# ACHIEVING OPTIMAL COMPLEXITY IN DECENTRALIZED LEARNING OVER ROW-STOCHASTIC NETWORKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A key challenge in decentralized optimization is determining the optimal convergence rate and designing algorithms that can achieve it. While this issue has been thoroughly addressed for doubly-stochastic and column-stochastic mixing matrices, the row-stochastic setting remains largely unexplored. This study establishes the first convergence lower bound for decentralized learning over row-stochastic networks. However, developing algorithms to achieve this lower bound is highly challenging due to several factors: (i) the widely used ROW-ONLY gossip protocol, PULL-DIAG, suffers from significant instability in achieving average consensus; (ii) PULL-DIAG-based algorithms are sensitive to data heterogeneity; and (iii) there has been no analysis in nonconvex and stochastic settings to date. This work addresses these deficiencies by proposing and analyzing a new gossip protocol called PULL-SUM, along with its gradient tracking extension, Pull-Sum-GT. The PULL-SUM protocol mitigates the instability issues of PULL-DIAG, while PULL-SUM-GT achieves the first linear speedup convergence rate without relying on data heterogeneity assumptions. Additionally, we introduce a multi-step strategy that enables PULL-SUM-GT to match the established lower bound up to logarithmic factors, demonstrating its near-optimal performance and the tightness of our established lower bound. Experiments validate our theoretical results.

## 1 INTRODUCTION

Scaling machine learning tasks to large datasets and models requires efficient distributed computing across multiple nodes. This paper investigates decentralized stochastic optimization over a network of $n$ nodes:

$$\min_{x\in\mathbb{R}^d} \quad f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{where} \quad f_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F(x;\xi_i)]. \tag{1}$$

Here, $\xi_i$ is a random data vector supported on $\Xi_i \subseteq \mathbb{R}^q$ with some distribution $\mathcal{D}_i$, and $F : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$ is a Borel measurable function. Each loss function $f_i$ is accessible only by node $i$ and is assumed to be smooth and potentially non-convex. Note that data heterogeneity typically exists, i.e., the local data distributions $\{\mathcal{D}_i\}_{i=1}^n$ vary across nodes. The decentralized communication among nodes is represented as a strongly connected *directed graph*, which is practically valuable in real applications. For example, bidirectional communication may be infeasible due to nodes having different power ranges (Yang et al., 2019) or experiencing channel disruptions. In distributed deep learning model training, well-designed directed topologies often result in sparser and faster communication compared to undirected ones, thus accelerating training in terms of wall-clock time (Bottou et al., 2018; Assran et al., 2019; Yuan et al., 2021).

**Network topology and mixing matrix.** A key challenge in decentralized optimization is to determine the optimal convergence rate and design algorithms that achieve it. Addressing this challenge requires a theoretical characterization of how network topologies influence decentralized algorithms. For a given connected network, we represent the topology using a mixing matrix that follows its connectivity pattern, serving as an effective tool for evaluating the network's impact. For undirected networks, a symmetric and doubly-stochastic matrix can be easily constructed. However, in directed networks, constructing a doubly-stochastic mixing matrix is generally impossible. Instead, mixing matrices are typically either column-stochastic (Nedić & Olshevsky, 2014; Nedić et al., 2017) or row-stochastic (Sayed, 2014; Mai & Abed, 2016), but not both.

**Optimal complexity over doubly-stochastic networks is well-established.** The connectivity of a doubly-stochastic mixing matrix can be effectively evaluated through a metric called spectral gap, which measures how closely the decentralized network approximates a fully connected network. Building on this metric, a series of works have established the optimal convergence rates for decentralized algorithms. For instance, the studies in (Scaman et al., 2017; 2018; Sun & Hong, 2019; Kovalev et al., 2021) provide optimal convergence rates for convex or non-stochastic decentralized optimization. Lu & De Sa (2021) establishes the optimal complexity for non-convex and stochastic decentralized optimization over a specific type of linear networks, while (Yuan et al., 2022) extends this optimal complexity to a much broader class of networks.

**Optimal complexity over column-stochastic networks is established recently.** If out-degree information is available prior to communication, a column-stochastic matrix can be easily constructed. When only column-stochastic matrices are used in decentralized algorithms, this is referred to as the COL-ONLY setting. The foundation of COL-ONLY algorithms is the PUSH-SUM gossip protocol (Kempe et al., 2003; Tsianos et al., 2012). Many algorithms based on PUSH-SUM achieve superior convergence rates, e.g., Nedić & Olshevsky (2015); Tsianos et al. (2012); Zeng & Yin (2017); Xi & Khan (2017); Xi et al. (2017); Nedić et al. (2017); Assran et al. (2019); Qureshi et al. (2020). However, these works do not precisely capture the influence of column-stochastic networks and, therefore, cannot clarify the optimal complexity in the COL-ONLY setting. This open question has been addressed in a recent study by Liang et al. (2023), which establishes effective metrics to evaluate the influence of column-stochastic networks and provides the optimal lower bound for the COL-ONLY setting. Additionally, it proposes algorithms that achieve this lower bound.

**Optimal complexity over row-stochastic networks remains unclear yet.** If out-degree information is unavailable, column-stochastic matrices cannot be directly constructed. However, row-stochastic matrices can be formed using in-degree information, which can be obtained by counting received messages. This is referred to as the ROW-ONLY setting. Similar to how PUSH-SUM serves as the basis for COL-ONLY algorithms, the foundation of ROW-ONLY methods is the PULL-DIAG gossip protocol (Mai & Abed, 2016). Building on PULL-DIAG, Mai & Abed (2016) adapted the distributed gradient descent (DGD) algorithm for the ROW-ONLY setting, while Li et al. (2019); Xin et al. (2019c) extended gradient tracking methods, and Ghaderyan et al. (2021); Lü et al. (2020); Xin et al. (2019a) introduced momentum-based ROW-ONLY gradient tracking. However, the convergence analysis for ROW-ONLY algorithms is still quite limited. Current analyses focus only on deterministic and strongly convex loss functions, leaving the performance of ROW-ONLY algorithms in non-convex and stochastic settings unknown. More importantly, the impact of row-stochastic networks on the convergence rate of ROW-ONLY algorithms remains unclear. These gaps present significant obstacles to determining the optimal complexity in the ROW-ONLY setting. The following fundamental open problems then naturally arise:

- Q1. What are the effective metrics that can fully capture the impact of row-stochastic networks on decentralized stochastic optimization, and how do they influence the convergence of prevalent ROW-ONLY algorithms?
- Q2. Given these metrics, what is the lower bound on the convergence rate for ROW-ONLY algorithms in the non-convex and stochastic setting?
- Q3. Can existing ROW-ONLY algorithms readily achieve the optimal convergence rate? If not, what limitations do they face?
- Q4. Can we develop new ROW-ONLY algorithms that overcome the limitations of existing algorithms and attain the aforementioned lower bound?

**Main contributions.** This paper provides an in-depth understanding of decentralized optimization over row-stochastic networks by addressing the above open questions. Our contributions are:

- C1. We find that the metrics *generalized spectral gap* and *equilibrium skewness*, proposed by Liang et al. (2023) to characterize the influence of column-stochastic networks, can also effectively capture the impact of row-stochastic networks on decentralized algorithms.
- C2. Using these metrics, we establish the *first* lower bound on the convergence rate for any non-convex decentralized stochastic first-order algorithm with a row-stochastic mixing matrix.

This bound reflects the optimal influence of gradient noise, the mixing matrix, the number of nodes, and problem smoothness on the algorithm.

C3. We find that existing ROW-ONLY algorithms cannot attain the aforementioned lower bound due to two limitations. First, the PULL-DIAG protocol involves the inversion of small values during its operation, leading to instability in ROW-ONLY algorithms. Second, improper algorithmic construction makes current ROW-ONLY algorithms highly sensitive to data heterogeneity. However, neither the instability nor data heterogeneity affects our lower bound, suggesting that these issues can be eliminated with improved algorithm design.

C4. We develop novel ROW-ONLY algorithms to achieve the established lower bound. First, we propose a PULL-SUM gossip protocol that avoids the inversion of small values. Next, we introduce a new row-stochastic gradient tracking structure that removes the impact of data heterogeneity. Together with a multi-step gossip protocol, these techniques will yield an effective algorithm that nearly attains the established lower bound, demonstrating its near-optimal performance and the tightness of the established lower bound.

**Notations.** Let $\mathbb{1}_n$ denote the $n$-dimensional all-ones vector, and $I_n \in \mathbb{R}^{n\times n}$ the identity matrix. We let matrix $A$ denote the row-stochastic matrix ($A\mathbb{1}_n = \mathbb{1}_n$) and $B$ denote the column-stochastic ($\mathbb{1}_n^\top B = \mathbb{1}_n^\top$) matrix. The set $[n]$ represents the indices $\{1, 2, \ldots, n\}$. $\mathrm{Diag}(A)$ refers to the diagonal matrix formed by $A$'s diagonal entries, and $\mathrm{diag}(v)$ is the diagonal matrix formed from vector $v$. The Perron vectors of $A$ and $B$ are $\pi_A$ and $\pi_B$, respectively, with $\Pi_A = \mathrm{diag}(\pi_A)$ and $\Pi_B = \mathrm{diag}(\pi_B)$. We define $\|v\|_{\pi_A} = \|\Pi_A^{1/2} v\|$ and $\|v\|_{\pi_B} = \|\Pi_B^{-1/2} v\|$, with corresponding induced matrix norm $\|W\|_{\pi_A} = \|\Pi_A^{1/2} W \Pi_A^{-1/2}\|_2$ and $\|W\|_{\pi_B} = \|\Pi_B^{-1/2} W \Pi_B^{1/2}\|_2$. We define $A_\infty = \mathbb{1}_n \pi_A^\top$ and $B_\infty = \pi_B \mathbb{1}_n^\top$. Vector $\boldsymbol{x}_i^{(k)} \in \mathbb{R}^d$ denotes the local model at node $i$ at iteration $k$. We also define

$$\mathbf{x}^{(k)} := [(\boldsymbol{x}_1^{(k)})^\top; (\boldsymbol{x}_2^{(k)})^\top; \cdots; (\boldsymbol{x}_n^{(k)})^\top] \in \mathbb{R}^{n\times d},$$

$$\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) := [\nabla F_1(\boldsymbol{x}_1^{(k)}; \xi_1^{(k)})^\top; \cdots; \nabla F_n(\boldsymbol{x}_n^{(k)}; \xi_n^{(k)})^\top] \in \mathbb{R}^{n\times d},$$

by stacking all local variables. The upright bold symbols (e.g. $\mathbf{x}, \mathbf{w}, \mathbf{g} \in \mathbb{R}^{n\times d}$) always denote stacked network-level quantities.

## 2 EFFECTIVE METRICS FOR ROW-STOCHASTIC NETWORKS

We consider a directed network with $n$ computing nodes that is associated with a mixing matrix $A = [a_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ where $a_{ij} \in (0, 1)$ if node $j$ can send information to node $i$ otherwise $a_{ij} = 0$. Decentralized optimization is built upon partial averaging $\boldsymbol{z}_i^+ = \sum_{j\in\mathcal{N}_i} a_{ij} \boldsymbol{z}_j$ in which $\boldsymbol{z}_i \in \mathbb{R}^d$ is a local vector held by node $i$ and $\mathcal{N}_i$ denotes the in-neighbors of node $i$, including node $i$ itself. Since every node conducts partial averaging simultaneously, we have

$$\mathbf{z} \triangleq [\boldsymbol{z}_1^\top; \boldsymbol{z}_2^\top; \cdots; \boldsymbol{z}_n^\top] \xmapsto{\text{A-protocol}} \mathbf{z}^+ = A\mathbf{z} = [\sum_{j\in\mathcal{N}_1} a_{1j} \boldsymbol{z}_j^\top; \cdots; \sum_{j\in\mathcal{N}_n} a_{nj} \boldsymbol{z}_j^\top] \tag{2}$$

where $A$-protocol represents partial averaging with mixing matrix $A$. Evidently, the algebraic characteristics of $A$ substantially affects the convergence of partial averaging and the corresponding decentralized optimization. This section explores metrics that capture the characteristics of $A$.

### 2.1 ROW-STOCHASTIC MIXING MATRIX

This paper focuses on a static directed network $\mathcal{G}$ associated with a row-stochastic matrix $A$.

**Assumption 1** (PRIMITIVE AND ROW-STOCHASTIC MIXING MATRIX)**.** *The mixing matrix $A$ is non-negative, primitive, and satisfies $A\mathbb{1}_n = \mathbb{1}_n$.*

If $\mathcal{G}$ is strongly-connected, *i.e.*, there exists a directed path from each node to every other node, and $A$ has a positive trace, then $A$ is primitive. It is straightforward to make $A$ row-stochastic by setting $a_{ij} = 1/(1 + d_i^{\text{in}})$ if $(i, j) \in \mathcal{E}$ or $j = i$ otherwise $a_{ij} = 0$, where $\mathcal{E}$ is the set of directed edges and $d_i^{\text{in}}$ is the in-degree of node $i$ excluding the self-loop. With Assumption 1, Perron-Frobenius theorem (Perron, 1907) ensures a unique equilibrium vector $\pi_A \in \mathbb{R}^n$ with positive entries so that

$$\pi_A^\top A = \pi_A^\top, \quad \mathbb{1}_n^\top \pi_A = 1, \quad \text{and} \quad \lim_{k\to\infty} A^k = \mathbb{1}_n \pi_A^\top.$$

### 2.2 Effective metrics to capture the impact of row-stochastic weight matrix

Most decentralized algorithms rely on gossip protocols like (2), where local variables are partially mixed to approximate the global average. The properties of the mixing matrix $A$ are crucial in determining whether this gossip mixing can achieve the global average and how efficiently this process occurs. These properties will translate into effective metrics for evaluating the influence of $A$ on algorithmic performance.

Now we examine the gossip process with a row-stochastic mixing matrix $A$. Suppose that each node $i$ has a local variable $\boldsymbol{z}_i \in \mathbb{R}^d$, we let $\mathbf{z} = [\boldsymbol{z}_1^\top; \boldsymbol{z}_2^\top; \cdots; \boldsymbol{z}_n^\top] \in \mathbb{R}^{n\times d}$ and initialize $\mathbf{x}^{(0)} = \mathbf{z}$. Following the gossip protocol as in (2), we have the following recursions:

$$\mathbf{x}^{(k)} = A\mathbf{x}^{(k-1)} = A^k\mathbf{x}^{(0)} \xmapsto{k\to\infty} \mathbb{1}_n\pi_A^\top\mathbf{x}^{(0)} = \mathbb{1}_n\pi_A^\top\mathbf{z}, \tag{3}$$

where we utilize the property that $\lim_{k\to\infty} A^k = \mathbb{1}_n\pi_A^\top$. It is evident that the matrix $A$ influences both whether and how quickly $\mathbf{x}^{(k)}$ approaches the global average $(1/n)\mathbb{1}_n\mathbb{1}_n^\top\mathbf{z}$. Inspired by (3), we propose the following two metrics to capture the impact of the row-stochastic matrix $A$:

- The **equilibrium skewness**

  $$\kappa_A := \max(\pi_A)/\min(\pi_A) \in [1, +\infty)$$

  captures the disagreement between the equilibrium vector $\pi_A$ and the uniform vector $n^{-1}\mathbb{1}_n$. When $\kappa_A \to 1$, the weighted average $\mathbb{1}_n\pi_A^\top\mathbf{z}$ aligns better with the global average $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\mathbf{z}$.
- The **generalized spectral gap** $1-\beta_A$ of the row-stochastic matrix $A$, where

  $$\beta_A := \left\|A - \mathbb{1}_n\pi_A^\top\right\|_{\pi_A} = \|A - A_\infty\|_{\pi_A} \in [0, 1)$$

  quantifies the convergence rate of $\mathbf{x}^{(k)}$ to the weighted global average $\mathbb{1}_n\pi_A^\top\mathbf{z}$ in (3). As $\beta_A$ approaches 0, the iterates $\mathbf{x}^{(k)}$ converge more rapidly to the weighted global average.

It is important to note that these two metrics are not new; they were proposed in (Xin et al., 2019b; Liang et al., 2023) to assess the influence of column-stochastic mixing matrices. Our contribution lies in demonstrating that these metrics are also applicable to row-stochastic mixing matrices.

Another remark is that the standard gossip protocol using a row-stochastic matrix $A$ in (3) cannot achieve the global average. However, this issue can be resolved with an enhanced gossip protocol called Pull-Diag (Mai & Abed, 2016; Xi et al., 2018), which will be discussed in Section 4.

## 3 Convergence Lower Bounds over Row-Stochastic Networks

### 3.1 Assumptions

This subsection specifies the category of decentralized algorithms to which our lower bound applies.

**Function class.** We define the function class $\mathcal{F}_{\Delta,L}$ as the set of functions that satisfy Assumption 2, for any given dimension $d \in \mathbb{N}_+$ and any initialization point $x^{(0)} \in \mathbb{R}^d$.

**Assumption 2** (Smoothness). *There exists a constant $L, \Delta \geq 0$ such that*

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|,$$

*for all $1 \leq i \leq n, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, and $f(\boldsymbol{x}^{(0)}) - \inf_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \leq \Delta$.*

**Gradient oracle class.** We assume that each node $i$ processes its local cost function $f_i$ using a stochastic gradient oracle $\nabla F(\boldsymbol{x}; \xi_i)$, which provides unbiased estimates of the exact gradient $\nabla f_i$ with bounded variance. Specifically, we define the stochastic gradient oracle class $O_{\sigma^2}$ as the set of all oracles $\nabla F(\cdot; \xi_i)$ that satisfy Assumption 3.

**Assumption 3** (Gradient oracles). *There exists a constant $\sigma \geq 0$ such that*

$$\mathbb{E}[\nabla F(\boldsymbol{x}; \xi_i)] = \nabla f_i(\boldsymbol{x}),\ \mathbb{E}[\|\nabla F(\boldsymbol{x}; \xi_i) - \nabla f_i(\boldsymbol{x})\|^2] \leq \sigma^2,\ \forall\, \boldsymbol{x} \in \mathbb{R}^d,\ \forall\, 1 \leq i \leq n.$$

**Algorithm class description.** We focus on decentralized algorithms where each node $i$ maintains a local solution $\boldsymbol{x}_i^{(k)}$ at iteration $k$ and communicates using the $A$-protocol defined in (2). These algorithms also adhere to the linear-spanning property, as defined in prior works (Carmon et al., 2020; 2021; Yuan et al., 2022; Lu & De Sa, 2021). Informally, this property ensures that each local solution $\boldsymbol{x}_i^{(k)}$ resides within the linear space spanned by $\boldsymbol{x}_i^{(0)}$, its local stochastic gradients, and interactions with neighboring nodes. Upon completing $K$ iterations, the final output $\hat{\boldsymbol{x}}^{(K)}$ can be any variable in $\mathrm{span}(\{\{\boldsymbol{x}_i^{(k)}\}_{i=1}^n\}_{k=0}^K)$. Let $\mathcal{A}_A$ denote the set of all algorithms that adhere to partial averaging via mixing matrix $A$ and satisfy the linear-spanning property.

### 3.2 Lower bound

With $\beta_A$ and $\kappa_A$ at hand, we show, for the first time, that the convergence rate of any non-convex decentralized stochastic first-order algorithm with a row-stochastic mixing matrix is lower bounded by the following theorem.

**Theorem 1** (Lower bound)**.** *For any given $L \geq 0$, $n \geq 2$, $\sigma \geq 0$, and $\tilde{\beta} \in [\Omega(1), 1 - 1/n]$, there exists a set of loss functions $\{f_i\}_{i=1}^n \in \mathcal{F}_{\Delta,L}$, a set of stochastic gradient oracles in $\mathcal{O}_{\sigma^2}$, and a row-stochastic matrix $A \in \mathbb{R}^{n\times n}$ with $\beta_A = \tilde{\beta}$ and $\ln(\kappa_A) = \Omega(n(1-\beta_A))$, such that the convergence of any algorithm $\mathscr{A} \in \mathcal{A}_A$ starting from $\boldsymbol{x}_i^{(0)} = \boldsymbol{x}^{(0)}, i \in [n]$ with $K$ iterations is lower bounded by*

$$\mathbb{E}[\|\nabla f(\hat{\boldsymbol{x}}^{(K)})\|^2] = \Omega\left(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}} + \frac{(1+\ln(\kappa_A))L\Delta}{(1-\beta_A)K}\right). \tag{4}$$

where $K$, $\sigma$, $L$, and $\Delta$ represent the total number of iterations, gradient variance, smoothness parameter of the functions, and the initial gap in function values, respectively. The lower bound in (4) explicitly demonstrates the combined influence of the generalized spectral gap $\beta_A$ and the equilibrium skewness $\kappa_A$ on decentralized algorithms employing row-stochastic weight matrices.

## 4 Limitations in existing Row-Only algorithms

This section examines the convergence of several existing Row-Only algorithms and identifies the limitations that prevent them from reaching the established lower limit.

### 4.1 Pull-Diag protocol suffers from instability

**Algorithm review.** According to (3), for a row-stochastic matrix $A$, the convergence $A^k \to \mathbb{1}_n\pi_A^\top$ results in a biased weighted average during gossiping, *i.e.*,

$$[\text{Gossip}]: \qquad \mathbf{x}^{(k)} = A\mathbf{x}^{(k-1)} = A^k\mathbf{x}^{(0)} \xmapsto{k\to\infty} \mathbb{1}_n\pi_A^\top\mathbf{x}^{(0)} \neq \bar{\mathbf{x}}^{(0)}, \tag{5}$$

where $\bar{\mathbf{x}}^{(0)} := n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\mathbf{x}^{(0)}$ is the desired global average. The Pull-Diag protocol, commonly used in Row-Only optimization (Mai & Abed, 2016; Xi et al., 2018; Li et al., 2019; Xin et al., 2019c; Ghaderyan et al., 2021), corrects this bias by utilizing the diagonal entries of $A^k$, *i.e.*,

$$[\text{Pull-Diag}]: \quad \mathbf{x}^{(k)} = n^{-1}A^k\mathrm{Diag}(A^k)^{-1}\mathbf{x}^{(0)} \xmapsto{k\to\infty} n^{-1}\mathbb{1}_n\pi_A^\top\mathrm{diag}(\pi_A^{-1})\mathbf{x}^{(0)} = \bar{\mathbf{x}}^{(0)}. \tag{6}$$

It is evident that the inversion of the diagonal entries of $A^k$ plays a crucial role in correcting the bias inherent in the vanilla gossip protocol.

**Limitation.** A key limitation of Pull-Diag is its instability when diagonal entries of $A^k$ approaches zero. To ensure the protocol remains well-defined, an additional assumption is required to provide a lower bound on the diagonal entries of $A^k$ across all iterations. We present this assumption along with the following lemma to ensure the convergence of Pull-Diag.

**Proposition 1** (Pull-Diag convergence)**.** *For a row-stochastic and primitive matrix $A$, if $\max_k \|\mathrm{Diag}(A^k)^{-1}\|_2 = \theta_A > 0$, then* Pull-Diag *converges at the following rate:*

$$\|n^{-1}\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(0)}\|_F \leq \min\{1+\theta_A, 2\theta_A\kappa_A^{1.5}\beta_A^k\}\|\mathbf{x}^{(0)}\|_F. \tag{7}$$

The convergence rate of PULL-DIAG is influenced by $\theta_A$, which arises from the inversion of small values. Notably, $\theta_A$ is independent of $\beta_A$ and $\kappa_A$ and can become arbitrarily large, making PULL-DIAG highly unstable when $\theta_A$ is substantial, as illustrated in Figure 1. Consequently, all algorithms built upon PULL-DIAG (Mai & Abed, 2016; Xi et al., 2018; Li et al., 2019; Xin et al., 2019c; Ghaderyan et al., 2021) suffer from this instability issue. However, our established lower bound in Theorem 1 is unaffected by $\theta_A$, suggesting that this impact can be eliminated. In Section 5, we introduce a new protocol PULL-SUM to address this issue.

### 4.2 PULL-DIAG-GT SUFFERS FROM DATA HETEROGENEITY

**Algorithm review.** Gradient tracking (Xu et al., 2015; Di Lorenzo & Scutari, 2016; Qu & Li, 2017; Nedic et al., 2017) is among the state-of-the-art algorithms in decentralized optimization. Initially designed for undirected networks, it has been extended by (Li et al., 2019; Xin et al., 2019c; Ghaderyan et al., 2021; Lü et al., 2020; Xin et al., 2019a) to the ROW-ONLY setting using the PULL-DIAG protocol. We revisit a representative algorithm FROST (Li et al., 2019; Xin et al., 2019c):

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} - \alpha \mathbf{y}^{(k)} \tag{8a}$$

$$\mathbf{y}^{(k+1)} = A\mathbf{y}^{(k)} + D_{k+1}^{-1}\mathbf{g}^{(k+1)} - D_k^{-1}\mathbf{g}^{(k)} \tag{8b}$$

where $D_k = \mathrm{Diag}(A^k)$, $\mathbf{x}^{(k)}$ is the variables, and $\mathbf{y}^{(k)}$ denotes the gradient tracking term. Specifically, $\mathbf{g}^{(k)}$ is the stochastic gradient, defined as $\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})$, with $\mathbf{y}^{(0)} = \mathbf{g}^{(0)}$. We refer to algorithms with structures similar to (8) as belonging to the PULL-DIAG-GT family.

**Algorithm insight.** The fundamental reason PULL-DIAG-GT is effective for row-stochastic networks is that it essentially functions as an asymptotic global gradient descent. To illustrate this, we can first check the gradient tracking part by left-multiplying $\pi_A^\top$ on both sides of (8b):

$$\pi_A^\top \mathbf{y}^{(k+1)} - \pi_A^\top D_{k+1}^{-1}\mathbf{g}^{(k+1)} \overset{(a)}{=} \pi_A^\top \mathbf{y}^{(k)} - \pi_A^\top D_k^{-1}\mathbf{g}^{(k)} = \cdots = \pi_A^\top \mathbf{y}^{(0)} - \pi_A^\top \mathbf{g}^{(0)} \overset{(b)}{=} 0, \tag{9}$$

where equality (a) holds because $\pi_A^T A = \pi_A^T$, and equality (b) holds due to $\mathbf{y}^{(0)} = \mathbf{g}^{(0)}$. The above equality indicates that $\pi_A^\top \mathbf{y}^{(k)} = \pi_A^T D_k^{-1}\mathbf{g}^{(k)}$. Similarly, we left-multiply $\pi_A^\top$ on both sides of (8a) and denote the weighted parameter $\boldsymbol{w}^{(k)} = \pi_A^\top \mathbf{x}^{(k)}$ :

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \alpha \pi_A^\top \mathbf{y}^{(k)} \overset{(9)}{=} \boldsymbol{w}^{(k)} - \alpha \pi_A^T D_k^{-1}\mathbf{g}^{(k)}. \tag{10}$$

Noting that $\pi_A^\top D_k^{-1} \to \mathbb{1}_n^\top, k \to \infty$, denote $\bar{\boldsymbol{g}}^{(k)} = n^{-1}\mathbb{1}_n^\top \mathbf{g}^{(k)}$, the iteration can be asymptotically written as $\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - n\alpha\bar{\boldsymbol{g}}^{(k)}$. As $\mathbf{x}^{(k)}$ achieve consensual $\boldsymbol{x}^{(k)}$ at each node at the end, this becomes $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - n\alpha\bar{\boldsymbol{g}}^{(k)}$, making it a centralized parallel SGD.

**Limitation.** While PULL-DIAG-GT is simple and effective, it suffers from two limitations:

- It builds on PULL-DIAG, which introduces instability due to the inversion of the diagonal entries of $A^k$. Consequently, its convergence is influenced by $\theta_A = \max_k \|\mathrm{Diag}(A^k)^{-1}\|_2$.
- It suffers from data heterogeneity. As seen from (9), the effectiveness of PULL-DIAG-GT stems from the fact that $n^{-1}\pi_A^\top \mathbf{y}^{(k)}$ asymptotically approaches the globally averaged gradient $\bar{\boldsymbol{g}}^{(k)} = n^{-1}\sum_{i=1}^n \boldsymbol{g}_i^{(k)}$. To quantify the discrepancy between $\pi_A^\top \mathbf{y}^{(k)}$ and $\bar{\boldsymbol{g}}^{(k)}$, we have:

$$n^{-1}\pi_A^\top \mathbf{y}^{(k)} - \bar{\boldsymbol{g}}^{(k)} = (\sum_{i=1}^n \frac{[\pi_A]_i}{[A^k]_{ii}} - 1)\bar{\boldsymbol{g}}^{(k)} + \sum_{i=1}^n \frac{[\pi_A]_i}{[A^k]_{ii}} \underbrace{(\boldsymbol{g}_i^{(k)} - \bar{\boldsymbol{g}}^{(k)})}_{\text{gradient dissimlarity}} . \tag{11}$$

  The first term is a global gradient and it is naturally bounded in the gradient descent process. However, the second term is bounded only if we have assumed that the gradient dissimilarity or data heterogeneity is bounded.

To illustrate the limitations of PULL-DIAG-GT, we analyze its convergence in the non-convex and stochastic setting. Prior to our work, its convergence has only been examined in strongly-convex and deterministic settings by Li et al. (2019); Xin et al. (2019c).

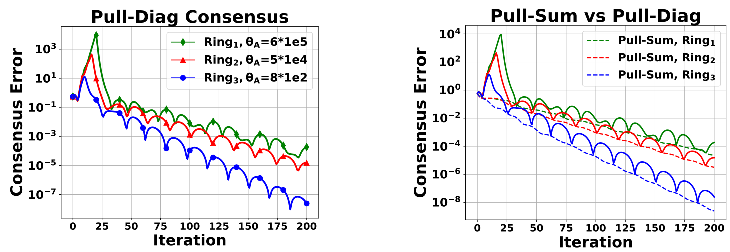


Figure 1: The left plot illustrates the PULL-DIAG consensus on three different networks with varying $\theta_A$, while other parameters remain approximately constant, with $\kappa_A \approx 2$ and $\beta_A \approx 0.989$. It is observed that PULL-DIAG exhibits a larger initial spike for networks with higher $\theta_A$. The right plot compares the consensus error of PULL-SUM (dashed lines) and PULL-DIAG (solid lines). PULL-SUM consistently outperforms PULL-DIAG across all cases. Detailed experimental setup is referred to Appendix G.

**Assumption 4** (Bounded data heterogeneity)**.** *There exists constant $b > 0$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x}) - \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\boldsymbol{x})\|^2 \le b^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^q.$$

**Theorem 2** (PULL-DIAG-GT convergence)**.** *Under assumptions 1, 2, 3 and 4, when total iteration $K > \frac{\sqrt{n\kappa_A}\theta_A^3}{1-\beta_A}$, there exists a learning rate $\alpha$ (see Theorem1 in Appendix F) such that*

$$\min_{k=0,1,\dots,K} \mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] = \mathcal{O}\left(\frac{\sigma}{\sqrt{nK}} + \frac{(n\kappa_A^2\theta_A^4\sigma^2)^{\frac{1}{3}}}{nK^{\frac{2}{3}}(1-\beta_A^2)^{\frac{4}{3}}} + \frac{\kappa_A\theta_A^3(\sigma^{\frac{2}{3}} + b^{\frac{2}{3}})}{n^{\frac{1}{3}}K(1-\beta_A)^{\frac{5}{3}}}\right).$$

*Here we omit constant coefficients including $\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2], \Delta, L$. We define $\mathbf{w}^{(k)} = \pi_A^\top \mathbf{x}^{(k)}$.*

The convergence rate of PULL-DIAG-GT is influenced by $\theta_A$ due to its reliance on the PULL-DIAG protocol. Additionally, the rate is affected by the data heterogeneity metric $b^2$; greater data heterogeneity results in slower convergence, as shown in Figure 3. However, our established lower bound in Theorem 1 is unaffected by both $\theta_A$ and $b^2$, indicating that their impacts can be eliminated.

## 5 ACHIEVING OPTIMAL COMPLEXITY WITH NEW ALGORITHM DESIGNS

The instability issues in PULL-DIAG protocol and the sensitivity to data heterogeneity in PULL-DIAG-GT hinder existing algorithms from achieving the convergence lower bound. This section develops new algorithms to address these limitations and attain the established lower bound.

### 5.1 PULL-SUM ADDRESSES THE INSTABILITY ISSUE

A key insight into how the PULL-DIAG protocol (6) corrects the bias in the vanilla gossip protocol (5) is that $\mathrm{Diag}(A^k) \to \mathrm{diag}(\pi_A)$ as $k \to \infty$. To address the instability arising from $\mathrm{Diag}(A^k)^{-1}$, we propose a novel PULL-SUM protocol:

$$[\text{PULL-SUM}]: \ \mathbf{w}^{(k)} = A^k \mathrm{diag}(\mathbb{1}_n^\top A^k)^{-1}\mathbf{x}^{(0)} \xmapsto{k\to\infty} \mathbb{1}_n \pi_A^\top \mathrm{diag}(n\pi_A)^{-1}\mathbf{x}^{(0)} = \bar{\mathbf{x}}^{(0)}, \qquad (12)$$

where we utilize the fact that $A^k \to \mathbb{1}_n \pi_A^\top$ as $k \to \infty$. Notably, the inversion of the column sum of $A$ is significantly more stable than the inversion of the diagonal entries, as illustrated below:

**Proposition 2.** *Let $D_k = \mathrm{diag}(\mathbb{1}_n^\top A^k)$, it holds that $\|D_k^{-1}\|_2 \le \kappa_A, \forall k \ge 0$.*

With this result, we achieve the convergence rate of the PUSH-SUM protocol:

**Proposition 3** (PULL-SUM convergence)**.** *For a row-stochastic and primitive matrix $A$,* PULL-SUM *converges at the following rate: $\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(0)}\|_F \le \max\{1 + n\kappa_A, \kappa_A^{1.5}\beta_A^k\}\|\mathbf{x}^{(0)}\|_F$.*

Compared to the convergence of PULL-DIAG shown in Proposition 1, PULL-SUM eliminates the influence of $\theta_A$, as illustrated in Figure 1. Because PULL-SUM is unaffected by $\theta_A$, algorithms built upon PULL-SUM have the potential to approach the lower bound established in Theorem 1.

### 5.2 PULL-SUM-GT ADDRESSES THE DATA HETEROGENEITY ISSUE

The primary reason for PULL-DIAG-GT's susceptibility to data heterogeneity lies in its weighted average $\boldsymbol{w}$, which performs global gradient descent asymptotically rather than non-asymptotically; that is, $\pi_A^\top \mathbf{y}^{(k)} \to \bar{\boldsymbol{g}}^{(k)}$ as $k \to \infty$, but $\pi_A^\top \mathbf{y}^{(k)} \neq \bar{\boldsymbol{g}}^{(k)}$ in the non-asymptotic phase, see the illustration in (9) and (10). To address this issue, we develop a new gradient tracking method based on the PULL-SUM protocol, termed PULL-SUM-GT, as follows. For $\forall k \geq 0$,

$$\tilde{D}_{k+1} = A\tilde{D}_k \quad (13a) \qquad D_{k+1} = \mathrm{diag}(\mathbb{1}_n^\top \tilde{D}_{k+1}) \quad (13b)$$

$$\mathbf{x}^{(k+1)} = A(\mathbf{x}^{(k)} - \alpha D_{k+1}^{-1}\mathbf{y}^{(k)}) \quad (13c) \qquad \mathbf{y}^{(k+1)} = B(\mathbf{y}^{(k)} + \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}) \quad (13d)$$

where $B = A\mathrm{diag}(\mathbb{1}_n^\top A)^{-1}, \mathbf{y}^{(0)} = \mathbf{g}^{(0)}$, and $\tilde{D}_0 = A^\ell$, with $\ell$ representing the number of warm-up iterations, *i.e.*, we perform $\ell$ rounds of communication before starting the optimization to obtain $\tilde{D}_0 = A^\ell$. We refer the implementation details to Appendix D. A critical strategy in PULL-SUM-GT is the construction of a column-stochastic matrix $B$ from the row-stochastic matrix $A$:

**Lemma 4.** *For any non-negative integers u, $B = \mathrm{diag}(\mathbb{1}_n^\top A^u)\, A\, \mathrm{diag}(\mathbb{1}_n^\top A^{u+1})^{-1}$ is a column-stochastic matrix, i.e., $\mathbb{1}_n^\top B = \mathbb{1}_n^\top$. Specifically, we take $u = 0$ in our* PULL-SUM-*GT.*

**Algorithm insight.** We now provide insights into why PULL-SUM-GT is robust to data heterogeneity. If we left-multiply $n^{-1}\mathbb{1}_n^\top$ on both sides of (13d), we obtain $\bar{\boldsymbol{y}}^{(k+1)} - \bar{\boldsymbol{g}}^{(k+1)} = \bar{\boldsymbol{y}}^{(k)} - \bar{\boldsymbol{g}}^{(k)} = \cdots = \bar{\boldsymbol{y}}^{(0)} - \bar{\boldsymbol{g}}^{(0)} = 0$, where we use the property $\mathbb{1}_n^\top B = \mathbb{1}_n^\top$. This result indicates that $\bar{\boldsymbol{y}}^{(k)} = \bar{\boldsymbol{g}}^{(k)}$. Next, we denote $\boldsymbol{d}_k^\top = n^{-1}\mathbb{1}_n^\top A^k$ and left-multiply $\boldsymbol{d}_{k+\ell-1}^\top$ on both sides of (13c):

$$\boldsymbol{d}_{k+\ell-1}^\top \mathbf{x}^{(k+1)} = \boldsymbol{d}_{k+\ell}^\top \mathbf{x}^{(k)} - \alpha \bar{\boldsymbol{y}}^{(k)} = \boldsymbol{d}_{k+\ell}^\top \mathbf{x}^{(k)} - \alpha \bar{\boldsymbol{g}}^{(k)}. \quad (14)$$

As $\mathbf{x}^{(k)}$ approaches consensus as $k$ increases, *i.e.*, $\mathbf{x}^{(k)} \to \mathbb{1}_n \boldsymbol{x}^{(k)}$, it holds that $\boldsymbol{d}_{k+\ell}^\top \mathbf{x}^{(k)} = \boldsymbol{x}^{(k)}$. The above recursion becomes $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \bar{\boldsymbol{g}}^{(k)}$, the centralized parallel SGD. Unlike PULL-DIAG-GT (11), recursion (14) is unaffected by gradient dissimilarity or data heterogeneity. Our next theorem establishes the convergence of PULL-SUM-GT, which achieves the first linear speedup convergence rate in ROW-ONLY decentralized learning in the stochastic and non-convex settings, without assuming data heterogeneity:

**Theorem 3** (PULL-SUM-GT convergence)**.** *Under assumptions 1, 2 and 3, when $\ell \geq \ell_0 := \lceil \frac{4+\ln(LKn^2\kappa_A^2) - \ln(1-\beta_A)}{1-\beta_A} \rceil = \tilde{\mathcal{O}}(\frac{1}{1-\beta_A})$, with proper $\alpha$ shown in Theorem 1 of Appendix D, we have the following convergence result: $\forall K \geq 0$,*

$$\frac{1}{K}\sum_{k=0}^{K} \mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] = \mathcal{O}\left( \frac{\sqrt{L\Delta}\sigma}{\sqrt{nK}} + \left(\frac{nL\Delta q_A q_B \sigma}{K}\right)^{2/3} + \frac{L\Delta n^{3/2}\kappa_A q_A q_B}{K} + \frac{n^3 q_A^2 q_B^2 \sigma^2}{K^2} \right)$$

*where $\kappa_B := \frac{\max(\pi_B)}{\min(\pi_B)}$, $\beta_B := \|B - B_\infty\|_{\pi_B}$, $q_A := \frac{1+\ln(\kappa_A)}{1-\beta_A}, q_B := \frac{1+\ln(\kappa_B)}{1-\beta_B}$, $\boldsymbol{w}^{(k)} = \boldsymbol{d}_k^\top \mathbf{x}^{(k)}$, $\boldsymbol{d}_k^\top = n^{-1}\mathbb{1}_n^\top A^{k+\ell+2}$. Absolute constants and $\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2]$ are omitted.*

**Remark 1** Although it is generally challenging to establish an exact relationship between $\kappa_B$ and $\kappa_A$, as well as $\beta_B$ and $\beta_A$, in practice we often observe that the quantities $\beta_B$ and $\kappa_B$ closely resemble $\beta_A$ and $\kappa_A$, respectively. We will further eliminate the impact of $B$ through multi-gossip strategy in next subsection.

**Remark 2** The convergence rate of PULL-SUM-GT effectively addresses the instability and heterogeneity issues encountered by existing algorithms; it is not influenced by the inversion of the small value $\theta_A$ or by data heterogeneity $b^2$. Numerical experiments confirm that PULL-SUM-GT exhibits greater stability than PULL-DIAG-GT in the presence of data heterogeneity, see Figure 3. This gives PULL-SUM-GT based algorithms a chance to achieve the lower bound in Theorem 1.

### 5.3 ACHIEVING OPTIMAL CONVERGENCE RATE

Building on the PULL-SUM protocol and PULL-SUM-GT, we now develop an algorithm to achieve the convergence lower bound established in Theorem 1. Inspired by the optimal algorithm development for doubly-stochastic mixing matrices demonstrated in Lu & De Sa (2021); Yuan et al. (2022), we introduce two additional components to PULL-SUM-GT: gradient accumulation and multi-gossip (MG) communication, resulting in the algorithm named as MG-PULL-SUM-GT. For $t \in [T]$:

$$\tilde{D}_{t+1} = A^M \tilde{D}_t, D_{t+1} = \mathrm{diag}(\mathbb{1}_n^\top \tilde{D}_{t+1}) \quad (15a) \qquad \mathbf{x}^{(t+1)} = A^M(\mathbf{x}^{(t)} - \alpha D_{t+1}^{-1}\mathbf{y}^{(t)}) \quad (15b)$$

$$\mathbf{g}^{(t+1)} = M^{-1}\textstyle\sum_{r=1}^{M}\nabla F(\mathbf{x}^{(t+1)};\boldsymbol{\xi}^{(t+1,r)}) \quad (15c) \qquad \mathbf{y}^{(t+1)} = B(\mathbf{y}^{(t)} + \mathbf{g}^{(t+1)} - \mathbf{g}^{(t)}) \quad (15d)$$

Here, $B = A^M \mathrm{diag}(\mathbb{1}_n^\top A^M)^{-1}$, $\tilde{D}_0 = A^\ell$, and $\mathbf{y}^{(0)} = \mathbf{g}^{(0)} = \frac{1}{M}\sum_{r=1}^{M}\nabla F(\mathbf{x}^{(0)};\boldsymbol{\xi}^{(0,r)})$. The detail of implementation can be found in Appendix E. In contrast to the vanilla PULL-SUM-GT algorithm (13), which performs one gossip communication and one gradient computation per iteration, MG-PULL-SUM-GT conducts $M$ gossip communications and $M$ gradient computations per iteration. To maintain the same communication and computation budgets, we run the vanilla PULL-SUM-GT for $K$ iterations while executing MG-PULL-SUM-GT for $T = K/M$ iterations. The following theorem shows that MG-PULL-SUM-GT achieves optimal convergence rate.

**Theorem 4** (Convergence of MG-PULL-SUM-GT)**.** *Under Assumptions 1, 2 and 3, When $\ell \geq \ell_0$, $M = \lceil \frac{1+\ln(n^2\kappa_A^2)+2|\ln(\sigma)|+|\ln(L)|+|\ln(\Delta)|}{1-\beta_A} \rceil = \tilde{\mathcal{O}}(\frac{1}{1-\beta_A})$, MG-PULL-SUM-GT converges as:*

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] = \tilde{\mathcal{O}}\left(\frac{\sqrt{L\Delta}\sigma}{\sqrt{nK}} + \frac{(1+\ln(\kappa_A))L\Delta}{(1-\beta_A)K}\right) \tag{16}$$

*where $K = MT$ is the total rounds of communication, $M$ is the multi-gossip number, $\tilde{\mathcal{O}}(\cdot)$ absorbs logarithmic factors and absolute constants. $\boldsymbol{w}^{(k)} = \boldsymbol{d}_{kM+\ell+1}^T\mathbf{x}^{(k)}$, $\boldsymbol{d}_t = n^{-1}\mathbb{1}_n^\top A^t$.*

The detailed proof of Theorem 4 is referred to Appendix E. Remarkably, the rate (16) aligns with the lower bound (4) up to logarithmic factors that are independent of $\kappa_A$ and $\beta_A$. This demonstrates the near-optimality of MG-PULL-SUM-GT and the tightness of the lower bound (4).

## 6 EXPERIMENTS

In this section, we numerically compare PULL-SUM-GT with the state-of-the-art ROW-ONLY gradient tracking algorithms, including PULL-DIAG-GT (Xin et al., 2019d; Li et al., 2019), FRSD (Ghaderyan et al., 2021) and FROZEN (Xin et al., 2019a). We mainly exhibit that i) PULL-SUM-GT is able to overcome the influence of $\theta_A$, ii) PULL-SUM-GT is able to overcome the influence of data heterogeneity, iii) Multi-Gossip strategy can bring significant improvement.

**Network Design.** We construct four topologies with 20 nodes, labeled $\mathrm{Ring}_i$ for $i = 1, 2, 3, 4$, as they represent directed ring graphs with $i$ additional connections. The mixing matrices are defined by $a_{ij} = 1/(1 + d_i^{\mathrm{in}})$ if $(i, j) \in \mathcal{E}$ or $j = i$, and $a_{ij} = 0$ otherwise. These topologies exhibit significant differences in $\theta_A$ but have similar $\beta_A$ and $\kappa_A$. Further details are in Appendix G.

**Synthetic Dataset: Influence of $\theta_A$.** We solve a decentralized logistic regression problem with non-convex regularization using synthetic data (Xin et al., 2021; Alghunaim & Yuan, 2022). Consider

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \rho r(x) \quad \text{where} \quad f_i(x) = \frac{1}{M}\sum_{l=1}^{M}\ln(1+\exp(-y_{i,l}h_{i,l}^\top x).$$

The function $r(x) = \sum_{j=1}^{d}[x]_j^2/(1+[x]_j^2)$ is a non-convex regularizer, and $\rho > 0$ is the regularization coefficient. The training dataset at node $i$, $\{h_{i,l}, y_{i,l}\}_{l=1}^{M}$, consists of feature vectors $h_{i,l} \in \mathbb{R}^d$ and corresponding labels $y_{i,l} \in \{+1, -1\}$. Detailed hyper-parameter settings are provided in Appendix G. We compare PULL-SUM-GT with other PULL-DIAG-GT methods on topologies $\mathrm{Ring}_{1,2,3,4}$. PULL-SUM-GT remains unaffected by $\theta_A$, while PULL-DIAG-based methods are significantly influenced. As shown in Figure 2, when $\theta_A$ is small ($\mathrm{Ring}_{3,4}$), the performance of all algorithms is similar. However, for larger $\theta_A$ ($\mathrm{Ring}_{1,2}$), PULL-SUM-GT remains robust, whereas PULL-DIAG-based methods deteriorate significantly.

**Real-World Dataset: Influence of heterogeneity.** We train a four-layer fully connected neural network in a decentralized setting to solve the handwritten digit classification task on the MNIST dataset (Deng, 2012). Two scenarios are evaluated: (i) Uniformly distributed data, where each node

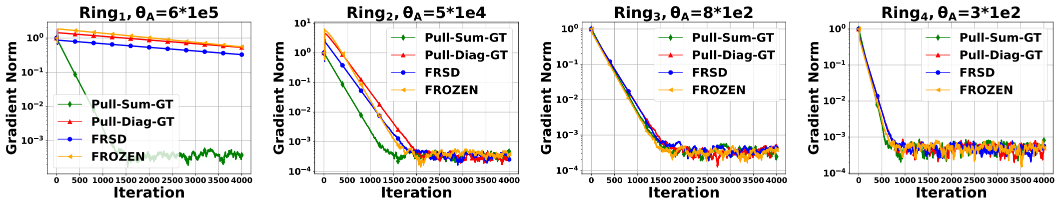


Figure 2: Comparison on synthetic dataset.

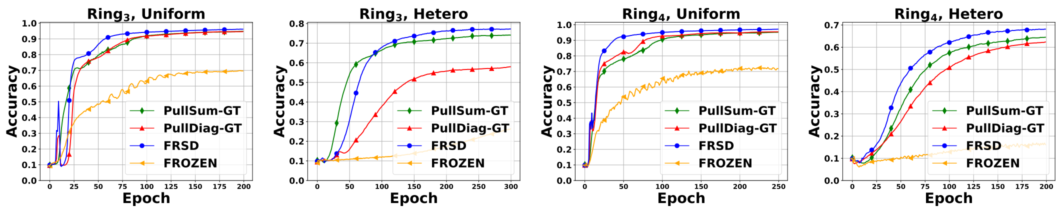


Figure 3: Comparison on MNIST dataset. "Uniform" denotes evenly distributed data, "Hetero" denotes heterogeneous data.

holds a shuffled partition of the dataset, and (ii) Heterogeneous data, where each node contains images from only one digit class. As shown in Figure 3, PULL-DIAG-GT performs comparably to PULL-SUM-GT in the uniform setting. However, PULL-SUM-GT is more robust to data heterogeneity, outperforming PULL-DIAG-GT and FROZEN. Notably, FRSD achieves the best performance, as expected, due to its integration with momentum.

**Benefit of Multiple Gossip.** In the third set of experiments, we illustrate the performance of MG-PULL-SUM-GT on two tasks described in the above two paragraphs. As demonstrated in Figure 4, multiple rounds of gossip help mitigate the impact of the network and allow for a larger learning rate, leading to faster convergence. Note that all curves are compared fairly, with each iteration involving a single gradient computation and one communication round.

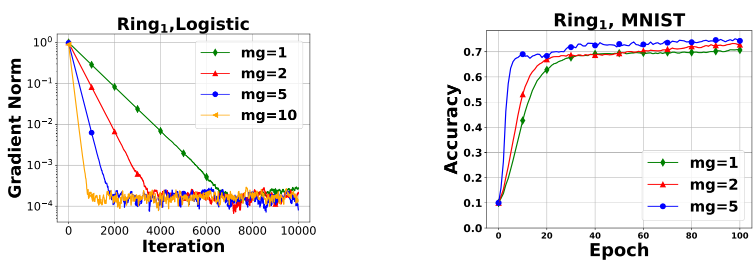


Figure 4: Performance of MG-PULL-SUM-GT on synthetic dataset (the left plot) and highly heterogeneous MNIST dataset (the right plot).

## 7 CONCLUSION AND LIMITATIONS

This paper establishes a tight lower bound and identifies optimal algorithms for decentralized optimization using row-stochastic mixing matrices. Our analysis shows that existing PULL-DIAG-based methods are sensitive to algorithmic instability and data heterogeneity, preventing them from reaching the lower bound. We propose the PULL-SUM protocol and (MG-)PULL-SUM-GT, which mitigate dependence on network properties and heterogeneity, achieving the lower bound up to logarithmic factors. Experimental results support our findings. A limitation of this work is that, while the method performs well empirically without a warm-up, the convergence guarantees for PULL-SUM-GT are currently provided only with a warm-up stage. Addressing unconditional convergence guarantees will be left for future work.

## REFERENCES

Sulaiman A Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 70:3264–3279, 2022.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2019. URL `https://api.semanticscholar.org/CorpusID:208637439`.

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pp. 344–353. PMLR, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1-2):315–355, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

P. Di Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Diyako Ghaderyan, Necdet Serhat Aybat, A Pedro Aguiar, and Fernando Lobo Pereira. A fast row-stochastic decentralized optimization method over directed graphs. *arXiv preprint arXiv:2112.13257*, 2021.

Xinmeng Huang, Yiming Chen, Wotao Yin, and Kun Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. *Advances in Neural Information Processing Systems*, 35:18955–18969, 2022.

David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 482–491. IEEE, 2003.

Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Huaqing Li, Jinmeng Wang, and Zheng Wang. Row-stochastic matrices based distributed optimization algorithm with uncoordinated step-sizes. In *2019 6th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)*, pp. 124–131. IEEE, 2019.

Liyuan Liang, Xinmeng Huang, Ran Xin, and Kun Yuan. Towards better understanding the influence of directed networks on decentralized stochastic optimization. *arXiv preprint arXiv:2312.04928*, 2023.

Qingguo Lü, Xiaofeng Liao, Huaqing Li, and Tingwen Huang. A nesterov-like gradient tracking algorithm for distributed optimization over directed networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(10):6258–6270, 2020.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pp. 7111–7123. PMLR, 2021.

Van Sy Mai and Eyad H Abed. Distributed optimization over weighted directed graphs using row stochastic matrix. In *2016 American Control Conference (ACC)*, pp. 7165–7170. IEEE, 2016.

A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017. doi: 10.1137/16M1084316.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015. doi: 10.1109/TAC.2014.2364096.

Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907. doi: 10.1007/BF01449896. URL https://doi.org/10.1007/BF01449896.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Muhammad I Qureshi, Ran Xin, Soummya Kar, and Usman A Khan. S-addopt: Decentralized stochastic first-order optimization over directed graphs. *IEEE Control Systems Letters*, 5(3):953–958, 2020.

Ali H Sayed. Adaptive networks. *Proceedings of the IEEE*, 102(4):460–497, 2014.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pp. 3027–3036. PMLR, 2017.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2740–2749, 2018.

Haoran Sun and Mingyi Hong. Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms. *IEEE Transactions on Signal processing*, 67(22):5912–5928, 2019.

Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Push-sum distributed dual averaging for convex optimization. In *2012 ieee 51st ieee conference on decision and control (cdc)*, pp. 5453–5458. IEEE, 2012.

Chenguang Xi and Usman A Khan. Dextra: A fast algorithm for optimization over directed graphs. *IEEE Transactions on Automatic Control*, 62(10):4980–4993, 2017.

Chenguang Xi, Ran Xin, and Usman A Khan. Add-opt: Accelerated distributed directed optimization. *IEEE Transactions on Automatic Control*, 63(5):1329–1339, 2017.

Chenguang Xi, Van Sy Mai, Ran Xin, Eyad H Abed, and Usman A Khan. Linear convergence in optimization over directed graphs with row-stochastic matrices. *IEEE Transactions on Automatic Control*, 63(10):3558–3565, 2018.

Ran Xin, Dušan Jakovetić, and Usman A Khan. Distributed nesterov gradient methods over arbitrary graphs. *IEEE Signal Processing Letters*, 26(8):1247–1251, 2019a.

Ran Xin, Anit Kumar Sahu, Usman A Khan, and Soummya Kar. Distributed stochastic optimization with gradient tracking over strongly-connected networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 8353–8358. IEEE, 2019b.

Ran Xin, Chenguang xi, and Usman Khan. Frost—fast row-stochastic optimization with uncoordinated step-sizes. *EURASIP Journal on Advances in Signal Processing*, 2019, 01 2019c. doi: 10.1186/s13634-018-0596-y.

Ran Xin, Chenguang Xi, and Usman A Khan. Frost—fast row-stochastic optimization with uncoordinated step-sizes. *EURASIP Journal on Advances in Signal Processing*, 2019(1):1–14, 2019d.

Ran Xin, Usman Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 01 2021. doi: 10.1109/TSP.2021.3062553.

Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060, Osaka, Japan, 2015.

Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Decentlam: Decentralized momentum sgd for large-batch deep training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3029–3039, 2021.

Kun Yuan, Xinmeng Huang, Yiming Chen, Xiaohan Zhang, Yingya Zhang, and Pan Pan. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. *Advances in Neural Information Processing Systems*, 35:36382–36395, 2022.

Jinshan Zeng and Wotao Yin. Extrapush for convex smooth decentralized optimization over directed networks. *Journal of Computational Mathematics*, pp. 383–396, 2017.

# A Lower Bound

## A.1 A Matrix Example

**Proposition 5.** *For any $n \geq 2$, there exists a row-stochastic, primitive matrix $A \in \mathbb{R}^{n\times n}$ satisfying $\beta_A = \frac{\sqrt{2}}{2}$ but $\kappa_A = 2^{n-1}$.*

*Proof.* Proposition 2.5 of Liang et al. (2023) tells us that For any $n \geq 2$, there exists a column-stochastic, primitive matrix $W \in \mathbb{R}^{n\times n}$ satisfying $\beta_W = \frac{\sqrt{2}}{2}$ but $\kappa_W = 2^{n-1}$. By taking $A = B^\top$, their Perron vectors are the same, i.e., $\pi_A = \pi_W$. Therefore, $\kappa_A = \kappa_W$. By definition of $\pi$-norm, we know that $\beta_A = \|A - A_\infty\|_{\pi_A} = \|\Pi_A^{1/2}(A - A_\infty)\Pi_A^{-1/2}\|_2 = \|(\Pi_W^{-1/2}(W - W_\infty)\Pi_W^{1/2})^\top\|_2 = \|W\|_{\pi_W} = \beta_W$. $\square$

## A.2 Proof of Theorem 1

The core idea of the proof is derived from Liang et al. (2023). The first complexity term, $\Omega(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}})$, is standard, and its proof can be found in works such as Lu & De Sa (2021) and Yuan et al. (2022). Therefore, we concentrate on proving the second term, $\Omega((1 + \ln(\kappa_A))L\Delta/K)$.

To proceed, let $[x]_j$ represent the $j$-th coordinate of a vector $x \in \mathbb{R}^d$ for $1 \leq j \leq d$, and define:

$$\mathrm{prog}(x) := \begin{cases} 0 & \text{if } x = 0; \\ \max_{1\leq j\leq d}\{j : [x]_j \neq 0\} & \text{otherwise.} \end{cases}$$

We also introduce several important lemmas, which have been established in previous research.

**Lemma 6** (Lemma 2 of Arjevani et al. (2019))**.** *Consider the function*

$$h(x) := -\psi(1)\phi([x]_1) + \sum_{j=1}^{d-1}\Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big)$$

*where for any $z \in \mathbb{R}$,*

$$\psi(z) = \begin{cases} 0 & z \leq 1/2; \\ \exp\left(1 - \frac{1}{(2z-1)^2}\right) & z > 1/2, \end{cases} \qquad \text{and} \qquad \phi(z) = \sqrt{e}\int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\mathrm{d}t.$$

*The function $h(x)$ has the following properties:*

1. *$h$ is zero-chain, i.e.,* $\mathrm{prog}(\nabla h(x)) \leq \mathrm{prog}(x) + 1$ *for all $x \in \mathbb{R}^d$.*
2. *$h(x) - \inf_x h(x) \leq \Delta_0 d$, for all $x \in \mathbb{R}^d$ with $\Delta_0 = 12$.*
3. *$h$ is $L_0$-smooth with $L_0 = 152$.*
4. *$\|\nabla h(x)\|_\infty \leq G_\infty$, for all $x \in \mathbb{R}^d$ with $G_\infty = 23$.*
5. *$\|\nabla h(x)\|_\infty \geq 1$ for any $x \in \mathbb{R}^d$ with $[x]_d = 0$.*

**Lemma 7** (Lemma 4 of Huang et al. (2022))**.** *Letting functions*

$$h_1(x) := -2\psi(1)\phi([x]_1) + 2\sum_{j \text{ even},\, 0<j<d}\Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big)$$

*and*

$$h_2(x) := 2\sum_{j \text{ odd},\, 0<j<d}\Big(\psi(-[x]_j)\phi(-[x]_{j+1}) - \psi([x]_j)\phi([x]_{j+1})\Big),$$

*then $h_1$ and $h_2$ satisfy the following properties:*

1. *$\frac{1}{2}(h_1 + h_2) = h$, where $h$ is defined in Lemma 6.*
2. *$h_1$ and $h_2$ are zero-chain,* i.e., $\mathrm{prog}(\nabla h_i(x)) \leq \mathrm{prog}(x) + 1$ *for all $x \in \mathbb{R}^d$ and $i = 1, 2$. Furthermore, if* $\mathrm{prog}(x)$ *is odd, then* $\mathrm{prog}(\nabla h_1(x)) \leq \mathrm{prog}(x)$*; if* $\mathrm{prog}(x)$ *is even, then* $\mathrm{prog}(\nabla h_2(x)) \leq \mathrm{prog}(x)$.
3. *$h_1$ and $h_2$ are also $L_0$-smooth with $L_0 = 152$.*

We are now ready to prove our lower bound. This proceeds in three steps. Without loss of generality, we assume $n$ can be divided by 3.

(Step 1.) We let $f_i = L\lambda^2 h_1(x/\lambda)/L_0, \forall\, i \in E_1 \triangleq \{j : 1 \le j \le n/3\}$ and $f_i = L\lambda^2 h_2(x/\lambda)/L_0$, $\forall\, i \in E_2 \triangleq \{j : 2n/3 \le j \le n\}$, where $h_1$ and $h_2$ are defined in Lemma 7, and $\lambda > 0$ will be specified later. By the definitions of $h_1$ and $h_2$, we have that $f_i$, $\forall\, 1 \le i \le n$, is zero-chain and $f(x) = n^{-1}\sum_{i=1}^n f_i(x) = 2L\lambda^2 h(x/\lambda)/3L_0$. Since $h_1$ and $h_2$ are also $L_0$-smooth, $\{f_i\}_{i=1}^n$ are $L$-smooth. Furthermore, since

$$f(0) - \inf_x f(x) = \tfrac{2L\lambda^2}{3L_0}(h(0) - \inf_x h(x)) \le \tfrac{L\lambda^2\Delta_0 d}{L_0},$$

to ensure $\{f_i\}_{i=1}^n$ satisfy L-smooth Assumption, it suffices to let

$$\tfrac{L\lambda^2\Delta_0 d}{L_0} \le \Delta, \quad \textit{i.e.}, \quad \lambda \le \sqrt{\tfrac{L_0\Delta}{L\Delta_0 d}}. \tag{17}$$

With the functions defined above, we have $f(x) = n^{-1}\sum_{i=1}^n f_i(x) = L\lambda^2 l(x/\lambda)/(3L_0)$ and $\mathrm{prog}(\nabla f_i(x)) = \mathrm{prog}(x)+1$ if $\mathrm{prog}(x)$ is even and $i \in E_1$ or $\mathrm{prog}(x)$ is odd and $i \in E_2$, otherwise $\mathrm{prog}(\nabla f_i(x)) \le \mathrm{prog}(x)$. Therefore, to make progress (*i.e.*, to increase $\mathrm{prog}(x)$), for any gossip algorithm $\mathbb{A} \in \mathcal{A}_W$, one must take the gossip communication protocol to transmit information between $E_1$ and $E_2$ alternatively.

(Step 2.) We consider the noiseless gradient oracles and the constructed mixing matrix $W$ in Subsection 5 with $\epsilon = 2\beta_A^2 - 1$ so that $\frac{1+\ln(\kappa_A)}{1-\beta_A} = O(n)$. Note the directed distance from $E_1$ to $E_2$ is $n/3$. Consequently, starting from $x^{(0)} = 0$, it takes of at least $n/3$ communications for any possible algorithm $\mathbb{A} \in \mathcal{A}_A$ to increase $\mathrm{prog}(\hat{x})$ by 1 if it is odd. Therefore, we have $\lceil \mathrm{prog}(\hat{x}^{(k)})/2 \rceil \le \left\lfloor \frac{k}{2n/3} \right\rfloor, \forall\, k \ge 0$. This further implies

$$\mathrm{prog}(\hat{x}^{(k)}) \le 2\left\lfloor \tfrac{k}{2n/3} \right\rfloor + 1 \le 3k/n + 1, \quad \forall\, k \ge 0. \tag{18}$$

(Step 3.) We finally show the error $\mathbb{E}[\|\nabla f(x)\|^2]$ is lower bounded by $\Omega\left(\frac{(1+\ln(\kappa_A))L\Delta}{(1-\beta_A)K}\right)$, with any algorithm $\mathbb{A} \in \mathcal{A}_W$ with $K$ communication rounds. For any $K \ge n$, we set $d = 2\left\lfloor \frac{K}{2n/3} \right\rfloor + 2 \le 3K/n + 2 \le 5K/n$ and $\lambda = \left(\frac{nL_0\Delta}{5L\Delta_0 K}\right)^{1/2}$. Then (17) naturally holds. Since $\mathrm{prog}(\hat{x}^{(K)}) < d$ by (18), using the last point of Lemma 6 and the value of $\lambda$, we obtain

$$\mathbb{E}[\|\nabla f(\hat{x})\|^2] \ge \min_{[\hat{x}]_d = 0} \|\nabla f(\hat{x})\|^2 \ge \tfrac{L^2\lambda^2}{9L_0^2} = \Omega\left(\tfrac{nL\Delta}{K}\right).$$

By finally using $n = \Omega((1+\ln(\kappa_A))/(1-\beta_A))$, we complete the proof.

# B PULL-DIAG CONVERGENCE

We need the following assumption.

**Assumption 5** (Upper bound for inverse diagonal entries.)**.** *We assume that the diagonal entries of $A^k$ is lower bounded, in other words,* $\theta_A = \sup_k \|\mathrm{Diag}(A^k)^{-1}\|_2$.

**Lemma 8.** *Under assumption 5,* PULL-DIAG *converges by:*

$$\|n^{-1}\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(0)}\|_F \le \min\{1+\theta_A, 2\theta_A\kappa_A^{1.5}\beta_A^k\}\|\mathbf{x}^{(0)}\|_F \tag{19}$$

*Proof.* On the one hand,

$$\begin{aligned}\|A^k\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top\|_2 &\le \|A^k\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top\|_F \\ &\le n\max_{i,j}\{|[A^k\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top]_{ij}|\} \\ &\le n + n\theta_A\end{aligned}$$

On the other hand,

$$
\begin{aligned}
&\|A^k\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top\|_2 \\
&\le \|(A^k - A_\infty)\mathrm{Diag}(A^k)^{-1}\|_2 + \|A_\infty\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top\|_2 \\
&\le \|A^k - A_\infty\|_2\|\mathrm{Diag}(A^k)^{-1}\|_2 + \|\mathbb{1}_n(\pi_A^\top - \mathrm{diag}(A^k)^\top)\|_2\|\mathrm{Diag}(A^k)^{-1}\|_2 \\
&\le \theta_A\kappa_A^{1.5}\beta_A^k + \theta_A n^{0.5}\|\pi_A - \mathrm{diag}(A^k)\| \le \theta_A\kappa_A^{1.5}\beta_A^k(1+n)
\end{aligned}
$$

The last inequality comes from $\|\pi_A - \mathrm{diag}(A^k)\| \le \|A^k - A_\infty\|_F \le n^{0.5}\kappa_A^{1.5}\beta_A^k$. Therefore, we have

$$
\begin{aligned}
\|n^{-1}\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(0)}\|_F &= n^{-1}\|(A^k\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top)\mathbf{x}^{(0)}\|_F \\
&\le n^{-1}\|A^k\mathrm{Diag}(A^k)^{-1} - \mathbb{1}_n\mathbb{1}_n^\top\|_2\|\mathbf{x}^{(0)}\|_F \\
&\le \min\{1+\theta_A, 2\theta_A\kappa_A^{1.5}\beta_A^k\}\|\mathbf{x}^{(0)}\|_F,
\end{aligned}
$$

which finishes our proof. □

# C PULL-SUM CONVERGENCE

**Lemma 9.** PULL-SUM *converges by:*

$$
\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(0)}\|_F \le \min\{1+n\kappa_A, \kappa_A^{1.5}\beta_A^k\}\|\mathbf{x}^{(0)}\|_F \tag{20}
$$

*Proof.* We define $D_k = \mathrm{diag}(\mathbb{1}_n^\top A^k)$. Note that $A^\top$ is a column-stochastic matrix, we can apply the third statement of Lemma 2.4 of Liang et al. (2023) and obtain that $\|D_k^{-1}\|_2 \le \kappa_A$. On the one hand,

$$
\|A^kD_k^{-1} - n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\|_2 \le n\max_{i,j}\{|[A^kD_k^{-1} - n^{-1}\mathbb{1}_n\mathbb{1}_n^\top]_{ij}|\} \le 1+n\kappa_A.
$$

On the other hand, we have

$$
\|A^kD_k^{-1} - \mathbb{1}_n\mathbb{1}_n^\top\|_2 \le \|(I-R)(A^k - A_\infty)\|_2\|D_k^{-1}\|_2 \le \kappa_A^{1.5}\|A - A_\infty\|_{\pi_A}^k = \kappa_A^{1.5}\beta_A^k.
$$

Therefore, we have

$$
\begin{aligned}
\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(0)}\|_F = \|(A^kD_k^{-1} - n^{-1}\mathbb{1}_n\mathbb{1}_n^\top)\mathbf{x}^{(0)}\|_F &\le \|A^kD_k^{-1} - n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\|_2\|\mathbf{x}^{(0)}\|_F \\
&\le \min\{1+\kappa_A, \kappa_A^{1.5}\beta_A^k\}\|\mathbf{x}^{(0)}\|_F,
\end{aligned}
$$

which finishes our proof. □

# D PULL-SUM-GT CONVERGENCE

## D.1 NOTATIONS

We denote $\mathbb{1}_n$ as an $n$-dimensional all-ones vector. We define $I_n \in \mathbb{R}^{n\times n}$ as the identity matrix. Throughout the paper, $A$ is always a row-stochastic matrix, *i.e.*, $A\mathbb{1}_n = \mathbb{1}_n$ and $B$ is always a column-stochastic matrix, *i.e.*, $\mathbb{1}_n^\top B = \mathbb{1}_n^\top$. We denote $[n]$ as the index set $\{1, 2, \ldots, n\}$. We denote $\mathrm{Diag}(A)$ as the the diagonal matrix generated from the diagonal entries of $A$. We denote $\mathrm{diag}(v)$ as the diagonal matrix whose diagonal entries comes from vector $v$. We denote $\pi_A$ as the left Perron vector of $A$ and $\pi_B$ as the right Perron vector of $B$. We denote $\Pi_A = \mathrm{diag}(\pi_A)$, $\Pi_B = \mathrm{diag}(\pi_B)$ We define the $\pi_A$-vector norm $\|v\|_{\pi_A} = \|\Pi_A^{1/2}v\|$ and the induced $\pi_A$-matrix norm as $\|W\|_{\pi_A} = \|\Pi_A^{1/2}W\Pi_A^{-1/2}\|_2$. We define the $\pi_B$-vector norm $\|v\|_{\pi_B} = \|\Pi_B^{-1/2}v\|$ and the induced $\pi_B$-matrix norm as $\|W\|_{\pi_A} = \|\Pi_B^{-1/2}W\Pi_B^{1/2}\|_2$. We define $A_\infty = \mathbb{1}_n\pi_A^\top$ and $B_\infty = \pi_B\mathbb{1}_n^\top$. We define $\beta_A = \|A - A_\infty\|_{\pi_A}$, $\beta_B = \|B - B_\infty\|_{\beta_B}$, $\kappa_A = \max(\pi_A)/\min(\pi_B)$, $\kappa_B = \max(\pi_B)/\min(\pi_B)$, $q_A = \max_{k\ge 0} p_A$. Throughout the paper, we let $\boldsymbol{x}_i^{(k)} \in \mathbb{R}^d$ denote the local model copy at node $i$ at iteration $k$. Furthermore, we define the matrices

$$\mathbf{x}^{(k)} := [(\boldsymbol{x}_1^{(k)})^\top; (\boldsymbol{x}_2^{(k)})^\top; \cdots; (\boldsymbol{x}_n^{(k)})^\top] \in \mathbb{R}^{n\times d},$$
$$\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) := [\nabla F_1(\boldsymbol{x}_1^{(k)}; \xi_1^{(k)})^\top; \cdots; \nabla F_n(\boldsymbol{x}_n^{(k)}; \xi_n^{(k)})^\top] \in \mathbb{R}^{n\times d},$$
$$\nabla f(\mathbf{x}^{(k)}) := [\nabla f_1(\boldsymbol{x}_1^{(k)})^\top; \nabla f_2(\boldsymbol{x}_2^{(k)})^\top; \cdots; \nabla f_n(\boldsymbol{x}_n^{(k)})^\top] \in \mathbb{R}^{n\times d},$$

by stacking all local variables. The upright bold symbols (e.g. $\mathbf{x}, \mathbf{w}, \mathbf{g} \in \mathbb{R}^{n\times d}$) always denote stacked network-level quantities. Throughout the proof, we define $\overline{\nabla f(\mathbf{x}^{(k)})} = n^{-1}\mathbb{1}_n^\top \nabla f(\mathbf{x}^{(k)})$, $\Delta_x^{(k)} := (I - RA^{k+l})\mathbf{x}^{(k)}$, $\Delta_g^{(k)} = \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}$. We define $K$ as the total rounds of communication, $T$ as the iteration number of model parameter $\mathbf{x}$ and $M$ as the multi-gossip number, i.e., $K = MT$. We define $\ell_0 = \lceil \frac{4+\ln(L)+\ln(T)+2\ln(\kappa_A)+2\ln(n)-2\ln(1-\beta_A)}{1-\beta_A} \rceil$.

## D.2 Useful Inequalities

**Lemma 10.** *1.* $\|RA^k\|_2 \le 1$ *and* $\|A - I\|_2 \le \sqrt{n}$.

*2.* $\underline{\pi}_A\|v\|^2 \le \|v\|_{\pi_A}^2 \le \overline{\pi}_A\|v\|^2$, $\overline{\pi}_B^{-1}\|v\|^2 \le \|v\|_{\pi_B}^2 \le \underline{\pi_B}^{-1}\|v\|^2$.

*3.* $\|D_k^{-1}\|_2 \le \kappa_A, \forall k \ge 0$.

*4. When* $\ell \ge 1 + \lceil \frac{\ln(LK)+2\ln(\kappa_A)+2\ln(n)-\ln(1-\beta_A)}{1-\beta_A} \rceil$, $\|\mathbb{1}_n^\top A^{k+\ell} - n\pi_A^\top\| \le \frac{1}{24nLs_A\kappa_A K}$, $\forall k \ge 0$.

*Proof.* First, $\forall v \in \mathbb{R}^n$, we have $\|RA^k v\| = \frac{1}{\sqrt{n}}\langle d_k, v\rangle \le \frac{1}{\sqrt{n}}\|d_k\|\|v\|$. Therefore, $\|RA^k\|_2 \le \frac{\|d_k\|}{\sqrt{n}} \le 1$. $|[(A-I)v]_i| = |\sum_{j=1}^n a_{ij}v_j - v_i| \le \max_i |v_i|$, which means $\|(A-I)v\|_F \le \sqrt{n}\max_i |v_i| \le \sqrt{n}\|v\|$ and $\|A - I\|_2 \le \sqrt{n}$. The second lemma can be verified straight forward. The third inequality comes from our Appendix C. The fourth inequality can derived by

$$\begin{aligned}
&\|\mathbb{1}_n^\top A^{k+\ell} - n\pi_A^\top\| = \|\mathbb{1}_n^\top (A^{k+\ell} - A_\infty)\| \le \sqrt{n\kappa_A}\|A^{k+\ell} - A_\infty\|_{\pi_A} \\
&\le \sqrt{n\kappa_A}\beta_A^\ell \le \exp(-4 - ln(K) - 2\ln(\kappa_A) - 1 - \ln(n) + \ln(1-\beta_A)) \cdot \sqrt{n\kappa_A} \\
&= e^{-4}\frac{1-\beta_A}{n^{1.5}\kappa_A^{1.5}K} \le \frac{1}{24ns_A\kappa_A K}.
\end{aligned}$$

□

**Lemma 11** (Rolling Sum Lemma 1)**.** *We have the following rolling sum lemmas.*

1. *If $l \ge 1$ and $A \in \mathbb{R}^{n\times n}$ is a primitive and row-stochastic matrix, the following estimation holds for $\forall T \ge 0$.*

$$\sum_{k=0}^T \|\sum_{i=0}^k (A^{k+l-i} - A_\infty)\Delta^{(i)}\|_F^2 \le s_A^2 \sum_{i=0}^T \|\Delta^{(i)}\|_F^2, \tag{21}$$

*where $\Delta^{(i)} \in \mathbb{R}^{n\times d}$ are arbitrary matrices, and $s_A$ is defined by:*

$$s_A := \max_{k\ge 1}\|A^k - A_\infty\|_2 \cdot \frac{1 + \frac{1}{2}\ln(\kappa(\pi_A))}{1-\beta_A}. \tag{22}$$

2. *If $l \ge 1$ and $B \in \mathbb{R}^{n\times n}$ is a primitive and column-stochastic matrix, the following estimation holds for $\forall T \ge 0$.*

$$\sum_{k=0}^T \|\sum_{i=0}^k (B^{k+1-i} - B_\infty)\Delta^{(i)}\|_F^2 \le s_B^2 \sum_{i=0}^T \|\Delta^{(i)}\|_F^2, \tag{23}$$

*where $s_B$ is defined by:*

$$s_B := \max_{k\ge 1}\|B^k - B_\infty\|_2 \cdot \frac{1 + \frac{1}{2}\ln(\kappa(\pi_B))}{1-\beta_B}. \tag{24}$$

*Proof.* First, we prove that

$$\|A^i - A_\infty\|_2 \le \sqrt{\kappa(\pi_A)}\beta_A^i, \forall i \ge 0. \tag{25}$$

Notice that $\beta_A := \|A - A_\infty\|_{\pi_A}$ and

$$\|A^i - A_\infty\|_{\pi_A} = \|(A - A_\infty)^i\|_{\pi_A} \le \|A - A_\infty\|_{\pi_A}^i = \beta_A^i,$$

we have

$$\|(A^i - A_\infty)v\| = \|\Pi_A^{-1/2}(A^i - A_\infty)v\|_{\pi_A} \le \sqrt{\underline{\pi_A}}\beta_A^i\|v\|_{\pi_A} \le \sqrt{\kappa(\pi_A)}\beta_A^i\|v\|,$$

which proves (25).

Second, we want to prove that for all $k \ge 0$, we have

$$\sum_{i=0}^{k} \|A^{k+l-i} - A_\infty\|_2 \le s_A. \tag{26}$$

Towards this end, we define $M_A := \max_{k\ge 1} \|A^k - A_\infty\|_2$. According to (25), $M_A$ is well-defined. We also define $p = \max\left\{\frac{\ln(\sqrt{\kappa(\pi_A)}) - \ln(M_A)}{-\ln(\beta_A)}, 0\right\}$, then we can verify that $\|A^i - A_\infty\|_2 \le \min\{M_A, M_A\beta_A^{i-p}\}, \forall i \ge 1$. With this inequality, we can bound $\sum_{i=0}^k \|A^{k+1-i} - A_\infty\|_2$ as follows:

$$\begin{aligned}
&\sum_{i=0}^{k} \|A^{k+1-i} - A_\infty\|_2 = \sum_{i=1}^{\min\{\lfloor p\rfloor,k\}} \|A^i - A_\infty\|_2 + \sum_{i=\min\{\lfloor p\rfloor,k\}+1}^{k+1} \|A^i - A_\infty\|_2 \\
\le& \sum_{i=0}^{\min\{\lfloor p\rfloor,k\}} M_A + \sum_{i=\min\{\lfloor p\rfloor,k\}+1}^{k+1} M_A\beta_A^{i-p} \\
\le& M_A \cdot (1 + \min\{\lfloor p\rfloor, k\}) + M_A \cdot \frac{1}{1-\beta_A}\beta_A^{\min\{\lfloor p\rfloor,k\}+1-p}.
\end{aligned} \tag{27}$$

If $p = 0$, (27) is simplified to $\sum_{i=0}^k \|A^{k+1-i} - A_\infty\|_2 \le M_A \cdot \frac{1}{1-\beta_A}$ and (26) is naturally satisfied. If $p > 0$, let $x = \min\{\lfloor p\rfloor, k\} + 1 - p \in [0, 1)$, (26) is simplified to

$$\sum_{i=0}^{k} \|A^{k-i} - A_\infty\|_2 \le M_A(x + p + \frac{\beta_A^x}{1-\beta_A}) \le M_A(p + \frac{1}{1-\beta_A}).$$

Noting that $p \le \frac{\frac{1}{2}\ln(\kappa(\pi_A))}{1-\beta_A}$, we finish the proof of (26).

Finally, to obtain (21), we use Jensen's inequality. For positive numbers $a_i, i \in [k+1]$ satisfying $\sum_{i=1}^{k+1} a_i = 1$, we have

$$\begin{aligned}
&\|\sum_{i=0}^{k}(A^{k+1-i} - A_\infty)\Delta^{(i)}\|_F^2 = \|\sum_{i=0}^{k} a_{k+1-i} \cdot a_{k+1-i}^{-1}(A^{k-i} - A_\infty)\Delta^{(i)}\|_F^2 \\
\le& \sum_{i=0}^{k} a_{k+1-i}\|a_{k+1-i}^{-1}(A^{k+1-i} - A_\infty)\Delta^{(i)}\|_F^2 \le \sum_{i=0}^{k} a_{k+1-i}^{-1}\|A^{k+1-i} - A_\infty\|_2^2\|\Delta^{(i)}\|_F^2.
\end{aligned} \tag{28}$$

By choosing $a_{k+1-i} = (\sum_{i=0}^k \|A^{k+1-i} - A_\infty\|_2)^{-1}\|A^{k+1-i} - A_\infty\|_2$ in (28), we obtain that

$$\|\sum_{i=0}^{k}(A^{k+1-i} - A_\infty)\Delta^{(i)}\|_F^2 \le \sum_{i=0}^{k} \|A^{k+1-i} - A_\infty\|_2 \cdot \sum_{i=0}^{k} \|A^{k+1-i} - A_\infty\|_2\|\Delta^{(i)}\|_F^2. \tag{29}$$

By summing up (29) from $k = 0$ to $T$, we obtain that

$$\begin{aligned}
&\sum_{k=0}^{T}\|\sum_{i=0}^{k}(A^{k+1-i} - A_\infty)\Delta^{(i)}\|_F^2 \le s_A \sum_{k=0}^{T}\sum_{i=0}^{k} \|A^{k+1-i} - A_\infty\|_2\|\Delta^{(i)}\|_F^2 \\
\le& s_A \sum_{i=0}^{T}(\sum_{k=i}^{T} \|A^{k+1-i} - A_\infty\|_2)\|\Delta^{(i)}\|_F^2 \le s_A^2 \sum_{i=0}^{T} \|\Delta^{(i)}\|_F^2,
\end{aligned}$$

**Algorithm 1:** PULL-SUM-GT

```
1  Initialize ψ_i^(0) = e_i, x_i^(0) = x^(0), y_i^(0) = g_i^(0) = ∇F_i(x_i^(0), ξ_i^(0)), ℓ ≥ 1;
2  for s = 0, 1, ..., ℓ−1, each node i in parallel do
3      ψ_i^(s+1) = Σ_{j∈N_i^in} a_ij ψ_j^(s);
4  end
5  Let d_i^(0) = ψ_i^(ℓ) and v_i = Σ_{r=1}^n ψ_ir^(1) at each node i;
6  for k = 0, 1, ..., K, each node i in parallel do
7      x_i^(k+1) = Σ_{j∈N_i^in} a_ij (x_j^(k) − α(Σ_{r=1}^n d_jr^(k))^{-1} y_j^(k));
8      Locally calculate g_i^(k+1) = ∇F_i(x_i^(k+1), ξ_i^(k+1));
9      Locally calculate z_i^(k) = v_i^{-1}(y_i^(k) + g_i^(k+1) − g_i^(k));
10     y_i^(k+1) = Σ_{j∈N_i^in} a_ij z_j^(k);
11     d_i^(k+1) = Σ_{j∈N_i^in} a_ij d_j^(k);
12 end
```

which finishes the proof of this lemma. The proof can be applied in the same way when $B$ is column-stochastic.

□

### D.3 ALGORITHM FORMULATION

We also provide a denser form of Pull-Sum GT:

$$\mathbf{x}^{(k+1)} = A(\mathbf{x}^{(k)} - \alpha D_{k+\ell+1}^{-1}\mathbf{y}^{(k)}) \tag{30a}$$
$$\mathbf{y}^{(k+1)} = B(\mathbf{y}^{(k)} + \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}) \tag{30b}$$

Hereon we define $D_k = \mathrm{diag}(\mathbb{1}_n^\top A^k)$, $\ell$ is a non-negative integer.

### D.4 BASIC TRANSFORMATIONS

1. $\Delta_x^{(k+1)} = R(A^{k+\ell} - A^{k+\ell+2})\Delta_x^{(k+1)} + (A - RA^{k+l})\Delta_x^{(k)} - \alpha(A - RA^{k+l})D_{k+l}^{-1}y^{(k)}$.

2. $\Delta_x^{(k+1)} = -\alpha\sum_{i=0}^{k}(A^{k+l-i} - RA^{2k+l-i})D_{i+l}^{-1}\mathbf{y}^{(i)}$.

3. $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = (A - I)\Delta_x^{(k)} - \alpha AD_{k+l}^{-1}\mathbf{y}^{(k)}$.

4. $\mathbf{y}^{(k)} = B_\infty\mathbf{y}^{(k)} + \sum_{i=0}^{k-1}(B^{k-i} - B_\infty)\Delta_g^{(i)}$

### D.5 CONSENSUS LEMMA

**Lemma 12** (Consensus Lemma)**.**

$$\sum_{k=0}^{T}\|\Delta_x^{(k+1)}\|_F^2 \le 4\alpha^2 s_A^2\kappa_A^2\sum_{k=0}^{T}\|\mathbf{y}^{(k)}\|_F^2 \tag{31}$$

*Proof.*

$$
\begin{aligned}
\|\Delta_x^{(k+1)}\|_F^2 &= \alpha^2 \|(I - RA^{k+l}) \sum_{i=0}^{k} (A^{k+l-i} - A_\infty) D_{i+l}^{-1} \mathbf{y}^{(i)}\|_F^2 \\
&\leq 4\alpha^2 \| \sum_{i=0}^{k} (A^{k+l-i} - A_\infty) D_{i+l}^{-1} \mathbf{y}^{(i)}\|_F^2 \\
&\leq 4\alpha^2 s_A^2 \sum_{k=0}^{T} \|D_{i+l}^{-1} \mathbf{y}^{(k)}\|_F^2 \leq 4\alpha^2 s_A^2 \kappa_A^2 \sum_{k=0}^{T} \|\mathbf{y}^{(k)}\|_F^2
\end{aligned} \tag{32}
$$

The first equality in (32) uses $\|I - RA^{k+l}\|_2 \leq 2$. The second inequality uses the rolling sum Lemma 11. ☐

### D.6 Gradient Tracking Analysis

**Lemma 13.** *When the learning rate satisfies $\alpha \leq \min\{\frac{1}{3L}, \frac{1}{10 s_A s_B \|A-I\|_2 L}, \frac{1}{10 s_B \kappa_A p_A L}\}$, the following inequality holds $\forall T \geq 1$:*

$$
\sum_{k=1}^{T} \mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2] \leq 63 s_B^2 n T \sigma^2 + 9n^2 \sum_{k=1}^{T} \mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2]. \tag{33}
$$

*Proof.* Using the fourth transformation in Sec. D.4, we have

$$
\begin{aligned}
\|\mathbf{y}^{(k)}\|_F^2 &= \|B_\infty \mathbf{y}^{(k)} + \sum_{i=0}^{k-1} (B^{k-i} - B_\infty) \Delta_g^{(i)}\|_F^2 \\
&\leq 2\|B_\infty \mathbf{y}^{(k)}\|_F^2 + 2\| \sum_{i=0}^{k-1} (B^{k-i} - B_\infty) \Delta_g^{(i)}\|_F^2.
\end{aligned} \tag{34}
$$

Note that $\Delta_g^{(k)}$ can be further decomposed as follows:

$$
\Delta_g^{(k)} = \mathbf{g}^{(k+1)} - \nabla f(\mathbf{x}^{(k+1)}) + \nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)}) - \mathbf{g}^{(k)}.
$$

Therefore, we can apply Cauchy-Schwarz inequality and obtain that

$$
\begin{aligned}
&\mathbb{E}[\|\Delta_g^{(k)}\|_F^2] \leq 3\mathbb{E}[\|\mathbf{g}^{(k+1)} - \nabla f(\mathbf{x}^{(k+1)})\|_F^2] + 3\mathbb{E}[\|\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)})\|_F^2] \\
&\quad + 3\mathbb{E}[\|\nabla f(\mathbf{x}^{(k)}) - \mathbf{g}^{(k)}\|_F^2] \\
&\leq 6n\sigma^2 + 3L^2 \mathbb{E}[\|(A - I)\Delta_x^{(k)} - \alpha A D_{k+\ell+1}^{-1} \mathbf{y}^{(k)}\|_F^2] \\
&\leq 6n\sigma^2 + 9L^2 \|A - I\|_2^2 \mathbb{E}[\|\Delta_x^{(k)}\|_F^2] \\
&\quad + 9\alpha^2 L^2 \|A D_{k+\ell+1}^{-1} - R\|_2^2 \mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2] + 9n\alpha^2 L^2 \mathbb{E}[\|\bar{g}^{(k)}\|_F^2].
\end{aligned} \tag{35}
$$

where the second inequality uses the boundness of noise and $\|RA^k\|_2 \le 1$, the last inequality uses Cauchy-Schwarz inequality. Now we sum up $\mathbb{E}[\|\mathbf{y}^{(k)}\|^2]$ from $k=1$ to $T$ and obtain that

$$\sum_{k=1}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|^2] \le 2\sum_{k=1}^{T}\mathbb{E}[\|B_\infty \mathbf{y}^{(k)}\|_F^2] + 2\sum_{k=1}^{T}\mathbb{E}[\|\sum_{i=0}^{k-1}(B^{k-i}-B_\infty)\Delta_g^{(i)}\|_F^2]$$
$$\le 2n^2\sum_{k=1}^{T}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + 2s_B^2\sum_{k=0}^{T}\mathbb{E}[\|\Delta_g^{(i)}\|^2]$$
$$\overset{(35)}{\le} (2n^2+9n\alpha^2L^2)\sum_{k=1}^{T}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + 18s_B^2nT\sigma^2 + 18s_B^2L^2\|A-I\|_2^2\sum_{k=0}^{T}\mathbb{E}[\|\Delta_x^{(k)}\|_F^2$$
$$+ 18\alpha^2L^2s_B^2\kappa_A^2q_A^2\sum_{k=0}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2]$$
$$\overset{(31),\alpha\le\frac{1}{3L}}{\le} 3n^2\sum_{k=1}^{T}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + 18s_B^2nT\sigma^2 + C_{yy}\sum_{k=0}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2]. \tag{36}$$

where $q_A = \sup_{k\ge 0}\|A-RA^{k+\ell+1}\|_2$, $C_{yy} = 72\alpha^2s_A^2s_B^2L^2\|A-I\|_2^2 + 18\alpha^2L^2s_B^2\kappa_A^2p_A^2$. When we set $\alpha \le \min\{\frac{1}{10s_As_B\|A-I\|_2L}, \frac{1}{10s_B\kappa_Aq_AL}\}$, $C_{yy} \le 2/3$. Therefore, we can subtract $\frac{2}{3}\sum_{k=0}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2]$ from the both sides of (36). Finally, using the fact that $\mathbb{E}[\|\bar{g}^{(k)}\|^2] \le \frac{\sigma^2}{n} + \mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2]$, we have

$$\sum_{k=1}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2] \le 9n^2\sum_{k=1}^{T}\mathbb{E}[\|\bar{g}^{(k)}\|^2] + 54s_B^2nT\sigma^2$$
$$\le 63s_B^2nT\sigma^2 + 9n^2\sum_{k=1}^{T}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2]$$

□

### D.7 Descent Lemma

**Lemma 14** (Preparation for Descent Lemma)**.** *If $\ell \ge \ell_0$, we have the following inequality:*

$$\sum_{k=0}^{T}\mathbb{E}[\|\mathbb{1}_n^\top(A^{k+\ell+2}-A^{k+\ell})\Delta_x^{(k+1)}\|^2] \le \frac{\alpha^2}{36T^2}\sum_{k=0}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2] \tag{37}$$

*Proof.* Our selection of $\ell$ can guarantee that $\|\mathbb{1}_n^\top A^k - n\pi_A^\top\| \le \frac{1}{24ns_A\kappa_ALT}, \forall k \ge \ell$ (The proof can be found in Sec. D.2 ). Therefore, the second part can be bounded as:

$$\sum_{k=0}^{T}\mathbb{E}[\|\mathbb{1}_n^\top(A^{k+\ell+2}-A^{k+\ell})\Delta_x^{(k+1)}\|^2]$$
$$\le \sum_{k=0}^{T}\|(\mathbb{1}_n^\top A^{k+\ell+2}-\pi_A^\top)-(\mathbb{1}_n^\top A^{k+\ell}-\pi_A^\top)\|^2\mathbb{E}[\|\Delta_x^{(k+1)}\|_F^2$$
$$\le \frac{1}{144s_A^2\kappa_A^2T^2}\sum_{k=0}^{T}\mathbb{E}[\|\Delta_x^{(k+1)}\|_F^2] \overset{(31)}{\le} \frac{\alpha^2}{36n^2L^2T^2}\sum_{k=0}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2]. \tag{38}$$

□

**Lemma 15** (Descent Lemma)**.** *When $\alpha \leq \min\{\frac{1}{3L}, \frac{1}{10Ls_A\kappa_A s_B\sqrt{n}}, \frac{1}{10Ls_B\kappa_A p_A}\}$ and $T \geq 30ns_A$, we have the following inequality:*

$$\frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)}\|^2] \leq \frac{4(f(x^{(0)})-f^*)}{\alpha T} + \frac{8\alpha L\sigma^2}{n} + 600\alpha^2L^2s_A^2s_B^2\sigma^2 + \frac{16000ns_A^2s_B^2\sigma^2}{3T^2} + \frac{1}{5n^2T}\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2] \tag{39}$$

*Proof.* Left-multiply $\mathbb{1}_n^\top A^{k+\ell}$ on both sides of (30a), we have

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \alpha\bar{g}^{(k)} + \mathbb{1}_n^\top(A^{k+\ell+2} - A^{k+\ell})\Delta_x^{(k+1)} \tag{40}$$

Further apply L-smoothness inequality using (40), we have

$$\begin{aligned}
&\mathbb{E}f(\boldsymbol{w}^{(k+1)}) - \mathbb{E}f(\boldsymbol{w}^{(k)}) \leq \mathbb{E}\left\langle \mathbb{1}_n^\top(A^{k+\ell+2} - A^{k+\ell})\Delta_x^{(k+1)}, \nabla f(\boldsymbol{w}^{(k)})\right\rangle \\
&\quad - \alpha\mathbb{E}\left\langle \bar{g}^{(k)}, \nabla f(\boldsymbol{w}^{(k)})\right\rangle + 0.5L\mathbb{E}[\|\mathbb{1}_n^\top(A^{k+\ell+2} - A^{k+\ell})\Delta_x^{(k+1)} - \alpha\bar{g}^{(k)}\|^2] \\
&\leq -\frac{\alpha - 4\alpha^2L}{2}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2] - (\frac{\alpha}{2} - \frac{\alpha}{4})\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] + \frac{\alpha L^2}{2n}\mathbb{E}[\|\Delta_x^{(k)}\|_F^2] \\
&\quad + (\alpha^{-1} + L)\mathbb{E}[\|\mathbb{1}_n^\top(A^{k+\ell+2} - A^{k+\ell})\Delta_x^{(k+1)}\|^2] + \frac{2\alpha^2L\sigma^2}{n} \\
&\leq -\frac{\alpha}{4}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2] - \frac{\alpha}{4}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] + \frac{\alpha L^2}{2n}\mathbb{E}[\|\Delta_x^{(k)}\|_F^2] \\
&\quad + \frac{4}{3\alpha}\mathbb{E}[\|\mathbb{1}_n^\top(A^{k+\ell+2} - A^{k+\ell})\Delta_x^{(k+1)}\|^2] + \frac{2\alpha^2L\sigma^2}{n},
\end{aligned} \tag{41}$$

where the first inequality uses assumption 2, and the last inequality uses $\alpha \leq \frac{1}{3L}$. Notice that $\boldsymbol{w}^{(0)} = x^{(0)}$, we sum up (41) from $k=0$ to $T$ and obtain that

$$\begin{aligned}
&\frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)}\|^2] + \frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2] \\
&\leq \frac{4}{\alpha T}\mathbb{E}[f(x^{(0)}) - f(\boldsymbol{w}^{(T+1)})] + \frac{8\alpha L\sigma^2}{n} \\
&\quad + \frac{16}{3\alpha^2T}\sum_{k=0}^{T}\|\mathbb{1}_n^\top(A^{k+\ell+2} - A^{k+\ell})\Delta_x^{(k+1)}\|^2 + \frac{2L^2}{nT}\sum_{k=0}^{T}\|\Delta_x^{(k)}\|^2 \\
&\overset{(37)}{\leq} \frac{4(f(x^{(0)})-f^*)}{\alpha T} + \frac{8\alpha L\sigma^2}{n} + C_{yg}\sum_{k=0}^{T}\mathbb{E}[\|\mathbf{y}^{(k)}\|_F^2] \\
&\overset{(33)}{\leq} \frac{4(f(x^{(0)})-f^*)}{\alpha T} + \frac{8\alpha L\sigma^2}{n} + 63C_{yg}s_B^2nT\sigma^2 \\
&\quad + 9n^2C_{yg}\sum_{k=1}^{T}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2] + C_{yg}\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2]
\end{aligned} \tag{42}$$

Where we define $C_{yg} = \frac{1}{36n^2T^3} + \frac{8\alpha^2L^2s_A^2\kappa_A^2}{nT}$. When $\alpha \leq \frac{1}{4L\sqrt{ns_A\kappa_A}}$, we have $9n^2C_{yg} \leq \frac{1}{T}$. Therefore, we can subtract $\frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2]$ from both sides of (42) and finish the proof of this lemma.

□

### D.8 Main theorem

**Theorem 5.** *When $\ell \geq 1 + \lceil \frac{\ln(LTn^2\kappa_A^2) - \ln(1-\beta_A)}{1-\beta_A} \rceil$, by setting $\alpha = (\alpha_1^{-1} + \alpha_2^{-1} + \alpha_3^{-1} + \alpha_4^{-1} + \alpha_5^{-1})^{-1}$, we have the following inequality:*

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] \leq \frac{8\sqrt{2L\Delta}\sigma}{\sqrt{nT}} + 6\left(\frac{140L^2\Delta^2 s_A^2 s_B^2\sigma^2}{T^2}\right)^{1/3} + \frac{1}{5n^2T}\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2] \\ + \frac{40L\Delta s_B\kappa_A(s_A\sqrt{n}+p_A)}{T} + \frac{16000ns_A^2s_B^2\sigma^2}{3T^2} + \frac{12L\Delta}{T}\end{aligned} \tag{43}$$

*where $\Delta := f(x^{(0)}) - f^*$, $\alpha_1 = \left(\frac{n\Delta}{2LT\sigma^2}\right)^{1/2}$, $\alpha_2 = \left(\frac{\Delta}{140L^2Ts_A^2s_B^2}\right)^{1/3}$, $\alpha_3 = \frac{1}{10Ls_A\kappa_A s_B\sqrt{n}}$, $\alpha_4 = \frac{1}{10s_B\kappa_A p_A L}$, $\alpha_5 = \frac{1}{3L}$.*

*Proof.* Our selection of $\alpha$ ensures that $\alpha \leq \min\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$. The learning rate $\alpha_1$, $\alpha_2$ further satisfies $\frac{4\Delta}{\alpha_1 T} = \frac{8\alpha_1 L\sigma^2}{n}$, $\frac{2\Delta}{\alpha_2 T} = 280\alpha_2^2L^2s_A^2s_B^2$. Plug $\alpha$ into (39), we have

$$\begin{aligned}&\frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] - (\frac{1}{5n^2T}\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2] + \frac{8960ns_A^2s_B^2\sigma^2}{3T^2}) \\ &\leq \frac{4\Delta}{T}(\alpha_1^{-1} + \alpha_2^{-1} + \alpha_3^{-1} + \alpha_4^{-1} + \alpha_5^{-1}) + \frac{8\alpha_1 L\sigma^2}{n} + 280\alpha_2^2L^2s_A^2s_B^2 \\ &= \frac{16\alpha_1 L\sigma^2}{n} + 840\alpha_2^2L^2s_A^2s_B^2 + \frac{40L\Delta s_B(s_A\sqrt{n}+\kappa_A p_A)}{T} + \frac{12L\Delta}{T} \\ &= \frac{8\sqrt{2L\Delta}\sigma}{\sqrt{nT}} + 6\left(\frac{140L^2\Delta^2s_A^2s_B^2\sigma^2}{T^2}\right)^{1/3} + \frac{40L\Delta s_A s_B\sqrt{n}}{T} + \frac{12L\Delta}{T},\end{aligned}$$

which finishes the proof of our main theorem. □

**Corollary 16.** *we have the following coarse estimate which only involves $\beta_A$, $\beta_B$, $\kappa_A$, $\kappa_B$ for demonstrating the convergence of* PULL-SUM-*GT*

$$\begin{aligned}&\frac{1}{T}\sum_{k=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] \\ &= \mathcal{O}\left(\frac{\sqrt{L\Delta}\sigma}{\sqrt{nT}} + \left(\frac{nL\Delta q_A q_B\sigma}{T}\right)^{2/3} + \frac{L\Delta n^{3/2}\kappa_A q_A q_B}{T} + \frac{n^3q_A^2q_B^2\sigma^2}{T^2}\right),\end{aligned} \tag{44}$$

*where $q_A := \frac{1+\ln(\kappa_A)}{1-\beta_A}$, $q_B := \frac{1+\ln(\kappa_B)}{1-\beta_B}$.*

*Proof.* Note that $\|A - RA^{k+l+1}\|_2 \leq \sqrt{n}$, $s_A \leq \frac{\sqrt{n}(2+\ln(\kappa_A))}{1-\beta_A} \leq 2\sqrt{n}q_A$, $s_B \leq \frac{\sqrt{n}(2+\ln(\kappa_B))}{1-\beta_B} \leq 2\sqrt{n}q_B$, by taking these estimate into (43), we obtain the corollary when the constant $\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2]$ is omitted. □

## E MG-Pull-Sum-GT Convergence

Implementation details of MG-PULL-SUM-GT are as follows:

**Algorithm 2:** MG-PULL-SUM-GT

**Initialize** $\boldsymbol{\psi}_i^{(0)} = \boldsymbol{e}_i$, $\boldsymbol{x}_i^{(0)} = x^{(0)}$, $\boldsymbol{y}_i^{(0)} = \boldsymbol{g}_i^{(0)} = \frac{1}{M}\sum_{m=1}^{M} \nabla F_i(\boldsymbol{x}_i^{(0)}; \xi_i^{(0,m)})$, $\ell \geq M$, $K = MT$ ;

**for** $s = 0, 1, \ldots, \ell - 1$, *each node $i$ in parallel* **do**

  $\boldsymbol{\psi}_i^{(s+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{\psi}_j^{(s)}$;

**end**

Let $\boldsymbol{d}_i^{(0)} = \boldsymbol{\psi}_i^{(\ell)}$, $v_i = \sum_{r=1}^{n} \psi_{ir}^{(M)}$ at each node $i$;

**for** $t = 0, 1, \ldots, T - 1$, *each node $i$ in parallel* **do**

  Let $\boldsymbol{\phi}_i^{(t+1,0)} = \boldsymbol{x}_j^{(t)} - \alpha(\sum_{r=1}^{n} d_{jr}^{(t)})^{-1} \boldsymbol{y}_j^{(k)}$;

  **for** $m = 0, 1, \ldots, M - 1$, *each node $i$ in parallel* **do**

    Update $\boldsymbol{\phi}_i^{(k+1,m+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{\phi}_j^{(t+1,m)}$;

    Update $\boldsymbol{d}_i^{(tM+m+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{d}_j^{(tM+m)}$;

  **end**

  Update $\boldsymbol{x}_i^{(t+1)} = \phi_i^{(t+1,M)}$ and compute $\boldsymbol{g}_i^{(t+1)} = \frac{1}{M}\sum_{m=1}^{M} \nabla F_i(\boldsymbol{x}_i^{(t+1)}; \xi_i^{(t+1,m)})$;

  Let $\boldsymbol{z}_i^{(t+1,0)} = v_i^{-1}(\boldsymbol{y}_i^{(t)} + \boldsymbol{g}_i^{(t+1)} - \boldsymbol{g}_i^{(t)})$;

  **for** $m = 0, 1, \ldots, M - 1$, *each node $i$ in parallel* **do**

    Update $\boldsymbol{z}_i^{(t+1,m+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{z}_j^{(t+1,m)}$;

  **end**

  Update $\boldsymbol{y}_i^{(t+1)} = (\sum_{r=1}^{n} \psi_{ir}^{(u)}) \boldsymbol{z}_i^{(t+1,M)}$ ;

**end**

We also provide a denser form of MG-PULL-SUM-GT:

$$\mathbf{x}^{(t+1)} = A'(\mathbf{x}^{(t)} - \alpha D_{(t+1)M}^{-1} \mathbf{y}^{(t)}) \tag{45a}$$

$$\mathbf{g}^{(t+1)} = \frac{1}{M}\sum_{m=1}^{M} \nabla F(\mathbf{x}^{(t+1)}, \xi^{(t,m)}) \tag{45b}$$

$$\mathbf{y}^{(t+1)} = B'(\mathbf{y}^{(t)} + \mathbf{g}^{(t+1)} - \mathbf{g}^{(t)}) \tag{45c}$$

Here $A' := A^M$, $B' := A^M D_M^{-1}$, $D_t := \mathrm{diag}(\mathbb{1}_n^\top A^t)$.

**Lemma 17.** *When $\ell \geq M \geq \lceil \frac{1+2\ln(\kappa_A)+2\ln(n)}{1-\beta_A} \rceil$, we have $n^2 s_{A'} \frac{\sigma^2}{L\Delta}, s_{B'}, np_{A'} \cdot \kappa_A = \mathcal{O}(1)$.*

*Proof.* Easy to verify that $\kappa_A = \kappa_{A'}$. note that $\beta_A^{\frac{1}{1-\beta_A}} \leq 1/e$, we have $\beta_A^M \leq exp(-1 - 2\ln(\kappa_A) - 2\ln(n)) = 1/(en^2\kappa_A^2)$. From the proof of Lemma 11 we know that $\|A^t - A_\infty\|_2 \leq \beta_A^t \kappa_A, \forall t \geq 0$. First we show that $\beta_{A'} \cdot \kappa_A^2 \leq \frac{1}{e}$. This can be derived from $\beta_{A'} = \|A^M - A_\infty\|_{\pi_A} \leq \|A - A_\infty\|_{\pi_A}^M = \beta_A^M \leq \frac{1}{en^2\kappa_A^2}$.

Then, we have $ns_{A'} = n\max_{k\geq 1} \|A'^k - A_\infty\|_2 \cdot \frac{1+\ln(\kappa_A)}{1-\beta_{A'}} \leq n\beta_A^M \kappa_A \cdot \frac{1+\ln(\kappa_A)}{1-\beta_A'} \leq \mathcal{O}(\frac{1}{ne}\frac{1+\ln(\kappa_A)}{\kappa_A}\min\{1, \frac{L\Delta}{\sigma^2}\}) = \mathcal{O}(\frac{L\Delta}{\sigma^2})$. $p_{A'} = \sup_{k\geq 0} \|(A^M - A_\infty)(I - RA^{kM+\ell-M+1})\| \leq \sqrt{n}\beta_A^M \kappa_A \leq 1/(en\kappa_A)$, so $np_{A'} \cdot \kappa_A = \mathcal{O}(1)$.

To understand $B$, we estimate $\|B' - R\|_F \leq \|A^M D_M^{-1} - R\|_F \leq \kappa_A \|(I - R)(A - A_\infty)^M\|_F \leq \kappa_A \beta_A^M \sqrt{n} \leq 1/(e\kappa_A n)$. This indicates that $\min_{i,j} b'_{ij} \geq \frac{1}{n} - \|B' - R\|_F \geq \frac{3}{5n}$, $\max_{i,j} b'_{ij} \leq \frac{1}{n} + \|B' - R\|_F \leq \frac{7}{5n}$. Therefore, $[\pi_{B'}]_i = \sum_{j=1} b_{ij}[\pi_{B'}]_j \in [\frac{3}{5n}, \frac{7}{5n}]$, $\kappa_{B'} \leq \frac{7}{3} = \mathcal{O}(1)$.

$\beta_{B'} = \|\Pi_{B'}^{-1/2}(A^M D_M^{-1} - \pi_{B'}\mathbb{1}_n^\top)\Pi_{B'}^{1/2}\|_2 \leq \kappa_{B'}\|(A^M D_M^{-1} - R) + (n^{-1}\mathbb{1}_n - \pi_{B'})\mathbb{1}_n^\top\|_2 \leq \kappa_{B'}(\frac{2}{5} + \kappa_A^2\beta_A^M) \leq \kappa'_B = \mathcal{O}(1)$.

Finally, $s_{B'} = \max_{k\geq 1} \|B'^k - B_\infty\|_2 \cdot \frac{1+\ln(\kappa_{B'})}{1-\beta_{B'}} \leq \beta_{B'}\kappa_{B'} \cdot \frac{1+\ln(\kappa_{B'})}{1-\beta_{B'}} = \mathcal{O}(1)$. $\square$

**Theorem 6.** *When $\ell \geq M = \frac{1+2\ln(n)+2\ln(\kappa_A)+|\ln(\frac{\sigma^2}{L\Delta})|}{1-\beta_A}$, we have the following convergence result for MG-*PULL-SUM-*GT:*

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] = \tilde{\mathcal{O}}\left(\frac{\sqrt{L\Delta}\sigma}{\sqrt{nK}} + \frac{(1+\ln(\kappa_A))L\Delta}{(1-\beta_A)K}\right) \tag{46}$$

*where $K = MT$ is the total rounds of communication. $\tilde{\mathcal{O}}(\cdot)$ absorbs some logarithmic factors including $\ln(\sigma), \ln(n), \ln(L), \ln(\Delta)$ and absolute constants including $\mathbb{E}[\|\mathbf{g}^{(0)}\|_F^2]$.*

*Proof.* We replace $s_A$, $s_B$, $p_A$, $\sigma$ with $s_{A'}$, $s_{B'}$, $p_{A'}$, $\hat{\sigma}$ in Theorem (43) respectively. Using the conclusion of Lemma 17, we obtain that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] = \mathcal{O}\left(\frac{\sqrt{L\Delta}\hat{\sigma}}{\sqrt{nT}} + \left(\frac{L^2\Delta^2\hat{\sigma}^2}{n^2T^2}\right)^{1/3} + \frac{L\Delta}{T} + \frac{\min\{L\Delta, 1\}}{Mn^2T^2}\right)\\
&= \mathcal{O}\left(\frac{\sqrt{L\Delta}\sigma}{\sqrt{MnT}} + \left(\frac{L^2\Delta^2\sigma^2}{n^2MT^2}\right)^{1/3} + \frac{ML\Delta}{MT} + \frac{L\Delta}{n^2MT^2}\right)\\
&= \mathcal{O}\left(\frac{\sqrt{L\Delta}\sigma}{\sqrt{nK}} + \left(\frac{ML^2\Delta^2\sigma^2}{n^2K^2}\right)^{1/3} + \frac{ML\Delta}{K} + \frac{ML\Delta}{n^2K^2}\right)
\end{aligned}$$

Note that $\left(\frac{ML^2\Delta^2\sigma^2}{n^2K^2}\right)^{1/3} \leq \frac{2}{3}\cdot\frac{\sqrt{L\Delta}\sigma}{\sqrt{nK}} + \frac{1}{3}\cdot\frac{ML\Delta}{nK}$ and $\frac{ML\Delta}{n^2K^2} \leq \frac{ML\Delta}{K}$, we thus obtain (46). $\square$

# F PULL-DIAG-GT CONVERGENCE

## F.1 NOTATION

We denote $\mathbb{1}_n$ as an $n$-dimensional all-ones vector. We define $I_n \in \mathbb{R}^{n\times n}$ as the identity matrix. Throughout the paper, $A$ is always a row-stochastic matrix, *i.e.*, $A\mathbb{1}_n = \mathbb{1}_n$. We denote $[n]$ as the index set $\{1, 2, \ldots, n\}$. We denote $\mathrm{Diag}(A)$ as the the diagonal matrix generated from the diagonal entries of $A$. We denote $\mathrm{diag}(v)$ as the diagonal matrix whose diagonal entries comes from vector $v$. We denote $\pi_A$ as the left Perron vector of $A$. We denote $\Pi_A = \mathrm{diag}(\pi_A)$, $\pi_A$-vector norm $\|v\|_{\pi_A} = \|\Pi_A^{1/2}v\|$ and the induced $\pi_A$-matrix norm as $\|W\|_{\pi_A} = \|\Pi_A^{1/2}W\Pi_A^{-1/2}\|_2$. We define $A_\infty = \mathbb{1}_n\pi_A^\top$, $\beta_A = \|A - A_\infty\|_{\pi_A}$, $\kappa_A = \max(\pi_A)/\min(\pi_A)$, $q_A = \max_{k\geq 0} p_A$. Throughout the paper, we let $\boldsymbol{x}_i^{(k)} \in \mathbb{R}^d$ denote the local model copy at node $i$ at iteration $k$. Furthermore, we define the matrices

$$\begin{aligned}
\mathbf{x}^{(k)} &:= [(\boldsymbol{x}_1^{(k)})^\top; (\boldsymbol{x}_2^{(k)})^\top; \cdots; (\boldsymbol{x}_n^{(k)})^\top] \in \mathbb{R}^{n\times d},\\
\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) &:= [\nabla F_1(\boldsymbol{x}_1^{(k)}; \xi_1^{(k)})^\top; \cdots; \nabla F_n(\boldsymbol{x}_n^{(k)}; \xi_n^{(k)})^\top] \in \mathbb{R}^{n\times d},\\
\nabla \mathbf{f}_k &:= [\nabla f_1(\boldsymbol{x}_1^{(k)})^\top; \nabla f_2(\boldsymbol{x}_2^{(k)})^\top; \cdots; \nabla f_n(\boldsymbol{x}_n^{(k)})^\top] \in \mathbb{R}^{n\times d},
\end{aligned}$$

by stacking all local variables. The upright bold symbols (e.g. $\mathbf{x}, \mathbf{w}, \mathbf{g} \in \mathbb{R}^{n\times d}$) always denote stacked network-level quantities. Throughout the proof, we define $\overline{\nabla f(\mathbf{x}^{(k)})} = n^{-1}\mathbb{1}_n^\top \nabla f(\mathbf{x}^{(k)})$, $\Delta_x^{(k)} := (I - A_\infty)\mathbf{x}^{(k)}$, $\Delta_g^{(k)} = \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}$.

## F.2 ASSUMPTION

**Assumption 6** (First-order Lipschitz continuity.)**.** *There exists a constant $L$ such that $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$, $\forall i = 1, 2 \ldots n$.*

**Assumption 7** (Heterogeneity Bound.)**.** *There exists some constant $b$ such that $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\boldsymbol{x}) - \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\boldsymbol{x})\|^2 \leq b^2$ for every $x \in \mathbb{R}^d$.*

**Assumption 8** (Gradient Oracle.)**.** *There exists some constant $\sigma$ such that $\mathbb{E}[\|\nabla F_i(\boldsymbol{x}, \xi_i) - f_i(\boldsymbol{x})\|^2] \leq \sigma^2$, $\forall i = 1, 2 \ldots n$.*

### F.3 Some Useful Equations and Inequalities

**Lemma 18** (PROPERTIES OF $\pi_A$ NORM). *By definition of $\pi_A$ norm, we have*

1. $\|A - A_\infty\|_{\pi_A} = \beta_A < 1$.
2. $\|I - A_\infty\|_{\pi_A} = 1$

**Lemma 19** (RELATIONSHIP BETWEEN F-NORM AND $\pi_A$-NORM.). *The following inequalities hold*

1. $\|(A - A_\infty)^i \mathbf{z}\|_F^2 \le \kappa_A \beta_A^{2i} \|\mathbf{z}\|_F^2$.
2. $\|UV\|_F^2 \le \kappa_A \|U\|_{\pi_A}^2 \|V\|_F^2$

*Proof.* Denote that $V = [\boldsymbol{v}_i]_{i=1}^n$ and $\underline{\pi_A} = \min \pi_A$. By definition

$$
\begin{aligned}
\|UV\|_F^2 =& \sum_{i=1}^n \|\text{diag}(\pi)^{-0.5} U \boldsymbol{v}_i\|_{\pi_A}^2 \\
\le& \frac{1}{\underline{\pi_A}} \sum_{i=1}^n \|U\boldsymbol{v}_i\|_{\pi_A}^2 = \frac{1}{\underline{\pi_A}} \sum_{i=1}^n \|\text{diag}(\pi) U \text{diag}(\pi)^{0.5} \text{diag}(\pi) \boldsymbol{v}_i\|^2 \\
\le& \frac{1}{\underline{\pi_A}} \|U\|_{\pi_A}^2 \sum_{i=1}^n \|\text{diag}(\pi)\boldsymbol{v}_i\|^2 \le \kappa_A \|U\|_{\pi_A}^2 \|V\|_F^2.
\end{aligned}
$$

The first inequality can be derived from the second one. $\square$

**Lemma 20** (CONVERGENCE OF DIAGONAL MATRIX). *The following inequalities hold for all $k \ge 1$.*

1. $\|D_k^{-1} - \mathit{diag}(A_\infty)^{-1}\|_2 \le \theta_A^2 \sqrt{\kappa_A n} \beta_A^k$.
2. $\|D_k^{-1} - D_{k+1}^{-1}\|_2 \le 2\theta_A^2 \sqrt{\kappa_A n} \beta_A^k$.

*Proof.* Denote that $\Pi = \text{diag}(A_\infty)^{-1}$ and $\underline{\pi_A} = \min \pi_A$. Then we have $\|D_k^{-1} - \Pi^{-1}\|_2 = \|D_k^{-1}(D_k - \Pi)\Pi^{-1}\|_2 \le \theta_A^2 \|D_k - \Pi\|_2$. It is sufficient to estimate $\|D_k - \Pi\|_2$.

$$
\begin{aligned}
\|D_k - \Pi\|_2^2 \le& \max_i \|(A^k - A_\infty)\mathbf{e}_i\|_2^2 \le \sum_{i=1}^n \|(A^k - A_\infty)\mathbf{e}_i\|_2^2 \\
=& \sum_{i=1}^n \|\text{diag}(\pi)^{-0.5}(A^k - A_\infty)\mathbf{e}_i\|_{\pi_A}^2 \le \frac{1}{\underline{\pi_A}} \beta_A^k \sum_{i=1}^n \|\mathbf{e}_i\|_\pi^2 \\
\le& \kappa_A n \beta_A^{2k}.
\end{aligned}
$$

$\square$

**Lemma 21.** *$\|\nabla \mathbf{f}(\mathbf{w}^{(k)})\|_F^2$ is bounded by $\|\nabla f(\boldsymbol{w}^{(k)})\|$. Because of the heterogeneity bound, we obtain the inequality that $\|\nabla \mathbf{f}(\mathbf{w}^{(k)})\|_F^2 \le 2n\|\nabla f(\boldsymbol{w}^{(k)})\|^2 + 2nb^2$.*

*Proof.* Because of the heterogeneity bound, we have

$$
\begin{aligned}
\|\nabla \mathbf{f}(\mathbf{w}^{(k)})\|_F^2 =& \sum_{i=1}^n \|\nabla f_i(\boldsymbol{w}^{(k)})\|^2 \\
\le& 2\sum_{i=1}^n \|\nabla f_i(\boldsymbol{w}^{(k)}) - \nabla f(\boldsymbol{w}^{(k)})\|^2 + 2n\|\nabla f(\boldsymbol{w}^{(k)})\|^2 \\
\le& 2n\|\nabla f(\boldsymbol{w}^{(k)})\|^2 + 2nb^2.
\end{aligned}
$$

$\square$

**Algorithm 3:** PULL-DIAG-GT

1 **Initialize** $\boldsymbol{d}_i^{(0)} = \boldsymbol{e}_i$, $\boldsymbol{x}_i^{(0)} = x^{(0)}$, $\boldsymbol{y}_i^{(0)} = \boldsymbol{g}_i^{(0)} = \mathbb{1}_d^\top$;
2 **for** $k = 0, 1, \ldots, K$, *each node* $i$ *in parallel* **do**
3 $\quad \boldsymbol{x}_i^{(k+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{x}_j^{(k)} - \alpha(\boldsymbol{y}_j^{(k)};$
4 $\quad \boldsymbol{d}_i^{(k+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{d}_j^{(k)};$
5 $\quad$ Locally calculate $\boldsymbol{r}_i^{(k+1)} = (d_{ii}^{(k+1)})^{-1} \nabla F_i(\boldsymbol{x}_i^{(k+1)}, \xi_i^{(k+1)});$
6 $\quad$ Locally calculate $\boldsymbol{z}_i^{(k)} = \boldsymbol{r}_i^{(k+1)} - \boldsymbol{r}_i^{(k)}$ ;
7 $\quad \boldsymbol{y}_i^{(k+1)} = \sum_{j \in \mathcal{N}_i^{in}} a_{ij} \boldsymbol{y}_j^{(k)} + \boldsymbol{z}_i^{(k)};$
8 **end**

**Lemma 22.** *When we have the initial setting,* $\mathbf{y}^{(0)} = \mathbf{g}_0$*, for a given* $k > 0$*, we have* $\pi_A^T \mathbf{y}^{(k)} = \pi_A^T D_k^{-1} \mathbf{g}_k$*. In this way, we take the expectation of both sides* $\mathbb{E}[\pi^T \mathbf{y}^{(k)}] = \pi^T D_k^{-1} \nabla \mathbf{f}_k$*.*

*Proof.* Since $D_0 = I$, the proposition holds true when $k = 0$. Given the proposition holds when $n \leq k$, for $n = k + 1$:

$$\begin{aligned}
\pi^T \mathbf{y}^{(k+1)} =& \pi^T \mathbf{y}^{(k)} + \pi^T D_{k+1}^{-1} \mathbf{g}_{k+1} - \pi^T D_k^{-1} \mathbf{g}_k \\
=& \pi^T D_{k+1}^{-1} \mathbf{g}_{k+1}.
\end{aligned}$$

Then it holds for $n = k + 1$. By induction, the proposition holds for all $k \geq 0$. $\square$

### F.4 ALGORITHM FORMULATION

Pull-diag-AGT:

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} - \alpha \mathbf{y}^{(k)} \tag{47a}$$
$$V_{k+1} = AV_k \tag{47b}$$
$$\mathbf{y}^{(k+1)} = A\mathbf{y}^{(k)} + D_{k+1}^{-1} \mathbf{g}_{k+1} - D_k^{-1} \mathbf{g}_k \tag{47c}$$

hereon we set $D_0 = I_n$, $\mathbf{y}^{(0)} = \mathbf{g}_0$.

### F.5 BASIC TRANSFORMATIONS

1. $\Delta_x^{(k+1)} = -\alpha \sum_{i=0}^k (A - \mathbb{1}_n \pi^T)^i (I_n - \mathbb{1}_n \pi^T) \mathbf{y}^{(k-i)}$.
2. $\mathbf{y}^{(k+1)} = A_\infty \mathbf{y}^{(k+1)} + \sum_{i=0}^k [(A - A_\infty)^{k-i} (D_{i+1}^{-1} - A_\infty D_{i+1}^{-1}) \Delta_g^i - (A - A_\infty)^{k+1-i} \delta_i \mathbf{g}_i]$.

### F.6 CONSENSUS

**Lemma 23** (CONSENSUS LEMMA)**.**

$$\sum_{k=0}^T \|\Delta_x^{(i+1)}\|_F^2 \leq \alpha^2 \frac{\kappa_A}{(1 - \beta_A)^2} \sum_{k=0}^T \|\mathbf{y}^{(k)}\|_F^2. \tag{48}$$

*Proof.* we can use Jensen inequality and apply 19

$$\begin{aligned}
\|\Delta_x^{(i+1)}\|_F^2 \leq& \alpha^2 \sum_{i=0}^k \frac{1}{(1 - \beta_A)\beta_A^i} \|(A - \mathbb{1}_n \pi^T)^i (I_n - \mathbb{1}_n \pi^T) \mathbf{y}^{(k-i)}\|_F^2 \\
\leq& \alpha^2 \sum_{i=0}^k \frac{\kappa_A \beta_A^i}{(1 - \beta_A)} \|\mathbf{y}^{(k-i)}\|_F^2.
\end{aligned} \tag{49}$$

By summing up (49) from 0 To $T$, we have

$$\sum_{k=0}^{T}\|\Delta_x^{(i+1)}\|_F^2 \le \alpha^2\sum_{k=0}^{T}\sum_{i=0}^{k}\frac{\kappa_A\beta_A^i}{(1-\beta_A)}\|\mathbf{y}^{(k-i)}\|_F^2 \le \frac{\alpha^2\kappa_A}{(1-\beta_A)^2}\sum_{k=0}^{T}\|\mathbf{y}^{(k)}\|_F^2.$$

□

### F.7 GRADIENT TRACKING

**Lemma 24** (GRADIENT TRACKING ). *We have*

$$\begin{aligned}&\sum_{k=0}^{T+1}\|\mathbf{y}^{(k+1)}\|_F^2\\ \le& 4\sum_{k=0}^{T+1}\|A_\infty\mathbf{y}^{(k+1)}\|_F^2 + (C_{01}(T+1)+C_{02})\sigma^2 + C_{03}\sum_{k=0}^{T}\|\nabla f(\mathbf{w}^{(k)})\|_F^2.\end{aligned}$$

*Where* $\alpha \le \alpha_1 = \left[\frac{(1-\beta_A)^4}{48L^2\theta_A^2\kappa_A(2n\theta_A^4\kappa^2(\pi_A)\beta_A+\kappa_A\|A-I\|_F^2+(1-\beta_A)^2)}\right]^{0.5}$ *and*

1. $C_{01} = \frac{48n\kappa_A\theta_A^2}{(1-\beta_A)^2}$.
2. $C_{02} = \frac{96n^2\kappa^2(\pi_A)\theta_A^6\beta_A^2}{(1-\beta_A)^3}$.
3. $C_{03} = \frac{96n\kappa^2(\pi_A)\theta_A^6\beta_A^2}{(1-\beta_A)^2}$

*Proof.* Firstly, applying lemma 19 and noticing $\|I - A_\infty\|_{\pi_A} = 1$ we have

$$\begin{aligned}&\|(A-A_\infty)^{k-i}(D_{i+1}^{-1} - A_\infty D_{i+1}^{-1})\Delta_g^i - (A-A_\infty)^{k+1-i}\delta_i\mathbf{g}_i\|_F^2\\ \le& 2\|(A-A_\infty)^{k-i}(D_{i+1}^{-1} - A_\infty D_{i+1}^{-1})\Delta_g^i\|_F^2 + 2\|(A-A_\infty)^{k+1-i}\delta_i\mathbf{g}_i\|_F^2\\ \le& 2\kappa_A\beta_A^{2(k-i)}\|D_{i+1}^{-1}\Delta_g^i\|_F^2 + 2\kappa_A\beta_A^{2(k-i+1)}\|\delta_i\mathbf{g}_i\|_F^2.\end{aligned} \tag{50}$$

Use the second transformation in F.5 ,(50) and Jensen's inequality, we have

$$\begin{aligned}&\|\mathbf{y}^{(k+1)}\|_F^2 \le 2\|A_\infty\mathbf{y}^{(k+1)}\|_F^2\\ &+2\|\sum_{i=0}^{k}[(A-A_\infty)^{k-i}(D_{i+1}^{-1} - A_\infty D_{i+1}^{-1})\Delta_g^i - (A-A_\infty)^{k+1-i}\delta_i\mathbf{g}_i]\|_F^2\\ &\le 2\|A_\infty\mathbf{y}^{(k+1)}\|_F^2 + 4\sum_{i=0}^{k}\frac{\beta_A^{k-i}\kappa_A}{(1-\beta_A)}\|D_{i+1}^{-1}\Delta_g^i\|_F^2 + 4\sum_{i=0}^{k}\frac{\beta_A^{k-i+2}\kappa_A}{(1-\beta_A)}\|\delta_i\mathbf{g}_i\|_F^2\end{aligned}$$

Note that $\Delta_g^{(k)}$ can be further decomposed as follows:

$$\Delta_g^{(k)} = \mathbf{g}^{(k+1)} - \nabla f(\mathbf{x}^{(k+1)}) + \nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)}) + \nabla f(\mathbf{x}^{(k)}) - \mathbf{g}^{(k)}.$$

Therefore, we can apply Cauchy-Schwarz inequality and obtain that

$$\begin{aligned}\|\Delta_g^{(k)}\|_F^2 &\le 3\|\mathbf{g}^{(k+1)} - \nabla f(\mathbf{x}^{(k+1)})\|_F^2 + 3\|\nabla f(\mathbf{x}^{(k+1)}) - \nabla f(\mathbf{x}^{(k)})\|_F^2\\ &\quad + 3\|\nabla f(\mathbf{x}^{(k)}) - \mathbf{g}^{(k)}\|_F^2\\ &\le 6n\sigma^2 + 3L^2\|(A-I)\Delta_x^{(k)} - \alpha\mathbf{y}^{(k)}\|_F^2\\ &\le 6n\sigma^2 + 6L^2\|A-I\|_F^2\|\Delta_x^{(k)}\|_F^2 + 6\alpha^2L^2\|\mathbf{y}^{(k)}\|_F^2\end{aligned} \tag{51}$$

Applying lemma 20, $\|\delta_i\mathbf{g}_i\|_F^2$ can be estimated as

$$\begin{aligned}&\|\delta_i\mathbf{g}_i\|_F^2 \le 4\theta_A^4\kappa_A n\beta_A^{2i}\|\mathbf{g}_i\|_F^2\\ \le& 4\theta_A^4\kappa_A n\beta_A^{2i}[3n\sigma^2 + 3\|\nabla f(\mathbf{x}^{(i)}) - \nabla f(\mathbf{w}^{(i)})\|_F^2 + 3\|\nabla f(\mathbf{w}^{(i)})\|_F^2]\\ \le& 4\theta_A^4\kappa_A n\beta_A^{2i}[3n\sigma^2 + 3L^2\|\Delta_x^i\|_F^2 + 3\|\nabla f(\mathbf{w}^{(i)})\|_F^2]\end{aligned} \tag{52}$$

Applying (51), (52) and and consensus lemma 23, we obtain

$$\begin{aligned}
&\sum_{k=0}^{T}\|\mathbf{y}^{(k+1)}\|_F^2\\
\leq&2\sum_{k=0}^{T}\|A_\infty\mathbf{y}^{(k+1)}\|_F^2+[\frac{24n(T+1)\kappa_A\theta_A^2}{(1-\beta_A)^2}+\frac{48n^2\kappa^2(\pi_A)\theta_A^6\beta_A^2}{(1-\beta_A)^3}]\sigma^2\\
&+[\frac{24\alpha^2L^2\kappa^2(\pi_A)\theta_A^2\|A-I\|_F^2}{(1-\beta_A)^4}+\frac{48\alpha^2nL^2\kappa^3(\pi_A)\theta_A^6\beta_A^2}{(1-\beta_A)^4}]\sum_{k=0}^{T}\|\mathbf{y}^{(k)}\|_F^2\\
&+\frac{24\alpha^2L^2\kappa_A\theta_A^2}{(1-\beta_A)^2}\sum_{k=0}^{T}\|\mathbf{y}^{(k)}\|_F^2+\frac{48n\kappa^2(\pi_A)\theta_A^6\beta_A^2}{(1-\beta_A)^2}\sum_{k=0}^{T}\|\nabla f(\mathbf{w}^{(k)})\|_F^2.
\end{aligned}$$

Since $\alpha\leq\left[\frac{(1-\beta_A)^4}{48L^2\theta_A^2\kappa_A(2n\theta_A^4\kappa^2(\pi_A)\beta_A+\kappa_A\|A-I\|_F^2+(1-\beta_A)^2)}\right]^{0.5}$ holds, we obtain

$$\begin{aligned}
&\sum_{k=0}^{T+1}\|\mathbf{y}^{(k+1)}\|_F^2\\
\leq&4\sum_{k=0}^{T+1}\|A_\infty\mathbf{y}^{(k+1)}\|_F^2+(C_{01}(T+1)+C_{02})\sigma^2+C_{03}\sum_{k=0}^{T}\|\nabla f(\mathbf{w}^{(k)})\|_F^2.
\end{aligned}$$

□

### F.8 DESCENT

**Lemma 25** (PREPARATION FOR DESCENT LEMMA 1)**.** *We have the equality that*

$$\begin{aligned}
&-\alpha\mathbb{E}\left[\left\langle\nabla f(\boldsymbol{w}^{(k)}),\pi_A^TD_k^{-1}\mathbf{f}_k\right\rangle\right]+\frac{\alpha^2L}{2}\mathbb{E}\left[\|\pi_A^T\mathbf{y}^{(k)}\|^2\right]\\
\leq&-\frac{\alpha d_k}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right]+\frac{\alpha}{2d_k}\mathbb{E}\left[\|d_k\nabla f(\boldsymbol{w}^{(k)})-\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right]\\
&+2\alpha^2nL\sigma^2+2\alpha^2L\theta_A^6\kappa_An\beta_A^{2k}\sigma^2+(\alpha^2L-\frac{\alpha}{2d_k})\mathbb{E}\left[\|\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right].
\end{aligned}$$

*Where* $d_k=\sum_{i=1}^n\frac{\pi_i}{D_{k,i}}$.

*Proof.* Applying the gradient oracle assumption, we obtain

$$\begin{aligned}
&\frac{\alpha^2L}{2}\mathbb{E}\left[\|\pi_A^TD_{k+i}^{-1}\mathbf{g}_k\|^2\right]\\
\leq&\alpha^2L\mathbb{E}\left[\|\pi_A^TD_k^{-1}(\mathbf{g}_k-\nabla\mathbf{f}_k)\|^2\right]+\alpha^2L\mathbb{E}\left[\|\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right]\\
\leq&2\alpha^2nL\sigma^2+2\alpha^2L\mathbb{E}\left[\|(\pi_A^TD_k^{-1}-\mathbb{1}_n^T)(\mathbf{g}_k-\nabla\mathbf{f}_k)\|_F^2\right]+\alpha^2L\mathbb{E}\left[\|\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right]\\
\leq&2\alpha^2nL\sigma^2+2\alpha^2L\theta_A^6\kappa_An\beta_A^{2k}\sigma^2+\alpha^2L\mathbb{E}\left[\|\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right].
\end{aligned}$$

Therefore we have

$$\begin{aligned}
&-\alpha d_k^{-1}\mathbb{E}\left[\left\langle d_k\nabla f(\boldsymbol{w}^{(k)}),\pi_A^TD_k^{-1}\mathbf{f}_k\right\rangle\right]+\frac{\alpha^2L}{2}\mathbb{E}\left[\|\pi_A^T\mathbf{y}^{(k)}\|^2\right]\\
\leq&-\frac{\alpha d_k}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right]-\frac{\alpha}{2d_k}\mathbb{E}\left[\|\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right]\\
&+\frac{\alpha}{2d_k}\mathbb{E}\left[\|d_k\nabla f(\boldsymbol{w}^{(k)})-\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right]+\frac{\alpha^2L}{2}\mathbb{E}\left[\|\pi_A^T\mathbf{y}^{(k)}\|^2\right]\\
=&-\frac{\alpha d_k}{2}\mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right]+\frac{\alpha}{2d_k}\mathbb{E}\left[\|d_k\nabla f(\boldsymbol{w}^{(k)})-\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right]\\
&+2\alpha^2nL\sigma^2+2\alpha^2L\theta_A^6\kappa_An\beta_A^{2k}\sigma^2+(\alpha^2L-\frac{\alpha}{2d_k})\mathbb{E}\left[\|\pi_A^TD_k^{-1}\nabla\mathbf{f}_k\|^2\right].
\end{aligned}$$

□

**Lemma 26** (PREPARATION FOR DESCENT LEMMA 2). *When heterogeneity bound assumption holds, we have the equality that*

$$\|d_k \nabla f(\boldsymbol{w}^{(k)}) - \pi_A^T D_k^{-1} \nabla \mathbf{f}_k\|^2 \leq L^2 \theta_A^2 \|\Delta_x^{(k)}\|_F^2.$$

*Proof.* Since the heterogeneity bound, we have

$$\|d_k \nabla f(\boldsymbol{w}^{(k)}) - \pi_A^T D_k^{-1} \nabla \mathbf{f}_k\|^2 = \|\pi_A^T D_k^{-1} (\nabla F(\mathbf{w}_k) - \nabla \mathbf{f}_k)\|^2 \leq L^2 \theta_A^2 \|\Delta_x^{(k)}\|_F^2.$$

□

### F.9 MAIN THEOREM

**Theorem 7** (CONVERGENCE OF PULL-DIAG-GT). *We can prove the linear speedup convergence rate under a proper choice of $T$ and $\alpha$,i.e.*

$$\sum_{k=0}^{T} \frac{d_k}{S_T} \mathbb{E}[\|\nabla f(\boldsymbol{w}^{(k)})\|^2] \leq \frac{16\sqrt{2L\Delta_0}\sigma}{\sqrt{n(T+1)}} + \frac{3(C_{12} 16 \Delta_0^2)^{\frac{1}{3}}}{n\sigma^{\frac{4}{3}}(T+1)^{\frac{2}{3}}}$$
$$+ \frac{2\sqrt{\Delta_0 C_{11}}\sigma + 3(16\Delta_0^2 C_{13}\sigma^2)^{\frac{1}{3}} + 3(16\Delta_0^2 C_{14} b^2)^{\frac{1}{3}}}{n(T+1)}.$$

*Where $C_{1i}$ are constants and $S_k = \sum_{k=0}^{T} d_k$.*

*Proof.* Since $\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \alpha \pi_A \mathbf{y}^{(k)}$ and $L$-smoothness of $f(\boldsymbol{x})$, we have

$$\begin{aligned}
&\mathbb{E}\left[f(\boldsymbol{w}^{(k+1)})\right] \leq \mathbb{E}\left[f(\boldsymbol{w}^{(k)})\right] - \alpha d_k^{-1} \mathbb{E}\left[\left\langle d_k \nabla f(\boldsymbol{w}^{(k)}), \pi_A^T D_k^{-1} \nabla \mathbf{f}_k \right\rangle\right] \\
&+ \frac{\alpha^2 L}{2} \mathbb{E}\left[\|\pi_A \mathbf{y}^{(k)}\|^2\right] \\
&\overset{25}{\leq} \mathbb{E}\left[f(\boldsymbol{w}^{(k)})\right] - \frac{\alpha d_k}{2} \mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right] + 2\alpha^2 n L (1 + \theta_A^6 \kappa_A \beta_A^{2k}) \sigma^2 \\
&+ \frac{\alpha}{2d_k} \mathbb{E}\left[\|d_k \nabla f(\boldsymbol{w}^{(k)}) - \pi_A^T D_k^{-1} \nabla \mathbf{f}_k\|^2\right] + (\alpha^2 L - \frac{\alpha}{2d_k}) \mathbb{E}\left[\|\pi_A^T D_k^{-1} \nabla \mathbf{f}_k\|^2\right] \\
&\overset{26}{\leq} \mathbb{E}\left[f(\boldsymbol{w}^{(k)})\right] - \frac{\alpha d_k}{2} \mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right] + (\alpha^2 L - \frac{\alpha}{2d_k}) \mathbb{E}\left[\|\pi^T D_k^{-1} \nabla \mathbf{f}_k\|^2\right] \\
&+ \frac{L^2 \theta_A^2 \alpha}{2d_k} \mathbb{E}\left[\|\Delta_x^{(k)}\|_F^2\right] + 2\alpha^2 n L \sigma^2 + 2\alpha^2 L \theta_A^6 \kappa_A n \beta_A^{2k} \sigma^2
\end{aligned} \tag{53}$$

Since $1 \leq \frac{1}{D_{k,i}} \leq \theta_A$, it holds that $1 \leq d_k \leq \theta_A$. Summing up (53) for 0 to $T$ we obtain

$$\begin{aligned}
&\frac{\alpha}{2} \sum_{k=0}^{T} d_k \mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right] \leq \Delta_0 + (\alpha^2 L - \frac{\alpha}{2\theta_A}) \sum_{k=0}^{T} \mathbb{E}\left[\|\pi^T D_k^{-1} \nabla \mathbf{f}_k\|^2\right] \\
&+ \frac{L^2 \theta_A^2 \alpha}{2} \sum_{k=0}^{T} \mathbb{E}\left[\|\Delta_x^{(k)}\|_F^2\right] + 2\alpha^2 n L (T+1) \sigma^2 + \frac{2\alpha^2 L \theta_A^6 \kappa_A n}{1 - \beta_A^2} \sigma^2.
\end{aligned}$$

Where $\Delta_0$ is defined as $f(w^{(0)}) - f^*$, notice that lemma 23 and lemma 24 hold, it holds that

$$\begin{aligned}
&\frac{\alpha}{2} \sum_{k=0}^{T} d_k \mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right] \\
&\leq \Delta_0 + \left[\alpha^2 L - \frac{\alpha}{2\theta_A} + \frac{2\alpha^3 L^2 \theta_A^2 n \kappa_A}{(1-\beta_A)^2}\right] \sum_{k=0}^{T} \mathbb{E}\left[\|\pi^T D_k^{-1} \nabla \mathbf{f}_k\|^2\right] \\
&+ \left[2\alpha^2 n L (T+1) + \frac{2\alpha^2 L \theta_A^6 \kappa_A n}{1 - \beta_A^2} + \frac{\alpha^3 L^2 \theta_A^2 \kappa_A}{2(1-\beta_A)^2} (C_{01}(T+1) + C_{02})\right] \sigma^2 \\
&+ \frac{\alpha^3 L^2 \theta_A^2 \kappa_A}{2(1-\beta_A)^2} C_{03} \left[2n \sum_{k=0}^{T} d_k \mathbb{E}\left[\|\nabla f(\boldsymbol{w}^{(k)})\|^2\right] + 2nb^2\right].
\end{aligned}$$

Applying lemma 19, we can estimate $d_k$ as $d_k \geq n - n\sqrt{n\kappa_A}\theta_A^3\beta_A^k$. In this way, we define $S_k = \sum_{i=0}^k d_i$, then we have $S_k \geq n(k+1) - \frac{n\sqrt{n\kappa_A}\theta_A^3}{1-\beta_A}$.
Denote that

1. $\alpha_2 = \frac{(1-\beta_A)\left[-L(1-\beta_A)+\sqrt{L^2(1-\beta_A)^2+4nL^2\theta_A\kappa_A}\right]}{4nL^2\theta_A^2\kappa_A}$.

2. $\alpha_3 \leq \frac{(1-\beta_A)}{2\sqrt{n\kappa_A}L\theta_A}$.

Let $\alpha \leq \alpha_2$ and $\alpha \leq \alpha_3$ ,Finally, we have

$$\begin{aligned}\sum_{k=0}^{T}\frac{d_k}{S_T}\mathbb{E}\|\nabla f(\boldsymbol{w}^{(k)})\|^2 \leq \frac{4\Delta_0}{\alpha S_T} + \frac{8\alpha Ln(T+1)\sigma^2}{S_T} + \frac{8\alpha L\theta_A^6\kappa_A n}{(1-\beta_A^2)S_T}\sigma^2 \\ +\frac{2\alpha^2L^2\theta_A^2\kappa_A}{(1-\beta_A)^2S_T}(C_{01}(T+1)+C_{02})\sigma^2 + \frac{4n\alpha^2L^2\theta_A^2\kappa_A}{(1-\beta_A)^2S_T}C_{03}b^2.\end{aligned} \tag{54}$$

Let $T \geq \frac{2\sqrt{n\kappa_A}\theta_A^3}{1-\beta_A}$, then $S_T \geq \frac{n(T+1)}{2}$. Define that

1. $C_{11} = \frac{16L\theta_A^6\kappa_A n}{(1-\beta_A^2)}$

2. $C_{12} = \frac{4L^2\theta_A^2\kappa_A}{(1-\beta_A)^2}C_{01}$

3. $C_{13} = \frac{4L^2\theta_A^2\kappa_A}{(1-\beta_A)^2}C_{02}$

4. $C_{14} = \frac{8nL^2\theta_A^2\kappa_A}{(1-\beta_A)^2}C_{03}$

Then (54) can be reformulated as

$$\begin{aligned}\sum_{k=0}^{T}\frac{d_k}{S_T}\mathbb{E}\|\nabla f(\boldsymbol{w}^{(k)})\|^2 \leq \frac{8\Delta_0}{\alpha n(T+1)} + 16\alpha L\sigma^2 + \frac{\alpha C_{11}}{n(T+1)}\sigma^2 \\ +\frac{\alpha^2C_{12}}{n}\sigma^2 + \frac{\alpha^2C_{13}}{n(T+1)}\sigma^2 + \frac{\alpha^2C_{14}}{n(T+1)}b^2.\end{aligned} \tag{55}$$

We define

1. $\alpha_4 = \left[\frac{\Delta_0}{2Ln(T+1)\sigma^2}\right]^{\frac{1}{2}}$.

2. $\alpha_5 = \left[\frac{\Delta_0}{C_{11}\sigma^2}\right]^{\frac{1}{2}}$.

3. $\alpha_6 = \left[\frac{4\Delta_0}{C_{12}\sigma^2(T+1)}\right]^{\frac{1}{3}}$.

4. $\alpha_7 = \left[\frac{4\Delta_0}{C_{13}\sigma^2}\right]^{\frac{1}{3}}$.

5. $\alpha_8 = \left[\frac{4\Delta_0}{C_{14}b^2}\right]^{\frac{1}{3}}$.

In (55), we let $\alpha \leq \frac{1}{\sum_{i=1}^{8} \alpha_i^{-1}}$ and $T \geq \frac{2\sqrt{n\kappa_A}\theta_A^3}{1-\beta_A}$, we finally prove the linear speedup of PULL-DIAG.

$$\sum_{k=0}^{T} \frac{d_k}{S_T} \mathbb{E}\|\nabla f(\boldsymbol{w}^{(k)})\|^2 \leq \frac{16\sqrt{2L\Delta_0}\sigma}{\sqrt{n(T+1)}} + \frac{3(3072n\kappa_A^2\theta_A^4 L^2\Delta_0^2\sigma^2)^{\frac{1}{3}}}{n(T+1)^{\frac{2}{3}}(1-\beta_A^2)^{\frac{4}{3}}}$$
$$+\frac{8\sqrt{n\kappa_A}\theta_A^3\sqrt{L\Delta_0}\sigma}{n(T+1)\sqrt{1-\beta_A^2}} + \frac{3(6144n^2\kappa_A^3\theta_A^8 L^2\beta_A^2\Delta_0^2\sigma^2)^{\frac{1}{3}}}{n(T+1)(1-\beta_A)^{\frac{5}{3}}}$$
$$+\frac{3(12288n^2\kappa_A^3\theta_A^8 L^2\beta_A^2\Delta_0^2 b^2)^{\frac{1}{3}}}{n(T+1)(1-\beta_A)^{\frac{4}{3}}}.$$

□

# G EXPERIMENT DETAILS

In this section we provide unexplained details of experiment setup in the article.

## G.1 NETWORK DESIGN

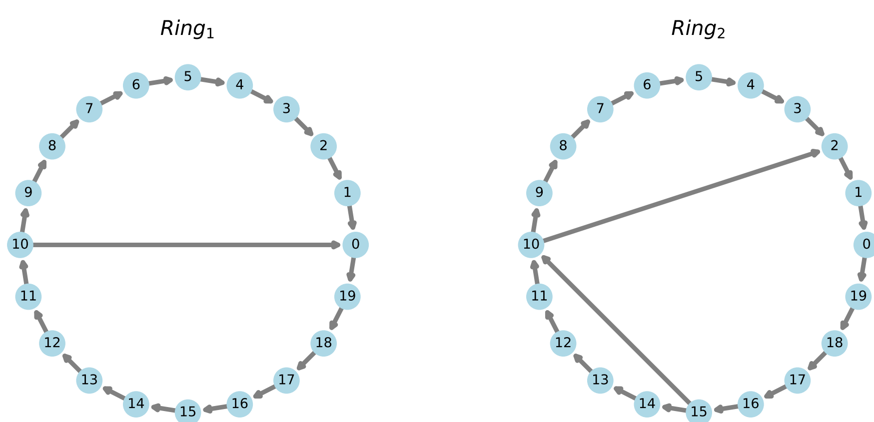


Figure A1: Left plot: $\mathrm{Ring}_1$, a directed ring graph with one extra connection. $\theta_A \approx 629145.6$, $\kappa_A \approx 2, \beta_A \approx 0.989$. Right plot: $\mathrm{Ring}_2$, a directed ring graph with two extra connections. $\theta_A \approx 53387, \kappa_A \approx 2.5, \beta_A \approx 0.989$.

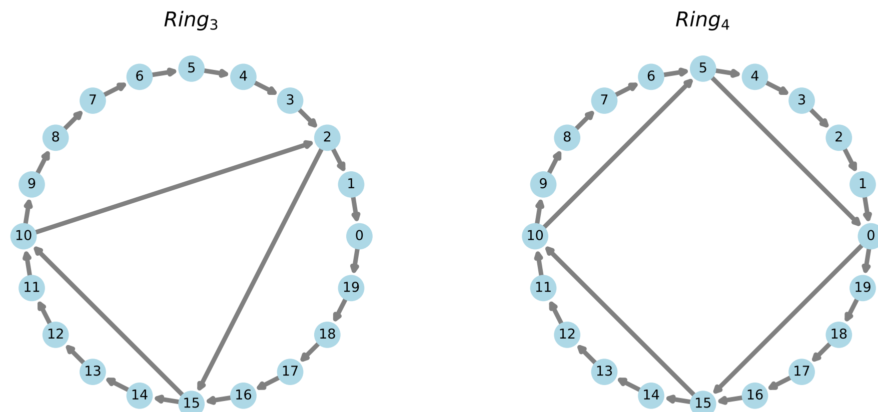


Figure A2: Left plot: $\mathrm{Ring}_3$, a directed ring graph with three extra connections. $\theta_A \approx 844$, $\kappa_A \approx 2$, $\beta_A \approx 0.989$. Right plot: $\mathrm{Ring}_4$, a directed ring graph with four extra connections. $\theta_A \approx 313$, $\kappa_A \approx 1.6$, $\beta_A \approx 0.970$.

### G.2 Figure 1 in Section 4

**Setup.** We randomly generate an initial vector $\boldsymbol{x}^{(0)} \in \mathbb{R}^{20}$, and then run either Pull-Diag and Pull-Sum consensus on a network of size 20 to estimate $\bar{\boldsymbol{x}}^{(0)}$. The estimate at iteration $k$ is denoted by $\boldsymbol{w}^{(k)}$, and the consensus error is computed as $\|\boldsymbol{w}^{(k)} - \bar{\boldsymbol{x}}^{(0)}\|/\|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}^{(0)}\|$.

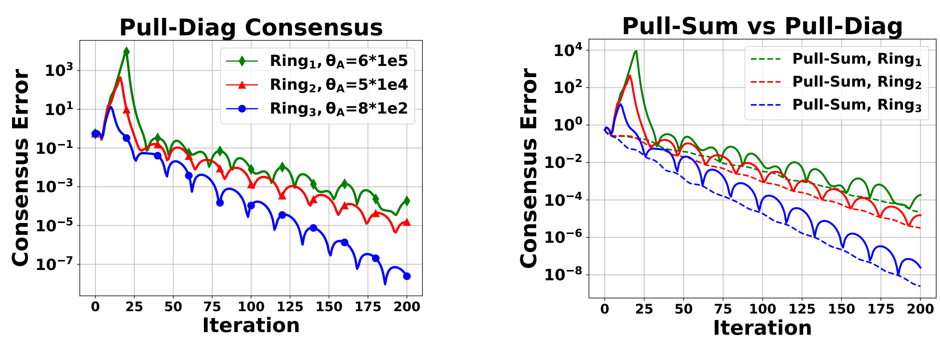


Figure A3: The left plot illustrates the Pull-Diag consensus on three different networks with varying $\theta_A$, while other parameters remain approximately constant, with $\kappa_A \approx 2$ and $\beta_A \approx 0.989$. It is observed that Pull-Diag exhibits a larger initial spike for networks with higher $\theta_A$. The right plot compares the consensus error of Pull-Sum (dashed lines) and Pull-Diag (solid lines). Pull-Sum consistently outperforms Pull-Diag across all cases.

### G.3 Figure 2 in Section 6

**Setup.** Our experiment on synthetic dataset focuses on a decentralized logistic regression with non-convex regularization. The objective is formulated as minimizing $\min_{x \in \mathbb{R}^d} n^{-1} \sum_{i=1}^{n} (f_i(x) + \rho r(x))$, where

$$f_i(x) = \frac{1}{M} \sum_{l=1}^{M} \ln(1 + \exp(-y_{i,l} h_{i,l}^\top x)) \quad \text{and} \quad r(x) = \sum_{j=1}^{d} \frac{[x]_j^2}{1 + [x]_j^2}.$$

Here, $\{h_{i,l}, y_{i,l}\}_{l=1}^{M}$ represents the training dataset held by node $i$, where $h_{i,l} \in \mathbb{R}^d$ denotes feature vectors and $y_{i,l} \in \{+1, -1\}$ signifies labels. The regularization term $r(x)$, which is non-convex, is controlled by $\rho > 0$.

We fix $d = 10$, $M = 20$, and $\rho = 0.001$. To accommodate varying data characteristics across nodes, each node $i$ computes a local solution $x_i^\star$. This solution is computed as $x_i^\star = x^\star + v_i$, where $x^\star \sim \mathcal{N}(0, I_d)$ is a shared, randomly generated vector, and $v_i \sim \mathcal{N}(0, \sigma_h^2 I_d)$ introduces variability among local solutions.

To generate local data reflecting diverse distributions, we sample each feature vector $h_{i,l} \sim \mathcal{N}(0, I_d)$ at node $i$. The corresponding label $y_{i,l}$ is determined by a random variable $z_{i,l} \sim \mathcal{U}(0,1)$, with $y_{i,l}$ set to 1 if $z_{i,l} < 1/(1 + \exp(-y_{i,l} h_{i,l}^\top x_i^\star))$ and $-1$ otherwise, thereby controlled by $x_i^\star$.

To introduce controlled gradient noise, we modify the actual gradient using Gaussian noise $\varepsilon_i$ such that $\nabla f_i(x) = \nabla f(x) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2 I_d)$. The parameter $\sigma_n^2$ governs the intensity of gradient noise.

In our experiments, we set $\sigma_h = 20$ and $\sigma_n = 0.001$. Our primary performance metric of interest across all experiments is $\|\nabla f(\bar{\mathbf{x}}^{(k)})\|$.

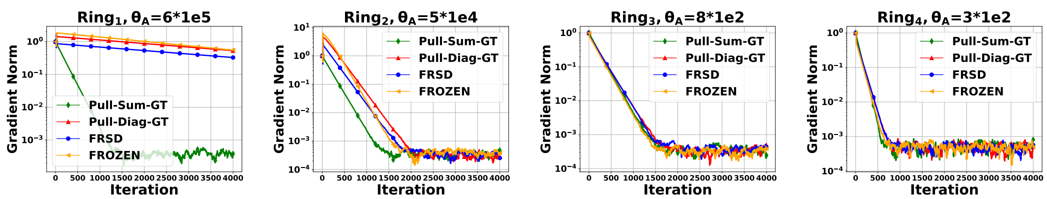


Figure A4: When $\theta_A$ is significant (Ring$_{1,2}$), PULL-SUM-GT is not influenced while PULL-DIAG based methods suffer a lot. When $\theta_A$ is small (Ring$_{3,4}$), their performance are similar.