# Towards Better Understanding the Influence of Directed Networks on Decentralized Stochastic Optimization

Liyuan Liang[*†] Xinmeng Huang[*‡] Ran Xin[§] Kun Yuan[¶]

**Abstract**

This paper investigates the influence of directed networks on the convergence of smooth non-convex stochastic decentralized optimization associated with column-stochastic mixing matrices. We find that the canonical spectral gap, a widely-used metric in undirected networks, alone fails to adequately characterize the impact of directed networks. Through a new analysis of the Push-Sum strategy, a fundamental building block for decentralized algorithms over directed graphs, we identify another novel metric called the equilibrium skewness. Next, we establish the first convergence lower bound for non-convex stochastic decentralized algorithms over directed networks, which explicitly manifests the impact of both the spectral gap and equilibrium skewness and justifies the imperative need for both metrics in analysis. Moreover, by jointly considering the spectral gap and equilibrium skewness, we present the state-of-the-art convergence rate for the Push-DIGing algorithm. Our findings reveal that these two metrics exert a more pronounced negative impact on Push-DIGing than suggested by our lower bound. We further integrate the technique of multi-round gossip to Push-DIGing to obtain MG-Push-DIGing, which nearly achieves the established lower bound, demonstrating its convergence optimality, best-possible resilience to directed networks, and the tightness of our lower bound. Experiments verify our theoretical findings.

## 1 Introduction

### 1.1 Decentralized optimization

Decentralized optimization has emerged as a promising paradigm for mitigating communication costs in large-scale signal processing and machine learning. In decentralized optimization, all computing nodes are interconnected through a specific network topology (*e.g.*, ring, grid, hypercube). Each node communicates merely with its immediate neighbors, avoiding communication across the entire network and hence leading

[*]Equal Contribution.

[†]Peking University. 2000010643@stu.pku.edu.cn.

[‡]University of Pennsylvania. xinmengh@sas.upenn.edu.

[§]ByteDance. ran.xin@bytedance.com

[¶]Corresponding author. Peking University. kunyuan@pku.edu.cn.

to substantially lower per-iteration communication overhead. Compared to centralized approaches, decentralized algorithms have demonstrated enhanced scalability and robustness to large numbers of computing nodes. They have found wide applications in wireless sensor networks, robotic control, and deep neural network training, see, *e.g.*, [1, 2, 3, 4].

Decentralized stochastic optimization over a connected network consisting of $n$ nodes can be formulated as

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

Here, the random variable $\xi_i$ represents the local data maintained at node $i$ following the distribution $\mathcal{D}_i$. Each node can locally evaluate its stochastic gradient $\nabla F(x;\xi_i)$ by sampling $\xi_i \overset{iid}{\sim} \mathcal{D}_i$, but it has to communicate to access information from other nodes in the network. Throughout this paper, we assume data heterogeneity exists, meaning the local data distributions $\{\mathcal{D}_i\}_{i=1}^n$ differ across nodes.

## 1.2 Influence of undirected network is well studied

The fundamental distinction between decentralized and centralized optimization lies in the utilization of decentralized network topologies. Consequently, a primary challenge in decentralized algorithms is elucidating the influence of network topology on their convergence performance. This challenge has been partially addressed for algorithms over *undirected* networks. Existing literature typically associates an undirected network with a doubly-stochastic

mixing matrix $W = [w_{ij}]_{i,j=1}^n$, where $w_{ij} \in (0,1)$ is the weight to scale information flowing from node $j$ to node $i$. If node $j$ cannot send information to node $i$, it holds that $w_{ij} = 0$. The *spectral gap* of $W$ is defined as $1-\beta$ where

$$\beta := \left\|W - \mathbb{1}_n\mathbb{1}_n^\top/n\right\|_2 \in [0,1) \quad \text{and} \quad \mathbb{1}_n := (1,\ldots,1)^\top \in \mathbb{R}^n.$$

The *spectral gap* is a prevalent metric used in literature to gauge the connectivity of an undirected network, with $1-\beta \to 1$ (implying $W \to \mathbb{1}_n\mathbb{1}_n^\top/n$) denoting a well-connected network and $1-\beta \to 0$ (implying $W \to I_n$) a poorly-connected one.

Utilizing the concept of the spectral gap, numerous results have been established to clarify the impact of undirected networks on decentralized algorithms. References such as [5, 6, 7, 8] demonstrate that decentralized stochastic gradient descent (SGD) can asymptotically achieve the same convergence rate as parallel SGD[1] after a specific number of transient iterations. The number of transient iterations is quantified in [6, 9] to be on the order of $\mathcal{O}(n^3/(1-\beta)^4)$, a value significantly influenced by the network topology. In the case of large and sparse networks where $1-\beta \to 0$, the transient stage is notably prolonged, potentially causing decentralized SGD to lag behind parallel SGD throughout the entire practical timeframe in implementations. To mitigate this, references [10, 11, 12, 13] propose Exact-Diffusion/D$^2$ and stochastic gradient tracking

[1]Parallel SGD involves a central server collecting information across the entire network.

strategies to shorten the transient stage by eliminating the influence of data heterogeneity, while [9] suggests the use of periodic global averaging to compensate the impact of sparse undirected network topology.

### 1.3 Influence of directed network is unclear

*Directed* network is a more general class of topology that encompasses undirected networks as a special case. Directed networks commonly manifest in scenarios where communication is permitted in one direction but not the other. For example, within a social network, one user can follow a celebrity without necessarily being followed back. Similarly, in a wireless sensor network, data may flow from sensor nodes toward an upper-level sink node without flowing back down to the sensors.

Extensive algorithms have been developed to solve problem (1.1) over directed networks associated with column-stochastic weight matrices, where the push-sum technique [14, 15] is used to achieve consensus among the nodes. When the true gradient $\nabla f_i(x)$ is available at each node, the seminal subgradient-push algorithm [16, 15] can converge to the desired solution, but with slow sublinear convergence even if each $f_i(x)$ is strongly-convex. Advanced algorithms such as EXTRA-Push [17], DEXTRA [18], ADD-OPT [19], and Push-DIGing [20] have been developed to achieve an enhanced linear convergence by mitigating the impact of data heterogeneity. In scenarios where only stochastic gradient $\nabla F(x;\xi_i)$ is accessible to each node, Stochastic Gradient Push [3] and S-ADDOPT [21] can asymptotically achieve the same convergence as parallel SGD. However, none of these works have clarified how the directed network topology influences the convergence of decentralized algorithms.

### 1.4 Challenges, main results, and contributions

Understanding the impact of directed networks poses two primary challenges: (1) the identification of metrics that can aptly capture the characteristics of directed networks, and (2) the derivation of a precise convergence rate that distinctly elucidates the influence of these metrics. This paper reveals the influence of directed network topology on decentralized algorithms by addressing these two challenges.

**Novel metrics for directed networks.** We employ the average consensus algorithm

$$x^{(k+1)} = Wx^{(k)} \quad \text{where} \quad x^{(0)} = x$$

to motivate metrics that capture the characteristics of directed networks. Given a column-stochastic mixing matrix $W \in \mathbb{R}^{n\times n}$ satisfying $\mathbb{1}_n^\top W = \mathbb{1}_n^\top$, the Perron-Frobenius theorem [22] guarantees a unique equilibrium vector $\pi \in \mathbb{R}^n$ with positive entries such that $W\pi = \pi$ and $\mathbb{1}_n^\top \pi = 1$. This implies that the power iterations starting from $x^{(0)} = x$ converges

$$x^{(k+1)} = Wx^{(k)} = W^k x^{(0)} = W^k x \;\longrightarrow\; \pi\mathbb{1}_n^\top x \quad \text{as} \quad k \to \infty,$$

which deviates from the desired global average $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top x$ due to using a column-stochastic mixing matrix[2].

[2]This deviation can be corrected by the Push-Sum strategy [14, 15], see details in Section 2

Inspired by relation (1.4), this paper identifies two effective metrics to capture the influence of directed networks on decentralized stochastic optimization by quantifying how close the iterate $x^{(k)}$ is to the desired global average $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top x$:

- The *spectral gap* $1-\beta_\pi$ of the column-stochastic $W$ in which

$$\beta_\pi := \left\|W-\pi\mathbb{1}_n^\top\right\|_\pi \in [0,1)$$

measures the rate at which $x^{(k)}$ converges to the weighted global average $\pi\mathbb{1}_n^\top x$. Apparently, it holds that $\|x^{(k)}-\pi\mathbb{1}_n^\top x\|_\pi = \|(W-\pi\mathbb{1}_n^\top)^k x\|_\pi \le \beta_\pi^k\|x\|_\pi$ where $\|\cdot\|_\pi$ is a weighted norm defined in Section 1.6. As $\beta_\pi$ approaches to 0, the iterate $x^{(k)}$ will converge faster to weighted global average $\pi\mathbb{1}_n^\top x$.

- The *equilibrium skewness* (in which $\pi_i$ is the $i$-th entry in $\pi$)

$$\kappa_\pi := \max_i \pi_i / \min_i \pi_i \in [1,+\infty)$$

captures the disagreement between the equilibrium vector $\pi$ and the uniform vector $n^{-1}\mathbb{1}_n$. When $\kappa_\pi \to 1$, the weighted average $\pi\mathbb{1}_n^\top x$ aligns better with the desired global average $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top x$.

This paper demonstrates that these two metrics together effectively reflect the influence of directed networks on decentralized algorithms. Omitting either of them would result in an inadequate understanding of the impact of directed networks.

**Tight analysis on the influence of directed networks.** To justify the efficacy of the spectral gap $1-\beta_\pi$ and the equilibrium skewness $\kappa_\pi$ in capturing the characteristics of directed networks, we demonstrate, *for the first time*, that the convergence rate of any first-order non-convex decentralized stochastic algorithm with a column-stochastic weight matrix is lower bounded by

$$\mathbb{E}[\|\nabla f(x^{(K)})\|_2^2] = \Omega\left(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}} + \frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}\right),$$

where notations $L$, $\Delta$, $K$, and $\sigma$ are referred to Section 3.1. It is noteworthy that both $\beta_\pi$ and $\kappa_\pi$ are present in this lower bound, underscoring that neither of these two metrics can be omitted from convergence rates, irrespective of algorithmic designs.

As no prior works have elucidated the impact of directed networks on non-convex decentralized stochastic optimization, it remains uncertain whether the established lower bound (1.4) can be attained by any existing algorithms. Consequently, we reexamine the seminal Push-DIGing algorithm [20] and present a novel convergence rate explicitly revealing the influence of spectral gap and equilibrium skewness, as detailed in Table 2. The observation indicates that these two metrics exert a more pronounced adverse effect on Push-DIGing than suggested by the established lower bound, highlighting a substantial disparity between the lower and upper bounds. To mitigate the adverse effects of directed networks, we propose Push-DIGing with multiple

Table 1: Upper bounds and Lower bounds for non-convex decentralized optimization based over directed networks. Here, $\sigma^2$ denotes variances of gradient noises, $L$ denotes the smoothness constant and $n$ is network size. "Rate (A.)" represents the asymptotic convergence rate excluding network influence (*i.e.*, $K \to \infty$). "Rate (F.T.)" represents the finite-time convergence rate influenced by networks explicitly. "N.A." means not available in existing literature.

| Algorithm | Rate (A.) | Rate (F.T.) | Transient Stage |
|---|---|---|---|
| Gradient-Push [3] | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}$ | N.A. | N.A. |
| Push-DIGing [23] | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}$ | N.A. | N.A. |
| Push-DIGing Theorem 2 | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}$ | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}} + \frac{\beta_\pi\kappa_\pi^3(1+\kappa_\pi\beta_\pi)}{(1-\beta_\pi)^2K}$ | $\frac{n\kappa_\pi^8}{(1-\beta_\pi)^4}$ |
| MG-Push-DIGing Theorem 3 | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}$ | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}} + \frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}$ | $\frac{n(1+\ln(\kappa_\pi))^2}{(1-\beta_\pi)^2}$ |
| Lower Bound Theorem 1 | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}$ | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}} + \frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}$ | $\frac{n(1+\ln(\kappa_\pi))^2}{(1-\beta_\pi)^2}$ |

gossip communications, a new algorithm that achieves a rate on the order of

$$\widetilde{\mathcal{O}}\left(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}} + \frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}\right).$$

Here, we absorb logarithmic factors that are independent of $\kappa_\pi$ and $\beta_\pi$ in $\widetilde{\mathcal{O}}(\cdot)$. Our new algorithm nearly attains the lower bound (1.4), demonstrating the tightness of the established lower bound. Our novel algorithm achieves a nearly optimal rate, showcasing the best possible resilience to directed networks' influence.

**Contributions.** The contributions in this paper can be summarized as follows.

- Through a novel convergence analysis of the push-sum strategy, we identify two key metrics crucial for understanding the impact of directed topologies on decentralized algorithms: the spectral gap that quantifies the speed of information diffusion and the skewness of the equilibrium vector $\pi$. As orthogonal notions, the combination of the two measures provides deeper insights into the influence of directed networks.

- We prove the first convergence lower bound of non-convex decentralized stochastic algorithms with column-stochastic mixing matrices. This lower bound underscores that neither spectral gap nor equilibrium skewness can be omitted from convergence rates. Furthermore, our lower bound clarifies the best possible resilience to directed networks' adverse influence on convergence rates.

- We revisit the Push-DIGing algorithm and provide the first convergence analysis revealing how directed networks affect its convergence rate. Our results show that the spectral gap and equilibrium skewness

have a more significant negative impact on Push-DIGing than our lower bound suggests. This emphasizes that Push-DIGing is far from optimal in its resilience to directed networks.

- To enhance the network resilience of Push-DIGing, we incorporate a multi-gossip communication strategy into Push-DIGing. The resulting algorithm, named MG-Push-DIGing, nearly achieves the lower bound we have established. This implies that our lower bound is tight, and MG-Push-DIGing is nearly optimal for non-convex stochastic decentralized optimization and attains the best possible resilience against the influence of directed networks.

### 1.5 More related works

The spectral gap of directed networks was first introduced by [24] and later utilized by [25, 26] to facilitate convergence analyses of decentralized algorithms over directed networks. However, these studies did not explicitly elucidate how the spectral gap influences convergence rates. A recent work [27] sheds light on the impact of the spectral gap on decentralized algorithms over directed networks without rigorous proof and is limited to deterministic and strongly convex optimization. To the best of our knowledge, no existing findings have jointly considered both the spectral gap and equilibrium skewness in analysis to clarify their combined effects on convergence rates.

Although this paper addresses decentralized optimization over column-stochastic directed networks, a few other studies such as [5, 28, 29, 30] focus on row-stochastic directed networks. It has been found in [25, 26] that algorithms relying solely on either a column-stochastic or row-stochastic mixing matrix converge relatively slowly. To significantly accelerate convergence, a novel $\mathcal{AB}$ algorithm [25] (also known as push-pull algorithm [26]) was proposed that alternates between using column-stochastic and row-stochastic mixing matrices. However, the influence of directed network topologies remains unclear in all these existing works. We conjecture the equilibrium skewness can play similar roles in these settings but leave the systematic study for future work.

### 1.6 Notations

Throughout the paper, we let $x_i^{(k)} \in \mathbb{R}^d$ denote the local model copy at node $i$ at iteration $k$. Furthermore, we define the matrices

$$\begin{aligned}
\mathbf{x}^{(k)} &= [(x_1^{(k)})^\top; (x_2^{(k)})^\top; \cdots; (x_n^{(k)})^\top] \in \mathbb{R}^{n\times d},\\
\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) &= [\nabla F_1(x_1^{(k)}; \xi_1^{(k)})^\top; \cdots; \nabla F_n(x_n^{(k)}; \xi_n^{(k)})^\top] \in \mathbb{R}^{n\times d},\\
\nabla f(\mathbf{x}^{(k)}) &= [\nabla f_1(x_1^{(k)})^\top; \nabla f_2(x_2^{(k)})^\top; \cdots; \nabla f_n(x_n^{(k)})^\top] \in \mathbb{R}^{n\times d},
\end{aligned}$$

by stacking all local variables. Here, upright bold symbols (*e.g.*, $\mathbf{x}, \mathbf{f}, \mathbf{w}$) are used to denote augmented network-level quantities. We similarly denote $\frac{1}{n}\sum_{i=1}^n x_i^{(k)} \in \mathbb{R}^d$ as $\bar{x}^{(k)}$ and denote $[(\bar{x}^{(k)})^\top; (\bar{x}^{(k)})^\top; \cdots; (\bar{x}^{(k)})^\top] \in \mathbb{R}^{n\times d}$ as $\bar{\mathbf{x}}^{(k)}$.

We use $\mathrm{col}\{a_1,\ldots,a_n\}$ and $\mathrm{diag}\{a_1,\ldots,a_n\}$ to denote a column vector and a diagonal matrix formed from scalars $a_1,\ldots,a_n$. We let $\mathbb{1}_n = \mathrm{col}\{1,\ldots,1\} \in \mathbb{R}^n$ and $I_n \in \mathbb{R}^{n\times n}$ denote the identity matrix. Given two matrices $\mathbf{x},\mathbf{y} \in \mathbb{R}^{n\times d}$, we define inner product $\langle \mathbf{x},\mathbf{y}\rangle = \mathrm{tr}(\mathbf{x}^\top\mathbf{y})$. For a matrix $A$, we let $\|A\|$ denote its $\ell_2$ norm and $\|A\|_F$ denote its Frobenius norm and refer its $j$-th column to $A_{\cdot,j}$. We use $\|\cdot\|_p$ to denote the $\ell_p$ vector norm for $p \geq 1$. For a vector $\pi \in \mathbb{R}^n$ with positive entries, we represent the entry-wise $\alpha$-th power by $\pi^\alpha$, where $\alpha > 0$. The vector/matrix $\pi$-norm [24] is defined as $\|v\|_\pi := \|\mathrm{diag}(\sqrt{\pi})^{-1/2}v\|_2$ for $v \in \mathbb{R}^n$ and $\|A\|_\pi = \|\mathrm{diag}(\sqrt{\pi})^{-1}A\ \mathrm{diag}(\sqrt{\pi})\|_2$ for $A \in \mathbb{R}^{n\times n}$, respectively. We use $\gtrsim$ and $\lesssim$ to indicate inequalities that hold up to absolute constants.

## 2 Effective metrics on directed networks

In decentralized optimization, a network consisting of $n$ computing nodes is associated with a mixing matrix $W = [w_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ where $w_{ij} \in (0,1)$ if node $j$ can send information to node $i$ otherwise $w_{ij} = 0$. Decentralized algorithms are built upon partial averaging $z_i^+ = \sum_{j\in\mathcal{N}_i} w_{ij}z_j$ in which $\mathcal{N}_i$ denotes the direct in-neighbors of node $i$ (including node $i$ itself). Since every node needs to conduct partial averaging simultaneously, we have

$$\mathbf{z} \triangleq [z_1^\top; z_2^\top; \cdots; z_n^\top] \xrightarrow{\text{W-protocol}} \mathbf{z}^+ = W\mathbf{z} = \left[\sum_{j=1}^n w_{1j}z_j^\top; \cdots; \sum_{j=1}^n w_{nj}z_j^\top\right]$$

where $W$-protocol indicates partial averaging with mixing matrix $W$. Evidently, the characteristics of $W$, such as sparsity and connectivity, will substantially impact the efficiency and efficacy of partial averaging and consequently the convergence rate of decentralized algorithms. This section will identify key metrics crucial for understanding the impact of directed topologies on decentralized stochastic algorithms.

### 2.1 Column-stochastic mixing matrix

Throughout this paper, we consider a static directed network $\mathcal{G}$ associated with a column-stochastic matrix $W$:

**Assumption 1** (Column-stochastic mixing matrix)**.** *The mixing matrix $W$ is entry-wisely non-negative, primitive (*i.e.*, all entries of $W^{k_0}$ are positive for sufficiently large $k_0 \in \mathbb{N}_+$), and satisfies $\mathbb{1}_n^\top W = \mathbb{1}_n^\top$.*

Under Assumption 1, the Perron-Frobenius theorem [22] guarantees a unique equilibrium vector $\pi \in \mathbb{R}^n$ with positive entries such that

$$W\pi = \pi \quad \text{and} \quad \mathbb{1}_n^\top\pi = 1.$$

The equilibrium $\pi$ is also referred to as the right Perron vector in literature [22]. A column-stochastic mixing matrix is easy to construct, see the example below where $d_i^{\text{out}}$ is the out-degree of node $i$ excluding the self-loop.

**Example 1.** *Given a network $\mathcal{G}$ with $n$ nodes and a set of directed edges $\mathcal{E}$, one can induce a column-stochastic matrix $W$ by setting weights as*

$$w_{ij} = \begin{cases} 1/(1+d_i^{\text{out}}) & \text{if directed edge } (j,i) \in \mathcal{E} \text{ or } j = i \\ 0 & \text{otherwise} \end{cases}$$

*If $\mathcal{G}$ is strongly connected (*i.e.*, each pair of distinct nodes are connected through directed paths), then $W$ is primitive.*

As discussed in Section 1.4, the vanilla averaging approach (also known as power iterations), $\mathbf{z}^{k+1} = W\mathbf{z}^{(k)}$ for $k \geq 0$ with $\mathbf{z}^{(0)} = \mathbf{z}$, ultimately drives $\mathbf{z}^{(k)}$ to the fixed point $\pi\mathbb{1}_n^\top\mathbf{z}$ which deviates from the desired global average $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\mathbf{z}$. For this reason, the average consensus algorithm cannot be directly utilized in decentralized algorithms over directed networks. A correction mechanism must be introduced into the algorithm to address the mismatch in decentralized consensus.

## 2.2 Push-sum strategy

With the vanilla average consensus update $\mathbf{z}^{k+1} = W\mathbf{z}^{(k)}$ and a column-stochastic $W$, the iterate $\mathbf{z}^{(k)}$ will converge to $\pi\mathbb{1}_n^\top\mathbf{z}$. The mismatch between this limit point $\pi\mathbb{1}_n^\top\mathbf{z}$ and the desired average $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\mathbf{z}$ can be easily eliminated by applying one-shot reweighting with $\text{diag}(n\pi)^{-1}$:

$$\text{diag}(n\pi)^{-1}\mathbf{z}^{(k)} \longrightarrow \text{diag}(n\pi)^{-1}\pi\mathbb{1}_n^\top\mathbf{z} = n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\mathbf{z} = \bar{\mathbf{z}}.$$

Often, the equilibrium vector $\pi$ is not known by the nodes a priori. One can thus conduct auxiliary power iterations $v^{(k+1)} = Wv^{(k)}$, starting from some $v^{(0)} \in \mathbb{R}^n$ with $\mathbb{1}_n^\top v^{(0)} = n$, to asymptotically attain $v^{(k)} \to \pi\mathbb{1}_n^\top v^{(0)} = n\pi$. This leads to the push-sum iterations [14, 15]:

$$\begin{aligned} \mathbf{z}^{(k+1)} &= W\mathbf{z}^{(k)} \\ v^{(k+1)} &= Wv^{(k)} \\ V^{(k+1)} &= \text{diag}(v^{(k+1)}) \\ \mathbf{w}^{(k+1)} &= {V^{(k+1)}}^{-1}\mathbf{z}^{(k+1)}. \end{aligned}$$

The push-sum strategy has become a fundamental pillar, underpinning decentralized algorithms over directed networks. Analyzing the influence of $W$ on push-sum will shed light on its impact on decentralized algorithms.

## 2.3 Spectral gap

To evaluate how $W$ impacts push-sum, we need to identify effective metrics that can capture the characteristics of directed networks. The spectral gap measures the rate at which $\mathbf{z}^{(k)}$ converges to the weighted average $\pi\mathbb{1}_n^\top\mathbf{z}$. It also indicates the rate at which information diffuses across the entire network. However,

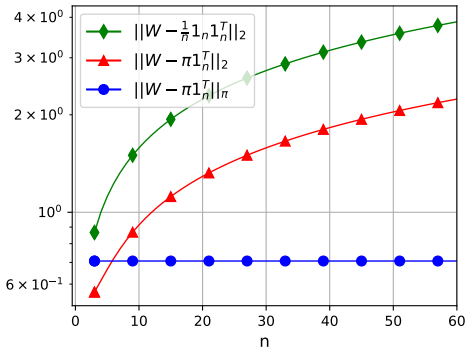


Figure 1: Comparison of the canonical spectral gap and the generalized spectral gap of the column-stochastic instance in Section 5.1.

the canonical spectral gap $1-\beta$ with $\beta := \|W - \mathbb{1}_n\mathbb{1}_n^\top/n\|_2$, inspired by the formulation in (1.2) for undirected networks, becomes ineffective for column-stochastic matrices. This is because $\beta$ can exceed 1 and grow rapidly for large-scale directed networks, as illustrated in Section 2.3.

Fortunately, it is proved in [31] that quantity $\|W - \pi\mathbb{1}_n^\top\|_\pi$ is below 1 for matrix $W$ satisfying Assumption 1, where $\|A\|_\pi := \|\mathrm{diag}(\sqrt{\pi})^{-1} A\ \mathrm{diag}(\sqrt{\pi})\|_2$ for $A \in \mathbb{R}^{n\times n}$. Notably, when $\pi = \mathbb{1}_n/n$ (holds for *e.g.*, doubly-stochastic matrices), $\|A\|_\pi \equiv \|A\|_2$ and thus $\|W - \pi\mathbb{1}_n^\top\|_\pi = \|W - \mathbb{1}_n\mathbb{1}_n^\top/n\|_2$. Therefore, we can naturally view $1 - \|W - \pi\mathbb{1}_n^\top\|_\pi$ as the generalization of the canonical spectral gap $1-\beta$ defined for doubly-stochastic matrices. Formally, we define

**Definition 1** (Generalized Spectral gap [31]). *For matrix $W$ satisfying Assumption 1 associated with its equilibrium vector $\pi$, we define the generalized spectral gap as $1-\beta_\pi$ (abbreviated as spectral gap hereafter) where*

$$\beta_\pi := \|W - \pi\mathbb{1}_n^\top\|_\pi \in [0, 1).$$

As $\beta_\pi$ approaches to 0, the iterate $\mathbf{z}^{(k)}$ will converge faster to $\pi\mathbb{1}_n^\top\mathbf{z}$.

## 2.4 Equilibrium skewness

In addition to the information diffusion rate captured by the spectral gap, the disagreement between $\pi$ and $n^{-1}\mathbb{1}_n$ also imposes additional challenges in characterizing the convergence of the push-sum strategy. As observed in the left plot of Figure 2.4, push-sum using a matrix associated with a small spectral gap $1-\beta_\pi$ can counterintuitively attain faster averaging than one with a large spectral gap. This phenomenon contrasts with the prior findings on doubly stochastic matrices, which would predict faster convergence with a larger spectral gap. This demonstrates that the spectral gap alone is insufficient to fully characterize the convergence performance of push-sum.

To identify additional relevant metrics, we consider an idealized version of push-sum iterations where

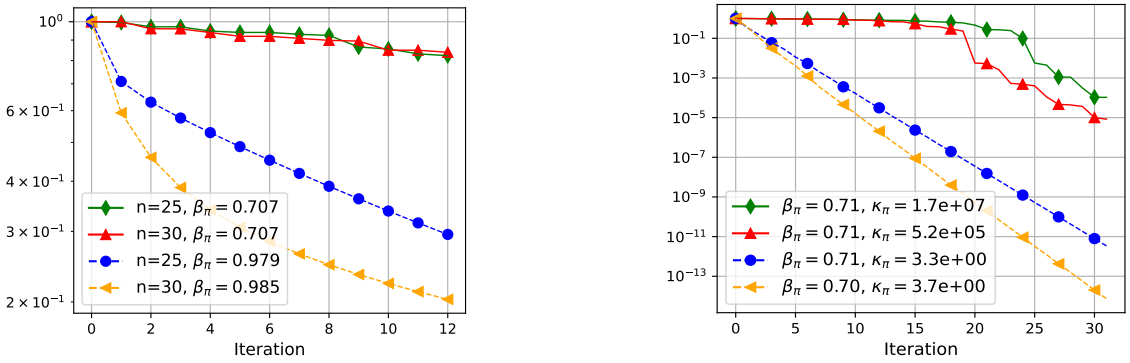


Figure 2: Push-sum iterations with matrices of different spectral gaps and skewness. The y-axis representing the relative error, $\|\mathbf{z}^{(k)}-\bar{\mathbf{z}}\|_F/\|\mathbf{z}^{(0)}-\bar{\mathbf{z}}\|_F$. The left figure illustrates that push-sum with a smaller spectral gap $1-\beta_\pi$ can counter-intuitively achieve faster consensus. The second figure illustrates Push-sum's performance with (nearly) identical $\beta_\pi$ but varying $\kappa_\pi$.

$v^{(0)} = n\pi$ and thus $v^{(k)} = n\pi$, $\forall, k \geq 0$. In this idealized case, push-sum iterations can be summarized as:

$$\mathbf{z} \xmapsto{\text{power iterations}} W^k\mathbf{z} \xmapsto{\text{one-shot reweighting}} \operatorname{diag}(n\pi)^{-1}W^k\mathbf{z} \triangleq \mathbf{w}^{(k)}.$$

With $\mathbf{w}^{(k)} - \bar{\mathbf{z}} = \operatorname{diag}(n\pi)^{-1}(W^k\mathbf{z} - \pi\mathbb{1}_n^\top\mathbf{z}) = \operatorname{diag}(n\pi)^{-1}(W - \pi\mathbb{1}_n^\top)^k\mathbf{z}$, we have

$$\begin{aligned}
\|\mathbf{w}^{(k)} - \bar{\mathbf{z}}\|_F^2 &= \sum_{j=1}^{d} \left\|\operatorname{diag}(n\pi)^{-1}(W - \pi\mathbb{1}_n^\top)^k\mathbf{z}_{\cdot,j}\right\|_2^2 \\
&\leq \left\|\operatorname{diag}(n\pi)^{-1}\operatorname{diag}(\sqrt{\pi})\right\|_2^2 \sum_{j=1}^{d} \|(W - \pi\mathbb{1}_n^\top)^k\mathbf{z}_{\cdot,j}\|_\pi^2 \\
&\leq \frac{\beta_\pi^{2k}\sum_{j=1}^{d}\|\mathbf{z}_{\cdot,j}\|_\pi^2}{n^2\min_i \pi_i} \leq \frac{\max_i \pi_i^2}{\min_i \pi_i^2}\beta_\pi^{2k}\|\mathbf{z}\|_F^2.
\end{aligned}$$

From (2.4), we observe that the consensus error of push-sum iterations stems from both the power iterations (governed by $\beta_\pi^k$) and the reweighting step that corrects the disparity between the $\pi$-norm and $\ell_2$ norm (governed by $\max_i \pi_i / \min_i \pi_i$). The constant $\max_i \pi_i / \min_i \pi_i$ captures the disagreement between $\pi$ and $\mathbb{1}_n/n$, and represents the skewness of the equilibrium vector $\pi$. We thus formally define it as:

**Definition 2** (Equilibrium skewness). *For matrix $W$ satisfying Assumption 1 associated with its equilibrium $\pi$, we define the equilibrium skewness as*

$$\kappa_\pi := \frac{\max_i \pi_i}{\min_i \pi_i} \in [1, +\infty).$$

It is worth noting that the equilibrium skewness does not manifest in the convergence rate of decentralized algorithms over undirected networks because $\pi = n^{-1}\mathbb{1}_n$ and $\kappa_\pi = 1$ therein.

## 2.5 Network influence on push-sum

By jointly considering both the spectral gap and equilibrium skewness, we are able to provide an analysis for push-sum which explicitly reflects the influence of directed networks in Lemma 1. This result also

facilitates our subsequent analysis of decentralized optimization algorithms.

**Lemma 1.** *Under Assumption 1, it holds for push-sum iterations (2.2) with initialization* $\mathbf{z}^{(0)}$ *and* $v^{(0)}$ *such that* $\mathbb{1}_n^\top v^{(0)} = n$,

1. $\min_i\{v_i^{(k)}/\pi_i\}$ *is a non-decreasing sequence with respect to* $k$ *and thus* $\forall k \geq 0$, $\min_i\{v_i^{(k)}/\pi_i\} \geq \min_i\{\pi_i^{-1}\}$.
2. $\max_i\{v_i^{(k)}/\pi_i\}$ *is a non-increasing sequence with respect to* $k$ *and thus* $\forall k \geq 0$, $\max_i\{v_i^{(k)}/\pi_i\} \leq \max_i\{\pi_i{}^{-1}\}$.
3. $\left\|{V^{(k)}}^{-1}\right\|_2 \leq \kappa_\pi$.
4. $\|\mathbf{w}^{(k)} - \bar{\mathbf{z}}^{(0)}\|_F \leq \kappa_\pi^{1.5}\beta_\pi^k\|\mathbf{z}^{(0)}\|_F$.

**Remark 1.** *In essence, the push-sum strategy shows that* ${V^{(k)}}^{-1}W^k\mathbf{z}$ *provides a good estimate of the average* $n^{-1}\mathbb{1}_n\mathbb{1}_n^\top\mathbf{z}$ *when* $k$ *is sufficiently large. The third point in Lemma 1 gives a precise upper bound on the norm of the inverse matrix* ${V^{(k)}}^{-1}$. *This bound is critically important in analyzing decentralized algorithms based on the push-sum strategy. Prior work [16, 20, 32] typically bounds* $\left\|{V^{(k)}}^{-1}\right\|_2$ *using an extremely conservative value,* e.g., $\left\|{V^{(k)}}^{-1}\right\|_2 \leq n^n$. *This loose bound provides limited practical utility in understanding the influence of networks on push-sum.*

## 2.6 Influence of $\kappa_\pi$ can be substantial

As illustrated in Lemma 1, achieving $\|\mathbf{w}^{(k)} - \bar{\mathbf{z}}\|_F = O(\epsilon)$ with push-sum necessitates $O(\ln(\kappa_\pi/\epsilon)/(1-\beta_\pi))$ communication rounds. Given that the influence of the equilibrium skewness $\kappa_\pi$ is only logarithmic in the convergence, one might naturally speculate whether its effect is negligible in comparison to $(1-\beta_\pi)^{-1}$. Surprisingly, we negate the speculation by showing that the skewness $\kappa_\pi$ can be exponentially large with respect to the network size $n$.

**Proposition 2.** *For any* $n \geq 1$, *there exists a column-stochastic matrix* $W \in \mathbb{R}^{n\times n}$ *such that* $\beta_\pi = \dfrac{\sqrt{2}}{2}$ *but* $\kappa_\pi = 2^{n-1}$.

It is revealed in Proposition 2 that the impact of the equilibrium skewness can dominate that of the spectral gap, even after taking the logarithm, since $\ln(\kappa_\pi) = \Omega(n)$ while $(1-\beta_\pi)^{-1} = O(1)$. This highlights the central role of the equilibrium skewness in decentralized optimization over directed networks. These theoretical insights are verified by the simulation results in the right plot of Figure 2.4, where the convergence of push-sum varies significantly for matrices with (nearly) identical spectral gaps but different skewness.

Another important implication in Proposition 2 is that the metrics $\kappa_\pi$ and $1-\beta_\pi$ can be orthogonal, with $1-\beta_\pi$ being a constant while $\kappa_\pi$ grows exponentially. This again necessitates the need to consider both metrics when evaluating the influence of directed networks on decentralized algorithms.

# 3 Lower bound for decentralized optimization

Introducing the equilibrium skewness $\kappa_\pi$ facilitates a sharp characterization of push-sum's convergence. However, it remains unclear if such a notion is necessary for general decentralized optimization algorithms. We show an affirmative answer by establishing a lower bound for smooth and non-convex decentralized stochastic optimization using column-stochastic mixing matrices.

## 3.1 Assumptions

We first state the decentralized setup to which our lower bound applies.

**Function class.** We let the function class $\mathcal{F}_{\Delta,L}$ denote the set of functions satisfying Assumption 2 for any underlying dimension $d \in \mathbb{N}_+$ and initialization $x^{(0)} \in \mathbb{R}^d$.

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

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$$

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

**Gradient oracle class.** We assume each node $i$ has access to its local gradients via a stochastic gradient oracle $\nabla F(x;\xi_i)$ subject to independent randomness $\xi_i$. We further assume that $\nabla F(x;\xi_i)$ is an unbiased estimate of $\nabla f_i(x)$ with a bounded variance. Formally, we let the stochastic gradient oracle class $O_{\sigma^2}$ denote the set of all oracles $\nabla F(\cdot;\xi_i)$ satisfying Assumption 3.

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

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

**Algorithm class.** Suppose a directed network of $n$ nodes and the column-stochastic mixing matrix $W = [w_{ij}]_{i,j=1}^n \in \mathbb{R}^{n\times n}$ are given. We consider algorithms in which each node $i$ accesses an unknown local function $f_i$ via the stochastic gradient oracle $\nabla F(x;\xi_i)$. Each node $i$ will maintain a local model copy $x_i^{(k)}$ at iteration $k$ in algorithms. We assume algorithms to follow the partial averaging policy, *i.e.*, nodes communicate simultaneously via the $W$-protocol as (2), where $\mathbf{z}$ and $W\mathbf{z}$ are the input and output variables of the communication protocol. In addition, we assume $A$ to follow the linear-spanning property [33, 34, 35, 36]. Informally speaking, the linear-spanning property requires that the local model $x_i^{(k)}$ lies the linear space spanned by $x_i^{(0)}$ and its local stochastic gradients or interactions with the neighboring nodes. Furthermore, upon the end of $K$ algorithmic iterations, we allow the ultimate output $x^{(K)}$ to be any variable in $\text{span}(\{\{x_i^{(k)}\}_{i=1}^n\}_{k=0}^K)$. The linear-spanning policy is met by all methods in Table 1 and most first-order optimization methods. We let $\mathcal{A}_W$ be the set of all algorithms following the partial averaging (through matrix $W$) and the linear-spanning property.

Table 2: Comparison of lower bounds in non-convex stochastic decentralized optimization with a certain type of mixing matrix $W$.

| Literature | Lower Bound | Mixing Matrix |
|---|---|---|
| [35, 36] | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}+\frac{L\Delta}{\sqrt{1-\beta}K}$ | Doubly-stochastic Static |
| [37] | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}+\frac{L\Delta}{(1-\beta)K}$ | Doubly-stochastic Time-varying |
| Theorem 1 | $\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}+\frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}$ | Column-stochastic Static |

### 3.2 Lower Bound

With all the interested classes introduced above, we are ready to state our lower bound.

**Theorem 1.** *For any given $L \geq 0$, $n \geq 2$, $\sigma \geq 0$, and $\tilde{\beta} \in [1/\sqrt{2}, 1-1/n]$, 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 column-stochastic matrix $W \in \mathbb{R}^{n\times n}$ with $\beta_\pi = \tilde{\beta}$ and $\ln(\kappa_\pi) = \Omega(n(1-\beta_\pi))$, such that the convergence of any $A \in \mathcal{A}_W$ starting form $x_i^{(0)} = x^{(0)}$, $\forall\, 1 \leq i \leq n$ with $K$ iterations is lower bounded by*

$$\mathbb{E}[\|\nabla f(x^{(K)})\|_2^2] = \Omega\left(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}}+\frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}\right).$$

**Necessity of $\kappa_\pi$ for directed networks.** The lower bound reveals that the equilibrium skewness s $\kappa_\pi$ is a *necessary notion beyond the spectral gap* $1-\beta_\pi$ in measuring the performance of decentralized algorithms over directed networks. Moreover, while $\kappa_\pi$ itself can be exponentially large with respect to the network size $n$ as justified in Proposition 2, the lower bound implies its influence can be possibly mitigated logarithmically to $\ln(\kappa_\pi)$, with proper algorithmic designs, despite that $\ln(\kappa_\pi)$ can still be polynomially large.

**Comparison with other lower bounds.** The lower bound in Theorem 1 is novel compared to the existing lower bounds in decentralized optimization with doubly-stochastic mixing weights as it reflects the impacts of the spectral gap $1-\beta_\pi$ and the equilibrium skewness $\kappa_\pi$ simultaneously. See the comparison in Table 2.

**Linear-speedup.** The first term $\sigma/\sqrt{nK}$ dominates (2) when $K$ is sufficiently large, which will require $K = O(1/(n\epsilon^2))$ iterations, inversely proportional to the number of nodes $n$, to reach accuracy $\epsilon$. Therefore, if the term $\sigma/\sqrt{nK}$ is dominating the rate for some $K$, we say the algorithm is in its linear speedup stage. Linear speedup is a fundamental property that distributed algorithms enjoy.

**Transient time.** Due to the decentralization-incurred overhead in convergence rate, algorithms have to experience a transient stage to achieve its linear-speedup stage. Transient time is thus referred to as the number of iterations when $K$ is relatively small so non-$\sigma/\sqrt{nK}$ terms still dominate the rate. The shorter the transient time is, the fewer iterations the algorithm requires to achieve linear-speedup stage. For example,

if an algorithm can achieve the lower bound established in (1), it requires $(\frac{L\Delta\sigma^2}{nK})^{\frac{1}{2}} \gtrsim \frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}$, *i.e.*, $K \gtrsim n(1+\ln(\kappa_\pi))/(1-\beta_\pi)$ iterations to achieve linear-speedup. Here we mainly care about the factors concerning the network size $n$, the generalized spectral gap $1-\beta$, and the equilibrium skewness $\ln(\kappa_\pi)$ in transient time to evaluate how sensitive the algorithm is to decentralization. Transient time is widely used in decentralized learning and other domains [7, 8, 38, 39].

# 4 Push-DIGing algorithms

In this section, we revisit the seminal Push-DIGing algorithm [20] and give a refined analysis building upon the generalized spectral gap and the equilibrium skewness . We then enhance the performance of Push-DIGing to attain optimal convergence by leveraging multi-round push-sum iterations.

## 4.1 Vanilla Push-DIGing algorithm

The seminal Push-DIGing algorithm incorporates the push-sum module into the backbone of gradient tracking [20] and can be formulated as

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= W(\mathbf{x}^{(k)} - \gamma \mathbf{y}^{(k)}) \\
v^{(k+1)} &= W v^{(k)} \\
V^{(k+1)} &= \mathrm{diag}(v^{(k+1)}) \\
\mathbf{w}^{(k+1)} &= {V^{(k+1)}}^{-1} \mathbf{x}^{(k+1)} \\
\mathbf{y}^{(k+1)} &= W(\mathbf{y}^{(k)} + \nabla F(\mathbf{w}^{(k+1)}; \boldsymbol{\xi}^{(k+1)}) - \nabla F(\mathbf{w}^{(k)}; \boldsymbol{\xi}^{(k)}))
\end{aligned}$$

where $\mathbf{w}^{(0)} = \mathbf{x}^{(0)}$, $\mathbf{y}^{(0)} = \nabla F(\mathbf{x}^{(0)}; \boldsymbol{\xi}^{(0)})$, and $v^{(0)} = \mathbb{1}_n$. The following theorem presents the convergence of Push-DIGing with the spectral gap and equilibrium skewness, which explicitly clarifies the dependence of the network.

**Theorem 2** (Push-DIGing Convergence)**.** *Under Assumptions 1, 2, and 3, by setting the learning rate $\gamma$ as (4), the convergence rate of Push-DIGing satisfies*

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|_2^2] \lesssim \frac{\sigma}{\sqrt{nK}} + \frac{\beta_\pi^{\frac{2}{3}}\kappa_\pi^{\frac{5}{3}}\sigma^{\frac{2}{3}}}{(1-\beta_\pi)K^{\frac{2}{3}}} + \frac{\beta_\pi\kappa_\pi^3(1+\kappa_\pi\beta_\pi)}{(1-\beta_\pi)^2K} + \frac{1}{K}.$$

*where constants $L$, $\Delta$, and $\sum_{i=1}^{n}\|\nabla f(x^{(0)}) - \nabla f_i(x^{(0)})\|^2/n$ are absorbed to highlight the network dependence.*

**Comparison with existing analyses.** Our theorem is the first to present a convergence result for Push-DIGing in the non-convex case. To depict the impact of the network, we introduce two constants, $\kappa_\pi$ and $\beta_\pi$, with explicit meanings. Previous analyses of Push-DIGing [20, 40] have also defined a quantity similar to $1/(1-\beta_\pi)$, but they often overlook the impact of the network by incorporating it into hidden constants.

In contrast, our work introduces the skewness measure $\kappa_\pi$ to explicitly capture the hidden impact of the network. When the weight matrix $W$ is doubly-stochastic so that $\kappa_\pi = 1$, our rate in Theorem 2 almost recovers the convergence of gradient tracking which targets doubly-stochastic matrices.

**Transient time.** To reach linear-speedup rate $\sigma/\sqrt{nK}$, Push-DIGing requires the number of iterations $K$ to satisfy

$$\max\left\{\frac{\beta_\pi^{\frac{2}{3}}\kappa_\pi^{\frac{4}{3}}}{(1-\beta_\pi)K^{\frac{2}{3}}}, \frac{\beta_\pi\kappa_\pi^3(1+\kappa_\pi\beta_\pi)}{K}, \frac{1}{K}\right\} \lesssim \frac{1}{\sqrt{nK}}$$

and hence $K \gtrsim n\kappa_\pi^6/(1-\beta_\pi)^4$. Compared to the order $n(1+\ln(\kappa_\pi))/(1-\beta_\pi)$ induced by Theorem 1, we find a significant margin.

## 4.2 MG-Push-DIGing algorithm

We next present a new algorithm to attain the lower bound established in Section 3, by enhancing the vanilla Push-DIGing. Inspired by the algorithm development in [36, 37], we add two additional components to Push-DIGing: gradient accumulation and multiple-gossip communication. We call the new algorithm as MG-Push-DIGing where "MG" indicates "multiple gossips".

Compared to the original Push-DIGing (4.1), we enhance the communication efficiency by replacing every $W$ with $W^R$. Let $K$ be the total budget for gradient evaluations and decentralized communications at each node. Specifically, each node takes $2R$ rounds of communication at $k$-th stage instead of two rounds of communication in vanilla Push-DIGing. Thus, we reduce the variances of stochastic gradients by enlarging the sampling batch by $R$ times. Consequently, MG-Push-DIGing costs $T = KR$ gradient queries and $2T$ rounds of communication and when it runs $K$ stages. The following theorems clarify the convergence rate of MG-Push-DIGing.

**Theorem 3** (MG-Push-DIGing convergence)**.** *Under Assumptions 1, 2, and 3, by setting* $R = \dfrac{(1+\sqrt{\ln(\kappa_\pi)})^2}{1-\beta_\pi}$, $T = KR$, *and* $\gamma$ *as* (4)*, the convergence of MG-Push-DIGing satisfies*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|_2^2] = \widetilde{\mathcal{O}}\left(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nT}} + \frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)T}\right),$$

*where* $\widetilde{\mathcal{O}}(\cdot)$ *absorbs logarithmic factors independent of* $\kappa_\pi$ *and* $\beta_\pi$*.*

Notably, while the computation/communication cost of each stage of MG-Push-DIGing is $R$ times more than the one of a vanilla Push-DIGing iteration, it is expected to only run $K/R$ stages, as opposed to $K$ iterations. This ensures a fair comparison of the eventual convergence rates. The rate (3) matches with the lower bound (1) up to logarithm factors independent of $\kappa_\pi$ and $\beta_\pi$. This reveals of the optimality of MG-Push-DIGing and the tightness of the lower bound (1). The comparison between MG-Push-DIGing with other state-of-the-art algorithms is listed in Table 1.

**Algorithm 1** MG-Push-DIGing: Push-DIGing with multi-round gossip

**Require:** Initialize $v^{(0,0)} = \mathbb{1}_n$, $\mathbf{w}^{(0)} = \mathbf{x}^{(0)}$, $g_i^{(0)} = y_i^{(0)} = \frac{1}{R}\sum_{r=1}^{R} \nabla F(x^{(0)}; \xi_i^{(0,r)})$, the mixing matrix $W$, the multi-round number $R$.

**for** $s = 0, 1, \ldots, S-1$, each node $i$ **do**

Update $\phi^{(s+1,0)} = x_i^{(s)} - \gamma y_i^{(s)}$;

**for** $r = 0, 1, \ldots, R-1$, each node $i$ **do**

Update $\phi_i^{(s+1,r+1)} = \sum_{j \in \mathcal{N}_i^{\text{in}}} w_{ij} \phi_j^{(s+1,r)}$;

Update $v_i^{(s,r+1)} = \sum_{j \in \mathcal{N}_i^{\text{in}}} w_{ij} v_j^{(s,r)}$;

**end for**

Update $x_i^{(s+1)} = \phi_i^{(s+1,R)}$;

Update $v_i^{(s+1,0)} = v_i^{(s,R)}$ and $w_i^{(s+1)} = \phi_i^{(s+1,0)} / v_i^{(s+1,0)}$;

Evaluate new gradient $g_i^{(s+1)} = \frac{1}{R}\sum_{r=1}^{R} \nabla F(w_i^{(s+1)}; \xi_i^{(s+1,r)})$;

Let $\psi_i^{(s+1,0)} = y_i^{(s)} + g_i^{(s+1)} - g_i^{(s)}$;

**for** $r = 0, 1, \ldots, R-1$, each node $i$ **do**

Update $\psi_i^{(s+1,r+1)} = \sum_{j \in \mathcal{N}_i^{\text{in}}} w_{ij} \psi_j^{(s+1,r)}$;

**end for**

Update $y_i^{(s+1)} = \psi^{(s+1,R)}$;

**end for**

# 5 Experiments

In this section, we conduct a series of experiments that verify our theoretical findings.

## 5.1 Examples of highly skewed network

In this subsection, we present a column-stochastic matrix whose skewness $\kappa_\pi$ is large while the spectral gap $1 - \beta_\pi$ is small. We consider the following column-stochastic matrix

$$W = \begin{bmatrix} 1/2 & 1/2 & \cdots & 1/2 & 1 \\ 1/2 & 0 & & & \\ & \ddots & \ddots & & \\ & & 1/2 & 0 & \\ & & & 1/2 & 0 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

The right Perron vector of $W$ is $\pi \propto \left(2^{n-1}, 2^{n-2}, \ldots, 1\right)^\top$, resulting in $\kappa_\pi = 2^{n-1}$. Moreover, it can be shown that $\beta_\pi \equiv 1/\sqrt{2}$ for $W$, see Appendix D for the proof. The underlying network associated with $W$ requires

$n-1$ steps to transmit information from node 1 to node $n$, see the left plot of Figure 3 for the illustration of $n = 7$. For such a matrix, its equilibrium skewness $\kappa_\pi$ (even taking the logarithm $\ln(\kappa_\pi)$) is significantly larger than $(1-\beta_\pi)^{-1}$, see the right plot of Figure 3.

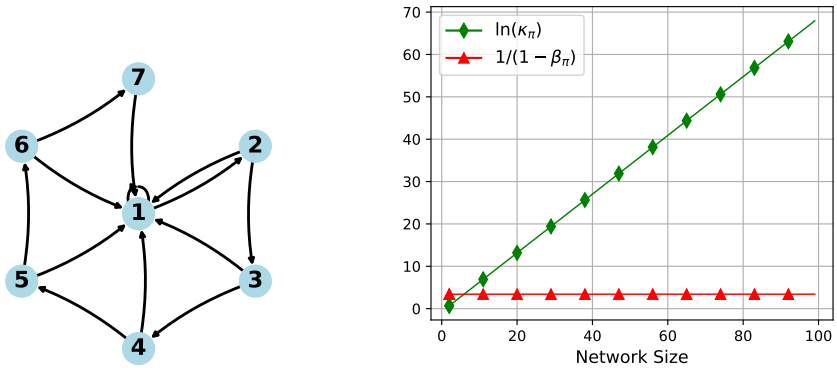


Figure 3: (Left.) The network topology associated with (5.1) when $n = 7$. (Right.) $\kappa_\pi$ increases exponentially with the network size $n$ while the spectral gap remains constant for the mixing matrix (5.1).

## 5.2 Influence of network on Push-DIGing

In this subsection, we examine the performance of Push-DIGing. Our numerical experiments are based on the distributed logistic regression problem with a non-convex regularization term [41, 42, 10] over a directed graph. The problem formulation is given by $\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(f_i(x)+\rho r(x))$, where

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

where $[x]_j$ denotes the $j$-th entry of $x$, $\{h_{i,l}, y_{i,l}\}_{l=1}^{L}$ is the training dataset held by node $i$ in which $h_{i,l}\in\mathbb{R}^d$ is a feature vector while $y_{i,l}\in\{+1,-1\}$ is the corresponding label. The regularization $r(x)$ is non-convex and the parameter $\rho > 0$ controls the regularization impact.

### 5.2.1 Experiment Settings

We set $d = 10, L = 2000$ and $\rho = 0.001$. To manage the heterogeneity of data across nodes, each node $i$ is associated with a local solution $x_i^*$. Such $x_i^*$ is generated by $x_i^* = x^* + v_i$ where the shared $x^*\sim\mathcal{N}(0, I_d)$ is randomly generated while $v_i\sim\mathcal{N}(0,\sigma_h^2 I_d)$ controls the similarity between each local solutions. Intuitively, a higher value of $\sigma_h^2$ leads to more diverse local solutions. With $x_i^*$ at hand, we can generate local data that follows distinct distributions. At node $i$, we generate each feature vector $h_{i,l}\sim\mathcal{N}(0, I_d)$. To produce the corresponding label $y_{i,l}$, we generate a random variable $z_{i,l}\sim\mathcal{U}(0,1)$. If $z_{i,l} < 1/(1+\exp(-y_{i,l}h_{i,l}^\top x_i^*))$, we set $y_{i,l} = 1$. Otherwise we set $y_{i,l} = -1$. Clearly, solution $x_i^*$ controls the distribution of the labels. In this way, we can easily control data heterogeneity by adjusting $\sigma_h^2$. To facilitate the management of gradient noise, we introduce Gaussian noise to the actual gradient to achieve stochastic gradients. That is $\nabla f_i(x) = \nabla f(x) + \varepsilon_i$,

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2 I_d)$. We can control the magnitude of the gradient noise by adjusting $\sigma_n^2$. Throughout the experiments, we set $\sigma_h = 1$, $\sigma_n = 0.001$. The metric for all simulations is $\|\mathbf{y}^{(k)}\|_F$, where $\mathbf{y}^{(k)}$ is the gradient tracking term defined in Push-DIGing algorithm.

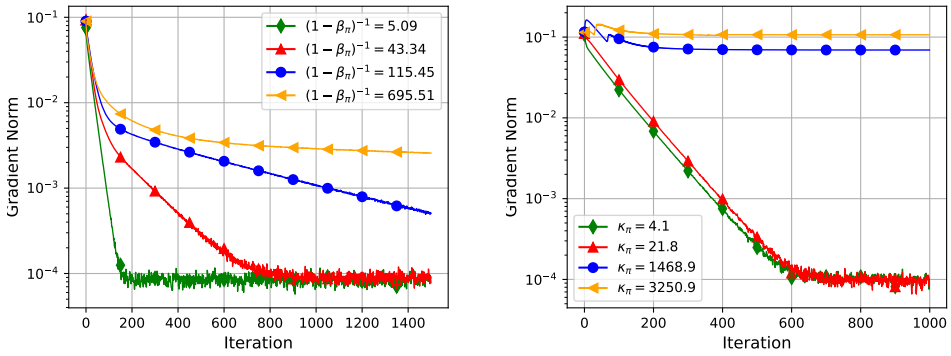


Figure 4: The performance of Push-DIGing under different setups. (Left.) $1 - \beta_\pi$ varies while $\kappa_\pi \approx 50$ is fixed. (Right.) $\kappa_\pi$ varies while $\beta_\pi \approx 0.8$ is fixed.

#### 5.2.2 Convergence with varying $\beta_\pi$

To explore how $1 - \beta_\pi$ affects convergence, we create multiple column-stochastic matrices with the same topology as in Figure 3. We carefully choose these matrices to have the equilibrium skewness $\kappa_\pi$ in $(50, 51)$, while $1/(1 - \beta_\pi)$ varies from 5.0 to 695.5. Notably, the equilibrium vector $\pi$ can be different across these matrices. By running Push-DIGing with these mixing matrices, we gain direct insight into how the spectral gap $1 - \beta_\pi$ influences convergence rates. The results are depicted in the left plot of Figure 4, which shows that the convergence slows down as $\beta_\pi$ gets closer to 1, *i.e.*, the spectral gap gets larger.

#### 5.2.3 Convergence with varying $\kappa_\pi$

Likewise, we generate multiple column-stochastic matrices with the same topology shown in Figure 3 with nearly the same inverse spectral gap $(1 - \beta_\pi)^{-1}$ in $(5, 5.1)$. In contrast, their skewness $\kappa_\pi$ varies from 4.1 to 3250.9. By running Push-DIGing with these mixing matrices, we observe that $\kappa_\pi$ affects convergence rates significantly. The results are depicted in the right plot of Figure 4, which demonstrates that the convergence slows down as $\kappa_\pi$ increases.

### 5.3 Comparison of MG-Push-DIGing and Push-DIGing

Here, we numerically justify the benefit of multi-round gossip by comparing MG-Push-DIGing and the vanilla Push-DIGing. According to Theorem 3, the MG technique effectively mitigates the influences of both $1 - \beta$ and $\kappa_\pi$. To illustrate this, we devise two sets of experiments. In the first one, depicted in the left plot of Figure 5, we generate a matrix $W_1$ from the topology shown in Figure 3, exhibiting a noteworthy

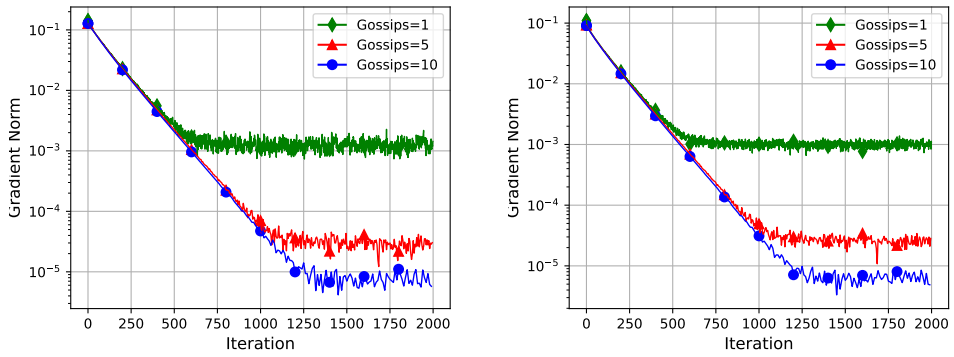


Figure 5: MG-Push-DIGing. For $W_1$ used in the left plot, $\kappa_\pi = 4804.49$, $(1-\beta_\pi)^{-1} = 2.34$. For $W_2$ used in the right plot, $\kappa_\pi = 6.33$, $(1-\beta_\pi)^{-1} = 51.24$.

value of skewness $\kappa_\pi$ but negligible $(1-\beta_\pi)^{-1}$. In the second one, depicted in the right plot of Figure 5, we pick a matrix $W_2$ similarly but with a significant $(1-\beta_\pi)^{-1}$ and a mild value of skewness $\kappa_\pi$. It is worth noting that, as per our analysis in Appendix C, when employing $R$ rounds of gossip, the learning rate is supposed to amplify by $R$ times in a feasible regime. Therefore, we consistently set the learning rate of Push-DIGing to 0.01, while setting the learning rate of $R$-gossip Push-DIGing to $0.01R$. Our numerical results corroborate the conclusion of Theorem 3, showing that MG-Push-DIGing alleviates the dependence on network properties in terms of both $1-\beta_\pi$ and $\kappa_\pi$.

# 6 Conclusion

This paper investigates smooth and non-convex stochastic decentralized optimization over directed networks using column-stochastic mixing matrices. We introduce a novel metric, named equilibrium skewness $\kappa_\pi$, to facilitate explicit characterization of algorithmic convergence and the network dependence, along with the generalized spectral gap. We prove a tight lower bound for general decentralized algorithms using column-stochastic matrices that explicitly manifest the influence of the equilibrium skewness and spectral gap. We present the state-of-the-art convergence rates of the Push-DIGing algorithm. We also combine the technique of multi-round gossip with Push-DIGing to obtain MG-Push-DIGing that nearly attains the established lower bound. Experiments conducted verify our theoretical findings.

# References

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

[2] 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.

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

[4] 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*, pages 3029–3039, 2021.

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

[6] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[7] Shi Pu, Alex Olshevsky, and Ioannis Ch Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 67(11):5900–5915, 2021.

[8] Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34:13975–13987, 2021.

[9] Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip sgd with periodic global averaging. In *International Conference on Machine Learning*, pages 1791–1802. PMLR, 2021.

[10] 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.

[11] Kun Yuan, Sulaiman A Alghunaim, and Xinmeng Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *Journal of Machine Learning Research*, 24(280):1–53, 2023.

[12] Kun Huang and Shi Pu. Improving the transient times for distributed stochastic gradient methods. *IEEE Transactions on Automatic Control*, 2022.

[13] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. $D^2$: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018.

[14] 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.*, pages 482–491. IEEE, 2003.

[15] 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)*, pages 5453–5458. IEEE, 2012.

[16] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] Oskar Perron. Zur theorie der matrices. *Mathematische Annalen*, 64(2):248–263, 1907.

[23] Vyacheslav Kungurtsev, Mahdi Morafah, Tara Javidi, and Gesualdo Scutari. Decentralized asynchronous non-convex stochastic optimization on directed graphs. *IEEE Transactions on Control of Network Systems*, 2023.

[24] 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)*, pages 8353–8358. IEEE, 2019.

[25] Ran Xin and Usman A Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

[26] Shi Pu, Wei Shi, Jinming Xu, and Angelia Nedić. Push–pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 66(1):1–16, 2020.

[27] Zhuoqing Song, Lei Shi, Shi Pu, and Ming Yan. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, pages 1–53, 2023.

[28] 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, 2019.

[29] 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.

[30] Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H Sayed. Exact diffusion for distributed optimization and learning—part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2018.

[31] 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)*, pages 8353–8358. IEEE, 2019.

[32] Songtao Lu and Chai Wah Wu. Decentralized stochastic non-convex optimization over weakly connected time-varying digraphs. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5770–5774, 2020.

[33] 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.

[34] 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.

[35] 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.

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

[37] Xinmeng Huang and Kun Yuan. Optimal complexity in non-convex decentralized learning over time-varying networks. *arXiv preprint arXiv:2211.00533*, 2022.

[38] 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.

[39] Xinmeng Huang, Ping Li, and Xiaoyun Li. Stochastic controlled averaging for federated learning with communication compression. *arXiv preprint arXiv:2308.08165*, 2023.

[40] Hongzhe Liu, Wenwu Yu, Wei Xing Zheng, Angelia Nedić, and Yanan Zhu. Distributed constrained optimization algorithms with linear convergence rate over time-varying unbalanced graphs. *Automatica*, 159:111346, 2024.

[41] Anestis Antoniadis, Irène Gijbels, and Mila Nikolova. Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63:585–615, 2011.

[42] 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.

[43] 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.

# A Proof of Lemma 1

*Proof.* We prove the first statement. Let $a^{(k)} \triangleq \min_i\{v_i^{(k+1)}/\pi_i\}$. Notice that

$$a^{(k+1)} = \min_i \left\{ \frac{\sum_{j=1}^n w_{ij} v_j^{(k)}}{\pi_i} \right\} = \min_i \left\{ \frac{1}{\pi_i} \sum_{j=1}^n w_{ij}\pi_j \frac{v_j^{(k)}}{\pi_j} \right\}$$
$$\overset{(a)}{\geq} \min_i \left\{ \frac{1}{\pi_i} \left( \sum_{j=1}^n w_{ij}\pi_j \right) a^{(k)} \right\} \overset{(b)}{=} a^{(k)}$$

where (a) holds because both $v_j^{(k)} > 0$ and $\pi_j > 0$ for any $1 \leq j \leq n$ and $k \geq 0$, and (b) holds because of Assumption 1. The above inequality implies that $\min_i\{\frac{v_i^{(k)}}{\pi_i}\}$ is a non-decreasing sequence over $k$ and hence

$$\min_i \left\{ \frac{v_i^{(k)}}{\pi_i} \right\} \geq \min_i \left\{ \frac{v_i^{(0)}}{\pi_i} \right\} = \min_i \left\{ \frac{1}{\pi_i} \right\}.$$

The second statement follows the same argument and is omitted here. To prove the third statement, note that $\forall\, k = 0, 1, \cdots,$

$$\min_i\{v_i^{(k)}\} = \min_i \left\{ \frac{v_i^{(k)}\pi_i}{\pi_i} \right\} \geq a^{(k)} \min_i \pi_i \overset{(c)}{\geq} \min_i \pi_i^{-1} \min_i \pi_i = \frac{\min_i \pi_i}{\max_i \pi_i},$$

where inequality (c) holds because of the first statement. With the above relation, we have

$$\|{V^{(k)}}^{-1}\|_2 = \frac{1}{\min_i\{v_i^{(k)}\}} \leq \frac{\max_i \pi_i}{\min_i \pi_i} = \kappa_\pi.$$

For the fourth statement, let $c_\pi^2 = \max_i \pi_i, d_\pi^2 = \min_i \pi_i$, notice that $\bar{x}^{(k)} = \bar{x}^{(0)} = \frac{1}{n}\mathbb{1}_n^\top x^{(0)}$, we have

$$\begin{aligned}
\|\mathbf{w}^{(k)} - \bar{\mathbf{z}}^{(k)}\|_F^2 &= \sum_{j=1}^d \left\| {V^{(k)}}^{-1} (W^k - \frac{1}{n} v^{(k)} \mathbb{1}_n^\top) \mathbf{z}_{\cdot,j} \right\|_2^2 \\
&= \sum_{j=1}^d \left\| \mathrm{diag}(\pi)^{\frac{1}{2}} {V^{(k)}}^{-1} (W^k - \frac{1}{n} v^{(k)} \mathbb{1}_n^\top) \mathbf{z}_{\cdot,j} \right\|_\pi^2 \\
&\leq \left\| \mathrm{diag}(\pi)^{\frac{1}{2}} {V^{(k)}}^{-1} \right\|_\pi^2 \sum_{j=1}^d \left\| (W^k - \frac{1}{n} v^{(k)} \mathbb{1}_n^\top) \mathbf{z}_{\cdot,j} \right\|_\pi^2 \\
&\overset{d}{\leq} \kappa_\pi c_\pi^2 \left( \sum_{j=1}^d \|(W - \pi\mathbb{1}_n^\top)^k \mathbf{z}_{\cdot,j}\|_\pi^2 + \sum_{j=1}^d \left\| \left( \pi\mathbb{1}_n^\top - \frac{1}{n} v^{(k)} \mathbb{1}_n^\top \right) \mathbf{z}_{\cdot,j} \right\|_\pi^2 \right) \\
&\leq \kappa_\pi c_\pi^2 \beta_\pi^{2k} \left( \|\mathbf{z}\|_\pi^2 + \left\| \pi\mathbb{1}_n^\top - \frac{1}{n} v^{(0)} \mathbb{1}_n^\top \right\|_\pi^2 \|\mathbf{z}\|_\pi^2 \right) \\
&\overset{(e)}{\leq} \kappa_\pi \frac{c_\pi^2}{d_\pi^2} \beta_\pi^{2k} (1+\delta^2) \|\mathbf{z}\|_F^2 = \kappa_\pi^2 \beta_\pi^{2k} (1+\delta^2) \|\mathbf{z}\|_F^2
\end{aligned}$$

where (d) holds because

$$\min_i \left\{ \frac{v_i^{(k)}}{\sqrt{\pi_i}} \right\} \geq \min_i \sqrt{v_i^{(k)}} \min_i \left\{ \sqrt{\frac{v_i^{(k)}}{\pi_i}} \right\} \geq \frac{1}{\sqrt{\kappa_\pi}} \sqrt{\min_i \left\{ \frac{1}{\pi_i} \right\}} = \frac{1}{c_\pi \sqrt{\kappa_\pi}},$$

thus $\left\|\text{diag}(\pi)^{\frac{1}{2}}{V^{(k)}}^{-1}\right\|_\pi^2 = (1/\min_i\left\{v_i^{(k)}/\sqrt{\pi_i}\right\})^2 \leq \kappa_\pi c_\pi^2$. (e) holds because we define $\delta := \|\pi\mathbb{1}_n^\top - \frac{1}{n}v^{(0)}\mathbb{1}_n^\top\|_\pi$ and

$$\begin{aligned}\left\|\pi\mathbb{1}_n^\top - \frac{1}{n}v^{(k)}\mathbb{1}_n^\top\right\|_\pi &= \left\|(W-\pi\mathbb{1}_n^\top)^k\left(\pi\mathbb{1}_n^\top - \frac{1}{n}v^{(0)}\mathbb{1}_n^\top\right)\right\|_\pi \\ &\leq \|(W-\pi\mathbb{1}_n^\top)\|_\pi^k\left\|\pi\mathbb{1}_n^\top - \frac{1}{n}v^{(0)}\mathbb{1}_n^\top\right\|_\pi = \beta_\pi^k\delta.\end{aligned}$$

We can further upper bound $\delta$ with $n$ and $\kappa_\pi$ as

$$\begin{aligned}\delta &= \left\|\pi\mathbb{1}_n^\top - \frac{1}{n}v^{(0)}\mathbb{1}_n^\top\right\|_\pi = \left\|\sqrt{\pi}\sqrt{\pi}^\top - \frac{1}{n}\Pi^{-1/2}v^{(0)}\sqrt{\pi}^\top\right\|_2 \\ &\leq \sqrt{\sum_{i=1}^n \frac{\left(v_i^{(0)}\right)^2}{n^2\pi_i} - 1} \leq \sqrt{\max_i\left\{\frac{v_i^{(0)}}{n\pi_i}\right\}\sum_{i=0}^n\frac{v_i^{(0)}}{n} - 1} \\ &= \sqrt{\max_i\left\{\frac{v_i^{(0)}}{n\pi_i}\right\} - 1} \leq \sqrt{\kappa_\pi - 1}.\end{aligned}$$

□

# B Proof of Section 5

To avoid heavy notations, we use the following simplified notations.

$\mathbf{g}^{(k)} = \nabla F({V^{(k)}}^{-1}\mathbf{x}^{(k)};\xi^{(k+1)}) \in \mathbb{R}^{n\times p}$. The bar symbol above the letters indicates the average vector. $\bar{x}^k = \frac{1}{n}\mathbb{1}_n^\top\mathbf{x}^{(k)}$, $\bar{y}^{(k)} = \frac{1}{n}\mathbb{1}_n^\top\mathbf{y}^{(k)}$, $\bar{g}^{(k)} = \frac{1}{n}\mathbb{1}_n^\top\mathbf{g}^{(k)}$, $\overline{\nabla f}^{(k)} = \frac{1}{n}\mathbb{1}_n^\top\nabla f(\mathbf{w}^{(k)})$. We also define the averaging matrix $R = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^\top$ and the weighted averaging matrix $J^{(k)} = \frac{1}{n}v^{(k)}\mathbb{1}_n^\top$. Let $C_k = \sum_{i=0}^k(\beta_\pi^i + \delta\beta_\pi^k)^2$.

We start by presenting standard results in gradient tracking algorithms [42].

**Lemma B.1.** *Under Assumption 1,Assumption 3, we have the following statement:*

1. $J^{(k)}W = J^{(k)}, WJ^{(k)} = J^{(k+1)}, (W-J^{(k+1)})J^{(k)} = 0, J^{(l)}(W-J^{(k)}) = 0, \forall k,l \geq 0.$
2. $\bar{y}^{(k+1)} = \bar{g}^{(k)}, \forall k \geq 0.$
3. $W\mathbf{x} - J^{(k+1)}\mathbf{x} = (W-J^{(k)})(\mathbf{x}-J^{(k)}\mathbf{x}), \forall\mathbf{x}\in\mathbb{R}^{n\times p}, \forall k\geq 0$
4. $\mathbf{w}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)} = ({V^{(k+1)}}^{-1}WV^{(k)} - RV^{(k)})(\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}) - \gamma({V^{(k+1)}}^{-1}W - R)(\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)})$
5. $\|\mathbf{w}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_\pi \leq \kappa_\pi\beta(1+\delta)\|\mathbf{x}^{(k)} - J^{(k)}\mathbf{x}^{(k)}\|_\pi + \kappa_\pi\beta\gamma(1+\delta)\|\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)}\|_\pi$

The first four statements can be verified directly. We give a proof to the last statement of Lemma B.1.

*Proof.* According to lemma Lemma B.1(d), we have

$$\begin{aligned}\mathbf{w}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)} &= ({V^{(k+1)}}^{-1} W - R)\mathbf{x}^{(k+1)} \\ =&({V^{(k+1)}}^{-1} W V^{(k)} - R V^{(k)})(\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}) \\ &- \gamma({V^{(k+1)}}^{-1} W - R)(\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)})\end{aligned}$$

Thus, we can apply triangle inequality to obtain:

$$\begin{aligned}\|\mathbf{w}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|_\pi \leq& \|{V^{(k+1)}}^{-1}\|_\pi \|W - J^{(k+1)}\|_\pi \|V^{(k)}\|_\pi \|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_\pi \\ &+ \gamma \|{V^{(k+1)}}^{-1}\|_\pi \|W - J^{(k+1)}\|_\pi \|\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)}\|_\pi \\ \leq& \kappa_\pi \beta (1+\delta)\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_\pi + \gamma \kappa_\pi \beta (1+\delta)\|\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)}\|_\pi\end{aligned}$$

□

The following Lemma is a variation of Lemma 4 in [42].

**Lemma B.2.**

$$\begin{aligned}\|\mathbf{y}^{(k+1)} - J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2 &= \sum_{i=0}^{k} \|(W^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\|_\pi^2 \\ + 2\sum_{i=0}^{k} \langle (W^{k+1-i} - J^{(k+1)})(\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}), &(W^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\rangle_\pi\end{aligned}$$

*Proof.* Note that $(W - J^{(k+1)})J^{(k)} = 0$, we have the following transformation:

$$\begin{aligned}&\|\mathbf{y}^{(k+1)} - J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2 \\ =&\|W(\mathbf{y}^{(k+1)} + \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)}) - J^{(k+1)}(\mathbf{y}^{(k+1)} + \mathbf{g}^{(k+1)} - \mathbf{g}^{(k)})\|_\pi^2 \\ =&\|(W - J^{(k+1)})(\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)})\|_\pi^2 + \|(W - J^{(k+1)})(\mathbf{g}^{(k+1)} - \mathbf{g}^{(k)})\|_\pi^2 \\ &+ 2\langle (W - J^{(k+1)})(\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)}), (W - J^{(k+1)})(\mathbf{g}^{(k+1)} - \mathbf{g}^{(k)})\rangle_\pi\end{aligned}$$

We can continue to decompose the first term in the last equation of (B) with the same manner, :

$$\begin{aligned}&\|(W - J^{(k+1)})(\mathbf{y}^{(k)} - J^{(k)}\mathbf{y}^{(k)})\|_\pi^2 \\ =&\|(W^2 - J^{(k+1)})(\mathbf{y}^{(k-1)} - J^{(k-1)}\mathbf{y}^{(k-1)})\|_\pi^2 + \|(W^2 - J^{(k+1)})(\mathbf{g}^{(k)} - \mathbf{g}^{(k-1)})\|_\pi^2 \\ &+ 2\langle (W^2 - J^{(k+1)})(\mathbf{y}^{(k-1)} - J^{(k-1)}\mathbf{y}^{(k-1)}), (W^2 - J^{(k+1)})(\mathbf{g}^{(k)} - \mathbf{g}^{(k-1)})\rangle_\pi\end{aligned}$$

where we use $(W - J^{(k+1)})(W - J^{(k)}) = W^2 - J^{(k+1)}$. By repeating this decomposition, we obtain the lemma. Next we proceed by bounding each term appeared in the decomposition. □

**Lemma B.3.** *The first term in Lemma B.2 can be bounded as follow:*

$$\begin{aligned}&\mathbb{E}\|(W^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\|_\pi^2 \\ \leq& \beta_\pi^{2(k+1-i)}(1+\delta)^2 \left(3nc_\pi^2\sigma^2 + 12L^2\kappa_\pi^2\beta_\pi^2\gamma^2\mathbb{E}\|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\right) \\ &+ \beta_\pi^{2(k+1-i)}(1+\delta)^2 \left(6L^2\gamma^2\mathbb{E}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + C_2\mathbb{E}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2\right)\end{aligned}$$

*Proof.*

$$\mathbb{E}\|(W^{k+1-i}-J^{(k+1)})(\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)})\|_\pi^2 \le \beta_\pi^{2(k+1-i)}(1+\delta)^2\mathbb{E}\|\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)}\|_\pi^2, \forall i\ge 0.$$

Since both $\nabla f^{(i+1)}$ and $\mathbf{g}^{(i)}$ are $\mathcal{F}^{(i+1)}$-measurable and $\mathbb{E}[\mathbf{g}^{(i+1)}|\mathcal{F}^{(i+1)}]=\nabla f^{(i+1)}$, we have:

$$\begin{aligned}
&\mathbb{E}\|\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)}\|_\pi^2 \le \mathbb{E}\|\mathbf{g}^{(i+1)}-\nabla f^{(i+1)}\|_\pi^2+\mathbb{E}\|\nabla f^{(i+1)}-\mathbf{g}^{(i)}\|_\pi^2\\
\le& nc_\pi^2\sigma^2+\mathbb{E}\|\nabla f^{(i+1)}-\mathbf{g}^{(i)}\|_\pi^2\\
\le& nc_\pi^2\sigma^2+2\mathbb{E}\|\nabla f^{(i+1)}-\nabla f^{(i)}\|_\pi^2+2\mathbb{E}\|\nabla f^{(i)}-\mathbf{g}^{(i)}\|_\pi^2\\
\le& 3nc_\pi^2\sigma^2+2L^2\mathbb{E}\|\mathbf{w}^{(i+1)}-\mathbf{w}^{(i)}\|_\pi^2,
\end{aligned}$$

where the first inequality uses Assumption 3 and the last inequality uses Assumption 2.

$$\begin{aligned}
&\|\mathbf{w}^{(i+1)}-\mathbf{w}^{(i)}\|_\pi^2 \le 3\|\mathbf{w}^{(i+1)}-\bar{\mathbf{x}}^{(i+1)}\|_\pi^2+3\|\bar{\mathbf{x}}^{(i+1)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2+3\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2\\
=&3\|\mathbf{w}^{(i+1)}-\bar{\mathbf{x}}^{(i+1)}\|_\pi^2+3\gamma^2\|\bar{\mathbf{g}}^{(i)}\|_\pi^2+3\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}(i)\|_\pi^2\\
\le& 6\kappa_\pi^2\beta_\pi^2\gamma^2\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2+3\gamma^2\|\bar{\mathbf{g}}^{(i)}\|_\pi^2+(3+6\kappa_\pi^2\beta_\pi^2(1+\delta)^2)\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2,
\end{aligned}$$

where the second inequality uses Jensen inequality and the last inequality uses Lemma B.1(4). The proof follows by using (B) in (B). □

**Lemma B.4.** *The second term in Lemma B.2 can be bounded as follow:*

$$\begin{aligned}
&\mathbb{E}\langle(W^{k+1-i}-J^{(k+1)})(\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}),(W^{k+1-i}-J^{(k+1)})(\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)})\rangle_\pi\\
\le&\sigma^2\beta_\pi^{k+1-i}(1+\delta)+\beta_\pi^{2(k-i)}\left(\frac{1-\beta}{6}+L\kappa_\pi\beta\gamma(1+\delta)^3\right)\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\\
&+\beta_\pi^{2(k-i)}((1+\delta)^4\frac{3\kappa_\pi^2\beta_\pi^2L^2\gamma^2}{1-\beta}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2+\frac{3}{1-\beta}L^2(1+\kappa_\pi\beta(1+\delta))^2\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi)
\end{aligned}$$

*Proof.* Notice that $\mathbb{E}[\mathbf{g}^{(i+1)}-\mathbf{g}^{(i)}|\mathcal{F}^{(i)}]=\mathbb{E}[(\nabla f^{(i+1)}-\nabla f^{(i)})+(\nabla f^{(i)}-\mathbf{g}^{(i)})|\mathcal{F}^{(i)}]$, the term in the left hand of (B.4) can be decomposed to two terms of inner product. For the first term $D_1$, we can use Cauchy-Schwarz inequality: $\forall k\ge 0$,

$$\begin{aligned}
&\langle(W^{k+1-i}-J^{(k+1)})(\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}),(W^{k+1-i}-J^{(k+1)})(\nabla f^{(i+1)}-\nabla f^{(i)})\rangle_\pi\\
\le&\beta_\pi^{2(k+1-i)}(1+\delta)^2L\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi\|\mathbf{w}^{(i+1)}-\mathbf{w}^{(i)}\|_\pi
\end{aligned}$$

where we use $\|W^{k+1-i}-J^{(k+1)}\|_\pi\le\beta_\pi^{k+1-i}(1+\delta)$ and Assumption 3. Note that, $\forall k\ge 0$,

$$\begin{aligned}
&\|\mathbf{w}^{(i+1)}-\mathbf{w}^{(i)}\|_\pi \le \|\mathbf{w}^{(i+1)}-\bar{\mathbf{x}}^{(i+1)}\|_\pi+\|\bar{\mathbf{x}}^{(i+1)}-\bar{\mathbf{x}}^{(i)}\|_\pi+\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi\\
\le&\gamma\|\bar{\mathbf{g}}^{(i)}\|_\pi+(1+\kappa_\pi\beta(1+\delta))\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi+\kappa_\pi\beta\gamma(1+\delta)\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi
\end{aligned}$$

where the last inequality uses Lemma B.1(5). We use (B) in (B) to obtain: $\forall k \geq 0$,

$$\begin{aligned}
&\langle (W^{k+1-i} - J^{(k+1)})(\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}), (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i+1)} - \nabla f^{(i)})\rangle_\pi \\
\leq & \kappa_\pi \beta L \gamma \beta_\pi^{2(k+1-i)} (1+\delta)^3 \|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2 \\
&+ \beta_\pi^{2(k+1-i)}(1+\delta)^2 (\underbrace{\kappa_\pi \beta \gamma L (1+\delta) \|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi \|\bar{\mathbf{g}}^{(i)}\|_\pi}_{C_1} \\
&+ \underbrace{(1 + \kappa_\pi \beta L (1+\delta)) \|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi \|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi)}_{C_2}
\end{aligned}$$

By Young inequality, we obtain that:

$$C_1 \leq 0.5\eta_1 \kappa_\pi^2 \beta_\pi^2 (1+\delta)^2 \|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2 + 0.5\eta_1^{-1}\gamma^2 L^2 \|\bar{\mathbf{g}}^{(i)}\|_\pi^2,$$

where $\eta_1 > 0$ is arbitrary, and that,

$$C_2 \leq 0.5(1 + \kappa_\pi \beta L(1+\delta))(\eta_2 \|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2 + 0.5\eta_2^{-1}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2),$$

where $\eta_1 > 0$ is arbitrary. Later we will set $\eta_1 = \dfrac{1-\beta}{6\kappa_\pi^2 \beta_\pi^2 (1+\delta)^2}$ and $\eta_2 = \dfrac{1-\beta}{6(1+\kappa_\pi \beta L(1+\delta))}$.

Note that $(W^{k+1-i} - J^{(k+1)})J^{(i)} = 0$, the second term $D_2$ can be written as follow:

$$D_2 = \mathbb{E}[\langle (W^{k+1-i} - J^{(k+1)})\mathbf{y}^{(i)}, (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i)}]$$

The second term can be bounded as follow:

$$\begin{aligned}
D_2 =& \mathbb{E}[\langle (W^{k+1-i} - J^{(k+1)})\mathbf{y}^{(i)}, (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i)}] \\
=& \mathbb{E}[\langle W^{k+1-i}\mathbf{y}^{(i)}, (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i+1)}] \\
=& \mathbb{E}[\langle W^{k+2-i}(\mathbf{y}^{(i-1)} - \mathbf{g}^{(i)} - \mathbf{g}^{(i-1)}), (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i)}] \\
=& \mathbb{E}[\langle W^{k+2-i}\mathbf{g}^{(i)}, (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i)}] \\
=& \mathbb{E}[\langle W^{k+2-i}(\mathbf{g}^{(i)} - \nabla f^{(i)}), (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i+1)}] \\
=& \mathbb{E}\left[\mathrm{tr}\left((\mathbf{g}^{(i)} - \nabla f^{(i)})^\top (J^{(k+1)} - W^{k+1-i})^\top \Pi^{-1} W^{k+2-i}(\mathbf{g}^{(i)} - \nabla f^{(i)})\right) | \mathcal{F}^{(i)}\right] \\
=& \mathbb{E}\left[\mathrm{tr}\left((\mathbf{g}^{(i)} - \nabla f^{(i)})(\mathbf{g}^{(i)} - \nabla f^{(i)})^\top (J^{(k+1)} - W^{k+1-i})^\top \Pi^{-1} W^{k+2-i}\right) | \mathcal{F}^{(i)}\right]
\end{aligned}$$

where $\Pi = \mathrm{diag}(\pi)$. In light of Assumption 3, (B) reduces to

$$\begin{aligned}
&\mathbb{E}[\langle (W^{k+1-i} - J^{(k+1)})(\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}), (W^{k+1-i} - J^{(k+1)})(\nabla f^{(i)} - \mathbf{g}^{(i)})\rangle_\pi | \mathcal{F}^{(i)}] \\
&= \sigma^2 \mathrm{tr}\left((J^{(k+1)} - W^{k+1-i})^\top \mathrm{diag}(\pi)^{-1} W^{k+2-i}\right) \\
&= \sigma^2 \mathrm{tr}\left(\mathrm{diag}(\pi)^{-1} W^{k+2-i}(J^{(k+1)} - W^{k+1-i})\right) \\
&\leq \sigma^2 \langle (J^{(k+1)} - W^{k+1-i}, W^{k+2-i}\rangle)_\pi \\
&\leq \sigma^2 \|J^{(k+1)} - W^{k+1-i}\|_\pi \|W^{k+2-i}\|_\pi \\
&\leq \sigma^2 \beta_\pi^{k+1-i}(1+\delta)
\end{aligned}$$

The last inequality holds because $\|W\|_\pi = 1$. Now combine the two parts and we finish the proof of this lemma. □

**Lemma B.5.** *When* $\gamma \le \dfrac{1-\beta}{12L\kappa_\pi\gamma(1+\delta)^3}$, $\mathbb{E}\sum_{k=0}^{T}\|\mathbf{y}^{(k+1)} - J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2$ i.e.*the accumulated consensus error of* $\mathbf{y}$ *, can be bounded with the following inequality:*

$$\begin{aligned}
&\mathbb{E}\sum_{k=0}^{T}\|\mathbf{y}^{(k+1)} - J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2 \\
\le&\frac{8n(T+1)c_\pi^2(1+\delta)^2\beta_\pi^2\sigma^2}{1-\beta} + \frac{36L^2(1+\delta)^6(1+\kappa_\pi\beta)^2}{(1-\beta_\pi)^2}\mathbb{E}\sum_{i=0}^{T}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2 \\
&+ \frac{24L^2\kappa_\pi^2\beta_\pi^2\gamma^2(1+\delta)^4}{(1-\beta_\pi)^2}\mathbb{E}\sum_{i=0}^{T}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + 2\mathbb{E}\|\mathbf{y}^{(0)} - J^{(0)}\mathbf{y}^{(0)}\|_\pi^2
\end{aligned}$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}\|\mathbf{y}^{(k+1)} - J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2 \overset{(B.2)}{=} \sum_{i=0}^{k}\mathbb{E}\|(W^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\|_\pi^2 \\
&+ 2\sum_{i=0}^{k}\mathbb{E}\langle W^{k+1-i}(\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}), (W^{k+1-i} - J^{(k+1)})(\mathbf{g}^{(i+1)} - \mathbf{g}^{(i)})\rangle_\pi \\
&\overset{(B.3),(B.4)}{\le} \sum_{i=0}^{k}\beta_\pi^{2(k+1-i)}(1+\delta)^2\left(3nc_\pi^2\sigma^2 + 12L^2\kappa_\pi^2\beta_\pi^2\gamma^2\mathbb{E}\|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\right) \\
&+ \sum_{i=0}^{k}\beta_\pi^{2(k+1-i)}(1+\delta)^2\left(6L^2\gamma^2\mathbb{E}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + C_2\mathbb{E}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2\right) \\
&+ 2\beta_\pi^2\sigma^2\sum_{i=0}^{k}\beta_\pi^{k+1-i}(1+\delta) \\
&+ 2\mathbb{E}\sum_{i=0}^{k}\beta_\pi^{2(k-i)}\left(\frac{1-\beta}{6} + L\kappa_\pi\beta\gamma(1+\delta)^3\right)\|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2 \\
&+ \mathbb{E}\sum_{i=0}^{k}\beta_\pi^{2(k-i)}(1+\delta)^4\left(\frac{6\kappa_\pi^2\beta_\pi^2L^2\gamma^2}{1-\beta}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 + \frac{C_2}{1-\beta}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2\right) \\
\le&\left(3nc_\pi^2(1+\delta)^2 + 2(1+\delta)\right)\frac{\beta_\pi^2\sigma^2}{1-\beta_\pi^2} \\
&+ \left(6L^2\beta_\pi^2(1+\delta)^2\gamma^2 + (1+\delta)^4\frac{6\kappa_\pi^2\beta_\pi^2L^2\gamma^2}{1-\beta}\right)\mathbb{E}\sum_{i=0}^{k}\beta_\pi^{2(k-i)}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2 \\
&+ \left((1+\delta)^2C_2 + \frac{(1+\delta)^4C_2}{1-\beta}\right)\mathbb{E}\sum_{i=0}^{k}\beta_\pi^{2(k+1-i)}\|\mathbf{w}^{(i)} - \bar{\mathbf{x}}^{(i)}\|_\pi^2 \\
&+ C_3\mathbb{E}\sum_{i=0}^{k}\beta_\pi^{2(k-i)}\|\mathbf{y}^{(i)} - J^{(i)}\mathbf{y}^{(i)}\|_\pi^2
\end{aligned}$$

where $C_2 := (12L^2 + 12L^2\kappa_\pi^2\beta_\pi^2(1+\delta)^2), C_3 := (12L^2\kappa_\pi^2\beta_\pi^2\gamma^2(1+\delta)^2 + \dfrac{1-\beta}{3} + 2L\kappa_\pi\beta\gamma(1+\delta)^3)$. The last inequality uses $\sum_{i=0}^{k}\beta_\pi^{2(k+1-i)} \le \dfrac{\beta_\pi^2}{1-\beta_\pi^2}$.

To further bound $\mathbb{E}\sum_{k=0}^{T}\|\mathbf{y}^{(k+1)}-J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2$, we can sum up the inequality above from $k=0$ to $T$ and calculate the coefficient term by term:

$$\begin{aligned}\sigma^2:\sum_{k=0}^{T}\left(3nc_\pi^2(1+\delta)^2+2(1+\delta)\right)\frac{\beta_\pi^2\sigma^2}{1-\beta}&=(T+1)\left(3nc_\pi^2(1+\delta)^2+2(1+\delta)\right)\frac{\beta_\pi^2\sigma^2}{1-\beta}\\&\leq\frac{4nc_\pi^2(1+\delta)^2(T+1)\beta_\pi^2\sigma^2}{1-\beta}\end{aligned}$$

$$\begin{aligned}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2:\left(6L^2\beta_\pi^2(1+\delta)^2\gamma^2+(1+\delta)^4\frac{6\kappa_\pi^2\beta_\pi^2L^2\gamma^2}{1-\beta}\right)\sum_{k=0}^{T}\sum_{i=0}^{k}\beta_\pi^{2(k-i)}\|\bar{\mathbf{g}}^{(i)}\|\\\leq\frac{12L^2\kappa_\pi^2\beta_\pi^2\gamma^2(1+\delta)^4}{(1-\beta_\pi)^2}\mathbb{E}\sum_{i=0}^{T}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2\end{aligned}$$

$$\begin{aligned}\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2:&((1+\delta)^2(C_2+\frac{(1+\delta)^4C_2}{(1-\beta_\pi)})\sum_{k=0}^{T}\mathbb{E}\sum_{i=0}^{k}\beta_\pi^{2(k+1-i)}\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2\\\leq&2(1+\delta)^6L^2\left(\frac{1+2\kappa_\pi^2\beta_\pi^2}{1-\beta}+\frac{(1+\kappa_\pi\beta)^2}{(1-\beta_\pi)^2}\right)\mathbb{E}\sum_{i=0}^{T}\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2\\\leq&\frac{18L^2(1+\delta)^6(1+\kappa_\pi\beta)^2}{(1-\beta_\pi)^2}\mathbb{E}\sum_{i=0}^{T}\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2\end{aligned}$$

For $\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2$, when $\gamma\leq\frac{1-\beta}{12L\kappa_\pi\beta(1+\delta)^3}$:

$$\begin{aligned}&C_3\mathbb{E}\sum_{k=0}^{T}\sum_{i=0}^{k}\beta_\pi^{2(k-i)}\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\\\leq&\left(12L^2\kappa_\pi^2\beta_\pi^2\gamma^2\frac{(1+\delta)^2}{1-\beta}+\frac{1}{3}+\frac{2L\kappa_\pi\beta\gamma(1+\delta)^3}{1-\beta}\right)\mathbb{E}\sum_{i=0}^{T}\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\\\leq&\frac{1}{2}\mathbb{E}\sum_{i=0}^{T}\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\end{aligned}$$

Now we can merge the same terms to obtain the lemma. □

**Lemma B.6.** *Under Assumption 1–Assumption 2, when* $\gamma\leq\frac{(1-\beta_\pi)^2}{12Lr(1+\delta)^4}$ *we have the following consensus lemma.*

$$\begin{aligned}&\mathbb{E}\sum_{k=0}^{T+1}\|\mathbf{x}^{(k+1)}-J^{(k+1)}\mathbf{x}^{(k+1)}\|_\pi^2\\\leq&\left(\frac{6nc_\pi^2\kappa_\pi^2\beta_\pi^2\gamma^2(T+1)(1+\delta)^4}{(1-\beta_\pi)^3}+\frac{48L^2c_\pi^2\kappa_\pi^4\beta_\pi^4\gamma^4(1+\delta)^6}{(1-\beta_\pi)^4}\right)\sigma^2\\&+\frac{48L^2\kappa_\pi^4\beta_\pi^4\gamma^4(1+\delta)^6}{(1-\beta_\pi)^4}\mathbb{E}\sum_{i=0}^{T}\|\overline{\nabla f}^{(i)}\|_\pi^2+\frac{4(1+\delta)^2\kappa_\pi^2\beta_\pi^2\gamma^2}{(1-\beta_\pi)^2}\mathbb{E}\|\mathbf{y}^{(0)}-J^{(0)}\mathbf{y}^{(0)}\|_\pi^2\end{aligned}$$

*Proof.* Notice that

$$\begin{aligned}\mathbf{w}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}&=V^{(k+1)^{-1}}WV^{(k)}(\mathbf{w}^{(k)}-\bar{\mathbf{x}}^{(k)})-\gamma(V^{(k+1)^{-1}}W-R)\mathbf{y}^{(k+1)}\\&=V^{(k+1)^{-1}}WV^{(k)}(\mathbf{w}^{(k)}-\bar{\mathbf{x}}^{(k)})-\gamma(V^{(k+1)^{-1}}W-R)(\mathbf{y}^{(k)}-J^{(k)}\mathbf{y}^{(k)})\end{aligned}$$

Repeat (B) from k to 1, we have

$$\mathbf{w}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}=-\gamma\sum_{i=0}^{k}(V^{(k+1)^{-1}}W^{k+1-i}-R)(\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)})$$

Thus, apply Jensen inequality in (B) :

$$\begin{aligned}&\|\mathbf{w}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_\pi^2\\\leq&\gamma^2\sum_{i=0}^{k}\frac{\|V^{(k+1)^{-1}}W^{k+1-i}-R\|_\pi^2}{\beta_\pi^{k-i}(1-\beta_\pi)}\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\\\leq&\frac{(1+\delta)^2\gamma^2\kappa_\pi^2\beta_\pi^2}{1-\beta}\sum_{i=0}^{k}\beta_\pi^{k-i}\|\mathbf{y}^{(i)}-J^{(i)}\mathbf{y}^{(i)}\|_\pi^2\end{aligned}$$

Sum up (B) and we obtain

$$\begin{aligned}&\mathbb{E}\sum_{k=0}^{T+1}\|\mathbf{w}^{(k+1)}-\bar{\mathbf{x}}^{(k+1)}\|_\pi^2\\\leq&\frac{(1+\delta)^2\gamma^2\kappa_\pi^2\beta_\pi^2}{1-\beta}\mathbb{E}\sum_{k=0}^{T+1}\sum_{i=-1}^{k-1}\beta_\pi^{k-i}\|\mathbf{y}^{(i+2)}-J^{(i+1)}\mathbf{y}^{(i+2)}\|_\pi^2\\\leq&\frac{(1+\delta)^2\gamma^2\kappa_\pi^2\beta_\pi^2}{(1-\beta_\pi)^2}\mathbb{E}\sum_{k=-1}^{\top}\|\mathbf{y}^{(k+1)}-J^{(k+1)}\mathbf{y}^{(k+1)}\|_\pi^2\\\overset{(B.5)}{\leq}&\frac{3n(T+1)c_\pi^2(1+\delta)^4\gamma^2\kappa_\pi^2\beta_\pi^2}{(1-\beta_\pi)^3}\sigma^2\\&+\frac{36L^2\kappa_\pi^2\beta_\pi^2(1+\kappa_\pi\beta)^2\gamma^2(1+\delta)^8}{(1-\beta_\pi)^4}\mathbb{E}\sum_{i=0}^{T}\|\mathbf{w}^{(i)}-\bar{\mathbf{x}}^{(i)}\|_\pi^2\\&+\frac{24L^2\kappa_\pi^4\beta_\pi^4\gamma^4(1+\delta)^6}{(1-\beta_\pi)^4}\mathbb{E}\sum_{i=0}^{T}\|\bar{\mathbf{g}}^{(i)}\|_\pi^2+\frac{2(1+\delta)^2\kappa_\pi^2\beta_\pi^2\gamma^2}{(1-\beta_\pi)^2}\mathbb{E}\|\mathbf{y}^{(0)}-J^{(0)}\mathbf{y}^{(0)}\|_\pi^2\end{aligned}$$

When $\gamma\leq\frac{(1-\beta_\pi)^2}{9L\kappa_\pi\beta(1+\kappa_\pi\beta)(1+\delta)^4}$, we have $\frac{36L^2\kappa_\pi^2\gamma^2(1+\delta)^8}{(1-\beta_\pi)^4}\leq\frac{1}{2}$ . Further note that

$$\mathbb{E}\|\bar{\mathbf{g}}^{(k)}\|_\pi^2=\mathbb{E}\|\overline{\nabla f}^{(k)}\|_\pi^2+\mathbb{E}\|\bar{\mathbf{g}}^{(k)}-\overline{\nabla f}^{(k)}\|_\pi^2\leq\mathbb{E}\|\overline{\nabla f}^{(k)}\|_\pi^2+c_\pi^2\sigma^2,$$

we can replace $\mathbb{E}\|\bar{\mathbf{g}}^{(k)}\|_\pi^2$ in (B) , combine like terms and the consensus lemma is achieved. □

**Lemma B.7.** *Under Assumptions Assumption 3–Assumption 1 and step-size $\gamma<\frac{1}{2L}$, it holds for any k that*

$$\begin{aligned}&\mathbb{E}[f(\bar{x}^{(k+1)})|\mathcal{F}_k]\\&\leq f(\bar{x}^{(k)})-\frac{\gamma}{2}\|\nabla f(\bar{x}^{(k)})\|^2-\frac{\gamma}{4}\|\overline{\nabla f}^{(k)}\|^2+\frac{\gamma L^2}{2n}\|\mathbf{w}^{(k)}-\bar{\mathbf{x}}^{(k)}\|_F^2+\frac{\gamma^2L\sigma^2}{2}\end{aligned}$$

*Proof.* Since $f$ is $L$-smooth, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

With Lemma B.1(b), we have $\bar{x}^{(k+1)} = \bar{x}^{(k)} - \gamma \bar{y}^{(k+1)} = \bar{x}^{(k)} - \gamma \bar{g}^{(k)}$. Thus, setting $y = \bar{x}^{(k+1)}$ and $x = \bar{x}^{(k)}$ in (B) to obtain:

$$f(\bar{x}^{(k+1)}) \leq f(\bar{x}^{(k)}) - \gamma \langle \nabla f(\bar{x}^{(k)}), \bar{g}^{(k)} \rangle + \frac{\gamma^2 L}{2}\|\bar{g}^{(k)}\|^2$$

conditioning on $\mathcal{F}_k$ , since $\mathbb{E}[\bar{g}^{(k)}|\mathcal{F}_k] = \overline{\nabla f}^{(k)}$ , we have

$$\begin{aligned}
&\mathbb{E}[f(\bar{x}^{(k+1)})|\mathcal{F}_k] - f(\bar{x}^{(k)}) \leq -\gamma \langle \nabla f(\bar{x}^{(k)}), \overline{\nabla f}^{(k)} \rangle + \frac{\gamma^2 L}{2}\mathbb{E}[\|\bar{g}^{(k)}\|^2|\mathcal{F}_k] \\
=& -\frac{\gamma}{2}\|\nabla f(\bar{x}^{(k)})\|^2 - \frac{\gamma}{2}\|\overline{\nabla f}^{(k)}\|^2 + \frac{\gamma}{2}\|\nabla f(\bar{x}^{(k)}) - \overline{\nabla f}^{(k)}\|^2 + \frac{\gamma^2 L}{2}\mathbb{E}[\|\bar{g}^{(k)}\|^2|\mathcal{F}_k] \\
\leq& -\frac{\gamma}{2}\|\nabla f(\bar{x}^{(k)})\|^2 - \frac{\gamma}{2}\|\overline{\nabla f}^{(k)}\|^2 + \frac{\gamma L^2}{2n}\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + \frac{\gamma^2 L}{2}(\|\overline{\nabla f}^{(k)}\|^2 + \frac{\sigma^2}{n}) \\
\leq& -\frac{\gamma}{2}\|\nabla f(\bar{x}^{(k)})\|^2 - \frac{\gamma}{4}\|\overline{\nabla f}^{(k)}\|^2 + \frac{\gamma L^2}{2n}\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + \frac{\gamma^2 L \sigma^2}{2n}
\end{aligned}$$

where the last inequality holds when $\gamma \leq \frac{1}{2L}$. Now we get the descent lemma.

□

Now we present the complete version of Theorem 2.

**Theorem 4.** *Under Assumption 1 – Assumption 3 ,* $f^* := \min_{x \in \mathbb{R}^d} f(x)$, $\Delta := f(\bar{x}^{(0)}) - f^*, \Delta_1 := \mathbb{E}\|\mathbf{y}^{(0)} - J^{(0)}\mathbf{y}^{(0)}\|_F^2/n$ . *Besides, if the learning rate is set as*

$$\gamma = \frac{1}{\frac{1}{\gamma_1} + \frac{1}{\gamma_2} + \frac{1}{\gamma_3} + \frac{1}{\gamma_4} + \frac{1}{\gamma_5} + \frac{1}{\gamma_6}}$$

*where*

$$\begin{aligned}
\gamma_1 &= \left(\frac{\Delta(1-\beta_\pi)^2}{2\Delta_1(1+\delta)^2 L^2 \kappa_\pi^3 \beta_\pi^2}\right)^{\frac{1}{3}}, \quad \gamma_2 = \left(\frac{2n\Delta}{(K+2)L\sigma^2}\right)^{\frac{1}{2}}, \\
\gamma_3 &= \left(\frac{\Delta(1-\beta_\pi)^3}{3L^2\kappa_\pi^3\beta_\pi^2\sigma^2(K+2)(1+\delta)^4}\right)^{\frac{1}{3}}, \quad \gamma_4 = \left(\frac{n\Delta(1-\beta_\pi)^4}{24L^4\kappa_\pi^5\beta_\pi^4\sigma^2(1+\delta)^6}\right)^{\frac{1}{5}}, \\
\gamma_5 &= \frac{(1-\beta_\pi)^2}{9L\kappa_\pi\beta(1+\kappa_\pi\beta)(1+\delta)^4}, \quad \gamma_6 = \frac{1}{2L}.
\end{aligned}$$

*It then follows*

$$\begin{aligned}
\frac{1}{K+2}\sum_{i=0}^{K+1}\mathbb{E}[\|\nabla f(\bar{x}^{(k)})\|^2] \leq\ & 2\left(\frac{2L\Delta\sigma^2}{n(K+2)}\right)^{\frac{1}{2}} \\
&+ \left(\frac{256L^2\kappa_\pi^4\beta_\pi^2\Delta^2\Delta_1}{(1-\beta_\pi)^2(K+2)^3}\right)^{\frac{1}{3}} + \left(\frac{96L^2\kappa_\pi^5\beta_\pi^2\Delta^2\sigma^2}{(1-\beta_\pi)^3(K+2)^2}\right)^{\frac{1}{3}} \\
&+ \frac{72L\kappa_\pi^3\beta(1+\kappa_\pi\beta)\Delta}{(K+2)(1-\beta_\pi)^2} + \frac{4L\Delta}{K+2}
\end{aligned}$$

*With this selected* $\gamma$*, The transient time is* $K > \frac{nL^2\kappa_\pi^2(1+\delta)^8\Delta}{(1-\beta_\pi)^4} = \mathcal{O}\left(\frac{nL^2}{(1-\beta_\pi)^4}\right)$

*Proof.* Rearrange (B.7) and using that $f$ is lower-bounded by $f^*$ we obtain: if $\gamma \le \min\{\frac{1}{2L}, \frac{(1-\beta_\pi)^2}{6Lr(1+\delta)^2}\}$, then

$$
\begin{aligned}
&\sum_{i=0}^{K+1} \mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \\
\le&\frac{2}{\gamma}(f(\bar{x}^{(0)}) - f^*) - \frac{1}{2}\sum_{i=0}^{K+1}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2] + \frac{L^2}{nd_\pi^2}\sum_{i=0}^{K+1}\mathbb{E}[\|\mathbf{w}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_\pi^2] + \frac{(K+2)\gamma L\sigma^2}{n} \\
\le&\frac{2}{\gamma}(f(\bar{x}^{(0)}) - f^*) - \left(\frac{1}{2} - \frac{48c_\pi^2 L^4\kappa_\pi^4\beta_\pi^4\gamma^4(1+\delta)^6}{nd_\pi^2(1-\beta_\pi)^4}\right)\sum_{i=0}^{K-1}\mathbb{E}[\|\overline{\nabla f}^{(k)}\|^2] + \frac{(K+2)\gamma L\sigma^2}{n} \\
&+ \frac{L^2}{nd_\pi^2}\left(\frac{6nc_\pi^2\kappa_\pi^2\beta_\pi^2\gamma^2(K+1)(1+\delta)^4}{(1-\beta_\pi)^3} + \frac{48L^2c_\pi^2\kappa_\pi^4\beta_\pi^4\gamma^4(1+\delta)^6}{(1-\beta_\pi)^4}\right)\sigma^2 \\
&+ \frac{L^2}{nd_\pi^2}\frac{4(1+\delta)^2\kappa_\pi^2\beta_\pi^2\gamma^2}{(1-\beta_\pi)^2}\mathbb{E}\|\mathbf{y}^{(0)} - J^{(0)}\mathbf{y}^{(0)}\|_\pi^2 \\
\le&\frac{2}{\gamma}(f(\bar{x}^{(0)}) - f^*) + \frac{L^2}{nd_\pi^2}\frac{4c_\pi^2(1+\delta)^2\kappa_\pi^2\beta_\pi^2\gamma^2}{(1-\beta_\pi)^2}\mathbb{E}\|\mathbf{y}^{(0)} - J^{(0)}\mathbf{y}^{(0)}\|_F^2 \\
&+ \frac{(K+2)\gamma L\sigma^2}{n} + \frac{L^2c_\pi^2}{nd_\pi^2}\left(\frac{6n\kappa_\pi^2\beta_\pi^2\gamma^2(K+1)(1+\delta)^4}{(1-\beta_\pi)^3} + \frac{48L^2\kappa_\pi^4\beta_\pi^4\gamma^4(1+\delta)^6}{(1-\beta_\pi)^4}\right)\sigma^2
\end{aligned}
$$

where the first inequality comes from Lemma B.7, the second inequality comes from Lemma B.6. In Lemma B.6 we need $\gamma \le \frac{(1-\beta_\pi)^2}{9L\kappa_\pi\beta(1+\kappa_\pi\beta)(1+\delta)^4} = \gamma_5$. And it is easy to verify that the last inequality in (B) holds when $\gamma \le \gamma_5$. In Lemma B.7 we need $\gamma \le \frac{1}{2L} = \gamma_6$. Then it is easy to verify our $\gamma$ in (4) meets these constraints. With the selection of $\gamma$ in (4), we have

$$
\begin{aligned}
&\frac{1}{K+2}\sum_{i=0}^{K+1}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 \\
\le&\frac{2\Delta}{K+2}\sum_{i=1}^{6}\frac{1}{\gamma_i} + \frac{4L^2\kappa_\pi^3\beta_\pi^2\gamma_1^2(1+\delta)^2}{(K+2)(1-\beta_\pi)^2}\Delta_1 \\
&+ \frac{\gamma_2 L\sigma^2}{n} + \frac{L^2\kappa_\pi}{n}\left(\frac{9n\kappa_\pi^2\gamma_3^2(1+\delta)^4}{(1-\beta_\pi)^3} + \frac{72L^2\kappa_\pi^4\gamma_4^4(1+\delta)^6}{(1-\beta_\pi)^4(K+2)}\right)\sigma^2 \\
\le&2\left(\frac{2L\Delta\sigma^2}{n(K+2)}\right)^{\frac{1}{2}} + \left(\frac{128L^2\kappa_\pi^3\beta_\pi^2\Delta^2\Delta_1(1+\delta)^2}{(K+2)^3(1-\beta_\pi)^2}\right)^{\frac{1}{3}} + 2\left(\frac{3L^2\kappa_\pi^3\beta_\pi^2\Delta^2\sigma^2(1+\delta)^4}{(1-\beta_\pi)^3(K+2)^2}\right)^{\frac{1}{3}} \\
&+ \frac{18L\kappa_\pi\beta(1+\kappa_\pi\beta)\Delta(1+\delta)^4}{(K+2)(1-\beta_\pi)^2} + \frac{4L\Delta}{K+2}
\end{aligned}
$$

The first term is the standard error term for centralized SGD while other parts are decentralized network effect. Finally we replace $\frac{c_\pi^2}{d_\pi^2}$ and $1+\delta$ with $\kappa_\pi$ and $\sqrt{2\kappa_\pi}$ to obtain (4). □

# C Proof of section 6

The only difference between MG-Push-DIGing and original Push-DIGing is that every $W$ is replaced with $W^b$. Note that $\beta$ only arises from $\|W - \pi\mathbb{1}_n^\top\|_\pi = \beta$ in our proof, and $\|W^b - \pi\mathbb{1}_n^\top\|_\pi = \|(W - \pi\mathbb{1}_n^\top)^b\|_\pi \le$

$\|W-\pi\mathbb{1}_n^\top\|_\pi^b = \beta_\pi^b$, now $\beta$ appearing in the analysis of Appendix B are all replaced by $\hat{\beta} := \beta_\pi^b$. Besides, since we have used a $b$ times larger batch size, the noise is reduced by $b$ times, *i.e.*$\hat{\sigma} = \sigma^2/b$. Taking $b = \dfrac{p}{1-\beta}$, where $p$ is any number larger than $(1+\sqrt{3\ln(\kappa_\pi)})^2$. Then we have

$$p\hat{\beta} \le p(\beta_\pi^{1/(1-\beta_\pi)})^p \le pe^{-p} \le \frac{1+\ln(\kappa_\pi)+\sqrt{3\ln(\kappa_\pi)}}{e^{1+2\sqrt{3\ln(\kappa_\pi)}}\kappa_\pi^3} \le \frac{1}{\kappa_\pi^3}.$$

*Proof.*

$$\begin{aligned}
\frac{1}{K}\sum_{i=0}^{K-1}\mathbb{E}\|\nabla f(\bar{x}^{(k)})\|^2 &\overset{(B)}{\le} \mathcal{O}\left(\frac{L\Delta\hat{\sigma}^2}{nK}\right)^{\frac{1}{2}} + \mathcal{O}\left(\frac{\Delta^2L^2\kappa_\pi^4\hat{\beta}^2}{K^3}\right)^{\frac{1}{3}} \\
&\quad + \mathcal{O}\left(\frac{\Delta^2L^2\kappa_\pi^4\hat{\beta}^2\hat{\sigma}^2}{K^2}\right)^{\frac{1}{3}} + \mathcal{O}\left(\frac{\Delta^4L^4\kappa_\pi^8\hat{\beta}^4\hat{\sigma}^2}{nK^5}\right)^{\frac{1}{5}} + \mathcal{O}\left(\frac{\Delta L(1+\kappa_\pi^4\hat{\beta}^2)}{K}\right) \\
&\overset{T=bK}{=} \mathcal{O}\left(\frac{L\Delta\sigma^2}{nT}\right)^{\frac{1}{2}} + \mathcal{O}\left(\frac{\Delta^2L^2p\kappa_\pi^4\hat{\beta}^2}{T^3(1-\beta_\pi)^3}\right)^{\frac{1}{3}} + \mathcal{O}\left(\frac{\Delta^2L^2p\kappa_\pi^4\hat{\beta}^2\sigma^2}{T^2(1-\beta_\pi)}\right)^{\frac{1}{3}} \\
&\quad + \mathcal{O}\left(\frac{\Delta^4L^4p^4\kappa_\pi^8\hat{\beta}^4\sigma^2}{nT^5(1-\beta_\pi)^4}\right)^{\frac{1}{5}} + \mathcal{O}\left(\frac{\Delta Lp(1+\kappa_\pi^4\hat{\beta}^2)}{T(1-\beta_\pi)}\right) \\
&\overset{(C)}{\le} \mathcal{O}\left(\frac{L\Delta\sigma^2}{nT}\right)^{\frac{1}{2}} + \mathcal{O}\left(\frac{\Delta^2L^2}{T^3(1-\beta_\pi)^3}\right)^{\frac{1}{3}} + \mathcal{O}\left(\frac{\Delta^2L^2\sigma^2}{T^2(1-\beta_\pi)}\right)^{\frac{1}{3}} \\
&\quad + \frac{1}{T}\mathcal{O}\left(\frac{\Delta^4L^4\sigma^2}{n(1-\beta_\pi)^4}\right)^{\frac{1}{5}} + \mathcal{O}\left(\frac{\Delta Lp}{T(1-\beta_\pi)}\right) \\
&\sim \mathcal{O}\left(\frac{L\Delta\sigma^2}{nT}\right)^{\frac{1}{2}} + \mathcal{O}\left(\frac{L\Delta(1+\ln(\kappa_\pi)+\sqrt{3\ln(\kappa_\pi)})}{(1-\beta_\pi)T}\right)
\end{aligned}$$

where $1-\hat{\beta}, \Delta_1$ are ignored in the first inequality because they are $\mathcal{O}(1)$ constants. □

# D Proof of Proposition 2

For $n \ge 2$, we more generally consider the following matrix

$$W = \begin{bmatrix} (1-\epsilon)/2 & (1-\epsilon)/2 & \cdots & (1-\epsilon)/2 & 1 \\ (1+\epsilon)/2 & 0 & & & \\ & \ddots & \ddots & & \\ & & (1+\epsilon)/2 & 0 & \\ & & & (1+\epsilon)/2 & 0 \end{bmatrix} = \frac{1+\epsilon}{2}J + \frac{1-\epsilon}{2}e_1\mathbb{1}_n^\top$$

for some $\epsilon \in [0,1)$, whose right Perron vector is

$$\pi \propto \left((2/(1+\epsilon))^{n-1}, (2/(1+\epsilon))^{n-2}, \ldots, 2/(1+\epsilon), 1\right)^\top,$$

which directly implies $\ln(\kappa_\pi) = (n-1)\ln(2/(1+\epsilon))$. On can also easily verify that

$$\pi_1 = \frac{1-\epsilon}{2(1-((1+\epsilon)/2)^{n-1})} > \frac{1-\epsilon}{2}.$$

Following the same workflow, by denoting $\mathbb{1}_n^\top v$ as $s$, we have

$$\begin{aligned}
&\|Mv\|_\pi^2 = (Mv)^\top \Pi^{-1}(Mv) \\
=&\left(\frac{1+\epsilon}{2}Jv + \left(\frac{1-\epsilon}{2}e_1 - \pi\right)s\right)^\top \Pi^{-1}\left(\frac{1+\epsilon}{2}Jv + \left(\frac{1-\epsilon}{2}e_1 - \pi\right)s\right) \\
=&\frac{(1+\epsilon)^2}{4}(Jv)^\top\Pi^{-1}Jv + (1+\epsilon)(Jv)^\top\Pi^{-1}\left(\frac{1-\epsilon}{2}e_1 - \pi\right)s \\
&+\left(\frac{1-\epsilon}{2}e_1 - \pi\right)^\top \Pi^{-1}\left(\frac{1-\epsilon}{2}e_1 - \pi\right)s^2
\end{aligned}$$

Using $J^\top\Pi^{-1}J = \frac{2}{1+\epsilon}\Pi^{-1} + \left(\pi_1^{-1} - \frac{2}{1+\epsilon}\pi_n^{-1}\right)e_n e_n^\top$, and $\Pi^{-1}e_1 = e_1/\pi_1$, $\Pi^{-1}\pi = \mathbb{1}_n$, we further have

$$\begin{aligned}
\|Mv\|_\pi^2 =& \frac{(1+\epsilon)^2}{4}v^\top\left(\frac{2}{1+\epsilon}\Pi^{-1} + \left(\pi_1^{-1} - \frac{2}{(1+\epsilon)\pi_n}\right)e_n e_n^\top)\right)v \\
&+(1+\epsilon)\left(\frac{1-\epsilon}{2\pi_1}v_n - s\right)s + \left(\frac{(1-\epsilon)^2}{4\pi_1} + \epsilon\right)s^2 \\
=&\frac{1+\epsilon}{2}\|v\|_\pi^2 - \left(\frac{(1+\epsilon)^2}{4}\left(\frac{2}{(1+\epsilon)\pi_n} - \frac{1}{\pi_1}\right)v_n^2 - \frac{1-\epsilon^2}{2\pi_1}v_n s + \left(1 - \frac{(1-\epsilon)^2}{4\pi_1}\right)s^2\right).
\end{aligned}$$

Considering the determinant of the quadratic function with respect to $v_n$ and $s$ in (D), we have

$$\begin{aligned}
&\left(\frac{1-\epsilon^2}{2\pi_1}\right)^2 - 4\times\frac{(1+\epsilon)^2}{4}\left(\frac{2}{(1+\epsilon)\pi_n} - \frac{1}{\pi_1}\right)\times\left(1 - \frac{(1-\epsilon)^2}{4\pi_1}\right) \\
=&\frac{(1+\epsilon)^2}{4\pi_1^2}\left((1-\epsilon)^2 - (1+\epsilon)^2\left(\frac{2\pi_1}{(1+\epsilon)\pi_n} - 1\right)\left(4\pi_1 - (1-\epsilon)^2\right)\right) \\
\le&\frac{(1+\epsilon)^2}{4\pi_1^2}\left((1-\epsilon)^2 - (1+\epsilon)^2\frac{1-\epsilon}{1+\epsilon}(1-\epsilon^2)\right) = \frac{(1-(1+\epsilon)^2)(1-\epsilon)^2}{4\pi_1^2} \le 0
\end{aligned}$$

where (D) is due to $\pi_1 \ge \pi_n$ and $\pi_1 > (1-\epsilon)/2$. Since the determinant is non-positive, combining this with (D), we conclude $\|Mv\|_\pi^2 \le \frac{1+\epsilon}{2}\|v\|_\pi^2$ and the equality holds for $v$ with $v_n = 0$ and $s = \sum_{i=1}^{n-1} v_i = 0$. Therefore, we have $\beta = \sqrt{(1+\epsilon)/2} \in [1/\sqrt{2}, 1)$, $\ln(\kappa_\pi) = -2(n-1)\ln(\beta)$, and consequently

$$\frac{1+\ln(\kappa_\pi)}{1-\beta} = \frac{1-2(n-1)\ln(1-(1-\beta_\pi))}{1-\beta} = O(n)$$

provided with $\beta \le 1 - 1/n$ so that $-n\ln(\beta) = \Omega(1)$.

# E Proof of Theorem 1

The first complexity $\Omega(\frac{\sigma\sqrt{L\Delta}}{\sqrt{nK}})$ is customary, whose proof can be found in *e.g.*, [36, 35]. We thus focus on proving the second term $\Omega((1+\ln(\kappa_\pi))L\Delta/K)$. To proceed, we denote the $j$-th coordinate of a vector $x \in \mathbb{R}^d$ by $[x]_j$ for $1 \le j \le d$, and define

$$\mathrm{prog}(x) := \begin{cases} 0 & \text{if } x = 0; \\ \max\limits_{1\le j\le d}\{j : [x]_j \ne 0\} & \text{othwewise.} \end{cases}$$

We also present several key lemmas, which have appeared in prior literature.

**Lemma E.1** (Lemma 2 of [43]). *Let 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* $\forall\, z \in \mathbb{R}$,

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

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

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

**Lemma E.2** (Lemma 4 of [38]). *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 E.1.*
2. $h_1$ *and* $h_2$ *are zero-chain,* i.e., $\mathrm{prog}(\nabla h_i(x)) \le \mathrm{prog}(x) + 1$ *for all* $x \in \mathbb{R}^d$ *and* $i = 1, 2$. *Furthermore, if* $\mathrm{prog}(x)$ *is odd, then* $\mathrm{prog}(\nabla h_1(x)) \le \mathrm{prog}(x)$*; if* $\mathrm{prog}(x)$ *is even, then* $\mathrm{prog}(\nabla h_2(x)) \le \mathrm{prog}(x)$.
3. $h_1$ *and* $h_2$ *are also* $L_0$*-smooth with* $L_0 = 152$.

Our proof proceeds in 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 E.2, 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 $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = 2L\lambda^2 h(x/\lambda)/3L_0$. Since $h_1$ and $h_2$ are also $L_0$-smooth, $\{f_i\}_{i=1}^n$ are $L$-smooth. Furthermore, since

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

to ensure $\{f_i\}_{i=1}^n$ satisfy Assumption 2, it suffices to let

$$\frac{L\lambda^2\Delta_0 d}{L_0} \leq \Delta, \quad i.e., \quad \lambda \leq \sqrt{\frac{L_0\Delta}{L\Delta_0 d}}.$$

With the functions defined above, we have $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) = L\lambda^2\ell(x/\lambda)/(3L_0)$ and $\text{prog}(\nabla f_i(x)) = \text{prog}(x)+1$ if $\text{prog}(x)$ is even and $i \in E_1$ or $\text{prog}(x)$ is odd and $i \in E_2$, otherwise $\text{prog}(\nabla f_i(x)) \leq \text{prog}(x)$. Therefore, to make progress (*i.e.*, to increase $\text{prog}(x)$), for any gossip algorithm $A \in \mathcal{A}_W$, 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 Appendix D with $\epsilon = 2\beta_\pi^2 - 1$ so that $\frac{1+\ln(\kappa_\pi)}{1-\beta} = O(n)$. 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 $A \in \mathcal{A}_W$ to increase $\text{prog}(\hat{x})$ by 1 if it is odd. Therefore, we have

$$\left\lceil \text{prog}(\hat{x}^{(k)})/2 \right\rceil \leq \left\lfloor \frac{k}{2n/3} \right\rfloor, \quad \forall\, k \geq 0.$$

which further implies

$$\text{prog}(\hat{x}^{(k)}) \leq 2\left\lfloor \frac{k}{2n/3} \right\rfloor + 1 \leq 3k/n + 1, \quad \forall\, k \geq 0.$$

(Step 3.) We finally show the error $\mathbb{E}[\|\nabla f(x)\|^2]$ is lower bounded by $\Omega\left(\frac{(1+\ln(\kappa_\pi))L\Delta}{(1-\beta_\pi)K}\right)$, with any algorithm $A \in \mathcal{A}_W$ with $K$ communication rounds. For any $K \geq n$, we set

$$d = 2\left\lfloor \frac{K}{2n/3} \right\rfloor + 2 \leq 3K/n + 2 \leq 5K/n.$$

and

$$\lambda = \left(\frac{nL_0\Delta}{5L\Delta_0 K}\right)^{\frac{1}{2}}.$$

Then (E) naturally holds. Since $\text{prog}(\hat{x}^{(K)}) < d$ by (E), using the last point of Lemma E.1 and (E), we have

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

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