Tree Push-Pull" (BTPP), inherited from the Push-Pull method in [22, 33]. Rather than relaying the messages over one graph at every iteration, BTPP uses two B-ary trees ($\mathcal{G}_\mathcal{R}$ and $\mathcal{G}_\mathcal{C}$) to spread the information about the parameters and the stochastic gradients, respectively. Each agent assigned in the B-ary tree acts as a worker on an assembly line. The model parameters are transmitted through the graph $\mathcal{G}_\mathcal{R}$ from the parent nodes to the child nodes. Meanwhile, the stochastic gradients are computed under the current model parameters and accumulated through the inverse graph of $\mathcal{G}_\mathcal{R}$ denoted as $\mathcal{G}_\mathcal{C}$. BTPP can be viewed as a semi-(de)centralized approach given the critical role of node 1. Notably, the corresponding mixing matrices of $\mathcal{G}_\mathcal{R}$ and $\mathcal{G}_\mathcal{C}$ only consist of 0's and 1's, which together with the B-ary Tree topology design, results in high algorithmic efficiency. We show BTPP achieves an $\tilde{\mathcal{O}}(n)$ transient time under smooth nonconvex objective functions with $\Theta(1)$ per-iteration communications for each agent. By comparison, the state-of-the-art transient time is $\mathcal{O}(n^3)$ (see Table 1).

1.1   Related Works

Decentralized Learning   Decentralized Stochastic Gradient Descent (DSGD) type algorithms are increasingly popular for accelerating the training of large-scale machine learning models [14, 34, 10] . These algorithms have been adapted under a range of practical settings, including those discussed in [1, 16]. However, DSGD suffers from data heterogeneity [9], which triggers more advanced techniques such as EXTRA [25], Exact-Diffusion/D^2 [13], and gradient tracking [19]. The Push-Pull method [22, 33] which enjoys broad topological requirements was introduced for deterministic decentralized optimization under strongly convex objectives. This work particularly takes advantage of the flexibility in the network design of Push-Pull, utilizing the B-ary tree family, while considering stochastic gradients for minimizing smooth nonconvex objectives.

Topology Design   Decentralized stochastic gradient algorithms often rely on gossip averaging over various topologies such as rings, grids, and tori [18]. The hypercube graph [30] strikes a balance between the communication efficiency and the consensus rate, but the network size is constrained to be the power of two. The work in [34] re-examined the static exponential graph with $\Theta(\ln(n))$

2

ALGORITHM	PER-ITER COMM.	SIZE n	BASED GRAPH	TRANS. ITER.
D^2 (RING) [29]	$\Theta(1)$	ARBITRARY	1	$\mathcal{O}(n^{11})$
DSGD (RING) [18]	$\Theta(1)$	ARBITRARY	1	$\mathcal{O}(n^7)$
HYPERCUBE [30]	$\Theta(\ln(n))$	POWER OF 2	1	$\tilde{\mathcal{O}}(n^3)$
STATIC EXP. [34]	$\Theta(\ln(n))$	ARBITRARY	1	$\tilde{\mathcal{O}}(n^3)$
O.-P. EXP. [34]	1	POWER OF 2	$\Theta(\ln(n))$	$\tilde{\mathcal{O}}(n^3)$
RELAYSGD [31]	$\Theta(1)$	ARBITRARY	1	$\mathcal{O}(n^3)$
OD(OU)-EQUIDYN [26]	1	ARBITRARY	$\Theta(n)$	$\mathcal{O}(n^3)$
DSGD-CECA [4]	$\Theta(1)$	ARBITRARY	$\Theta(\ln(n))$	$\tilde{\mathcal{O}}(n^3)$
BASE-$(k+1)$ [28]	$\Theta(1)$	ARBITRARY	$\Theta(\ln(n))$	$\tilde{\mathcal{O}}(n^3)$
BTPP (OURS)	$\mathbf{\Theta(1)}$	**ARBITRARY**	**2**	$\tilde{\mathcal{O}}(n)$

Table 1: Comparison of different algorithms for distributed stochastic optimization under smooth nonconvex objectives. "Per-iter Comm." denotes the number of per-iteration communications or neighbors (degree) for each agent. "Based Graph" represents the number of required graph topologies during the entire training procedure. "Trans. Iter." represents the number of transient iterations. The notation $\tilde{\mathcal{O}}(\cdot)$ hides all the polylogarithmic factors.

ALGORITHM	PER-ITER COMM.	SIZE n	BASED GRAPH	TRANS. ITER.
DSGD (RING) [18]	$\Theta(1)$	ARBITRARY	1	$\tilde{\mathcal{O}}(n^5)$
STATIC EXP. [34]	$\Theta(\ln(n))$	ARBITRARY	1	$\tilde{\mathcal{O}}(n)$
O.-P. EXP. [34]	1	POWER OF 2	$\Theta(\ln(n))$	$\tilde{\mathcal{O}}(n)$
RELAYSGD [31]	$\Theta(1)$	ARBITRARY	1	$\tilde{\mathcal{O}}(n)$
OD(OU)-EQUIDYN [26]	1	ARBITRARY	$\Theta(n)$	$\tilde{\mathcal{O}}(n)$
BTPP (OURS)	$\mathbf{\Theta(1)}$	**ARBITRARY**	**2**	$\tilde{\mathcal{O}}(1)$

Table 2: Comparison of different algorithms for distributed stochastic optimization under smooth strongly convex objectives. The notation $\tilde{\mathcal{O}}(\cdot)$ hides all the polylogarithmic factors inheriting from [26, 9].

degree and introduced a one-peer exponential graph with $\Theta(1)$ degree while preserving the consensus properties under the specific requirement of n. The paper [28] proposed a base-$(k + 1)$ graph as an enhancement that achieves similar convergence rate as in [34] under arbitrary network size by sequentially employing multiple graph topologies (splitting an all-connected graph into $\Theta(\ln(n))$ different subgraphs). DSGD-CECA [4] requires roughly $\lceil \log_2(n) \rceil$ rounds of message passing for global averaging with $\Theta(n)$ network topologies. OD(OU)-EquiDyn [26] introduces algorithms that employ various topologies to achieve network-size independent consensus rates. RelaySGD [31] offers a relay-based algorithm that ensures $\Theta(1)$ per-iteration communication across different topologies.

The above-mentioned methods all enjoy comparable convergence rates with centralized SGD (and thus achieves linear speedup) when the number of iterations T is large enough. The transient times are generally in the order of $\tilde{\mathcal{O}}(n^3)$ under smooth nonconvex objectives (see Table 1) and $\tilde{\mathcal{O}}(n)$ under smooth strongly convex objectives (see Table 2).

Note that the above works and this paper generally consider training machine learning modes within high-performance data-center clusters, in which the network topology can be fully controlled: any two nodes can directly communicate over the network when necessary. By comparison, in some other scenarios, the underlying network topology is fixed, and the communication between two nodes is constrained (e.g., in wireless sensor networks, internet of vehicles, etc).

1.2 Main Contribution

This paper introduces a novel distributed stochastic gradient algorithm, termed "B-ary Tree Push-Pull" (BTPP), which is provably efficient for solving the distributed learning problem (1) under arbitrary network size. The main contribution is summarized as follows:

3

- BTPP incurs a $\Theta(1)$ communication overhead per-iteration for each agent. Specifically, any agent in the network communicates with at most $(B + 1)$ neighbors, where B can be freely chosen to fit different settings. Generally speaking, larger B increases the per-iteration communication cost but reduces the transient time at the same time.

- We show BTPP enjoys an $\tilde{\mathcal{O}}(n)$ transient time or iteration complexity under smooth non-convex objectives and an $\tilde{\mathcal{O}}(1)$ transient time or iteration complexity under smooth strongly convex objectives. Such results outperform the baselines: see Table 1 and Table 2. The improvement is significant since the transient time greatly impacts the algorithmic performance, especially under large n.

- The convergence analysis for BTPP is non-trivial, partly due to the fact that the algorithm admits two different network topologies for communicating the model parameters and the (stochastic) gradient trackers respectively. Instead of constructing the induced matrix norms $\|\cdot\|_{\mathcal{R}}$ and $\|\cdot\|_{\mathcal{C}}$ as in [22], the analysis is performed under $\|\cdot\|_2$ and $\|\cdot\|_F$ only by carefully treating the matrix products and related terms.

1.3 Notation and Preliminaries

Throughout the paper, vectors default to columns if not otherwise specified. Let each agent i hold a local copy $x_i \in \mathbb{R}^p$ of the decision variable and an auxiliary variable $y_i \in \mathbb{R}^p$. Their values at iteration k are denoted by $x_i^{(k)}$ and $y_i^{(k)}$, respectively. We let $\mathbf{X} = [x_1, x_2, \cdots, x_n]^\top \in \mathbb{R}^{n \times p}$, $\mathbf{Y} = [y_1, y_2, \cdots, y_n]^\top \in \mathbb{R}^{n \times p}$, and $\mathbf{1}$ denotes the column vector with all entries equal to 1. We also define the aggregated gradients at the local variables as $\nabla F(\mathbf{X}) := [\nabla f_1(x_1), \nabla f_2(x_2), \cdots, \nabla f_n(x_n)]^\top \in \mathbb{R}^{n \times p}$, where $F(\mathbf{X}) := \sum_{i=1}^n f_i(x_i)$. In addition, denote $\boldsymbol{\xi} := [\xi_1, \xi_2, \cdots, \xi_n]^\top$, $\mathbf{G}(\mathbf{X}, \boldsymbol{\xi}) := [g_1(x_1, \xi_1), g_2(x_2, \xi_2), \cdots, g_n(x_n, \xi_n)]^\top \in \mathbb{R}^{n \times p}$. For the conciseness of presentation, we also use $\mathbf{G}^{(t)}$ to represent $\mathbf{G}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)})$. The term $\langle a, b \rangle$ stands for the inner product of two vectors $a, b \in \mathbb{R}^p$. For matrices, $\|\cdot\|_2$ and $\|\cdot\|_F$ represent the spectral norm and the Frobenius norm respectively, which degenerate to the Euclidean norm for vectors. For simplicity, any square matrix with power 0 is the unit matrix $\mathbf{I}$ with the same dimension if not otherwise specified.

We assume each agent i is able to obtain noisy gradient samples of the form $g_i(x, \xi_i)$ that satisfies the following assumption.

Assumption 1.1. For all $i \in \mathcal{N}$ and $x \in \mathbb{R}^p$, each random vector ξ_i is independent and

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[g_i(x, \xi_i)|x\right] = \nabla f_i(x), \ \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\|g_i(x, \xi_i) - \nabla f_i(x)\|^2 \big| x\right] \leq \sigma^2$$

for some $\sigma^2 > 0$.

Regarding the individual objective functions f_i, we make the following standard assumption.

Assumption 1.2. Each $f_i(x) : \mathbb{R}^p \to \mathbb{R}$ is lower bounded with L-Lipschitz continuous gradients, i.e., for any $x, x' \in \mathbb{R}^p$,

$$\|\nabla f_i(x) - \nabla f_i(x')\| \leq L \|x - x'\|.$$

We also consider the following standard assumption regarding strongly convexity.

Assumption 1.3. For any $x, y \in \mathbb{R}^p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2.$$

Denote $f^* := \min_{x \in \mathbb{R}^p} f(x)$. Let $x^* = \arg\min_x f(x)$ if Assumption 1.3 holds.

A directed graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ consists of a set of n nodes $\mathcal{N}$ and a set of directed edges $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$, where an edge $(j, i) \in \mathcal{E}$ indicates that node j can directly send information to node i. To facilitate the local averaging procedure, each graph can be associated with a non-negative weight matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$, whose element w_{ij} is non-zero only if $(j, i) \in \mathcal{E}$. Similarly, a non-negative weight matrix $\mathbf{W}$ corresponds to a directed graph denoted by $\mathcal{G}_{\mathbf{W}}$. For a given graph $\mathcal{G}_{\mathbf{W}}$, the in-neighborhood and out-neighborhood of node $i \in \mathcal{N}$ are given by $\mathcal{N}_{\mathbf{W},i}^{in} := \{j \in \mathcal{N} : (j, i) \in \mathcal{E}\}$ and $\mathcal{N}_{\mathbf{W},i}^{out} := \{j \in \mathcal{N} : (i, j) \in \mathcal{E}\}$, respectively. The degree of node i is the number of its in-neighbors or out-neighbors. For example, in a one-peer graph, the degree of each node is at most 1.

4

1.4 Organization of the Paper

The rest of this paper is organized as follows. In Section 2, we introduce the B-ary Tree Push-Pull algorithm and present its main convergence results. The sketch of analysis is presented in Section 3, and numerical experiments are provided in Section 4. We conclude the paper in Section 5.

2 B-ary Tree Push-Pull Method

2.1 Communication Graphs

The proposed B-ary Tree Push-Pull method makes use of two spanning trees as communication graphs: $\mathcal{G}_{\mathcal{R}}$ and $\mathcal{G}_{\mathcal{C}}$, which correspond to two mixing matrices $\mathcal{R}$ and $\mathcal{C}$, respectively. Specifically, we consider B-ary tree graphs with arbitrary number of nodes n and depth d. The root node is labeled as 1 for convenience, and we index the nodes layer-by-layer. The additional nodes are placed at the last layer if the tree is not full. Figure 1 illustrates the assignment of 10 nodes when $B = 2$. In the Pull Tree $\mathcal{G}_{\mathcal{R}}$ (the left ones), each node has 1 parent node and B child nodes (except the ones in the last layer). The root node 1 has no parent node. In the Push Tree $\mathcal{G}_{\mathcal{C}}$ (the right ones), each node has 1 child node and B parent nodes (except the ones in the last layer). It can be seen that the tree $\mathcal{G}_{\mathcal{C}}$ is identical to $\mathcal{G}_{\mathcal{R}}$ with all the edges reversing directions. Note that only node 1 has a self-loop.
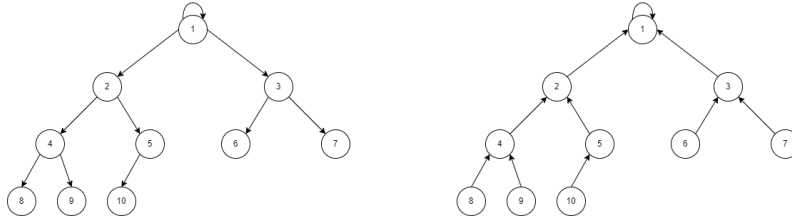
Figure 1: Two spanning trees with 10 nodes when $B = 2$. On the left is $\mathcal{G}_{\mathcal{R}}$, and the right one is $\mathcal{G}_{\mathcal{C}}$.

2.2 Algorithm

We consider the following distributed stochastic gradient method (Algorithm 1) for solving problem (1). At every iteration t, each agent i pulls the state information from its in-neighborhood $\mathcal{N}_{\mathcal{R},i}^{in}$, pushes its (stochastic) gradient tracker y_i to the out-neighborhood $\mathcal{N}_{\mathcal{C},i}^{out}$, and updates its local variables x_i and y_i based on the received information. The agents aim to find the ϵ-stationary point jointly by performing local computation and exchanging information through two spanning trees.

More specifically, in the pull tree $\mathcal{G}_{\mathcal{R}}$, each node i pulls the updated model from its parent node along the tree. Note that $\mathcal{N}_{\mathcal{R},i}^{in}$ consists of only one node, the parent node. The Push Tree $\mathcal{G}_{\mathcal{C}}$ is the inverse of the Pull Tree, in which each node collects and aggregates the gradient trackers from its parent nodes. Due to the tree structure, only y_1^t aggregates and tracks the stochastic gradients across the entire network, which will be made clear from the analysis. The implementation of the algorithm is rather simple. Taking node 2 in Figure 1 as an example, we have $x_2^{(t+1)} = x_1^{(t)} - \gamma y_1^{(t)}$ and $y_2^{(t+1)} = y_4^{(t)} + y_5^{(t)} + g_2(x_2^{(t+1)}; \xi_2^{(t+1)}) - g_2(x_2^{(t)}; \xi_2^{(t)})$.

We can write Algorithm 1 in the following compact form:

$$\begin{aligned}
\mathbf{X}^{(t+1)} &= \mathcal{R}\left(\mathbf{X}^{(t)} - \gamma\mathbf{Y}^{(t)}\right) \\
\mathbf{Y}^{(t+1)} &= \mathcal{C}\mathbf{Y}^{(t)} + \mathbf{G}(\mathbf{X}^{(t+1)}, \boldsymbol{\xi}^{(t+1)}) - \mathbf{G}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)})
\end{aligned} \tag{2}$$

where $\mathbf{Y}^{(0)} = \mathbf{G}(\mathbf{X}^{(0)}, \boldsymbol{\xi}^{(0)})$, and $\mathcal{R}, \mathcal{C} \in \mathbb{R}^{n \times n}$ are non-negative matrices whose elements are given by

$$[\mathcal{R}]_{i,j} = \begin{cases} 1 & \text{if } i \in \{Bj + 1 - B + [B]\} \cap [n] \text{ or } i = j = 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{C} = \mathcal{R}^{\top}$ which corresponds to $\mathcal{G}_{\mathcal{C}}$, the inverse tree of $\mathcal{G}_{\mathcal{R}}$. It can be seen that $\mathcal{R}$ is a row-stochastic matrix that only consists of 0's and 1's, and $\mathcal{C}$ is column stochastic. For example, the mixing matrices corresponding to the graphs in Figure 1 are given by

5

Algorithm 1 B-ary Tree Push-Pull Method (BTPP)

1: Each agent i initializes with any arbitrary but identical $x_i^{(0)} = x^{(0)} \in \mathbb{R}^p$, $y_i^{(0)} = g_i(x_i^{(0)}, \xi_i^{(0)}) \in \mathbb{R}^p$ after drawing a random sample $\xi_i^{(0)}$, stepsize $\gamma > 0$ and number of iterations T.

2: **for** iteration $t = 0, 1, 2, \ldots, T - 1$ **do**

3:      **for** agent i **in parallel do**

4:          Pull $x_j^{(t)} - \gamma y_j^{(t)}$ from each $j \in \mathcal{N}_{\mathcal{R},i}^{in}$

5:          Push $y_i^{(t)}$ to each $j \in \mathcal{N}_{\mathcal{C},i}^{out}$

6:          Independently draw a random sample $\xi_i^{(t+1)}$

7:          Update parameters through

$$x_i^{(t+1)} = \sum_{j \in \mathcal{N}_{\mathcal{R},i}^{in}} \left(x_j^{(t)} - \gamma y_j^{(t)} \right)$$

$$y_i^{(t+1)} = \sum_{j \in \mathcal{N}_{\mathcal{C},i}^{in}} y_j^{(t)} + g_i(x_i^{(t+1)}; \xi_i^{(t+1)}) - g_i(x_i^{(t)}; \xi_i^{(t)})$$

8:      **end for**

9: **end for**

10: Output $x_1^{(T)}$.

$$\mathcal{R} = \begin{pmatrix} 1 & & & & & & & & & \\ 1 & & & & & & & & & \\ 1 & & & & & & & & & \\ & & & 1 & & & & & & \\ & & & 1 & & & & & & \\ & & & & 1 & & & & & \\ & & & & & 1 & & & & \\ & & & & & & 1 & & & \\ & & & & & & & 1 & & \\ & & & & & & & & 1 & \end{pmatrix}, \quad \mathcal{C} = \begin{pmatrix} 1 & 1 & 1 & & & & & & & \\ & & & 1 & 1 & & & & & \\ & & & & & 1 & 1 & & & \\ & & & & & & & 1 & 1 & \\ & & & & & & & & & 1 \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \end{pmatrix},$$

where the unspecified elements are zeros.

2.3 Main Results

The main convergence properties of BTPP are summarized in the following two theorems, where the second result assumes strongly convexity on f.

Theorem 2.1. *For the BTPP algorithm outlined in Algorithm 1 implemented on B-ary tree graphs $\mathcal{G}_{\mathcal{R}}$ and $\mathcal{G}_{\mathcal{C}}$, assume Assumption 1.1 and Assumption 1.2 hold. Let $\gamma = \min\{\left(\frac{\Delta_f}{3\sigma^2 Ln(T+1)}\right)^{\frac{1}{2}}, \left(\frac{\Delta_f}{1500n^2 d^6 \sigma^2 L^2 (T+1)}\right)^{\frac{1}{3}}, \frac{1}{100nd^3 L}\}$. The following convergence result holds:*

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2 \leq \frac{32\sqrt{\Delta_f \sigma^2 L}}{\sqrt{n(T+1)}} + \frac{240 d^2 \left(\sigma^2 L^2 \Delta_f^2 \right)^{\frac{1}{3}}}{\left(\sqrt{n}(T+1) \right)^{\frac{2}{3}}} + \frac{800 d^3 L \Delta_f}{T+1} + \frac{\left\| \nabla \mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2}{n(T+1)},$$

(3)

where $\Delta_f := f(\mathbf{x}_1^{(0)}) - f^*$ and $d = \lfloor \log_B(n) \rfloor$ represents the diameter of the graphs.

Remark 2.2. Based on the convergence rate in (3) of BTPP, we can derive that when $T = \Theta(n \log^{12}(n))$, the term $\mathcal{O}(\frac{1}{\sqrt{nT}})$ dominates the remaining terms up to a constant scalar. This implies that BTPP achieves linear speedup after $\mathcal{O}\left(n \log^{12}(n)\right)$ transient iterations.

Remark 2.3. The convergence rate in (3) is related to the branch size B. For larger B, the diameter $d = \lfloor \log_B(n) \rfloor$ becomes smaller, which results in more efficient transmission of information and fewer transient iterations. However, the per-iteration communication cost is relatively larger. When B is smaller, the communication burden for each agent at every iteration is lighter, but the transient time is larger. Therefore, in practice, the communication cost and convergence rate can be balanced by considering a proper B.

6

Theorem 2.4. *For the BTPP algorithm outlined in Algorithm 1 implemented on B-ary tree graphs $\mathcal{G}_{\mathcal{R}}$ and $\mathcal{G}_{\mathcal{C}}$, assume Assumption 1.1, Assumption 1.2 and Assumption 1.3 hold. Let $\gamma = \min\left\{\frac{1}{100nd^2\kappa L}, \frac{16\log(n(T+1)^2)}{n(T+1)\mu}\right\}$ and $T \geq 2d$. The following convergence result holds:*

$$
\begin{aligned}
\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 &\leq \frac{2240\sigma^2 \log(n(T+1)^2)}{n(T+1)\mu^2} + \frac{26880000d^6\kappa^2\sigma^2\left(\log(n(T+1)^2)\right)^2}{n(T+1)^2\mu^2} \\
&\quad + \max\left\{\exp(-\frac{T}{800d^2\kappa^2}), \frac{40}{n(T+1)^2}\right\}\left(\left\|x_1^{(0)} - x^*\right\|^2 + \frac{1}{nL^2}\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2\right).
\end{aligned}
\tag{4}
$$

Remark 2.5. The convergence rate in (4) implies that $\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 \leq \tilde{\mathcal{O}}\left(\frac{1}{nT} + \frac{1}{nT^2} + \exp(-T)\right)$, where $\tilde{\mathcal{O}}$ hides the constants and polylogarithmic factors. The transient time is thus $\tilde{O}(1)$, i.e., the number of iterations before the term $\mathcal{O}(\frac{1}{nT})$ dominates the remaining terms. Such a transient time also outperforms the state-of-the-art results.

3 Analysis of B-ary Tree Push-Pull

In this section, we study the convergence of BTPP and prove Theorem 2.1 by analyzing the properties of the weight matrices $\mathcal{R}$ and $\mathcal{C}$, the evolution of the aggregated consensus error $\sum_{t=0}^{T}\mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2$, and the expected inner products of the stochastic gradients between different layers. The approach is different from those employed in [22, 19, 26], where the analysis considers two special matrix norms related to $\mathcal{R}$ and $\mathcal{C}$, respectively. Such a distinction is because BTPP works with two B-ary trees and iterates in a layer-wise manner, while most other works consider connected graphs.

Our analysis starts with characterizing the weight matrices $\mathcal{R}$ and $\mathcal{C}$, as delineated in the following lemmas. It is important to note that for any given n and a specific integer B, we can determine an integer d satisfying $\frac{B^d-1}{B-1} < n \leq \frac{B^{d+1}-1}{B-1}$ which is the diameter of the graphs.

Notice that $\mathcal{R}$ has a unique non-negative left eigenvector $\mathbf{u}^\top$ (w.r.t. eigenvalue 1) with $\mathbf{u}^\top\mathbf{1} = n$. More specifically, $\mathbf{u} = [n, 0, \cdots, 0]^\top$, which is also the unique right eigenvector of $\mathcal{C}$ (w.r.t. eigenvalue 1), denoted by $\mathbf{v}$ for the clarity of presentation. Following the above observations, it is revealed in Lemma 3.1 that the 2-norm of the matrix $\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top$ with exponent k remains bounded by $\sqrt{n}$ and equals zero when k exceeds $d-1$.

Lemma 3.1. *Given a positive integer k, the 2-norm of the matrix $\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top$ satisfies*

$$
\|\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\|_2 \begin{cases} \leq \sqrt{n} & k \leq d-1 \\ = 0 & k \geq d \end{cases}
$$

Similar result applies to the matrix $\mathcal{C}^k - \frac{1}{n}\mathbf{v}\mathbf{1}^\top$. Consequently, we introduce the mixing matrices $\mathbf{\Pi_u}, \mathbf{\Pi_v}$ based on the eigenvectors $\mathbf{u}, \mathbf{v}$, which play a crucial role in the follow-up analysis.

$$
\mathbf{\Pi_u} := \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top, \quad \mathbf{\Pi_v} := \mathbf{I} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top.
$$

The following lemmas delineate the critical elements for constraining the average expected norms of the objective function as formulated in (1), i.e., $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2$. Lemma 3.2 and Lemma 3.3 provide bounds on the expressions $\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\|_F^2$ and $\sum_{t=0}^{T}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2$, where $\bar{\mathbf{X}}^{(t)} := \frac{1}{n}\mathbf{1}\mathbf{u}^\top\mathbf{X}^{(t)}$.

Lemma 3.2. *Suppose Assumption 1.1 holds and $\gamma \leq \frac{1}{10ndL}$, we have the following inequality:*

$$
\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\|_F^2 &\leq 6\gamma^2n^2\sigma^2(T+1) + 50\gamma^2n^2d^2L^2\sum_{t=0}^{T}\mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2 \\
&\quad + 6\gamma^2n^2d\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2 + 15\gamma^2n^3\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2.
\end{aligned}
$$

7

Lemma 3.3. *Suppose Assumption 1.1 holds and $\gamma \leq \frac{1}{40nd^2L}$, we have for $d \geq 2$ that*

$$\sum_{t=0}^{T} \mathbb{E} \left\| \mathbf{\Pi_u X}^{(t)} \right\|_F^2 \leq 300\gamma^2 n^2 d^4 (T+1)\sigma^2 + 20\gamma^2 n^3 d^2 \sum_{t=0}^{T} \mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2$$

$$+ 6nd \left\| \mathbf{\Pi_u X}^{(0)} \right\|_F^2 + 40\gamma^2 n^2 d^3 \left\| \nabla \mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2.$$

From the design of BTPP, there is an inherent delay in the transmission of information from layer k to layer 1. As information traverses through the B-ary trees, the delay becomes evident. Specifically, for nodes at layer k, their information requires an additional k iterations to successfully reach and impact node 1, as demonstrated in Lemma 3.4.

Lemma 3.4. *For any integer $t > 1$, we have*

$$\sum_{m=1}^{\min\{t,d\}} \mathbb{E} \left\langle \nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right) \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right) \right\rangle = 0,$$

where $\mathbf{A}_m = \mathcal{C}^m - \mathcal{C}^{m-1}$ and $\mathbf{A}_1 = \mathcal{C} - \mathbf{I}$.

Building on the preceding lemmas, we are in a position to establish the main convergence result for the BTPP algorithm. This involves upper bounding the expected norm for the gradient of the objective function evaluated at $x_1^{(t)}$. To show the result, we integrate the findings from Lemma 3.2, Lemma 3.3, and Lemma 3.4, as detailed in Lemma 3.5.

Lemma 3.5. *Suppose Assumption 1.1 and Assumption 1.2 hold and $\gamma \leq \frac{1}{100nd^3L}$, we have*

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2 \leq \frac{8\Delta_f}{\gamma n(T+1)} + 24\gamma\sigma^2 L + 20000\gamma^2 nd^6\sigma^2 L^2$$

$$+ \frac{400d^3 L^2 \left\| \mathbf{\Pi_u X}^{(0)} \right\|_F^2}{T+1} + \frac{56\gamma d^3 L \left\| \nabla\mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2}{T+1}.$$

Remark 3.6. Lemma 3.5 implies that the transient time of BTPP under Assumption 1.2 is influenced by the fourth term in the upper bound: $\frac{400d^3L^2\left\|\mathbf{\Pi_u X}^{(0)}\right\|_F^2}{T+1}$ which is related to the initial consensus error. Therefore, we initialize all the agents with the same solution $x^{(0)}$.

Under strong convexity of f, we have the following key lemma.

Lemma 3.7. *Suppose Assumption 1.1, Assumption 1.2 and Assumption 1.3 hold, and $\gamma \leq \frac{1}{100nd^2\kappa L}$, we have*

$$\mathbb{E} \left\| x_1^{(T)} - x^* \right\|^2 \leq \left(1 - \frac{n\gamma\mu}{4}\right)^T \left\| x_1^{(0)} - x^* \right\|^2$$

$$+ 7\gamma^2 n\sigma^2 (T+1) + 21000\gamma^3 n^2 d^6 \kappa L\sigma^2 (T+1)$$

$$+ 80\gamma^2 nd^3 \left(1 - \frac{n\gamma\mu}{4}\right)^{T-d} \left\| \nabla\mathbf{F}(\mathbf{X}^{(0)}) \right\| + 420\gamma^3 n^2 d^3 \kappa L \left\| \mathbf{\Pi_u X}^{(0)} \right\|_F^2,$$

where $\kappa := L/\mu$ is the conditional number.

4 Numerical Results

This section presents experimental results to compare the B-ary Tree Push-Pull method with other popular algorithms on logistic regression with synthetic data and deep learning tasks with real data.

4.1 Logistic Regression

We compare the performance of BTPP against other algorithms listed in Table 1 for logistic regression with non-convex regularization [26]. The objective functions $f_i : \mathbb{R}^p \to \mathbb{R}$ are given by

$$f_i(x) := \frac{1}{J} \sum_{j=1}^{J} \ln\left(1 + \exp(-y_{i,j} h_{i,j}^\top x)\right) + R \sum_{k=1}^{p} \frac{x_{[k]}^2}{1 + x_{[k]}^2},$$

8

where $x_{[k]}$ is the k-th element of x, and $\{(h_{i,j}, y_{i,j})\}_{j=1}^{J}$ represent the local data kept by node i. To control the data heterogeneity across the nodes, we first let each node i be associated with a local logistic regression model with parameter $\tilde{x}_i$ generated by $\tilde{x}_i = \tilde{x} + v_i$, where $\tilde{x} \sim \mathcal{N}(0, \mathbf{I}_p)$ is a common random vector, and $v_i \sim \mathcal{N}(0, \sigma_h^2 \mathbf{I}_p)$ are random vectors generated independently. Therefore, $\{v_i\}$ decide the dissimilarities between $\tilde{x}_i$, and larger σ_h generally amplifies the difference. After fixing $\{\tilde{x}_i\}$, local data samples are generated that follow distinct distributions. For node i, the feature vectors are generated as $h_{i,j} \sim \mathcal{N}(0, \mathbf{I}_p)$, and $z_{i,j} \sim \mathcal{U}(0, 1)$. Then, the labels $y_{i,j} \in \{-1, 1\}$ are set to satisfy $z_{i,j} \le 1 + \exp(-y_{i,j} h_{i,j}^\top \tilde{x}_i)$.

In the simulations, the parameters are set as follows: $n = 100$, $p = 500$, $J = 1000$, $R = 0.01$, and $\sigma_h = 0.8$. All the algorithms initialize with the same stepsize $\gamma = 0.3$, except BTPP, which employs a modified stepsize γ/n. Such an adjustment is due to BTPP's update mechanism, which incorporates a tracking estimator that effectively accumulates n times the averaged stochastic gradients as the number of iterations increases. This can also be seen from the stepsize choice in Theorem 2.1.[2] Additionally, we implement a stepsize decay of 60% every 100 iterations to facilitate convergence.

In Figure 2, the gradient norm is used as a metric to gauge the algorithmic performance of each algorithm. The left panel of Figure 2 illustrates the comparative performance of various algorithms, highlighting that BTPP (in red) achieves faster convergence than the other algorithms with $\Theta(1)$ degree and closely approximates the performance of the centralized SGD algorithm (i.e., DSGD-FullyConnected). The right panel of Figure 2 demonstrates the behavior of BTPP when increasing the branch size B. It is observed that with larger B, the convergence trajectory of BTPP more closely aligns with that of centralized SGD, corroborating the prediction of the theoretical analysis.
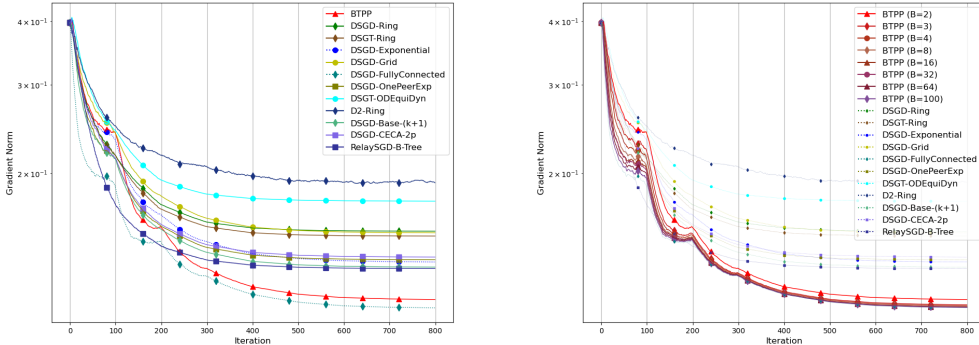
Figure 2: Left: performance of algorithms for logistic regression with nonconvex regularization, where the dotted lines correspond to algorithms whose degrees are not $\Theta(1)$. We let the branch size $B = 2$ in BTPP, $\eta = 0.5$ in OD-EquiDyn, $k = 2$ in Base-$(k+1)$, and perform RelaySGD on a binary tree graph for fairness. Right: performance of BTPP with different branch size B.

4.2 Deep Learning

We apply BTPP and the other algorithms to solve the image classification task with CNN over **MNIST** dataset [11]. We run all experiments on a server with eight Nvidia RTX 3090 GPUs. The network contains two convolutional layers with max pooling and ReLu and two feed-forward layers. In particular, we consider a heterogeneous data setting, where data samples are sorted based on their labels and partitioned among the agents. The local batch size is set to 8 with 24 agents in total. The learning rate is 0.01 for all the algorithms except BTPP (which employs a modified stepsize γ/n) for fairness. Additionally, the starting model is enhanced by pre-training using the SGD optimizer on the entire MNIST dataset for several iterations. Figure 3 illustrates the training loss and the test accuracy curves. Comparing the performance of different algorithms, it can be seen that BTPP (in red) and DSGT with ODEquiDyn (based on $\Theta(n)$ graphs) achieve faster convergence than the

---

[2]Note that this particular configuration results in slower convergence for BTPP during the initial $\mathcal{O}(d)$ iterations, which can be improved by using larger initial stepsizes.

9

other algorithms with $\Theta(1)$ degree and closely approximate the performance of centralized SGD(i.e., DSGD-FullyConnected).
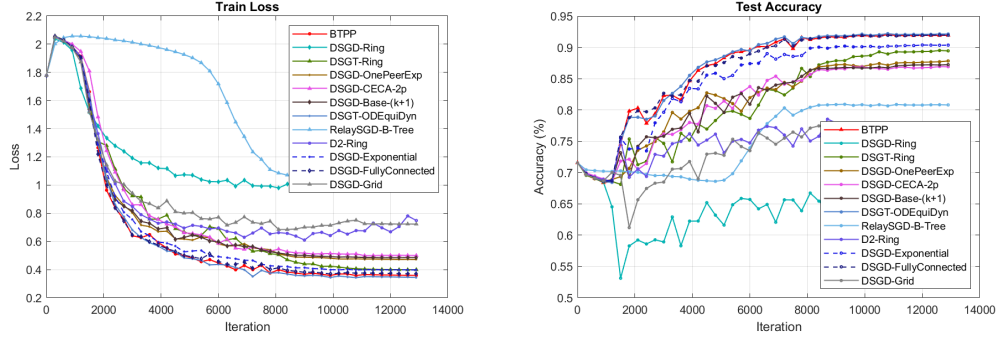
Figure 3: Train loss and test accuracy of different algorithms for training CNN on MNIST, where the dotted lines correspond to the algorithms whose degrees are not $\Theta(1)$. We perform BTPP with $B = 2$, ODEquiDyn with $\eta = 0.5$, Base-$(k + 1)$ with $k = 2$, and RelaySGD on a binary tree graph for fairness.

Remark 4.1. Higher accuracy can be achieved for BTPP and other methods when using the momentum technique, or when the data heterogeneity is removed, meaning that samples are randomly assigned to each agent. Additional experiments demonstrating the performance of various algorithms across different tasks and scenarios are provided in Appendix B.

5 Conclusions

This paper proposes a novel algorithm for distributed learning over heterogeneous data, named BTPP. The convergence is theoretically analyzed for smooth non-convex stochastic optimization. The results demonstrate that, at the minimal communication cost per iteration, BTPP achieves linear speedup in the number of nodes n, and the transient times behaves as $\tilde{O}(n)$ and $\tilde{O}(1)$ respectively for smooth nonconvex and strongly convex objectives, outperforming the state-of-the-art results. Numerical experiments further validate the efficiency of BTPP.

10

References

[1] M. ASSRAN, N. LOIZOU, N. BALLAS, AND M. RABBAT, *Stochastic gradient push for distributed deep learning*, in International Conference on Machine Learning, PMLR, 2019, pp. 344–353.

[2] L. BOTTOU, *Stochastic gradient descent tricks*, in Neural Networks: Tricks of the Trade: Second Edition, Springer, 2012, pp. 421–436.

[3] J. DEAN, G. CORRADO, R. MONGA, K. CHEN, M. DEVIN, M. MAO, M. RANZATO, A. SENIOR, P. TUCKER, K. YANG, ET AL., *Large scale distributed deep networks*, Advances in neural information processing systems, 25 (2012).

[4] L. DING, K. JIN, B. YING, K. YUAN, AND W. YIN, *Dsgd-ceca: Decentralized sgd with communication-optimal exact consensus algorithm*, arXiv preprint arXiv:2306.00256, (2023).

[5] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, Advances in neural information processing systems, 27 (2014).

[6] D. P. KINGMA, M. WELLING, ET AL., *An introduction to variational autoencoders*, Foundations and Trends® in Machine Learning, 12 (2019), pp. 307–392.

[7] J. P. KLEIJNEN, *Design and analysis of simulation experiments*, Springer, 2018.

[8] A. KOLOSKOVA, T. LIN, AND S. U. STICH, *An improved analysis of gradient tracking for decentralized machine learning*, Advances in Neural Information Processing Systems, 34 (2021), pp. 11422–11435.

[9] A. KOLOSKOVA, N. LOIZOU, S. BOREIRI, M. JAGGI, AND S. STICH, *A unified theory of decentralized sgd with changing topology and local updates*, in International Conference on Machine Learning, PMLR, 2020, pp. 5381–5393.

[10] A. KOLOSKOVA, S. STICH, AND M. JAGGI, *Decentralized stochastic optimization and gossip algorithms with compressed communication*, in International Conference on Machine Learning, PMLR, 2019, pp. 3478–3487.

[11] Y. LECUN, C. CORTES, C. BURGES, ET AL., *Mnist handwritten digit database*, 2010.

[12] M. LI, D. G. ANDERSEN, J. W. PARK, A. J. SMOLA, A. AHMED, V. JOSIFOVSKI, J. LONG, E. J. SHEKITA, AND B.-Y. SU, *Scaling distributed machine learning with the parameter server*, in 11th USENIX Symposium on operating systems design and implementation (OSDI 14), 2014, pp. 583–598.

[13] Z. LI, W. SHI, AND M. YAN, *A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates*, IEEE Transactions on Signal Processing, 67 (2019), pp. 4494–4506.

[14] X. LIAN, C. ZHANG, H. ZHANG, C.-J. HSIEH, W. ZHANG, AND J. LIU, *Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent*, Advances in neural information processing systems, 30 (2017).

[15] T. P. LILLICRAP, J. J. HUNT, A. PRITZEL, N. HEESS, T. EREZ, Y. TASSA, D. SILVER, AND D. WIERSTRA, *Continuous control with deep reinforcement learning*, arXiv preprint arXiv:1509.02971, (2015).

[16] T. LIN, S. P. KARIMIREDDY, S. U. STICH, AND M. JAGGI, *Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data*, arXiv preprint arXiv:2102.04761, (2021).

[17] V. MNIH, K. KAVUKCUOGLU, D. SILVER, A. GRAVES, I. ANTONOGLOU, D. WIERSTRA, AND M. RIEDMILLER, *Playing atari with deep reinforcement learning*, arXiv preprint arXiv:1312.5602, (2013).

11

[18] A. NEDIĆ, A. OLSHEVSKY, AND M. G. RABBAT, *Network topology and communication-computation tradeoffs in decentralized optimization*, Proceedings of the IEEE, 106 (2018), pp. 953–976.

[19] S. PU AND A. NEDIĆ, *Distributed stochastic gradient tracking methods*, Mathematical Programming, 187 (2021), pp. 409–457.

[20] S. PU, A. OLSHEVSKY, AND I. C. PASCHALIDIS, *Asymptotic network independence in distributed stochastic optimization for machine learning: Examining distributed and centralized stochastic gradient descent*, IEEE signal processing magazine, 37 (2020), pp. 114–122.

[21] S. PU, A. OLSHEVSKY, AND I. C. PASCHALIDIS, *A sharp estimate on the transient time of distributed stochastic gradient descent*, IEEE Transactions on Automatic Control, 67 (2021), pp. 5900–5915.

[22] S. PU, W. SHI, J. XU, AND A. NEDIĆ, *Push–pull gradient methods for distributed optimization in networks*, IEEE Transactions on Automatic Control, 66 (2020), pp. 1–16.

[23] B. RECHT, C. RE, S. WRIGHT, AND F. NIU, *Hogwild!: A lock-free approach to parallelizing stochastic gradient descent*, Advances in neural information processing systems, 24 (2011).

[24] S. RESNICK, *A probability path*, Springer, 2019.

[25] W. SHI, Q. LING, G. WU, AND W. YIN, *Extra: An exact first-order algorithm for decentralized consensus optimization*, SIAM Journal on Optimization, 25 (2015), pp. 944–966.

[26] Z. SONG, W. LI, K. JIN, L. SHI, M. YAN, W. YIN, AND K. YUAN, *Communication-efficient topologies for decentralized learning with $o(1)$ consensus rate*, Advances in Neural Information Processing Systems, 35 (2022), pp. 1073–1085.

[27] N. SRIVASTAVA, G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER, AND R. SALAKHUTDINOV, *Dropout: a simple way to prevent neural networks from overfitting*, The journal of machine learning research, 15 (2014), pp. 1929–1958.

[28] Y. TAKEZAWA, R. SATO, H. BAO, K. NIWA, AND M. YAMADA, *Beyond exponential graph: Communication-efficient topologies for decentralized learning via finite-time convergence*, arXiv preprint arXiv:2305.11420, (2023).

[29] H. TANG, X. LIAN, M. YAN, C. ZHANG, AND J. LIU, D^2: *Decentralized training over decentralized data*, in International Conference on Machine Learning, PMLR, 2018, pp. 4848–4856.

[30] L. TREVISAN, *Lecture notes on graph partitioning, expanders and spectral methods*, University of California, Berkeley, https://people. eecs. berkeley. edu/luca/books/expanders-2016. pdf, (2017).

[31] T. VOGELS, L. HE, A. KOLOSKOVA, S. P. KARIMIREDDY, T. LIN, S. U. STICH, AND M. JAGGI, *Relaysum for decentralized deep learning on heterogeneous data*, Advances in Neural Information Processing Systems, 34 (2021), pp. 28004–28015.

[32] L. XIAO AND S. BOYD, *Fast linear iterations for distributed averaging*, Systems & Control Letters, 53 (2004), pp. 65–78.

[33] R. XIN AND U. A. KHAN, *A linear algorithm for optimization over directed graphs with geometric convergence*, IEEE Control Systems Letters, 2 (2018), pp. 315–320.

[34] B. YING, K. YUAN, Y. CHEN, H. HU, P. PAN, AND W. YIN, *Exponential graph is provably efficient for decentralized deep training*, Advances in Neural Information Processing Systems, 34 (2021), pp. 13975–13987.

[35] K. YUAN, S. A. ALGHUNAIM, AND X. HUANG, *Removing data heterogeneity influence enhances network topology dependence of decentralized sgd*, Journal of Machine Learning Research, 24 (2023), pp. 1–53.

12

A Convergence Analysis of BTPP

In this section, we aim to demonstrate the convergence results of BTPP through a three-step process. First, we explore the key properties of matrices $\mathcal{R}$ and $\mathcal{C}$, acquainting readers with several operations that will be frequently utilized in the subsequent parts. Then, we introduce various technical tools essential for the analysis. Finally, we delve into proving the convergence results supported by a series of pertinent lemmas.

A.1 Properties of the Weight Matrices

In this part, we first demonstrate that $\mathcal{R} \in \mathbb{R}^{n \times n}$ possesses a set of properties (the matrix $\mathcal{C} = \mathcal{R}^\top$ shares similar properties). Then, we utilize the established tools to prove the crucial result presented in Lemma 3.1. Lastly, we provide clarifications on certain matrix operations that will be frequently employed in deriving the convergence results.

It is important to note that for any given n and specific integer B, the diameter of the corresponding B-ary tree graph d (the distance from the last layer node to node 1) satisfies $\frac{B^d - 1}{B - 1} < n \leq \frac{B^{d+1} - 1}{B - 1}$. To investigate the properties of $\mathcal{R}$ and $\mathcal{C}$, we will introduce the column vector $\mathbf{e}_\mathcal{I} \in \mathbb{R}^n$, where each element of $\mathbf{e}_\mathcal{I}$ is equal to 1 for indices $i \in \mathcal{I}$ and 0 otherwise. Define the index sets

$$\mathcal{I}_{1,k} = \left\{ 1 : \frac{B^{k+1} - 1}{B - 1} \right\},$$

$$\mathcal{I}_{i,k} = \left\{ \left(\frac{B^{k+1} - 1}{B - 1} + (i-2)B^k + 1 \right) : \left(\frac{B^{k+1} - 1}{B - 1} + (i-1)B^k \right) \right\},$$

where $k_1 : k_2$ is the arithmetic progression from k_1 to k_2 with difference 1. We can then define the matrix $\mathbf{Z}_k \in \mathbb{R}^{n \times n}$ as a composite of several column vectors arranged in the following format:

$$\mathbf{Z}_k = \left[\mathbf{e}_{\mathcal{I}_{1,k}}, \mathbf{e}_{\mathcal{I}_{2,k}}, \cdots, \mathbf{e}_{\mathcal{I}_{n,k}} \right].$$

This closed-form expression of $\mathcal{R}$ with any power k is shown in Lemma A.1 which aids in developing further properties.

Lemma A.1. *For the pull matrix $\mathcal{R}$ corresponding to the B-ary tree $\mathcal{G}_\mathcal{R}$, given any positive index k, we have*

$$\mathcal{R}^k = \mathbf{Z}_k.$$

Proof. We prove the lemma by induction. First, it is obvious that $\mathcal{R} = \mathbf{Z}_1$ by the definition of $\mathcal{R}$:

$$\mathcal{R}_{ij} = 1$$
$$\text{iff } i \in \{Bj + 1 - B + [B]\} \cap [n] \text{ or } i = j = 1$$
$$\text{iff } B(j-1) + 2 \leq i \leq Bj + 1, \ i \in [n] \text{ or } i = j = 1$$
$$\text{iff } [Z_1]_{ij} = 1.$$

Now assume the statement is true for $k = j$. Then, for $k = j + 1$, we have

$$\mathcal{R}^{j+1} = \mathcal{R}^j * \mathcal{R} = \mathbf{Z}_j \mathbf{Z}_1.$$

Denote $[\mathbf{Z}_j \cdot \mathbf{Z}_1]_i$ as the i-th column of $\mathbf{Z}_j \cdot \mathbf{Z}_1$. To establish the result, we only need to demonstrate that the two matrices, $\mathcal{R}^{k+1}$ and $\mathbf{Z}_{k+1}$, have the same column entries. For $i = 1$,

$$[\mathbf{Z}_j \mathbf{Z}_1]_1 = \mathbf{Z}_j [\mathbf{Z}_1]_1 = \sum_{i=1}^{B+1} [\mathbf{Z}_j]_i = \sum_{i=1}^{B+1} \mathbf{e}_{\mathcal{I}_{i,j}} = \mathbf{e}_{\cup_{i=1}^{B+1} \mathcal{I}_{i,j}} = \mathbf{e}_{\mathcal{I}_{1,j+1}}.$$

For $i > 1$, we have

$$[\mathbf{Z}_j \mathbf{Z}_1]_i = \mathbf{Z}_j [\mathbf{Z}_1]_i = \sum_{m \in \mathcal{I}_{i,1}} [\mathbf{Z}_j]_m = \sum_{m=(i-1)B+2}^{iB+1} [\mathbf{Z}_j]_m$$

$$= \sum_{m=(i-1)B}^{iB-1} \mathbf{e}_{\mathcal{I}_{m,j}} = \mathbf{e}_{\cup_{m=(i-1)B}^{iB-1} \mathcal{I}_{m,j}} = \mathbf{e}_{\mathcal{I}_{i,j+1}}.$$

Thus, we conclude that $\mathcal{R}^{k+1} = \mathbf{Z}_{k+1}$. $\qquad\square$

13

Corollary A.2 below reveals that when the power k exceeds d, $\mathcal{R}^k$ transforms into a matrix where the first column is entirely composed of ones, while all the other columns consist of zeros.

Corollary A.2. *For $k = d$, we have*

$$\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top = \mathbf{0}$$

where $\mathbf{0}$ is the matrix with all entries equal 0.

Proof. From Lemma A.1, we have for the i-th column of $\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top$ that

$$\left[\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right]_i = \begin{cases} -\mathbf{e}_{\frac{B^{k+1}-1}{B-1}+1:n} & i = 1 \\ \mathbf{e}_{\mathcal{I}_{i,k}} & i > 1 \end{cases}$$

For $k = d$, the first n elements of all the columns remain 0, which implies the desired result. $\square$

Now, we are ready to prove Lemma 3.1:

Proof of Lemma 3.1. For any integer $k \leq d - 1$, define

$$n_0 := \left\lfloor \frac{n - \frac{B^{k+1}-1}{B-1}}{B^k} \right\rfloor.$$

This ensures that $\frac{B^{k+1}-1}{B-1} + n_0 B^k \leq n$ and $\frac{B^{k+1}-1}{B-1} + (n_0 + 1)B^k > n$, so that only the first $(n_0 + 2)$-th columns of $\mathcal{R}^k$ consist of non-zero elements. Note that

$$\max_{\|\mathbf{x}\|_2 = 1} \left\{ \left\| \left(\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top \right) \mathbf{x} \right\|_2^2 \right\} = \max_{\mathbf{x}} \left\{ \frac{\left\| \left(\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top \right) \mathbf{x} \right\|_2^2}{\|\mathbf{x}\|_2^2} \right\}.$$

Then, we focus on the non-zero elements of the matrix $\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top$.

$$\frac{\left\| \left(\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top \right) \mathbf{x} \right\|_2^2}{\|\mathbf{x}\|_2^2} = \frac{1}{\|\mathbf{x}\|_2^2} \left(\sum_{j=2}^{n_0+1} B^k (x_j - x_1)^2 + \left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right](x_{n_0+2} - x_1)^2 \right)$$

$$:= \frac{\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}}{\|\mathbf{x}\|^2},$$

where $\tilde{\mathbf{x}} = (x_1, \cdots, x_{n_0+2})$ is the truncated $\mathbf{x}$, and

$$\Sigma = \begin{pmatrix} n - \frac{B^{k+1}-1}{B-1} & -B^k & \cdots & -\left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right] \\ -B^k & B^k & & \\ \vdots & & \ddots & \\ -\left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right] & & & \left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right] \end{pmatrix},$$

where the unspecified elements are all zeros. Since Σ is symmetric, all the eigenvalues are real. We show by contradiction that any eigenvalue λ of Σ is upper bounded by n. Otherwise, if there exists $\lambda > n$, we denote $\mathbf{x}$ as the corresponding eigenvector of λ. Then, we have from $\Sigma \mathbf{x} = \lambda \mathbf{x}$ that

$$\lambda x_1 = \left(n - \frac{B^{k+1}-1}{B-1}\right)x_1 - B^k x_2 - \cdots - \left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]x_{n_0+2}$$

$$\lambda x_2 = -B^k x_1 + B^k x_2$$

$$\lambda x_3 = -B^k x_1 + B^k x_3$$

$$\cdots$$

$$\lambda x_{n_0+2} = -\left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]x_1 + \left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]x_{n_0+2}.$$

14

Without loss of generality, assume $x_1 \neq 0$. Then, by substituting the other relations into the first one, we have

$$\lambda = n - \frac{B^{k+1}-1}{B-1} + \sum_{i=1}^{n_0} \frac{B^{2k}}{\lambda - B^k} + \frac{\left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]^2}{\lambda - \left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]}.$$

With the fact that $\lambda > n$, there holds

$$\begin{aligned}
\lambda \leq\; & n - \frac{B^{k+1}-1}{B-1} + \frac{n_0 B^{2k}}{n - B^k} + \frac{\left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]^2}{n - \left[n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right]} \\
=\; & n - \frac{B^{k+1}-1}{B-1} + \frac{n_0 B^{2k}}{n - B^k} + \frac{n^2}{\frac{B^{k+1}-1}{B-1} + B^k n_0} - 2n + \frac{B^{k+1}-1}{B-1} + B^k n_0 \\
=\; & \frac{n B^k n_0}{n - B^k} + \frac{n\left(n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right)}{\frac{B^{k+1}-1}{B-1} + B^k n_0} \\
\leq\; & \frac{n B^k n_0}{n - B^k} + \frac{n}{n - B^k}\left(n - \frac{B^{k+1}-1}{B-1} - B^k n_0\right) \\
=\; & n \frac{n - \frac{B^{k+1}-1}{B-1}}{n - B^k} < n,
\end{aligned}$$

which is a contradiction. Thus, we have $\lambda \leq n$. It follows that

$$\frac{\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}}{\|\mathbf{x}\|^2} \leq \frac{\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|^2} \leq \lambda_{\max}(\Sigma) \leq n.$$

From the fact that the square root function is monotonically increasing on $[0, \infty)$, we have

$$\left\|\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right\|_2^2 = \max_{\|\mathbf{x}\|_2=1}\left\{\left\|\left(\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)\mathbf{x}\right\|_2^2\right\} \leq n,$$

which implies that $\left\|\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right\|_2 \leq \sqrt{n}$ for $k \leq d-1$ and $\left\|\mathcal{R}^k - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right\|_2 = 0$ otherwise by Corollary A.2. □

The transformations described in Corollary A.3 below are straightforward.

Corollary A.3. *For any integer $m > 0$, we have*

$$\begin{aligned}
\mathbf{\Pi}_\mathbf{u}\mathcal{R} &= \mathbf{\Pi}_\mathbf{u}\left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right) = \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)\mathbf{\Pi}_\mathbf{u}, \\
\mathbf{\Pi}_\mathbf{u}\mathcal{R}^m &= \mathbf{\Pi}_\mathbf{u}\left(\mathcal{R}^m - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right) = \mathbf{\Pi}_\mathbf{u}\left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^m = \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^m\mathbf{\Pi}_\mathbf{u}.
\end{aligned}$$

To simplify the convergence analysis, we introduce the matrix $\mathbf{A}_i$ defined as follows:

$$\mathbf{A}_i = \mathcal{C}^i - \mathcal{C}^{i-1}$$

for $i = 1, 2, \cdots, d$. Specifically, $\mathbf{A}_1 = \mathcal{C} - \mathbf{I}$. Consequently, Corollary A.4 below can be directly deduced from Lemma A.1 and Corollary A.3.

Corollary A.4. *For $i = 1, \cdots, d$, we have*

$$\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{A}_i = \begin{cases} \mathbf{e}^\top_{\frac{B^i-1}{B-1}+1:\frac{B^{i+1}-1}{B-1}} & i \leq d-1 \\ \mathbf{e}^\top_{\frac{B^d-1}{B-1}+1:n} & i = d \end{cases}$$

Intuitively, Corollary A.4 illustrates that $\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{A}_i$ serves as an indicator vector representing the $(i+1)$-th layer of the graph.

15

A.2 Supporting Inequalities and Lemmas

Lemma A.5 and Lemma A.6 below are frequently employed for bounding the norms of matrix summations and multiplications. Their proofs rely on the Cauchy-Schwartz inequality and the definitions of matrix norms $\|\cdot\|_2$ and $\|\cdot\|_F$.

Lemma A.5. *For an arbitrary set of m matrices $\{\mathbf{A}_i\}_{i=1}^m$ with the same size, we have*

$$\left\|\sum_{i=1}^m \mathbf{A}_i\right\|_F^2 \leq m\sum_{i=1}^m \|\mathbf{A}_i\|_F^2.$$

Proof. By the definition of Frobenius norm, we have

$$\left\|\sum_{i=1}^m \mathbf{A}_i\right\|_F \leq \sum_{i=1}^m \|\mathbf{A}_i\|_F.$$

Taking squares on both sides and invoking the Cauchy-Schwarz inequality, we have

$$\left\|\sum_{i=1}^m \mathbf{A}_i\right\|_F^2 \leq \left(\sum_{i=1}^m \|\mathbf{A}_i\|_F\right)^2 \leq m\sum_{i=1}^m \|\mathbf{A}_i\|_F^2.$$

$\square$

Lemma A.6. *Let $\mathbf{A}$, $\mathbf{B}$ be two real matrices whose sizes match. Then,*

$$\|\mathbf{A}\mathbf{B}\|_F \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_F.$$

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ be the singular value decomposition of $\mathbf{A}$, with the largest singular value $\sigma_{\max}$ and hence $\|\mathbf{A}\|_2 = \sigma_{\max}$. Then, we have

$$\begin{aligned}
\|\mathbf{A}\mathbf{B}\|_F^2 = \left\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{B}\right\|_F^2 &= \text{trace}\left(\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{B}\right)^H \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{B}\right)\right) \\
&= \text{trace}\left(\left(\mathbf{\Sigma}\mathbf{V}^H\mathbf{B}\right)^H \left(\mathbf{\Sigma}\mathbf{V}^H\mathbf{B}\right)\right) = \left\|\mathbf{\Sigma}\mathbf{V}^H\mathbf{B}\right\|_F^2 \\
&\leq \sigma_{\max}^2 \|\mathbf{V}^H\mathbf{B}\|_F^2 = \sigma_{\max}^2 \text{trace}\left(\mathbf{B}^\top \mathbf{V}\mathbf{V}^H\mathbf{B}\right) \\
&= \sigma_{\max}^2 \text{trace}\left(\mathbf{B}^\top \mathbf{B}\right) = \sigma_{\max}^2 \|\mathbf{B}\|_F^2 \\
&= \|\mathbf{A}\|_2^2 \|\mathbf{B}\|_F^2,
\end{aligned}$$

which implies the desired result. $\square$

Lemma A.7 below will be used in the last step for deriving the convergence rate of BTPP.

Lemma A.7. *Let A, B, C and α be positive constants and T be a positive integer. Define function*

$$g(\gamma) = \frac{A}{\gamma(T+1)} + B\gamma + C\gamma^2.$$

Then,

$$\inf_{\gamma \in (0,\frac{1}{\alpha}]} g(\gamma) \leq 2\left(\frac{AB}{T+1}\right)^{\frac{1}{2}} + 2C^{\frac{1}{3}}\left(\frac{A}{T+1}\right)^{\frac{2}{3}} + \frac{\alpha A}{T+1},$$

where the upper bound can be achieved by choosing $\gamma = \min\left\{\left(\frac{A}{B(T+1)}\right)^{\frac{1}{2}}, \left(\frac{A}{C(T+1)}\right)^{\frac{1}{3}}, \frac{1}{\alpha}\right\}$.

Proof. See Lemma 26 in [8] for a reference. $\square$

Lemma A.8 is a technical result related to random variables.

Lemma A.8. *Consider three random variables X, Y, and Z. Assume that Z is independent with (X, Y). Let h and g be functions such that the conditional expectation $\mathbb{E}[g(Y, Z) \mid Y] = 0$. We have*

$$\mathbb{E}\left(h(X)g(Y, Z)\right) = 0.$$

16

Proof. It implies by the condition $Z \perp\!\!\!\perp (X, Y)$ that $\sigma(Z) \perp\!\!\!\perp \sigma(X, Y)$. Then,

$$
\mathbb{E}\left[h(X)g(Y, Z)|Y\right] = \mathbb{E}\left\{\mathbb{E}\left[h(X)g(Y, Z)|X, Y\right]|Y\right\}
$$
$$
= \mathbb{E}\left\{h(X)\mathbb{E}\left[g(Y, Z)|X, Y\right]|Y\right\}.
$$

It suffices to show

$$
\mathbb{E}\left[g(Y, Z)|X, Y\right] = \mathbb{E}\left[g(Y, Z)|Y\right](= 0).
$$

Let $f_g(y) = \mathbb{E}\left(g(y, Z)\right)$. Since $\sigma(Z) \perp\!\!\!\perp \sigma(X, Y)$, we have $\sigma(Z) \perp\!\!\!\perp \sigma(Y)$. Then,

$$
f_g(Y) = \mathbb{E}\left(g(Y, Z)|Y\right) = \mathbb{E}\left[g(Y, Z)|X, Y\right],
$$

which follows directly from (10.17) in [24]. Thus, by the Tower Rule, we reach the statement as follows:

$$
\mathbb{E}\left(h(X)g(Y, Z)\right) = \mathbb{E}\left\{\mathbb{E}\left(h(X)g(Y, Z)|Y\right)\right\} = 0.
$$

$\square$

A.3 Proofs of Key Lemmas

In this section, we prove several key lemmas for proving the main convergence result of BTPP.

A.3.1 Preparation

Algorithm 1, as encapsulated by the equations in (2), can be succinctly expressed in the following matrix form:

$$
\begin{pmatrix} \mathbf{X}^{(t+1)} \\ \mathbf{Y}^{(t+1)} \end{pmatrix} = \begin{pmatrix} \mathcal{R} & -\gamma\mathcal{R} \\ \mathbf{0} & \mathcal{C} \end{pmatrix} \begin{pmatrix} \mathbf{X}^{(t)} \\ \mathbf{Y}^{(t)} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{G}^{(t+1)} - \mathbf{G}^{(t)} \end{pmatrix}. \tag{5}
$$

By repeatedly applying equation (5) starting from time step t and going back to time step 0, we arrive at the following relation:

$$
\begin{pmatrix} \mathbf{X}^{(t)} \\ \mathbf{Y}^{(t)} \end{pmatrix} = \begin{pmatrix} \mathcal{R} & -\gamma\mathcal{R} \\ \mathbf{0} & \mathcal{C} \end{pmatrix}^t \begin{pmatrix} \mathbf{X}^{(0)} \\ \mathbf{Y}^{(0)} \end{pmatrix} + \sum_{m=0}^{t-1} \begin{pmatrix} \mathcal{R} & -\gamma\mathcal{R} \\ \mathbf{0} & \mathcal{C} \end{pmatrix}^{t-m-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{G}^{(m+1)} - \mathbf{G}^{(m)} \end{pmatrix}.
$$

For the sake of clarity, we start with introducing some simple definitions. Any matrix raised to the power of 0 is defined as the identity matrix $\mathbf{I}$, which matches the original matrix in dimension. The only exceptions are $\left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^0 := \mathbf{\Pi_u}$ and $\left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^0 := \mathbf{\Pi_v}$ for convenience. Furthermore, we introduce the following terms:

$$
\bar{\mathbf{X}}^{(t)} := \frac{1}{n}\mathbf{1}\mathbf{u}^\top\mathbf{X}^{(t)}, \quad \bar{\mathbf{Y}}^{(t)} := \frac{1}{n}\mathbf{v}\mathbf{1}^\top\mathbf{Y}^{(t)}.
$$

Note that, for any given integer $m > 0$,

$$
\begin{pmatrix} \mathcal{R} & -\gamma\mathcal{R} \\ \mathbf{0} & \mathcal{C} \end{pmatrix}^m = \begin{pmatrix} \mathcal{R}^m & -\gamma\sum_{j=1}^m \mathcal{R}^j\mathcal{C}^{m-j} \\ \mathbf{0} & \mathcal{C}^m \end{pmatrix}.
$$

As a result, given the initial condition $\mathbf{Y}^{(0)} = \mathbf{G}^{(0)}$, we can deduce the outcomes of $\mathbf{X}^{(t)}$ and $\mathbf{Y}^{(t)}$ as follows.

$$
\mathbf{X}^{(t)} = \mathcal{R}^t\mathbf{X}^{(0)} - \gamma\sum_{m=0}^{t-2}\sum_{j=1}^{t-m-1} \mathcal{R}^j\mathcal{C}^{t-m-1-j}\left[\mathbf{G}^{(m+1)} - \mathbf{G}^{(m)}\right] - \gamma\sum_{j=1}^{t} \mathcal{R}^j\mathcal{C}^{t-j}\mathbf{G}^{(0)}, \tag{6}
$$

$$
\mathbf{Y}^{(t)} = \sum_{m=0}^{t-1} \mathcal{C}^{t-m-1}\left[\mathbf{G}^{(m+1)} - \mathbf{G}^{(m)}\right] + \mathcal{C}^t\mathbf{G}^{(0)}. \tag{7}
$$

17

Then, after multiplying $\mathbf{\Pi_u}$ and $\mathbf{\Pi_v}$ to equation (6) and equation (7) respectively, and invoking Corollary A.3, we have

$$\mathbf{\Pi_u X}^{(t)} = \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^t \mathbf{X}^{(0)} - \gamma \sum_{j=1}^{\min\{d-1,t\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^j \mathcal{C}^{t-j}\mathbf{G}^{(0)}$$

$$- \gamma \sum_{m=0}^{t-2} \sum_{j=1}^{\min\{t-m-1,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^j \mathcal{C}^{t-m-1-j} \left[\mathbf{G}^{(m+1)} - \mathbf{G}^{(m)}\right], \tag{8}$$

$$\mathbf{\Pi_v Y}^{(t)} = \sum_{m=\max\{0,t-d\}}^{t-1} \left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^{t-m-1} \left[\mathbf{G}^{(m+1)} - \mathbf{G}^{(m)}\right] + \left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^t \mathbf{G}^{(0)}$$

$$= \sum_{m=0}^{\min\{t,d\}-1} \left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^m \left(\mathbf{G}^{(t-m)} - \mathbf{G}^{(t-m-1)}\right) + \left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^t \mathbf{G}^{(0)} \tag{9}$$

$$= \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \mathbf{G}^{(t-m)} + \mathbf{\Pi_v G}^{(t)}.$$

A.3.2   Analysis of the Variance

Denote by $\mathcal{F}_t$ the σ-algebra generated by $\{\boldsymbol{\xi}_0, \cdots, \boldsymbol{\xi}_{t-1}\}$, and define $\mathbb{E}\left[\cdot|\mathcal{F}_t\right]$ as the conditional expectation given $\mathcal{F}_t$. Lemma A.9 provides an estimate for the variance of the gradient estimator $G(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)})$.

Lemma A.9. *Under Assumption 1.1, for any given power $k \leq d-1$, we have for all $t \geq 0$ that*

$$\mathbb{E}\left[\left\|\left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^k \left(\mathbf{G}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)}) - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\right\|_F^2 \,\Big|\, \mathcal{F}_t\right] \leq 2n\sigma^2.$$

Proof. For any given t and $i \neq j$, due to the independently drawn sample $\xi_i^{(t)}$, we have that $\xi_i^{(t)}$ is independent of $(\mathcal{F}_t, \xi_j^{(t)})$, and thus $\xi_i^{(t)}$ is independent of $\sigma(x_i^{(t)}, x_j^{(t)}, \xi_j^{(t)})$. Hence, invoking Lemma A.8 and Assumption 1.1 yields

$$\mathbb{E}\left[\nabla F(x_i^{(t)}; \xi_i^{(t)}) - \nabla f_i(x_i^{(t)}) \Big| x_i^{(t)}\right] = \mathbb{E}\left[\nabla F(x_i^{(t)}; \xi_i^{(t)}) - \nabla f_i(x_i^{(t)}) \Big| \mathcal{F}_t\right] = 0,$$

$$\mathbb{E}\left\langle \nabla F(x_i^{(t)}; \xi_i^{(t)}) - \nabla f_i(x_i^{(t)}), \nabla F(x_j^{(t)}; \xi_j^{(t)}) - \nabla f_j(x_j^{(t)})\right\rangle = 0.$$

Then, for any index set $\mathcal{I} \subseteq \{1, 2, \cdots, n\}$, we have $\mathbb{E}\|\mathbf{e}_\mathcal{I}^\top \left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\|_2^2 \leq |\mathcal{I}|\sigma^2$. Notice that

$$\left\|\left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^k \left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\right\|_F^2 = \left\|\mathbf{e}_{\frac{B^{k+1}-1}{B-1}+1:n} \left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\right\|_2^2$$

$$+ \sum_{i=2}^{n} \left\|\mathbf{e}_{\mathcal{I}_{i,k}} \left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\right\|_2^2.$$

Thus, we obtain the desired result by invoking Lemma A.1 and Corollary A.2 after taking expectation on both sides of the above relation:

$$\mathbb{E}\left\|\left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^j \left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\right\|_F^2 \leq 2\left(n - \frac{B^{k+1}-1}{B-1}\right)\sigma^2 \leq 2n\sigma^2.$$

$\square$

18

Under Assumption 1.1 and the randomly selected samples, Lemma A.9 and Corollary A.10 below provide an initial estimation for the variance terms. The proof of Corollary A.10 is directly from the analysis in Appendix A.3.2 and Corollary A.4.

Corollary A.10. *Under Assumption 1.1, we have for all $t \geq 0$ that*

$$\sum_{k=1}^{d} \mathbb{E}\| \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) A_k \left(\mathbf{G}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)}) - \nabla\mathbf{F}(\mathbf{X}^{(t)}) \right) \|_2^2 \leq (n-1)\sigma^2.$$

A.3.3 Proof of Lemma 3.2

Proof. Notice that

$$\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} = -\gamma \frac{1}{n} \mathbf{1}\mathbf{u}^{\top} \mathbf{Y}^{(t)} = -\gamma \frac{1}{n} \mathbf{1}\mathbf{u}^{\top} \left[\Pi_{\mathbf{v}} \mathbf{Y}^{(t)} + \frac{1}{n} \mathbf{v}\mathbf{1}^{\top} \mathbf{Y}^{(t)} \right]$$

$$= -\gamma \frac{1}{n} \mathbf{1}\mathbf{u}^{\top} \Pi_{\mathbf{v}} \mathbf{Y}^{(t)} - \gamma \mathbf{1}\mathbf{1}^{\top} \mathbf{Y}^{(t)} = -\gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \Pi_{\mathbf{v}} \mathbf{Y}^{(t)} - \gamma \mathbf{1}\mathbf{1}^{\top} \mathbf{Y}^{(t)}.$$

Multiplying $\mathbf{1}^{\top}$ on both sides of equation (7), we have $\mathbf{1}^{\top}\mathbf{Y}^{(t)} = \mathbf{1}^{\top}\mathbf{G}^{(t)}$ for any integer t. Thus, in light of equation (9), we have

$$\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} = -\gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \Pi_{\mathbf{v}} \mathbf{Y}^{(t)} - \gamma \mathbf{1}\mathbf{1}^{\top}\mathbf{G}^{(t)}$$

$$= -\gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \mathbf{G}^{(t-m)} - \gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \mathbf{G}^{(t)} - \gamma \mathbf{1}\mathbf{1}^{\top}\mathbf{G}^{(t)}$$

$$= -\gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)}) \right)$$

$$- \gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \nabla\mathbf{F}(\mathbf{X}^{(t-m)}) - \gamma \mathbf{1} \frac{\mathbf{u}^{\top}}{n} \mathbf{G}^{(t)}$$

$$= -\gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)}) \right)$$

$$- \gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \sum_{m=\max\{0,t-d\}}^{t-1} \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^{\top} \right)^{t-m-1} \left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)}) - \nabla\mathbf{F}(\mathbf{X}^{(m)}) \right]$$

$$- \gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^{\top} \right)^{t} \nabla\mathbf{F}(\mathbf{X}^{(0)}) + \gamma \mathbf{1} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \nabla\mathbf{F}(\mathbf{X}^{(t)}) - \gamma \mathbf{1} \frac{\mathbf{u}^{\top}}{n} \mathbf{G}^{(t)}.$$

Hence, taking the F-norm and expectation on both sides, we have from Lemma A.5 that

$$\mathbb{E}\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\|_F^2 \leq 5\gamma^2 n \mathbb{E} \left\| \sum_{m=1}^{\min\{t,d\}} \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)}) \right) \right\|_F^2$$

$$+ 5\gamma^2 n \mathbb{E} \left\| \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \sum_{m=\max\{0,t-d\}}^{t-1} \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^{\top} \right)^{t-m-1} \left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)}) - \nabla\mathbf{F}(\mathbf{X}^{(m)}) \right] \right\|_F^2$$

$$+ 5\gamma^2 n \mathbb{E} \left\| \left(\frac{\mathbf{u}^{\top}}{n} - \mathbf{1}^{\top} \right) \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^{\top} \right)^{t} \nabla\mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2$$

$$+ 5\gamma^2 n \mathbb{E} \left\| \frac{\mathbf{u}^{\top}}{n} \left(\nabla\mathbf{F}(\mathbf{X}^{(t)}) - \mathbf{G}^{(t)} \right) \right\|_F^2 + 5\gamma^2 n \mathbb{E} \left\| \mathbf{1}^{\top}\mathbf{F}(\mathbf{X}^{(t)}) \right\|_F^2.$$

$$(10)$$

19

Note that, invoking Lemma A.9 and Corollary A.10 yields

$$\mathbb{E}\left\|\sum_{m=1}^{\min\{t,d\}}\left(\frac{\mathbf{u}^\top}{n}-\mathbf{1}^\top\right)\mathbf{A}_m\left(\mathbf{G}^{(t-m)}-\nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\|_F^2$$

$$=\sum_{m=1}^{\min\{t,d\}}\mathbb{E}\left\|\left(\frac{\mathbf{u}^\top}{n}-\mathbf{1}^\top\right)\mathbf{A}_m\left(\mathbf{G}^{(t-m)}-\nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\|_F^2$$

$$\leq(n-1)\sigma^2.$$

From Assumption 1.2, Lemma A.6 and Lemma A.5, we have

$$\mathbb{E}\left\|\left(\frac{\mathbf{u}^\top}{n}-\mathbf{1}^\top\right)\sum_{m=\max\{0,t-d\}}^{t-1}\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1}\left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)})-\nabla\mathbf{F}(\mathbf{X}^{(m)})\right]\right\|_F^2$$

$$\leq d\sum_{m=\max\{0,t-d\}}^{t-1}\mathbb{E}\left\|\left(\frac{\mathbf{u}^\top}{n}-\mathbf{1}^\top\right)\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1}\left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)})-\nabla\mathbf{F}(\mathbf{X}^{(m)})\right]\right\|_F^2$$

$$\leq d\sum_{m=\max\{0,t-d\}}^{t-1}\mathbb{E}\left\|\left(\frac{\mathbf{u}^\top}{n}-\mathbf{1}^\top\right)\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1}\right\|_2^2\left\|\nabla\mathbf{F}(\mathbf{X}^{(m+1)})-\nabla\mathbf{F}(\mathbf{X}^{(m)})\right\|_F^2$$

$$\leq ndL^2\sum_{m=\max\{0,t-d\}}^{t-1}\mathbb{E}\left\|\mathbf{X}^{(m+1)}-\mathbf{X}^{(m)}\right\|_F^2$$

$$\leq ndL^2\sum_{m=\max\{0,t-d\}}^{t-1}3\left(\mathbb{E}\left\|\mathbf{X}^{(m+1)}-\bar{\mathbf{X}}^{(m+1)}\right\|_F^2+\mathbb{E}\left\|\mathbf{X}^{(m)}-\bar{\mathbf{X}}^{(m)}\right\|_F^2+\mathbb{E}\left\|\bar{\mathbf{X}}^{(m+1)}-\bar{\mathbf{X}}^{(m)}\right\|_F^2\right).$$

Thus, summing over t in (10) from 0 to T, combining all the inequalities above, and invoking Assumption 1.1 and Assumption 1.2, we have

$$\sum_{t=0}^{T}\mathbb{E}\|\bar{\mathbf{X}}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\|_F^2\leq5\gamma^2n(n-1)\sigma^2(T+1)+30\gamma^2n^2d^2L^2\sum_{t=0}^{T}\mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2$$

$$+15\gamma^2n^2d^2L^2\sum_{t=0}^{T}\mathbb{E}\left\|\bar{\mathbf{X}}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2$$

$$+5\gamma^2n^2\sum_{t=0}^{\min\{t,d-1\}}\mathbb{E}\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2+5\gamma^2n\sigma^2(T+1)$$

$$+5\gamma^2n\sum_{t=0}^{T}\left(2\mathbb{E}\left\|\mathbf{1}^\top\nabla\mathbf{F}(\mathbf{X}^{(t)})-\mathbf{1}^\top\nabla\mathbf{F}(\bar{\mathbf{X}}^{(t)})\right\|_2^2+2n^2\mathbb{E}\left\|\frac{1}{n}\mathbf{1}^\top\nabla\mathbf{F}(\bar{\mathbf{X}}^{(t)})\right\|_2^2\right)$$

$$\leq5\gamma^2n^2\sigma^2(T+1)+40\gamma^2n^2d^2L^2\sum_{t=0}^{T}\mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2+15\gamma^2n^2d^2L^2\sum_{t=0}^{T}\mathbb{E}\left\|\bar{\mathbf{X}}^{(t+1)}-\bar{\mathbf{X}}^{(t)}\right\|_F^2$$

$$+5\gamma^2n^2d\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2+10\gamma^2n^3\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2.$$

$$(11)$$

Since $\gamma\leq\frac{1}{10ndL}$, we have $15\gamma^2n^2d^2L^2\leq\frac{1}{6}$, and the desired result follows. $\qquad\square$

A.3.4   Proof of Lemma 3.3

Proof. We show the upper bound for $\mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2$ by studying equation (8) and bound the F-norm of each term respectively. From Corollary A.2, we can change the power of $\mathcal{R}-\frac{1}{n}\mathbf{1}\mathbf{u}^\top$ to at most

20

$d-2$:

$$\Pi_{\mathbf{u}}\mathbf{X}^{(t)} =(\mathcal{R}^t - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)\mathbf{X}^{(0)} - \gamma\sum_{m=0}^{t-2}\sum_{j=1}^{\min\{t-m-1,d-1\}} (\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)^j \mathcal{C}^{t-m-1-j}\left[\mathbf{G}^{(m+1)} - \mathbf{G}^{(m)}\right]$$

$$- \gamma\sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \mathcal{C}^{t-j}\mathbf{G}^{(0)}.$$

$$(12)$$

Then, we derive the following decomposition by pairing the gradients with each of the stochastic gradients in order to use Assumption 1.1.

$$\Pi_{\mathbf{u}}\mathbf{X}^{(t)} =(\mathcal{R}^\top - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)\mathbf{X}^{(0)} - \gamma\sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-j}\left(\mathbf{G}^{(0)} - \nabla\mathbf{F}(\mathbf{X}^{(0)})\right)$$

$$- \gamma\sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \frac{1}{n}\mathbf{v}\mathbf{1}^\top\left(\mathbf{G}^{(0)} - \nabla\mathbf{F}(\mathbf{X}^{(0)})\right)$$

$$- \gamma\sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-j}\nabla\mathbf{F}(\mathbf{X}^{(0)})$$

$$- \gamma\sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \frac{1}{n}\mathbf{v}\mathbf{1}^\top\nabla\mathbf{F}(\mathbf{X}^{(0)})$$

$$- \gamma\sum_{m=0}^{t-2}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}} (\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)^j \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1-j}$$

$$\left[\mathbf{G}^{(m+1)} - \nabla\mathbf{F}(\mathbf{X}^{(m+1)}) + \nabla\mathbf{F}(\mathbf{X}^{(m)}) - \mathbf{G}^{(m)}\right]$$

$$- \gamma\sum_{m=0}^{t-2}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}} (\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)^j \left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1-j} \left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)}) - \nabla\mathbf{F}(\mathbf{X}^{(m)})\right]$$

$$- \gamma\sum_{m=0}^{t-2}\sum_{j=1}^{\min\{t-m-1,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \frac{\mathbf{v}\mathbf{1}^\top}{n} \left[\mathbf{G}^{(m+1)} - \nabla\mathbf{F}(\mathbf{X}^{(m+1)}) + \nabla\mathbf{F}(\mathbf{X}^{(m)}) - \mathbf{G}^{(m)}\right]$$

$$- \gamma\sum_{m=0}^{t-2}\sum_{j=1}^{\min\{t-m-1,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \frac{\mathbf{v}\mathbf{1}^\top}{n} \left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)}) - \nabla\mathbf{F}(\mathbf{X}^{(m)})\right]$$

$$:= \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^t \mathbf{X}^{(0)} - \gamma\mathbf{Q}_{0,t,1} - \gamma\mathbf{Q}_{0,t,2} - \gamma\mathbf{Q}_{0,t,3} - \gamma\mathbf{Q}_{0,t,4}$$

$$- \gamma\mathbf{Q}_{1,t} - \gamma\mathbf{Q}_{2,t} - \gamma\mathbf{Q}_{3,t} - \gamma\mathbf{Q}_{4,t},$$

$$(13)$$

where the terms from $\mathbf{Q}_{0,t,1}$ to $\mathbf{Q}_{0,t,4}$ and from $\mathbf{Q}_{1,t}$ to $\mathbf{Q}_{4,t}$ correspond to each term following $\left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^\top \mathbf{X}^{(0)}$ one-by-one.

We assume that $d \geq 2$, since $d = 1$ makes the summation illegal (summing over j from a positive number to a non-positive number), in which case $\Pi_{\mathbf{u}}\mathbf{X}^{(t)}$ degenerates to $(\mathcal{R}^t - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)\mathbf{X}^{(0)}$ and hence by Corollary A.2, there is no consensus error, i.e.

$$\sum_{t=0}^{T}\left\|\Pi_{\mathbf{u}}\mathbf{X}^{(t)}\right\|_F^2 = 0.$$

For $d \geq 2$, Lemma A.11 - Lemma A.13 below introduce the upper bounds for the F-norms of $\mathbf{Q}_{1,t}$, $\mathbf{Q}_{2,t}$ and $\mathbf{Q}_{3,t} + \mathbf{Q}_{0,t,2}$, summing from $t = 0$ to T. Lemma A.14 establishes a similar upper bound for the F-norm of $\mathbf{Q}_{0,t,1} + \mathbf{Q}_{0,t,3}$. Furthermore, Lemma A.15 provides the upper bound for the F-norm of $\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t}$.

21

Lemma A.11. *For any iteration number T, we have*

$$\sum_{t=0}^{T}\mathbb{E}\|\mathbf{Q}_{1,t}\|_F^2 \le 32n^2d^4(T+1)\sigma^2.$$

Proof. To make the summation legal given $d \ge 2$, we need $d-1 \ge t-m-d$, which implies that $m \ge t+1-2d$. Then,

$$\mathbf{Q}_{1,t} = \sum_{m=0}^{t-2}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}}\left(\mathcal{R}-\frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1-j}$$
$$\left[\mathbf{G}^{(m+1)}-\nabla\mathbf{F}(\mathbf{X}^{(m+1)})+\nabla\mathbf{F}(\mathbf{X}^{(m)})-\mathbf{G}^{(m)}\right]$$
$$= \sum_{m=\max\{t+1-2d,0\}}^{t-2}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}}\left(\mathcal{R}-\frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j$$
$$\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1-j}\left[\mathbf{G}^{(m+1)}-\nabla\mathbf{F}(\mathbf{X}^{(m+1)})+\nabla\mathbf{F}(\mathbf{X}^{(m)})-\mathbf{G}^{(m)}\right].$$

Invoking Lemma A.5, Lemma 3.1 and Lemma A.9, we have

$$\mathbb{E}\|\mathbf{Q}_{1,t}\|_F^2 \le 2(d-1)^2\sum_{m=\max\{t+1-2d,0\}}^{t-2}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}}\mathbb{E}\left\|\left(\mathcal{R}-\frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j\right\|_2^2 .$$
$$\left\|\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1-j}\left[\mathbf{G}^{(m+1)}-\nabla\mathbf{F}(\mathbf{X}^{(m+1)})+\nabla\mathbf{F}(\mathbf{X}^{(m)})-\mathbf{G}^{(m)}\right]\right\|_F^2$$
$$\le 2n(d-1)^2\sum_{m=\max\{t+1-2d,0\}}^{t-1}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}}8n\sigma^2$$
$$\le 32n^2d^4\sigma^2.$$

Summing over t from 0 to T, we get the desired result. $\qquad\square$

Lemma A.12. *For any $\gamma \le \frac{1}{20ndL}$, we have*

$$\sum_{t=0}^{T}\mathbb{E}\|\mathbf{Q}_{2,t}\|_F^2 \le 52n^2d^4L^2\sum_{t=0}^{T}\mathbb{E}\left\|\mathbf{\Pi}_{\mathbf{u}}\mathbf{X}^{(t)}\right\|_F^2 + 144\gamma^2n^4d^4(T+1)L^2\sigma^2$$
$$+ 360\gamma^2n^5d^4L^2\sum_{t=0}^{T}\mathbb{E}\left\|\frac{1}{n}\nabla\mathbf{F}(\bar{\mathbf{X}}^{(t)})\right\|_2^2 + 360\gamma^2n^4d^5L^2\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2 .$$

Proof. Similar to the proof of Lemma A.11, we have

$$\mathbf{Q}_{2,t} = \sum_{m=\max\{t+1-2d,0\}}^{t-1}\sum_{j=\max\{1,t-m-d\}}^{\min\{t-m-1,d-1\}}\left(\mathcal{R}-\frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j\left(\mathcal{C}-\frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-m-1-j}\left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)})-\nabla\mathbf{F}(\mathbf{X}^{(m)})\right].$$

Invoking Lemma A.5 and Assumption 1.2, we obtain

$$\mathbb{E}\|\mathbf{Q}_{2,t}\|_F^2 \le 4n^2d^3\sum_{m=\max\{t+1-2d,0\}}^{t-1}\mathbb{E}\left[\|\nabla\mathbf{F}(\mathbf{X}^{(m+1)})-\nabla\mathbf{F}(\mathbf{X}^{(m)})\|_F^2\right]$$
$$\le 4n^2d^3L^2\sum_{m=\max\{t+1-2d,0\}}^{t-1}\mathbb{E}\left\|\mathbf{X}^{(m+1)}-\bar{\mathbf{X}}^{(m+1)}+\bar{\mathbf{X}}^{(m)}-\mathbf{X}^{(m)}+\bar{\mathbf{X}}^{(m+1)}-\bar{\mathbf{X}}^{(m)}\right\|_F^2$$
$$\le 24n^2d^3L^2\sum_{m=\max\{t+1-2d,0\}}^{t}\mathbb{E}\left\|\mathbf{\Pi}_{\mathbf{u}}\mathbf{X}^{(t)}\right\|_F^2 + 12n^2d^3L^2\sum_{m=\max\{t+1-2d,0\}}^{t-1}\mathbb{E}\left\|\bar{\mathbf{X}}^{(m+1)}-\bar{\mathbf{X}}^{(m)}\right\|_F^2 .$$

22

It follows by summing over t from 0 to T and applying Lemma 3.2 that

$$\sum_{t=0}^{T} \mathbb{E}\|\mathbf{Q}_{2,t}\|_F^2 \leq 48n^2 d^4 L^2 \sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{\Pi_u X}^{(t)}\right\|_F^2 + 24n^2 d^4 L^2 \sum_{t=0}^{T} \mathbb{E}\left\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2$$

$$\leq \left(48n^2 d^4 L^2 + 1200\gamma^2 n^4 d^6 L^4\right) \sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{\Pi_u X}^{(t)}\right\|_F^2 + 144\gamma^2 n^4 d^4 (T+1) L^2 \sigma^2$$

$$+ 360\gamma^2 n^5 d^4 L^2 \sum_{t=0}^{T} \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 + 144\gamma^2 n^4 d^5 L^2 \left\|\nabla \mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2.$$

Hence, under the condition that $\gamma \leq \frac{1}{20ndL}$, there holds $1200\gamma^2 n^4 (d-1)^6 L^4 \leq 4n^2(d-1)^4 L^2$, which implies the desired result. $\qquad\square$

Lemma A.13. *For any T, we have*

$$\sum_{t=0}^{T} \mathbb{E}\|\mathbf{Q}_{3,t} + \mathbf{Q}_{0,t,2}\|_F^2 \leq d^2 n^2 (T+1)\sigma^2.$$

Proof. By definition, we have

$$\mathbf{Q}_{3,t} + \mathbf{Q}_{0,t,2} = \sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^j \frac{1}{n}\mathbf{v1}^\top \left(\mathbf{G}^{(0)} - \nabla \mathbf{F}(\mathbf{X}^{(0)})\right) +$$

$$\sum_{m=0}^{t-2} \sum_{j=1}^{\min\{t-m-1,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^j \frac{\mathbf{v1}^\top}{n} \left[\mathbf{G}^{(m+1)} - \nabla \mathbf{F}(\mathbf{X}^{(m+1)}) + \nabla \mathbf{F}(\mathbf{X}^{(m)}) - \mathbf{G}^{(m)}\right]$$

$$= \sum_{m=\max\{t-d,0\}}^{t-1} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^{t-m} \frac{\mathbf{v1}^\top}{n} \left[\mathbf{G}^{(m)} - \nabla \mathbf{F}(\mathbf{X}^{(m)})\right].$$

Thus, invoking Lemma A.5 and Lemma 3.1, we have

$$\mathbb{E}\|\mathbf{Q}_{3,t}\|_F^2 \leq nd \sum_{m=\max\{t-d,0\}}^{t-1} \mathbb{E}\|\frac{\mathbf{v}}{n}\|_2^2 \cdot \|\mathbf{1}^\top \left(\mathbf{G}^{(m+1)} - \nabla \mathbf{F}(\mathbf{X}^{(m+1)})\right)\|_F^2 \leq d^2 n^2 \sigma^2.$$

After summing over t from 0 to T, we get the desired result. $\qquad\square$

Lemma A.14. *For any T, we have*

$$\sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{Q}_{0,t,1} + \mathbf{Q}_{0,t,3}\right\|_F^2 \leq 4n^2 d^3 \sigma^2 + 4n^2 d^3 \left\|\nabla \mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2.$$

Proof. Note that
$$\left\|\mathbf{Q}_{0,t,1} + \mathbf{Q}_{0,t,3}\right\|_F^2 \leq 2\left\|\mathbf{Q}_{0,t,1}\right\|_F^2 + 2\left\|\mathbf{Q}_{0,t,3}\right\|_F^2.$$
We show the upper bounds for $\sum_{t=0}^{T} \|\mathbf{Q}_{0,t,i}\|_F^2$, where $i = 1, 3$ respectively. Based on Corollary A.2, Lemma A.5 and Lemma A.9, we have the following result:

$$\|\mathbf{Q}_{0,t,1}\|_F^2 = \left\|\sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^j \left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^{t-j} \left(\mathbf{G}^{(0)} - \nabla \mathbf{F}(\mathbf{X}^{(0)})\right)\right\|_F^2$$

$$= \left\|\sum_{j=\max\{1,t-d+1\}}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1u}^\top\right)^j \left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^{t-j} \left(\mathbf{G}^{(0)} - \nabla \mathbf{F}(\mathbf{X}^{(0)})\right)\right\|_F^2$$

$$\leq nd \sum_{j=\max\{1,t-d+1\}}^{\min\{t,d-1\}} \left\|\left(\mathcal{C} - \frac{1}{n}\mathbf{v1}^\top\right)^{t-j} \left(\mathbf{G}^{(0)} - \nabla \mathbf{F}(\mathbf{X}^{(0)})\right)\right\|_F^2.$$

23

Then, recall that the summation is legal only when $t \leq 2(d-2)$. We have

$$\sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{Q}_{0,t,1}\right\|_F^2 \leq \sum_{t=0}^{\min\{T,2(d-2)\}} nd \sum_{j=\max\{1,t-d+1\}}^{\min\{t,d-1\}} \left\|\left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-j}\left(\mathbf{G}^{(0)} - \nabla\mathbf{F}(\mathbf{X}^{(0)})\right)\right\|_F^2$$

$$\leq 2n^2 d^3 \sigma^2.$$

Similarly,

$$\sum_{t=0}^{\top} \mathbb{E}\left\|\mathbf{Q}_{0,t,3}\right\|_F^2$$

$$\leq \sum_{t=0}^{\min\{T,2(d-1)\}} d \sum_{j=\max\{1,t-d+1\}}^{\min\{t,d-1\}} \left\|\left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j\right\|_2^2 \left\|\left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{t-j}\right\|_2^2 \left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2$$

$$\leq 2n^2 d^3 \left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2.$$

Combining the above upper bounds leads to the final result. $\qquad\square$

Lemma A.15. *For any T, we have*

$$\sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t}\right\|_F^2 \leq 2n^2 d^2 L^2 \sum_{t=0}^{T} \mathbb{E}\|\mathbf{\Pi}_{\mathbf{u}}\mathbf{X}^{(t)}\|_F^2 + 2n^3 d^2 \sum_{t=0}^{T} \mathbb{E}\|f(x_1^{(m)})\|_2^2.$$

Proof. Note that

$$\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t} = \sum_{m=0}^{t-2} \sum_{j=1}^{\min\{t-m-1,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \frac{\mathbf{v}\mathbf{1}^\top}{n}\left[\nabla\mathbf{F}(\mathbf{X}^{(m+1)}) - \nabla\mathbf{F}(\mathbf{X}^{(m)})\right]$$

$$+ \sum_{j=1}^{\min\{t,d-1\}} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^j \frac{1}{n}\mathbf{v}\mathbf{1}^\top \nabla\mathbf{F}(\mathbf{X}^{(0)})$$

$$= \sum_{m=\max\{t-d,0\}}^{t-1} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^{t-m} \frac{\mathbf{v}\mathbf{1}^\top}{n} \nabla\mathbf{F}(\mathbf{X}^{(m)}),$$

where the last equality comes from extending the summation in the first line and telescoping the summation. Consequently, we have

$$\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t} = \sum_{m=\max\{t-d,0\}}^{t-1} \left(\mathcal{R} - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^{t-m} \frac{\mathbf{v}\mathbf{1}^\top}{n}\left(\nabla\mathbf{F}(\mathbf{X}^{(m)}) - \nabla\mathbf{F}(\bar{\mathbf{X}}^{(m)}) + \nabla\mathbf{F}(\bar{\mathbf{X}}^{(m)})\right).$$

Then, taking the F-norm on both sides and invoking Lemma A.6, Lemma A.5 and Lemma 3.2 as before, we have

$$\|\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t}\|_F^2 \leq dn \sum_{m=\max\{t-d,0\}}^{t-1} \left(2\|\frac{\mathbf{v}\mathbf{1}^\top}{n}\left[\nabla\mathbf{F}(\mathbf{X}^{(m)}) - \nabla\mathbf{F}(\bar{\mathbf{X}}^{(m)})\right]\|_F^2 + 2\|\frac{\mathbf{v}\mathbf{1}^\top}{n}\nabla\mathbf{F}(\bar{\mathbf{X}}^{(m)})\|_F^2\right)$$

$$\leq 2dn^2 L^2 \sum_{m=\max\{t-d,0\}}^{t-1} \|\mathbf{\Pi}_{\mathbf{u}}\mathbf{X}^{(m)}\|_F^2 + 2dn^3 \sum_{m=\max\{t-d,0\}}^{t-1} \|\nabla f(x_1^{(m)})\|_2^2.$$

Taking expectation on both sides and summing over t from 0 to T, we get

$$\sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t}\right\|_F^2 \leq 2n^2 d^2 L^2 \sum_{t=0}^{T} \mathbb{E}\|\mathbf{\Pi}_{\mathbf{u}}\mathbf{X}^{(t)}\|_F^2 + 2n^3 d^2 \sum_{t=0}^{T} \mathbb{E}\|\nabla f(x_1^{(m)})\|_2^2.$$

$\qquad\square$

24

Back to equation (13), note that

$$\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2 \le 6\left\|(\mathcal{R}^\top - \frac{1}{n}\mathbf{1}\mathbf{u}^\top)^t\mathbf{X}^{(0)}\right\|_F^2 + 6\gamma^2\left\|\mathbf{Q}_{0,t,1} + \mathbf{Q}_{0,t,3}\right\|_F^2$$
$$+ 6\gamma^2\left\|\mathbf{Q}_{1,t}\right\|_F^2 + 6\gamma^2\left\|\mathbf{Q}_{2,t}\right\|_F^2 + 6\gamma^2\left\|\mathbf{Q}_{3,t} + \mathbf{Q}_{0,t,2}\right\|_F^2 + 6\gamma^2\left\|\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t}\right\|_F^2.$$

Taking full expectation on both sides, summing over t from 0 to T and combining Lemma A.11 to Lemma A.14, we have

$$\sum_{t=0}^T \mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2 \le 6\sum_{t=0}^{\min\{T,d-1\}} \mathbb{E}\left\|\left(\mathcal{R}^\top - \frac{1}{n}\mathbf{1}\mathbf{u}^\top\right)^t\mathbf{X}^{(0)}\right\|_F^2 + 6\gamma^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{Q}_{1,t}\right\|_F^2 + 6\gamma^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{Q}_{2,t}\right\|_F^2$$

$$+ 6\gamma^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{Q}_{3,t} + \mathbf{Q}_{0,t,2}\right\|_F^2 + 6\gamma^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{Q}_{0,t,4} + \mathbf{Q}_{4,t}\right\|_F^2 + 6\gamma^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{Q}_{0,t,1} + \mathbf{Q}_{0,t,3}\right\|_F^2$$

$$\le 6nd\left\|\mathbf{\Pi_u}\mathbf{X}^{(0)}\right\|_F^2 + 192\gamma^2 n^2 d^4(T+1)\sigma^2 + 312\gamma^2 n^2 d^4 L^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2$$

$$+ 1000\gamma^4 n^4 d^4(T+1)L^2\sigma^2 + 2160\gamma^4 n^5 d^4 L^2\sum_{t=0}^T \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2$$

$$+ 2160\gamma^4 n^4 d^5 L^2\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2 + 6\gamma^2 n^2 d^2(T+1)\sigma^2 + 24\gamma^2 n^2 d^3\sigma^2$$

$$+ 24\gamma^2 n^2 d^3\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2 + 12\gamma^2 n^2 d^2 L^2\sum_{t=0}^T \mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2 + 12\gamma^2 n^3 d^2\sum_{t=0}^T \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2.$$
(14)

For $\gamma \le \frac{1}{40nd^2 L}$, which implies that $312\gamma^2 n^2 d^4 L^2 + 12\gamma^2 n^2 d^2 L^2 \le \frac{1}{4}$, we can simplify equation (14) as follows:

$$\sum_{t=0}^T \left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2 \le 300\gamma^2 n^2 d^4(T+1)\sigma^2 + 20\gamma^2 n^3 d^2\sum_{t=0}^T \mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2$$
$$+ 6nd\left\|\mathbf{\Pi_u}\mathbf{X}^{(0)}\right\|_F^2 + 40\gamma^2 n^2 d^3\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2,$$

which implies the desired result.

$\square$

A.3.5 Proof of Lemma 3.4

Proof. Notice that

$$x_1^{(t)} = x_1^{(t-1)} - \gamma y_1^{(t-1)} = x_1^{(t-1)} - \gamma\sum_{i\in\mathcal{I}_{1,1}} y_1^{(t-2)} - \gamma g_1(x_1^{(t-1)}, \xi_1^{(t-1)}) + \gamma g_1(x_1^{(t-2)}, \xi_1^{(t-2)}).$$

Therefore, $x_1^{(t)}$ does not depend on ξ_i^{t-1} for $i \ne 1$. We iterate the above procedure to get

$$x_1^{(t)} = x_1^{(t-1)} - \gamma g_1(x_1^{(t-1)}, \xi_1^{(t-1)}) + \gamma g_1(x_1^{(t-2)}, \xi_1^{(t-2)})$$
$$- \gamma\sum_{i\in\mathcal{I}_{1,2}} y_i^{(t-3)} - \gamma\sum_{i\in\mathcal{I}_{1,1}} g_i(x_1^{(t-2)}, \xi_1^{(t-2)}) + \gamma\sum_{i\in\mathcal{I}_{1,1}} g_i(x_1^{(t-3)}, \xi_1^{(t-3)}).$$

Similar to $x_1^{(t)}$, we know that $x_1^{(t-1)}$ does not depend on $\xi_i^{(t-2)}$, $i \ne 1$. Hence $x_1^{(t)}$ is independent with $\xi_i^{(t-2)}$ for $i \notin \mathcal{I}_{1,1}$. By iterating the procedure, we conclude that $x_1^{(t)}$ is independent with $\xi_i^{(t-k)}$ for $i \notin \mathcal{I}_{1,k}$.

Consequently, by choosing $Z = \{\xi_i^{(t-k)}, i\in\mathcal{I}_{1,k}, i\notin\mathcal{I}_{1,k-1}\}$, $X = x_1^{(t)}$, $Y = \{x_i^{(t-k)}, i\in[n]\}$ in Lemma A.8, Z is independent with (X,Y), we get

$$\mathbb{E}\left\langle \nabla f(x_1^{(t)}), \mathbf{e}_{\mathcal{I}_{1,k}/\mathcal{I}_{1,k-1}}\left(\mathbf{G}^{(t-k)} - \nabla\mathbf{F}(\mathbf{X}^{(t-k)})\right)\right\rangle = 0.$$

25

Then, invoking Corollary A.4, we have

$$
\sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left\langle \nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right) \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle
$$

$$
= \sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left\langle \nabla f(x_1^{(t)}), \mathbf{e}_{\mathcal{I}_{1,m}/\mathcal{I}_{1,m-1}} \left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle = 0.
$$

$\square$

A.3.6 Proof of Lemma 3.5

Proof. By Assumption 1.2, the function $f := \frac{1}{n}\sum_{i=1}^n f_i$ is L-smooth. Then,

$$
\mathbb{E}f(x_1^{(t+1)}) \le \mathbb{E}f(x_1^{(t)}) + \mathbb{E}\langle\nabla f(x_1^{(t)}), x_1^{(t+1)} - x_1^{(t)}\rangle + \frac{L}{2}\mathbb{E}\|x_1^{(t+1)} - x_1^{(t)}\|_2^2. \tag{15}
$$

For the last term, we have

$$
\mathbb{E}\|x_1^{(t+1)} - x_1^{(t)}\|_2^2 = \mathbb{E}\|\frac{1}{n}\mathbf{u}^\top\left(\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right)\|_2^2
$$

$$
= \frac{1}{n}\mathbb{E}\|\frac{1}{n}\mathbf{1}\mathbf{u}^\top\left(\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right)\|_F^2 = \frac{1}{n}\mathbb{E}\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\|_F^2.
$$

For the second last term, we have

$$
\mathbb{E}\langle\nabla f(x_1^{(t)}), x_1^{(t+1)} - x_1^{(t)}\rangle = \mathbb{E}\langle\nabla f(x_1^{(t)}), \frac{\mathbf{u}^\top}{n}\left(\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right)\rangle = \mathbb{E}\langle\nabla f(x_1^{(t)}), -\gamma\frac{\mathbf{u}^\top}{n}\mathbf{Y}^{(t)}\rangle
$$

$$
= \mathbb{E}\langle\nabla f(x_1^{(t)}), -\gamma\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{Y}^{(t)}\rangle + \mathbb{E}\langle\nabla f(x_1^{(t)}), -\gamma\mathbf{1}^\top\mathbf{Y}^{(t)}\rangle.
$$

$$\tag{16}$$

We now bound the two terms in the above equation. Firstly,

$$
\mathbb{E}\langle\nabla f(x_1^{(t)}), -\gamma\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{Y}^{(t)}\rangle = \gamma\mathbb{E}\langle\nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{\Pi_v}\mathbf{Y}^{(t)}\rangle.
$$

Recall that by equation (9),

$$
\gamma\mathbb{E}\langle\nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{\Pi_v}\mathbf{Y}^{(t)}\rangle
$$

$$
= \gamma\mathbb{E}\left\langle\nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\sum_{m=1}^{\min\{t,d\}}\mathbf{A}_m\mathbf{G}^{(t-m)} - \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{G}^{(t)}\right\rangle
$$

$$
= -\gamma\mathbb{E}\left\langle\nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\right\rangle
$$

$$
- \gamma\mathbb{E}\left\langle\nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\sum_{m=1}^{\min\{t,d\}}\mathbf{A}_m\left(\mathbf{G}^{(t-m)} - \nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle \tag{17}
$$

$$
- \gamma\mathbb{E}\left\langle\nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\nabla\mathbf{F}(\mathbf{X}^{(t)})\right\rangle
$$

$$
- \gamma\mathbb{E}\left\langle\nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\sum_{m=1}^{\min\{t,d\}}\mathbf{A}_m\nabla\mathbf{F}(\mathbf{X}^{(t-m)})\right\rangle.
$$

We bound the four terms above one by one. For the first term,

$$
\mathbb{E}\langle\nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\left(\mathbf{G}^{(t)} - \nabla\mathbf{F}(\mathbf{X}^{(t)})\right)\rangle = 0.
$$

26

For the second one, invoking Lemma 3.4, we have

$$\mathbb{E}\left\langle \nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla \mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle$$

$$= \sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left\langle \nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right) \mathbf{A}_m \left(\mathbf{G}^{(t-m)} - \nabla \mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle = 0.$$

For the last two terms in (17), we have as:

$$- \mathbb{E}\left\langle \nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right) \nabla \mathbf{F}(\mathbf{X}^{(t)})\right\rangle - \mathbb{E}\left\langle \nabla f(x_1^{(t)}), \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \nabla \mathbf{F}(\mathbf{X}^{(t-m)})\right\rangle$$

$$= \sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left\langle \nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{m-1} \left(\nabla \mathbf{F}(\mathbf{X}^{(t-m+1)}) - \nabla \mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle.$$

By the Cauchy-Schwartz inequality, we have

$$\sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left\langle \nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{m-1} \left(\nabla \mathbf{F}(\mathbf{X}^{(t-m+1)}) - \nabla \mathbf{F}(\mathbf{X}^{(t-m)})\right)\right\rangle$$

$$\leq \sum_{m=1}^{\min\{t,d\}} \left\{ \frac{n}{2d}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 + \frac{d}{2n}\mathbb{E}\left\|\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\left(\mathcal{C} - \frac{1}{n}\mathbf{v}\mathbf{1}^\top\right)^{m-1}\right\|_2^2 \left\|\nabla \mathbf{F}(\mathbf{X}^{(t-m+1)}) - \nabla \mathbf{F}(\mathbf{X}^{(t-m)})\right\|_F^2\right\}$$

$$\leq \frac{n}{2}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 + \frac{dL^2}{2}\sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left\|\mathbf{X}^{(t-m+1)} - \mathbf{X}^{(t-m)}\right\|_F^2$$

$$\leq \frac{n}{2}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 + \frac{3dL^2}{2}\sum_{m=1}^{\min\{t,d\}} \mathbb{E}\left(\left\|\mathbf{\Pi_u}\mathbf{X}^{(t-m+1)}\right\|_F^2 + \left\|\mathbf{\Pi_u}\mathbf{X}^{(t-m)}\right\|_F^2 + \left\|\bar{\mathbf{X}}^{(t-m+1)} - \bar{\mathbf{X}}^{(t-m)}\right\|_F^2\right).$$

Thus, combining the above inequalities together yields

$$\sum_{t=0}^{T} \gamma \mathbb{E}\langle \nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{\Pi_v}\mathbf{Y}^{(t)}\rangle \leq \frac{n\gamma}{2}\sum_{t=0}^{T} \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2$$

$$+ 3\gamma d^2 L^2 \sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2 + 3\gamma d^2 L^2 \sum_{t=0}^{T} \mathbb{E}\left\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2.$$

Secondly, for the second term in (16),

$$\mathbb{E}\langle \nabla f(x_1^{(t)}), -\gamma \mathbf{1}^\top \mathbf{Y}^{(t)}\rangle = \mathbb{E}\langle \nabla f(x_1^{(t)}), -\gamma \mathbf{1}^\top \mathbf{G}^{(t)}\rangle = \mathbb{E}\langle \nabla f(x_1^{(t)}), -\gamma \mathbf{1}^\top \nabla \mathbf{F}(\mathbf{X}^{(t)})\rangle$$

$$= -n\gamma \mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2 - n\gamma \mathbb{E}\langle \nabla f(x_1^{(t)}), \frac{1}{n}\mathbf{1}^\top \nabla \mathbf{F}(\mathbf{X}^{(t)}) - \frac{1}{n}\mathbf{1}^\top \nabla \mathbf{F}(\bar{\mathbf{X}}^{(t)})\rangle$$

$$\leq -n\gamma \mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2 + \gamma \frac{n}{4}\mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2 + 2\gamma \mathbb{E}\|\nabla \mathbf{F}(\mathbf{X}^{(t)}) - \nabla \mathbf{F}(\bar{\mathbf{X}}^{(t)})\|_F^2$$

$$\leq -n\gamma \mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2 + \gamma \frac{n}{4}\mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2 + 2\gamma L^2 \mathbb{E}\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\|_F^2.$$

Summing over t from 0 to T, we have

$$\sum_{t=0}^{T} \mathbb{E}\langle \nabla f(x_1^{(t)}), -\gamma \mathbf{1}^\top \mathbf{Y}^{(t)}\rangle \leq -\frac{3n\gamma}{4}\sum_{t=0}^{T} \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 + 2\gamma L^2 \sum_{t=0}^{T} \mathbb{E}\left\|\mathbf{\Pi_u}\mathbf{X}^{(t)}\right\|_F^2.$$

It yields by summing over t from 0 to T on both sides of equation (15) that

$$\mathbb{E}f(x_1^{(T+1)}) - \mathbb{E}f(x_1^0) \leq \sum_{t=0}^{T} \mathbb{E}\langle \nabla f(x_1^{(t)}), x_1^{(t+1)} - x_1^{(t)}\rangle + \frac{L}{2n}\sum_{t=0}^{T} \mathbb{E}\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\|_F^2$$

$$\leq \sum_{t=0}^{T} \mathbb{E}\langle \nabla f(x_1^{(t)}), -\gamma \mathbf{1}^\top \mathbf{Y}^{(t)}\rangle + \sum_{t=0}^{T} \gamma \mathbb{E}\langle \nabla f(x_1^{(t)}), -\left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top\right)\mathbf{\Pi_v}\mathbf{Y}^{(t)}\rangle + \frac{L}{2n}\sum_{t=0}^{T} \mathbb{E}\|\bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)}\|_F^2.$$

27

With the results above, given $\Delta_f = f(x^0) - f^*$, we have

$$-\Delta_f \leq -\frac{n\gamma}{4} \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2 + \frac{L}{2n} \sum_{t=0}^{T} \mathbb{E} \| \bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} \|_F^2$$

$$+ 5\gamma d^2 L^2 \sum_{t=0}^{T} \mathbb{E} \left\| \mathbf{\Pi_u X}^{(t)} \right\|_F^2 + 3\gamma d^2 L^2 \sum_{t=0}^{T} \mathbb{E} \left\| \bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} \right\|_F^2.$$

Invoking Lemma 3.2, we have for $\gamma \leq \frac{1}{100nd^3L}(\leq \frac{1}{10ndL})$ that

$$-\Delta_f \leq n\gamma \left(45\gamma^2 n^2 d^2 L^2 + 8\gamma nL - \frac{1}{4} \right) \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2$$

$$+ \left(150\gamma^3 n^2 d^4 L^4 + 25\gamma^2 nd^2 L^3 + 5\gamma d^2 L^2 \right) \sum_{t=0}^{T} \mathbb{E} \left\| \mathbf{\Pi_u X}^{(t)} \right\|_F^2$$

$$+ 3\gamma^2 n(T+1)L\sigma^2 + 18\gamma^3 n^2 d^2 L^2(T+1)\sigma^2$$

$$+ \left(18\gamma^3 n^2 d^3 L^2 + 3\gamma^2 nL \right) \left\| \nabla \mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2,$$

where it holds that $150\gamma^3 n^2 d^4 L^4 + 25\gamma^2 nd^2 L^3 + 5\gamma d^2 L^2 \leq 6\gamma d^2 L^2$.

Invoking Lemma 3.3, we have for $\gamma \leq \frac{1}{100nd^3L}(\leq \frac{1}{40nd^2L})$ that

$$-\Delta_f \leq n\gamma \left(120\gamma^2 n^2 d^4 L^2 + 45\gamma^2 n^2 d^2 L^2 + 8\gamma nL - \frac{1}{4} \right) \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2$$

$$+ 3\gamma^2 n(T+1)L\sigma^2 + 18\gamma^3 n^2 d^2 L^2(T+1)\sigma^2 + 1800\gamma^3 n^2 d^6 (T+1)\sigma^2 L^2$$

$$+ \left(240\gamma^3 n^2 d^5 L^2 + 18\gamma^3 n^2 d^3 L^2 + 3\gamma^2 nL \right) \left\| \nabla \mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2$$

$$+ 36\gamma nd^3 L^2 \left\| \mathbf{\Pi_u X}^{(0)} \right\|_F^2.$$

Thus, for $\gamma \leq \frac{1}{100nd^3L}$, we have

$$120\gamma^2 n^2 d^4 L^2 + 45\gamma^2 n^2 d^2 L^2 + 8\gamma nL - \frac{1}{4} \leq -\frac{1}{8}$$

$$\gamma n \left(240\gamma^2 nd^5 L^2 + 18\gamma^2 nd^3 L^2 + 3\gamma L \right) \leq 7\gamma^2 nd^2 L.$$

After re-arranging the terms, we conclude that

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2 \leq \frac{8\Delta_f}{\gamma n(T+1)} + 24\gamma \sigma^2 L + 20000\gamma^2 nd^6 \sigma^2 L^2$$

$$+ \frac{400d^3 L^2 \left\| \mathbf{\Pi_u X}^{(0)} \right\|_F^2}{T+1} + \frac{56\gamma d^3 L \left\| \nabla \mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2}{T+1}.$$

$\square$

A.3.7 Proof of Lemma 3.7

Let $x^* = \arg\min_x f(x)$. We start with analyzing the behavior of $\|x_1^{(t)} - x^*\|^2$ after obtaining Lemma 3.4. It holds that

$$\left\| x_1^{(t+1)} - x^* \right\|^2 = \left\| x_1^{(t)} - x^* \right\|^2 + 2 \left\langle x_1^{(t)} - x^*, x_1^{(t+1)} - x_1^{(t)} \right\rangle + \left\| x_1^{(t+1)} - x_1^{(t)} \right\|^2. \quad (18)$$

To deal with the critical inner product, similar to the decomposition in Equation (16), we have, by replacing $\nabla f(x_1^{(t)})$ with $x_1^{(t)} - x^*$ in Equation (17), and invoking Lemma 3.4 as we have done in Lemma 3.5, that

$$\left\langle x_1^{(t)} - x^*, x_1^{(t+1)} - x_1^{(t)} \right\rangle$$

$$= - \left\langle x_1^{(t)} - x^*, \gamma \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top \right) \mathbf{\Pi_v Y}^{(t)} \right\rangle - \left\langle x_1^{(t)} - x^*, \gamma \mathbf{1}^\top \mathbf{Y}^{(t)} \right\rangle. \quad (19)$$

28

and

$$- \mathbb{E} \left\langle x_1^{(t)} - x^*, \gamma \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top \right) \boldsymbol{\Pi}_{\mathbf{v}} \mathbf{Y}^{(t)} \right\rangle$$

$$= - \gamma \mathbb{E} \left\langle x_1^{(t)} - x^*, \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top \right) \nabla \mathbf{F}(\mathbf{X}^{(t)}) \right\rangle$$

$$- \gamma \mathbb{E} \left\langle x_1^{(t)} - x^*, \left(\frac{\mathbf{u}^\top}{n} - \mathbf{1}^\top \right) \sum_{m=1}^{\min\{t,d\}} \mathbf{A}_m \nabla \mathbf{F}(\mathbf{X}^{(t)}) \right\rangle$$

$$\leq \frac{n\gamma\mu}{4} \mathbb{E} \left\| x_1^{(t)} - x^* \right\|^2 + \frac{3dL^2}{\mu} \gamma \sum_{m=1}^{\min\{t,d\}} \mathbb{E} \left(\left\| \boldsymbol{\Pi}_{\mathbf{u}} \mathbf{X}^{t-m+1} \right\|^2 + \left\| \boldsymbol{\Pi}_{\mathbf{u}} \mathbf{X}^{t-m} \right\|^2 + \left\| \bar{\mathbf{X}}^{t-m+1} - \bar{\mathbf{X}}^{t-m} \right\|^2 \right).$$

(20)

Notice that, first by strong convexity of f (Assumption 1.3) and then by L-smoothness (Assumption 1.2), there holds

$$- \left\langle x_1^{(t)} - x^*, \nabla f(x_1^{(t)}) \right\rangle \leq f(x^*) - f(x_1^{(t)}) - \frac{\mu}{2} \left\| x_1^{(t)} - x^* \right\|^2$$

$$\leq - \frac{1}{2L} \left\| \nabla f(x_1^{(t)}) \right\|^2 - \frac{\mu}{2} \left\| x_1^{(t)} - x^* \right\|^2.$$

(21)

Then,

$$- \gamma \mathbb{E} \left\langle x_1^{(t)} - x^*, \mathbf{1}^\top \mathbf{Y}^{(t)} \right\rangle = - \gamma \mathbb{E} \left\langle x_1^{(t)} - x^*, \mathbf{1}^\top \nabla \mathbf{F}(\mathbf{X}^{(t)}) \right\rangle$$

$$= - n\gamma \mathbb{E} \left\langle x_1^{(t)} - x^*, \nabla f(x_1^{(t)}) \right\rangle - \gamma \mathbb{E} \left\langle x_1^{(t)} - x^*, \mathbf{1}^\top \nabla \mathbf{F}(\mathbf{X}^{(t)}) - \mathbf{1}^\top \nabla \mathbf{F}(\bar{\mathbf{X}}^{(t)}) \right\rangle$$

$$\leq - \frac{n\gamma\mu}{2} \mathbb{E} \left\| x_1^{(t)} - x^* \right\|^2 - \frac{n\gamma}{2L} \left\| \nabla f(x_1^{(t)}) \right\|^2 + \frac{n\gamma\mu}{8} \mathbb{E} \left\| x_1^{(t)} - x^* \right\|^2 + \frac{2L^2\gamma}{\mu} \mathbb{E} \left\| \boldsymbol{\Pi}_{\mathbf{u}} \mathbf{X}^{(t)} \right\|^2.$$

(22)

Thus, plugging the above results into Equation (18), with $\kappa := L/\mu$ as the conditional number, it holds that

$$\mathbb{E} \left\| x_1^{(t+1)} - x^* \right\|^2 \leq \left(1 - \frac{n\gamma\mu}{4} \right) \mathbb{E} \left\| x_1^{(t)} - x^* \right\|^2 - \frac{n\gamma}{L} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|^2 + \frac{1}{n} \mathbb{E} \left\| \bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} \right\|_F^2$$

$$+ 6d\kappa L \gamma \sum_{m=1}^{\min\{t,d\}} \mathbb{E} \left\| \bar{\mathbf{X}}^{(t-m+1)} - \bar{\mathbf{X}}^{(t-m)} \right\|_F^2 + 20d\kappa L \gamma \sum_{m=0}^{\min\{t,d\}+1} \mathbb{E} \left\| \boldsymbol{\Pi}_{\mathbf{u}} \mathbf{X}^{t-m} \right\|_F^2.$$

(23)

We derive the convergence result by several standard steps as follows.

Step 1. Unwinding the above recursion, it follows by $\frac{1}{2} \leq 1 - \frac{n\gamma\mu}{4} \leq 1$ for $\gamma \leq \frac{1}{10nd^2\kappa L}$ that

$$\mathbb{E} \left\| x_1^{(T)} - x^* \right\|^2 \leq \left(1 - \frac{n\gamma\mu}{4} \right)^T \left\| x_1^{(0)} - x^* \right\|^2 - \frac{n\gamma}{2L} \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|^2$$

$$+ \left(\frac{1}{n} + 6d^2\kappa L\gamma \right) \sum_{t=0}^{T} \left(1 - \frac{n\gamma\mu}{4} \right)^{T-t} \mathbb{E} \left\| \bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} \right\|_F^2$$

(24)

$$+ 60d^2\kappa L\gamma \sum_{t=0}^{T} \left(1 - \frac{n\gamma\mu}{4} \right)^{T-t} \mathbb{E} \left\| \boldsymbol{\Pi}_{\mathbf{u}} \mathbf{X}^t \right\|_F^2.$$

Step 2. To refine Lemma 3.2, we start with Equation (11) multiplied by the coefficient $\left(1 - \frac{n\gamma\mu}{4} \right)^{T-t}$ before summing over t in Equation (10) from 0 to T. Then, we have, for $\gamma \leq \frac{1}{10nd^2\kappa L}$ ($\frac{1}{2} \leq 1 - \frac{n\gamma\mu}{4} \leq 1$ for $\gamma \leq \frac{1}{10nd^2\kappa L}$), that

$$\sum_{t=0}^{T} \left(1 - \frac{n\gamma\mu}{4} \right)^{T-t} \mathbb{E} \| \bar{\mathbf{X}}^{(t+1)} - \bar{\mathbf{X}}^{(t)} \|_F^2 \leq 6\gamma^2 n^2 \sigma^2 (T+1) + 50\gamma^2 n^2 d^2 L^2 \sum_{t=0}^{T} \left(1 - \frac{n\gamma\mu}{4} \right)^{T-t} \mathbb{E} \left\| \boldsymbol{\Pi}_{\mathbf{u}} \mathbf{X}^{(t)} \right\|_F^2$$

$$+ 6\gamma^2 n^2 \sum_{t=0}^{d} \left(1 - \frac{n\gamma\mu}{4} \right)^{T-t} \left\| \nabla \mathbf{F}(\mathbf{X}^{(0)}) \right\|_F^2 + 15\gamma^2 n^3 \sum_{t=0}^{T} \mathbb{E} \left\| \nabla f(x_1^{(t)}) \right\|_2^2.$$

(25)

29

where we get the result which only modifies the coefficient of the term $\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2$.

Implementing the above result, we have, for $\gamma \leq \frac{1}{10nd^2\kappa L}\left(\leq \frac{1}{10ndL}\right)$,

$$
\begin{aligned}
\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 \leq{}& \left(1 - \frac{n\gamma\mu}{4}\right)^T \left\|x_1^{(0)} - x^*\right\|^2 \\
&+ \left(-\frac{n\gamma}{2L} + 15\gamma^2 n^2 + 90\gamma^3 n^3 d^2 \kappa L\right) \sum_{t=0}^T \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|^2 \\
&+ 6\gamma^2 n\sigma^2 (T+1) + 36\gamma^3 n^2 d^2 \kappa L\sigma^2 (T+1) \\
&+ \left(60\gamma d^2 \kappa L + 50\gamma^2 nd^2 L^2 + 300\gamma^3 n^2 d^4 \kappa L^3\right) \sum_{t=0}^T \left(1 - \frac{n\gamma\mu}{4}\right)^{T-t} \mathbb{E}\left\|\mathbf{\Pi_u X}^t\right\|_F^2 \\
&+ \left(6\gamma^2 nd + 36\gamma^3 n^2 d^3 \kappa L\right)\left(1 - \frac{n\gamma\mu}{4}\right)^{T-d} \left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|.
\end{aligned}
$$

Step 3. Similarly, we refine Lemma 3.3 as follows. By multiplying $\left(1 - \frac{n\gamma\mu}{4}\right)^{T-t}$ before summing over t in Equation (14), we have

$$
\begin{aligned}
\sum_{t=0}^T \left(1 - \frac{n\gamma\mu}{4}\right)^{T-t} \left\|\mathbf{\Pi_u X}^{(t)}\right\|_F^2 \leq{}& 300\gamma^2 n^2 d^4 (T+1)\sigma^2 + 20\gamma^2 n^3 d^2 \sum_{t=0}^T \mathbb{E}\|\nabla f(x_1^{(t)})\|_2^2 \\
&+ 6nd\left\|\mathbf{\Pi_u X}^{(0)}\right\|_F^2 + 40\gamma^2 n^2 d^3 \left(1 - \frac{n\gamma\mu}{4}\right)^{T-d} \left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2,
\end{aligned}
$$

Implementing the above result, we have for $\gamma \leq \frac{1}{40nd^2\kappa L}$ that

$$
60\gamma d^2 \kappa L + 50\gamma^2 nd^2 L^2 + 300\gamma^3 n^2 d^4 \kappa L^3 \leq 70\gamma d^2 \kappa L,
$$

and

$$
\begin{aligned}
\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 \leq{}& \left(1 - \frac{n\gamma\mu}{4}\right)^T \left\|x_1^{(0)} - x^*\right\|^2 \\
&+ \left(-\frac{n\gamma}{2L} + 15\gamma^2 n^2 + 90\gamma^3 n^3 d^2 \kappa L + 1400\gamma^3 n^3 d^4 \kappa L\right) \sum_{t=0}^T \mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|^2 \\
&+ 6\gamma^2 n\sigma^2 (T+1) + 36\gamma^3 n^2 d^2 \kappa L\sigma^2 (T+1) + 21000\gamma^3 n^2 d^6 \kappa L\sigma^2 (T+1) \\
&+ \left(6\gamma^2 nd + 36\gamma^3 n^2 d^3 \kappa L + 2800\gamma^3 n^2 d^5 \kappa L\right)\left(1 - \frac{n\gamma\mu}{4}\right)^{T-d} \left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\| + 420\gamma^3 n^2 d^3 \kappa L\left\|\mathbf{\Pi_u X}^{(0)}\right\|_F^2.
\end{aligned}
$$

Step 4. Note that the coefficient of the term $\sum_{t=0}^T \left\|\nabla f(x_1^{(t)})\right\|^2$ is smaller than 0 for $\gamma \leq \frac{1}{100nd^2\kappa L}$, derived as follows:

$$
\begin{aligned}
&-\frac{n\gamma}{2L} + 15\gamma^2 n^2 + 90\gamma^3 n^3 d^2 \kappa L + 1400\gamma^3 n^3 d^4 \kappa L \\
={}&\frac{n\gamma}{2L}\left(-1 + 30\gamma nL + 180\gamma^2 n^2 d^2 \kappa L^2 + 2800\gamma^2 nd^4 \kappa L\right) \\
\leq{}&\frac{n\gamma}{2L}\left(-1 + \frac{30}{100} + \frac{180}{10^4} + \frac{2800}{10^4}\right) \leq -\frac{\gamma n}{20L} < 0.
\end{aligned}
$$

Thus, it holds that

$$
\begin{aligned}
\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 \leq{}& \left(1 - \frac{n\gamma\mu}{4}\right)^T \left\|x_1^{(0)} - x^*\right\|^2 \\
&+ 7\gamma^2 n\sigma^2 (T+1) + 21000\gamma^3 n^2 d^6 \kappa L\sigma^2 (T+1) \\
&+ 80\gamma^2 nd^3 \left(1 - \frac{n\gamma\mu}{4}\right)^{T-d} \left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\| + 420\gamma^3 n^2 d^3 \kappa L\left\|\mathbf{\Pi_u X}^{(0)}\right\|_F^2.
\end{aligned}
$$

30

A.4 Proof of the Convergence Results

A.4.1 Proof of Theorem 2.1

Invoking Lemma 3.5, with identical initial values $x_i^{(0)}$ that implies $\left\|\mathbf{\Pi_u X}^{(0)}\right\|_F^2 = 0$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 \leq \frac{8\Delta_f}{\gamma n(T+1)} + 24\gamma\sigma^2 L + 2\cdot 10^4\gamma^2 nd^6\sigma^2 L^2 + \frac{56\gamma d^3 L\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2}{T+1}.$$

Referring to Lemma 26 in [8], as stated in Lemma A.7, by taking $A = \frac{8\Delta_f}{n}$, $B = 24\sigma^2 L$, $C = 20000nd^6\sigma^2 L^2$ and $\alpha = 100nd^3 L$, when considering $\gamma = \min\{\left(\frac{\Delta_f}{3\sigma^2 Ln(T+1)}\right)^{\frac{1}{2}}, \left(\frac{\Delta_f}{1500n^2 d^6\sigma^2 L^2(T+1)}\right)^{\frac{1}{3}}, \frac{1}{100nd^3 L}\}$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(x_1^{(t)})\right\|_2^2 \leq \frac{32\sqrt{\Delta_f\sigma^2 L}}{\sqrt{n(T+1)}} + \frac{240d^2\left(\sigma^2 L^2\Delta_f^2\right)^{\frac{1}{3}}}{\left(\sqrt{n(T+1)}\right)^{\frac{2}{3}}} + \frac{800d^3 L\Delta_f}{T+1} + \frac{\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2}{n(T+1)}.$$

A.4.2 Proof of Theorem 2.4

Invoking Lemma 3.7, with identical values $x_i^{(0)}$ that implies $\left\|\mathbf{\Pi_u X}^{(0)}\right\|_F^2 = 0$, we have

$$\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 \leq \left(1 - \frac{n\gamma\mu}{4}\right)^T\left\|x_1^{(0)} - x^*\right\|^2$$
$$+ 7\gamma^2 n\sigma^2(T+1) + 21000\gamma^3 n^2 d^6\kappa L\sigma^2(T+1)$$
$$+ 80\gamma^2 nd^3\left(1 - \frac{n\gamma\mu}{4}\right)^{T-d}\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|.$$

Considering $\gamma = \min\left\{\frac{1}{100nd^2\kappa L}, \frac{16\log(n(T+1)^2)}{n(T+1)\mu}\right\}$, for $T \geq 2d$ we get that

$$\left(1 - \frac{n\mu\gamma}{4}\right)^T \leq \max\left\{\left(1 - \frac{1}{400d^2\kappa^2}\right)^T, \left(1 - \frac{4\log(n(T+1)^2)}{T+1}\right)^T\right\}$$

$$\leq \max\left\{\exp(-\frac{T}{400d^2\kappa^2}), \frac{40}{n(T+1)^2}\right\},$$

and

$$\left(1 - \frac{n\mu\gamma}{4}\right)^{T-d} \leq \max\left\{\left(1 - \frac{1}{400d^2\kappa^2}\right)^{T/2}, \left(1 - \frac{4\log(n(T+1)^2)}{T+1}\right)^{T/2}\right\}$$

$$\leq \max\left\{\exp(-\frac{T}{800d^2\kappa^2}), \frac{10}{n(T+1)^2}\right\},$$

where we use the fact $(1 - \frac{1}{x})^x \leq e^{-1}$ and $1 - x \leq \exp(-x)$ for any $x \in \mathbb{R}_+$. Then,

$$\mathbb{E}\left\|x_1^{(T)} - x^*\right\|^2 \leq \frac{2240\sigma^2\log(n(T+1)^2)}{n(T+1)\mu^2} + \frac{26880000d^6\kappa^2\sigma^2\left(\log(n(T+1)^2)\right)^2}{n(T+1)^2\mu^2}$$

$$+ \max\left\{\exp(-\frac{T}{800d^2\kappa^2}), \frac{40}{n(T+1)^2}\right\}\left(\left\|x_1^{(0)} - x^*\right\|^2 + \frac{1}{nL^2}\left\|\nabla\mathbf{F}(\mathbf{X}^{(0)})\right\|_F^2\right).$$

B Additional Experiments

For the problem of training a CNN on the MNIST dataset, we have further compared the real-time performance of BTPP with other representative methods. The experiments are conducted on a server equipped with eight Nvidia RTX 3090 GPUs and two Intel Xeon Gold 4310 CPUs, where the

31
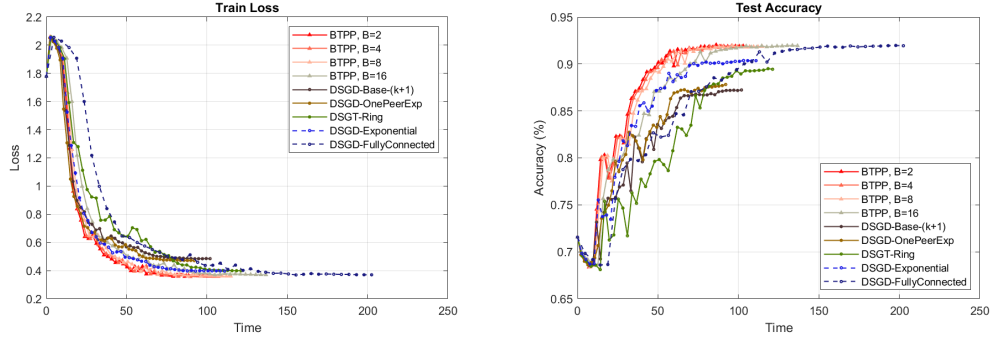
Figure 4: Real-time performance of BTPP (with different branch size B) compared with related methods when training CNN over MNIST.

communication between GPUs follows the topology requirement of each algorithm. We measure the running time including GPU computation and communication for 13,000 iterations. The experimental settings are consistent with those described in Section 4.2. From Figure 4, BTPP outperforms the other algorithms concerning the running time. Additionally, we evaluate BTPP with various branch sizes B, concluding that for relatively small values of n, a branch size of $B = 2$ is most effective.

Furthermore, we consider training VGG13 on the CIFAR10 dataset, with $n = 8$ and a batch size of 16. The learning rate and topology configurations are consistent with those described in Section 4.2. Additionally, the case of BTPP with $B = 8$ is equivalent to DSGD in a fully connected setting, meaning that they produce identical outputs when using the same random seed. Figure 5 and Figure 6 illustrate that BTPP beats competing algorithms in terms of the convergence rate (against iteration number) and running time. Moreover, a branch size of $B = 2$ is optimal.
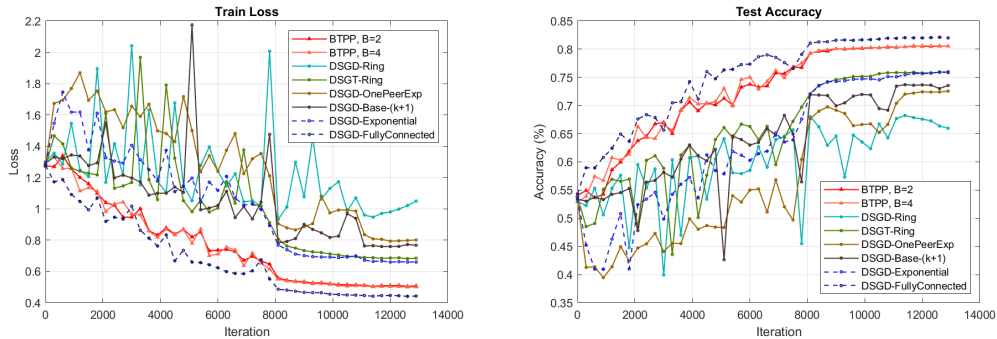
Figure 5: Performance of BTPP (with different branch size B) compared with related methods for training VGG13 over CIFAR10.

We further demonstrate that the performance of BTPP can be improved by incorporating a momentum term (with momentum parameter set to 0.9) when data heterogeneity exists or by removing the data heterogeneity, which involves randomly assigning samples to each agent; see Figure 7 and Figure 8.

32
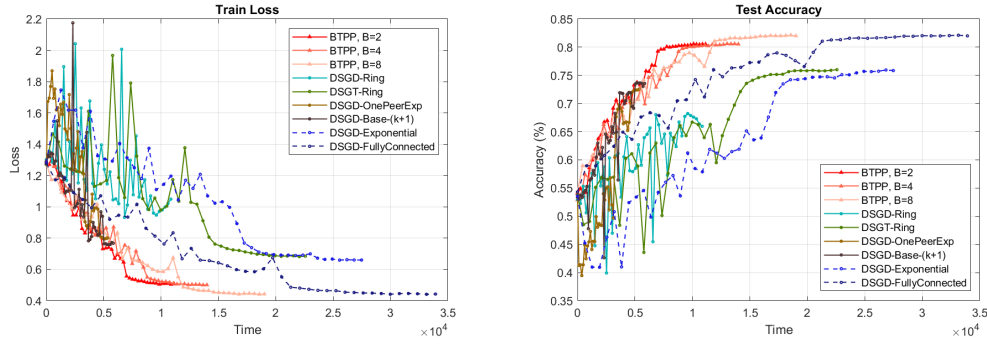
Figure 6: Real-time performance of BTPP (with different branch size B) compared with related methods when training VGG13 over CIFAR10.
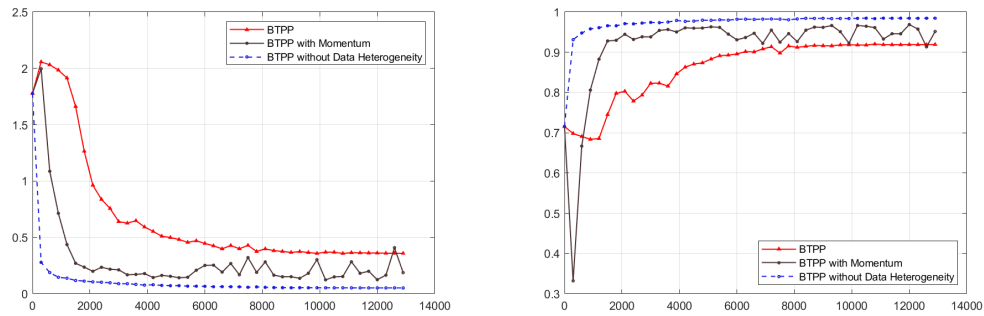
Figure 7: Performance of BTPP with branch size $B = 2$ under various configurations when training CNN over MNIST.
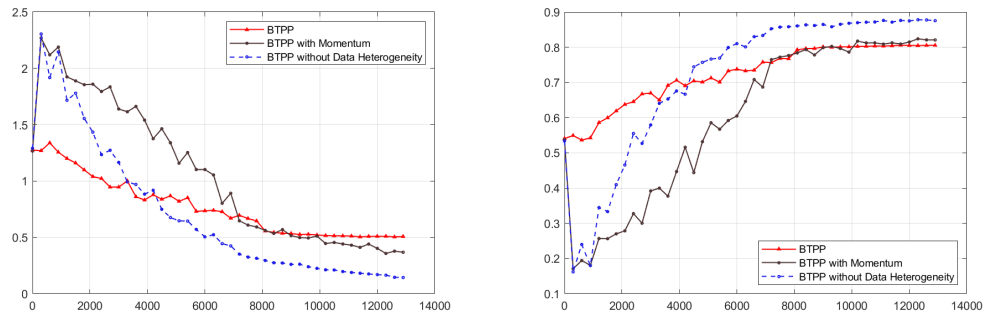
Figure 8: Performance of BTPP with branch size $B = 2$ under various configurations when training VGG13 over CIFAR10.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our paper meticulously delineates its primary contributions in both the abstract (lines 8-9) and the introduction (subsection 1.2, lines 105-121).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

33

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: At the end of section 1.1, lines 100-104.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are in subsection 1.3, lines 129-132. The proofs of all theoretical results are in the Appendix and subsection 2.1.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

34

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We present the experiment details in section 4 and upload our code in the link shown in the abstract.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code with the link shown in lines 10-11.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

35

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the experimental details and settings in Section 4 and our code link.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We have repeated some experiments with different seeds and reported the averaged performance. See Section 4 for details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

36

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See line 255 for reference.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

37

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited the original paper that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer:[NA]

Justification: The paper does not release new assets.

38

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

39