OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Decentralized Stochastic Optimization with Client Sampling

Ziwei Liu ZIWEI.LIU@EPFL.CH
Anastasia Koloskova ANASTASIA.KOLOSKOVA@EPFL.CH
Martin Jaggi MARTIN.JAGGI@EPFL.CH
Tao Lin LINTAO@WESTLAKE.EDU.CN
EPFL, Switzerland
Westlake University, China

Abstract

Decentralized optimization is a key setting toward enabling data privacy and on-device learning over networks. Existing research primarily focuses on distributing the objective function across n nodes/clients, lagging behind the real-world challenges such as i) node availability—not all n nodes are always available during the optimization—and ii) slow information propagation (caused by a large number of nodes n).

In this work, we study Decentralized Stochastic Gradient Descent (D-SGD) with node subsampling, i.e. when only s ($s \leq n$) nodes are randomly sampled out of n nodes per iteration. We provide the theoretical convergence rates in smooth (convex and non-convex) problems with heterogeneous (non-identically distributed data) functions. Our theoretical results capture the effect of node subsampling and choice of the topology on the sampled nodes, through a metric termed *the expected consensus rate*. On a number of common topologies, including ring and torus, we theoretically and empirically demonstrate the effectiveness of such a metric.

1. Introduction

Decentralized stochastic optimization methods have advantages over their centralized counterpart like cheap per iteration communication cost, data locality, and communication efficiency, and thus have attracted a lot of attention. [5] introduced a convergence analysis to unify a large variety of decentralized Stochastic Gradient Descent (SGD) methods, covering the scenarios that all n nodes/clients actively compute gradient updates in each iteration and exchange updates via arbitrary communication topologies. [8] studied the convergence rate for a completely different setting, where only one worker will be selected in each iteration for gradient update and only communicate with another node.

However, these decentralized learning settings may not meet the requirements from the real-world constraints: not all workers are always available in every iteration, while the case in [8] is too extreme. As the first attempt to model the decentralized learning challenges with realistic constraints, we consider a relaxed setting, namely *decentralized learning with node sampling*, in which in each iteration, only $s \leq n$ nodes are randomly sampled out of n nodes for gradient update and update synchronization.

Our main contributions are

- We propose a unified algorithmic framework to capture the realistic decentralized learning scenarios, where only s out of n workers are sampled per iteration for gradient computa-

tion and decentralized communication. Such framework covers synchronous and pairwise gossip updates on time-varying network topology, as well as a large variety of decentralized SGD methods on various communication topologies that are developed separately in various communities.

- We provide convergence rates for the proposed framework, for smooth (convex and non-convex) problems on heterogeneous (non-identically distributed data) and homogeneous (i.i.d.) data settings.
- We empirically verify the tightness of our theoretical results on strongly convex functions and explain the impact of noise and data diversity on the convergence. When the noise level is small and node sampling is applied, random rings and random torus provide much better convergence rates compared to fixed rings and fixed torus.
- We further identify the influence of communication topology on the convergence rate, which can be represented by the expected consensus rate. We additionally provide a theorem with weak assumptions (i.e. symmetry and doubly stochasticity of the mixing matrix) to measure the expected consensus rate of various topologies, including the time-varying topologies. The tightness of these theoretical results is empirically justified by our experiments.

2. Related Work

[5] provides a unified convergence analysis that covers a large variety of decentralized SGD methods, yet does not cover the realistic scenario of node sampling. Moreover, some works provide convergence analysis for individual decentralized SGD methods of their choices. [8] gives the convergence rate of its proposed asynchronous decentralized parallel SGD algorithm (AD-PSGD). [15] provides a convergence rate for its SGD with a matching decomposition sampling algorithm. [3] provides a tight bound for the FedAvg algorithm [10], and also its proposed Stochastic Controlled Averaging (Scaffold) method. Also, [2, 9, 12] study the problem client sampling, which is orthogonal to the problem we study as the impact of communication topology does not take into account.

To the best of our knowledge, while convergence rates of different algorithms have been analyzed individually, there has not been a convergence analysis that covers a large variety of decentralized SGD methods which are developed separately in various communities and takes account of node sampling. Here we provide a framework for such convergence analysis, which covers convex and non-convex problems and also heterogeneous and iid-data settings.

3. A Unified Framework for Decentralized Learning with Client Sampling

We study the distributed stochastic optimization problem following the setting in [5]:

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^d} \left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right], \tag{1}$$

where the components $f_i \colon \mathbb{R}^d \to \mathbb{R}$ are distributed among n nodes and are given in the stochastic form: $f_i(\mathbf{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i} F_i(\mathbf{x},\xi_i)$, where $\mathcal{D}_i$ denotes the distribution of ξ_i over parameter space Ω_i on node i. We do not make any assumptions about the distributions $\mathcal{D}_i$.

Preliminaries and Assumptions We define $X^{(t)} := \left[\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right] \in \mathbb{R}^{d\times n}$, and $\bar{X}^{(t)} := \left[\bar{\mathbf{x}}^{(t)}, \ldots, \bar{\mathbf{x}}^{(t)}\right] \equiv X^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Our theoretical results rely on assumptions about the smoothness and convexity of f_i, F_i, and the data's noise and variance, which are included in the Appendix C.

The gossip averaging protocol can be compactly written in matrix notation, with $\mathcal{N}_i^{(t)} := \{j\colon w_{ij}^{(t)} > 0\}$ denoting the neighbors of node i at iteration t: $X^{(t+1)} = X^{(t)}W^{(t)} \Longleftrightarrow \mathbf{x}_i^{(t+1)} = \sum_{j\in\mathcal{N}_i^{(t)}} w_{ij}^{(t)}\mathbf{x}_j^{(t)}$, where the mixing matrix $W^{(t)} \in [0,1]^{n\times n}$ encodes the network structure at time t and the averaging weights (nodes i and j are connected if $w_{ij}^{(t)} > 0$). Our scheme shows great flexibility as the mixing matrices can change over iterations and can be selected from the (changing) distributions.

Definition 1 (Mixing matrix) *A symmetric ($W = W^\top$) doubly stochastic ($W\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top W = \mathbf{1}^\top$) matrix $W \in [0,1]^{n\times n}$.*

3.1. Decentralized (Gossip) SGD

This section introduces the generalized decentralized SGD framework that accommodates the case that s ($s \leq n$) workers are sampled out of n workers to update gradients and communicate in each iteration. This is adapted from the algorithm in [5].

Similar to existing works [6, 7, 14] our proposed method allows only decentralized communications. That is, the exchange of information (through *gossip* averaging) can only occur between connected nodes (neighbors). The algorithm is outlined in Algorithm 1.

Algorithm 1: Decentralized SGD with Node Sampling

Input: $X^{(0)}$, number of sampled nodes per iteration s, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T, mixing matrix distributions $\mathcal{W}^{(t)}$ for $t \in [0, T]$

for t ***in*** $0 \ldots T$ **do**
- Sample s workers $\mathcal{S}^{(t)}$ out of n workers
- Sample $W^{(t)} \sim \mathcal{W}^{(t)}$
- **for** i *in* $\mathcal{S}^{(t)}$ **do**
  - Sample $\xi_i^{(t)}$, compute $\mathbf{g}_i^{(t)} := \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$
  - $\mathbf{x}_i^{(t+\frac{1}{2})} = \mathbf{x}_i^{(t)} - \eta_t \mathbf{g}_i^{(t)}$ ▷ stochastic gradient updates
  - $\mathbf{x}_i^{(t+1)} := \sum_{j\in\mathcal{N}_i^t} w_{ij}^{(t)}\mathbf{x}_j^{(t+\frac{1}{2})}$ ▷ gossip averaging
- **end**

end

In each iteration in Algorithm 1, a new mixing matrix $W^{(t)}$ is sampled from a possibly time-varying distribution $\mathcal{W}^{(t)}$, $t \in \{0, \ldots, T\}$. For randomized gossip averaging with a randomly sampled mixing matrix $W \sim \mathcal{W}$ it holds

$$\mathbb{E}_W \left\|XW - \bar{X}\right\|_F^2 \leq (1-p)\left\|X - \bar{X}\right\|_F^2, \tag{2}$$

for a value $p \geq 0$ (related to the spectrum of $\mathbb{E}\, W^\top W$), that is, the averaging step brings the values in the columns of $X \in \mathbb{R}^{d\times n}$ closer to their row-wise average $\bar{X} := X \cdot \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ in expectation [1].

Random ring and torus After a set of s nodes $\mathcal{S}$ is selected, with fixed ring/torus topology, the ring/torus remains the same at different iterations. However, with random ring/torus topology, each possible connection has equal probability. For example, three different rings with the same $\mathcal{S}$ are shown in Figure 1.

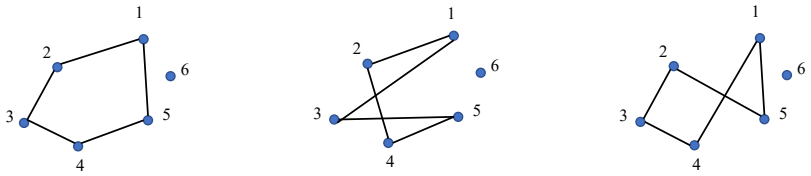


Figure 1: 3 possible random rings when $\mathcal{S} = 1, 2, 3, 4, 5$ and $n = 6$.

4. Theoretical Results and Empirical Justification

4.1. Theoretical Results

In this section, we introduce our convergence results and justify that it can recover the baseline [5] for the case of $s = n$ (i.e. w/o client sampling).

Theorem 2 *For any target accuracy $\epsilon > 0$ and Algorithm 1 with mixing matrices e.g. in (2), there exists a (constant) stepsize (potentially depending on ϵ) such that the accuracy can be reached after at most the following number of iterations T.*
Non-Convex: *Under Assumption 2 and Assumption 5, it holds $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \leq \epsilon$ after*

$$T := \mathcal{O}\left(\frac{\hat{\sigma}^2 + (1-\frac{s}{n})\hat{\zeta}^2}{s\epsilon^2} + \frac{n}{s}\frac{\hat{\zeta}^2\sqrt{(M+1)} + \hat{\sigma}\sqrt{p}}{p\epsilon^{3/2}} + \frac{\sqrt{(P+1)(M+1)}}{p\epsilon}\right) Lr_0$$

iterations. If we in addition assume convexity and $\mu > 0$,
Strongly-Convex: *Similarly, it holds $\sum_{t=0}^{T}\frac{w_t}{W_T}(\mathbb{E}\, f(\bar{\mathbf{x}}^{(t)}) - f^\star) + \mu\,\mathbb{E}\,\|\bar{\mathbf{x}}^{(T+1)} - \mathbf{x}^\star\|^2 \leq \epsilon$ after*[1]

$$T := \tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu s\epsilon} + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{\mu s\epsilon} + \frac{\sqrt{L}(\bar{\zeta} + \sqrt{p}\bar{\sigma})}{\mu p\sqrt{\epsilon}}\sqrt{\frac{n}{s}} + \frac{n}{s}\frac{L}{\mu p}\log\left(\frac{1}{\epsilon}\right)\right)$$

iterations, for positive weights w_t and $r_0 := f(\mathbf{x}_0) - f^\star$ denote the initial errors.

We also show that our result recovers [5] when $s = n$ in Appendix F.

4.2. Empirical Justification of Theoretical Results

In this section, we verify that the numerical performance of decentralized stochastic optimization algorithms coincides with the rates predicted by theory, focusing on the strongly convex case for now.

1. $\tilde{\mathcal{O}}/\tilde{\Omega}$-notation hides constants and polylogarithmic factors.

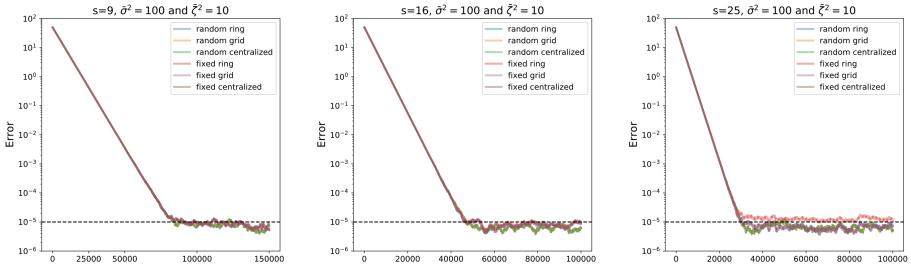


Figure 2: Convergence of $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\right\|_2^2$ to target accuracy $\epsilon = 10^{-5}$ for the same problem difficulty ($\bar{\sigma}^2 = 100, \bar{\zeta}^2 = 10$), and different topologies on $n = 25$, $s = 9, 16, 25$ nodes, $d = 50$. Stepsizes were tuned to be the same for all experiments.

We consider a distributed least squares objective with $f_i(\mathbf{x}) := \frac{1}{2}\left\|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\right\|_2^2$, for fixed Hessian $\mathbf{A}_i^2 = \frac{i^2}{n}\cdot\mathbf{I}_d$ and sample each $\mathbf{b}_i \sim \mathcal{N}(0, \bar{\zeta}^2/i^2\mathbf{I}_d)$ for a parameter $\bar{\zeta}^2$, which controls the similarity of the functions and coincides with the parameter in Assumption 4. We control the stochastic noise $\bar{\sigma}^2$ by adding Gaussian noise to every stochastic gradient.

Discussion Our convergence rate with convex $f_i(\mathbf{x})$ is:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu s\epsilon} + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{\mu s\epsilon} + \frac{\sqrt{L}(\bar{\zeta}+\sqrt{p}\bar{\sigma})}{\mu p\sqrt{\epsilon}}\sqrt{\frac{n}{s}} + \frac{n}{s}\frac{L}{\mu p}\log\left(\frac{1}{\epsilon}\right)\right)$$

When the noise level is high. We can see from Figure 2 that the difference in topology does not have a significant influence on the convergence rate and that increasing the number of sampled nodes reduces the number of iterations until convergence, as expected from the theoretical analysis.

Furthermore, when $\bar{\sigma}^2 = 100, \bar{\zeta}^2 = 10$, the convergence rate is dominated by $\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu s\epsilon} + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{\mu s\epsilon}\right)$, and therefore is expected to be proportional to $t(s) = \left(\frac{\bar{\sigma}^2}{s} + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{s}\right)$ when other parameters remain the same. This is validated by experimental results in Figure 6 and Figure 7 in Appendix G.

When the noise level is low. Here we consider the extreme case of $\bar{\sigma}^2 = \bar{\zeta}^2 = 0$, where topologies should have a more significant influence on the convergence rate. This is illustrated by experimental results in Figure 8 in Appendix G, where we can see that fixed ring and torus need more iterations to reach the error $\epsilon = 10^{-5}$ compared to random ring and torus. However, random rings and torus have similar numerical performances to the centralized topology. Therefore, we guess that fixed ring and torus have significantly larger, i.e. worse expected consensus rate compared to random ring and grid. We verify this hypothesis experimentally in section 5.

5. Expected Consensus Rate of Sampled Time-varying Topology

In our experiments in Figure 8 in subsection 4.2, we have seen that random ring and torus topologies achieve similar convergence rates as centralized topology but have a much smaller communication cost for both sub-sampled cases and all n nodes/clients update cases. While it's always been proven that in all n nodes/clients update cases random rings and toruses achieve a better convergence rate

and expected consensus rate than fixed rings and toruses, in this section, we theoretically and empirically prove that this also stands true in sub-sampled cases.

Recall the definition of expected consensus rate $(1-p)$ in (2). Here we provide a formula to calculate $(1-p)$ for cases with or without node samplings.

Lemma 3 (Expected Consensus Rate)
The consensus rate decrease over T steps can be bounded by:

$$\mathbb{E}_W \left\| X \prod_{t=1}^{T} W_t - \bar{X} \right\|_F^2 \leq |\lambda_2(\mathbb{E}\left[W^T W\right])|^T \left\|(X-\bar{X})\right\|_F^2 \tag{3}$$

where W_t is the mixing matrix W at step t. When $T=1$,

$$\mathbb{E}_W \left\|XW-\bar{X}\right\|_F^2 \leq |\lambda_2(\mathbb{E}\left[W^T W\right])| \left\|(X-\bar{X})\right\|_F^2 \tag{4}$$

$\mathbb{E}$ is taken over the distributions $W \sim \mathcal{W}$. We set the expected consensus rate $(1-p) = |\lambda_2(\mathbb{E}\left[W^T W\right])|$.

Using Lemma 3, we further derive lemmas of consensus rate for sub-sampled centralized topology, sub-sampled random ring topology, and sub-sampled centralized topology. We present the proof for random ring topology in Appendix H as an example.

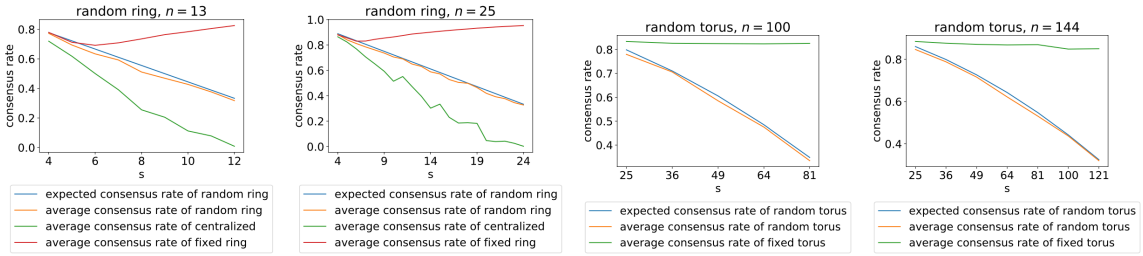
Theorem 4 (Expected Consensus Rate of Sub-sampled Centralized Topology) *When s nodes are randomly sampled out of n nodes each time and communicate with each other using centralized/random ring/random torus topology, the expected consensus rate is*

$$1-p = \begin{cases} \frac{n-s}{n-1}, & \textit{with centralized topology}; \\ 1-\frac{2s}{3(n-1)}, & \textit{with random ring topology}; \\ 1-\frac{4s}{5n}-\frac{12s}{25n(n-1)}, & \textit{with random torus topology}. \end{cases}$$

We then experimentally validate Theorem 4. We consider the case where in each step, a set $\mathcal{S}$ of s $(i.e.\ 2 \leq s \leq (n-1))$ workers are sampled out of n nodes and communicate with each other with centralized, ring, or torus topology. For or random ring topology, the case where $n=25, d=50$ and the case where $n=13, d=20$ is shown in Figure 3. For 2D random torus topology, the case where $n=100, d=100$ and the case where $n=144, d=200$ are both shown in Figure 3. For centralized topology, the case where $n=25, d=50$ and the case where $n=13, d=20$ is shown in Figure 5 in Appendix A.

6. Conclusion

We present a framework for the analysis of decentralized SGD methods with node sampling and provide convergence guarantees. Our results show that with the presence of node sampling, decentralized SGD methods can achieve linear speedup in the number of workers n and the convergence rate does only weakly depend on the graph topology, the number of local steps, or the data heterogeneity when the noise level is high. With node sampling, the effect of those parameters becomes more pronounced when the noise is small, and especially function diversity can hamper the convergence of decentralized SGD methods.



(a) $n = 13$ (b) $n = 25$ (c) $n = 100$ (d) $n = 144$

Figure 3: Expected consensus rate and average consensus rate of the random ring, fixed ring, random torus, and fixed torus when sampling $s(2 \leq s < n)$ nodes out of n nodes during each update. It's easy to see that the expected consensus rate serves as a tight upper bound for the average consensus rate, which validates the correctness of our theory.

To quantify the influence of communication topology on the convergence rate, we additionally provide a theorem with weak assumptions (i.e. symmetry and doubly stochasticity of the mixing matrix) to measure the expected consensus rate of various topologies, including the time-varying topologies. The tightness of these theoretical results on the centralized, random ring, and random torus topologies are empirically justified by our experiments. We also show that the random ring and torus topology achieve much better consensus rates compared to their fixed counterparts.

References

[1] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

[2] Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning, 2021. URL https://openreview.net/forum?id=l0BP1lHpPW.

[3] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[4] Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019.

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

[6] 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, 2019.

[7] 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. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[9] Bing Luo, Wenli Xiao, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1739–1748. IEEE, 2022.

[10] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[11] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[12] Monica Ribero and Haris Vikalo. Communication-efficient federated learning via optimal client sampling. *arXiv preprint arXiv:2007.15197*, 2020.

[13] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21:1–36, 2020.

[14] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

[15] Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. Matcha: Speeding up decentralized sgd via matching decomposition sampling. In *2019 Sixth Indian Control Conference (ICC)*, pages 299–300. IEEE, 2019.

Appendix: Decentralized Stochastic Optimization with Client Sampling

The appendix is organized as follows: In Section A In Section B, Algorithm 1 is rewritten in matrix notation as Algorithm 2 and gives a sketch of the proof using this new notation. In Section C we state a few auxiliary technical lemmas, before giving all details for the proof of the theorem in Sections D and E. We also provide a discussion to show that our results recover the results of [5] in F. We conclude the appendix in Section G by presenting additional numerical results that confirm the tightness of our theoretical analysis in the strongly convex case.

We also provide proof of the expected consensus rate formula, i.e. Theorem 4 for random ring topology. The expected consensus rate formula for centralized and random torus topologies follows similarly.

Appendix A. Additional experiments for consensus rate

we validate our theoretical expected consensus rate derived in section 5 on a few examples, including fixed full/broken ring, alternating full/broken ring, and centralized.

Here we introduce our experimental settings:

- **Initialization**. X is initialized using normal initialization.
- **Notations**: We denote the initial X as X_0, and $X_0 \prod_{t=1}^{T}(W_t)$ as X_n. In other words, X_n denotes the X_0 after T steps of consensus decrease.
- **One step consensus rate**. In experiments, we calculate the *one-step consensus* rates as $\frac{\|XW_t-\bar{X}\|_F^2}{\|(X-\bar{X})\|_F^2}$. Lemma 3, Equation 4 can also be written as $\frac{\mathbb{E}\|X_tW_t-\bar{X}\|_F^2}{\mathbb{E}\|(X-\bar{X})\|_F^2} \leq (1-p) = |\lambda_2(\mathbb{E}\,[W_t^TW_t])|$. In the following section, we compare the experimental one-step consensus rate with the expected consensus rate $(1-p)$.
- **Average consensus rate**. Equation 3 in Lemma 3 can also be written as

$$\sqrt[L]{\frac{\mathbb{E}\left\|X\prod_{l=1}^{L}(W_t)-\bar{X}\right\|_F^2}{\mathbb{E}\left\|(X-\bar{X})\right\|_F^2}} \leq (1-p) = |\lambda_2(\mathbb{E}\,[W_t^TW_t])|.$$

  Therefore, we calculate the average consensus rate as $\sqrt[T]{\frac{\mathbb{E}\|X\prod_{t=1}^{T}(W_t)-\bar{X}\|_F^2}{\mathbb{E}\|(X-\bar{X})\|_F^2}}$, and compare it with the expected consensus rate $(1-p)$.
- **Selection of** T. In each of our example, after a number of steps U, $X_t = \bar{X}$, the consensus distance $\left\|X_t-\bar{X}\right\| = \left\|X_tW_t-\bar{X}\right\| = 0, (t \geq U)$, and $\frac{\left\|X_t-\bar{X}\right\|}{\left\|X_tW_t-\bar{X}\right\|} = 1$. Therefore, when calculating the *average consensus rate*, we set $\frac{3}{4}U \leq T \leq U$ for the accuracy.

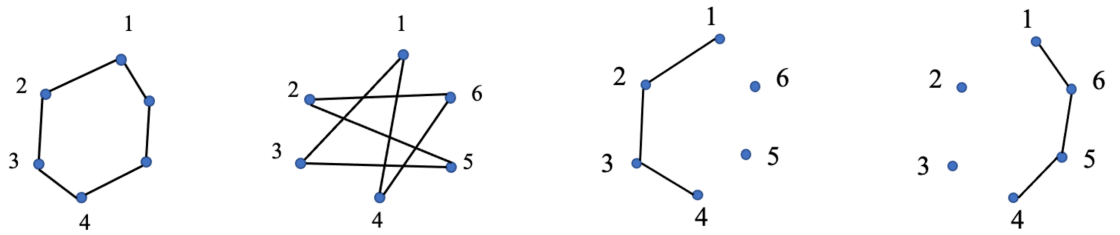
For all examples, we calculate the theoretical expected consensus rate $(1-p)$ using Lemma Theorem 3, Equation 4, and the experimental average consensus rate. We further compare the average consensus rate and proved that it is close to and smaller than the corresponding expected consensus rate, as shown in Table 1.

topology	expected consensus rate	average consensus rate
alternating full rings	0.33	0.22
fixed full ring	0.44	0.43
alternating broken rings	0.78	0.68
fixed broken ring	1	0.94

Table 1: Comparison between expected consensus rate and average consensus rate for examples.

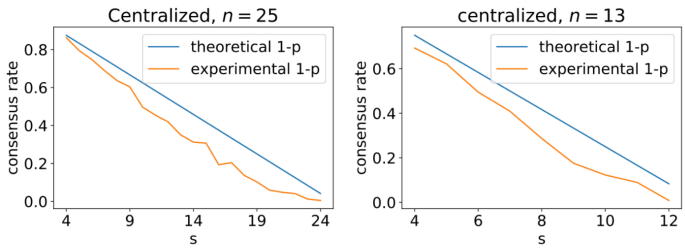
Two Alternating Full Rings and Broken Rings. Here we show that Lemma Theorem 3 captures the improvement brought by alternating between two different full rings or alternating broken rings over their fixed counterparts in Table Table 1. Alternating full rings indicates alternating between topology 4(a) and 4(a) with equal probability, while fixed full ring indicates sticking to 4(a). Similarly, alternating broken rings indicates alternating between topology 4(c) and 4(c) with equal probability, while fixed broken ring indicates sticking to topology 4(c).

Furthermore, with the random ring, a bigger s leads to a better consensus rate, yet with fixed rings, a bigger s doesn't necessarily lead to a better consensus rate. We explain that for fixed rings when s is small, there are more possible $\mathcal{S}$ thus more chances for different nodes to have direct communication.



(a) Full ring 1 (b) Full ring 2 (c) Broken ring 1 (d) Broken ring 2

Figure 4: Topologies of full and broken rings



(a) $n = 25$ (b) $n = 13$

Figure 5: Expected consensus rate and average consensus rate of centralized topology when sampling $s(2 \leq s < n)$ nodes out of n nodes each round.

Appendix B. Proof of Theorem 2

B.1. Matrix Notation for Decentralized SGD

We rewrite the Algorithm with node sampling using the following matrix notation, extending the definition used in the main text:

$$\begin{aligned} X^{(t)} &:= \left[\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right] \in \mathbb{R}^{d\times n}, \\ \bar{X}^{(t)} &:= \left[\bar{\mathbf{x}}^{(t)}, \ldots, \bar{\mathbf{x}}^{(t)}\right] \in \mathbb{R}^{d\times n}, \\ \partial F(X^{(t)}, \xi^{(t)}, \mathcal{S}^{(t)}) &:= \left[\ldots, 0, \nabla F_{i_1}(\mathbf{x}_{i_1}^{(t)}, \xi_n^{(t)}), 0, \ldots, 0, \nabla F_{i_2}(\mathbf{x}_{i_2}^{(t)}, \xi_n^{(t)}), 0, \ldots\right] \in \mathbb{R}^{d\times n}, \\ \partial f(X^{(t)}, \mathcal{S}^{(t)}) &:= \left[\ldots, 0, \nabla f_{i_1}(\mathbf{x}_{i_1}^{(t)}), 0, \ldots, 0, \nabla f_{i_2}(\mathbf{x}_{i_2}^{(t)}), 0, \ldots, 0, \nabla f_{i_3}(\mathbf{x}_{i_3}^{(t)}), 0, \ldots\right] \in \mathbb{R}^{d\times n}. \end{aligned} \tag{5}$$

where nonzero entries exist only for $i \in \mathcal{S}$.

Algorithm 2: Decentralized SGD with Node Sampling

Input: $X^{(0)}$, number of sampled nodes per iteration s, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T, mixing matrix distributions $\mathcal{W}^{(t)}$ for $t \in [0, T]$

for t ***in*** $0 \ldots T$ **do**

- Sample s workers $\mathcal{S}^{(t)}$ out of n workers
- Sample $W^{(t)} \sim \mathcal{W}^{(t)}$
- $X^{(t+\frac{1}{2})} = X^{(t)} - \eta_t \partial F(X^{(t)}, \xi^{(t)}, \mathcal{S}^{(t)})$ ▷ stochastic gradient updates
- $X^{(t+1)} = X^{(t+\frac{1}{2})} W^{(t)}$ ▷ gossip averaging

end

B.2. Proof Sketch—Combining Consensus Progress (Gossip) and Optimization Progress (SGD)

In this section, we sketch the proof for Theorem 2. As a first step in the proof, we will derive an upper bound on the expected progress, measured as distance to the optimum, $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$ for the convex cases, and function suboptimality $r_t = \mathbb{E}\, f(\bar{\mathbf{x}}^{(t)}) - f^\star$ in the non-convex case. These bounds will have the following form:

$$r_{t+1} \leq (1 - a\eta_t) r_t - b\eta_t e_t + c\eta_t^2 + \eta_t B \Xi_t\,, \tag{6}$$

with $\Xi_t = \frac{1}{n}\mathbb{E}_t \sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ and

- for strongly convex case $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$, $e_t = f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)$, $a = \frac{\mu}{2}$, $b = 1$, $c = \frac{\bar{\sigma}^2}{n}$, $B = 3L$ (Lemma 16);
- for the non-convex case $r_t = \mathbb{E}\, f(\bar{\mathbf{x}}^{(t)}) - f^\star$, $e_t = \left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2$, $a = 0$, $b = \frac{1}{4}$, $c = \frac{L\hat{\sigma}^2}{n}$, $B = L^2$ (Lemma 19).

We will then bound the consensus distance Ξ_t as detailed in Section D; Lemmas 17 and 20 by a recursion of the form

$$\Xi_t \leq \left(1-\frac{p}{2}\right)\Xi_{t-1} + \frac{p}{64}\Xi_{t-1} + D\eta_{t-1}^2 e_{t-1} + A\eta_{t-1}^2, \tag{7}$$

for convex cases $A = \frac{s}{n}(4\bar{\sigma}^2 + \frac{9}{p}\bar{\zeta}^2)$, $D = \frac{s}{n}\frac{34}{p}L$ (Lemma 17) and for non-convex case $A = \left(\hat{\sigma}^2 + 2(\frac{3}{p}+M)\hat{\zeta}^2\right)$, $D = 2P\frac{n}{s}(\frac{3}{p}+M)$ (Lemma 20).

Next, we simplify this recursive equation (7) using Lemma 21 and some positive weights $\{w_t\}_{t\geq 0}$ (see Lemma 21 for the definition of the weights w_t) to

$$B\cdot\sum_{t=0}^{T} w_t\Xi_t \;\leq\; \frac{b}{2}\cdot\sum_{t=0}^{T} w_t e_t + 64AB\frac{1}{p}\cdot\sum_{t=0}^{T} w_t\eta_t^2\,, \tag{8}$$

where again $\Xi_t = \frac{1}{n}\,\mathbb{E}_t\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$.

Then we combine (6) and (8). Firstly rearranging (6), multiplying by w_t and dividing by η_t, we get

$$bw_t e_t \leq \frac{(1-a\eta_t)}{\eta_t}w_t r_t - \frac{w_t}{\eta_t}r_{t+1} + cw_t\eta_t + Bw_t\Xi_t\,,$$

Now summing up and dividing by $W_T = \sum_{t=0}^T w_t$,

$$\begin{aligned}\frac{1}{W_T}\sum_{t=0}^{T} bw_t e_t &\leq \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{(1-a\eta_t)w_t}{\eta_t}r_t - \frac{w_t}{\eta_t}r_{t+1}\right) + \frac{c}{W_T}\sum_{t=0}^{T} w_t\eta_t + \frac{1}{W_T}B\sum_{t=0}^{T} w_t\Xi_t \\ &\overset{(8)}{\leq} \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{(1-a\eta_t)w_t}{\eta_t}r_t - \frac{w_t}{\eta_t}r_{t+1}\right) + \frac{c}{W_T}\sum_{t=0}^{T} w_t\eta_t + \frac{1}{2W_T}\sum_{t=0}^{T} w_t e_t + \frac{64BA}{W_T}\frac{\tau}{p}\sum_{t=0}^{T} w_t\eta_t^2\,,\end{aligned}$$

Therefore,

$$\frac{1}{2W_T}\sum_{t=0}^{T} bw_t e_t \leq \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{(1-a\eta_t)w_t}{\eta_t}r_t - \frac{w_t}{\eta_t}r_{t+1}\right) + \frac{c}{W_T}\sum_{t=0}^{T} w_t\eta_t + \frac{64BA}{W_T}\frac{\tau}{p}\sum_{t=0}^{T} w_t\eta_t^2 \tag{9}$$

Finally, to solve this main recursion (9) and obtain the final convergence rates of Theorem 2, we will use the following Lemmas, which will be presented in Section E:

- Lemma 22 for strongly convex case when $a > 0$.
- Lemma 23 for non-convex cases as $a = 0$.

B.3. How the Proof of Theorem 2 Follows

In this section we summarize how the proof of Theorem 2 follows from the results that we establish in Sections D and E below. Note that for convex cases we require both f_i and F_i to be convex as in Lemma 17.

Proof [Proof of Theorem 2, strongly convex case] The proof follows by applying the result of Lemma 22 to the equation (9) (obtained with Lemmas 16, 17, 21) with $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$, $e_t = f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)$, $a = \frac{\mu s}{2n}$, $b = \left(\frac{3}{2}\frac{s}{n} - \frac{2s^2}{3n^2}\right)$, $c = \frac{s}{n^2}(3\bar{\sigma}^2 + 8(1-\frac{s}{n})\bar{\zeta}^2)$, $d = \frac{10L}{p}$, $A = \frac{s}{n}(4\bar{\sigma}^2 + \frac{9}{p}\bar{\zeta}^2)$, $B = \left(\frac{9s}{4n} + \frac{s^2}{6n^2}\right)L$, $D = \frac{s}{n}\frac{34}{p}L$. It is only left to show that chosen weights w_t stepsizes η_t in Lemma 22 satisfy conditions of Lemmas 16, 17, 21. It is shown in Proposition 14 that $\{\eta_t\}$ is $\frac{4}{p}$-slow decreasing and $\{w_t\}$ is $\frac{8}{p}$-slow increasing (condition in Lemma 21). Moreover the stepsize $\eta_t := \eta < \frac{1}{d}$ is smaller than conditions on η_t in Lemmas 16, 17, 21. ■

Proof [Proof of Theorem 2, non-convex case] applying the result of Lemma 23 to the equation (9) (obtained with Lemmas 19, 20, 21) with $r_t = \mathbb{E}\, f(\bar{\mathbf{x}}^{(t)}) - f^\star$, $\frac{s}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2$, $a = 0$, $b = \frac{1}{4}$, $c = L\frac{s}{n}\frac{\hat{\sigma}^2}{n} + L\frac{s^2}{n^2}\frac{2}{s}(1-\frac{s}{n})\hat{\zeta}^2$, $d = 32L\sqrt{2\max\{P,1\}\left(\frac{6\tau}{p}+M\right)\frac{\tau}{p}}$, $A = \left(\hat{\sigma}^2 + 2(\frac{3}{p}+M)\hat{\zeta}^2\right)$, $B = \frac{s}{n}L^2$, $D = 2P\frac{n}{s}(\frac{3}{p}+M)$. Weights w_t stepsizes η_t chosen in Lemma 23 satisfy conditions of Lemmas 19, 20, 21, as shown in Proposition 14. ■

Appendix C. Technical Preliminaries

C.1. Assumptions

Assumptions on the objective function f For all our theoretical results we assume that f is smooth.

Assumption 1 (L-smoothness) *Each function $F_i(\mathbf{x},\xi)\colon \mathbb{R}^d \times \Omega_i \to \mathbb{R}$, $i \in [n]$ is differentiable for each $\xi \in \operatorname{supp}(\mathcal{D}_i)$ and there exists a constant $L \geq 0$ such that for each $\mathbf{x},\mathbf{y} \in \mathbb{R}^d, \xi \in \operatorname{supp}(\mathcal{D}_i)$:*

$$\|\nabla F_i(\mathbf{y},\xi) - \nabla F_i(\mathbf{x},\xi)\| \leq L\|\mathbf{x}-\mathbf{y}\| \,. \tag{10}$$

Sometimes it will be enough to just assume smoothness of f_i instead.

Assumption 2 (L-smoothness) *Each function $f_i(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$ is differentiable and there exists a constant $L \geq 0$ such that for each $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$:*

$$\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\| \leq L\|\mathbf{x}-\mathbf{y}\| \,. \tag{11}$$

Remark 5 *Clearly, Assumption 2 is more general than Assumption 1. Moreover, for convex $F(\mathbf{y},\xi)$ Assumption 1 implies Assumption 2 [11].*

For some of the derived results, we need in addition convexity. Specifically, μ-convexity for a parameter $\mu \geq 0$.

Assumption 3 (μ-convexity) *Each function $f_i\colon \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$ is μ-(strongly) convex for constant $\mu \geq 0$. That is, for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$:*

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \langle \nabla f_i(\mathbf{x}), \mathbf{x}-\mathbf{y}\rangle \,. \tag{12}$$

Assumptions on the noise We now formulate our assumptions on the noise.

Assumption 4 (Bounded noise at the optimum) *Let* $\mathbf{x}^\star = \arg\min f(\mathbf{x})$ *and define*

$$\zeta_i^2 := \|\nabla f_i(\mathbf{x}^\star)\|_2^2 \,, \qquad \bar{\zeta}^2 := \tfrac{1}{n}\textstyle\sum_{i=1}^n \zeta_i^2 \,. \tag{13}$$

Further, define

$$\sigma_i^2 := \mathbb{E}_{\xi_i} \left\|\nabla F_i(\mathbf{x}^\star, \xi_i) - \nabla f_i(\mathbf{x}^\star)\right\|_2^2 \,, \tag{14}$$

and similarly as above, $\bar{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. *We assume that* $\bar{\sigma}^2$ *and* $\bar{\zeta}^2$ *are bounded.*

Here, $\bar{\sigma}^2$ measures the noise level, and $\bar{\zeta}^2$ the diversity of the functions f_i. If all functions are identical, $f_i = f_j$, for all i, j, then $\bar{\zeta}^2 = 0$.

For the non-convex case—where a unique $\mathbf{x}^\star$ does not necessarily exist—we generalize Assumption 4 to:

Assumption 5 (Bounded noise) *We assume that there exists constants* P, $\hat{\zeta}$ *such that* $\forall \mathbf{x} \in \mathbb{R}^d$,

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\nabla f_i(\mathbf{x})\|_2^2 \leq \hat{\zeta}^2 + P\,\|\nabla f(\mathbf{x})\|_2^2 \,, \tag{15}$$

and constants M, $\hat{\sigma}$ *such that* $\forall \mathbf{x}_1, \ldots \mathbf{x}_n \in \mathbb{R}^d$

$$\Psi \leq \hat{\sigma}^2 + \tfrac{M}{n}\textstyle\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i)\|_2^2 \,, \tag{16}$$

where $\Psi := \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi_i} \left\|\nabla F_i(\mathbf{x}_i, \xi_i) - \nabla f_i(\mathbf{x}_i)\right\|_2^2$.

We see that Assumption 4 is weaker than Assumption 5 as it only needs ho hold for $\mathbf{x}_i = \mathbf{x}^\star$.

C.2. Implications of the assumptions

Proposition 6 *One step of gossip averaging with the mixing matrix* W *(def. 1) preserves the average of the iterates, i.e.*

$$XW\frac{\mathbf{1}\mathbf{1}^\top}{n} = X\frac{\mathbf{1}\mathbf{1}^\top}{n} \,.$$

Proposition 7 (Implications of the smoothness Assumption 1) *If for functions* $F_i(\mathbf{x}, \xi)$ *Assumption 1 holds, then it also holds that*

$$F_i(\mathbf{x}, \xi) \leq F_i(\mathbf{y}, \xi) + \langle \nabla F_i(\mathbf{y}, \xi), \mathbf{x} - \mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \,, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \Omega_i \tag{17}$$

If functions $f_i(\mathbf{x}) = \mathbb{E}_\xi F_i(\mathbf{x}, \xi)$, *then*

$$f_i(\mathbf{x}) \leq f_i(\mathbf{y}) + \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \,, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{18}$$

Moreover, if in addition, F_i *are convex functions, then*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq L\,\|\mathbf{x} - \mathbf{y}\|_2 \,, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{19}$$

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\|_2^2 \leq 2L\left(g(\mathbf{x}) - g(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla g(\mathbf{y})\rangle\right) \,, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{20}$$

where $g(\mathbf{x})$ *is either* F_i *or* f_i.

Proposition 8 (Implications of the smoothness Assumption 2) *From Assumption 2 it follows that*

$$f_i(\mathbf{x}) \leq f_i(\mathbf{y}) + \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \,, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \,. \tag{21}$$

C.3. Useful Inequalities

Lemma 9 *For arbitrary set of n vectors $\{\mathbf{a}_i\}_{i=1}^n$, $\mathbf{a}_i \in \mathbb{R}^d$*

$$\left\| \sum_{i=1}^n \mathbf{a}_i \right\|^2 \leq n \sum_{i=1}^n \|\mathbf{a}_i\|^2 \,. \tag{22}$$

Lemma 10 *For given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$*

$$2\langle \mathbf{a}, \mathbf{b} \rangle \leq \gamma \|\mathbf{a}\|^2 + \gamma^{-1} \|\mathbf{b}\|^2 \,, \qquad \forall \gamma > 0 \,. \tag{23}$$

Lemma 11 *For given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$*

$$\|\mathbf{a} + \mathbf{b}\|^2 \leq (1+\alpha) \|\mathbf{a}\|^2 + (1+\alpha^{-1}) \|\mathbf{b}\|^2 \,, \qquad \forall \alpha > 0 \,. \tag{24}$$

This inequality also holds for the sum of two matrices $A, B \in \mathbb{R}^{n \times d}$ in the Frobenius norm.

Remark 12 *For $A \in \mathbb{R}^{d \times n}$, $B \in \mathbb{R}^{n \times n}$*

$$\|AB\|_F \leq \|A\|_F \|B\|_2 \,. \tag{25}$$

C.4. τ-slow Sequences

Definition 13 (τ-slow sequences [13]) *The sequence $\{a_t\}_{t \geq 0}$ of positive values is* τ-slow decreasing *for parameter $\tau > 0$ if*

$$a_{t+1} \leq a_t, \quad \forall t \geq 0 \qquad \textit{and,} \qquad a_{t+1}\left(1 + \frac{1}{2\tau}\right) \geq a_t, \quad \forall t \geq 0 \,.$$

The sequence $\{a_t\}_{t \geq 0}$ is τ-slow increasing *if $\{a_t^{-1}\}_{t \geq 0}$ is τ-slow decreasing.*

Proposition 14 (Examples)

1. *The sequence $\{\eta_t^2\}_{t \geq 0}$ with $\eta_t = \frac{a}{b+t}$, $b \geq \frac{32}{p}$ is $\frac{4}{p}$-slow decreasing.*
2. *The sequence of constant stepsizes $\{\eta_t^2\}_{t \geq 0}$ with $\eta_t = \eta$ is τ-slow decreasing for any τ.*

Appendix D. Descent Lemmas and Consensus Recursions

In this section, according to our proof sketch we derive descent (6) and consensus recursions (8) for both convex and also non-convex cases.

D.1. Convex Cases

We require both f_i and F_i to be convex. We do not need Assumption 3 to hold for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and we could weaken it to hold only for $\mathbf{x} = \mathbf{x}^\star$ and for all $\mathbf{y} \in \mathbb{R}^d$.

Proposition 15 (Mini-batch variance with node sampling) *Let functions $F_i(\mathbf{x}, \xi)$, $i \in [n]$ be L-smooth (Assumption 1) with bounded noise at the optimum (Assumption 4). Then for any $\mathbf{x}_i \in \mathbb{R}^d, i \in [n]$ and $\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ it holds*

$$\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}} \left[\left\|\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x}_i^{(t)}) - \sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2\right]^2 \leq 3s\Big(L^2 \left\|\mathbf{x}_i - \bar{\mathbf{x}}\right\|^2 + 2L\left(f(\bar{\mathbf{x}}) - f(\mathbf{x}^\star)\right) + \sigma_i^2\Big).$$

Proof

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}} \left\|\sum_{i\in\mathcal{S}} (\nabla f_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i, \xi_i))\right\|^2 &= \mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}} \sum\nolimits_{i\in\mathcal{S}} \left\|\nabla f_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i, \xi_i)\right\|^2 \\
&= s\mathbb{E}_{i^{(t)},\boldsymbol{\xi}_i^{(t)}} \left\|\nabla f_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i, \xi_i)\right\|^2 \\
&\leq 3s\mathbb{E}_{i^{(t)},\boldsymbol{\xi}_i^{(t)}} \Big(\left\|\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla F_i(\bar{\mathbf{x}}, \xi_i)\right\|^2 \\
&\qquad + \left\|\nabla F_i(\bar{\mathbf{x}}, \xi_i) - \nabla F_i(\mathbf{x}^\star, \xi_i)\right\|^2 + \left\|\nabla F_i(\mathbf{x}^\star, \xi_i) - \nabla f_i(\mathbf{x}^\star)\right\|^2 \Big) \\
&\overset{(10),(20),(14)}{\leq} 3s\left(L^2 \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}\right\|^2 + 2L\left(f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)\right) + \bar{\sigma}^2\right)
\end{aligned}$$

where we used that $\mathbb{E}\left\|Y - a\right\|^2 = \mathbb{E}\left\|Y\right\|^2 - \left\|a\right\|^2 \leq \mathbb{E}\left\|Y\right\|^2$ if $a = \mathbb{E}\, Y$. ■

Lemma 16 (Descent lemma for convex cases) *Under Assumptions 1, 3, 4 and Equation 2, the averages $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ of the iterates of Algorithm with the stepsize $\eta_t \leq \frac{1}{12L}$ satisfy*

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}_1^{(t)},\ldots,\boldsymbol{\xi}_n^{(t)}} \left\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^\star\right\|^2 \leq \left(1 - \frac{\mu s}{2n}\eta_t\right) \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2 + \eta_t^2 \frac{s}{n^2}(3\bar{\sigma}^2 + 8(1 - \frac{s}{n})\bar{\zeta}^2) \\
- \eta_t \left(\frac{3}{2}\frac{s}{n} - \frac{2s^2}{3n^2}\right)\left(f(\bar{\mathbf{x}}^{(t)}) - f^\star\right) \\
+ \eta_t \left(\frac{9s}{4n} + \frac{s^2}{6n^2}\right) L\mathbb{E}_{i^{(t)},\boldsymbol{\xi}_i^{(t)}} \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2,
\end{aligned} \tag{26}$$

Proof Because all mixing matrices preserve the average (Proposition 6), we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}_i^{(t)}} \left\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^\star\right\|^2 \\
&= \mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}} \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star - \frac{\eta_t}{n}\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x}_i^{(t)})\right\| + \frac{\eta_t}{n}\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x}_i^{(t)}) - \frac{\eta_t}{n}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})^2 \\
&\quad + 2\frac{\eta_t}{n}\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}} \left[\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star - \frac{\eta_t}{n}\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x}_i^{(t)}), \sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x}_i^{(t)}) - \sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\rangle\right]
\end{aligned}$$

where $i \in [n]$. The last term is zero in expectation, as $\mathbb{E}_{\xi_i^{(t)}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) = \nabla f_i(\mathbf{x}_i^{(t)})$. The second term is estimated using Proposition 15.

The first term can be written as:

$$\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}}\left[\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^\star-\frac{\eta_t}{n}\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2\right]$$
$$=\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^\star\right\|^2+\frac{\eta_t^2}{n^2}\underbrace{\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}}\left\|\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2}_{=:T_1}-\frac{1}{n}\underbrace{2\eta_t\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}}\left\langle\bar{\mathbf{x}}^{(t)}-\mathbf{x}^\star,\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}_i^{(t)})\right\rangle}_{=:T_2}.$$

We can estimate

$$T_1\leq s^2\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}}\left\|(\nabla f_i(\mathbf{x}_i^{(t)})-\nabla f_i(\bar{\mathbf{x}}^{(t)})+\nabla f_i(\bar{\mathbf{x}}^{(t)})-\nabla f_i(\mathbf{x}^\star)+\nabla f_i(\mathbf{x}^\star))\right\|^2$$
$$\overset{(22),(24)}{\leq}\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}}\left[2s^2\left\|\nabla f_i(\mathbf{x}_i^{(t)})-\nabla f_i(\bar{\mathbf{x}}^{(t)})\right\|^2+4s^2\left\|\nabla f_i(\bar{\mathbf{x}}^{(t)})-\nabla f_i(\mathbf{x}^\star)\right\|^2+4\left\|\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}^\star)\right\|^2\right]$$
$$\overset{(19),(20)}{\leq}s^2\left(2L^2\mathbb{E}_{i^{(t)}}\left\|\mathbf{x}_i^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2+8L\left(f(\bar{\mathbf{x}}^{(t)})-f(\mathbf{x}^\star)\right)\right)+4\left\|\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}^\star)\right\|^2$$

Using the property of *sampling without replacement*, we have

$$\mathbb{E}_{r-1}\left\|\frac{1}{S}\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x})\right\|^2\leq 2(1-\tfrac{S}{N})\frac{1}{SN}\sum_i\|\nabla f_i(\mathbf{x})\|^2+\mathbb{E}_{r-1}\|\nabla f(\mathbf{x})\|^2$$

where $\nabla f(\mathbf{x})=\mathbb{E}\,\frac{1}{S}\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x})$.

$$4\frac{1}{N^2}\left\|\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}^\star)\right\|^2\leq\frac{\mathcal{S}^2}{N^2}8(1-\tfrac{S}{N})\frac{1}{SN}\sum_i\|\nabla f_i(\mathbf{x}^\star)\|^2+4\,\mathbb{E}_{r-1}\|\nabla f(\mathbf{x}^\star)\|^2=8\frac{\mathcal{S}}{N^2}(1-\tfrac{S}{N})\bar{\zeta}^2$$

And for the remaining T_2 term:

$$-\frac{1}{\eta_t}\mathbb{E}_{\mathcal{S}^{(t)},\boldsymbol{\xi}_i^{(t)}}T_2=-2s\mathbb{E}_{i^{(t)},\boldsymbol{\xi}_i^{(t)}}\left[\left\langle\bar{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)},\nabla f_i(\mathbf{x}_i^{(t)})\right\rangle+\left\langle\mathbf{x}_i^{(t)}-\mathbf{x}^\star,\nabla f_i(\mathbf{x}_i^{(t)})\right\rangle\right]$$
$$\overset{(18),(12)}{\leq}-2s\mathbb{E}_{i^{(t)},\boldsymbol{\xi}_i^{(t)}}\left[f_i(\bar{\mathbf{x}}^{(t)})-f_i(\mathbf{x}_i^{(t)})-\frac{L}{2}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)}\right\|^2+f_i(\mathbf{x}_i^{(t)})-f_i(\mathbf{x}^\star)+\frac{\mu}{2}\left\|\mathbf{x}_i^{(t)}-\mathbf{x}^\star\right\|^2\right]$$
$$\overset{(24)}{\leq}s\left[-2\left(f(\bar{\mathbf{x}}^{(t)})-f(\mathbf{x}^\star)\right)+(L+\mu)\mathbb{E}_{i^{(t)},\boldsymbol{\xi}_i^{(t)}}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)}\right\|^2-\frac{\mu}{2}\left\|\bar{\mathbf{x}}^{(t)}-\mathbf{x}^\star\right\|^2\right],$$

Putting everything together, and using that $\mu\leq L$ and $\eta_t\leq\frac{1}{12L}$ we get the statement of the lemma.
■

Lemma 17 (Recursion for consensus distance) *Under Assumptions 1, 3, 4 and Equation 2, if in addition functions F_i are convex and if stepsizes $\eta_t \leq \frac{p}{10L}$, then*

$$\Xi_t \leq \left(1 - \frac{p}{2} + \frac{s}{n}\frac{p}{4}\right)\Xi_{t-1} + \frac{s}{n}\frac{34}{p}L\eta_{t-1}^2\,\mathbb{E}\left(f(\bar{\mathbf{x}}^{(t-1)}) - f(\mathbf{x}^\star)\right) + \frac{s}{n}(4\bar{\sigma}^2 + \frac{9}{p}\bar{\zeta}^2)\eta_{t-1}^2$$

where $\Xi_t = \frac{1}{n}\,\mathbb{E}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ is a consensus distance.

Proof Unrolling $X^{(t)}$ up to $X^{(t-1)}$ using lines 3–4 of the Algorithm 2,

$$\begin{aligned}
n\Xi_t &= \mathbb{E}\left\|X^{(t-1)}W^{(t-1)} - \eta_{t-1}\partial F(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)})W^{(t-1)} - (\bar{X}^{(t-1)} - \eta_{t-1}\partial\bar{F}(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}))\right\|_F^2\\
&= \mathbb{E}\left\|X^{(t-1)}W^{(t-1)} - \bar{X}^{(t-1)} - \eta_{t-1}(\partial f(X^{(t-1)},\mathcal{S}^{(t-1)})W^{(t-1)} - \partial\bar{f}(X^{(t-1)},\mathcal{S}^{(t-1)}))\right\|_F^2\\
&\quad + \eta_{t-1}^2\,\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial f(X^{(t-1)},\mathcal{S}^{(t-1)})\right)W^{(t-1)}\right.\\
&\quad \left. - (\partial\bar{F}(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial\bar{f}(X^{(t-1)},\mathcal{S}^{(t-1)}))\right\|_F^2
\end{aligned}$$

where we used that $\mathbb{E}\,\partial F(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) = \partial f(X^{(t-1)},\mathcal{S}^{(t-1)})$ and that $\xi^{(t-1)}$ is independent of the rest.
Taking $\beta = \frac{p}{2}$ and using (2) to bound the first term we get that

$$\begin{aligned}
n\Xi_t &\leq \left(1 + \frac{p}{2}\right)(1-p)\,\mathbb{E}\left\|X^{(t-1)} - \bar{X}^{(t-1)}\right\|_F^2 + \frac{3}{p}\eta_{t-1}^2\,\mathbb{E}\left\|\partial f(X^{(t-1)},\mathcal{S}^{(t-1)})W^{(t-1)} - \partial\bar{f}(X^{(t-1)},\mathcal{S}^{(t-1)})\right\|_F^2\\
&\quad + \eta_{t-1}^2\,\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial f(X^{(t-1)},\mathcal{S}^{(t-1)})\right)W^{(t-1)}\right.\\
&\quad \left. - (\partial\bar{F}(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial\bar{f}(X^{(t-1)},\mathcal{S}^{(t-1)}))\right\|_F^2\\
&\leq \left(1 - \frac{p}{2}\right)\mathbb{E}\left\|X^{(t-1)} - \bar{X}^{(t-1)}\right\|_F^2 + \frac{3}{p}\eta_{t-1}^2\,\underbrace{\mathbb{E}\left\|\partial f(X^{(t-1)},\mathcal{S}^{(t-1)})\right\|_F^2}_{:=T_1}\\
&\quad + \eta_{t-1}^2\,\underbrace{\mathbb{E}\left\|\partial F(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial f(X^{(t-1)},\mathcal{S}^{(t-1)})\right\|_F^2}_{:=T_2},
\end{aligned}$$

Estimating separately the last two terms, and using the notation $\pm a = a - a = 0\ \forall a$,

$$\begin{aligned}
T_1 &= \mathbb{E}\left\|\partial f(X^{(t-1)},\mathcal{S}^{(t-1)}) \pm \partial f(\bar{X}^{(t-1)},\mathcal{S}^{(t-1)}) \pm \partial f(X^\star,\mathcal{S}^{(t-1)})\right\|_F^2\\
&= 3s\left(\mathbb{E}_i\left\|\left(\nabla f_i(\mathbf{x}_i^{(t-1)}) - \nabla f_i(\bar{\mathbf{x}}^{(t-1)})\right)\right\|_2^2 + \mathbb{E}_i\left\|\left(\nabla f_i(\bar{\mathbf{x}}_i^{(t-1)}) - \nabla f_i(\mathbf{x}^\star)\right)\right\|_2^2 + \mathbb{E}_i\left\|\nabla f_i(x^\star)\right\|_{22}^{22}\right)\\
&\overset{(10),(20),(13)}{\leq} 3s\left(L^2\,\mathbb{E}_i\left\|x_i^{(t-1)} - \bar{x}^{(t-1)}\right\|_2^2 + 2L\,\mathbb{E}_i\left(f_i(\bar{\mathbf{x}}^{(t-1)}) - f_i(\mathbf{x}^\star)\right) + \bar{\zeta}^2\right)
\end{aligned}$$

$$
\begin{aligned}
T_2 \leq 4\,\mathbb{E}_{\mathcal{S}}\Bigg[&\sum_{i\in\mathcal{S}}\left\|\left(\nabla F_i(\mathbf{x}_i^{(t-1)},\xi_n^{(t-1)})-\nabla F_i(\bar{\mathbf{x}}^{(t-1)},\xi_n^{(t-1)})+\nabla f_i(\mathbf{x}_i^{(t-1)})-\nabla f_i(\bar{\mathbf{x}}^{(t-1)}),i^{(t-1)})\right)\right\|_2^2\\
&+\sum_{i\in\mathcal{S}}\left\|\left(\nabla F_i(\bar{\mathbf{x}}^{(t-1)},\xi_n^{(t-1)})-\nabla F_i(\mathbf{x}^\star,\xi_n^{(t-1)})\right)\right\|_2^2+\sum_{i\in\mathcal{S}}\left\|\left(\nabla f_i(\bar{\mathbf{x}}^{(t-1)})-\nabla f_i(\mathbf{x}^\star)\right)\right\|_2^2\\
&+\sum_{i\in\mathcal{S}}\left\|\left(\nabla F_i(\mathbf{x}^\star,\xi_n^{(t-1)})-\nabla f_i(\mathbf{x}^\star)\right)\right\|_2^2\Bigg]\\
\overset{(22),(20)}{\leq}\;& 4s\left(4L^2\,\mathbb{E}\left\|x_i^{(t-1)}-\bar{x}^{(t-1)}\right\|_2^2+4L\,\mathbb{E}\left(f_i(\bar{x}^{(t-1)})-f_i(x^\star)\right)+\bar{\sigma}^2\right)
\end{aligned}
$$

Putting back estimates for T_1 and $T_2$and using that $\eta_t \leq \frac{p}{10L}$ we arrive to the statement of the lemma. ■

D.2. Non-convex case

Here we derive descent recursive equation (6) and recursion for consensus distance (7) for the non-convex case.

Proposition 18 (Mini-batch variance with node sampling) *Let functions $F_i(\mathbf{x},\xi)$, $i\in[n]$ be L-smooth (Assumption 1) with bounded noise as in Assumption 5. Then for any $\mathbf{x}_i\in\mathbb{R}^d, i\in[n]$ and $\bar{\mathbf{x}}:=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i$ it holds*

$$
\mathbb{E}_{t+1}\left\|\frac{1}{n}\sum_{i\in\mathcal{S}}\left(\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})-\nabla f_i(\mathbf{x}_i^{(t)})\right)\right\|_2^2\leq\frac{s}{n}\left(\frac{\hat{\sigma}^2}{n}+\frac{2M}{n^2}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2+\frac{2M}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|\right)
\tag{27}
$$

Proof

$$
\begin{aligned}
\mathbb{E}_{t+1}\left\|\frac{1}{n}\sum_{i\in\mathcal{S}}\left(\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})-\nabla f_i(\mathbf{x}_i^{(t)})\right)\right\|_2^2&\overset{(16)}{\leq}\frac{s}{n}\left(\frac{\hat{\sigma}^2}{n}+\frac{M}{n^2}\sum_{i=1}^n\left\|\nabla f(\mathbf{x}_i^{(t)})\pm\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2\right)\\
&\overset{(22)}{\leq}\frac{s}{n}\left(\frac{\hat{\sigma}^2}{n}+2\frac{M}{n^2}\sum_{i=1}^n\left\|\nabla f(\mathbf{x}_i^{(t)})-\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2+2\frac{M}{n^2}\sum_{i=1}^n\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2\right)\\
&\overset{(11)}{\leq}\frac{s}{n}\left(\frac{\hat{\sigma}^2}{n}+\frac{2M}{n^2}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|^2+\frac{2M}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2\right)
\end{aligned}
$$

■

Lemma 19 (Descent lemma for non-convex case) *Under Assumptions 2, 5 and Equation 2, the averages $\bar{\mathbf{x}}^{(t)}:=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i^{(t)}$ of the iterates of Algorithm 1 with the constant stepsize $\eta\leq\frac{1}{4L(M+1+2p)}$ satisfy*

$$
\begin{aligned}
&\mathbb{E}_{t+1}\,f(\bar{\mathbf{x}}^{(t+1)})\\
&\leq f(\bar{\mathbf{x}}^{(t)})+\frac{s}{n}\frac{\eta L^2}{n}\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2-\frac{s}{n}\frac{\eta}{4}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2+L\eta^2\frac{s}{n}\frac{\hat{\sigma}^2}{n}+L\eta^2\frac{s^2}{n^2}\frac{2}{s}(1-\frac{s}{n})\hat{\zeta}^2
\end{aligned}
\tag{28}
$$

Proof Because all mixing matrices preserve the average (Proposition 6) and function f is L-smooth, we have

$$\begin{aligned}\mathbb{E}_{t+1}\, f(\bar{\mathbf{x}}^{(t+1)}) &= \mathbb{E}_{t+1}\, f\left(\bar{\mathbf{x}}^{(t)} - \frac{\eta}{n}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right) \\ &\leq f(\bar{\mathbf{x}}^{(t)}) - \underbrace{\mathbb{E}_{t+1}\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{\eta}{n}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\rangle}_{:=T_1} + \mathbb{E}_{t+1}\,\frac{L}{2}\eta^2 \underbrace{\left\|\frac{1}{n}\sum_{i\in\mathcal{S}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|_2^2}_{:=T_2}\end{aligned}$$

To estimate the second term, we add and subtract $\sum_{i\in\mathcal{S}} \nabla f_i(\bar{\mathbf{x}}^{(t)})$.

$$\begin{aligned}T_1 &\leq \mathbb{E}_{t+1}\, -\eta\frac{s}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{\eta}{2}\frac{s}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{\eta}{2n}\frac{s}{n}\sum_{i=1}^n \left\|\left(\nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right)\right\|_2^2 \\ &\overset{(23),\gamma=1;(22)}{\leq} -\frac{\eta}{2}\frac{s}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta}{2n}\frac{s}{n}L^2\sum_{i=1}^n \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|_2^2\end{aligned}$$

For the last term, add and subtract $\frac{1}{n}\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x}_i^{(t)})$,

$$\begin{aligned}T_2 &= \mathbb{E}_{t+1}\left\|\frac{1}{n}\sum_{i\in\mathcal{S}}\left(\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right)\right\|_2^2 + \mathbb{E}_{t+1}\left\|\frac{1}{n}\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}_i^{(t)})\right\|_2^2 \\ &\overset{\text{Proposition 18}}{\leq} \frac{s}{n}\left(\frac{\hat{\sigma}^2}{n} + \frac{2M}{n^2}L^2\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + \frac{2M}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2\right) + \frac{s^2}{n^2}\underbrace{\mathbb{E}_{t+1}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\nabla f_i(\mathbf{x}_i^{(t)})\right\|_2^2}_{:=T_3}\end{aligned}$$

The inequality derived from sampling from replacement property (D.1) for the convex case only holds when $\mathbf{x}$ is independent of i. However, $\mathbf{x}_i^{(t)}$ in T_3 is dependent on i. We therefore transform T_3 to $\mathbb{E}_{t+1}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\nabla f_i(\bar{\mathbf{x}})\right\|_2^2$, where $\bar{\mathbf{x}}^{(t)}$ is independent on i:

$$\begin{aligned}T_3 &= \left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\left(\nabla f_i(\mathbf{x}_i^{(t)}) - f_i(\bar{\mathbf{x}}^{(t)})\right) + \frac{1}{s}\sum_{i\in\mathcal{S}}\nabla f_i(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &\overset{(24)}{\leq} 2\,\underbrace{\mathbb{E}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\left(\nabla f_i(\mathbf{x}_i^{(t)}) - f_i(\bar{\mathbf{x}}^{(t)})\right)\right\|_2^2}_{:=T_4} + 2\,\underbrace{\mathbb{E}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\nabla f_i(\bar{\mathbf{x}}^{(t)})\right\|_2^2}_{:=T_5}\end{aligned}$$

$$T_4 \overset{(22)}{\leq} \mathbb{E}_{\mathcal{S}}\,\frac{1}{s^2}s\sum_{i\in\mathcal{S}}\left\|\left(\nabla f_i(\mathbf{x}_i^{(t)}) - f_i(\bar{\mathbf{x}}^{(t)})\right)\right\|_2^2 \overset{(11)}{\leq} \frac{1}{n}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2$$

Using sampling without replacement,

$$
\begin{aligned}
T_5 &= \mathbb{E}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}}\nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f(\bar{\mathbf{x}}^{(t)}) + \nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\
&\leq \mathbb{E}\left(1-\frac{s-1}{n-1}\right)\frac{\left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(\bar{\mathbf{x}}^{(t)})^2\right) - \nabla f(\bar{\mathbf{x}}^{(t)})^2}{s} + \left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 \\
&\overset{(15)}{\leq} \mathbb{E}\,\frac{2}{s}\left(1-\frac{s}{n}\right)\left(\zeta^2 + P\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2\right) + \left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 \\
&= \frac{2}{s}\left(1-\frac{s}{n}\right)\zeta^2 + \left(\frac{2}{s}(1-\frac{s}{n})P + 1\right)\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2
\end{aligned}
$$

Combining this,

$$
\begin{aligned}
T_2 \leq{}& \frac{s}{n}\left(\frac{\hat{\sigma}^2}{n} + \frac{2M}{n^2}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + \frac{2M}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|\right) \\
&+ \frac{s^2}{n^2}\left(\frac{2}{n}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{4}{s}\left(1-\frac{s}{n}\right)\zeta^2 + 2\left(\frac{2}{s}(1-\frac{s}{n})P+1\right)\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2\right)
\end{aligned}
$$

When $s = n$, the first term becomes $\left(\frac{\hat{\sigma}^2}{n} + \frac{2M}{n^2}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + \frac{2M}{n}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|\right)$, the second term becomes $\frac{2}{n}L^2\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2$, the third term becomes 0, the last term becomes $2\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2$. Therefore, the T_2 for $s = n$ case in [5] is recovered.
When $\frac{s}{n} = 1$, T_1 for $s = n$ case in [5] is also recovered.
Putting back estimations and using that $\eta \leq \frac{1}{4L(M+1+2p)}$ we arrive to the statement of this lemma.
■

Lemma 20 (Recursion for consensus distance) *Under Assumptions 2, 5 and Equation 2, if in addition functions F_i are convex and if stepsizes $\eta_t \leq \frac{p}{2L}\sqrt{\frac{1}{2(3+pM)}}$, then*

$$
\Xi_t \leq \left(1-\frac{p}{4}\right)\mathbb{E}\,\Xi_{t-1} + 2P(1-p)(\frac{3}{p}+M)\eta_{t-1}^2\left\|\nabla f(\bar{\mathbf{x}}^{(t-1)})\right\|_2^2 + (1-p)\left(\hat{\sigma}^2 + 2(\frac{3}{p}+M)\hat{\zeta}^2\right)\eta_{t-1}^2 \tag{29}
$$

where $\Xi_t = \frac{1}{n}\mathbb{E}\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ is a consensus distance.

Proof Using matrix notation (5), for $t \geq 1$, $n\Xi_t = \mathbb{E}\left\|X^{(t)} - \bar{X}^{(t)}\right\|_F^2$. Unrolling $X^{(t)}$ up to $X^{(t-1)}$ using lines 3–4 of the Algorithm 2,

$$
\begin{aligned}
n\Xi_t = \mathbb{E}\Big\|&X^{(t-1)}W^{(t-1)} - \bar{X}^{(t-1)} - \eta_{t-1}(\partial f(X^{(t-1)},\mathcal{S}^{(t-1)})W^{(t-1)} - \partial\bar{f}(X^{(t-1)},\mathcal{S}^{(t-1)}))\Big\|_F^2 \\
&+ \eta_{t-1}^2\,\mathbb{E}\Big\|\left(\partial F(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial f(X^{(t-1)},\mathcal{S}^{(t-1)})\right)W^{(t-1)} \\
&- (\partial\bar{F}(X^{(t-1)},\xi^{(t-1)},\mathcal{S}^{(t-1)}) - \partial\bar{f}(X^{(t-1)},\mathcal{S}^{(t-1)}))\Big\|_F^2
\end{aligned}
$$

Taking $\beta = \frac{p}{2}$, we then have $1 + \frac{1}{\beta} < \frac{3}{p}$, further using (2) to bound the first term we get that

$$\begin{aligned} n\Xi_t &\leq \left(1 + \frac{p}{2}\right)(1-p)\,\mathbb{E}\left\|X^{(t-1)} - \bar{X}^{(t-1)}\right\|_F^2 + \frac{3}{p}\eta_{t-1}^2\,\mathbb{E}\left\|\partial f(X^{(t-1)})\right\|_F^2 \\ &\quad + \eta_{t-1}^2\,\mathbb{E}\left\|\partial F(X^{(t-1)}, \xi^{(t-1)}) - \partial f(X^{(t-1)})\right\|_F^2 \\ &\leq \left(1 - \frac{p}{2}\right)\mathbb{E}\left\|X^{(t-1)} - \bar{X}^{(t-1)}\right\|_F^2 + \frac{3}{p}\eta_{t-1}^2\,\mathbb{E}\underbrace{\left\|\partial f(X^{(t-1)})\right\|_F^2}_{:=T} \\ &\quad + \eta_{t-1}^2\underbrace{\mathbb{E}\left\|\partial F(X^{(t-1)}, \xi^{(t-1)}) - \partial f(X^{(t-1)})\right\|_F^2}_{:=T_2}, \end{aligned}$$

where we used that $\left\|X^{(t)}W - \overline{X}^{(t)}\right\| = \left\|X^{(t)}(W - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\right\| \leq \left\|X^{(t)}\right\|$ from [4].
From Assumption 5, $T_2 = \sum_{i=1}^n \mathbb{E}_{\xi_i}\left\|\nabla F_i(\mathbf{x}_i, \xi_i) - \nabla f_i(\mathbf{x}_i)\right\|_2^2 \leq n\hat{\sigma}^2 + M\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_i)\right\|_2^2$.

Estimating T,

$$\begin{aligned} T &\overset{(24)}{\leq} 2\left\|\partial f(X^{(t-1)}) - \partial f(\bar{X}^{(t-1)})\right\|_F^2 + 2\left\|\partial f(\bar{X}^{(t-1)})\right\|_F^2 \\ &\overset{(11),(15)}{\leq} 2L^2\left\|X^{(t-1)} - \bar{X}^{(t-1)}\right\|_F^2 + 2n\hat{\zeta}^2 + 2Pn\left\|\nabla f(\bar{\mathbf{x}}^{(t-1)})\right\|_2^2 \end{aligned}$$

Putting back estimate for T and using that $\eta_t \leq \frac{p}{2L}\sqrt{\frac{1}{2(3+pM)}}$ we arrive to the statement of this lemma. ■

D.3. Simplifying the Consensus Recursion

In Lemmas 17, 20 we obtained the consensus recursive equation (7) for both convex and non-convex cases. In this section we simplify it to be able to easily combine it later with (6).

Lemma 21 *If non-negative sequences $\{\Xi_t\}_{t\geq 0}$, $\{e_t\}_{t\geq 0}$ and $\{\eta_t\}_{t\geq 0}$ satisfy (7) for some constants $0 < p \leq 1, A, D \geq 0$, moreover if the stepsizes $\{\eta_t^2\}_{t\geq 0}$ is $\frac{8}{p}$-slow decreasing sequence (Definition 13), and if $\{w_t\}_{t\geq 0}$ is $\frac{16}{p}$-slow increasing non-negative sequence of weights, then it holds that*

$$B\sum_{t=0}^T w_t\Xi_t \leq \frac{b}{2}\sum_{t=0}^T w_t e_t + \frac{16}{p}AB\sum_{t=0}^T w_t\eta_t^2,$$

for some constant $B > 0$ with the constraint that stepsizes $\eta_t \leq \frac{1}{8}\sqrt{\frac{pb}{DB}}$.

Proof Set $H = 1 - \frac{p}{4}$ and recursively substituting every Ξ_t with Ξ_{t-1} we get

$$\Xi_t \leq H\Xi_{t-1} + D\eta_{t-1}^2 e_{t-1} + A\eta_{t-1} \leq H^2\Xi_{t-2} + D\sum_{j=t-2}^{t-1} H^{t-1-j}\eta_j^2 e_j + A\sum_{j=t-2}^{t-1} H^{t-1-j}\eta_j^2$$

Unrolling Ξ_t recursively up to 0 we get,

$$\Xi_t \leq D\sum_{j=0}^{t-1} H^{t-1-j}\eta_j^2 e_j + A\sum_{j=0}^{t-1} H^{t-1-j}\eta_j^2$$

Now using that η_t^2 is $\frac{4}{p}$-slow decreasing, i.e. $\eta_j^2 \leq \eta_t^2\left(1+\frac{p}{8}\right)^{t-j}$,

$$\begin{aligned}\Xi_t &\leq \frac{4}{4-p}D\sum_{j=0}^{t-1}(1-\frac{p}{4})^{t-j}\eta_j^2 e_j + \frac{4}{4-p}A\sum_{j=0}^{t-1}(1-\frac{p}{4})^{t-j}\eta_j^2 \\ &\leq \frac{4}{4-p}D\eta_t^2\sum_{j=0}^{t-1}\left(1-\frac{p}{8}\right)^{t-j}e_j + \frac{4}{4-p}\frac{8}{p}A\eta_t^2\end{aligned}$$

Now averaging Ξ_t with weights w_t,

$$B\sum_{t=0}^{T} w_t\Xi_t \leq \frac{4}{4-p}(DB\sum_{t=0}^{T} w_t\eta_t^2\sum_{j=0}^{t-1}\left(1-\frac{p}{8}\right)^{t-j}e_j + \frac{8}{p}AB\sum_{t=0}^{T} w_t\eta_t^2)$$

using that w_t is $\frac{8}{p}$-slow increasing sequence, i.e. $w_t \leq w_j\left(1+\frac{p}{16}\right)^{t-j}$ and $\eta_t \leq \frac{1}{8}\sqrt{\frac{pb}{DB}}$,

$$B\sum_{t=0}^{T} w_t\Xi_t \leq \frac{4}{4-p}(\underbrace{\frac{pb}{64}\sum_{t=0}^{T}\sum_{j=0}^{t-1} w_j\left(1-\frac{p}{16}\right)^{t-j}e_j}_{:=T_1} + \frac{8}{p}AB\sum_{t=0}^{T} w_t\eta_t^2)$$

$$T_1 = \frac{pb}{64}\sum_{j=0}^{T} w_j e_j\sum_{t=j+1}^{T}\left(1-\frac{p}{16}\right)^{t-j} \leq \frac{pb}{64}\sum_{j=0}^{T} w_j e_j\sum_{t=0}^{\infty}\left(1-\frac{p}{16}\right)^{t-j} \leq \frac{b}{4}\sum_{t=0}^{T} w_t e_t$$

Then we have

$$B\sum_{t=0}^{T} w_t\Xi_t \leq \frac{4}{4-p}(\frac{b}{4}\sum_{t=0}^{T} w_t e_t + \frac{8}{p}AB\sum_{t=0}^{T} w_t\eta_t^2) \leq \frac{b}{2}\sum_{t=0}^{T} w_t e_t + \frac{16}{p}AB\sum_{t=0}^{T} w_t\eta_t^2,$$

■

Appendix E. Solving the Main Recursion (9)

E.1. $a > 0$ (strongly convex case)

Lemma 22 *If non-negative sequences $\{r_t\}_{t\geq 0}, \{e_t\}_{t\geq 0}$ satisfy (9) for some constants $a, b, p > 0, c, A, B$, then there exists a constant stepsize $\eta_t = \eta < \frac{1}{d}$ such that for weights $w_t = (1-a\eta)^{-(t+1)}$ and $W_T := \sum_{t=0}^{T} w_t$ it holds:*

$$\frac{1}{2W_T}\sum_{t=0}^{T} be_t w_t + ar_{T+1} \leq \tilde{\mathcal{O}}\left(\frac{c}{aT} + \frac{BA}{a^2T^2}\frac{1}{p} + r_0 d\exp\left[-\frac{a(T+1)}{d}\right]\right),$$

where $\tilde{\mathcal{O}}$ hides polylogarithmic factors. [5]

Applying Lemma Theorem 22,

$$\frac{1}{2W_T}\sum_{t=0}^{T} e_t w_t + \frac{\mu}{2} r_{T+1} \leq \frac{1}{b}\left(\frac{1}{2W_T}\sum_{t=0}^{T} b e_t w_t + 2a r_{T+1}\right)$$
$$\leq \tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu T s} + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{\mu T s} + \frac{L(\bar{\zeta}^2 + p\bar{\sigma}^2)}{\mu^2 T^2 p^2}(\frac{n}{s}) + \frac{n}{s} r_0 d \exp\left[-\frac{a(T+1)}{d}\right]\right)$$

Thus, the final rate of convergence:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu s \epsilon} + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{\mu s \epsilon} + \frac{\sqrt{L}(\bar{\zeta} + \sqrt{p}\bar{\sigma})}{\mu p \sqrt{\epsilon}}\sqrt{\frac{n}{s}} + \frac{n}{s}\frac{L}{\mu p}\log\left(\frac{1}{\epsilon}\right)\right)$$

Lemma 23 *If non-negative sequences $\{r_t\}_{t\geq 0}, \{e_t\}_{t\geq 0}$ satisfy Lemma 19 and Lemma 21 with $a = 0, b > 0, c, A, B \geq 0$, then there exists a constant stepsize $\eta_t = \eta < \frac{1}{d}$ such that for weights $\{w_t = 1\}_{t\geq 0}$ it holds that [5]:*

$$\frac{1}{(T+1)}\sum_{t=0}^{T} e_t \leq \mathcal{O}\left(2\left(\frac{c r_0}{T+1}\right)^{\frac{1}{2}} + 2\left(\frac{BA\tau}{p}\right)^{1/3}\left(\frac{r_0}{T+1}\right)^{\frac{2}{3}} + \frac{d r_0}{T+1}\right).$$

Utilizing Lemma 23 and choosing $d = \mathcal{O}\left(\frac{L}{p}\sqrt{(P+1)(M+1)}\right)$, the final convergence rate:

$$\mathcal{O}\left(\frac{\hat{\sigma}^2 + (1-\frac{s}{n})\hat{\zeta}^2}{s\epsilon^2} + \frac{n}{s}\frac{\hat{\zeta}^2\sqrt{(M+1)} + \hat{\sigma}\sqrt{p}}{p\epsilon^{3/2}} + \frac{\sqrt{(P+1)(M+1)}}{p\epsilon}\right) L r_0$$

Appendix F. Discussion

Our rates recover the results of [5] for both convex and non-convex cases, by setting $s = n$.

Results for convex $f_i(\mathbf{x})$. The convergence rate we derived for convex $f_i(\mathbf{x})$:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu s}\epsilon + (1-\frac{s}{n})\frac{\bar{\zeta}^2}{\mu s \epsilon} + \frac{\sqrt{L}(\bar{\zeta} + \sqrt{p}\bar{\sigma})}{\mu p\sqrt{\epsilon}}\sqrt{\frac{n}{s}} + \frac{n}{s}\frac{L}{\mu p}\log\left(\frac{1}{\epsilon}\right)\right).$$

The convergence rate in [5] for convex $f_i(\mathbf{x})$:

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{\mu n}\epsilon + \frac{\sqrt{L}(\bar{\zeta}\tau + \bar{\sigma}\sqrt{p\tau})}{\mu p\sqrt{\epsilon}} + \frac{L\tau}{\mu p}\log\frac{1}{\epsilon}\right).$$

When $s = n$ and $\tau = 1$ (since we assumed there is no delay in our analysis), these two convergence rates are identical.

Results for non-convex $f_i(\mathbf{x})$. The convergence rate we derived for non-convex $f_i(\mathbf{x})$ is

$$\mathcal{O}\left(\frac{\hat{\sigma}^2+(1-\frac{s}{n})\hat{\zeta}^2}{s}\epsilon^2+\frac{n}{s}\frac{\hat{\zeta}^2\sqrt{(M+1)}+\hat{\sigma}\sqrt{p}}{p\epsilon^{3/2}}+\frac{\sqrt{(P+1)(M+1)}}{p\epsilon}\right)Lr_0\,,$$

while the convergence rate in [5] for convex case:

$$\mathcal{O}\left(\frac{\hat{\sigma}^2}{n}\epsilon^2+\frac{\hat{\zeta}\sqrt{M+1}+\hat{\sigma}\sqrt{p}}{p\epsilon^{3/2}}+\frac{\sqrt{(P+1)(M+1)}}{p\epsilon}\right)\cdot Lr_0$$

By setting $s=n$ and $\tau=1$, these two convergence rates are the same.

Appendix G. Additional Experiments

Setup. We consider three common network topologies, *ring*, 2-d *torus* and *fully-connected* graph and use the Metropolis-Hasting mixing matrix W, i.e. $w_{ij}=w_{ji}=\frac{1}{deg(i)+1}=\frac{1}{deg(j)+1}$ for $\{i,j\}\in E$. For each set of $\bar{\sigma}^2$ and $\bar{\zeta}^2$, we tune the step size for all these three topologies to reach the desired target accuracy ϵ with the fewest number of iterations.

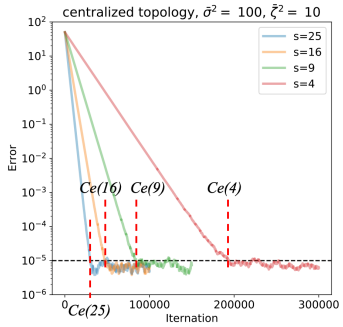


Figure 6: Convergence of $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)}-\mathbf{x}^\star\right\|_2^2$ to target accuracy $\epsilon=10^{-5}$ for $\bar{\sigma}^2=100$, $\bar{\zeta}^2=10$, and centralized topology on $n=25$ nodes, $s=9,16,25$, $d=50$. Stepsizes are the same for all experiments. We refer to iterations taken to reach $\epsilon=10^{-5}$ as $ce(s)$, where s denotes the number of nodes sampled out of n per iteration.

Appendix H. Theorem of Expected Consensus Rate for Sub-sampled Ring Topology

Here we use the Equation 4 to derive the expected consensus rate for random ring topology in Theorem 4. The proofs for centralized and random torus follow similarly.

In the random ring topology, the randomness is not only in choosing a set $\mathcal{S}$ of s workers but also in how to connect the s workers in $\mathcal{S}$ to be a ring. Here, we consider the case where every possible way of connecting $\mathcal{S}$ to a ring is of equal probability.

Here we derive the formula for $\mathbb{E}\left[W^TW\right]$ when $\mathcal{S}$ is fixed, i.e. we take expectation over possible ways to connect s workers to be a ring. When $s=2,3$, ring topology is the same as

n	s	$t = \left(\frac{\bar{\sigma}^2}{s} + \left(1 - \frac{s}{n}\right)\frac{\bar{\zeta}^2}{s}\right)$ $\bar{\sigma}^2 = 100, \bar{\zeta}^2 = 10$	centralized topology iteration taken to reach $\epsilon = 10^{-5}$ **(ce)**	grid topology iteration taken to reach $\epsilon = 10^{-5}$ (ge)	ring topology iteration taken to reach $\epsilon = 10^{-5}$ (re)
25	4	t(4) = 27.1	ce(4) = 192730	\	re(4) = 192730
25	9	t9 = 11.8	ce(9) = 82518	ge(9) = 82521	re(9) = 82525
25	16	t(16) = 6.475	ce(16) = 46472	ge(16) = 46473	re(16) = 46473
25	25	t25 = 4	ce(250 = 29599	ge(25) = 29599	re(25) = 29599

ratios	t(i) / t(j)	ce(i) / ce(j)	ge(i) / ge(j)	re(i) / re(j)
i = 4, j = 9	2.33	2.33	\	2.34
i = 4, j = 16	4.14	4.15	\	4.15
i = 4, j = 25	6.51	6.51	\	6.51
i = 9, j = 16	1.77	1.78	1.78	1.78
i = 9, j = 25	2.78	2.79	2.8	2.79
i = 16, j = 25	1.57	1.57	1.57	1.57

Figure 7: Ratios between theoretical convergence rate and numerical performances.

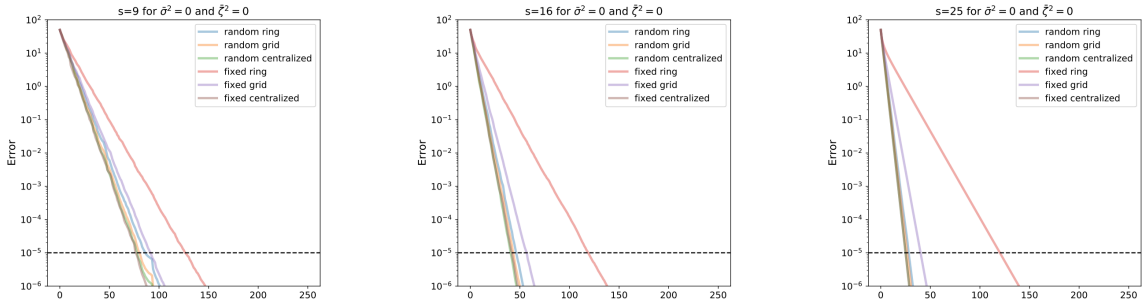


Figure 8: Convergence of $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\right\|_2^2$ to target accuracy $\epsilon = 10^{-5}$ for $\bar{\sigma}^2 = \bar{\zeta}^2 = 0$, and different topologies on $n = 25$ nodes, $s = 9, 16, 25$, $d = 50$. Stepsizes are the same for all experiments.

centralized topology. Therefore, in the discussion in this section, we only concern with the case where $s \geq 4$.

When $s = 4$, $\mathbb{E}\left[W^T W\right]$ is invariant to how nodes are connected. However, when $s \geq 5$, $\mathbb{E}\left[W^T W\right]$ does depend on how nodes are connected. According to the matrix multiplication formula,

$$(W_t^T W_t)_{(ij)} = \sum_{k=1}^{n} (W_t^T)_{(ik)} (W_t)_{(kj)} = \sum_{k=1}^{n} (W_t)_{(ki)} (W_t)_{(kj)}$$

where $(W_t)_{n\times n}$, $1 \leq i, j \leq n$ and if both node k and node m are sampled,

$$(W_t)_{km} = \begin{cases} \frac{1}{3}, & \text{if node } m \text{ and node } k \text{ are connected;} \\ 0, & \text{if node } m \text{ and node } j \text{ are not connected.} \end{cases}$$

Assuming i and j are both sampled, we summarized that

$$(W_t^T W_t)_{(ij)} = \begin{cases} \frac{1}{3}, & \text{if } i = j; \\ \frac{2}{9}, & \text{if } i \neq j, \text{ node } i \text{ and } j \text{ are connected directly;} \\ \frac{1}{9}, & \text{if } i \neq j, \text{ node } i \text{ and } j \text{ are both connected to the same node;} \\ 0, & \text{if } i \neq j, \text{ node } i \text{ and } j \text{ are neither connected directly nor both connected to the same node.} \end{cases} \tag{30}$$

The reason is listed as the following,

- If $i = j$, $(W_t^T W_t)_{(ii)} = \sum_{k=1}^n (W_t)_{ki}^2 = \frac{1}{3}$, since each node is connected to itself and 2 other nodes, i.e. there are three k $s.t.$ $W_{ki} = \frac{1}{3}$.
- If $i \neq j$ AND node i and node j are connected directly, $(W_t^T W_t)_{(ij)} = \sum_{k=1}^n (W_t)_{ki}(W_t)_{kj} = \frac{2}{9}$, since $(W_t)_{ki}(W_t)_{kj} = \frac{1}{3}$ when $k = i$ or $k = j$. There are no other nodes connected to both i and j, otherwise, there would be a ring formed by three nodes, which violates that $s \geq 5$.
- If $i \neq j$ AND node i and node j are not connected directly but are both connected to a node, $(W_t^T W_t)_{(ij)} = \sum_{k=1}^n (W_t)_{ki}(W_t)_{kj} = \frac{1}{9}$, since there is one and only one node connected to both i and j. There are no other nodes connected to both i and j, otherwise, there would be a ring formed by four nodes, which violates that $s \geq 5$.
- If $i \neq j$ AND node i and node j are neither connected directly nor both connected to a node, $(W_t^T W_t)_{(ij)} = \sum_{k=1}^n (W_t)_{ki}(W_t)_{kj} = 0$.

Then we derive the probability of each condition in Equation 30.

We state that when connecting s nodes to be a ring, there are $(s-1)!/2$ possible rings if we consider a clockwise connection the same as a counter-clockwise connection, i.e. we consider $1 \to 2 \to 3 \to 4 \to 5$ and $1 \to 5 \to 4 \to 3 \to 2$ as the same ring. We fix a node as the starting point, and the $(s-1)!$ permutations of other nodes include the clockwise and counter-clockwise connection versions of all possible rings.

When a node i and j are connected, there are $(s-2)!$ possible rings, which is the number of permutations of nodes except for nodes i and j. Node i and j can be seen as one node, yet in this case, clock-wise and counter-clockwise connections would be different, since $\cdots \to i \to j \to \cdots$ and $\cdots \to j \to i \to \cdots$ would be different.

Therefore, the possibility of node i and j connected directly is $\frac{(s-2)!}{(s-1)!/2} = \frac{2}{(s-1)}$.

When a node i and j are connected to a common node k, there are also $(s-2)!$ possible rings, which is (the number of permutations of nodes except for i, j, and k) $\times(s-2)$. Node i, j, and k can be seen as one node, and clock-wise and counter-clockwise connections are still different, so the number of permutations is $(s-3)!$. There are $(s-2)$ possible choices for k.

Therefore, the possibility of node i and j connected to a common node is also $\frac{(s-2)!}{(s-1)!/2} = \frac{2}{(s-1)}$.

We validated this on two examples. In the example of $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, (the prob of node 1 and 2 connected directly) = (the prob of node 1 and 2 connected to the same node) $= \frac{2}{5} = \frac{2}{n-1}$.

In the example of $\mathcal{S} = \{1, 2, 3, 4, 5\}$, (the prob of node 1 and 2 connected directly) = (the prob of node 1 and 2 connected to the same node) $= \frac{1}{2} = \frac{2}{n-1}$.

Therefore, we can summarize that, when $i \neq j$ and both node i and node j are selected to be in $\mathcal{S}$,

$$(W^T W)_{(ij)} = \begin{cases} \frac{2}{9}, & \text{with probability } \frac{2}{(s-1)}; \\ \frac{1}{9}, & \text{with probability } \frac{2}{(s-1)}; \\ 0, & \text{with probability} 1 - \frac{4}{(s-1)}. \end{cases}$$

Thus,

$$\mathbb{E}(W_t^T W_t)_{(ij)} = \begin{cases} \frac{2}{3(s-1)} & \text{if } i \neq j, i, j \in \mathcal{S}; \\ \frac{1}{3} & \text{if } i = j, i, j \in \mathcal{S}; \\ 0 & \text{if } i \neq j, i \text{ or j } \notin \mathcal{S}; \\ 1 & \text{if } i = j, i \notin \mathcal{S}. \end{cases}$$

The probability of a node i being selected is $\frac{s}{n}$, the probability of having node i and node j both selected is $\frac{s(s-1)}{n(n-1)}$. Therefore, the expectation of a diagonal element is $\frac{1}{3} \times \frac{s}{n} + 1 \times \left(1 - \frac{s}{n}\right) = 1 - \frac{2s}{3n}$; the expectation of a non-diagonal element is $\frac{2}{3(s-1)} \times \frac{s(s-1)}{n(n-1)} = \frac{2}{3(s-1)} \times \frac{s(s-1)}{n(n-1)} = \frac{2s}{3n(n-1)}$.

Therefore, the expected consensus rate of sub-sampled random ring topology is $1 - p = 1 - \frac{2s}{3n} - \frac{2s}{3n(n-1)} = 1 - \frac{2s}{3(n-1)}$.