Journal of Machine Learning Research 24 (2023) 1-53 Submitted 3/22; Revised 4/23; Published 9/23

Removing Data Heterogeneity Influence Enhances Network Topology Dependence of Decentralized SGD

Kun Yuan KUNYUAN@PKU.EDU.CN
Center for Machine Learning Research, Peking University
AI for Science Institute
Beijing 100871, P. R. China

Sulaiman A. Alghunaim SULAIMAN.ALGHUNAIM@KU.EDU.KW
Department of Electrical Engineering
Kuwait University
Safat 13060, Kuwait

Xinmeng Huang XINMENGH@SAS.UPENN.EDU
Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Suvrit Sra

Abstract

We consider decentralized stochastic optimization problems, where a network of n nodes cooperates to find a minimizer of the globally-averaged cost. A widely studied decentralized algorithm for this problem is the decentralized SGD (D-SGD), in which each node averages only with its neighbors. D-SGD is efficient in single-iteration communication, but it is very sensitive to the network topology. For smooth objective functions, the transient stage (which measures the number of iterations the algorithm has to experience before achieving the linear speedup stage) of D-SGD is on the order of $O(n/(1-\beta)^2)$ and $O(n^3/(1-\beta)^4)$ for strongly and generally convex cost functions, respectively, where $1-\beta \in (0,1)$ is a topology-dependent quantity that approaches 0 for a large and sparse network. Hence, D-SGD suffers from slow convergence for large and sparse networks.

In this work, we revisit the convergence property of the D^2/Exact-Diffusion algorithm. By eliminating the influence of data heterogeneity between nodes, D^2/Exact-diffusion is shown to have an enhanced transient stage that is on the order of $\tilde{O}(n/(1-\beta))$ and $O(n^3/(1-\beta)^2)$ for strongly and generally convex cost functions (where $\tilde{O}(\cdot)$ hides all logarithm factors), respectively. Moreover, when D^2/Exact-Diffusion is implemented with both gradient accumulation and multi-round gossip communications, its transient stage can be further improved to $\tilde{O}(1/(1-\beta)^{\frac{1}{2}})$ and $\tilde{O}(n/(1-\beta))$ for strongly and generally convex cost functions, respectively. To our knowledge, these established results for D^2/Exact-Diffusion have the best (*i.e.*, weakest) dependence on network topology compared to existing decentralized algorithms. Numerical simulations are conducted to validate our theories.

Keywords: Decentralized optimization, stochastic optimization, transient stage

1. Introduction

Large-scale optimization and learning has become an essential tool in many practical applications. State-of-the-art performances has been reported in various fields, such as signal

processing, control, reinforcement learning, and deep learning. The amount of data needed to achieve satisfactory results in these tasks is typically very large. Moreover, increasing the size of training data can significantly improve the ultimate performance in these tasks. For this reason, the scale of optimization and learning nowadays calls for efficient distributed solutions across multiple computing nodes (*e.g*, machines).

This work considers a network of n collaborative nodes connected through a given topology. Each node owns a private and local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$ and the goal of the network is to find a solution, denoted by $x^\star$, of the *stochastic optimization* problem

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

In this problem, ξ_i is a random variable that represents the local data available at node i, and follows a local distribution D_i. Each node i has access to the stochastic gradient $\nabla F(x_i;\xi_i)$ of its private cost, but has to communicate to exchange information with other nodes. In practice, the local data distribution D_i differs among nodes, and thus, $f_i(x) \neq f_j(x)$ holds for any node i and j. For *convex costs*, the data heterogeneity across the network can be characterized by $b^2 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^\star)\|^2$. If all local data samples follow the same distribution D, we have $f_i(x) = f_j(x)$ for any i, j and hence $b^2 = 0$.

One of the leading algorithms for solving problem (1) is parallel SGD (P-SGD) (Zinkevich et al., 2010). In P-SGD, each node computes its local stochastic gradient and then synchronizes across the entire network to find the globally averaged stochastic gradient used to update the model parameters (*i.e.*, solution estimates). The global synchronization step needed to compute the globally averaged stochastic gradient can be implemented via a Parameter Server (Smola and Narayanamurthy, 2010; Li et al., 2014) or Ring-Allreduce (Gibiansky, 2017), which suffers from either significant bandwidth cost or high latency, see (Ying et al., 2021c, Table I) for detailed discussion. Decentralized SGD (D-SGD) (Lopes and Sayed, 2008; Nedic and Ozdaglar, 2009; Chen and Sayed, 2012) is a promising alternative to P-SGD due to its ability to reduce the communication overhead (Lian et al., 2017; Assran et al., 2019; Lian et al., 2018; Yuan et al., 2021). D-SGD is based on *local averaging* (also known as *gossip averaging*) in which each node computes the locally averaged model parameters with its direct neighbors as opposed to the global average. Moreover, no global synchronization step is required in D-SGD. In a delicately-designed sparse topology such as one-peer exponential graph (Assran et al., 2019; Ying et al., 2021a), each node only needs to communicate with *one* neighbor per iteration in D-SGD. This results in significantly reduced communication costs compared to P-SGD, see the discussion in (Assran et al., 2019; Ying et al., 2021b; Chen et al., 2021) and (Ying et al., 2021c, Table I).

Apart from its efficient single-iteration communication, D-SGD can asymptotically achieve the same linear speedup as P-SGD (Lian et al., 2017, 2018; Assran et al., 2019; Koloskova et al., 2020). *Linear speedup* refers to a property in distributed algorithms where the number of iterations needed to reach an ϵ-accurate solution reduces linearly with the number of nodes. The *transient stage* (Pu et al., 2021), which refers to those iterations before an algorithm reaches its linear speedup stage, is an important metric to measure the convergence performance of decentralized algorithms. **The convergence rate and transient stage of D-SGD are very sensitive to the network topology.** For example, the transient stage of D-SGD for generally convex or non-convex objective functions is on the order of

$O(n^3/(1-\beta)^4)$ (Koloskova et al., 2020), where $1-\beta$ measures the network topology connectivity. For a large and sparse network in which $1-\beta$ approaches 0, D-SGD will suffer from an extremely long transient stage, and it may not be able to reach the linear speedup stage given limited training time and computing resource budget. For this reason, D-SGD may end up with a low-quality solution that is significantly worse than that obtained by P-SGD. As a result, improving the network topology dependence (*i.e.*, making the convergence rate less sensitive to network topology) in D-SGD is crucial to enhance its convergence rate and solution accuracy.

The data heterogeneity across each node is the main factor contributing to D-SGD strong dependence on the network topology as shown in (Koloskova et al., 2020; Yuan et al., 2020; Pu et al., 2021). This naturally motivates us to examine whether removing the influence of data heterogeneity (*i.e.*, b^2) can improve the dependence on the topology (*i.e.*, $1-\beta$) of D-SGD.

1.1 Main Results

In this work, we revisit the D^2 algorithm (Tang et al., 2018), also known as Exact-Diffusion (Yuan et al., 2018a,b, 2020) (or NIDS (Li et al., 2019b)). D^2/Exact-Diffusion is a decentralized optimization algorithm that can remove the influence of data heterogeneity (Yuan et al., 2018b; Li et al., 2019b), but it remains unclear whether D^2/Exact-Diffusion has an improved network topology dependence compared to D-SGD in the transient stage. In this work, we establish non-asymptotic convergence rates for D^2/Exact-Diffusion under both the generally-convex and strongly-convex settings. The established bounds show that D^2/Exact-Diffusion has the best known network topology dependence compared with existing results. In particular, we establish that D^2/Exact-Diffusion at iteration T converges with the following rate:

$$\text{(G-C)}\quad \frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)=O\Big(\frac{\sigma}{\sqrt{nT}}+\frac{\sigma^{2/3}}{(1-\beta)^{1/3}T^{2/3}}+\frac{1}{(1-\beta)T}\Big) \tag{2}$$

$$\text{(S-C)}\quad \frac{1}{H_T}\sum_{k=0}^{T}h_k\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)=\tilde{O}\Big(\frac{\sigma^2}{nT}+\frac{\sigma^2}{(1-\beta)T^2}+\frac{1}{1-\beta}\exp\{-(1-\beta)T\}\Big) \tag{3}$$

where $\bar{x}^{(k)}=\frac{1}{n}\sum_{i=1}^n x_i^{(k)}$, $x_i^{(k)}$ is the estimate of node i at iteration k, and σ^2 denotes the variance of the stochastic gradient $\nabla F(x;\boldsymbol{\xi}_i)$. The weights $h_k\geq 0$ are given in Lemma 17 and $H_T=\sum_{k=0}^T h_k$. The notation $\tilde{O}(\cdot)$ hides all logarithm factors. Moreover, (G-C) stands for the generally convex scenario while (S-C) stands for strongly convex one. Below, we compare this result with D-SGD.

Transient stage and linear speedup. When T becomes sufficiently large, the first term $\sigma^2/(nT)$ in (3) (or $\sigma/\sqrt{nT}$ in (2)) will dominate the rate. In this scenario, D^2/Exact-Diffusion requires $T=O(1/n\epsilon)$ in (3) (or $T=O(1/n\epsilon^2)$ in (2)) iterations to reach a desired ϵ-accurate solution for strongly convex (or generally convex) problems, which is inversely proportional to the network size n. An algorithm is said to have reached the linear-speedup stage if for some T, the term involving nT is dominating the rate. Rates (2) and (3) corroborate that D^2/Exact-Diffusion, similar to P-SGD, achieves linear speedup for sufficiently large T. We note that D^2/Exact-Diffusion can also achieves linear speedup in

Scenario	D^2/Exact-Diffusion	D-SGD
Generally-convex	(2)	(2) + $O\left(\frac{b^{2/3}}{(1-\beta)^{2/3}T^{2/3}}\right)$
Strongly-convex	(3)	(3) + $O\left(\frac{b^2}{(1-\beta)^2T^2}\right)$

Table 1: Convergence rate comparison with D-SGD from (Koloskova et al., 2020).

the non-convex setting (Tang et al., 2018; Alghunaim and Yuan, 2022). The *transient stage* is the number of iterations needed to reach the linear-speedup stage, which is an important metric to measure the scalability of distributed algorithms (Pu et al., 2021). Consider, for instance, the D^2/Exact-Diffusion in the strongly convex scenario as given by (3). For it to achieve linear speedup, T must satisfy the condition $(1-\beta)T^2 \leq nT$, that is, $T \geq n/(1-\beta)$. Therefore, the transient stage in D^2/Exact-Diffusion for strongly-convex scenario requires $\tilde{O}(n/(1-\beta))$ iterations.

Comparison with D-SGD. Table 1 lists the convergence rates for both D-SGD and D^2/Exact-Diffusion in the generally and strongly convex scenarios. Compared to D-SGD, it is observed in (2) and (3) that D^2/Exact-Diffusion has eliminated the data heterogeneity b^2 term. Note that the term related to data heterogeneity b^2 has the strongest topology dependence on $1-\beta$ for D-SGD. In Table 2 we list the transient stage of D^2/Exact-Diffusion and other existing algorithms. It is observed that D^2/Exact-Diffusion has an improved (*i.e.*, shorter) transient stage in terms of $1-\beta$ compared to D-SGD by removing the influence of data heterogeneity. Gradient tracking methods (Xu et al., 2015; Lorenzo and Scutari, 2016; Nedich et al., 2017; Qu and Li, 2018; Xin et al., 2021) can also remove the data heterogeneity, but their transient stage established in existing prior works still suffers from worse network topology dependence than D^2/Exact-Diffusion and even D-SGD. In essence, D^2/Exact-Diffusion enjoys the state-of-the-art topology dependence in the generally and strongly convex scenarios.

Further improvement with multi-round gossip communication. Another orthogonal approach to improve network topology dependence is to run multiple gossip steps per D^2/Exact-Diffusion update. By leveraging multiple gossip communications and gradient accumulation, the transient stage of D^2/Exact-Diffusion can be significantly improved in both generally- and strongly-convex scenarios, as seen in the last row in Table 2.

1.2 Contributions

This work makes the following contributions:

- We revisit the D^2/Exact-Diffusion algorithm (Tang et al., 2018; Yuan et al., 2018a,b; Li et al., 2019b; Yuan et al., 2020) and establish its non-asymptotic convergence rate under *generally-convex* settings. By removing the influence of data heterogeneity, D^2/Exact-Diffusion is shown to improve the transient stage of D-SGD from $O(n^3/(1-\beta)^4)$ to $O(n^3/(1-\beta)^2)$, which is less sensitive to the network topology.

- We also establish the non-asymptotic convergence rate of D^2/Exact-Diffusion under *strongly-convex* settings. We demonstrate that D^2/Exact-Diffusion improves the tran-

Scenario	Generally-convex	Strongly-convex
D-SGD (Pu et al., 2021)	$O\left(\frac{n^3}{(1-\beta)^4}\right)$	$\tilde{O}\left(\frac{n}{(1-\beta)^2}\right)$
Gradient Tracking (Pu and Nedić, 2020)	N.A.	$\tilde{O}\left(\frac{n}{(1-\beta)^3}\right)$
D^2/Exact-Diffusion (Ours)	$O\left(\frac{n^3}{(1-\beta)^2}\right)$	$\tilde{O}\left(\frac{n}{1-\beta}\right)$
D^2/ED-MG (Ours)	$\tilde{O}\left(\frac{n}{(1-\beta)}\right)$	$\tilde{O}\left(\frac{1}{(1-\beta)^{\frac{1}{2}}}\right)$

Table 2: Transient stage comparison between D-SGD, gradient tracking, D^2/Exact-Diffusion, and D^2/Exact-Diffusion with multi-round gossip communication (D^2/ED-MG for short) in the strongly convex and generally convex settings. Note that $1-\beta \in (0,1)$. Notation $\tilde{O}(\cdot)$ hides all logarithm factors. The smaller the transient stage is, the faster the algorithm will achieve linear speedup.

sient stage of D-SGD from $\tilde{O}(n/(1-\beta)^2)$ to $\tilde{O}(n/(1-\beta))$. Furthermore, we prove that the transient stage of D-SGD is lower bounded by $\tilde{\Omega}(n/(1-\beta))$ with homogeneous data (*i.e.*, $b^2=0$) in the strongly-convex scenario. This implies that the dependence of D-SGD on the network topology can only match that of D^2/Exact-Diffusion if the data is homogeneous[1], which is typically an impractical assumption.

- We further improve the dependence on network topology by integrating *multiple gossip* and *gradient accumulation* to D^2/Exact-Diffusion. With these two useful techniques, the transient stage of D^2/Exact-Diffusion improves from $O(n^3/(1-\beta)^2)$ to $\tilde{O}(n/(1-\beta))$ in the generally-convex scenario, and $\tilde{O}(n/(1-\beta))$ to $\tilde{O}(1/(1-\beta)^{1/2})$ in the strongly-convex scenario. D^2/Exact-Diffusion with multiple gossip communications has a significantly better dependence on topology and network size than existing algorithms (Koloskova et al., 2020; Pu et al., 2021; Pu and Nedić, 2020; Huang and Pu, 2021; Koloskova et al., 2021).

1.3 Related Works

Decentralized optimization. Distributed optimization algorithms can at least be traced back to the work (Tsitsiklis et al., 1986). Decentralized gradient descent (DGD) (Nedic and Ozdaglar, 2009; Lopes and Sayed, 2008; Chen and Sayed, 2012) and dual averaging (Duchi et al., 2011) are among the earliest decentralized optimization algorithms. DGD can have several forms depending on the combination/communication step order such as consensus or diffusion (Sayed, 2014) (diffusion is also called adapt-then-combine DGD or just DGD in many recent works). Both DGD and dual averaging suffer from a bias caused by data heterogeneity even under deterministic settings (*i.e.,* no gradient noise exists) (Nedic and Ozdaglar, 2009; Chen and Sayed, 2013; Yuan et al., 2016) – see more explanation in Sec. 2.4. Numerous algorithms have been proposed to address this issue, such as alternating direction method of multipliers (ADMM) methods (Wei and Ozdaglar, 2012; Shi et al., 2014), explicit

1. The transient stage of D-SGD is lower bounded by $\Omega(n/(1-\beta)^2)$ in the heterogeneous case (Koloskova et al., 2020; Pu et al., 2021)

bias-correction methods (such as EXTRA (Shi et al., 2015), Exact-Diffusion (Yuan et al., 2018a,b), NIDS (Li et al., 2019b), and gradient tracking (Xu et al., 2015; Lorenzo and Scutari, 2016; Nedich et al., 2017; Qu and Li, 2018) – see (Alghunaim et al., 2021)), and dual acceleration (Scaman et al., 2017, 2018; Uribe et al., 2020). These algorithms, in the deterministic setting, can converge to the exact solution without any bias. On the other hand, decentralized stochastic methods (in which the gradient is noisy) have also gained a lot of attention recently. Decentralized SGD (D-SGD) has the same asymptotic linear speedup as P-SGD (Chen and Sayed, 2012; Sayed, 2014; Lian et al., 2017; Koloskova et al., 2020; Pu et al., 2021; Yuan et al., 2021) but with a more efficient single-iteration communication, it has been extensively studied in the context of large-scale machine learning (such as deep learning).

Data heterogeneity and network dependence. It is well-known that D-SGD is largely affected by gradient noise, but it was unclear how the data heterogeneity influences the performance of D-SGD. The work (Yuan et al., 2020) clarified that the error caused by the data heterogeneity can be greatly amplified when the network topology is sparse, which can even exceed the error caused by gradient noise. The works (Pu et al., 2021) and (Koloskova et al., 2020) also showed that the error term caused by data heterogeneity exhibits the poorest dependence on the network topology in D-SGD. It is thus conjectured that removing the influence of data heterogeneity might improve the topology dependence of D-SGD. Both the D^2/Exact-Diffusion algorithm and gradient tracking methods have been studied under stochastic settings in (Tang et al., 2018) and (Pu and Nedić, 2020; Xin et al., 2021; Lu et al., 2019; Zhang and You, 2019; Xin et al., 2022), respectively. However, the analysis in these works does not reveal whether the removal of data heterogeneity can improve the dependence on network topology. The work (Yuan et al., 2020) studied D^2/Exact-Diffusion in the *steady-state* (asymptotic) regime and for *strongly-convex* costs. Under this setting, (Yuan et al., 2020) showed that D^2/Exact-Diffusion has an improved network topology dependence by removing data heterogeneity, but it is unclear whether the improved steady-state performance in (Yuan et al., 2020) carries over to D-SGD's *non-asymptotic performance* (convergence rate). This paper clarifies the improvement in the *non-asymptotic* convergence rate in D^2/Exact-Diffusion for *both* generally and strongly convex scenarios. These results demonstrate that removing influence of data heterogeneity improves the dependence on network topology for D-SGD. In addition, the established dependence on network topology for D^2/Exact-Diffusion in the strongly-convex scenario aligns with lower bound of D-SGD with *homogeneous* dataset.

Transient stage. As to the transient stage of D-SGD, the work (Pu et al., 2021) shows that it is $O(n/(1-\beta)^2)$ for strongly-convex settings, and the results from (Koloskova et al., 2020) imply that it is $O(n^3/(1-\beta)^4)$ for generally-convex settings. In comparison, we establish that D^2/Exact-Diffusion has an improved $O(n/(1-\beta))$ and $O(n^3/(1-\beta)^2)$ transient stage for strongly and generally convex scenarios, respectively. This work does not study the transient stage of D^2/Exact-Diffusion for the non-convex scenario as the current analysis cannot be directly extended to such settings. Note that (Tang et al., 2018) and (Xin et al., 2021) provide an $O(n^3/(1-\beta)^6)$ transient stage for D^2/Exact-Diffusion and stochastic gradient tracking under the non-convex setting, respectively. These transient analysis results, however, are worse than D-SGD in terms of network topology dependence.

There are some recent works (Yuan et al., 2021; Yu et al., 2019; Lin et al., 2021) that aim to alleviate the influence of data heterogeneity on decentralized stochastic *momentum* SGD, but they do not show an improved dependence on network topology.

Multi-round gossip communication. Multi-round gossip communication has been utilized in recent works to boost the performance of decentralized algorithms. For example, (Berahas et al., 2018) employs multi-round gossip to balance communication and computation burdens, (Scaman et al., 2017) develops an optimal decentralized algorithm based upon multi-round gossip for smooth and strongly convex problems in the deterministic scenario, and (Li et al., 2020) achieves near-optimal communication complexity with multi-round gossip and increasing penalty parameters. In decentralized and stochastic optimization, recent works (Lu and De Sa, 2021; Yuan et al., 2022; Huang and Yuan, 2022) employ multi-round gossip communication and gradient accumulation on stochastic gradient tracking to significantly improve the convergence rate in non-convex settings. However, these works do not demonstrate how multi-round gossip communication can improve the network topology dependence in the strongly and generally convex scenarios.

Parallel works. Simultaneously and independently[2], a parallel work (Huang and Pu, 2021) has established a result under strongly-convex scenarios that is partially similar to this work. However, it does not study the generally-convex scenario. Another parallel work (Koloskova et al., 2021) demonstrates that gradient tracking has an improved network topology dependence that nearly-matches D^2/Exact-Diffusion (up to a $\log(1-\beta)$ term)[3]. This implies D^2/Exact-Diffusion has a slightly better theoretical dependence on network topology. Moreover, D^2/Exact-Diffusion is more communication-efficient than gradient tracking since it only requires one communication round per iteration. A detailed comparison with (Koloskova et al., 2021) is listed in Table 3. We note that, unlike (Huang and Pu, 2021) and (Koloskova et al., 2021), additional results on lower bound of homogeneous D-SGD and transient stage on D^2/Exact-Diffusion with multi-round gossip communication are also established in our work. Moreover, our analysis techniques are also different from (Huang and Pu, 2021) and (Koloskova et al., 2021).

1.4 Notations

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

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

which collect all local variables/gradients across the network. Note that $\nabla f(\mathbf{x}^{(k)}) = \mathbb{E}[\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})]$. 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 $a_1, \ldots, a_n$. 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. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we let $\lambda_i(A)$ to

2. (Huang and Pu, 2021) appeared online on May 11, 2021 and the first version of our work appeared online on May 17, 2021.
3. (Koloskova et al., 2021) appeared online around November 15, 2021.

be the ith largest eigenvalue and $\rho(A) = \max_i |\lambda_i(A)|$ denote the spectral radius of matrix A. In addition, we let $[n] = \{1, \ldots, n\}$ for any positive integer n. Suppose that $A \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix with eigen-decomposition $A = U\Lambda U^T$ where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{n\times n}$ is a non-negative diagonal matrix. Then, we let $A^{\frac{1}{2}} = U\Lambda^{\frac{1}{2}}U^T \in \mathbb{R}^{n\times n}$ be the square root of the matrix A. Note that $A^{\frac{1}{2}}$ is also positive semidefinite and $A^{\frac{1}{2}} \times A^{\frac{1}{2}} = A$. For a vector a, we let $\|a\|$ denote its ℓ_2 norm. For a matrix A, we let $\|A\|$ denote its ℓ_2 norm and $\|A\|_F$ denote its Frobenius norm. We use $\lesssim$ and $\gtrsim$ to indicate inequalities that hold by omitting absolute constants. We define $\mathbb{N}$ as the set of non-negative integers.

2. Preliminaries and Assumptions

2.1 Weight Matrix

To model the decentralized communication, we let $w_{ij} \geq 0$ be the weight used by node i to scale information flowing from node j to node i. We use $\mathcal{N}_i$ to denote the neighbors of node i, including node i itself. We let $w_{ij} = 0$ if node $j \notin \mathcal{N}_i$. We define the weight matrix $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ and assume the following condition.

Assumption 1 (Weight matrix**)** *The network is strongly connected and the weight matrix W is doubly stochastic and symmetric,* i.e., *$W = W^T$ and $W\mathbb{1}_n = \mathbb{1}_n$.* □

Remark 1 (Spectral gap**)** *Under Assumption 1, it holds that $1 = \lambda_1(W) > \lambda_2(W) \geq \cdots \geq \lambda_n(W) > -1$. If we let $\beta = \rho(W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)$, it follows that $\beta = \max\{|\lambda_2(W)|, |\lambda_n(W)|\} \in (0, 1)$. The network quantity $1 - \beta$ is called the spectral gap of W, which reflects the connectivity of the network topology. The scenario $1 - \beta \to 1$ implies a well-connected topology (e.g., for fully connected topology, we can choose $W = \frac{1}{n}\mathbb{1}\mathbb{1}^T$ and hence $\beta = 0$). In contrast, the scenario $1 - \beta \to 0$ implies a badly-connected topology.* □

Assumption 2 (Minimum eigenvalue**)** *It is assumed that the minimum eigenvalue of W is bounded away from -1, i.e., there exists a constant $\epsilon > 0$ such that $\lambda_n(W) + 1 > \epsilon$. Moreover, we assume ϵ is independent of network size n and the spectral gap $1 - \beta$.* □

Remark 2 *It is easy to construct a weight matrix W satisfying Assumption 2. Given any weight matrix W' that satisfies Assumption 1, if we construct $W = (1 - \frac{\epsilon}{2})W' + \frac{\epsilon}{2}I$ for any desired $\epsilon > 0$, then it holds that $\lambda_n(W) + 1 > \epsilon$.* □

We introduce $\bar{W} = (I + W)/2$ in the D^2/Exact-Diffusion algorithm below. Under Assumption 2, it is easy to verify that the minimum eigenvalue of $\bar{W}$ is bounded away from 0, i.e., there exists a constant $\epsilon > 0$ such that $\lambda_n(\bar{W}) > \epsilon$. Moreover, constant ϵ is independent of n and $1 - \beta$.

2.2 D^2/Exact-Diffusion Algorithm

Recursions. Using the notation $\mathbf{x}^{(k)}$ and $\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})$ introduced in Sec. 1.4, the D^2 algorithm (Tang et al., 2018), also known as Exact-Diffusion in (Yuan et al., 2018a,b, 2020)

Algorithm 1: D^2/Exact-Diffusion

Require: Let $\bar{W} = (I+W)/2$ and initialize $x_i^{(0)} = 0$; let $\psi_i^{(0)} = x_i^{(0)}$.

for $k = 0, 1, 2, ...,$ every node i **do**

Sample $\xi_i^{(k)}$ and calculate $g_i^{(k)} = \nabla F(x_i^{(k)}; \xi_i^{(k)})$;

Update $\psi_i^{(k+1)} = x_i^{(k)} - \gamma g_i^{(k)}$; ▷ local gradient descent step

Update $\phi_i^{(k+1)} = \psi_i^{(k+1)} + x_i^{(k)} - \psi_i^{(k)}$; ▷ solution correction step

Update $x_i^{(k+1)} = \sum_{j\in\mathcal{N}_i} \bar{w}_{ij}\, \phi_j^{(k+1)}$; ▷ communication step

or NIDS in (Li et al., 2019a), can be expressed as

$$\mathbf{x}^{(k+1)} = \bar{W}\Big(2\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} - \gamma\big(\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) - \nabla F(\mathbf{x}^{(k-1)}; \boldsymbol{\xi}^{(k-1)})\big)\Big), \quad \forall\ k = 1, 2, \ldots \tag{5}$$

where $\bar{W} := (W+I)/2$ and γ is the learning rate (stepsize) parameter. The algorithm is initialized with $\mathbf{x}^{(1)} = \bar{W}\big(\mathbf{x}^{(0)} - \gamma\big(\nabla F(\mathbf{x}^{(0)}; \boldsymbol{\xi}^{(0)})\big)$ for any $\mathbf{x}^{(0)}$. Unlike the vanilla decentralized SGD (D-SGD), D^2/Exact-Diffusion exploits the last two consecutive iterates and stochastic gradients to update the current variable. This algorithm construction has been proven to remove the influence of data heterogeneity., see (Tang et al., 2018; Yuan et al., 2018b, 2020; Li et al., 2019b). Recursion (5) can be conducted in a decentralized manner as listed in Algorithm 1 (see (Yuan et al., 2018a, 2020) for details). A fundamental difference from D-SGD lies in the solution correction step. If $x_i^{(k)} - \psi_i^{(k)}$ is removed, then Algorithm 1 reduces to D-SGD.

Primal-dual form. For analysis purposes, we rewrite recursion (5) into the following primal-dual equivalent form (Yuan et al., 2018b; Li et al., 2019b):

$$\begin{cases} \mathbf{x}^{(k+1)} = \bar{W}\big(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})\big) - V\mathbf{y}^{(k)}, \\ \mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} + V\mathbf{x}^{(k+1)}, \quad \forall\ k = 0, 1, 2, \ldots \end{cases} \tag{6}$$

where $\mathbf{y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]^T \in \mathbb{R}^{n\times d}$, $\boldsymbol{y}_i \in \mathbb{R}^d$, and $V = (I - \bar{W})^{1/2} \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix. For initialization, we let $\mathbf{y}^{(0)} = 0$. The equivalence between (5) and (6) can be easily verified, see (Yuan et al., 2018a, 2020).

Intuition to reduce transient stage. This paper proves that D^2/Exact-Diffusion has a shorter transient stage by removing the influence of data heterogeneity. To illustrate the intuition, we recall the convergence rate of D-SGD for generally-convex scenarios as

$$\frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E} f(\bar{x}^{(k)}) - f(x^\star)\big) = O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{2/3}}{(1-\beta)^{1/3}T^{2/3}} + \frac{b^{2/3}}{(1-\beta)^{2/3}T^{2/3}} + \frac{1}{(1-\beta)T}\Big)$$

where $b^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2$ indicates the data heterogeneity. It is evident that the existence of b^2 will result in a slower convergence and hence a longer transient stage. If the influence of b^2 can be removed from the convergence rate, which can be achieved by D^2/Exact-Diffusion due to their robustness to data heterogeneity (Yuan et al., 2018a; Li et al., 2019b), then the convergence rate should be improved and the transient stage should reduce. Indeed, we will prove this argument rigorously in the sections below.

2.3 Optimality Condition

The primal-dual recursion (6) facilitates the following optimality condition of problem (1).

Lemma 3 (Optimality Condition**)** *Assume each cost function $f_i(x)$ in problem (1) is convex. Then, there exists some primal-dual pair $(\mathbf{x}^\star, \mathbf{y}^\star)$, in which $\mathbf{y}^\star$ lies in the range space of V, that satisfies*

$$\gamma \bar{W} \nabla f(\mathbf{x}^\star) + V\mathbf{y}^\star = 0, \tag{7a}$$

$$V\mathbf{x}^\star = 0, \tag{7b}$$

and it holds that $\boldsymbol{x}_1^\star = \cdots = \boldsymbol{x}_n^\star = \boldsymbol{x}^\star$ where $\boldsymbol{x}^\star$ is a global solution to problem (1). ■

The proof of the above lemma is simple and can be referred to (Shi et al., 2015, Lemma 3.1). It is worth noting that when there is no gradient noise, *i.e.*, $\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) = \nabla f(\mathbf{x}^{(k)})$, the fixed point $(\mathbf{x}^o, \mathbf{y}^o)$ of the primal-dual recursion (6) satisfies the optimality condition (7a)–(7b). This implies the iterates $\boldsymbol{x}^{(k)}$ generated by the D^2/Exact-Diffusion algorithm will converge to a global solution $\boldsymbol{x}^\star$ of problem (1) in expectation. Such conclusion holds without any assumption whether the data distribution is homogeneous or not.

2.4 D-SGD and Data-Heterogeneity Bias

The (adapt-then-combine) D-SGD algorithm (also known as diffusion in (Lopes and Sayed, 2008; Chen and Sayed, 2012; Sayed, 2014)) (Nedic and Ozdaglar, 2009; Lopes and Sayed, 2008; Chen and Sayed, 2012; Lian et al., 2017) has the update form:

$$\mathbf{x}^{(k+1)} = W\big(\mathbf{x}^{(k)} - \gamma \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)})\big). \tag{8}$$

Without any auxiliary variable to help correct the gradient direction like D^2/Exact-Diffusion recursion (6), D-SGD suffers from a solution deviation caused by data heterogeneity even if there is no gradient noise. Given that data heterogeneity exists, we have $b^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^\star)\|^2 > 0$, which implies there exists at least one node i such that $\nabla f_i(x^\star) \neq 0$. Suppose there is no gradient noise and $\mathbf{x}^{(k)}$ is initialized as a consensual solution, represented by $\mathbf{x}^\star = [\boldsymbol{x}^\star, \ldots, \boldsymbol{x}^\star]^T$ where $\boldsymbol{x}^\star$ is a solution to problem (1), which satisfies $\mathbb{1}_n^T \nabla f(\mathbf{x}^\star) = 0$. Following recursion (8), we have

$$\mathbf{x}^{k+1} = W\big(\mathbf{x}^\star - \gamma \nabla f(\mathbf{x}^\star)\big) = \mathbf{x}^\star - \gamma W \nabla f(\mathbf{x}^\star) \neq \mathbf{x}^\star, \tag{9}$$

in which the last inequality holds because $W\nabla f(\mathbf{x}^\star) \neq 0$ in general. Relation (9) implies that even if D-SGD starts from the optimal consensual solution, it can still jump away to a biased solution due to the data heterogeneity effect.

2.5 Assumptions

We now introduce some standard assumptions that will be used throughout the paper.

Assumption 3 (Convexity**)** *Each cost function $f_i(x)$ is convex.*

Assumption 4 (Smoothness**)** *Each local cost function $f_i(x)$ is differentiable, and there exists a constant L such that for each $x, y \in \mathbb{R}^d$*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|, \quad \forall\, i \in [n]. \tag{10}$$

Assumption 5 (Gradient noise**)** *Define the filtration $\mathcal{F}^{(k)} = \big\{\{\xi_i^{(k)}\}_{i=1}^n, \{x_i^{(k)}\}_{i=1}^n, \ldots, \{\xi_i^{(0)}\}_{i=1}^n, \{x_i^{(0)}\}_{i=1}^n\big\}$. Then, it is assumed that for any k and i that*

$$\mathbb{E}[\nabla F(x_i^{(k)}; \xi_i^{(k)}) - \nabla f_i(x_i^{(k)}) | \mathcal{F}^{(k-1)}] = 0,$$
$$\mathbb{E}[\|\nabla F(x_i^{(k)}; \xi_i^{(k)}) - \nabla f_i(x_i^{(k)})\|^2 | \mathcal{F}^{(k-1)}] \le \sigma^2,$$

for some $\sigma^2 \ge 0$. Moreover, we assume $\xi_i^{(k)}$ are independent of each other for any k and i.

3. Fundamental Transformation

In this section we transform the primal-dual update of the D^2/Exact-Diffusion algorithm (6) into another equivalent recursion. This transformation, inspired by (Yuan et al., 2018b, 2020), is fundamental to establish the refined convergence results for D^2/Exact-Diffusion.

Let $(\mathbf{x}^\star, \mathbf{y}^\star)$ be one pair of variables that satisfies the optimality conditions in Lemma 3. Subtracting (7a) and (7b) from the primal-dual recursion (6), we have

$$\mathbf{x}^{(k+1)} - \mathbf{x}^\star = \bar{W}\big(\mathbf{x}^{(k)} - \mathbf{x}^\star - \gamma(\nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) - \nabla f(\mathbf{x}^\star))\big) - V(\mathbf{y}^{(k)} - \mathbf{y}^\star), \tag{12a}$$
$$\mathbf{y}^{(k+1)} - \mathbf{y}^\star = \mathbf{y}^{(k)} - \mathbf{y}^\star + V(\mathbf{x}^{(k+1)} - \mathbf{x}^\star). \tag{12b}$$

To simplify the notation, we define $\mathbf{s}^{(k)} := \nabla F(\mathbf{x}^{(k)}; \boldsymbol{\xi}^{(k)}) - \nabla f(\mathbf{x}^{(k)})$ as the gradient noise at iteration k. By substituting (12a) into (12b), and recalling that $I - V^2 = \bar{W}$, we obtain

$$\begin{bmatrix} \mathbf{x}^{(k+1)} - \mathbf{x}^\star \\ \mathbf{y}^{(k+1)} - \mathbf{y}^\star \end{bmatrix} = \underbrace{\begin{bmatrix} \bar{W} & -V \\ V\bar{W} & \bar{W} \end{bmatrix}}_{B} \begin{bmatrix} \mathbf{x}^{(k)} - \mathbf{x}^\star \\ \mathbf{y}^{(k)} - \mathbf{y}^\star \end{bmatrix} - \gamma \begin{bmatrix} \bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \\ V\bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \end{bmatrix}. \tag{13}$$

The main difficulty analyzing the convergence of the above recursion (as well as (12a)–(12b)) is that terms $\mathbf{x}^{(k)} - \mathbf{x}^\star$ and $\mathbf{y}^{(k)} - \mathbf{y}^\star$ are entangled together to update $\mathbf{x}^{(k+1)} - \mathbf{x}^\star$ or $\mathbf{y}^{(k+1)} - \mathbf{y}^\star$. For example, the update of $\mathbf{x}^{(k+1)} - \mathbf{x}^\star$ relies on both $\bar{W}(\mathbf{x}^{(k)} - \mathbf{x}^\star)$ and $-V(\mathbf{y}^{(k)} - \mathbf{y}^\star)$. In the following, we identify a change of basis and transform (13) into another equivalent form so that the involved iterated variables can be "decoupled". To this end, we need to introduce a fundamental decomposition lemma. This lemma was first established in (Yuan et al., 2018b). We have improved this lemma by establishing an upper bound of an important term (see (17)) that is critical for our later analysis.

Lemma 4 (Fundamental decomposition**)** *Under Assumption 1, the matrix $B \in \mathbb{R}^{2n\times 2n}$ in* (13) *can be diagonalized as*

$$B = \underbrace{[r_1\ r_2\ cX_R]}_{X} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & D_1 \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} \ell_1^T \\ \ell_2^T \\ X_L/c \end{bmatrix}}_{X^{-1}} \tag{14}$$

for any constant $c > 0$ where $D \in \mathbb{R}^{2n\times 2n}$ is a diagonal matrix. Moreover, we have

$$r_1 = \begin{bmatrix} \mathbb{1}_n \\ 0 \end{bmatrix}, \quad r_2 = \begin{bmatrix} 0 \\ \mathbb{1}_n \end{bmatrix}, \quad \ell_1 = \begin{bmatrix} \frac{1}{n}\mathbb{1}_n \\ 0 \end{bmatrix}, \quad \ell_2 = \begin{bmatrix} 0 \\ \frac{1}{n}\mathbb{1}_n \end{bmatrix} \tag{15}$$

and $X_R \in \mathbb{R}^{2n\times 2(n-1)}$, $X_L \in \mathbb{R}^{2(n-1)\times 2n}$. Also, the matrix D_1 is a diagonal matrix with diagonal entries strictly less than 1 in magnitude and

$$\|D_1\| = \bar{\lambda}_2^{1/2}, \quad \textit{where} \quad \bar{\lambda}_2 = \frac{1+\lambda_2(W)}{2}. \tag{16}$$

Furthermore, it holds that

$$\|X_L\|\|X_R\| \le \bar{\lambda}_n^{-1/2} \tag{17}$$

where $\bar{\lambda}_n = (1+\lambda_n(W))/2$. (Proof is in Appendix B.1). ■

Left-multiplying X^{-1} to both sides of (13) and using Lemma 4, we can get the transformed recursion given in Lemma 5. Note that Lemma 5 is also an improved version of (Yuan et al., 2020, Lemma 3), which will facilitate our sharper convergence analysis of the D^2/Exact-Diffusion algorithm.

Lemma 5 (Transformed Recursion**)** *Under Assumptions 1, the D^2/Exact-Diffusion error recursion* (13) *can be transformed into*

$$\begin{bmatrix} \bar{\boldsymbol{z}}^{(k+1)} \\ \check{\mathbf{z}}^{(k+1)} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{z}}^{(k)} - \frac{\gamma}{n}\mathbb{1}^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)) \\ D_1\check{\mathbf{z}}^{(k)} - \frac{\gamma}{c}\check{\mathbf{g}}^{(k)} \end{bmatrix} - \gamma \begin{bmatrix} \bar{\mathbf{s}}^{(k)} \\ \frac{1}{c}\check{\mathbf{s}}^{(k)} \end{bmatrix}, \tag{18}$$

where $\check{\mathbf{g}}^{(k)}, \bar{\mathbf{s}}^{(k)}$ and $\check{\mathbf{s}}^{(k)}$ are defined as

$$\check{\mathbf{g}}^{(k)} \triangleq \big(X_{L,\ell}Q_R + X_{L,r}Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\big)\bar{\Lambda}_R Q_R^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)) \tag{19a}$$

$$\bar{\mathbf{s}}^{(k)} \triangleq \frac{1}{n}\mathbb{1}^T\mathbf{s}^{(k)} \tag{19b}$$

$$\check{\mathbf{s}}^{(k)} \triangleq \big(X_{L,\ell}Q_R + X_{L,r}Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\big)\bar{\Lambda}_R Q_R^T\mathbf{s}^{(k)}. \tag{19c}$$

Here, the matrices $X_{L,\ell}, X_{L,r} \in \mathbb{R}^{2(n-1)\times n}$ are the left and right part of the matrix $X_L = [X_{L,\ell} \;\; X_{L,r}]$, respectively, $\bar{\Lambda}_R = \mathrm{diag}\{\bar{\lambda}_2(W), \ldots, \bar{\lambda}_n(W)\} \in \mathbb{R}^{(n-1)\times(n-1)}$, and $Q_R \in \mathbb{R}^{n\times(n-1)}$ is defined in (59)*. The relation between the original and the transformed error vectors are*

$$\begin{bmatrix} \bar{\boldsymbol{z}}^{(k)} \\ \check{\mathbf{z}}^{(k)} \end{bmatrix} = \begin{bmatrix} \ell_1^T \\ X_L/c \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(k)} - \mathbf{x}^\star \\ \mathbf{y}^{(k)} - \mathbf{y}^\star \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} \mathbf{x}^{(k)} - \mathbf{x}^\star \\ \mathbf{y}^{(k)} - \mathbf{y}^\star \end{bmatrix} = \begin{bmatrix} r_1 \;\; cX_R \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{z}}^{(k)} \\ \check{\mathbf{z}}^{(k)} \end{bmatrix}. \tag{20}$$

Note that $\bar{\boldsymbol{z}}^{(k)} \in \mathbb{R}^{1\times d}$ and $\check{\mathbf{z}}^{(k)} \in \mathbb{R}^{2(n-1)\times d}$ (Proof is in Appendix B.2). ■

Remark 6 (recursion interpretation**)** *Using the left relation in* (20) *and the definition of ℓ_1 in* (15)*, it holds that*

$$[\bar{\boldsymbol{z}}^{(k)}]^T = \frac{1}{n}\mathbb{1}^T(\mathbf{x}^{(k)} - \mathbf{x}^\star) = \bar{x}^{(k)} - x^\star.$$

Therefore, $\bar{\boldsymbol{z}}^{(k)}$ *gauges the distance between the averaged variable* $\bar{x}^{(k)}$ *and the solution* $x^\star$*. On the other hand, using the right relation in* (20) *and the definition of* r_1 *in* (15)*, it holds that* $\mathbf{x}^{(k)} - \mathbf{x}^\star = \mathbb{1}_n \bar{\boldsymbol{z}}^{(k)} + cX_{R,u}\check{\mathbf{z}}^{(k)} = \bar{\mathbf{x}}^{(k)} - \mathbf{x}^\star + cX_{R,u}\check{\mathbf{z}}^{(k)}$*, where* $X_{R,u} \in \mathbb{R}^{n\times 2(n-1)}$ *is the upper part of matrix* $X_R = [X_{R,u}; X_{R,d}]$*. This implies*

$$cX_{R,u}\check{\mathbf{z}}^{(k)} = \mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}. \tag{21}$$

Hence, $\check{\mathbf{z}}^{(k)}$ *measures the consensus error,* i.e.*, the distance between* $\mathbf{x}^{(k)}$ *and* $\bar{\mathbf{x}}^{(k)}$*.* □

The following proposition establishes the magnitude of $\|\check{\mathbf{z}}^{(0)}\|^2$, which will be used in later derivations.

Proposition 7 (The magnitude of $\|\check{\mathbf{z}}^{(0)}\|_F^2$) *If we initialize* $\mathbf{x}^{(0)} = 0$*,* $\mathbf{y}^{(0)} = 0$ *and set* $c = \|X_L\|$*, it holds that* $\|\check{\mathbf{z}}^{(0)}\|_F^2 \le \frac{\gamma^2\bar{\lambda}_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2}{1-\bar{\lambda}_2}$*. If we further assume* $\|\nabla f(\mathbf{x}^\star)\|_F^2 = \sum_{i=1}^n \|\nabla f_i(x^\star)\|^2 = O(n)$*, it follows that* $\|\check{\mathbf{z}}^{(0)}\|_F^2 = O(\frac{n\gamma^2\bar{\lambda}_2^2}{1-\bar{\lambda}_2})$ *(Proof is in Appendix B.3).* ■

4. Convergence Results: Generally-Convex Scenario

With Assumption 5, it is easy to verify that

$$\mathbb{E}[\|\bar{\mathbf{s}}^{(k)}\|^2 | \mathcal{F}^{(k-1)}] \le \frac{\sigma^2}{n}, \quad k = 1, 2, \ldots \tag{22}$$

We first establish a descent lemma for the D^2/Exact-Diffusion algorithm in the generally-convex setting, which describes how $\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2$ evolves with iteration.

Lemma 8 (Descent Lemma) *Under Assumptions 3–5 and learning rate* $\gamma < \frac{1}{4L}$*, it holds for* $k = 0, 1, \ldots$ *that*

$$\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2 \le \mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2 - \gamma\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) + \frac{3L\gamma}{2n\bar{\lambda}_n}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma^2\sigma^2}{n}, \quad k = 0, 1, \ldots \tag{23}$$

where $\bar{\lambda}_n = \lambda_n(\bar{W}) = (1 + \lambda_n(W))/2$*. (Proof is in Appendix C.1)* ■

With inequality (23), we have for $T \ge 0$ that

$$\frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) \le \frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma(T+1)} + \frac{3L}{2n\bar{\lambda}_n(T+1)}\sum_{k=0}^{T}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n}. \tag{24}$$

We next bound the ergodic consensus term on the right-hand-side.

Lemma 9 (Consensus Lemma) *Under Assumptions 1-5 and learning rate* $\gamma \le \frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{4\bar{\lambda}_2 L}$*, it holds for any* $k = 0, 1, \cdots$ *that*

$$\mathbb{E}\|\check{\mathbf{z}}^{(k+1)}\|_F^2 \le \left(\frac{1+\beta_1}{2}\right)\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{16n\gamma^2\bar{\lambda}_2^2 L}{1-\beta_1}(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)) + 4n\gamma^2\bar{\lambda}_2^2\sigma^2, \tag{25}$$

where $\beta_1 = \bar{\lambda}_2^{1/2}$*,* $\bar{\lambda}_2 = (1+\lambda_2(W))/2$*, and* $\bar{\lambda}_n = (1+\lambda_n(W))/2$ *(Proof is in Appendix C.2).* ■

Lemma 10 (ERGODIC CONSENSUS LEMMA**)** *Under the same assumptions as Lemma 9, it holds that (Proof is in Appendix C.3)*

$$\begin{aligned}&\frac{1}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2\\ &\le \frac{32n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)^2(T+1)}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)+\frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}+\frac{3\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2}{(1-\beta_1)(T+1)}. \qquad (26)\end{aligned}$$

■

With inequalities (24) and (26), and the fact that $\beta_1^2=\bar{\lambda}_2=(1+\lambda_2(W))/2\le(1+\beta)/2$ where $\beta=\rho(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)$ is defined in Remark 1, we can show the following convergence result for D^2/Exact-Diffusion in the generally convex scenario.

Theorem 11 (CONVERGENCE PROPERTY**)** *Under Assumptions 1-5 and learning rate*

$$\gamma=\min\left\{\frac{1}{4L},\frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{10L\bar{\lambda}_2},\left(\frac{r_0}{r_1(T+1)}\right)^{\frac{1}{2}},\left(\frac{r_0}{r_2(T+1)}\right)^{\frac{1}{3}},\left(\frac{r_0}{r_3}\right)^{\frac{1}{3}}\right\}$$

where r_0, r_1, r_2 and r_2 are constants defined in (83)*, then it holds that*

$$\frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)=O\Big(\frac{\sigma}{\sqrt{nT}}+\frac{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{(1-\beta)^{\frac{1}{3}}T^{\frac{2}{3}}}+\frac{L^{\frac{1}{3}}}{(1-\beta)^{\frac{2}{3}}T}+\frac{L}{(1-\beta)T}\Big).$$

(Proof is in Appendix C.4) ■

Remark 12 *If we pay attentions to how σ, n, T, and β influence the convergence rate, the result in Theorem 11 can be simplified as*

$$\frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)=O\Big(\frac{\sigma}{\sqrt{nT}}+\frac{\sigma^{\frac{2}{3}}}{(1-\beta)^{\frac{1}{3}}T^{\frac{2}{3}}}+\frac{1}{(1-\beta)T}\Big). \qquad (27)$$

Corollary 13 (TRANSIENT STAGE**)** *Under the same assumptions as Theorem 11, the transient stage for D^2/Exact-Diffusion is on the order of $O\left(\frac{n^3}{(1-\beta)^2}\right)$.*

Proof Transient stage is defined as the *minimum* number of iterations that an algorithm has to experience to achieve the linear speedup stage. With this definition and the derived convergence rate in (27), the transient stage for D^2/Exact-Diffusion is given by

$$T_{\text{tran}}\le\min\left\{\quad T\in\mathbb{N}\quad\middle|\quad\frac{\sigma^{\frac{2}{3}}}{(1-\beta)^{\frac{1}{3}}T^{\frac{2}{3}}}\lesssim\frac{\sigma}{\sqrt{nT}}\quad\text{and}\quad\frac{1}{(1-\beta)T}\lesssim\frac{\sigma}{\sqrt{nT}}\quad\right\}. \qquad (28)$$

The inequalities in (28) exist to guarantee that the linear speedup term dominates convergence rate (27). With (28), we have $T_{\text{tran}}\lesssim\max\big\{\frac{n^3}{(1-\beta)^2\sigma^2},\frac{n}{(1-\beta)^2\sigma^2}\big\}=O\big(\frac{n^3}{(1-\beta)^2}\big)$. ■

Algorithm	Convergence rate	Transient stage
G-T	$O\Big(\frac{\sigma}{\sqrt{nT}}+\frac{\tau^{1/3}\sigma^{2/3}}{T^{2/3}}+\frac{\tau}{T}\Big)$	$O\Big(\frac{n^3}{(1-\beta)^2}\log^2(\frac{1}{1-\beta}(1+\log(\frac{1}{1-\beta})))\Big)$
E-D (Ours)	$O\Big(\frac{\sigma}{\sqrt{nT}}+\frac{\sigma^{2/3}}{(1-\beta)^{1/3}T^{2/3}}+\frac{1}{(1-\beta)T}\Big)$	$O\Big(\frac{n^3}{(1-\beta)^2}\Big)$

Table 3: Convergence rate and transient stage comparison in the generally-convex scenario between our results (D^2/Exact-Diffusion, D^2/E-D for short) and a parallel work on gradient tracking (G-T for short) (Koloskova et al., 2021). In the table, $\tau=\frac{2}{1-\beta}\log(\frac{50}{1-\beta}(1+\log(\frac{1}{1-\beta})))+1$ is in (Koloskova et al., 2021, Lemma 20).

Remark 14 *The result from (Koloskova et al., 2020) indicates that the transient stage of D-SGD for the generally convex scenario is on the order of $O\big(\frac{n^3}{(1-\beta)^4}\big)$. By removing the influence of the data heterogeneity, D^2/Exact-Diffusion improves the transient stage to $O\big(\frac{n^3}{(1-\beta)^2}\big)$, which has a better network topology dependence on $1-\beta$.* □

Remark 15 *An independent and parallel work in (Koloskova et al., 2021) shows that gradient tracking can also improve the transient stage of D-SGD. The convergence rate and transient stage in the generally-convex scenario established in (Koloskova et al., 2021) are listed in Table 3. It is observed that the transient stage of D^2/Exact-Diffusion is better than gradient tracking by a factor $\log^2(\frac{1}{1-\beta}(1+\log(\frac{1}{1-\beta})))$. Moreover, D^2/Exact-Diffusion is more communication-efficient than gradient tracking since it only requires one communication round per iteration.*

5. Convergence results: Strongly-Convex Scenario

5.1 Convergence Analysis of D^2/Exact-Diffusion

In this subsection we establish the convergence rate of D^2/Exact-Diffusion in the strongly convex scenario and examine its transient stage.

Assumption 6 (strongly convex) *Each $f_i(x)$ is strongly convex,* i.e.*, there exists a constant $\mu>0$ such that for any $x,y\in\mathbb{R}^d$ we have:*

$$f_i(y)\geq f_i(x)+\langle\nabla f_i(x),y-x\rangle+\frac{\mu}{2}\|y-x\|^2,\quad\forall\ i\in[n].$$

□

Lemma 16 (Descent Lemma) *When Assumptions 4–6 hold and learning rate $\gamma<\frac{1}{4L}$, it holds for $k=0,1,\ldots$ that*

$$\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2\leq\Big(1-\frac{\gamma\mu}{2}\Big)\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2-\gamma\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)+\frac{5L\gamma}{2n\bar{\lambda}_n}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2+\frac{\gamma^2\sigma^2}{n}.\quad(29)$$

(Proof is in Appendix D.1) ■

With inequality (29), we have

$$\mathbb{E}f(\bar{x}^{(k)})-f(x^\star) \le \Big(1-\frac{\gamma\mu}{2}\Big)\frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2}{\gamma} - \frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2}{\gamma} + \frac{5L}{2n\bar{\lambda}_n}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n}.$$

If we take the uniform average for both sides over $k=0,\ldots,T$, the term $(1-\frac{\gamma\mu}{2})\frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2}{\gamma}$ from the kth iteration cannot cancel the term $-\frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2}{\gamma}$ from the $(k+1)$th iteration. Inspired by (Stich, 2019a), we instead take the weighted average over $k=0,\ldots,T$ so that

$$\begin{aligned}
&\frac{1}{H_T}\sum_{k=0}^{T} h_k\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)\\
&\le \frac{1}{H_T}\sum_{k=0}^{T} h_k\Big(\frac{(1-\frac{\gamma\mu}{2})\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2}{\gamma} - \frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2}{\gamma}\Big) + \frac{5L}{2nH_T\bar{\lambda}_n}\sum_{k=0}^{T} h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n},
\end{aligned}$$

where $h_k \ge 0$ is some weight to be determined, and $H_T = \sum_{k=0}^{T} h_k$. If we let $h_k = (1-\frac{\gamma\mu}{2})h_{k+1}$ for $k=0,1,\ldots$, the above inequality becomes

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big) \le \frac{h_0\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{H_T\gamma} + \frac{5L}{2nH_T\bar{\lambda}_n}\sum_{k=0}^{T} h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n}. \quad (30)$$

We next bound the ergodic consensus term in the right-hand-side.

Lemma 17 (ERGODIC CONSENSUS LEMMA) *Under Assumptions 1, 4, 5, and 6 and if learning rate satisfies $\gamma \le \frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{4L\lambda_2}$, then it holds that*

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 \le \frac{4Ch_0}{H_T(1-\beta_1)} + \frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1} + \frac{128n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)^2H_T}\sum_{k=0}^{T} h_k\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big), \quad (31)$$

where $C=\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2$, the positive weights $\{h_k\}_{k=0}^{\infty}$ satisfy

$$h_k \le h_\ell\left(1+\frac{1-\beta_1}{4}\right)^{k-\ell} \quad \text{for any } k\ge 0 \text{ and } 0\le\ell\le k, \quad (32)$$

and $H_T=\sum_{k=0}^{T} h_k$. (Proof is in Appendix D.2) ■

With inequalities (30) and (31), and the fact that $\beta_1^2=\bar{\lambda}_2=(1+\lambda_2(W))/2\le(1+\beta)/2$ with $\beta=\rho(W-\frac{1}{n}\mathbb{1}\mathbb{1}^T)$, we can establish the convergence property of D^2/Exact-Diffusion as follows.

Theorem 18 (CONVERGENCE PROPERTY) *Under Assumptions 1, 2, 4, 5, and 6, if*

$$\gamma=\min\left\{\frac{1}{4L}, \frac{1-\beta_1}{26L}\Big(\frac{\bar{\lambda}_n^{1/2}}{\bar{\lambda}_2}\Big), \frac{2\ln(2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2T^2/[\sigma^2(1-\beta)])}{\mu T}\right\},$$

then it holds that

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) = \tilde{O}\left(\frac{\sigma^2}{\mu nT} + \frac{L\sigma^2}{\mu^2(1-\beta)T^2} + \frac{L\exp\{-\frac{\mu(1-\beta)}{L}T\}}{(1-\beta)}\right) \quad (33)$$

where h_k and H_T are defined in Lemma 17. Notation $\tilde{O}(\cdot)$ hides logarithm terms. (Proof is in Appendix D.3) ■

Remark 19 *If we pay attentions to how σ, n, T, and β influence the convergence rate, expression* (33) *can be simplified as*

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) = \tilde{O}\left(\frac{\sigma^2}{nT} + \frac{\sigma^2}{(1-\beta)T^2} + \frac{1}{1-\beta}\exp\{-(1-\beta)T\}\right). \quad (34)$$

Corollary 20 (Transient stage) *Under assumptions in Theorem 18, the transient stage for D^2/Exact-Diffusion in the strongly convex scenario is on the order of $\tilde{O}(\frac{n}{1-\beta})$.*

Proof The third term (34) decays exponentially fast and hence can be ignored compared to the first two terms. As a result, the transient stage for D^2/Exact-Diffusion for strongly-convex problems is given by

$$T_{\text{tran}} \le \min\left\{ \quad T\in\mathbb{N} \quad \middle| \quad \frac{\sigma^2}{(1-\beta)T^2} \lesssim \frac{\sigma^2}{nT} \quad \right\}. \quad (35)$$

With (35), we have $T_{\text{tran}} \lesssim \frac{n}{1-\beta}$ and $T_{\text{tran}} = \tilde{O}\big(\frac{n}{1-\beta}\big)$. We use $\tilde{O}(\cdot)$ rather than $O(\cdot)$ because some logarithm factors are hidden inside. ■

Remark 21 *It is established in (Koloskova et al., 2020; Pu et al., 2021) that the transient stage of D-SGD for the strongly-convex scenario is on the order of $O(\frac{n}{(1-\beta)^2})$. By removing the influence of the data heterogeneity, D^2/Exact-Diffusion improves the transient stage to $\tilde{O}(\frac{n}{1-\beta})$ which has an improved dependence on $1-\beta$. This improved transient stage is consistent with those established in parallel works (Huang and Pu, 2021; Koloskova et al., 2021).* □

5.2 Transient Stage Lower Bound of the Homogeneous D-SGD

In Sec. 5.1, we showed that D^2/Exact-Diffusion, by removing the influence of data heterogeneity, can enhace the transient stage of D-SGD improving it from $O(\frac{n}{(1-\beta)^2})$ to $O(\frac{n}{1-\beta})$. In this section, we question what is the optimal transient stage of D-SGD when data distributions are homogeneous (*i.e., there is no influence of data heterogeneity*)? Can D-SGD have a better network topology dependence than D^2/Exact-Diffusion in certain scenarios? The answer reveals that D-SGD dependence on the network topology can match D^2/Exact-Diffusion only under the homogeneous setting and is always worse in heterogeneous setting. *In any case, D-SGD cannot be more robust to network topology than D^2/Exact-Diffusion.*

To this end, we let $\mathcal{F}_{\mu,L} = \{f : f \text{ is } \mu\text{-strongly convex, } L\text{-smooth with } \nabla f(x^\star) = 0\}$ with some fixed global $x^\star$, $\mathcal{O}_{\sigma^2}$ be all possible gradient oracles with σ^2-bounded noise, $\mathcal{W}_\beta = \{W : W \text{ satisfies Assumption 1 and } \rho(W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T) \le \beta\}$, and $\mathcal{A}$ consists of D-SGD algorithms with all possible hyper-parameter choices (such as learning rate γ) but with homogeneous dataset, then we consider the following minimax lower bound for D-SGD:

$$T_{\text{trans}}^{\text{dsgd}} = \min_{A\in\mathcal{A}} \max_{W\in\mathcal{W}_\beta} \max_{\mathcal{O}_{\sigma^2}} \max_{f_i\in\mathcal{F}_{\mu,L}} \{\text{Transient time of (A)}\}.$$

This definition implies that D-SGD cannot have a shorter transient stage than $T_{\text{trans}}^{\text{dsgd}}$ without further assumptions. Note that this lower bound only applies to vanilla D-SGD with single-round gossip communication. There might be algorithms (e.g., D-SGD with multi-round gossip communications per update) that enjoy provably shorter transient stage.

Theorem 22 (Lower Bound) *The transient time for D-SGD in the homogeneous scenario, i.e., $b^2 = 0$, is lower bounded by*

$$T_{\text{trans}}^{\text{dsgd}} = \tilde{\Omega}\left(\frac{n}{1-\beta}\right).$$

(Proof is in Appendix E) ■

From Corollary 20 and Theorem 22, we observe that the transient stage of D^2/Exact-Diffusion in the strongly-convex scenario coincides with the lower bound of homogeneous D-SGD in terms of the dependence on network topology (*i.e.*, the influence of β) and network size n. This implies that D-SGD has the same transient stage as D^2/Exact-Diffusion under the impractical homogeneous case and worse dependence in the heterogeneous case (Koloskova et al., 2020; Pu et al., 2021). Hence, the dependence of D^2/Exact-Diffusion on network topology is no worse than D-SGD and always better under the practical heterogeneous case.

6. D^2/Exact-Diffusion with Multi-Round Gossip

In this section, we will demonstrate how the use of multi-round gossip communication in D^2/Exact-Diffusion can further enhance its dependence on network topology. Motivated by (Lu and De Sa, 2021), we propose the multi-step D^2/Exact-Diffusion described in Algorithm 2. There are two fundamental differences between Algorithm 2 and the (vanilla) D^2/Exact-Diffusion listed in Algorithm 1: *gradient accumulation* and *fast gossip averaging*. The details in the fast gossip averaging, inspired by (Liu and Morse, 2011), are listed in Algorithm 3. Note that Algorithm 3 includes a damping (or interpolation) step in its output. This step is crucial for ensuring the convergence of D^2/Exact-Diffusion with multi-round gossip communication.

6.1 Fast Gossip Averaging

Using $\mathbf{z}^{(r)} = [z_1^{(r)}, \ldots, z_n^{(r)}]^T \in \mathbb{R}^{n\times d}$, the fast gossip average update (Algorithm 3) can be described by

$$\mathbf{z}^{(r+1)} = (1+\eta)W\mathbf{z}^{(r)} - \eta\mathbf{z}^{(r-1)}, \quad \textbf{for } r = 0, 1, \ldots, R-1 \tag{36a}$$

Algorithm 2: D^2/Exact-Diffusion with multiple gossip steps

Require: Initialize $x_i^{(0)}=0$, $\psi_i^{(0)}=x_i^{(0)}$, the rounds of gossip steps R, and damping ratio $\tau\in(0,1)$.

for $k=0,1,2,...,$ every node i **do**

Sample $\{\xi_i^{(k,r)}\}_{r=1}^R$ independently and let $g_i^{(k)}=\frac{1}{R}\sum_{r=1}^R\nabla F(x_i^{(k)};\xi_i^{(k,r)})$; ▷ grad.accu.

Update $\psi_i^{(k+1)}=x_i^{(k)}-\gamma g_i^{(k)}$; ▷ local gradient descent step

Update $\phi_i^{(k+1)}=\psi_i^{(k+1)}+x_i^{(k)}-\psi_i^{(k)}$; ▷ solution correction step

Update $x_i^{(k+1)}=\textbf{FastGossipAverage}(\{\phi_i^{(k+1)}\}_{i=1}^n,W,R,\tau)$; ▷ multiple gossips

Algorithm 3: $x_i=\textbf{FastGossipAverage}(\{\phi_i\}_{i=1}^n,W,R,\tau)$

Require: $\{\phi_i\}_{i=1}^n$, W, R, $\tau\in(0,1)$; let $z_i^{(0)}=z_i^{(-1)}=\phi_i$ and $\eta=\frac{1-\sqrt{1-\beta^2}}{1+\sqrt{1+\beta^2}}$.

for $r=0,1,2,...,R-1,$ every node i **do**

Update $z_i^{(r+1)}=(1+\eta)\sum_{j\in\mathcal{N}_i}w_{ij}z_i^{(r)}-\eta z_i^{(r-1)}$; ▷ fast gossip averaging

Output: $x_i=(1-\tau)z_i^{(R)}+\tau z_i^{(0)}$; ▷ damping step

$$\mathbf{x}=(1-\tau)\mathbf{z}^{(R)}+\tau\mathbf{z}^{(0)}. \tag{36b}$$

Since $\mathbf{z}^{(-1)}=\mathbf{z}^{(0)}$, it holds from (36a)–(36b) that $\mathbf{z}^{(r)}=M^{(r)}\mathbf{z}^{(0)}$ where $M^{(r)}\in\mathbb{R}^{n\times n}$ is defined by:

$$\begin{aligned}&M^{(-1)}=M^{(0)}=I\\&M^{(r+1)}=(1+\eta)WM^{(r)}-\eta M^{(r-1)},\quad\textbf{for } r=0,1,\ldots,R-1.\end{aligned} \tag{37}$$

Since W is symmetric and doubly stochastic (Assumption 1), the matrix $M^{(r)}$ is also symmetric and doubly stochastic for each $r=0,\ldots,R$. Furthermore, the following result holds.

Proposition 23 *Under Assumption 1, it holds that*

$$M^{(r)}=(M^{(r)})^T,\quad M^{(r)}\mathbb{1}=\mathbb{1},\quad \textit{and}\quad \rho(M^{(r)}-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\le\sqrt{2}\Big(1-\sqrt{1-\beta}\Big)^r.$$

where $\beta=\max\{|\lambda_2(W)|,|\lambda_n(W)|\}$. Therefore, the non-unit eigenvalues of $M^{(r)}$ vanish to zero as $r\to\infty$. (Proof is in Appendix F.1). ■

When the rounds of gossip steps R are sufficiently large, we can achieve the following important proposition:

Proposition 24 *We let $\bar{M}=(1-\tau)M^{(R)}+\tau I$. If Assumption 1 holds and $R=\lceil\frac{\ln(n)+4}{\sqrt{1-\beta}}\rceil$ and $\tau=\frac{1}{2n}$, then it holds for any $2\le k\le n$ that*

$$\lambda_k(\bar{M})\in\Big[\tau-(1-\tau)\rho(M-\frac{1}{n}\mathbb{1}\mathbb{1}^T),\ \tau+(1-\tau)\rho(M-\frac{1}{n}\mathbb{1}\mathbb{1}^T)\Big]\subseteq\Big[\frac{1}{4n},\ \frac{3}{4n}\Big], \tag{38}$$

In other words, for $2 \le k \le n$, $\lambda_k(\bar{M})$ can vanish with respect to n while keeping a constant ratio $\frac{\lambda_2(\bar{M})}{\lambda_n(\bar{M})} \le 3$ (Proof is in Appendix F.2). ■

6.2 Reformulating D^2/Exact-Diffusion with Multiple Gossip steps

Primal recursion. With the above discussion, it holds that $\mathbf{x}^{(k+1)} = \bar{M}\phi^{(k+1)}$ after the fast gossip averaging step in Algorithm 2. Substituting this relation into Algorithm 2, we achieve the primal recursion for D^2/Exact-Diffusion with multiple gossip steps:

$$\mathbf{x}^{(k+1)} = \bar{M}\Big(2\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} - \gamma\big(\mathbf{g}^{(k)} - \mathbf{g}^{(k-1)}\big)\Big), \quad \forall\ k = 1, 2, \dots \tag{39}$$

where $\mathbf{g}^{(k)} = [g_1^{(k)}, \dots, g_n^{(k)}]^T \in \mathbb{R}^{n\times d}$ and $g_i^{(k)}$ is achieved by the gradient accumulation step in Algorithm 2, $\bar{M} = (1-\tau)M^{(R)} + \tau I$ and $M^{(R)}$ is achieved by recursions (36a) and (36b). The spectral properties of $\bar{M}$ is given in (38).

Primal-dual recursion. The primal recursion in (39) is equivalent to the following primal-dual updates

$$\begin{cases} \mathbf{x}^{(k+1)} = \bar{M}\big(\mathbf{x}^{(k)} - \gamma\mathbf{g}^{(k)}\big) - V\mathbf{y}^{(k)}, \\ \mathbf{y}^{(k+1)} = \mathbf{y}^{(k)} + \bar{V}\mathbf{x}^{(k+1)}, \quad \forall\ k = 0, 1, 2, \dots \end{cases} \tag{40}$$

where $\bar{V} = (I - \bar{M})^{\frac{1}{2}}$. Recursions (39) and (40) have two differences from the vanilla D^2/Exact-Diffusion recursions (5) and (6). First, the weight matrix $\bar{W}$ is replaced by $\bar{M}$. Second, the gradient $\mathbf{g}^{(k)}$ is achieved via gradient accumulation. This implies that the convergence analysis of D^2/Exact-Diffusion with multi-round gossip communication can follow that of vanilla D^2/Exact-Diffusion. We only need to pay attentions to the influence of $\bar{M}$ obtained by multi-round gossip steps and the $\mathbf{g}^{(k)}$ achieved by gradient accumulation.

6.3 Convergence Rate and Transient Stage

The following theorem establishes the convergence property of Algorithm 2 under general convexity.

Theorem 25 (Convergence under general convexity**)** *With Assumptions 1-5, $R = \lceil \frac{\ln(n)+4}{\sqrt{1-\beta}} \rceil$, and learning rate*

$$\gamma = \min\left\{ \frac{1}{4L}, \frac{(1-\tilde{\beta})\tilde{\lambda}_n^{\frac{1}{2}}}{10L\tilde{\lambda}_2}, \Big(\frac{\tilde{r}_0}{\tilde{r}_1(K+1)}\Big)^{\frac{1}{2}}, \Big(\frac{\tilde{r}_0}{\tilde{r}_2(K+1)}\Big)^{\frac{1}{3}}, \Big(\frac{\tilde{r}_0}{\tilde{r}_3}\Big)^{\frac{1}{3}} \right\},$$

where $\tilde{r}_0, \tilde{r}_1$, $\tilde{r}_2$ and $\tilde{r}_3$ are constants defined in (117), $\tilde{\beta}$, $\tilde{\lambda}_n$ and $\tilde{\lambda}_2$ are constants defined in (115), and K is the number of outer loop, Algorithm 2 converges at

$$\frac{1}{K+1}\sum_{k=0}^{K}\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) = O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{\sigma^{\frac{2}{3}}\ln(n)^{\frac{1}{3}}}{n^{\frac{1}{3}}(1-\beta)^{\frac{1}{6}}T^{\frac{2}{3}}} + \frac{\ln(n)}{(1-\beta)^{\frac{1}{2}}T}\Big). \tag{41}$$

where $T = KR$ is the total number of sampled data (or gossip communications) (Proof is in Appendix F.3).

■

Corollary 26 (Transient stage under general convexity) *Under the same assumptions as in Theorem 25, the transient stage for multi-step D^2/Exact-Diffusion is on the order of* $\tilde{O}\left(\frac{n}{1-\beta}\right)$.

Proof With convergence rate derived in (41), the transient stage for D^2/Exact-Diffusion with multi-round gossip is given by

$$T_{\text{tran}} \leq \min\left\{ \quad T\in\mathbb{N} \quad \middle| \quad \frac{\sigma^{\frac{2}{3}}\ln(n)^{\frac{1}{3}}}{n^{\frac{1}{3}}(1-\beta)^{\frac{1}{6}}T^{\frac{2}{3}}} \lesssim \frac{\sigma}{\sqrt{nT}} \quad \text{and} \quad \frac{\ln(n)}{(1-\beta)^{\frac{1}{2}}T} \lesssim \frac{\sigma}{\sqrt{nT}} \quad \right\}. \tag{42}$$

With (42), we have $T_{\text{tran}} \lesssim \frac{n[\ln(n)]^2}{\sigma^2(1-\beta)} = \tilde{O}\left(\frac{n}{1-\beta}\right)$. We use $\tilde{O}(\cdot)$ to hide some logarithm factors. ■

The following theorem establishes the convergence performance of Algorithm 2 with strong convexity.

Theorem 27 (Convergence under strong convexity) *With Assumptions 1, 4, 5, 6 and* $R = \lceil \frac{\ln(n)+4}{\sqrt{1-\beta}} \rceil$, *if the learning rate satisfies*

$$\gamma = \min\left\{ \frac{1}{4L}, \frac{1-\tilde{\beta}}{26L}\Big(\frac{\tilde{\lambda}_n^{1/2}}{\tilde{\lambda}_2}\Big), \frac{2\ln(2n\mu\mathbb{E}\|\tilde{\boldsymbol{z}}^{(0)}\|^2K^2/[\tilde{\sigma}^2(1-\tilde{\beta})])}{\mu K} \right\},$$

where $\tilde{\sigma}$, $\tilde{\beta}$, $\tilde{\lambda}_n$ *and* $\tilde{\lambda}_2$ *are constants defined in* (115), *Algorithm 2 converges at*

$$\frac{1}{H_K}\sum_{k=0}^{K} h_k\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) = \tilde{O}\left(\frac{\sigma^2}{nT} + \frac{\sigma^2}{n(1-\beta)^{\frac{1}{2}}T^2} + \exp\{-(1-\beta)^{\frac{1}{2}}T\}\right). \tag{43}$$

where h_k *and* H_K *are defined in Lemma 17. (Proof is in Appendix F.4).* ■

Corollary 28 (Transient stage) *Under the same assumptions as Theorem 18, the transient stage for multi-step D^2/Exact-Diffusion in the strongly convex scenario is on the order of* $\tilde{O}\big((1-\beta)^{-\frac{1}{2}}\big)$.

Proof The third term (43) decays exponentially fast and hence can be ignored compared to the first two terms as long as $(1-\beta)^{\frac{1}{2}}T = \tilde{O}(1)$, i.e., $T = \tilde{O}((1-\beta)^{-\frac{1}{2}})$, otherwise the exponential term remains a constant. As a result, the transient stage for D^2/Exact-Diffusion with multi-round gossip for strongly-convex problems is given by

$$T_{\text{tran}} \leq \min\left\{ \quad T\in\mathbb{N} \quad \middle| \quad \frac{\sigma^2}{n(1-\beta)^{\frac{1}{2}}T^2} \lesssim \frac{\sigma^2}{nT} \quad \right\}. \tag{44}$$

With (44), we have $T_{\text{tran}} = \tilde{O}\big((1-\beta)^{-\frac{1}{2}}\big)$. We use $\tilde{O}(\cdot)$ rather than $O(\cdot)$ because some logarithm factors are hidden inside. ■

Remark 29 *In Corollary 28, the transient stage of D^2/Exact-Diffusion with multi-round gossip communication has a significantly better (i.e., weaker) dependence on network topology connectivity $1-\beta$ and network size n compared to existing works (Koloskova et al., 2020; Pu et al., 2021; Pu and Nedić, 2020; Huang and Pu, 2021; Koloskova et al., 2021), see Table 2.* □

Remark 30 *While the transient stage derived in Corollary 28 seems to be independent of network size n, it essentially grows with n logarithmically. We omit the logarithm terms in the $\tilde{O}(\cdot)$ expression.* □

7. Numerical Simulation

In this section, we validate the established theoretical results with numerical simulations.

7.1 Strongly-Convex Scenario

Problem. We consider the following decentralized least-square problem

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{2n}\sum_{i=1}^{n}\|A_i x - b_i\|^2 \tag{45}$$

where $A_i \in \mathbb{R}^{M\times d}$ is the coefficient matrix, and $b_i \in \mathbb{R}^M$ is the measurement. Quantities A_i and b_i are associated with node i, and M is the size of local dataset.

Simulation settings. In our simulations, we set $d = 10$ and $M = 1000$. To control the data heterogeneity across the nodes, we first let each node i be associated with a local solution $x_i^\star$, and such $x_i^\star$ is generated by $x_i^\star = x^\star + v_i$ where $x^\star \sim \mathcal{N}(0, I_d)$ is a randomly generated vector while $v_i \sim \mathcal{N}(0, \sigma_h^2 I_d)$ controls the similarity between each local solution. Generally speaking, a large σ_h^2 results in local solutions $\{x_i^\star\}$ that are vastly different from each other. With $x_i^\star$ at hand, we can generate local data that follows distinct distributions. At node i, we generate each element in A_i following standard normal distribution. Measurement b_i is generated by $b_i = A_i x_i^\star + s_i$ where $s_i \sim \mathcal{N}(0, \sigma_s^2 I)$ is some white noise. Clearly, solution $x_i^\star$ controls the distribution of the measurements b. In this way, we can easily control data heterogeneity by adjusting σ_h^2. At each iteration k, each node will randomly sample a row in A_i and the corresponding element in b_i and use them to evaluate the stochastic gradient. The metric for all simulations in this subsection is $\frac{1}{n}\sum_{i=1}^n \|x_i^{(k)} - x^\star\|^2$ where $x^\star$ is the global simulation to problem (45) and it has a closed-form $x^\star = (\sum_{i=1}^n A_i^T A_i)^{-1}(\sum_{i=1}^n A_i^T A_i b_i)$.

Performance with heterogeneous data. We now compare the convergence performance of Parallel SGD (P-SGD), Decentralized SGD (D-SGD), D^2/Exact-Diffusion (D2/ED), and D^2/Exact-Diffusion with multi-round gossip communication (MG-D2/ED) when data heterogeneity exists. The target is to examine their robustness to the influence of network topology. To this end, we let $\sigma_h^2 = 0.2$ and organize $n = 32$ nodes into a cycle. The left plot in Fig. 1 lists the performances of all algorithms. Each algorithm utilizes the same learning rate which decays by half for every 2,000 gossip communications. In this plot, it is observed that all decentralized algorithms, after certain amounts of transient iterations, can match with P-SGD asymptotically. In addition, we find D-SGD is least robust while MG-D2/ED

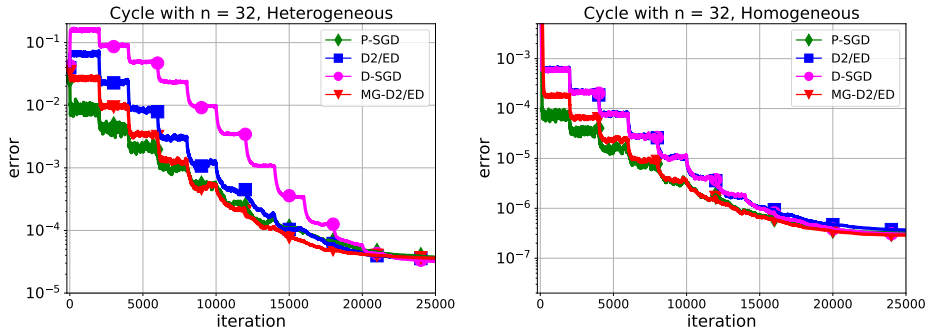

Figure 1: Performance of different stochastic algorithms to solve problem (45). The left plot is with heterogeneous data while the right is with homogeneous data.

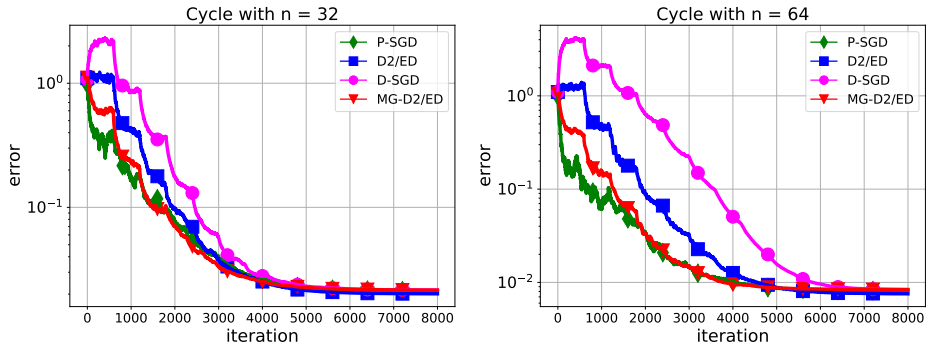

Figure 2: Performance of different stochastic algorithms to solve problem (46).

is most robust to network topology, which aligns with the theoretically established bounds for transient stage in Table 2.

Performance with homogeneous data. We next compare there algorithms with homogeneous data. To this end, we let $\sigma_h^2 = 0$ and organize $n = 32$ nodes into a cycle. The other settings are the same as in the heterogeneous scenario discussed in the above. The right plot in Fig. 1 lists the performances of all algorithms. For MG-D2/ED, the x-axis indicates $K \times R$ which facilitates the fair comparison with other algorithms where K is the number of outer loop and R is the inner gossip communication rounds. It is observed that D-SGD and D2/ED have almost the same convergence behaviours, which validates the conclusion in Theorem 22 that D-SGD can match with D2/ED in the homogeneous data scenario. In addition, we find MG-D2/ED requires less transient iterations than D-SGD and D2/ED to match with P-SGD, indicating that it is more robust to network topology even if in the homogeneous data scenario.

7.2 Generally-Convex Scenario

Problem. We consider the following decentralized logistic regression problem

$$\min_{x\in\mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x) \qquad \text{where} \qquad f_i(x) = \frac{1}{M}\sum_{m=1}^{M} \ln\left(1+\exp(-y_{i,m}h_{i,m}^{\mathsf{T}}x)\right) \tag{46}$$

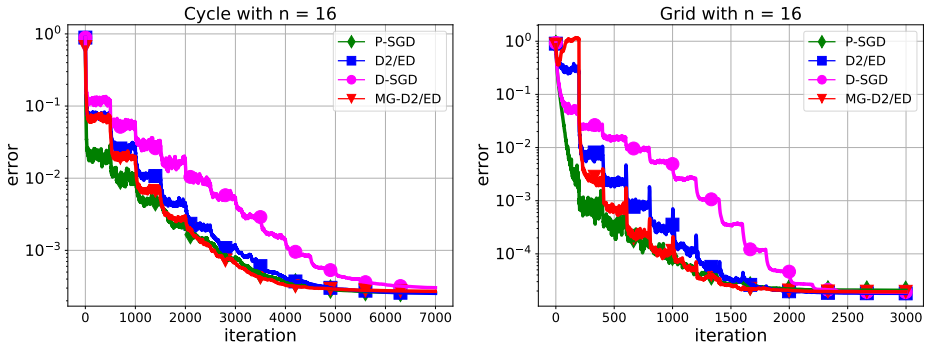

Figure 3: Performance of different stochastic algorithms to solve problem (47) with different real datasets and topologies. The left plot is with MNIST dataset, and the right is with COVTYPE dataset.

where $\{h_{i,m}, y_{i,m}\}_{m=1}^{M}$ is the training dateset held by node i in which $h_{i,m} \in \mathbb{R}^d$ is a feature vector while $y_{i,m} \in \{-1, +1\}$ is the corresponding label.

Simulation settings. Similar to the strongly-convex scenario, each node i is associated with a local solution $x_i^\star$. To generate local dataset $\{h_{i,m}, y_{i,m}\}_{m=1}^{M}$, we first generate each feature vector $h_{i,m} \sim \mathcal{N}(0, I_d)$. We label $y_{i,m} = 1$ with probability $1/(1+\exp(-y_{i,\ell} h_{i,m}^{\mathsf{T}} x_i^\star))$; otherwise $y_{i,m} = -1$. We can control data heterogeneity by adjusting σ_h^2.

Robustness to network topology. Fig. 2 lists the performances of all stochastic algorithms with different network sizes. When size of the cycle graph increases from 32 to 64 (the quantity $1/(1-\beta)$ increases from 78.07 to 311.51), it is observed that the performance of D-SGD is significantly deteriorated. In contrast, such change of the network topology just influences D2/ED slightly. Furthermore, it is observed that MG-D2/ED is always very close to P-SGD no matter the topology is well-connected or not. These phenomenons are consistent with the established transient stage for the generally-convex scenario in Table 2.

7.3 Simulation with Real Datasets

This subsection examines the performances of P-SGD, D-SGD, D2/ED, and MG-D2/ED with real datasets. We run experiments for the regularized logistic regression problem with

$$f_i(x) = \frac{1}{M} \sum_{m=1}^{M} \ln\left(1 + \exp(-y_{i,m} h_{i,m}^{\mathsf{T}} x)\right) + \frac{\rho}{2}\|x\|^2 \tag{47}$$

where $\rho > 0$ is a positive constant. We consider two real datasets: MNIST (Deng, 2012) and COVTYPE.binary (Rossi and Ahmed, 2015). The MNIST recognition task has been transformed into a binary classification problem by considering data with labels 2 and 4. In COVTYPE.binary, we use 50, 000 samples as training data and each data has dimension 54. In MNIST we use 10, 000 samples as training data and each data has dimension 784. The regularization coefficient $\rho = 0.001$ for all simulations. To promote data heterogeneity, we control the ratio of the sizes of positive and negative samples for each node. More specifically, in COVTYPE.binary, half of the nodes maintain 54% positive samples while the other half maintain 54% negative samples. Likewise, the ratio is fixed as 70% for MNIST

dataset, which is a bit larger since the heterogeneity between all grey-scale handwritten digits of 2 and 4 is relatively weak. Except for the fixed ratio of the positive samples to negative ones, all training data are distributed uniformly to each local node. The left plot in Fig. 3 illustrates the performance of various algorithms with MNIST dataset over the cycle graph while the right is with COVTYPE dataset over the grid graph. In all simulations, we find the transient stage as well as the robustness to network topology coincides those established in Table 2 well. D2/ED always converges better than D-SGD, and MG-D2/ED is least sensitive to network topology compared to D-SGD and D2/ED.

8. Conclusion and Discussion

In this work, we revisited the D^2/Exact-Diffusion algorithm (Tang et al., 2018; Yuan et al., 2018a,b; Li et al., 2019b; Yuan et al., 2020) and studied its non-asymptotic convergence rate under both the generally-convex and strongly-convex settings.By removing the influence of data heterogeneity, D^2/Exact-Diffusion is shown to improve the transient stage of D-SGD. For the generally-convex setting, the improvement is from $O(n^3/(1-\beta)^4)$ to $O(n^3/(1-\beta)^2)$. For the strongly-convex setting, the rate is improved from $O(n/(1-\beta)^2)$ to $\tilde{O}(n/(1-\beta))$. This result indicates that D^2/Exact-Diffusion (Tang et al., 2018; Yuan et al., 2018a,b; Li et al., 2019b; Yuan et al., 2020) is less sensitive to network topology. For the strongly-convex scenario, we also proved that our transient stage bound coincides with the lower bound of homogeneous D-SGD in terms of *network topology dependence*, which implies that D^2/Exact-Diffusion cannot have worse network dependence than D-SGD and has a better dependence in the heterogeneous setting. Moreover, when D^2/Exact-Diffusion is equipped with gradient accumulation and multi-round gossip communications, its transient stage can be further improved to $\tilde{O}(1/(1-\beta)^{\frac{1}{2}})$ and $\tilde{O}(n/(1-\beta))$ for strongly and generally convex cost functions, respectively.

There are still several open questions to answer for the family of data-heterogeneity-corrected methods such as EXTRA, D^2/Exact-Diffusion, and gradient-tracking. First, it remain uncertain whether these methods can still have improved dependence on network topology over time-varying topologies. Second, while data-heterogeneity-corrected methods are endowed with superior convergence properties in terms of robustness to heterogeneous data or network topology dependence, D-SGD can still *empirically* outperform them in deep learning applications, as seen in (Lin et al., 2021; Yuan et al., 2021). Significant efforts may still be required to bridge the gap between theory and practical implementation.

References

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

Sulaiman A. Alghunaim, Ernest K. Ryu, Kun Yuan, and Ali H. Sayed. Decentralized proximal gradient algorithms with linear convergence rates. *IEEE Transactions on Automatic Control*, 66(6):2787–2794, June 2021.

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

Albert S Berahas, Raghu Bollapragada, Nitish Shirish Keskar, and Ermin Wei. Balancing communication and computation in distributed optimization. *IEEE Transactions on Automatic Control*, 64(8):3141–3155, 2018.

Jianshu Chen and Ali H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

Jianshu Chen and Ali H. Sayed. Distributed pareto optimization via diffusion strategies. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):205–220, 2013.

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 (ICML)*, 2021.

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

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

Andrew Gibiansky. Bringing HPC techniques to deep learning. `https://andrew.gibiansky.com/blog/machine-learning/baidu-allreduce/`, 2017. Accessed: 2020-08-12.

Kun Huang and Shi Pu. Improving the transient times for distributed stochastic gradient methods. *arXiv preprint arXiv:2105.04851*, 2021.

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

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pages 1–12, 2020.

Anastasia Koloskova, Tao Lin, and Sebastian U. Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Huan Li, Cong Fang, Wotao Yin, and Zhouchen Lin. Decentralized accelerated gradient methods with increasing penalty parameters. *IEEE Transactions on Signal Processing*, 68:4855–4870, 2020.

Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pages 583–598, 2014.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019a.

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506, 2019b.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pages 3043–3052, 2018.

Tao Lin, Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. *arXiv preprint arXiv:2102.04761*, 2021.

Ji Liu and A Stephen Morse. Accelerated linear iterations for distributed averaging. *Annual Reviews in Control*, 35(2):160–165, 2011.

Cassio G Lopes and Ali H Sayed. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7): 3122–3136, 2008.

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

Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. GNSD: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *IEEE Data Science Workshop (DSW)*, pages 315–321. IEEE, 2019.

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

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

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

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, pages 1–49, 2020.

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.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

Ryan Rossi and Nesreen Ahmed. The network data repository with interactive graph analytics and visualization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Ali H Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7(ARTICLE):311–801, 2014.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning*, pages 3027–3036, 2017.

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

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

Sebastian U. Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019a.

Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019b.

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, 2018.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 1986.

César A. Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. *Optimization Methods and Software*, pages 1–40, 2020.

Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *IEEE Conference on Decision and Control (CDC)*, pages 5445–5450, Maui, HI, USA, 2012.

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

Ran Xin, Usman A. Khan, and Soummya Kar. Fast decentralized nonconvex finite-sum optimization with recursive variance reduction. *SIAM Journal on Optimization*, 32(1), 2022.

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

Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Bicheng Ying, Kun Yuan, Hanbin Hu, Yiming Chen, and Wotao Yin. BlueFog: Make decentralized algorithms practical for optimization and deep learning. `https://github.com/Bluefog-Lib/bluefog`, 2021b. Accessed: 2021-05-15.

Bicheng Ying, Kun Yuan, Hanbin Hu, Yiming Chen, and Wotao Yin. Bluefog: Make decentralized algorithms practical for optimization and deep learning. *arXiv preprint arXiv:2111.04287*, 2021c.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H. Sayed. Exact dffusion for distributed optimization and learning – Part I: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708 – 723, 2018a.

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H Sayed. Exact diffusion for distributed optimization and learning—Part II: Convergence analysis. *IEEE Transactions on Signal Processing*, 67(3):724–739, 2018b.

Kun Yuan, Sulaiman A. Alghunaim, Bicheng Ying, and Ali H. Sayed. On the influence of bias-correction on distributed stochastic optimization. *IEEE Transactions on Signal Processing*, 68:4352–4367, 2020.

Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. DecentLaM: Decentralized momentum sgd for large-batch deep training. pages 3029–3039, 2021.

Kun Yuan, Xinmeng Huang, Yiming Chen, Xiaohan Zhang, Yingya Zhang, and Pan Pan. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jiaqi Zhang and Keyou You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J. Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

Appendix

Appendix A. Notations and Preliminaries

We first review some notations and facts.

- $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ is a symmetric and doubly stochastic combination matrix
- $\bar{W} = (I + W)/2 \in \mathbb{R}^{n\times n}$
- $V = (I - \bar{W})^{1/2} = (\frac{I-W}{2})^{1/2}$ and hence $I - \bar{W} = V^2$
- $\lambda_i(W)$ is the ith largest eigenvalue of matrix W, and $\bar{\lambda}_i(W) = (1 + \lambda_i(W))/2$ is the ith largest eigenvalue of matrix $\bar{W}$. Note that $\lambda_i(W) \in (-1, 1)$ and $\bar{\lambda}_i(W) \in (0, 1)$ for $i = 2, \ldots, n$.
- Let $\Lambda = \mathrm{diag}\{\lambda_1(W), \ldots, \lambda_n(W)\} \in \mathbb{R}^{n\times n}$. It holds that $W = Q\Lambda Q^T$ where $Q = [q_1, q_2, \ldots, q_n] \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $q_1 = \frac{1}{\sqrt{n}}\mathbb{1}_n$.
- $\bar{W} = Q\bar{\Lambda}Q^T$ where $\bar{\Lambda} = (I + \Lambda)/2$.
- $V = Q(I - \bar{\Lambda})^{1/2}Q^T$
- If a matrix $A \in \mathbb{R}^{n\times n}$ is normal, *i.e.*, $AA^T = A^TA$, it holds that $A = UDU^*$ where D is a diagonal matrix and U is a unitary matrix.
- If a matrix $\Pi \in \mathbb{R}^{n\times n}$ is a permutation matrix, it holds that $\Pi^{-1} = \Pi^T$.

Smoothness. Since each $f_i(x)$ is assumed to be L-smooth in Assumption 4, it holds that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is also L-smooth. As a result, the following inequality holds for any $x, y \in \mathbb{R}^d$:

$$f_i(x) - f_i(y) - \frac{L}{2}\|x - y\|^2 \le \langle \nabla f_i(y), x - y\rangle$$

Smoothness and convexity. If each $f_i(x)$ is further assumed to be convex (see Assumption 3), it holds that $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is also convex. For this scenario, it holds for any $x, y \in \mathbb{R}^d$ that:

$$\|\nabla f(x) - \nabla f(x^\star)\|^2 \le 2L\big(f(x) - f(x^\star)\big)$$
$$f_i(x) - f_i(y) \le \langle \nabla f_i(x), x - y\rangle$$

Submultiplicativity of the Frobenius norm. Given matrices $W \in \mathbb{R}^{n\times n}$ and $\mathbf{y} \in \mathbb{R}^{n\times d}$, it holds that

$$\|W\mathbf{y}\|_F \le \|W\|_2\|\mathbf{y}\|_F.$$

To verify it, by letting y_j be the jth column of $\mathbf{y}$, we have $\|W\mathbf{y}\|_F^2 = \sum_{j=1}^d \|Wy_j\|_2^2 \le \sum_{j=1}^d \|W\|_2^2\|y_j\|_2^2 = \|W\|_2^2\|\mathbf{y}\|_F^2$.

Appendix B. The Fundamental Decomposition

B.1 Proof of Lemma 4

We now analyze the eigen-decomposition of matrix B:

$$B = \begin{bmatrix} \bar{W} & -V \\ V\bar{W} & \bar{W} \end{bmatrix}.$$

Proof. Using $\bar{W} = Q\bar{\Lambda}Q^T$, it holds that

$$B = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \bar{\Lambda} & -(I-\bar{\Lambda})^{1/2} \\ \bar{\Lambda}(I-\bar{\Lambda})^{1/2} & \bar{\Lambda} \end{bmatrix} \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix}.$$

Note that $\bar{\Lambda}(I-\bar{\Lambda})^{1/2} = (I-\bar{\Lambda})^{1/2}\bar{\Lambda}$ because both $\bar{\Lambda}$ and $I-\bar{\Lambda}$ are diagonal matrices. We next introduce

$$\begin{aligned} E_{(i)} &= \begin{bmatrix} \bar{\lambda}_i & -(1-\bar{\lambda}_i)^{1/2} \\ \bar{\lambda}_i(1-\bar{\lambda}_i)^{1/2} & \bar{\lambda}_i \end{bmatrix} \in \mathbb{R}^{2\times 2} \\ E &= \text{BlockDiag}\{E_{(1)}, \cdots, E_{(n)}\} \in \mathbb{R}^{2n\times 2n} \end{aligned} \tag{48}$$

where $\bar{\lambda}_i = \lambda_i(\bar{W})$, and E is a block diagonal matrix with each ith bloack diagonal matrix as $E_{(i)}$. It is easy to verify that there exists some permutation matrix Π such that

$$B = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \Pi\, E\, \Pi^T \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix}. \tag{49}$$

Next we focus on the matrix $E_{(i)}$ defined in (48). Note that $E_{(1)} = I$. For $i \geq 2$, it holds that

$$E_{(i)} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & \bar{\lambda}_i^{1/2} \end{bmatrix}}_{C_{(i)}} \underbrace{\begin{bmatrix} \bar{\lambda}_i & -[\bar{\lambda}_i(1-\bar{\lambda}_i)]^{1/2} \\ [\bar{\lambda}_i(1-\bar{\lambda}_i)]^{1/2} & \bar{\lambda}_i \end{bmatrix}}_{G_{(i)}} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & \bar{\lambda}_i^{-1/2} \end{bmatrix}}_{C_{(i)}^{(-1)}} \tag{50}$$

Since $G_{(i)}$ is normal, it holds that (see Appendix A)

$$G_{(i)} = U_{(i)} D_{(i)} U_{(i)}^*, \quad \text{where} \quad D_i = \text{diag}\{\sigma_1(G_{(i)}), \sigma_2(G_{(i)})\}, \tag{51}$$

In the above expression, $\sigma_1(G_{(i)})$ and $\sigma_2(G_{(i)})$ are complex eigenvalues of $G_{(i)}$. Moreover, it holds that $|\sigma_1(G_{(i)})| = |\sigma_2(G_{(i)})| = \bar{\lambda}_i^{1/2} < 1$. The quantity $U_{(i)} \in \mathbb{R}^{2\times 2}$ is a unitary matrix. Next we define $C = \text{BlockDiag}\{C_{(1)}, \ldots, C_{(n)}\}$, $U = \text{BlockDiag}\{U_{(1)}, \ldots, U_{(n)}\}$, and $D = \text{BlockDiag}\{D_{(1)}, \ldots, D_{(n)}\}$. By substituting (50) and (51) into (49), we have

$$B = d \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \Pi\, C\, U\, D\, U^*\, C^{-1}\, \Pi^T \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix} d^{-1} \tag{52}$$

where d is any positive constant. Next we define

$$X = d \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \Pi\, C\, U, \quad X^{-1} = U^*\, C^{-1}\, \Pi^T \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix} d^{-1}.$$

By letting $d = \sqrt{n}$ and considering the structure of Q, Π, C, and U, it is easy to verify that

$$X = [r_1\ r_2\ X_R] \quad \text{where} \quad r_1 = \begin{bmatrix} \mathbb{1}_n \\ 0 \end{bmatrix}, \quad r_2 = \begin{bmatrix} 0 \\ \mathbb{1}_n \end{bmatrix} \tag{53}$$

$$X^{-1} = [\ell_1\ \ell_2\ X_L^T]^T \quad \text{where} \quad \ell_1 = \begin{bmatrix} \frac{1}{n}\mathbb{1}_n \\ 0 \end{bmatrix}, \quad \ell_2 = \begin{bmatrix} 0 \\ \frac{1}{n}\mathbb{1}_n \end{bmatrix} \tag{54}$$

With (52)–(54), it holds that

$$B = XDX^{-1}$$

where X and X^{-1} take the form of (53) and (54), and

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & D_1 \end{bmatrix}$$

and D_1 is a diagonal matrix with complex entries. The magnitudes of the diagonal entries in D_1 are all strictly less than 1. Next we evaluate the quantity $\|X\|\|X^{-1}\|$:

$$\begin{aligned} \|X\|\|X^{-1}\| &\leq \left\| \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} \right\| \|\Pi\|\|C\|\|U\|\|U^*\|\|C^{-1}\|\|\Pi^T\| \left\| \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix} \right\| \\ &\overset{(a)}{=} \|C\|\|C^{-1}\| \\ &\leq \max_i \{\bar{\lambda}_i^{-1/2}\} = \bar{\lambda}_n^{-1/2} \end{aligned} \tag{55}$$

where (a) holds because Q is orthogonal, U is unitary, and $\Pi^T\Pi = I$. Note that

$$X_R = XS, \quad \text{and} \quad X_L = S^T X^{-1}$$

where $S = [e_3, \ldots, e_{2n}] \in \mathbb{R}^{2n \times 2(n-1)}$ and e_j is the jth column of the identity matrix I_{2n}. It then holds that

$$\|X_R\|\|X_L\| \leq \|X\|\|S\|\|S^T\|\|X^{-1}\| = \|X\|\|X^{-1}\| \overset{(55)}{\leq} \bar{\lambda}_n^{-1/2}$$

■

B.2 Proof of Lemma 5

Proof By left-multiplying X^{-1} to both sides of (13) and utilizing the decomposition in (14), we have

$$\begin{aligned} &\begin{bmatrix} \ell_1^T \\ \ell_2^T \\ X_L/c \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(k+1)} - \mathbf{x}^\star \\ \mathbf{y}^{(k+1)} - \mathbf{y}^\star \end{bmatrix} \\ = &\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & D_1 \end{bmatrix} \begin{bmatrix} \ell_1^T \\ \ell_2^T \\ X_L/c \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(k)} - \mathbf{x}^\star \\ \mathbf{y}^{(k)} - \mathbf{y}^\star \end{bmatrix} - \gamma \begin{bmatrix} \ell_1^T \\ \ell_2^T \\ X_L/c \end{bmatrix} \begin{bmatrix} \bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \\ V\bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \end{bmatrix}. \end{aligned} \tag{56}$$

With the definition of $\bar{\boldsymbol{z}}^{(k)}$ in (20), the structure of ℓ_1 in (15), and $\bar{W}\mathbb{1} = \mathbb{1}$, the first line in (56) becomes

$$\bar{\boldsymbol{z}}^{(k+1)} = \bar{\boldsymbol{z}}^{(k)} - \frac{\gamma}{n}\mathbb{1}^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)) - \gamma\bar{\mathbf{s}}^{(k)}. \tag{57}$$

where $\bar{\mathbf{s}}^{(k)}$ is defined in (19b). With the structure of ℓ_2 in (15) and $V\mathbb{1} = 0$, the second line in (56) becomes

$$\ell_2^T\begin{bmatrix}\mathbf{x}^{(k+1)} - \mathbf{x}^\star \\ \mathbf{y}^{(k+1)} - \mathbf{y}^\star\end{bmatrix} = \ell_2^T\begin{bmatrix}\mathbf{x}^{(k)} - \mathbf{x}^\star \\ \mathbf{y}^{(k)} - \mathbf{y}^\star\end{bmatrix} \quad \Longleftrightarrow \quad \frac{1}{n}\mathbb{1}^T(\mathbf{y}^{(k+1)} - \mathbf{y}^\star) = \frac{1}{n}\mathbb{1}^T(\mathbf{y}^{(k)} - \mathbf{y}^\star) \tag{58}$$

Since $\mathbf{y}^\star$ lies in the range space of V (see Lemma 3) and $\mathbf{y}^{(k)}$ also lies in the range space of V when $\mathbf{y}^{(0)} = 0$ (see the update of $\mathbf{y}$ in (6)), it holds that $\frac{1}{n}\mathbb{1}^T(\mathbf{y}^{(k)} - \mathbf{y}^\star) = 0$. As a result, the recursion (58) can be ignored since $\frac{1}{n}\mathbb{1}^T(\mathbf{y}^{(k)} - \mathbf{y}^\star) = 0$ holds for all iterations.

Finally we examine the third line in (56). To this end, we eigen-decompose $\bar{W}$ as

$$\bar{W} = \underbrace{\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbb{1} & Q_R\end{bmatrix}}_{:=Q}\begin{bmatrix}1 & 0 \\ 0 & \bar{\Lambda}_R\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbb{1}^T \\ Q_R^T\end{bmatrix} = \frac{1}{n}\mathbb{1}\mathbb{1}^T + Q_R\bar{\Lambda}_R Q_R^T \tag{59}$$

where $Q_R \in \mathbb{R}^{n\times(n-1)}$ satisfies $Q_R^TQ_R = I$ and $\bar{\Lambda}_R = \mathrm{diag}\{\lambda_2(\bar{W}), \ldots, \lambda_n(\bar{W})\}$. Since V shares the same eigen-space as $\bar{W}$, it holds that

$$V\bar{W} = \begin{bmatrix}\frac{1}{\sqrt{n}}\mathbb{1} & Q_R\end{bmatrix}\begin{bmatrix}0 & 0 \\ 0 & (I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{n}}\mathbb{1}^T \\ Q_R^T\end{bmatrix} = Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R Q_R^T \tag{60}$$

Next we rewrite $X_L = [X_{L,\ell}\ X_{L,r}]$ with $X_{L,\ell} \in \mathbb{R}^{2(n-1)\times n}$ and $X_{L,r} \in \mathbb{R}^{2(n-1)\times n}$. With the structure of X and X^{-1} in (14) and the equality $X^{-1}X = I$, it holds that

$$X_{L,\ell}\mathbb{1} = 0 \qquad X_{L,r}\mathbb{1} = 0. \tag{61}$$

With (59), (60), and (61), we have

$$\begin{aligned}
&X_L\begin{bmatrix}\bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \\ V\bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)})\end{bmatrix} \\
&= X_{L,\ell}\bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) + X_{L,r}V\bar{W}(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \\
&\overset{(a)}{=} X_{L,\ell}Q_R\bar{\Lambda}_R Q_R^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \\
&\quad + X_{L,r}Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R Q_R^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star) + \mathbf{s}^{(k)}) \\
&= \check{\boldsymbol{g}}^{(k)} + \check{\boldsymbol{s}}^{(k)}
\end{aligned}$$

where (a) holds because of (59) – (61), and $\check{\boldsymbol{g}}^{(k)}$ and $\check{\boldsymbol{s}}^{(k)}$ in the last equality are defined in (19a) and (19c), respectively. With the above equality and the definition of $\check{\mathbf{z}}^{(k)}$ in (20), the third line in (56) becomes

$$\check{\mathbf{z}}^{(k+1)} = D_1\check{\mathbf{z}}^{(k)} - \frac{\gamma}{c}\check{\mathbf{g}}^{(k)} - \frac{\gamma}{c}\check{\mathbf{s}}^{(k)}. \tag{62}$$

Combining (57) and (62), we achieve the result in (18). ■

B.3 Proof of Proposition 7

Proof We first evaluate the magnitude of $\|\mathbf{y}^\star\|_F^2$. Recall that $V = (I - \bar{W})^{\frac{1}{2}}$ and $\bar{W} = (I + W)/2$ is a symmetric and doubly-stochastic matrix. If we let $\bar{\lambda}_k = \lambda_k(\bar{W})$, it holds that $1 = \bar{\lambda}_1 > \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_n > 0$. We eigen-decompose $V = U\Lambda U^T$ with $\Lambda = \mathrm{diag}\{\lambda_k(V)\}$ and $\lambda_k(V) = (1-\bar{\lambda}_k)^{\frac{1}{2}}$. We next introduce $V^\dagger = U\Lambda^\dagger U^T$ with $\Lambda^\dagger = \mathrm{diag}\{\lambda_k(V^\dagger)\}$ in which $\lambda_1(V^\dagger) = 0$ and $\lambda_k(V^\dagger) = \lambda_k^{-1}(V) = (1-\bar{\lambda}_k)^{-\frac{1}{2}}$ for $2 \leq k \leq n$. Recall optimality condition (7a) that

$$\gamma\bar{W}\nabla f(\mathbf{x}^\star) + V\mathbf{y}^\star = 0 \quad \Longleftrightarrow \quad V\mathbf{y}^\star = -\gamma\bar{W}\nabla f(\mathbf{x}^\star). \tag{63}$$

Since $\mathbf{y}^\star$ lies in the range space of V, it holds that $V^\dagger V\mathbf{y}^\star = \mathbf{y}^\star$. This fact together with (63) leads to

$$\begin{aligned} & \mathbf{y}^\star = -\gamma V^\dagger\bar{W}\nabla f(\mathbf{x}^\star) \\ \Longrightarrow \quad & \|\mathbf{y}^\star\|_F^2 \leq \gamma^2\|V^\dagger\bar{W}\|_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2 \leq \frac{\gamma^2\bar{\lambda}_2^2}{1-\bar{\lambda}_2}\|\nabla f(\mathbf{x}^\star)\|_F^2 = O\left(\frac{n\gamma^2\bar{\lambda}_2^2}{1-\bar{\lambda}_2}\right) \end{aligned} \tag{64}$$

where we regard $\|\nabla f(\mathbf{x}^\star)\|_F^2 = \sum_{i=1}^n \|\nabla f_i(x^\star)\|^2 = O(n)$ in the last inequality.

Next we evaluate the magnitude of $\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2$. Recall from (20) that

$$\check{\mathbf{z}}^{(0)} = \frac{X_{L,\ell}}{c}(\mathbf{x}^{(0)} - \mathbf{x}^\star) + \frac{X_{L,r}}{c}(\mathbf{y}^{(0)} - \mathbf{y}^\star) = -\frac{X_{L,\ell}}{c}\mathbf{x}^\star - \frac{X_{L,r}}{c}\mathbf{y}^\star$$

where we utilized $\mathbf{x}^{(0)} = 0$ and $\mathbf{y}^{(0)} = 0$ in the last equality, and $X_L = [X_{L,\ell}, X_{L,r}]$. With (61) and the fact that $\mathbf{x}^\star = x^\star\mathbb{1}$, it holds that $X_{L,\ell}\mathbf{x}^\star = 0$ and

$$\|\check{\mathbf{z}}^{(0)}\|_F^2 \leq \frac{1}{c^2}\|X_{L,r}\|_2^2\|\mathbf{y}^\star\|_F^2 \overset{(a)}{\leq} \frac{1}{c^2}\|X_L\|_2^2\|\mathbf{y}^\star\|_F^2 \overset{(b)}{=} \|\mathbf{y}^\star\|_F^2 \overset{(64)}{=} O\left(\frac{n\gamma^2\bar{\lambda}_2^2}{1-\bar{\lambda}_2}\right)$$

where (a) holds because $\|X_{L,r}\| \leq \|X_L\|$ (see the detail derivation in (74)) and (b) holds by setting $c = \|X_L\|$).

■

Appendix C. Convergence Analysis for Generally-Convex Scenario

C.1 Proof of Lemma 8

Proof From (20) we have $[\bar{\boldsymbol{z}}^{(k)}]^T = [\frac{1}{n}\mathbb{1}^T(\mathbf{x}^{(k)} - \mathbf{x}^\star)]^T = \bar{x}^{(k)} - x^\star \in \mathbb{R}^d$, where $\bar{x}^{(k)} = \frac{1}{n}\mathbb{1}^T\mathbf{x}^{(k)}$ and $x^\star$ is the global solution to problem (1). With this relation, the first line of (18) becomes

$$\bar{x}^{(k+1)} - x^\star = \bar{x}^{(k)} - x^\star - \frac{\gamma}{n}\mathbb{1}^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)) - \gamma\bar{\mathbf{s}}^{(k)}.$$

The above equality implies that

$$\mathbb{E}[\|\bar{x}^{(k+1)} - x^\star\|^2|\mathcal{F}^{(k)}] \overset{(22)}{\leq} \|\bar{x}^{(k)} - x^\star - \frac{\gamma}{n}\mathbb{1}^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star))\|^2 + \frac{\gamma^2\sigma^2}{n} \tag{65}$$

Note that the first term can be expanded as follows.

$$\begin{aligned}
&\|\bar{x}^{(k)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n [\nabla f_i(x_i^{(k)}) - \nabla f_i(x^\star)]\|^2 \\
&= \|\bar{x}^{(k)} - x^\star\|^2 \underbrace{- \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{x}^{(k)} - x^\star, \nabla f_i(x_i^{(k)}) - \nabla f_i(x^\star)\rangle}_{(A)} + \underbrace{\gamma^2 \|\frac{1}{n}\sum_{i=1}^n [\nabla f_i(x_i^{(k)}) - \nabla f_i(x^\star)]\ \|^2}_{(B)}
\end{aligned} \tag{66}$$

We now bound the term (A):

$$\begin{aligned}
&\frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{x}^{(k)} - x^\star, \nabla f_i(x_i^{(k)}) - \nabla f_i(x^\star)\rangle \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{x}^{(k)} - x^\star, \nabla f_i(x_i^{(k)})\rangle \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{x}^{(k)} - x_i^{(k)}, \nabla f_i(x_i^{(k)})\rangle + \frac{2\gamma}{n}\sum_{i=1}^n \langle x_i^{(k)} - x^\star, \nabla f_i(x_i^{(k)})\rangle \\
&\overset{(a)}{\geq} \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\bar{x}^{(k)}) - f_i(x_i^{(k)}) - \frac{L}{2}\|\bar{x}^{(k)} - x_i^{(k)}\|^2\Big) + \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(x_i^{(k)}) - f_i(x^\star)\Big) \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\bar{x}^{(k)}) - f_i(x^\star)\Big) - \frac{\gamma L}{n}\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2 \\
&= 2\gamma\big(f(\bar{x}^{(k)}) - f(x^\star)\big) - \frac{\gamma L}{n}\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2
\end{aligned} \tag{67}$$

where (a) holds because of each $f_i(x)$ is convex and L-smooth. We next bound term (B):

$$\begin{aligned}
&\gamma^2 \|\frac{1}{n}\sum_{i=1}^n [\nabla f_i(x_i^{(k)}) - \nabla f_i(x^\star)]\ \|^2 \\
&= \gamma^2 \|\frac{1}{n}\sum_{i=1}^n [\nabla f_i(x_i^{(k)}) - \nabla f_i(\bar{x}^{(k)}) + \nabla f_i(\bar{x}^{(k)}) - \nabla f_i(x^\star)]\ \|^2 \\
&\overset{(10)}{\leq} \frac{2\gamma^2 L^2}{n}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + 2\gamma^2\|\nabla f(\bar{x}^{(k)}) - \nabla f(x^\star)\|^2 \\
&\leq \frac{2\gamma^2 L^2}{n}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + 4L\gamma^2\big(f(\bar{x}^{(k)}) - f(x^\star)\big).
\end{aligned} \tag{68}$$

where the last inequality holds because $f(x)$ is L-smooth. Substituting (68) and (67) into (66), we have

$$\begin{aligned}
&\|\bar{x}^{(k)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n \nabla f_i(x_i^{(k)})\|^2 \\
&\leq \|\bar{x}^{(k)} - x^\star\|^2 - 2\gamma(1 - 2L\gamma)\big(f(\bar{x}^{(k)}) - f(x^\star)\big) + \Big(\frac{\gamma L}{n} + \frac{2\gamma^2 L^2}{n}\Big)\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2
\end{aligned}$$

$$\leq \|\bar{x}^{(k)} - x^\star\|^2 - \gamma\big(f(\bar{x}^{(k)}) - f(x^\star)\big) + \frac{3\gamma L}{2n}\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2 \tag{69}$$

where the last inequality holds when $\gamma \leq \frac{1}{4L}$. Substituting (69) into (65) and taking expectation on the filtration $\mathcal{F}^{(k)}$, we achieve

$$\mathbb{E}\|\bar{x}^{(k+1)} - x^\star\|^2 \leq \mathbb{E}\|\bar{x}^{(k)} - x^\star\|^2 - \gamma\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) + \frac{3L\gamma}{2n}\mathbb{E}\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + \frac{\gamma^2\sigma^2}{n} \tag{70}$$

Substituting $\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)} = cX_{R,u}\check{\mathbf{z}}^{(k)}$ (see (21)) and $[\bar{\boldsymbol{z}}^{(k)}]^T = \bar{x}^{(k)} - x^\star$ into (70), we achieve

$$\begin{aligned}\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2 \leq& \mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2 - \gamma\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) + \frac{3L\gamma c^2}{2n}\|X_{R,u}\|^2\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma^2\sigma^2}{n}\\ \leq& \mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2 - \gamma\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) + \frac{3L\gamma c^2}{2n}\|X_R\|^2\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma^2\sigma^2}{n}\end{aligned}$$

where the last inequality holds because

$$\|X_{R,u}\| = \|\begin{bmatrix} I_n & 0 \end{bmatrix} X_R\| \leq \|\begin{bmatrix} I_n & 0 \end{bmatrix}\| \cdot \|X_R\| = \|X_R\|. \tag{71}$$

By setting $c^2 = \|X_L\|^2$ and recalling that $\|X_L\|\|X_R\| \leq \bar{\lambda}_n^{-1/2}$ from Lemma 4, we achieve (23).

■

C.2 Proof of Lemma 9

Proof From the second line in (18), it holds that

$$\check{\mathbf{z}}^{(k+1)} = D_1\check{\mathbf{z}}^{(k)} - \frac{\gamma}{c}\check{\mathbf{g}}^{(k)} - \frac{\gamma}{c}\check{\mathbf{s}}^{(k)}$$

We next introduce $\beta_1 = \|D_1\|$. With (16), we know that $\beta_1 = \bar{\lambda}_2^{1/2}$. By taking mean-square for both sides of the above recursion, we achieve

$$\begin{aligned}&\mathbb{E}[\|\check{\mathbf{z}}^{(k+1)}\|_F^2|\mathcal{F}^{(k)}]\\ =& \|D_1\check{\mathbf{z}}^{(k)} - \frac{\gamma}{c}\check{\mathbf{g}}^{(k)}\|_F^2 + \frac{\gamma^2}{c^2}\mathbb{E}\|\check{\mathbf{s}}^{(k)}\|_F^2\\ \overset{(a)}{\leq}& \frac{1}{t}\|D_1\|^2\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma^2}{(1-t)c^2}\|\check{\mathbf{g}}^{(k)}\|_F^2 + \frac{\gamma^2}{c^2}\|\check{\mathbf{s}}^{(k)}\|_F^2\\ \overset{(b)}{\leq}& \beta_1\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma^2}{(1-\beta_1)c^2}\|\check{\mathbf{g}}^{(k)}\|_F^2 + \frac{\gamma^2}{c^2}\|\check{\mathbf{s}}^{(k)}\|_F^2\end{aligned} \tag{72}$$

where inequality (a) holds because of the Jensen's inequality for any $t \in (0,1)$, and inequality (b) holds by letting $t = \beta_1 = \|D_1\|$. Next we bound $\|\check{\mathbf{g}}^{(k)}\|_F^2$ and $\|\check{\mathbf{s}}^{(k)}\|_F^2$. Recall the definition of $\check{\mathbf{g}}^{(k)}$ in (19a), we have

$$\|\check{\mathbf{g}}^{(k)}\|_F^2$$

$$
\begin{aligned}
&= \|\big(X_{L,\ell}Q_R + X_{L,r}Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\big)\bar{\Lambda}_R Q_R^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star))\|_F^2 \\
&\le 2\|X_{L,\ell}Q_R\bar{\Lambda}_R Q_R^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star))\|^2 \\
&\quad + 2\|X_{L,r}Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R Q_R^T(\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star))\|_F^2 \\
&\le 2\|X_{L,\ell}\|^2\|Q_R\bar{\Lambda}_R Q_R^T\|^2\|\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)\|_F^2 \\
&\quad + 2\|X_{L,r}\|^2\|Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R Q_R^T\|^2\|\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)\|_F^2 \qquad (73)
\end{aligned}
$$

First, it is easy to verify that $\|X_{L,\ell}\| \le \|X_L\|$ and $\|X_{L,r}\| \le \|X_L\|$. For example, it holds that

$$
\|X_{L,\ell}\| = \|X_L \begin{bmatrix} I_n \\ 0 \end{bmatrix}\| \le \|X_L\| \left\| \begin{bmatrix} I_n \\ 0 \end{bmatrix} \right\| = \|X_L\|. \qquad (74)
$$

Second, quantity $\|Q_R\bar{\Lambda}_R Q_R^T\|^2$ can be bounded as

$$
\begin{aligned}
\|Q_R\bar{\Lambda}_R Q_R^T\|^2 &= \lambda_{\max}(Q_R\bar{\Lambda}_R^2 Q_R^T) \\
&= \lambda_{\max}(0\cdot\frac{1}{n}\mathbb{1}\mathbb{1}^T + Q_R\bar{\Lambda}_R^2 Q_R^T) \\
&= \lambda_{\max}(Q\Lambda' Q^T) = \bar{\lambda}_2^2
\end{aligned}
$$

where $Q := [\frac{1}{\sqrt{n}}\mathbb{1}\ Q_R] \in \mathbb{R}^{n\times n}$ is defined in (59) and is an orthonormal matrix, and $\Lambda' = \mathrm{diag}\{0, \bar{\Lambda}_R^2\}$. Apparently, the largest eigenvalue of $Q\Lambda' Q^T$ is $\bar{\lambda}_2^2(W)$, which is denoted as $\bar{\lambda}_2^2$. Similarly, we can derive

$$
\|Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R Q_R^T\|^2 = \lambda_{\max}\{(1-\bar{\lambda}_i)\bar{\lambda}_i^2\} \le (1-\bar{\lambda}_n)\bar{\lambda}_2^2 \le \bar{\lambda}_2^2 \qquad (75)
$$

because $\bar{\lambda}_i := \lambda_i(\bar{W})$ and $\bar{\lambda}_2 \ge \cdots \ge \bar{\lambda}_n > 0$. Finally, quantity $\|\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)\|_F^2$ can be bounded as

$$
\begin{aligned}
\|\nabla f(\mathbf{x}^{(k)}) - \nabla f(\mathbf{x}^\star)\|_F^2 &= \|\nabla f(\mathbf{x}^{(k)}) - \nabla f(\bar{\mathbf{x}}^{(k)}) + \nabla f(\bar{\mathbf{x}}^{(k)}) - \nabla f(\mathbf{x}^\star)\|_F^2 \\
&\le 2\|\nabla f(\mathbf{x}^{(k)}) - \nabla f(\bar{\mathbf{x}}^{(k)})\|_F^2 + 2\|\nabla f(\bar{\mathbf{x}}^{(k)}) - \nabla f(\mathbf{x}^\star)\|_F^2 \\
&\overset{(a)}{\le} 2L^2\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}\|_F^2 + 4nL(f(\bar{x}^{(k)}) - f(x^\star)) \\
&\overset{(21)}{\le} 2c^2L^2\|X_R\|^2\|\check{\mathbf{z}}^{(k)}\|_F^2 + 4nL(f(\bar{x}^{(k)}) - f(x^\star)) \qquad (76)
\end{aligned}
$$

where (a) holds because $f(x)$ is convex and L-smooth. Substituting (74)–(76) into (73), we achieve

$$
\|\check{\mathbf{g}}^{(k)}\|_F^2 \le 8c^2L^2\bar{\lambda}_2^2\|X_L\|^2\|X_R\|^2\|\check{\mathbf{z}}^{(k)}\|_F^2 + 16nL\bar{\lambda}_2^2\|X_L\|^2(f(\bar{x}^{(k)}) - f(x^\star)). \qquad (77)
$$

Next we bound $\|\check{\mathbf{s}}^{(k)}\|_F^2$. Recalling the definition of $\check{\mathbf{s}}^{(k)}$ in (19c), we have

$$
\begin{aligned}
\|\check{\mathbf{s}}^{(k)}\|_F^2 &\le 2\|X_{L,\ell}\|^2\|Q_R\bar{\Lambda}_R Q_R^T\|^2\|\mathbf{s}^{(k)}\|_F^2 + 2\|X_{L,r}\|^2\|Q_R(I-\bar{\Lambda}_R)^{\frac{1}{2}}\bar{\Lambda}_R Q_R^T\|^2\|\mathbf{s}^{(k)}\|_F^2 \\
&\overset{(a)}{\le} 4\|X_L\|^2\bar{\lambda}_2^2\|\mathbf{s}^{(k)}\|_F^2
\end{aligned}
$$

$$\leq 4n\|X_L\|^2\bar{\lambda}_2^2\sigma^2 \tag{78}$$

where inequality (a) holds because of (74)–(75). Substituting (77) and (78) into (72), and taking expectation over the filtration $\mathcal{F}^{(k)}$, we achieve

$$\begin{aligned}
&\mathbb{E}\|\check{\mathbf{z}}^{(k+1)}\|_F^2\\
&\leq \big(\beta_1+\frac{8L^2\bar{\lambda}_2^2\gamma^2\|X_L\|^2\|X_R\|^2}{(1-\beta_1)}\big)\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2\\
&\quad+\frac{16nL\gamma^2\bar{\lambda}_2^2\|X_L\|^2}{c^2(1-\beta_1)}(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star))+\frac{4n\|X_L\|^2\bar{\lambda}_2^2\gamma^2\sigma^2}{c^2}\\
&\overset{(a)}{=}\big(\beta_1+\frac{8L^2\bar{\lambda}_2^2\gamma^2\|X_L\|^2\|X_R\|^2}{(1-\beta_1)}\big)\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2+\frac{16nL\gamma^2\bar{\lambda}_2^2}{1-\beta_1}(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star))+4n\bar{\lambda}_2^2\gamma^2\sigma^2\\
&\overset{(b)}{\leq}\big(\frac{1+\beta_1}{2}\big)\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2+\frac{16nL\gamma^2\bar{\lambda}_2^2}{1-\beta_1}(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star))+4n\bar{\lambda}_2^2\gamma^2\sigma^2
\end{aligned}$$

where (a) holds by setting $c^2=\|X_L\|^2$, and (b) holds by setting γ sufficiently small such that

$$\beta_1+\frac{8\bar{\lambda}_2^2\gamma^2L^2}{1-\beta_1}\|X_L\|^2\|X_R\|^2\leq\frac{1+\beta_1}{2}.$$

To satisfy the above inequality, it is enough to set (recall (17))

$$\gamma\leq\frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{4\bar{\lambda}_2L}.$$

■

C.3 Proof of Lemma 10

Proof Keep iterating (25) we achieve for $k=1,2,\ldots$ that

$$\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2\leq\big(\frac{1+\beta_1}{2}\big)^k\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2+\frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}+\frac{16n\gamma^2\bar{\lambda}_2^2L}{1-\beta_1}\sum_{\ell=0}^{k-1}\big(\frac{1+\beta_1}{2}\big)^{k-1-\ell}\big(\mathbb{E}f(\bar{x}^{(\ell)})-f(x^\star)\big) \tag{79}$$

We let $C=\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2$ be a positive constant. By taking average over $k=1,2,\ldots,T$, we achieve

$$\begin{aligned}
\frac{1}{T}\sum_{k=1}^T\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2\leq&\frac{C}{T}\sum_{k=1}^T\Big(\frac{1+\beta_1}{2}\Big)^k+\frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}\\
&+\frac{16n\gamma^2\bar{\lambda}_2^2L}{(1-\beta_1)T}\sum_{k=1}^T\sum_{\ell=0}^{k-1}\Big(\frac{1+\beta_1}{2}\Big)^{k-1-\ell}\big(\mathbb{E}f(\bar{x}^{(\ell)})-f(x^\star)\big)
\end{aligned}$$

$$
\begin{aligned}
\leq & \frac{C}{T}\sum_{k=1}^{T}\Big(\frac{1+\beta_1}{2}\Big)^k+\frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1} \\
& +\frac{16n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)T}\sum_{\ell=0}^{T-1}\Big[\sum_{k=\ell+1}^{T}\Big(\frac{1+\beta_1}{2}\Big)^{k-1-\ell}\Big]\big(\mathbb{E}f(\bar{x}^{(\ell)})-f(x^\star)\big) \\
\leq & \frac{2C}{T(1-\beta_1)}+\frac{32n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)^2 T}\sum_{k=0}^{T-1}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)+\frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1} \\
\leq & \frac{2C}{T(1-\beta_1)}+\frac{32n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)^2 T}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)+\frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}
\end{aligned}
$$

Since $\frac{1}{T+1}\sum_{k=0}^{T}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2=\big(\sum_{k=1}^{T}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2+\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2\big)/(T+1)$ and $1<\frac{1}{1-\beta_1}$, we achieve the result (26).

■

C.4 Proof of Theorem 11

Proof From (23), we have

$$
\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\leq\frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2-\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2}{\gamma}+\frac{3L}{2n\bar{\lambda}_n}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2+\frac{\gamma\sigma^2}{n}
$$

By taking average over $k=0,1,\ldots,T$, we have

$$
\begin{aligned}
& \frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big) \\
\leq & \frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma(T+1)}+\frac{3L}{2n\bar{\lambda}_n(T+1)}\sum_{k=0}^{T}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2+\frac{\gamma\sigma^2}{n} \\
\overset{(26)}{\leq} & \frac{48L^2\gamma^2\bar{\lambda}_2^2}{(T+1)(1-\beta_1)^2\bar{\lambda}_n}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)+\frac{12L\gamma^2\bar{\lambda}_2^2\sigma^2}{\bar{\lambda}_n(1-\beta_1)}+\frac{\gamma\sigma^2}{n} \\
& +\frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma(T+1)}+\frac{9L\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|^2}{2n(T+1)(1-\beta_1)\bar{\lambda}_n}.
\end{aligned}
\tag{80}
$$

If γ is sufficiently small such that

$$
\frac{48L^2\gamma^2\bar{\lambda}_2^2}{(1-\beta_1)^2\bar{\lambda}_n}\leq\frac{1}{2},
\tag{81}
$$

inequality (80) becomes

$$
\frac{1}{T+1}\sum_{k=0}^{T}\big(\mathbb{E}f(\bar{x}^{(k)})-f(x^\star)\big)
$$

$$\leq \frac{24L\gamma^2\bar{\lambda}_2^2\sigma^2}{(1-\beta_1)\bar{\lambda}_n} + \frac{2\gamma\sigma^2}{n} + \frac{2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma(T+1)} + \frac{9L\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|^2}{n(T+1)(1-\beta_1)\bar{\lambda}_n}$$
$$\overset{(a)}{\leq} \frac{24L\gamma^2\bar{\lambda}_2^2\sigma^2}{(1-\beta_1)\bar{\lambda}_n} + \frac{2\gamma\sigma^2}{n} + \frac{2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma(T+1)} + \frac{18L\gamma^2\bar{\lambda}_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2}{n(T+1)(1-\beta_1)(1-\beta)\bar{\lambda}_n} \tag{82}$$

where (a) holds because $\mathbb{E}\|\check{\mathbf{z}}^{(0)}\|^2 \leq \frac{\gamma^2\bar{\lambda}_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2}{1-\bar{\lambda}_2}$ (see Proposition 7) and $\bar{\lambda}_2 \leq (1+\beta)/2$. To satisfy (81), it is enough to let

$$\gamma \leq \frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{10L\bar{\lambda}_2}.$$

The way to choose step-size γ is adapted from (Koloskova et al., 2020, Lemma 15). For simplicity, we let $B^{(k)} = \mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)$ and

$$r_0 = 2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2,\ r_1 = \frac{2\sigma^2}{n},\ r_2 = \frac{24L\bar{\lambda}_2^2\sigma^2}{(1-\beta_1)\bar{\lambda}_n},\ r_3 = \frac{18L\bar{\lambda}_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2}{n(1-\beta_1)(1-\beta)\bar{\lambda}_n} \tag{83}$$

and inequality (82) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{r_0}{(T+1)\gamma} + r_1\gamma + r_2\gamma^2 + \frac{r_3\gamma^2}{T+1}. \tag{84}$$

Now we let

$$\gamma = \min\Big\{\frac{1}{4L}, \frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{10L\bar{\lambda}_2}, \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}, \Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}, \Big(\frac{r_0}{r_3}\Big)^{\frac{1}{3}}\Big\}. \tag{85}$$

- If $\frac{1}{4L}$ is the smallest, we let $\gamma = \frac{1}{4L}$. With $\gamma \leq \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$, $\gamma \leq \big(\frac{r_0}{r_2(T+1)}\big)^{\frac{1}{3}}$, and $\gamma \leq \big(\frac{r_0}{r_3}\big)^{\frac{1}{3}}$, (84) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{4Lr_0}{T+1} + \Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + \frac{r_3^{\frac{1}{3}}r_0^{\frac{2}{3}}}{T+1}.$$

- If $\frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{10L\bar{\lambda}_2}$ is the smallest, we let $\gamma = \frac{(1-\beta_1)\bar{\lambda}_n^{1/2}}{10L\bar{\lambda}_2}$. With $\gamma \leq \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$, $\gamma \leq \big(\frac{r_0}{r_2(T+1)}\big)^{\frac{1}{3}}$, and $\gamma \leq \big(\frac{r_0}{r_3}\big)^{\frac{1}{3}}$, (84) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{10L\bar{\lambda}_2r_0}{(1-\beta_1)(T+1)\bar{\lambda}_n^{1/2}} + \Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + \frac{r_3^{\frac{1}{3}}r_0^{\frac{2}{3}}}{T+1}. \tag{86}$$

- If $\Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$ is the smallest, we let $\gamma = \big(\frac{r_0}{r_1(T+1)}\big)^{\frac{1}{2}}$. With $\gamma \leq \big(\frac{r_0}{r_2(T+1)}\big)^{\frac{1}{3}}$ and $\gamma \leq \big(\frac{r_0}{r_3}\big)^{\frac{1}{3}}$, (84) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq 2\Big(\frac{r_0r_1}{T+1}\Big)^{\frac{1}{2}} + r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + \frac{r_3^{\frac{1}{3}}r_0^{\frac{2}{3}}}{T+1}. \tag{87}$$

- If $\Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}$ is the smallest, we let $\gamma = \Big(\frac{r_0}{r_2(T+1)}\Big)^{\frac{1}{3}}$. With $\gamma \leq \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$ and $\gamma \leq \big(\frac{r_0}{r_3}\big)^{\frac{1}{3}}$, (84) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq 2r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + \Big(\frac{r_0 r_1}{T+1}\Big)^{\frac{1}{2}} + \frac{r_3^{\frac{1}{3}} r_0^{\frac{2}{3}}}{T+1}. \tag{88}$$

- If $\big(\frac{r_0}{r_3}\big)^{\frac{1}{3}}$ is the smallest, we let $\gamma = \big(\frac{r_0}{r_3}\big)^{\frac{1}{3}}$. With $\gamma \leq \Big(\frac{r_0}{r_1(T+1)}\Big)^{\frac{1}{2}}$ and $\gamma \leq \big(\frac{r_0}{r_2(T+1)}\big)^{\frac{1}{3}}$, (84) becomes

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{2r_3^{\frac{1}{3}} r_0^{\frac{2}{3}}}{T+1} + \Big(\frac{r_0 r_1}{T+1}\Big)^{\frac{1}{2}} + r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}}.$$

Combining (88), (87) and (86), we have

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} \leq \frac{4Lr_0}{T+1} + \frac{10L\bar{\lambda}_2 r_0}{(1-\beta_1)(T+1)\bar{\lambda}_n^{1/2}} + 2\Big(\frac{r_0 r_1}{T+1}\Big)^{\frac{1}{2}} + 2r_2^{\frac{1}{3}}\Big(\frac{r_0}{T+1}\Big)^{\frac{2}{3}} + \frac{2r_3^{\frac{1}{3}} r_0^{\frac{2}{3}}}{T+1}.$$

Substituting constants r_0, r_1, r_2 and r_3 into the above inequality and regarding $\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2 = O(1)$ and $\|\nabla f(\mathbf{x})\|_F^2 = O(n)$, we achieve

$$\begin{aligned}\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} = O\Big(&\frac{4\sigma\|\bar{x}_0 - x^\star\|}{\sqrt{nT}} + \frac{16L^{\frac{1}{3}}\sigma^{\frac{2}{3}}\bar{\lambda}_2^{\frac{2}{3}}\|x_0 - x^\star\|^{\frac{4}{3}}}{(1-\beta_1)^{\frac{1}{3}}T^{\frac{2}{3}}\bar{\lambda}_n^{\frac{1}{3}}} + \frac{20L\bar{\lambda}_2\|\bar{x}_0 - x^\star\|^2}{(1-\beta_1)T\bar{\lambda}_n^{\frac{1}{2}}} \\ &+ \frac{6L^{\frac{1}{3}}\bar{\lambda}_2^{\frac{2}{3}}(\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^\star)\|^2)^{\frac{1}{3}}\|\bar{x}_0 - x^\star\|^{\frac{4}{3}}}{(1-\beta_1)^{\frac{1}{3}}(1-\beta)^{\frac{1}{3}}T\bar{\lambda}_n^{\frac{1}{3}}} + \frac{8L\|\bar{x}_0 - x^\star\|^2}{T}\Big)\end{aligned}$$

Since $\|\bar{x}_0 - x^\star\| = O(1)$ and $\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^\star)\|^2 = O(1)$, we ignore them to achieve the following clean convergence rate

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} = O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}\bar{\lambda}_2^{\frac{2}{3}}}{(1-\beta_1)^{\frac{1}{3}}T^{\frac{2}{3}}\bar{\lambda}_n^{\frac{1}{3}}} + \frac{L\bar{\lambda}_2}{(1-\beta_1)T\bar{\lambda}_n^{\frac{1}{2}}} + \frac{L^{\frac{1}{3}}\bar{\lambda}_2^{\frac{2}{3}}}{(1-\beta_1)^{\frac{1}{3}}(1-\beta)^{\frac{1}{3}}T\bar{\lambda}_n^{\frac{1}{3}}} + \frac{L}{T}\Big) \tag{89}$$

With $\beta_1^2 = \bar{\lambda}_2 \leq (1+\beta)/2$, we have

$$1 - \beta_1 = \frac{1-\beta_1^2}{1+\beta_1} \geq \frac{1-\beta}{2(1+\beta_1)} \geq \frac{1-\beta}{4}. \tag{90}$$

Substituting (90) to (85) and (89) and regarding $\bar{\lambda}_2$ and $\bar{\lambda}_n$ as constants (note that $\bar{\lambda}_n$ is bounded away from 0 under Assumption 2. For example, if $\bar{W} = (3I + W)/4$, we have $\bar{\lambda}_2 \geq \bar{\lambda}_n \geq 1/2$), we have the following result

$$\frac{1}{T+1}\sum_{k=0}^{T} B^{(k)} = O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{(1-\beta)^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{L}{(1-\beta)T} + \frac{L^{\frac{1}{3}}}{(1-\beta)^{\frac{2}{3}}T} + \frac{L}{T}\Big)$$

$$= O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{L^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{(1-\beta)^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{L^{\frac{1}{3}}}{(1-\beta)^{\frac{2}{3}}T} + \frac{L}{(1-\beta)T}\Big),$$

which is the result in Theorem 11. ■

Appendix D. Convergence Analysis for Strongly-Convex Scenario

D.1 Proof of Lemma 16

Proof Since each $f_i(x)$ is strongly convex, it holds that

$$f_i(x) - f_i(y) + \frac{\mu}{2}\|x-y\|^2 \leq \langle \nabla f_i(x), x-y\rangle, \quad \forall\ x, y \in \mathbb{R}^d$$

Let $x = x_i^{(k)}$ and $y = x^\star$, we have

$$f_i(x_i^{(k)}) - f_i(x^\star) + \frac{\mu}{2}\|x_i^{(k)} - x^\star\|^2 \leq \langle \nabla f_i(x_i^{(k)}), x_i^{(k)} - x^\star\rangle.$$

Following arguments from (66) to (68), and replacing the bound in (67) with

$$\begin{aligned}
&\frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{x}^{(k)} - x^\star, \nabla f_i(x_i^{(k)}) - \nabla f_i(x^\star)\rangle \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \langle \bar{x}^{(k)} - x_i^{(k)}, \nabla f_i(x_i^{(k)})\rangle + \frac{2\gamma}{n}\sum_{i=1}^n \langle x_i^{(k)} - x^\star, \nabla f_i(x^{(k)})\rangle \\
&\overset{(a)}{\geq} \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\bar{x}^{(k)}) - f_i(x_i^{(k)}) - \frac{L}{2}\|\bar{x}^{(k)} - x_i^{(k)}\|^2\Big) + \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(x_i^{(k)}) - f_i(x^\star) + \frac{\mu}{2}\|x_i^{(k)} - x^\star\|^2\Big) \\
&= \frac{2\gamma}{n}\sum_{i=1}^n \Big(f_i(\bar{x}^{(k)}) - f_i(x^\star)\Big) - \frac{\gamma L}{n}\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2 + \frac{\gamma\mu}{n}\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_F^2 \\
&\geq 2\gamma\big(f(\bar{x}^{(k)}) - f(x^\star)\big) - \frac{\gamma(L+\mu)}{n}\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2 + \frac{\gamma\mu}{2}\|\bar{x}^{(k)} - x^\star\|_F^2,
\end{aligned}$$

we achieve a slightly different bound from (69):

$$\begin{aligned}
&\|\bar{x}^{(k)} - x^\star - \frac{\gamma}{n}\sum_{i=1}^n \nabla f_i(x_i^{(k)})\|^2 \\
&\leq (1 - \frac{\gamma\mu}{2})\|\bar{x}^{(k)} - x^\star\|^2 - 2\gamma(1 - 2L\gamma)\big(f(\bar{x}^{(k)}) - f(x^\star)\big) + \Big(\frac{\gamma(L+\mu)}{n} + \frac{2\gamma^2L^2}{n}\Big)\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2 \\
&\leq (1 - \frac{\gamma\mu}{2})\|\bar{x}^{(k)} - x^\star\|^2 - \gamma\big(f(\bar{x}^{(k)}) - f(x^\star)\big) + \frac{5\gamma L}{2n}\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}\|_F^2 \qquad (91)
\end{aligned}$$

where the last inequality holds when $\gamma \leq \frac{1}{4L}$. With (91), we can follow arguments (70)-(71) to achieve the result in (29). ■

D.2 Proof of Lemma 17

Proof Recall from (79) that

$$\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 \le \Big(\frac{1+\beta_1}{2}\Big)^k \mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2 + \frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1} + \frac{16n\gamma^2\bar{\lambda}_2^2 L}{1-\beta_1}\sum_{\ell=0}^{k-1}\Big(\frac{1+\beta_1}{2}\Big)^{k-1-\ell}\big(\mathbb{E}f(\bar{x}^{(\ell)}) - f(x^\star)\big)$$

By taking the weighted sum over $k = 1, 2, \ldots, T$, we achieve

$$\sum_{k=1}^T h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 \le \mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2\sum_{k=1}^T h_k\Big(\frac{1+\beta_1}{2}\Big)^k + \frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}\sum_{k=1}^T h_k + \frac{16n\gamma^2\bar{\lambda}_2^2 L}{1-\beta_1}\sum_{k=1}^T h_k\sum_{\ell=0}^{k-1}\Big(\frac{1+\beta_1}{2}\Big)^{k-1-\ell}\big(\mathbb{E}f(\bar{x}^{(\ell)}) - f(x^\star)\big) \quad (92)$$

Since h_k satisfies condition (32), it holds (we define $B^{(\ell)} = \mathbb{E}f(\bar{x}^{(\ell)}) - f(x^\star)$) that

$$\begin{aligned}
\sum_{k=1}^T h_k\sum_{\ell=0}^{k-1}\Big(\frac{1+\beta_1}{2}\Big)^{k-1-\ell}B^{(\ell)} &\le \big(1+\frac{1-\beta_1}{4}\big)\sum_{k=1}^T\sum_{\ell=0}^{k-1} h_\ell\Big[\big(1+\frac{1-\beta_1}{4}\big)\big(\frac{1+\beta_1}{2}\big)\Big]^{k-1-\ell}B^{(\ell)} \\
&\le 2\sum_{k=1}^T\sum_{\ell=0}^{k-1} h_\ell\big(\frac{3+\beta_1}{4}\big)^{k-1-\ell}B^{(\ell)} \\
&\le 2\sum_{\ell=0}^{T-1} h_\ell\sum_{k=\ell+1}^{T}\big(\frac{3+\beta_1}{4}\big)^{k-1-\ell}B^{(\ell)} \\
&\le \frac{8}{1-\beta_1}\sum_{\ell=0}^{T-1} h_\ell B^{(\ell)} \le \frac{8}{1-\beta_1}\sum_{\ell=0}^{T} h_\ell B^{(\ell)}
\end{aligned}$$

Substituting the above inequality into (92), we achieve

$$\sum_{k=1}^T h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 \le C\sum_{k=1}^T h_k\Big(\frac{1+\beta_1}{2}\Big)^k + \frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}\sum_{k=0}^T h_k + \frac{128n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)^2}\sum_{\ell=0}^T h_\ell B^{(\ell)} \quad (93)$$

where $C = \mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2$. Adding $h_0 C$ to both sides of (93), we achieve

$$\sum_{k=0}^T h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 \le C\sum_{k=0}^T h_k\Big(\frac{1+\beta_1}{2}\Big)^k + \frac{8n\gamma^2\bar{\lambda}_2^2\sigma^2}{1-\beta_1}\sum_{k=0}^T h_k + \frac{128n\gamma^2\bar{\lambda}_2^2 L}{(1-\beta_1)^2}\sum_{\ell=0}^T h_\ell B^{(\ell)}. \quad (94)$$

Furthermore, with condition (32), we have $h_k \le h_0(1+\frac{1-\beta_1}{4})^k$ for any $k = 0, 1, \ldots$. This implies

$$\sum_{k=0}^T h_k\Big(\frac{1+\beta_1}{2}\Big)^k \le h_0\sum_{k=0}^T(1+\frac{1-\beta_1}{4})^k(\frac{1+\beta_1}{2})^k \le h_0\sum_{k=0}^T(\frac{3+\beta_1}{4})^k \le \frac{4h_0}{1-\beta_1} \quad (95)$$

Substituting (95) into (94) and dividing both sides by $H_T = \sum_{k=0}^{T} h_k$, we achieve the final result in (31). ■

D.3 Proof of Theorem 18

The following proof is inspired by (Stich, 2019b).

Proof With descent inequality (29), we have

$$\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star) \leq (1 - \frac{\gamma\mu}{2})\frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2}{\gamma} - \frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2}{\gamma} + \frac{5L}{2n\bar{\lambda}_n}\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n}$$

Taking the weighted average over k, it holds that (we let $B^{(k)} = \mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)$)

$$\begin{aligned}
&\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} \\
\leq &\frac{1}{H_T}\sum_{k=0}^{T} h_k\Big(\frac{(1-\frac{\gamma\mu}{2})\mathbb{E}\|\bar{\boldsymbol{z}}^{(k)}\|^2}{\gamma} - \frac{\mathbb{E}\|\bar{\boldsymbol{z}}^{(k+1)}\|^2}{\gamma}\Big) + \frac{5L}{2nH_T\bar{\lambda}_n}\sum_{k=0}^{T} h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n}.
\end{aligned}$$

If we let $h_k = (1 - \frac{\gamma\mu}{2})h_{k+1}$ for $k = 0, 1, \ldots$, the above inequality becomes

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} \leq \frac{h_0\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{H_T\gamma} + \frac{5L}{2nH_T\bar{\lambda}_n}\sum_{k=0}^{T} h_k\mathbb{E}\|\check{\mathbf{z}}^{(k)}\|_F^2 + \frac{\gamma\sigma^2}{n}. \tag{96}$$

Since $h_k = (1 - \frac{\gamma\mu}{2})h_{k+1}$, we have

$$h_k = h_\ell\big(\frac{1}{1-\frac{\gamma\mu}{2}}\big)^{k-\ell}, \text{ for any } k \geq 0 \text{ and } 0 \leq \ell \leq k.$$

If γ is sufficiently small such that

$$\frac{1}{1-\frac{\gamma\mu}{2}} \leq 1 + \frac{1-\beta_1}{4}, \quad (\text{it is enough to set } \gamma \leq \tfrac{1-\beta_1}{2\mu})$$

then $\{h_k\}_{k=0}^{\infty}$ satisfy condition (32). As a result, we can substitute inequality (31) into (96) to achieve

$$\begin{aligned}
\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} \leq &\frac{h_0\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{H_T\gamma} + \frac{\gamma\sigma^2}{n} + \frac{10LCh_0}{nH_T(1-\beta_1)\bar{\lambda}_n} \\
&+ \frac{20L\gamma^2\bar{\lambda}_2^2\sigma^2}{\bar{\lambda}_n(1-\beta_1)} + \frac{320\gamma^2\bar{\lambda}_2^2L^2}{(1-\beta_1)^2\bar{\lambda}_n}\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)}.
\end{aligned}$$

If γ is sufficiently small such that

$$\frac{320\gamma^2\bar{\lambda}_2^2L^2}{(1-\beta_1)^2\bar{\lambda}_n} \leq \frac{1}{2}, \quad (\text{it is enough to set } \gamma \leq \tfrac{1-\beta_1}{26L}\big(\tfrac{\bar{\lambda}_n^{1/2}}{\bar{\lambda}_2}\big)) \tag{97}$$

it holds that

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} \le \frac{2h_0\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{H_T\gamma} + \frac{20LCh_0}{nH_T(1-\beta_1)\bar{\lambda}_n} + \frac{2\gamma\sigma^2}{n} + \frac{40L\gamma^2\bar{\lambda}_2^2\sigma^2}{\bar{\lambda}_n(1-\beta_1)}.$$

Since $H_T \ge h_T = h_0(1-\frac{\gamma\mu}{2})^{-T}$, we have

$$\begin{aligned}
&\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} \\
\le&\frac{2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma}(1-\frac{\gamma\mu}{2})^T + \frac{20LC}{n(1-\beta_1)\bar{\lambda}_n}(1-\frac{\gamma\mu}{2})^T + \frac{2\gamma\sigma^2}{n} + \frac{40L\gamma^2\bar{\lambda}_2^2\sigma^2}{\bar{\lambda}_n(1-\beta_1)} \\
\le&\big(\frac{2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma} + \frac{20LC}{n(1-\beta_1)\bar{\lambda}_n}\big)\exp(-\frac{\gamma\mu T}{2}) + \frac{2\gamma\sigma^2}{n} + \frac{40L\gamma^2\bar{\lambda}_2^2\sigma^2}{\bar{\lambda}_n(1-\beta_1)} \\
\le&\big(\frac{2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2}{\gamma} + \frac{40L\gamma^2\bar{\lambda}_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2}{n(1-\beta_1)(1-\beta)\bar{\lambda}_n}\big)\exp(-\frac{\gamma\mu T}{2}) + \frac{2\gamma\sigma^2}{n} + \frac{40L\gamma^2\bar{\lambda}_2^2\sigma^2}{\bar{\lambda}_n(1-\beta_1)} \qquad (98)
\end{aligned}$$

where the last inequality holds because $C = \mathbb{E}\|\check{\mathbf{z}}^{(0)}\|_F^2 \le \frac{\gamma^2\bar{\lambda}_2^2\|\nabla f(\mathbf{x}^\star)\|_F^2}{1-\bar{\lambda}_2} = O(\frac{n\gamma^2\bar{\lambda}_2^2}{1-\bar{\lambda}_2})$ (see Proposition 7) and $\bar{\lambda}_2 \le (1+\beta)/2$, and γ needs to satisfy condition (97). Now we let

$$\gamma = \min\big\{\frac{1}{4L}, \frac{1-\beta_1}{26L}\big(\frac{\bar{\lambda}_n^{1/2}}{\bar{\lambda}_2}\big), \frac{2\ln(2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2T^2/[\sigma^2(1-\beta)])}{\mu T}\big\}. \qquad (99)$$

- If $\frac{2\ln(2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2T^2/[\sigma^2(1-\beta)])}{\mu T}$ is smallest, we set

$$\gamma = \frac{2\ln(2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2T^2/[\sigma^2(1-\beta)])}{\mu T} \quad \text{so that} \quad \exp(-\frac{\gamma\mu T}{2}) = \frac{\sigma^2(1-\beta)}{2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2T^2}$$

  Substituting the above γ into (98) and regarding $\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2 = O(1)$, we achieve

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} = \tilde{O}\left(\frac{\sigma^2}{\mu nT} + \frac{L\sigma^2\bar{\lambda}_2^2}{\mu^2(1-\beta_1)T^2\bar{\lambda}_n}\right). \qquad (100)$$

- If $\frac{1-\beta_1}{26L}\big(\frac{\bar{\lambda}_n^{1/2}}{\bar{\lambda}_2}\big)$ is smallest, we set $\gamma = \frac{1-\beta_1}{26L}\big(\frac{\bar{\lambda}_n^{1/2}}{\bar{\lambda}_2}\big)$. Since $\gamma \le \frac{2\ln(2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2T^2/[\sigma^2(1-\beta)])}{\mu T}$ and $\gamma \le \frac{1}{4L}$, (98) becomes

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} = \tilde{O}\left(\frac{L\bar{\lambda}_2}{(1-\beta_1)\bar{\lambda}_n^{\frac{1}{2}}}\exp\{-\frac{\mu(1-\beta_1)}{L}(\frac{\bar{\lambda}_n^{\frac{1}{2}}}{\bar{\lambda}_2})T\} + \frac{\sigma^2}{\mu nT} + \frac{L\sigma^2\bar{\lambda}_2^2}{\mu^2(1-\beta_1)T^2\bar{\lambda}_n}\right).$$

- If $\frac{1}{4L}$ is smallest, we set $\gamma = \frac{1}{4L}$. Since $\gamma \leq \frac{1-\beta_1}{26L}\Big(\frac{\bar{\lambda}_n^{1/2}}{\bar{\lambda}_2}\Big)$ and $\gamma \leq \frac{1}{4L}$ and $\gamma \leq \frac{2\ln(2n\mu\mathbb{E}\|\tilde{\boldsymbol{z}}^{(0)}\|^2T^2/[\sigma^2(1-\beta)])}{\mu T}$, (98) becomes

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} = \tilde{O}\left(L\exp\{-\frac{\mu T}{L}\} + \frac{\sigma^2}{\mu n T} + \frac{L\sigma^2\bar{\lambda}_2^2}{\mu^2(1-\beta_1)T^2\bar{\lambda}_n}\right). \tag{101}$$

Combining (100) – (101), substituting relation (90) to bound $1-\beta_1$, we achieve

$$\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} = \tilde{O}\left(\frac{\sigma^2}{\mu n T} + \frac{L\sigma^2\bar{\lambda}_2^2}{\mu^2(1-\beta)T^2\bar{\lambda}_n} + \frac{L\bar{\lambda}_2\exp\{-\frac{\mu(1-\beta)}{L}\big(\frac{\bar{\lambda}_n^{\frac{1}{2}}}{\bar{\lambda}_2}\big)T\}}{(1-\beta)\bar{\lambda}_n^{\frac{1}{2}}} + L\exp\{-\frac{\mu T}{L}\}\right).$$

Ignoring constants $\bar{\lambda}_2$ and $\bar{\lambda}_n$ (these quantities can be regarded as constants when $\bar{\lambda}_n$ is bounded away from zero. For example, if $\bar{W} = (3I+W)/4$, we have $\bar{\lambda}_2 \geq \bar{\lambda}_n \geq 1/2$.), and recalling the relation in (90), we achieve

$$\begin{aligned}\frac{1}{H_T}\sum_{k=0}^{T} h_k B^{(k)} &= \tilde{O}\left(\frac{\sigma^2}{\mu n T} + \frac{L\sigma^2}{\mu^2(1-\beta)T^2} + \frac{L\exp\{-\frac{\mu(1-\beta)}{L}T\}}{(1-\beta)} + L\exp\{-\frac{\mu T}{L}\}\right)\\ &= \tilde{O}\left(\frac{\sigma^2}{\mu n T} + \frac{L\sigma^2}{\mu^2(1-\beta)T^2} + \frac{L\exp\{-\frac{\mu(1-\beta)}{L}T\}}{(1-\beta)}\right),\end{aligned}$$

which is the result in Theorem 18.

■

Appendix E. Proof of Theorem 22

Proof We consider the minimization problem of the form (1) with $f_i(x) = \frac{1}{2}\|x\|^2$ where $x \in \mathbb{R}$ and with $W = \beta I_n + \frac{1}{n}(1-\beta)\mathbb{1}_n\mathbb{1}_n^T$. Note that the eigenvalues of W are $\lambda_1(W) = 1$ and $\lambda_k(W) = \beta$, $\forall\, 2 \leq k \leq n$, Under such setting, it holds that $f_i(x) = f_j(x)$ for any $i, j \in [n]$ and there is no heterogeneity, *i.e.*, $b^2 = 0$ and $\rho(W - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T) = \beta$. The D-SGD algorithm in this setting will iterate as follows:

$$\mathbf{x}^{(k+1)} = W(\mathbf{x}^{(k)} - \gamma\mathbf{x}^{(k)} - \gamma\mathbf{s}^{(k)}) = (1-\gamma)W\mathbf{x}^{(k)} - \gamma\bar{W}\mathbf{s}^{(k)} \tag{102}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a vector, and $\mathbf{s} \in \mathbb{R}^n$ is the gradient noise. With (102), we have

$$\bar{x}^{(k+1)} = (1-\gamma)\bar{x}^{(k)} - \gamma\bar{s}^{(k)} \tag{103}$$

and

$$\bar{\mathbf{x}}^{(k+1)} = (1-\gamma)\frac{1}{n}\mathbb{1}\mathbb{1}^T\mathbf{x}^{(k)} - \gamma\frac{1}{n}\mathbb{1}\mathbb{1}^T\mathbf{s}^{(k)}.$$

Moreover, we assume each element of the noise follows standard Gaussian distribution, *i.e.*, $s_i^{(k)} \sim \mathcal{N}(0, \sigma^2)$, and $s_i^{(k)}$ is independent of each other for any k and i. We also assume the gradient noise $\mathbf{s}^{(k)}$ is independent of $\mathbf{x}^{(\ell)}$ for any $\ell \leq k$. With these assumptions, it holds that $\mathbb{E}[\mathbf{s}^{(k)}(\mathbf{s}^{(k)})^T] = \sigma^2 I \in \mathbb{R}^{n\times n}$. Subtracting the above recursion from (102), we have

$$\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)} = (1-\gamma)(W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)}) - \gamma(W - \frac{1}{n}\mathbb{1}\mathbb{1}^T)\mathbf{s}^{(k)}. \tag{104}$$

We next define matrix $P = W - \frac{1}{n}\mathbb{1}\mathbb{1}^T$. Note that $\mathbf{s}^{(k)}$ is independent of $\mathbf{x}^{(k)}$. By taking the mean-square-expectation over both sides of the above equality, we have

$$\begin{aligned}
&\mathbb{E}\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\|^2 \\
=&\mathbb{E}\|(1-\gamma)P(\mathbf{x}^{(k)} - \bar{\mathbf{x}}^{(k)})\|^2 + \gamma^2\mathbb{E}\|P\mathbf{s}^{(k)}\|^2 \\
\overset{(104)}{=}&\mathbb{E}\|(1-\gamma)^2P^2(\mathbf{x}^{(k-1)} - \bar{\mathbf{x}}^{(k-1)}) - \gamma(1-\gamma)P^2\mathbf{s}^{(k-1)}\|^2 + \gamma^2\mathbb{E}\|P\mathbf{s}^{(k)}\|^2 \\
=&\mathbb{E}\|(1-\gamma)^2P^2(\mathbf{x}^{(k-1)} - \bar{\mathbf{x}}^{(k-1)})\|^2 + \gamma^2(1-\gamma)^2\mathbb{E}\|P^2\mathbf{s}^{(k-1)}\|^2 + \gamma^2\mathbb{E}\|P\mathbf{s}^{(k)}\|^2 \\
=&\cdots \\
=&\|(1-\gamma)^{k+1}P^{k+1}(\mathbf{x}^{(0)} - \bar{\mathbf{x}}^{(0)})\|^2 + \gamma^2\sum_{\ell=0}^{k}\mathbb{E}\|(1-\gamma)^\ell P^{\ell+1}\mathbf{s}^{(k-\ell)}\|^2
\end{aligned} \tag{105}$$

In the above derivations, we used the fact that $\mathbf{s}^{(k)}$ is independent of $\mathbf{x}^{(k)}$ for any k. Without loss of generality, we can assume $\gamma \leq \frac{2}{L} = 2$, otherwise the iteration explodes. Since $x^\star = 0$, by (103), we similarly have

$$\begin{aligned}
\mathbb{E}\|\bar{x}^{(k+1)} - \bar{x}^\star\|^2 =& (1-\gamma)^{2(k+1)}\|\bar{x}^{(0)} - \bar{x}^\star\|^2 + \gamma^2\sum_{\ell=0}^{k}\mathbb{E}\|(1-\gamma)^\ell\bar{s}^{(k-\ell)}\|^2 \\
\geq&(1-\gamma)^{2(k+1)}\|\bar{x}^{(0)} - \bar{x}^\star\|^2.
\end{aligned} \tag{106}$$

Next we examine $\mathbb{E}\|(1-\gamma)^\ell P^{\ell+1}\mathbf{s}^{(k-\ell)}\|^2$:

$$\begin{aligned}
&\mathbb{E}\|(1-\gamma)^\ell P^{\ell+1}\mathbf{s}^{(k-\ell)}\|^2 \\
=&\,(1-\gamma)^{2\ell}\mathbb{E}\{\mathrm{tr}([\mathbf{s}^{(k-\ell)}]^T P^{\ell+1}P^{\ell+1}\mathbf{s}^{(k-\ell)})\} \\
=&\,(1-\gamma)^{2\ell}\mathbb{E}\{\mathrm{tr}(P^{2(\ell+1)}\mathbf{s}^{(k-\ell)}[\mathbf{s}^{(k-\ell)}]^T)\} \\
=&\,(1-\gamma)^{2\ell}\mathrm{tr}\big(P^{2(\ell+1)}\mathbb{E}\{\mathbf{s}^{(k-\ell)}[\mathbf{s}^{(k-\ell)}]^T\}\big) \\
\overset{(a)}{=}&\,\sigma^2(1-\gamma)^{2\ell}\mathrm{tr}(P^{2(\ell+1)}) \\
\overset{(b)}{=}&\,\sigma^2(1-\gamma)^{2\ell}\mathrm{tr}(U\Lambda_P^{2(\ell+1)}U^T) \\
=&\,\sigma^2(1-\gamma)^{2\ell}\mathrm{tr}(\Lambda_P^{2(\ell+1)}U^TU) \\
=&\,\sigma^2(1-\gamma)^{2\ell}\mathrm{tr}(\Lambda_P^{2(\ell+1)}) \\
\overset{(c)}{=}&\,(n-1)\sigma^2(1-\gamma)^{2\ell}\beta^2
\end{aligned} \tag{107}$$

where (a) holds because $\mathbb{E}[\mathbf{s}^{(k)}(\mathbf{s}^{(k)})^T] = \sigma^2 I \in \mathbb{R}^{n\times n}$ for any k, and (b) holds because $\Lambda_P = \Lambda_{W-\frac{1}{n}\mathbb{1}\mathbb{1}^T} = \{0, \beta, \ldots, \beta\}$. With (107), we have

$$\begin{aligned}\gamma^2\sum_{\ell=0}^{k}\mathbb{E}\|(1-\gamma)^\ell P^{\ell+1}\mathbf{s}^{(k-\ell)}\|^2 \geq&(n-1)\gamma^2\sigma^2\beta^2\sum_{\ell=0}^{k}(1-\gamma)^{2\ell}\beta^{2\ell}\\ =&(n-1)\gamma^2\sigma^2\beta^2\frac{1-(1-\gamma)^{2(k+1)}\beta^{2(k+1)}}{1-(1-\gamma)^2\beta^2}.\end{aligned}\tag{108}$$

Substituting (108) into (105), we achieve

$$\mathbb{E}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|^2 \geq (n-1)\gamma^2\sigma^2\beta^2\frac{1-(1-\gamma)^{2k}\beta^{2k}}{1-(1-\gamma)^2\beta^2} \geq \frac{n}{2}\gamma^2\sigma^2\beta^2\frac{1-(1-\gamma)^{2k}\beta^{2k}}{1-(1-\gamma)^2\beta^2}.$$

Since $\frac{1}{n}\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^\star\|^2 = \mathbb{E}\|\bar{x}^{(k)}-x^\star\|^2 + \frac{1}{n}\mathbb{E}\|\mathbf{x}^{(k)}-\bar{\mathbf{x}}^{(k)}\|^2$, with (106) and (109), we have

$$\frac{1}{n}\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^\star\|^2 \geq (1-\gamma)^{2k}\|\bar{x}^{(0)}-x^\star\|^2 + \frac{\sigma^2\gamma^2\beta^2}{2}\frac{1-(1-\gamma)^{2k}\beta^{2k}}{1-(1-\gamma)^2\beta^2}\tag{109}$$

To guarantee D-SGD to achieve the linear speedup, we require that $\frac{1}{n}\mathbb{E}\|\mathbf{x}^{(k)}-\mathbf{x}^\star\|^2 \lesssim \frac{\sigma^2}{nk}$ holds for any sufficiently large k (note that P-SGD will achieve the linear speedup $\frac{\sigma^2}{nk}$ for the strongly-convex scenario). Thus, it is necessary to have that

$$\underbrace{(1-\gamma)^{2k}\|\bar{x}^{(0)}-x^\star\|^2}_{\text{I}} + \underbrace{\frac{\sigma^2\gamma^2\beta^2}{2}\frac{1-(1-\gamma)^{2k}\beta^{2k}}{1-(1-\gamma)^2\beta^2}}_{\text{II}} \leq \frac{\sigma^2}{nk}\tag{110}$$

up to some absolute constants. In other words, there exists transient time k_{trans} such that for all $k \geq k_{\text{trans}}$, the above inequality holds up to some absolute constants. We omit the potential constant factors for simplicity since our analysis can be easily adapted to the case with some absolute constants on the two sides of (110) and the rate remains the same.

Next we find $\gamma := \gamma(k)$ such that (110) holds and show that such γ exists only when $k_{\text{trans}} = \tilde{\Omega}\left(\frac{n\beta^2}{1-\beta}\right)$.

- If $\gamma \geq \frac{1}{2}$, then $\text{II} \geq \frac{\sigma^2\beta^2}{8}$ which means (110) can only holds when $k = O(1)$. Therefore, to let (110) holds for all sufficiently large k, one must consider $\gamma < \frac{1}{2}$.
- If $\gamma < \frac{1}{2}$, then by inequality $\exp(\frac{x}{1+x}) \leq 1+x$ for $x > -1$, we have

  $$\frac{\sigma^2}{nk} \geq (1-\gamma)^{2k}\|\bar{x}^{(0)}-x^\star\|^2 \geq \exp(-2k\gamma/(1-\gamma))\|\bar{x}^{(0)}-x^\star\|^2 \geq e^{-4k\gamma}\|\bar{x}^{(0)}-x^\star\|^2$$

  where the last inequality is due to the fact $\gamma < \frac{1}{2}$. Therefore, (110) implies

  $$\gamma \geq \frac{\ln(nk\|\bar{x}^{(0)}-x^\star\|^2/\sigma^2)}{4k} \triangleq \gamma^\star.\tag{111}$$

On the other hand, since (110) implies $(1-\gamma)^{2k} \leq \frac{\sigma^2}{nk\|\bar{x}^{(0)}-x^\star\|^2}$, we have

$$\text{II} \geq \frac{\sigma^2\gamma^2\beta^2}{2}\frac{1-\frac{\sigma^2}{nk\|\bar{x}^{(0)}-x^\star\|^2}\beta^{2k}}{1-(1-\gamma)^2\beta^2} \geq \frac{\sigma^2\gamma^2\beta^2}{2}\frac{1-\frac{\sigma^2}{nk\|\bar{x}^{(0)}-x^\star\|^2}}{1-(1-\gamma)^2\beta^2}$$

where we assume k sufficiently large such that $\frac{\sigma^2}{nk\|\bar{x}^{(0)}-x^\star\|^2} \leq 1$. Therefore, (110) also implies

$$\begin{aligned}
&\frac{\sigma^2}{nk} \geq \frac{\sigma^2\gamma^2\beta^2}{2}\frac{1-\frac{\sigma^2}{nk\|\bar{x}^{(0)}-x^\star\|^2}}{1-(1-\gamma)^2\beta^2} \\
\iff &\left(nk - \frac{\sigma^2}{\|\bar{x}^{(0)}-x^\star\|^2}+2\right)\gamma^2 - 4\gamma \leq \frac{2}{\beta^2}-2. \qquad (112)
\end{aligned}$$

Note that for $\gamma > 0$, $f(\gamma) \triangleq (nk - \frac{\sigma^2}{\|\bar{x}^{(0)}-x^\star\|^2}+2)\gamma^2 - 4\gamma$ decreases first then keeps increasing with respect to γ, so (111) and (112) are compatible only when

$$\left(nk - \frac{\sigma^2}{\|\bar{x}^{(0)}-x^\star\|^2}+2\right)(\gamma^\star)^2 - 4\gamma^\star \leq \frac{2}{\beta^2}-2. \qquad (113)$$

Considering k large enough such that

$$\frac{\sigma^2}{k\|\bar{x}^{(0)}-x^\star\|^2} \leq 1 \quad \text{and} \quad \ln(nk\|\bar{x}^{(0)}-x^\star\|^2/\sigma^2) \geq 1,$$

then (113) holds only when

$$\begin{aligned}
\frac{2}{\beta^2}-2 \geq &(nk-k)(\gamma^\star)^2 - 4\gamma^\star \\
\geq &\frac{(n-1)\ln(nk\|\bar{x}^{(0)}-x^\star\|^2/\sigma^2)^2 - 4\ln(nk\|\bar{x}^{(0)}-x^\star\|^2/\sigma^2)}{16k} \\
= &\frac{(n-5)\ln(nk\|\bar{x}^{(0)}-x^\star\|^2/\sigma^2)^2}{16k}
\end{aligned}$$

which leads to $k \geq \tilde{\Omega}\left(\frac{n\beta^2}{1-\beta^2}\right) = \tilde{\Omega}\left(\frac{n\beta^2}{1-\beta}\right)$. Therefore, we reach the conclusion that $k_{\text{trans}} = \tilde{\Omega}\left(\frac{n\beta^2}{1-\beta}\right)$.

■

Appendix F. Convergence of Algorithm 2

In this section we will establish the convergence of D^2/Exact-Diffusion with multiple gossip steps. As we have discussed in Sec. 6.2, there are two fundamental differences between the vanilla D^2/Exact-Diffusion and its variant with multiple gossip steps:

- **Gradient accumulation.** For each outer loop k, each node i in D^2/Exact-Diffusion with multiple gossip steps will draw R independent data samples $\{\xi_i^{(k,r)}\}_{r=1}^R$ to compute the stochastic gradient with $g_i^{(k)} = \frac{1}{R}\sum_{r=1}^R \nabla F(x_i^{(k)};\xi_i^{(k,r)})$. This will result in a reduced gradient noise:
$$\mathbb{E}[\|g_i^{(k)} - \nabla f_i(x_i^{(k)})\|^2 | \mathcal{F}^{(k-1)}] \le \frac{\sigma^2}{R}.$$
- **Fast gossip averaging.** With fast gossip averaging (i.e., Algorithm 3), the weight matrix utilized in D^2/Exact-Diffusion with multiple gossip steps is $\bar{M}$ instead of $\bar{W}$ (see recursions (39) and (40)). In addition, it is established in Proposition 24 that $\lambda_k(\bar{M}) \in [\frac{1}{4n}, \frac{3}{4n}]$ for $2 \le k \le n$ if $R = \lceil \frac{\ln(n)+4}{\sqrt{1-\beta}} \rceil$.

We will utilize these facts to facilitate the analysis for D^2/Exact-Diffusion with multiple gossip steps.

F.1 Proof of Proposition 23

This proposition directly follows the results of (Liu and Morse, 2011). We provide the proof for completeness. Note that (37) can be transformed into a first-order iteration as follows:
$$\begin{bmatrix} M^{(r+1)} \\ M^{(r)} \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\eta)W & -\eta I \\ I & 0 \end{bmatrix}}_{W_2} \begin{bmatrix} M^{(r)} \\ M^{(r-1)} \end{bmatrix}.$$
By (Liu and Morse, 2011, Proposition 3), the projection of augmented matrix W_2 on the subspace orthogonal to $\mathbb{1}_n$ is a contraction with spectral norm of $\frac{\beta}{1+\sqrt{1-\beta^2}}$. Since $\frac{\beta}{1+\sqrt{1-\beta^2}} \le 1-\sqrt{1-\beta}$ for any $0 \le \beta \le 1$, we have
$$\left\| \begin{bmatrix} M^{(r)} \\ M^{(r-1)} \end{bmatrix} \mathbf{z} \right\| \le \left(1-\sqrt{1-\beta}\right)^r \left\| \begin{bmatrix} M^{(0)} \\ M^{(-1)} \end{bmatrix} \mathbf{z} \right\| = \sqrt{2}\left(1-\sqrt{1-\beta}\right)^r \|\mathbf{z}\|$$
for any $\mathbf{z} \perp \mathbb{1}_n$. We thus have, for any $\mathbf{z} \perp \mathbb{1}_n$, that
$$\|M^{(r)}\mathbf{z}\| \le \sqrt{2}\left(1-\sqrt{1-\beta}\right)^r \|\mathbf{z}\| \quad \text{i.e.,} \quad \rho(M^{(r)} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T) \le \sqrt{2}\left(1-\sqrt{1-\beta}\right)^r.$$

F.2 Proof of Proposition 24

If we choose $R = \lceil \frac{\ln(n)+4}{\sqrt{1-\beta}} \rceil$, $\tau = \frac{1}{2n}$ and denote $M \triangleq M^{(R)}$ and $\bar{M} \triangleq (1-\tau)M + \tau I$. It follows from Proposition 23 that
$$\sqrt{2}\left(1-\sqrt{1-\beta}\right)^R = \sqrt{2}\exp(R\ln(1-\sqrt{1-\beta})) \overset{(a)}{\le} \sqrt{2}\exp(-R\sqrt{1-\beta}) \le \frac{1}{4n},$$
where (a) holds because of the inequality $\ln(1-x) \le -x$ for any $x \in (0,1)$. The above inequality implies
$$\max\{|\lambda_2(M)|, |\lambda_n(M)|\} = \rho(M - \frac{1}{n}\mathbb{1}\mathbb{1}^T) \le \frac{1}{4n}.$$
Since $\bar{M} \triangleq (1-\tau)M + \tau I$ and $\tau = \frac{1}{2n}$, we have $\lambda_k(\bar{M}) = (1-\frac{1}{2n})\lambda_k(M) + \frac{1}{2n}$, the spectrum estimates of $\bar{M}$ is given by (38).

F.3 Proof of Theorem 25

The gradient accumulation and the fast gossip averaging do not affect the convergence analysis of D^2/Exact-Diffusion. By following the analysis of Theorem 11, if the learning rate is set as

$$\gamma = \min\left\{\frac{1}{4L}, \frac{(1-\tilde{\beta})\tilde{\lambda}_n^{\frac{1}{2}}}{40L\tilde{\lambda}_2}, \Big(\frac{\tilde{r}_0}{\tilde{r}_1(K+1)}\Big)^{\frac{1}{2}}, \Big(\frac{\tilde{r}_0}{\tilde{r}_2(K+1)}\Big)^{\frac{1}{3}}, \left(\frac{\tilde{r}_0}{\tilde{r}_3}\right)^{\frac{1}{3}}\right\}, \tag{114}$$

it holds from (89) that

$$\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\Big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\Big)\\
\le& O\left(\frac{\tilde{\sigma}}{\sqrt{nK}} + \frac{\tilde{\sigma}^{\frac{2}{3}}\tilde{\lambda}_2^{\frac{2}{3}}}{(1-\tilde{\beta}_1)K^{\frac{2}{3}}\tilde{\lambda}_n^{\frac{1}{3}}} + \frac{\tilde{\lambda}_2}{(1-\tilde{\beta}_1)K\tilde{\lambda}_n^{\frac{1}{2}}} + \frac{\tilde{\lambda}_2^{\frac{2}{3}}}{(1-\tilde{\beta}_1)^{\frac{1}{3}}(1-\tilde{\beta})^{\frac{1}{3}}K\tilde{\lambda}_n^{\frac{1}{3}}} + \frac{1}{K}\right)
\end{aligned} \tag{115}$$

where K is the number of outer loops, and by the definition of M, we have

$$\begin{aligned}
&\tilde{\sigma}^2 = \sigma^2/R, \quad \tilde{\lambda}_2 = \lambda_2(\bar{M}) \in [\frac{1}{4n}, \frac{3}{4n}], \quad \tilde{\lambda}_n = \lambda_n(\bar{M}) \in [\frac{1}{4n}, \frac{3}{4n}],\\
&\tilde{\beta}_1 = \sqrt{\lambda_2(\bar{M})} \in [\frac{1}{2n^{\frac{1}{2}}}, \frac{\sqrt{3}}{2n^{\frac{1}{2}}}], \quad \tilde{\beta} = \max\{|\lambda_2(M)|, |\lambda_n(M)|\} \le \frac{3}{4n}
\end{aligned} \tag{116}$$

In addition, constants $\tilde{r}_0$, $\tilde{r}_1$, $\tilde{r}_2$, and $\tilde{r}_3$ in (114) are defined as follows

$$\tilde{r}_0 = 2\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2, \ \tilde{r}_1 = \frac{2\tilde{\sigma}^2}{n}, \ \tilde{r}_2 = \frac{24L\bar{\lambda}_2^2\tilde{\sigma}^2}{(1-\tilde{\beta}_1)\tilde{\lambda}_n}, \ \tilde{r}_3 = \frac{9L\tilde{\lambda}_2^2}{(1-\tilde{\beta}_1)(1-\tilde{\beta})\tilde{\lambda}_n}. \tag{117}$$

We let T be the total number of sampled data or gossip communications, it holds that $K = T/R$. Substituting $K = T/R$ and the facts in (116) into (115), and ignoring all constants, we achieve

$$\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\Big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\Big) =& O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{R^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{n^{\frac{1}{3}}T^{\frac{2}{3}}} + \frac{R}{n^{\frac{1}{2}}T} + \frac{R}{n^{\frac{1}{3}}T} + \frac{R}{T}\Big)\\
=& O\Big(\frac{\sigma}{\sqrt{nT}} + \frac{\ln(n)^{\frac{1}{3}}\sigma^{\frac{2}{3}}}{n^{\frac{1}{3}}T^{\frac{2}{3}}(1-\beta)^{\frac{1}{6}}} + \frac{\ln(n)}{T(1-\beta)^{\frac{1}{2}}}\Big)
\end{aligned}$$

where the last inequality holds by substituting $R = \lceil\frac{\ln(n)+4}{\sqrt{1-\beta}}\rceil = O\left(\frac{\ln(n)}{(1-\beta)^{\frac{1}{2}}}\right)$. Note that the third term is less than or equal to the last term, we achieve the result in (41).

F.4 Proof of Theorem 27

By following the analysis of Theorem 18, if the learning rate is set as ($T = KR$)

$$\gamma = \min\left\{\frac{1}{4L}, \frac{1-\tilde{\beta}_1}{26L}\Big(\frac{\tilde{\lambda}_n^{1/2}}{\tilde{\lambda}_2}\Big), \frac{2\ln(2n\mu\mathbb{E}\|\bar{\boldsymbol{z}}^{(0)}\|^2K^2/[\tilde{\sigma}^2(1-\tilde{\beta})])}{\mu K}\right\}$$

it holds from (99) that

$$\begin{aligned}&\frac{1}{H_K}\sum_{k=0}^{K} h_k\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big)\\ =&\tilde{O}\left(\frac{\tilde{\sigma}^2}{nK} + \frac{\tilde{\sigma}^2}{(1-\tilde{\beta})K^2}\big(\frac{\tilde{\lambda}_2^2}{\tilde{\lambda}_n}\big) + \frac{\tilde{\lambda}_2}{(1-\tilde{\beta})\tilde{\lambda}_n^{\frac{1}{2}}}\exp\{-(1-\tilde{\beta})\big(\frac{\tilde{\lambda}_n^{\frac{1}{2}}}{\tilde{\lambda}_2}\big)K\} + \exp(-K)\right). \quad (118)\end{aligned}$$

where K is the number of outer loop, and h_k and H_K are defined in Lemma 17. Notation $\tilde{O}(\cdot)$ hides all logarithm factors. Substituting $K = T/R$, and the facts in (116) into (118), we have

$$\begin{aligned}\frac{1}{H_K}\sum_{k=0}^{K} h_k\big(\mathbb{E}f(\bar{x}^{(k)}) - f(x^\star)\big) =&\tilde{O}\left(\frac{\sigma^2}{nT} + \frac{R\sigma^2}{nT^2} + n^{-\frac{1}{2}}\exp\{-n^{\frac{1}{2}}\frac{T}{R}\} + \exp(-\frac{T}{R})\right)\\ =&\tilde{O}\left(\frac{\sigma^2}{nT} + \frac{\sigma^2}{n(1-\beta)^{\frac{1}{2}}T^2} + \exp\{-(1-\beta)^{\frac{1}{2}}T\}\right).\end{aligned}$$