OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Data-heterogeneity-aware Mixing for Decentralized Learning

Yatin Dandi YATIN.DANDI@EPFL.CH
Anastasia Koloskova ANASTASIA.KOLOSKOVA@EPFL.CH
Martin Jaggi MARTIN.JAGGI@EPFL.CH
EPFL, Switzerland
Sebastian U. Stich STICH@CISPA.DE
CISPA, Germany

Abstract

Decentralized learning provides an effective framework to train machine learning models with data distributed over arbitrary communication graphs. However, most existing approaches towards decentralized learning disregard the interaction between data heterogeneity and graph topology. In this paper, we characterize the dependence of convergence on the relationship between the mixing weights of the graph and the data heterogeneity across nodes. We propose a metric that quantifies the ability of a graph to mix the current gradients. We further prove that the metric controls the convergence rate, particularly in settings where the heterogeneity across nodes dominates the stochasticity between updates for a given node. Motivated by our analysis, we propose an approach that periodically and efficiently optimizes the metric using standard convex constrained optimization and sketching techniques.

1. Introduction

Machine learning is gradually shifting from classical centralized training to decentralized data processing. For example, federated learning (FL) allows multiple parties to jointly train an ML model together without disclosing their personal data to others [18]. While FL training relies on a central coordinator, fully distributed learning methods instead use direct peer-to-peer communication between the parties (e.g. personal devices, organization, or compute nodes inside a datacenter) [3, 21, 27, 36]. In decentralized learning, communication is limited to the network topology. The nodes can only communicate with their direct neighbors in the network in each round of (one hop) communication [47].

The convergence in decentralized learning with distributed SGD [D-SGD, 27] crucially depends on two factors (i) the spread of information, i.e. many rounds (hops) of communication are required to spread information to all nodes in the network [3, 48], and (ii) the heterogeneity of the data sources, i.e. when local data on each node is drawn from different distributions [6, 16, 19, 52]. The influence of the first factor is usually expressed through the spectral gap of the mixing matrix [3, 13, 35, 39, 56]. However, spectral gap only provides a worst case diffusion rate independent of the data heterogeneity. The effect of data-heterogeneity is commonly incorporated through the variance of the gradients for the objectives across nodes [22], disregarding the relationship between nodes due to the underlying topology.

Instead of viewing these factors independently, we propose an analysis that incorporates the interaction between the graph topology and the data heterogeneity. Concretely, we refine the theo-

retical analysis of D-SGD to reveal precisely the tight interplay between the graph's mixing matrix and the time-varying distribution of gradients across nodes. We prove that the above interaction affects the convergence through a metric, that we call *relative heterogeneity*. The metric quantifies the error between the average of the gradient across all the nodes and the updates obtained by each node when using the given mixing weights to average the gradients of its neighbours at the same value of the parameters. Crucially, the metric doesn't depend on the magnitude of the consensus and the stochastic noise in the gradients, allowing us to disentangle the precise effect of data heterogeneity on the convergence.

Motivated by our theoretical analysis, we propose an approach to optimize the mixing matrix to have low *relative heterogeneity*. In contrast to existing works attempting to tackle data-heterogeneity through specialized algorithms [28, 30, 38, 46], we investigate how the performance of D-SGD can be improved by a *time-varying* and *data-aware* design of the mixing matrix (while respecting the network topology). We show that the design of an optimal data-dependent mixing matrix can be described as a quadratic program that can efficiently be solved.

Subsequently, we combine the above optimization with D-SGD to dynamically adapt the mixing matrix during training .While our approach only updates the mixing matrix, our metric can also be used to select different topologies for a given heterogeneity across nodes as well as different permutations of the nodes for a given network (see Appendix G).

Our main contributions are as follows:

1. We provide a tighter convergence analysis of DSGD by introducing a new metric that captures the interplay between the communication topology and data heterogeneity in decentralized (and federated) learning.

2. We propose a communication and computation efficient algorithm to design data-aware mixing matrices in practice and verify its effectiveness through a set of extensive experiments on synthetic and real data.

2. Related Work

Decentralized convex optimization over arbitrary network topologies has been studied in [13, 34, 47, 53] and decentralized versions of the stochastic gradient method (D-SGD) have been analyzed in [22, 26, 27, 49]. It was found that the convergence of D-SGD is strongly affected by heterogeneous data. Such impacts are not only observed in practice [16, 27], but also verified theoretically by theoretical complexity lower bounds [19, 22, 55].Concurrent to our work, [5] provided a similar analysis of the convergence rate of D-SGD for the smooth convex case using a metric quantifying the mixing error in gradients named "neighborhood heterogeneity". We discuss this work and other related works in more detail in Appendix D

3. Setup

We consider optimizing the sum structured minimization objective distributed over n nodes or workers/clients:

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

where the functions $f_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} F_i(\mathbf{x}, \xi_i)$ denote the stochastic objectives locally stored on every node i. In machine learning applications, this corresponds to minimizing an empirical loss f averaged over all local losses f_i, with $\mathcal{D}_i$ being a distribution over the local dataset on node i. We define a communication graph $G = (V, E)$ with $|V| = n$ nodes. Following the convention in the decentralized literature [e.g. 56], we define a mixing matrix $W \in \mathbb{R}^{n\times n}$ as a weighted adjacency matrix of $\mathcal{G}$ with $w_{ij} \in [0, 1]$, $w_{ij} > 0$ only if $(i, j) \in E$ and W being doubly stochastic, i.e. $\sum_{i=1}^n w_{ij} = 1$.

In D-SGD, every worker $i \in [n]$ maintains local parameters $\mathbf{x}_i^{(t)} \in \mathbb{R}^d$ that are updated in each iteration with a stochastic gradient update (computed on the local function f_i) and by averaging with neighbors in the communication graph. It is convenient to compactly write the gradients in matrix notation:

$$X^{(t)} := \left[\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right] \in \mathbb{R}^{d\times n}\,,\ \partial F(X^{(t)}, \xi^{(t)}) \ := \left[\nabla F_1(\mathbf{x}_1^{(t)}, \xi_1^{(t)}), \ldots, \nabla F_n(\mathbf{x}_n^{(t)}, \xi_n^{(t)})\right]\,,$$

where $\xi^{(t)}$ are independent random variables such that $\mathbb{E}\left[\partial F(X^{(t)}, \xi^{(t)})\right] = \partial f(X^{(t)})$. Similarly, we denote the mixing step as multiplication with the mixing matrix W. This is illustrated through the following update rule for D-SGD,

$$X^{(t+1)} = (X^{(t)} - \eta_t G^{(t)})W^{(t)}\,,$$

where the mixing matrix $W^{(t)}$ is sampled from a distribution $\mathcal{W}^{(t)}$ and $G^{(t)} = \partial F(X^{(t)}, \xi^{(t)})$.

3.1. Standard Assumptions

We use the following standard assumptions on objective functions, similar to several existing works [22, 27, 48, 50]. Following standard notation we let L and μ denote the smoothness and strong-convexity constants of $F_i(\mathbf{x}, \xi)$ and f_i respectively (defined in Appendix A), wherever applicable.

Assumption 1 (Bounded Variance)**.** *We assume that there exists a constant σ such that $\forall \mathbf{x} \in \mathbb{R}^d$*

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{\xi_i\sim\mathcal{D}_i} \left\|\nabla F_i(\mathbf{x}, \xi_i) - \nabla f_i(\mathbf{x})\right\|_2^2 \le \sigma^2\,. \tag{2}$$

For the convex case it suffices to assume a bound on the stochasticity at the optimum $\mathbf{x}^\star := \arg\min f(\mathbf{x})$. We assume there exists a constant $\sigma_\star^2 \le \sigma^2$, such that

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{\xi_i\sim\mathcal{D}_i} \left\|\nabla F_i(\mathbf{x}_\star, \xi_i) - \nabla f_i(\mathbf{x}_\star)\right\|_2^2 \le \sigma_\star^2 \tag{3}$$

Assumption 2 (Consensus Factor)**.** *We assume that there exists a constant $p \in (0, 1]$ such that for all $t \ge 0$:*

$$\mathbb{E}_{W\sim\mathcal{W}^{(t)}} \left\|XW - \overline{X}\right\|_F^2 \le (1 - p)\left\|X - \overline{X}\right\|_F^2\,, \tag{4}$$

for all $X \in \mathbb{R}^{d\times n}$ and $\overline{X} := X\frac{\mathbf{1}\mathbf{1}^\top}{n}$.

The factor p measures averaging abilities of the mixing matrix W: in case of the fully connected graph $p = 1$, as we obtain a perfect average in one step; for the disconnected graph $p = 0$.

3.2. Gradient Mixing

Prior work introduced various notions to measure the *dissimilarity* between the local objective functions [22, 27, 32, 46]. We instead introduce a new assumption that quantifies the interaction between the data heterogeneity and graph topology.

Assumption 3 (Relative Heterogenity)**.** *We define ζ' as the smallest positive constant such that $\forall X \in \mathbb{R}^{d\times n}$, $\forall t \geq 0$:*

$$\mathbb{E}_{W\sim\mathcal{W}^{(t)}} \tfrac{1}{n}\left\|\partial f(X)W - \overline{\partial f}(X)\right\|^2 \leq \zeta'^2 \,. \tag{5}$$

For the convex case, it suffices to assume a bound at the optimum $X_\star$ only. We assume define $\zeta_\star$ as the smallest positive constant, such that $\forall t \geq 0$

$$\mathbb{E}_{W\sim\mathcal{W}^{(t)}} \tfrac{1}{n}\left\|\partial f(X_\star)W - \overline{\partial f}(X_\star)\right\|^2 \leq \zeta_\star'^2 \,. \tag{6}$$

The above quantity measures the effectiveness of a mixing matrix in producing close to the global average of the gradients at each node.

4. Convergence Result

In this section, we present a refined analysis of the D-SGD algorithm [22, 27]. We state our main convergence results below, whose proofs can be found in Appendix B. In all cases, we assume L-smoothness of $F_i(\mathbf{x}, \xi_i)$. These results are stated in terms of the mean of the parameters across nodes $\overline{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$.

Theorem 1. *Let Assumptions 1, 2 and 3 hold. Then there exists a constant stepsize, such that D-SGD needs the following number of iterations to achieve an ε error:*

Non-Convex: *It holds $\frac{1}{T+1}\sum_{t=0}^T \mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \varepsilon$ after $\mathcal{O}\left(\frac{\sigma^2}{n\varepsilon^2} + \frac{\zeta'+\sigma\sqrt{p}}{p\varepsilon^{3/2}} + \frac{1}{p\varepsilon}\right)\cdot LF_0$ iterations.*

If we in addition assume convexity,

Convex: *Under convexity ($\mu \geq 0$), the error $\frac{1}{(T+1)}\sum_{t=0}^T(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^\star) \leq \varepsilon$ after $\mathcal{O}\left(\frac{\sigma^2}{n\varepsilon^2} + \frac{\sqrt{L}(\zeta'+\sigma\sqrt{p})}{p\varepsilon^{3/2}} + \frac{L}{p\varepsilon}\right)\cdot R_0^2$ iterations, and if strong-convexity $\mu > 0$,*

Strongly-Convex: *then $\sum_{t=0}^T \frac{w_t}{\sum_{t=0}^T w_t}(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^\star) + \mu\mathbb{E}\|\overline{\mathbf{x}}^{(T+1)} - \mathbf{x}^\star\|^2 \leq \varepsilon$ for[1] $\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu n\varepsilon} + \frac{\sqrt{L}(\zeta_\star'+\sigma_\star\sqrt{p})}{\mu p\sqrt{\varepsilon}} + \frac{L}{\mu p}\log\frac{1}{\varepsilon}\right)$ iterations, where w_t denote appropriately chosen positive weights, $F_0 := f(\mathbf{x}_0) - f^\star$ for $f^\star = \min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ and $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|$ denote the initial errors.*

4.1. Discussion

We note that our convergence rates in Theorem 1 resemble the ones in [22] but the old heterogeneity ζ (or $\zeta_\star$) described in Section 3.2 is replaced with the new relative heterogeneity measure ζ' (or $\zeta_\star'$ correspondingly). As $\zeta' \leq \zeta$ ($\zeta_\star' \leq \zeta_\star$), the convergence rates given in Theorem 1 are always tighter than previous works. In Appendix C, we explain that ζ' can be significantly smaller than ζ.

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

5. Heterogeneity-Aware Mixing

In this section, we build upon our novel theoretical insights developed above to improve the performance of D-SGD in practice.

5.1. Motivation

Theorem 1 predicts that small values of the relative heterogeneity parameter ζ' lead to improved convergence. More specifically, progress in each iteration is determined by the current data-dependent *gradient mixing error* $\left\|\partial f(\overline{X}^{(t)})W^{(t)} - \overline{\partial f}(\overline{X}^{(t)})\right\|^2$ which is upper bounded by ζ' (as defined in Assumption 3). This quantity depends both on the current iterate $X^{(t)}$ but also on the chosen mixing weights $W^{(t)}$, thus suggesting to continually update the mixing matrix such that the gradient mixing error remains low, while the gradients evolve during training.

Thus, we can write the following time-varying optimization problem for the mixing weights W. For current parameters $X \in \mathbb{R}^{d\times n}$, $\overline{X} = X\frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (we drop the time index) distributed over n nodes, we aim to solve

$$\min_{W\in\mathcal{M}_w} \left\|\partial f(\overline{X})W - \overline{\partial f}(\overline{X})\right\|_F^2 \tag{GME-exact}$$

where $\mathcal{M}_w = \{W : \mathbf{1}W = \mathbf{1}\,, \mathbf{1}^\top W = \mathbf{1}^\top; 0 \le w_{ij} \le 1 \;\; \forall i,j\,, \;\; w_{ij} = 0 \;\; \forall (i,j) \notin E\}$ is the set of allowed mixing matrices. The objective function comes from the definition of ζ'^2 in Equation (14). The first two conditions ensure double stochasticity of W, while the last condition respects edge constraints of the communication graph $\mathcal{G}$. Note that unlike the matrix corresponding to the optimal spectral gap, the optimal matrix obtained above could be asymmetric. We call this optimization problem the exact Gradient Mixing Error (GME-exact).

5.2. Proposed Algorithm

We can equivalently reformulate (GME-exact) as to more efficiently solve when the dimension d of the gradient vectors is large, compared to the number of nodes n. Defining the gram matrix $\Gamma := \left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right)^\top \left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right)$, we obtain:

$$\min_{W\in\mathcal{M}_w} \operatorname{Tr}\left[W^\top \Gamma W\right]\,. \tag{GME-opt-Γ}$$

This is a quadratic program with linear constraints. The minimizer, i.e. resulting mixing matrix, of (GME-opt-Γ) is the same as for (GME-exact). However, as the problem formulation depends only on the gram matrix $\Gamma \in \mathbb{R}^{n\times n}$ it can be solved more efficiently [9].

Since the direct optimization of (GME-opt-Γ) is infeasible due to the inability to access $\overline{\mathbf{x}}$ and the full gradients, we propose to *approximately* solve an approximation of the above objective, only once every $H \ge 1$ step. We summarise the resulting algorithm in Algorithm 1, which calls equation (GME-opt-Γ) as a subproblem. To make our approach communication-efficien, we utilize sketching techniques and intermittent communication. Sketching allows the nodes to communicate low dimensional projections of the gradients/parameters instead of the full parameter vectors. This is illustrated in Algorithm 2 (CE-GME), and further explained in Appendix E, along with theoretical guarantees in Appendix E.3. We provide additional justification for the design choices in Appendix E.

Algorithm 1 HETEROGENEITY-AWARE DECENTRALIZED SGD (HA-SGD)

$X^{(0)}$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T, communication graph G, GME optimization period H, initial mixing matrix $W^{(0)}$. **for** t **in** $0 \dots T$ **do in parallel on all workers**

$G^{(t)} = \partial F(X^{(t)}, \xi^{(t)})$ ▷ stochastic gradients

if $t \mod H = 0$ **then**

  $W^{(t)} = \text{CE-GME}(G^{(t)})$

else

  $W^{(t)} = W^{(t-1)}$

end if

$X^{(t+1)} = (X^{(t)} - \eta_t G^{(t)})W^{(t)}$ ▷ update & mixing

end parallel for

Algorithm 2 CE-GME: Communication Efficient GME

matrix $G \in \mathbb{R}^{d \times n}$, distributed column-wise across n nodes, random seed s, dimension k W, mixing matrix minimizing the GME **in parallel** on n nodes **do**

sample $A \in \mathbb{R}^{k \times d}$, $a_{ij} \sim \mathcal{N}(0,1)$ ▷ use the same random seed s on every node.

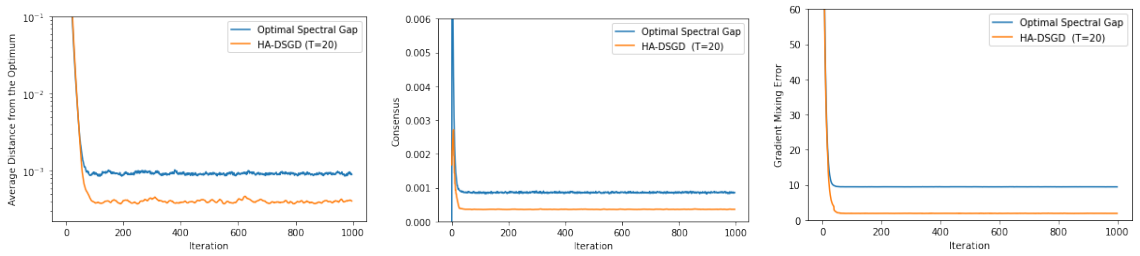
$S = \frac{1}{\sqrt{k}} AG \in \mathbb{R}^{k \times n}$ ▷ compute sketches

$\overline{S} = S\frac{\mathbf{1}\mathbf{1}^\top}{n}$ ▷ all-reduce-communication

$\Gamma = \left(S - \overline{S}\right)^\top \left(S - \overline{S}\right)$ ▷ sketched gram matrix

$W = \text{GME-opt}(\Gamma)$

end



(*a*) (*b*) (*c*)

Figure 1: Comparison of HA-DSGD to D-SGD. (a) Average distance from the optimum, (b) consensus distance $\frac{1}{n}\left\|X - \overline{X}\right\|_F^2$, and (c) gradient mixing error $\left\|\partial F(X,\xi)W - \overline{\partial F}(X,\xi)\right\|_F^2$ vs. the number of iterations for quadratic objectives. "Optimal Spectral Gap" denotes the DSGD algorithm with mixing matrix optimized using the SLEM objective defined in [8]. We report an average over a window of 5 iterations of corresponding quantity on each plot.

6. Experiments

For all our experiments, we use the CVXPY [12] convex optimization library to perform the constrained optimization defined in the section. We denote the number of nodes in the underlying topology and the time period for GME optimization by n and T respectively. HA-DSGD refers to our proposed Alg. 1 with updates alternating between the weights obtained by the GME optimization and the Metropolis-Hastings weights, similarly as discussed in Appendix E.

We consider a simple setting of random quadratic objectives, with the objective for the i_{th} client given by $f_i(\mathbf{x}) = \|A_i\mathbf{x} + b_i\|_2^2$ where $\mathbf{x}$ denotes a d dimensional parameter vector and both A_i and b_i contain entries sampled randomly from $\mathcal{N}(0,1)$ and fixed for each client. For our experiments, we set $d = 10$. We further introduce stochasticity to the gradients by adding random Gaussian

noise with variance 0.1. Figure 1 illustrates the improvements due to our approach across three metrics: the distance from the optimum, consensus error, as well as the gradient mixing error. In the Appendix J, we provide additional results for Deep Learning Benchmarks and quadratic objectives along with details of the experiments.

References

[1] Sulaiman A. Alghunaim and Ali H. Sayed. Linear convergence of primal–dual gradient methods and their performance in distributed optimization. *Automatica*, 117: 109003, 2020. doi: https://doi.org/10.1016/j.automatica.2020.109003. URL https://www.sciencedirect.com/science/article/pii/S0005109820302016.

[2] Sulaiman A. Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning, 2021.

[3] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 344–353. PMLR, 09–15 Jun 2019.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[5] B. Le Bars, A. Bellet, M. Tommasi, and AM. Kermarrec. Yes, topology matters in decentralized optimization: Refined convergence and topology learning under heterogeneous data, 2022. URL https://arxiv.org/abs/2204.04452.

[6] Aurélien Bellet, Anne-Marie Kermarrec, and Erick Lavoie. D-cliques: Compensating noniidness in decentralized federated learning with topology, 2021.

[7] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, 2013. URL https://hal.archives-ouvertes.fr/hal-00794821.

[8] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM REVIEW*, 46:667–689, 2003.

[9] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[10] Yatin Dandi, Luis Barba, and Martin Jaggi. Implicit gradient alignment in distributed and federated learning, 2021. URL https://arxiv.org/abs/2106.13897.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[12] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[13] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3): 592–606, 2012. doi: 10.1109/TAC.2011.2161027.

[14] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[15] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396, 2015.

[16] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning (ICML)*, pages 4387–4398. PMLR, 2020.

[17] Alexander Jung, Alfred O Hero III, Alexandru Cristian Mara, Saeed Jahromi, Ayelet Heimowitz, and Yonina C Eldar. Semi-supervised learning in network-structured data via total variation minimization. *IEEE Transactions on Signal Processing*, 67(24):6256–6269, 2019.

[18] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[19] Sai P. Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda T. Suresh. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. In *37th International Conference on Machine Learning (ICML)*. PMLR, 2020.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[21] Anastasia Koloskova, Tao Lin, Sebastian U. Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *International Conference on Learning Representations (ICLR)*, 2020.

[22] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5381–5393. PMLR, 13–18 Jul 2020.

[23] Anastasia Koloskova, Tao Lin, and Sebastian U. Stich. An improved analysis of gradient tracking for decentralized machine learning. In *NeurIPS*, 2021.

[24] Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian U. Stich. Consensus control for decentralized deep learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 5686–5696. PMLR, 2021. URL https://proceedings.mlr.press/v139/kong21a.html.

[25] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[26] Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication efficient decentralized training with multiple local updates. *arXiv preprint arXiv:1910.09126*, 2019.

[27] 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f75526659f31040afeb61cb7133e4e6d-Paper.pdf.

[28] Tao Lin, Sai Praneeth Karimireddy, Sebastian Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6654–6665. PMLR, 18–24 Jul 2021.

[29] Hanxiao Liu, Andy Brock, Karen Simonyan, and Quoc Le. Evolving normalization-activation layers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13539–13550. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9d4c03631b8b0c85ae08bf05eda37d0f-Paper.pdf.

[30] Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016. doi: 10.1109/TSIPN.2016.2524588.

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

[32] Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference*

on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 7111–7123. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/lu21a.html.

[33] Aritra Mitra, Rayana Jaafar, George Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34, 2021.

[34] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[35] A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[36] Angelia Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine, 37(3):92–101*, 2020.

[37] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27, 07 2016. doi: 10.1137/16M1084316.

[38] Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27, 07 2016. doi: 10.1137/16M1084316.

[39] Giovanni Neglia, Chuan Xu, Don Towsley, and Gianmarco Calbi. Decentralized gradient methods: does topology matter? In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2348–2358. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/neglia20a.html.

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

[41] Shaoqing Ren, Jian Sun, Kaiming He, and Xiangyu Zhang. Deep residual learning for image recognition. In *CVPR*, volume 2, page 4, 2016.

[42] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[43] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *ICML - Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3027–3036. PMLR, 2017. URL http://proceedings.mlr.press/v70/scaman17a.html.

[44] 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. doi: 10.1137/14096668X. URL https://doi.org/10.1137/14096668X.

[45] 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.

[46] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 4848–4856. PMLR, 2018. URL http://proceedings.mlr.press/v80/tang18a.html.

[47] John N. Tsitsiklis. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

[48] Thijs Vogels, Lie He, Anastasia Koloskova, Sai Praneeth Karimireddy, Tao Lin, Sebastian U Stich, and Martin Jaggi. Relaysum for decentralized deep learning on heterogeneous data. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Qo6kYy4SBI-.

[49] 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. URL http://arxiv.org/abs/1808.07576.

[50] Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. MATCHA: Speeding up decentralized SGD via matching decomposition sampling. *arXiv preprint arXiv:1905.09435*, 2019.

[51] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, 2020.

[52] Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

[53] E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. In *IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450, 2012.

[54] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer,

Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[55] Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/45713f6ff2041d3fdfae927b82488db8-Abstract.html.

[56] Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2004.02.022. URL https://www.sciencedirect.com/science/article/pii/S0167691104000398.

[57] Kun Yuan and Sulaiman A Alghunaim. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *arXiv preprint arXiv:2105.08023*, 2021.

[58] 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.

[59] Kun Yuan, Yiming Chen, Xinmeng Huang, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. DecentLaM: Decentralized momentum SGD for large-batch deep training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3029–3039, 2021.

[60] Jiong Zhang, Parameswaran Raman, Shihao Ji, Hsiang-Fu Yu, SVN Vishwanathan, and Inderjit Dhillon. Extreme stochastic variational inference: Distributed inference for large scale mixture models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 935–943. PMLR, 2019.

[61] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Appendix A. Preliminary Definitions

L-smoothness: Each local 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \operatorname{supp}(\mathcal{D}_i)$:

$$\|\nabla F_i(\mathbf{y}, \xi) - \nabla F_i(\mathbf{x}, \xi)\| \leq L \|\mathbf{x} - \mathbf{y}\| \ . \tag{7}$$

(μ-strong) 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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \leq \langle \nabla f_i(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle \ . \tag{8}$$

Appendix B. Proofs of Main Results

B.1. Preliminaries

We utilize the following set of standard useful inequalities:

Lemma 1. *Let g be an L-smooth convex function. Then we have:*

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\|_2^2 \leq 2L \left(g(\mathbf{x}) - g(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla g(\mathbf{y}) \rangle\right) , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{9}$$

Lemma 2. *Let $Y \in \mathbb{R}^{d \times n}$ be an arbitrary matrix and $\overline{Y}$ the matrix with each column containing the columnwise mean of Y i.e. $\overline{Y} = Y \frac{\mathbf{1}\mathbf{1}^\top}{n}$. Then we have:*

$$\left\|Y - \overline{Y}\right\|_F^2 = \|Y\|_F^2 - \left\|\overline{Y}\right\|_F^2 \leq \|Y\|_F^2 . \tag{10}$$

Lemma 3. *For arbitrary set of n vectors $\{\mathbf{a}_i\}_{i=1}^n$, $\mathbf{a}_i \in \mathbb{R}^d$*

$$\left\|\sum_{i=1}^n \mathbf{a}_i\right\|^2 \leq n \sum_{i=1}^n \|\mathbf{a}_i\|^2 \ . \tag{11}$$

Lemma 4. *For given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$*

$$\|\mathbf{a} + \mathbf{b}\|^2 \leq (1 + \alpha) \|\mathbf{a}\|^2 + (1 + \alpha^{-1}) \|\mathbf{b}\|^2 , \qquad \forall \alpha > 0 . \tag{12}$$

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

B.2. Recursion For Consensus

The recursion for consensus, analyzed in Lemmas 9 and 12 of [22] relies on the following inequalities:

$$n\Xi_t = \mathbb{E}\left\|X^{(t)} - \overline{X}^{(t)}\right\|_F^2 = \mathbb{E}\left\|X^{(t)} - \overline{X}^{(t-1)} - \left(\overline{X}^{(t)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \leq \mathbb{E}\left\|X^{(t)} - \overline{X}^{(t-1)}\right\|_F^2$$

The above inequality, however, discards the fact that it is desirable for the update at each node to be close to the update to the mean. Our analysis below instead incorporates the effect of the mixing of the gradient through the following lemma:

Lemma 5. *The update to $X^{(t-1)}$ at the t_{th} step of DSGD with mixing matrix $W^{(t-1)}$ can be reformulated as:*

$$X^{(t)}-\overline{X}^{(t)} = \Big(X^{(t-1)} - \overline{X}^{(t-1)}\Big)W^{(t-1)}-\eta_t\Big(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\Big)W^{(t-1)} \tag{13}$$

Proof

$$\begin{aligned}
&X^{(t)} - \overline{X}^{(t)} = (X^{(t-1)} - \eta_t \partial F(X^{(t-1)},\xi^{(t-1)}))(W^{(t-1)} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\\
&= X^{(t-1)}W^{(t-1)} - \overline{X}^{(t-1)} - \eta_t\Big(\partial F(X^{(t-1)},\xi^{(t-1)})W^{(t-1)} - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\Big)\\
&= \Big(X^{(t-1)} - \overline{X}^{(t-1)}\Big)W^{(t-1)} - \eta_t\Big(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\Big)W^{(t-1)}.
\end{aligned}$$

Where in the last step we used the identity $\frac{\mathbf{1}\mathbf{1}^\top}{n}W = \frac{\mathbf{1}\mathbf{1}^\top}{n}$, valid for any doubly stochastic matrix W, implying that $\overline{X}^{(t-1)}W^{(t-1)} = \overline{X}^{(t-1)}$ and $\overline{\partial F}(X^{(t-1)},\xi^{(t-1)})W^{(t-1)} = \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})$.

For the sake of generality and consistency with [22], we prove our result under a generalization of the Assumption 3 on the gradient mixing error

Assumption 4 (Relative Heterogenity with Growth)**.** *We assume that there exist constants ζ' and P', such that $\forall X \in \mathbb{R}^{d\times n}$:*

$$\mathbb{E}_{W\sim\mathcal{W}^{(t)}}\tfrac{1}{n}\left\|\partial f(\overline{X})W - \overline{\partial f}(\overline{X})\right\|^2 \leq \zeta'^2 + P'\left\|\overline{\partial f}(\overline{X})\right\|^2. \tag{14}$$

Assumption 3 corresponds to a special case of the above assumption with $P' = 0$.

We now prove the following consensus recursion:

Lemma 6. *Let $\Xi_t = \frac{1}{n}\mathbb{E}_t\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2$ denote the consensus distance at time t, and let $e_t = f(\overline{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)$ for the convex case and $e_t = \left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2$ for the non-convex case. Then:*

$$\Xi_t \leq \left(1 - \frac{p}{2}\right)\Xi_{t-1} + D\eta_{t-1}^2 e_{t-1} + A\eta_{t-1}^2. \tag{15}$$

Where $D = 36L(1-p) + 4L\frac{8-7p}{p}$ for the convex case , $\frac{8-7p}{p}P'$ for the nonconvex case and $A = \frac{8-7p}{p}(\zeta'^2) + 3(1-p)\sigma^2$ for the non-convex case and $\frac{16-14p}{p}(\zeta'^2) + 9(1-p)\sigma^2$ for the convex case.

Proof

$$\mathbb{E}_{W\sim\mathcal{W}}\tfrac{1}{n}\left\|\partial f(\overline{X})W - \overline{\partial f}(\overline{X})\right\|^2 \leq \zeta'^2 + P'\left\|\partial f(\overline{X})\right\|^2. \tag{16}$$

Let $\Xi_t = \frac{1}{n}\mathbb{E}_t\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2$ denote the consensus distance at time t. We have, using Lemma 5:

$$
\begin{aligned}
n\Xi_t &= \mathbb{E}\left\|X^{(t)} - \overline{X}^{(t)}\right\|_F^2 \\
&= \mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)W^{(t-1)} - \eta_t\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\right)W^{(t-1)}\right\|_F^2 \\
&= \underbrace{\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)W^{(t-1)} - \eta_t\left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)W^{(t-1)}\right\|_F^2}_{=:T_1} \\
&+ \underbrace{\eta_t^2\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\right)W^{(t-1)} - \left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)W^{(t-1)}\right\|_F^2}_{=:T_2}
\end{aligned}
$$

Where the last inequality follows from the fact that noise in the gradient is independent at each time step, and also from unbiased stochastic gradients $\mathbb{E}_{\xi^{(t-1)}}\partial F(X^{(t-1)},\xi^{(t-1)}) = \partial f(X^{(t-1)})$. We first observe that, using assumption 2, we have:

$$
\begin{aligned}
&\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\right)W^{(t-1)} - \left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)W^{(t-1)}\right\|_F^2 \\
&\leq (1-p)\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\right) - \left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)\right\|_F^2
\end{aligned}
$$

Furthermore, using equation (2), we have:

$$
\begin{aligned}
&\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\right) - \left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)\right\|_F^2 \\
&\leq \mathbb{E}\left\|\partial F(X^{(t-1)},\xi^{(t-1)}) - \partial f(X^{(t-1)})\right\|_F^2
\end{aligned}
$$

We then add and subtract the gradients at the mean point $\partial F(\overline{X}^{(t-1)},\xi^{(t-1)})$ and the corresponding mean $\partial F(\overline{X}^{(t-1)})$ to obtain:

$$
\begin{aligned}
&\mathbb{E}\left\|\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \partial f(X^{(t-1)})\right)\right\|_F^2 \\
&\overset{\text{Lemma 3}}{\leq} 3\left\|\partial F(X^{(t-1)},\xi^{(t-1)}) - \partial F(\overline{X}^{(t-1)},\xi^{(t-1)})\right\|^2 + 3\mathbb{E}\left\|\partial f(X^{(t-1)}) - \partial F(\overline{X}^{(t-1)}\right\|^2 \\
&+ 3\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)},\xi^{(t-1)}) - \partial F(\overline{X}^{(t-1)})\right\|_F^2
\end{aligned}
$$

Using the L smoothness of each node's objective, the first two terms can be bounded as follows:

$$
\mathbb{E}\left\|\partial F(X^{(t-1)},\xi^{(t-1)}) - \partial F(\overline{X}^{(t-1)},\xi^{(t-1)})\right\|^2 \leq L^2\mathbb{E}\left\|X^{(t-1)} - \overline{X}^{(t-1)}\right\|_F^2
$$

Similarly, we have:

$$\mathbb{E}\left\|\partial f(X^{(t-1)}) - \partial F(\overline{X}^{(t-1)})\right\|^2 \leq L^2\mathbb{E}\left\|X^{(t-1)} - \overline{X}^{(t-1)}\right\|_F^2. \tag{17}$$

Subsequently, we utilize the assumptions on stochasticity 1 to bound the third term.

We proceed separately for the Convex and non-convex cases:

Convex Case: We add and subtract $\partial F(X^\star, \xi^{(j)})$ and the corresponding mean $\partial F(X^\star)$ to obtain:

$$\begin{aligned}
&\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)}, \xi^{(t-1)}) - \partial F(\overline{X}^{(t-1)})\right\|_F^2 \\
&= \mathbb{E}\left\|\Big(\partial F(\overline{X}^{(t-1)}, \xi^{(t-1)}) - \partial F(X^\star, \xi^{(j)})\Big) - \Big(\partial F(\overline{X}^{(t-1)} - \partial F(X^\star)\Big) + \Big(\partial F(X^\star, \xi^{(j)}) - \partial F(X^\star))\Big)\right\|_F^2 \\
&\overset{\text{Lemma 3}}{\leq} 3\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)}, \xi^{(t-1)}) - \partial F(X^\star, \xi^{(j)})\right\|_F^2 + 3\left\|\partial F(\overline{X}^{(t-1)}) - \partial F(X^\star)\right\| \\
&+ 3\left\|\partial F(X^\star, \xi^{(j)}) - \partial F(X^\star))\right\|_F^2 \\
&\overset{\text{Lemma 1}}{\leq} 3 \cdot 2Ln(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 3 \cdot 2Ln(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 3n\overline{\sigma}^2 \\
&= 12Ln(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 3n\overline{\sigma}^2
\end{aligned} \tag{18}$$

Non-convex Case: We directly utilize the uniform bound on the stochasticity (assumption 1) to obtain:

$$\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)}, \xi^{(t-1)}) - \partial F(\overline{X}^{(t-1)})\right\|_F^2 \leq n\hat{\sigma}^2 \tag{19}$$

The final bound on T_2 is therefore given by:

Convex case:

$$\begin{aligned}
T_2 \leq{}& \eta_t^2 6(1-p)L^2\mathbb{E}\left\|\Big(X^{(t-1)} - \overline{X}^{(t-1)}\Big)\right\|_F^2 + 36(1-p)\eta_t^2 Ln(f(\mathbf{x}^{(t-1}) - f(\mathbf{x}^\star)) \\
&+ 9n(1-p)\eta_t^2\overline{\sigma}^2
\end{aligned}$$

Non-convex case:

$$T_2 \leq 6(1-p)\eta_t^2 L^2\mathbb{E}\left\|\Big(X^{(t-1)} - \overline{X}^{(t-1)}\Big)\right\|_F^2 + 3n(1-p)\eta_t^2\overline{\sigma}^2$$

We now bound T_1 as follows:

$$\begin{aligned}
&\mathbb{E}\left\|\left(X^{(t-1)}-\overline{X}^{(t-1)}\right)W^{(t-1)}-\eta_t\left(\partial f(X^{(t-1)})-\overline{\partial f}(X^{(t-1)})\right)W^{(t-1)}\right\|_F^2\\
&=\mathbb{E}\Bigg\|\left(X^{(t-1)}-\overline{X}^{(t-1)}\right)W^{(t-1)}-\eta_t\left(\partial F(\overline{X}^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\\
&-\eta_t\left(\left(\partial f(X^{(t-1)})-\partial F(\overline{X}^{(t-1)})\right)-\left(\overline{\partial f}(X^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)\right)W^{(t-1)}\Bigg\|_F^2\\
&\overset{\text{Lemma 4}}{\leq}(1+\beta_1)\mathbb{E}\Bigg\|\left(X^{(t-1)}-\overline{X}^{(t-1)}\right)W^{(t-1)}\\
&-\eta_t\left(\left(\partial f(X^{(t-1)})-\partial F(\overline{X}^{(t-1)})\right)-\left(\overline{\partial f}(X^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)\right)W^{(t-1)}\Bigg\|_F^2\\
&+(1+\beta_1^{-1})\mathbb{E}\left\|-\eta_t\left(\partial F(\overline{X}^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2\\
&\overset{\text{Lemma 4 ,}Assumption\ 2}{\leq}(1-p)(1+\beta_1)(1+\beta_2)\mathbb{E}\left\|\left(X^{(t-1)}-\overline{X}^{(t-1)}\right)\right\|_F^2\\
&+\eta_t^2(1-p)(1+\beta_1)(1+\beta_2^{-1})\mathbb{E}\Bigg\|\left(\partial f(X^{(t-1)})-\partial F(\overline{X}^{(t-1)})\right)\\
&-\left(\overline{\partial f}(X^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)\Bigg\|_F^2\\
&+(1+\beta_1^{-1})\mathbb{E}\left\|\eta_t\left(\partial F(\overline{X}^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2
\end{aligned}$$

The second term can be bounded by utilizing Equation 17 and Equation 2 as follows:

$$\begin{aligned}
&\mathbb{E}\Bigg\|\left(\partial f(X^{(t-1)})-\partial F(\overline{X}^{(t-1)})\right)-\left(\overline{\partial f}(X^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)\Bigg\|_F^2\\
&\overset{\text{Lemma 2}}{\leq}\mathbb{E}\left\|\left(\partial f(X^{(t-1)})-\partial F(\overline{X}^{(t-1)})\right)\right\|_F^2\\
&\overset{17}{\leq}L^2\left\|X^{(t-1)}-\overline{X}^{(t-1)}\right\|_F^2.
\end{aligned}$$

Therefore, we obtain:

$$\begin{aligned}
T_1\leq&\left((1-p)(1+\beta_1)(1+\beta_2)+\eta_t^2(1-p)(1+\beta_1)(1+\beta_2^{-1})L^2\right)\left\|X^{(t-1)}-\overline{X}^{(t-1)}\right\|_F^2\\
&+(1+\beta_1^{-1})\eta_t^2\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)})-\overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2
\end{aligned}$$

Finally, incorporating the bound on T_2, we obtain:

Convex Case:

$$
\begin{aligned}
n\Xi_t \leq & \left((1-p)(1+\beta_1)(1+\beta_2^{-1}) + \eta_t^2(1-p)(1+\beta_1)(1+\beta_2^{-1})\right)\left\|X^{(t-1)} - \overline{X}^{(t-1)}\right\|_F^2 \\
& + (1+\beta_1^{-1})\eta_t^2\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2 \\
& + 6(1-p)L^2\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
& + 36(1-p)\eta_t^2 Ln(f(\mathbf{x}) - f(\mathbf{x}^\star)) + 9n(1-p)\eta_t^2\overline{\sigma}^2
\end{aligned}
$$

Nonconvex Case:

$$
\begin{aligned}
n\Xi_t \leq & \left((1-p)(1+\beta_1)(1+\beta_2^{-1}) + \eta_t^2(1-p)(1+\beta_1)(1+\beta_2^{-1})\right)\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
& + (1+\beta_1^{-1})\eta_t^2\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2 \\
& + 6(1-p)L^2\eta_t^2\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
& + 3n(1-p)\eta_t^2\overline{\sigma}^2
\end{aligned}
$$

We now choose β_1 such that $(1-p)(1+\beta_1) = (1-\frac{7p}{8})$ i.e. $\beta_1 = \frac{p}{8(1-p)}$ Subsequently, we choose β_2 such that $((1-\frac{7p}{8})(1+\beta_2) = (1-\frac{3p}{4})$ i.e. $\beta_2 = \frac{p}{8-7p}$. Then, assuming that the step size η_t satisfies, $\eta_t^2 \leq \frac{\frac{p}{4}}{(1-p)(1+\beta_1)(1+\beta_2^{-1})L^2+6(1-p)L^2} = \frac{\frac{p}{4}}{\left((1-\frac{7p}{8})\frac{8-6p}{p}+6(1-p)\right)L^2}$, we obtain:

$$
\begin{aligned}
& n\Xi_t \leq (1-\frac{3p}{4})\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 + \frac{p}{4}\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
& + (1+\beta_1^{-1})\eta_t^2\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \overline{\partial F}(\overline{X}^{(t-1)}\right)W^{(t-1)}\right\|_F^2 + 6(1-p)L^2\eta_t^2\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
& + 3n(1-p)\eta_t^2\overline{\sigma}^2 \\
& \leq (1-\frac{p}{2})\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
& + \eta_t^2(1+\beta_1^{-1})\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2 + 3n(1-p)\eta_t^2\overline{\sigma}^2
\end{aligned}
$$

Since $\frac{\frac{p}{4}}{\left((1-\frac{7p}{8})\frac{8-6p}{p}+6(1-p)\right)L^2} \geq \frac{p^2}{80L^2}$, we only require the step size to be $\mathcal{O}(\frac{p}{L})$, same as [22]. Thus the consensus distance decreases linearly, along with an error dependent on the diffusion of the gradients across nodes. Finally, substituting the assumption 4 for the non-convex case, we obtain:

$$
\begin{aligned}
& n\Xi_t \leq (1-\frac{p}{2})\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 + \eta_t^2\frac{8-7p}{p}(\zeta'^2 + P'\left\|\overline{\partial F}(\overline{X})\right\|^2) \\
& = (1-\frac{p}{2})\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 + \eta_t^2(1+\beta_1^{-1})\zeta'^2 + \eta_t^2\frac{8-7p}{p}(1-p)P'\left\|\overline{\partial F}(\overline{X})\right\|^2 \\
& + 3n(1-p)\eta_t^2\sigma^2.
\end{aligned}
$$

For the convex case, we first bound the gradient mixing error at X in terms of that at X^* as follows:

$$
\begin{aligned}
&\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \overline{\partial F}(\overline{X}^{(t-1)})\right)W^{(t-1)}\right\|_F^2 \\
&= \mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \partial F(X^*) - (\overline{\partial F}(\overline{X}^{(t-1)}) - \overline{\partial F}(X^*))\right)W^{(t-1)} + \left(\partial F(X^*) - \overline{\partial F}(X^*)\right)W^{(t-1)}\right\|_F^2 \\
&\overset{\text{Lemma 3}}{\leq} 2\mathbb{E}\left\|\left(\partial F(\overline{X}^{(t-1)}) - \partial F(X^*) - (\overline{\partial F}(\overline{X}^{(t-1)}) - \overline{\partial F}(X^*))\right)W^{(t-1)}\right\|_2^2 \\
&+ 2\mathbb{E}\left\|\left(\partial F(X^*) - \overline{\partial F}(X^*)\right)W^{(t-1)}\right\|_F^2 \\
&\overset{\text{Assumption 2}}{\leq} 2(1-p)\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)}) - \partial F(X^*) - (\overline{\partial F}(\overline{X}^{(t-1)}) - \overline{\partial F}(X^*))\right\|_2^2 \\
&+ 2\mathbb{E}\left\|\left(\partial F(X^*) - \overline{\partial F}(X^*)\right)W^{(t-1)}\right\|_F^2 \\
&\overset{\text{Lemma 2}}{\leq} 2(1-p)\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)}) - \partial F(X^*)\right\|_2^2 + 2\mathbb{E}\left\|\left(\partial F(X^*) - \overline{\partial F}(X^*)\right)W^{(t-1)}\right\|_F^2 \\
&\leq 2(1-p)\mathbb{E}\left\|\partial F(\overline{X}^{(t-1)}) - \partial F(X^*)\right\|_2^2 + 2\mathbb{E}\left\|\left(\partial F(X^*) - \overline{\partial F}(X^*)\right)W^{(t-1)}\right\|_F^2 \\
&\leq 4(1-p)L\mathbb{E}\left(f(\overline{\mathbf{x}}^{(t-1)}) - f(\mathbf{x}^\star)\right) + 2\mathbb{E}\left\|\left(\partial F(X^*) - \overline{\partial F}(X^*)\right)W^{(t-1)}\right\|_F^2.
\end{aligned}
$$

Where in the last step, we used Equation 1. Therefore, we obtain:

$$
\begin{aligned}
n\Xi_t \leq\ & (1-\frac{p}{2})\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)\right\|_F^2 \\
&+ 4(1-p)\eta_t^2\frac{8-7p}{p}L\mathbb{E}\left(f(\overline{\mathbf{x}}^{(j)}) - f(\mathbf{x}^\star)\right) + 2\eta_t^2(1+\beta_1^{-1})n\overline{\zeta}^2 \\
&+ 36Ln(1-p)\eta_t^2(f(\mathbf{x}) - f(\mathbf{x}*)) + 9n(1-p)\eta_t^2\overline{\sigma}^2
\end{aligned}
$$

B.3. Convergence Rate

We utilize the consensus recursion in Lemma 6 to bound an appropriately weighted sum of the consensus iterates as follows:

$$
\sum_{t=0}^{T} w_t n\Xi_t \leq \sum_{t=1}^{T} w_t(1-\frac{p}{2})n\Xi_{t-1} + \sum_{t=1}^{T} w_t\eta_{t-1}^2 De_{t-1} + \sum_{t=1}^{T} w_t\eta_{t-1}^2 A
$$

Recursively substituting $n\Xi_{t-1}$ for t in $[1, \cdots, T]$, we then obtain:

$$\begin{aligned}
\sum_{t=0}^{T} w_t n\Xi_t &\leq \sum_{t=1}^{T}\sum_{j=0}^{t-1} w_t\eta_j^2(1-\frac{p}{2})^{t-j-1}(De_j + A) \\
&= \sum_{t=1}^{T}\sum_{j=0}^{t-1} w_t\eta_j^2(1-\frac{p}{2})^{t-j-1}(De_j + A) \\
&= \sum_{j=0}^{T-1}\sum_{t=j+1}^{T} \eta_j^2 w_t(1-\frac{p}{2})^{t-j-1}(De_j + A) \\
&\leq \sum_{j=0}^{T}\sum_{t=j+1}^{\infty} \eta_j^2 w_t(1-\frac{p}{2})^{t-j-1}(De_{t-1} + A) \\
&\leq \sum_{j=0}^{T} \eta_j^2 \frac{2}{p} w_j(De_j + A)\,.
\end{aligned}$$

Where in the last step we used $w_t \leq w_j$ for $j \geq t$
We thus obtain an Equation having the same form as Equation 18 of [22] :

$$B \cdot \sum_{t=0}^{T} w_t\Xi_t \;\leq\; \frac{b}{2} \cdot \sum_{t=0}^{T} w_t e_t + AB\frac{2}{p} \cdot \sum_{t=0}^{T} w_t\eta_t^2\,, \tag{20}$$

where η satisfies $\eta \leq \sqrt{\frac{pbD}{2B}}$ and the factor B is as defined in [21] for the different cases.

The rest of the proof involves utilizing the descent lemma in [22] and choosing the appropriate step size following exactly the use of Equation 18 in [22]. Finally, setting $P' = 0$ leads to the convergence rates provided in Theorem 1.

Appendix C. Comparision with Koloskova et al. [22]

In this section, we discuss how ζ' relates to ζ defined in Koloskova et al. [22].

Remark 1. *Using Assumption 4, we obtain:*

$$\begin{aligned}
\mathbb{E}_{W\sim\mathcal{W}^{(t)}}\tfrac{1}{n}\left\|\partial f(\overline{X})W - \overline{\partial f}(\overline{X})\right\|^2 &= \mathbb{E}_{W\sim\mathcal{W}^{(t)}}\tfrac{1}{n}\left\|(\partial f(\overline{X}) - \overline{\partial f}(\overline{X}))(W - \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top)\right\|^2 \\
&\leq \tfrac{1}{n}(1-p)\left\|\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right\|^2 \leq (1-p)\zeta^2\,.
\end{aligned}$$

This implies that $\zeta'^2 \leq (1-p)\zeta^2$ and $\zeta_\star'^2 \leq (1-p)\zeta_\star^2$. Often ζ' can even be much smaller

As a motivating example, we consider a ring topology with the Metropolis-Hasting mixing weights and a particular pattern on how the data is distributed across the nodes:

Example 1. *Consider a ring topology on $n = 3k$ nodes, $k \geq 1$, with uniform mixing among neighbors ($w_{i,i-1} = w_{i,i} = w_{i,i+1} = \frac{1}{3}$) and assume that $\mathcal{D}_i = \mathcal{D}_{i+3 \mod n}$ for all i and suppose there is an $\mathbf{x}'$ with $\nabla f(\mathbf{x}') = 0$, $\|\nabla f_1(\mathbf{x}')\| > 0$. Then $\zeta' = 0$ and $\zeta \neq 0$.*

It is easy to see that the relative heterogenity is $\zeta' = 0$. This holds, because uniform averaging of three neighboring gradients result in an unbiased gradient estimator:

$$\tfrac{1}{3}\sum_{j\in\{i-1,i,i+1\}} \nabla f_j(\mathbf{x}) = \nabla f(\mathbf{x})\,, \qquad \forall \mathbf{x} \in \mathbb{R}^d\,,$$

while in contrast

$$\zeta^2 \geq \tfrac{1}{3}\left(\left\|\nabla f_1(\mathbf{x}')\right\|^2 + \left\|\nabla f_2(\mathbf{x}')\right\|^2 + \left\|\nabla f_3(\mathbf{x}')\right\|^2\right) > 0.$$

In Appendix F, we discuss another example provided by [6], where mixing within interconnected cliques (i.e., locally fully connected sets of nodes) having global label distribution leads to $\zeta' = 0$.]

Appendix D. Additional Related Work

Several recent works have attempted to tackle the undesirable effects of data heterogeneity across nodes on the convergence of D-SGD through suitable modifications to the algorithm. D^2/Exact-diffusion [46, 57, 58] apply variance reduction on each node. Gradient Tracking [23, 30, 31, 37, 40] utilizes an estimate of the full gradient at each node, obtained by successive mixing of gradients along with corrections based on updates to the local gradients. However, these approaches have not been found to yield performances comparable to D-SGD in practice [28], despite superior theoretical properties [2, 23].

The undesirable effects of data heterogeneity persist also in the Federated Learning setting, which is a special case of the fully decentralized setting. Several algorithms have been designed to mitigate the undesirable effects of data heterogeneity [10, 19, 33, 51], yet extending them to the setup of decentralized learning remains challenging.

For optimizing convex functions, specialized variants such as EXTRA [45], decentralized primal-dual methods [1] have been developed. With a focus on deep learning applications, [28, 59] propose adaptations of momentum methods.

Bellet et al. [6] recently proposed utilizing a topology that minimizes the data-heterogeneity across cliques composed of clusters of nodes capturing the entire diversity of data distribution (D-Cliques). Our analysis does apply to their setting and can be used to theoretically explain the theoretical underpinnings behind D-Clique averaging (Appendix F)

Another line of work focuses on the design of (data-independent) mixing matrices with good spectral properties [56]. Another example is time-varying topologies such as the directed exponential graph [3] that allow for perfect mixing after multiple steps, or matchings [50]. Several theoretical works argue to perform multiple averaging steps between updates [24, 32, 43], though this introduces a noticeable overhead in practical DL applications. Vogels et al. [48] propose to replace gossip averaging with a new mechanism to spread information on embedded spanning trees.

Several works have analyzed the relationship between graph topology, data-heterogeneity, and computational efficiency in settings such as Network Lasso [15, 17], Clustered Federated Learning [42], and Extreme Variational Inference [60]. However, such an analysis for the optimization in decentralized SGD has been lacking.

Appendix E. Designing good mixing matrices

One of the main advantages of our theoretical analysis is that it allows a principled design of good mixing matrices. We identify in Theorem 1 two concurrent factors: on the one hand, the consensus

factor p should be close to 1, and on the other hand the relative heterogeneity parameter ζ' should be close to 0. Trying to find a mixing matrix satisfying both might seem a difficult task. However, one can combine matrices that are good for either of the tasks.

Example 2. *Suppose a mixing matrix W_p has consensus factor $p \leq 1$, and a mixing matrix $W_{\zeta'}$ has relative heterogeneity parameter ζ'. Then $W = W_{\zeta'}W_p$ has consensus factor at least p and relative heterogeneity at most ζ'.*

Proof By the mixing property of W_p,

$$\left\|XW - \overline{X}\right\|_F^2 \;=\; \left\|XW_{\zeta'}W_p - \overline{X}\right\|_F^2 \;\leq\; (1-p)\left\|XW_{\zeta'} - \overline{X}\right\|_F^2 \;\leq\; (1-p)\left\|X - \overline{X}\right\|_F^2,$$

and similarly, $\frac{1}{n}\left\|\partial f(\overline{X})W_{\zeta'}W_p - \overline{\partial f}(\overline{X})\right\|^2 \leq \frac{1}{n}\left\|\partial f(\overline{X})W_{\zeta'} - \overline{\partial f}(\overline{X})\right\|^2 \leq \zeta'^2$.

Where we used the double stochasticity of $W_{\zeta'}$ which implies that $\left\|(W - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\right\|_2 \leq 1$ (Proof in Proposition 5). In practice, we observe that two communication rounds are not necessary, alternating between mixing with W_p and $W_{\zeta'}$ works well and does not increase the communication costs.

E.1. Justifying the Design Choices

Next we analyze the relationship between Algorithm 2 and GME-exact.

Effect of Periodic Optimization. In Algorithm 1, we optimize the mixing matrix W only once every H steps in order to reduce the computational cost. Below we show that for small H, if at step $t+H$ we apply the matrix $W^{(t)}$ found by GME at the step t, then this matrix would still give a good error ζ'.

To isolate the effect of periodic optimization, we assume that every H steps we solve an original GME-exact problem. We perform optimization using the full gradients, moreover on line 4 of Algorithm 1 we solve an original (GME-exact) problem with full gradients on the averaged parameters, i.e. line 4 is replaced with $W^{(t)} = \mathrm{GME}(\partial f(\overline{X}^{(t)}))$.

Proposition 1. $\left\|\partial f\left(\overline{X}^{(t+H)}\right)W^{(t)} - \overline{\partial f}\left(\overline{X}^{(t+H)}\right)\right\|_F^2 \;\leq\; 2\left\|\partial f\left(\overline{X}^{(t)}\right)W^{(t)} - \overline{\partial f}\left(\overline{X}^{(t)}\right)\right\|_F^2$
$+ 2H\sum_{i=0}^{H-1}\eta_t^2 L^2\left\|\partial f(X^{(t+i)})\right\|_F^2$

For the proof refer to the appendix. Since the learning rate η_t is usually small, the relative heterogeniety does not increase much for a small number of steps H.

Effect of Stochastic Estimation. In practice the full gradients are too expensive to compute, so we will resort to stochastic gradients instead. The following proposition controls the error due to the selection of the mixing matrix using stochastic gradients.

Proposition 2. *Let $W^*(\xi)$ be any mixing matrix dependent on the noise parameters ξ satisfying the given edge constraints. Then, we have:*

$$\mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right)W^*(\xi)\right\|_F^2\right] \leq 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)W^*(\xi)\right\|_F^2\right] + 2n\sigma^2.$$

Proof can be found in the appendix. Setting $W^*(\xi) = \underset{W\in\mathcal{M}_w}{\arg\min}\left\|\partial f(\overline{X},\xi)W - \overline{\partial f}(\overline{X},\xi)\right\|_F^2$ reveals that minimizing GME with stochastic gradients would also lead to a small heterogeneity ζ up to additive stochastic noise.

Sketching for Gram Matrix Estimation. The original GME-exact formulation requires transmitting the entire gradients. We instead propose to calculate the Gram matrix using sketched gradients, for improved communication efficiency.

Let A denote a random matrix with Gaussian entries and let U be an arbitrary matrix. We observe that $\frac{1}{k}\mathbb{E}(UA)^\top UA = \frac{1}{k}\mathbb{E}U^\top A^\top AU = U^\top U$. Therefore, the above projection operation preserves the inner products in expectation. The approximation error of the above scheme can be bounded using standard concentration techniques. We provide precise theoretical guarantees for sketching and the use of local gradients in section **??** of the Appendix.

E.2. Optimizing mixing of updates of arbitrary algorithms

Our approach can be generalized to arbitrary additive updates to the parameters of the form $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \eta \mathbf{u}_i^{(t)}$. Concretely, let $U^{(t)}$ denote the matrix with its i_{th} column being $\mathbf{u}_i^{(t)}$. Then, the updates to X can be decoupled as follows:

$$X^{(t)} - \overline{X}^{(t)} = \left(X^{(t-1)} - \overline{X}^{(t-1)}\right)W^{(t-1)} - \eta_t\left(U^{(t)} - \overline{U^{(t)}}\right)W^{(t-1)}. \tag{21}$$

Therefore, the contribution to the deviation from the mean of the i_{th} node due to the mixing of the updates is again given by $\sum_{j=1}^n W_{ji}u_j - \overline{u}$.

For example, replacing the gradients in the Algorithm 1 by the updates of the Adam algorithm [20] results in the minimization of the mixing error involved in decentralized Adam updates. We empirically verify the effectiveness of such an algorithm for an NLP task as discussed in Section J.2. In Appendix I, we discuss other variations and extensions covered by our framework, such as directly optimizing the mixing of parameters.

E.3. Guarantees on Sketching and the use of Local Gradients

The bound on the error due to sketching is provided through the following extension of the Johnson–Lindenstrauss lemma:

Proposition 3. *Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\} \in \mathbb{R}^d$. Assume that the entries in $A \subset \mathbb{R}^{k\times d}$ are sampled independently from $\mathcal{N}(0,1)$. Then, for $k = \omega(\frac{log(\frac{m}{\delta})}{\varepsilon^2})$, with probability greater than $1-\delta$, we have:*

$$\left|\tfrac{1}{k}\left\langle A\mathbf{u}_i, A\mathbf{u}_j\right\rangle - \left\langle \mathbf{u}_i, \mathbf{u}_j\right\rangle\right| \leq \varepsilon \max_{i\in[m]} \|\mathbf{u}_i\|^2 \quad \textit{for all } \; i,j \in [m].$$

In our algorithm, the $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\} \in \mathbb{R}^d$ correspond to the gradients across nodes, and are compressed using a the random projection generated independently at each period, using shared seeds.

Use of local X. In our practical implementation we solve GME problem for gradients computed at the parameters X instead of $\overline{X}$ in GME-exact. We show that this leads to the minimization of the GME upto an additional term proportional to the consensus:

Proposition 4. $\left\|\partial f(\overline{X})W - \overline{\partial f}(\overline{X})\right\|_F^2 \;\leq\; 2\left\|\partial f(X)W - \overline{\partial f}(X)\right\|_F^2 + 2L^2\left\|X - \overline{X}\right\|_F^2$

Our analysis also provides an estimate of the decrease of consensus distance $\|X - \overline{X}\|_F^2$. Thus, the small right hand side ensures the small relative heterogeneity.

E.4. Proof of Proposition 1

We first note that

$$\overline{X}^{(t+H)} - \overline{X}^{(t)} = \sum_{i=0}^{H-1} -\eta_{t+i} \partial f(X^{(t+i)}) \tag{22}$$

We further have:

$$\begin{aligned}
&\left\| \left(\partial f\left(\overline{X}^{(t+H)}\right) - \overline{\partial f}\left(\overline{X}^{(t+H)}\right)\right) W^{(t)} \right\|_F^2 \overset{\text{Lemma 3}}{\leq} 2 \left\| \left(\partial f\left(\overline{X}^{(t)}\right) - \overline{\partial f}\left(\overline{X}^{(t)}\right)\right) W^{(t)} \right\|_F^2 \\
&+ 2 \left\| \left(\partial f\left(\overline{X}^{(t+H)}\right) - \overline{\partial f}\left(\overline{X}^{(t+H)}\right) - \left(\partial f\left(\overline{X}^{(t)}\right) - \overline{\partial f}\left(\overline{X}^{(t)}\right)\right)\right) W^{(t)} \right\|_F^2
\end{aligned}$$

Applying Lemma 2 to the second term in the RHS yields:

$$\begin{aligned}
\left\| \left(\partial f\left(\overline{X}^{(t+H)}\right) - \overline{\partial f}\left(\overline{X}^{(t+H)}\right)\right) W^{(t)} \right\|_F^2 &\overset{Lemma2}{\leq} 2 \left\| \left(\partial f\left(\overline{X}^{(t)}\right) - \overline{\partial f}\left(\overline{X}^{(t)}\right)\right) W^{(t)} \right\|_F^2 \\
&+ 2 \left\| \left(\partial f\left(\overline{X}^{(t+H)}\right) - \partial f\left(\overline{X}^{(t)}\right)\right) W^{(t)} \right\|_F^2
\end{aligned}$$

Finally, using Equation 22 and the L-smoothness of the objectives, we obtain:

$$\begin{aligned}
&\left\| \left(\partial f\left(\overline{X}^{(t+H)}\right) - \overline{\partial f}\left(\overline{X}^{(t+H)}\right)\right) W^{(t)} \right\|_F^2 \\
&\leq 2 \left\| \left(\partial f\left(\overline{X}^{(t)}\right) - \overline{\partial f}\left(\overline{X}^{(t)}\right)\right) W^{(t)} \right\|_F^2 + 2L^2 \left\| \overline{X}^{(t+H)} - \overline{X}^{(t)} \right\|^2 \\
&\leq 2 \left\| \left(\partial f\left(\overline{X}^{(t)}\right) - \overline{\partial f}\left(\overline{X}^{(t)}\right)\right) W^{(t)} \right\|_F^2 + 2H \sum_{i=0}^{H-1} \eta_t^2 L^2 \left\| \partial f(X^{(t+i)}) \right\|_F^2
\end{aligned}$$

E.5. Proof of Proposition 4

We start by adding and subtracting the corresponding gradients at the mean parameters X:

$$\begin{aligned}
&\left\| \left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right) W \right\|_F^2 \\
&= \left\| \left(\partial f(\overline{X}) - \partial f(X) - \left(\overline{\partial f}(\overline{X}) - \overline{\partial f}(X)\right)\right) W + \left(\partial f(X) - \overline{\partial f}(X)\right) W \right\|_F^2 \\
&\overset{\text{Lemma 3}}{\leq} \left\| \left(\partial f(\overline{X}) - \partial f(X) - \left(\overline{\partial f}(\overline{X}) - \overline{\partial f}(X)\right)\right) \right\|_F^2 + 2 \left\| \left(\partial f(X) - \overline{\partial f}(X)\right) W \right\|_F^2 \\
&\overset{\text{Lemma 2}}{\leq} 2 \left\| \left(\partial f(\overline{X}) - \partial f(X)\right) \right\|_F^2 + 2 \left\| \left(\partial f(X) - \overline{\partial f}(X)\right) W \right\|_F^2 \\
&\overset{\text{L-smoothness}}{\leq} L^2 \left\| X - \overline{X} \right\|_F^2 + 2 \left\| \left(\partial f(X) - \overline{\partial f}(X)\right) W \right\|_F^2 ,
\end{aligned}$$

E.6. Spectral Norm of Doubly Stochastic Matrices with Non-negative Entries

Proposition 5. *Let $W \in \mathbb{R}^{n \times n}$ be possibly asymmetric doubly stochastic matrix with non-negative entries. Then the spectral norm $\|W\|_2$ is bounded by* 1.

Proof We note that $W^T W$ is itself a symmetric doubly-stochastic matrix and therefore has an eigenvector $\frac{1}{\sqrt{n}}\mathbf{1}$ with eigenvalue 1. Perron-Frobenius theorem then implies that the largest eigenvalue of $(W^{(t)})^\top W^{(t)}$ is bounded by 1, completing the proof.

E.7. Proof of Proposition 2

The proof proceeds by introducing the stochastic gradients into the LHS as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right)W^*(\xi)\right\|_F^2\right] \\
&= \mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X}) - \left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right) + \left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)\right)W^*(\xi)\right\|_F^2\right] \\
&\overset{\text{Lemma 3}}{\leq} 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)W^*(\xi)\right\|_F^2\right] \\
&+ 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X}) - \left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)\right)W^*(\xi)\right\|_2^2\right] \\
&\overset{\text{Lemma 2}}{\leq} 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)W^*(\xi)\right\|_F^2\right] + 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \partial f(\overline{X},\xi)\right)W^*(\xi)\right\|_2^2\right].
\end{aligned}$$

Since $W^*(\xi)$ is doubly-stochastic, using Proposition 5, we obtain a bound on the spectral norm $\left\|W^{(*)}\right\|_2 \leq 1$. Combining the bound on the spectral norm with the assumption on the variance yields:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \overline{\partial f}(\overline{X})\right)W^*(\xi)\right\|_F^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)W^*(\xi)\right\|_F^2\right] + 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X}) - \partial f(\overline{X},\xi)\right)\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\left(\partial f(\overline{X},\xi) - \overline{\partial f}(\overline{X},\xi)\right)W^*(\xi)\right\|_F^2\right] + 2\overline{\sigma^2}
\end{aligned}$$

E.8. Proof of Proposition 3

We utilize the following compression bound, that arises as a consequence of the concentration of χ^2 random variables, as often utilized in the proof of the Johnson–Lindenstrauss lemma [7]:

Lemma 7. *Let $\{\mathbf{u}_1, \cdots, \mathbf{u}_m\} \in \mathbb{R}^d$. Assume that the entries in $A \subset \mathbb{R}^{k\times d}$ are sampled independently from $\mathcal{N}(0,1)$. Then, for $k \geq 100(\frac{log(\frac{m}{\delta})}{\varepsilon^2})$, with probability greater than $1-\delta$, we have,$\forall i, j \in [m]$:*

$$(1-\varepsilon)\left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2 \leq \frac{1}{k}\left\|A\mathbf{u}_i - A\mathbf{u}_j\right\|^2 \leq (1+\varepsilon)\left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2 \tag{23}$$

Slightly weaker bounds can be obtained in more general settings such as that of sub-Gaussian random variables but we restrict to the Gaussian case in the theory as well as implementations of our algorithm.

Now, adding $\{-\mathbf{u}_1, \cdots, -\mathbf{u}_m\}$ to the set of points and applying Lemma 7 yields, $\forall i, j \in [m]$:

$$(1-\varepsilon)\left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2 \leq \left\|A\mathbf{u}_i \pm A\mathbf{u}_j\right\|^2 \leq (1+\varepsilon)\left\|\mathbf{u}_i - \mathbf{u}_j\right\|^2 \tag{24}$$

Therefore, we bound the inner product as follows:

$$\begin{aligned}\frac{1}{k}\left\langle A\mathbf{u}_i, A\mathbf{u}_j\right\rangle &= \frac{1}{4k}\Big(\|A\mathbf{u}_i + A\mathbf{u}_j\|^2 - \|A\mathbf{u}_i - A\mathbf{u}_j\|^2\Big)\\ &\leq \frac{1}{4}\Big((1+\varepsilon)\|\mathbf{u}_i + \mathbf{u}_j\|^2 - (1-\varepsilon)\|\mathbf{u}_i - \mathbf{u}_j\|^2\Big)\\ &\leq \left\langle \mathbf{u}_i, \mathbf{u}_j\right\rangle + \frac{1}{2}\varepsilon\Big(\|\mathbf{u}_i + \mathbf{u}_j\|^2 + \|\mathbf{u}_i - \mathbf{u}_j\|^2\Big)\\ &\leq \left\langle \mathbf{u}_i, \mathbf{u}_j\right\rangle + \varepsilon \max_i \|\mathbf{u}_i\|^2\end{aligned}$$

Similarly, we obtain the lower bound:

$$\left\langle \mathbf{u}_i, \mathbf{u}_j\right\rangle - \varepsilon \max_i \|\mathbf{u}_i\|^2 \leq \frac{1}{k}\left\langle A\mathbf{u}_i, A\mathbf{u}_j\right\rangle$$

Appendix F. Using different matrices for Parameter and Gradient Mixing

An additional advantage of our analysis is that it decouples the effect of parameter and gradient mixing. This allows our analysis to be extended to the case of use of different mixing matrices W_p and W_g for mixing the parameters and gradients at each step respectively. Concretely, we consider the following algorithm:

Algorithm 3 DECENTRALIZED SGD WITH DECOUPLED MIXING

$X^{(0)}$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T, mixing matrix distributions $\mathcal{W}_p^{(t)}, \mathcal{W}_g^{(t)}$, $t \in [0, T]$ **for** t **in** $0 \dots T$ **do in parallel on all workers**
$G^{(t)} = \partial F(X^{(t)}, \xi^{(t)})$ ▷ stochastic gradients
$W_p^{(t)} \sim \mathcal{W}_p^{(t)}, W_g^{(t)} \sim \mathcal{W}_g^{(t)}$ ▷ sample mixing matrices
$X^{(t+1)} = X^{(t)} W_p^{(t)} - \eta_t G^{(t)} W_g^{(t)}$ ▷ update & mixing
end parallel for

We now show that the above algorithm leads to convergence rates having the same dependence on p and ζ' as Theorem 1 but with these parameters defined as above in terms of $\mathcal{W}_u^{(t)}$ and $\mathcal{W}_g^{(t)}$. For instance, for the Non-convex case, we obtain that $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \varepsilon$ after

$$\mathcal{O}\left(\frac{\sigma^2}{n\varepsilon^2} + \frac{\zeta' + \sigma\sqrt{p}}{p\varepsilon^{3/2}} + \frac{1}{p\varepsilon}\right)\cdot LF_0$$

iterations. Analogously, we obtain the corresponding convergence rates for the convex case with $\zeta'_\star$ defined at the optimum i.e. $\mathbb{E}_{W_g\sim\mathcal{W}_g^{(t)}}\frac{1}{n}\left\|\partial f(X_\star)W_g - \overline{\partial f}(X_\star)\right\|^2 \leq \zeta'^2$. Similar to Lemma 5, the update can then be expressed as

$$X^{(t)} - \overline{X}^{(t)} = \Big(X^{(t-1)} - \overline{X}^{(t-1)}\Big)W_p^{(t-1)} - \eta_t\Big(\partial F(X^{(t-1)}, \xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)}, \xi^{(t-1)})\Big)W_g \tag{25}$$

Subsequently, analogous to the proof of Theorem 1, we obtain the following decomposition of the consensus iterates:

$$n\Xi_t = \underbrace{\mathbb{E}\left\|\left(X^{(t-1)} - \overline{X}^{(t-1)}\right)W_u^{(t-1)} - \eta_t\left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)W_g^{(t-1)}\right\|_F^2}_{=:T_1}$$
$$+ \underbrace{\eta_t^2\mathbb{E}\left\|\left(\left(\partial F(X^{(t-1)},\xi^{(t-1)}) - \overline{\partial F}(X^{(t-1)},\xi^{(t-1)})\right) - \left(\partial f(X^{(t-1)}) - \overline{\partial f}(X^{(t-1)})\right)\right)W_g^{(t-1)}\right\|_F^2}_{=:T_2}$$

Now, for p and ζ' satisfying:

$$\mathbb{E}_{W_u\sim\mathcal{W}_u^{(t)}}\left\|XW_u - \overline{X}\right\|_F^2 \leq (1-p)\left\|X - \overline{X}\right\|_F^2, \tag{26}$$

and,

$$\mathbb{E}_{W_g\sim\mathcal{W}_g^{(t)}}\frac{1}{n}\left\|\partial f(X)W_g - \overline{\partial f}(X)\right\|^2 \leq \zeta'^2, \tag{27}$$

we obtain the analogous consensus recursion:

$$\Xi_t \leq \left(1 - \frac{p}{2}\right)\Xi_{t-1} + D\eta_{t-1}^2 e_{t-1} + A\eta_{t-1}^2\cdot, \tag{28}$$

where $D = 36L + 4L\frac{8-7p}{p}$ for the convex case , $\frac{8-7p}{p}P'$ for the nonconvex case and $A = \frac{8-7p}{p}(\zeta'^2) + 3\sigma^2$ for the non-convex case and $\frac{16-14p}{p}(\zeta'^2) + 9\sigma^2$ for the convex case.

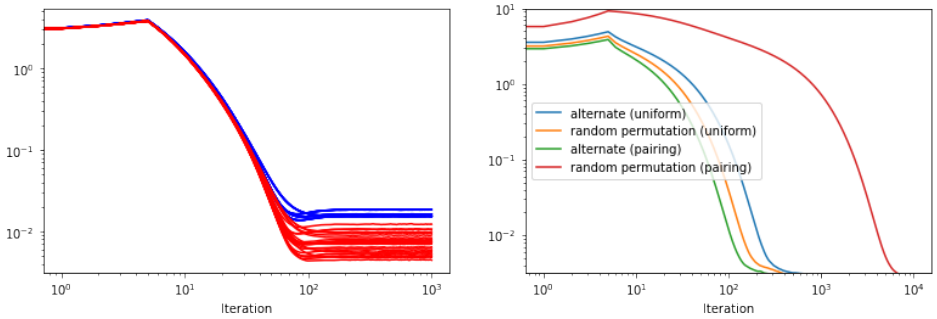
D-cliques [6]: Suppose that the graph can be divided into K cliques, such that the mean gradient for each clique equals the mean across the entire graph. Let the nodes be numbered such that the n_k nodes belonging to the k_{th} clique succeed the n_{k-1} nodes belonging to the $(k-1)_{th}$ clique. Then, we observe that utilizing a block matrix of the type

$$\begin{pmatrix} \frac{1}{n_1}\mathbf{1}\mathbf{1}^\top & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\mathbf{1}\mathbf{1}^\top & \cdots & \mathbf{0} \\ \vdots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \frac{1}{n_K}\mathbf{1}\mathbf{1}^\top \end{pmatrix}$$

leads to zero Gradient Mixing Error. This corresponds to the proposed algorithm in D-cliques [6] where $W_g^{(t)}$ is set to a matrix performing uniform averaging within each clique, while $W_u^{(t)}$ utilizes all the edges for mixing. For unbiased cliques, we obtain $\zeta' = 0$. Therefore, our analysis above provides an explanation for the improvements achieved by decoupled parameter mixing and clique-averaging [6] under the presence of unbiased cliques. We further note that, unlike the algorithm presented in [6], our algorithm HA-DSGD with random sampling of mixing matrices does not involve the additional communication overhead for separately mixing the gradients at each time-step.

Appendix G. Mixing Error under Permutations

We now demonstrate how the "Gradient Mixing Error" can be used to guide the choice of the arrangement of a given set of nodes over a graph. Given a set of nodes having fixed data distributions,



(*a*) (*b*)

Figure 2: Comparison of the distance from optimum vs number of iterations for different permutations of the nodes for (a) A random connected graph with 4 nodes (b) two-class ring topology setting with 16 nodes

the parameters controlling the Gradient Mixing Error (GME) is controlled by the choice of mixing weights as well as the graph topology. To illustrate the effects of the choice of topology on the convergence rates, we consider a toy setup of 4 nodes, having data distributions defined by quadratic objectives as in Section 6.

To further illustrate the benefits of selecting an appropriate permutation, we consider a setup of 16 nodes distributed on a ring topology with the data distributions of exactly half of the nodes belonging to each one of the following class of objectives:

$$f_1(\mathbf{x}) = \|A(\mathbf{x} - \mathbf{1})\|^2$$
$$f_2(\mathbf{x}) = \|A(\mathbf{x} + \mathbf{1})\|^2,$$

where A denotes a fixed matrix with entries from $\mathcal{N}(0, 1)$. We simulate the noise in SGD, by adding random vectors $\xi^{(t)} \sim \mathcal{N}(0, 0.001)$ to the gradient updates for each node We compare the performance of DSGD under the following two permutations and choices of the mixing matrices:

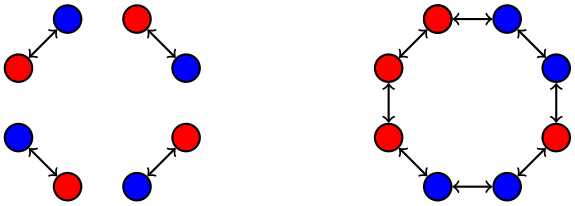
1. Heterogenous pairing: As illustrated in Figure 3, the nodes are ordered around the ring alternating between the data for objectives f_1 and f_2. Subsequently, every node is paired with exactly one of its neighbours such that the mixing steps involve averaging between the members of the pairs with equal weights of 0.5.

2. Random permutation: The nodes are randomly distributed on the ring with the mixing matrix corresponding to the maximal spectral gap.

We provide illustrations of the setup and the results in Figures 2 and 3 respectively, confirming the improvements in convergence due to the minimization of the GME.

Appendix H. Effect of Varying Data-Heterogeneity

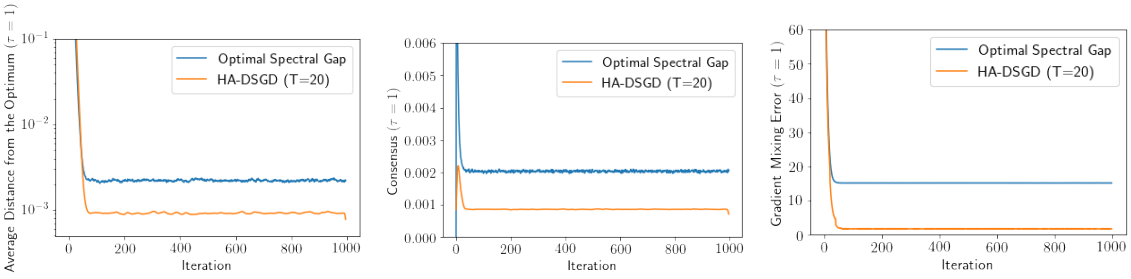
Since $\zeta' \leq \zeta$, reducing the heterogeneity of gradients across clients diminishes the role played by relative heterogeneity in the convergence rates. We verify this empirically on quadratic objectives under a similar setup as 6 with a random connected graph on 16 nodes having 60 edges.



(*a*) (*b*)

Figure 3: Different arrangements of data and mixing weights across a ring topology: (left) Heterogenous pairing between adjacent nodes having different data distributions, (right) Random permutation of nodes with uniform weights. The colors red and blue indicate two different classes of data distributions.

We consider local objectives of the form $f_i(\mathbf{x}) = \|((\mathbf{I} + A_i)\mathbf{x} - (\mathbf{1} + b_i)\|_2^2$ with $\mathbf{I}$ denoting the identity matrix, the dimension of the parameters $\mathbf{x}$ being $d = 10$ and A_i, b_i containing entries sampled randomly from $\mathcal{N}(0, \tau)$ and fixed for each client. The parameter τ therefore controls the level of heterogeneity acrosss clients. In Figures 4 and 5, we observe that reducing the level of heterogeneity across clients limits the improvements achieved by HA-DSGD in the distance to the optimum, while the consensus error and the GME are still lower than those for DSGD.



(*a*) (*b*) (*c*)

Figure 4: Comparison of HA-DSGD to D-SGD. (a) Average distance from the optimum, (b) consensus distance $\frac{1}{n}\left\|X - \overline{X}\right\|_F^2$, and (c) gradient mixing error $\left\|\partial F(X,\xi)W - \overline{\partial F}(X,\xi)\right\|_F^2$ vs. the number of iterations for quadratic objectives with heterogeneity parameter $\tau = 1$ We report an average over a window of 5 iterations of corresponding quantity on each plot.

Appendix I. Possible Extensions

Theorem 1 does not cover the just discussed case of alternating between two or more matrices. As our main focus in this work is on highlighting the benefits of relative heterogeneity, we just covered a simple case of time-varying mixing in the theorem (when all matrices are sampled from the same distribution). However, it is possible to extend our analysis to deterministic sequences (such as alternating) with the derandomization technique presented in [22, Assumption 4, Theorem 2]. Our

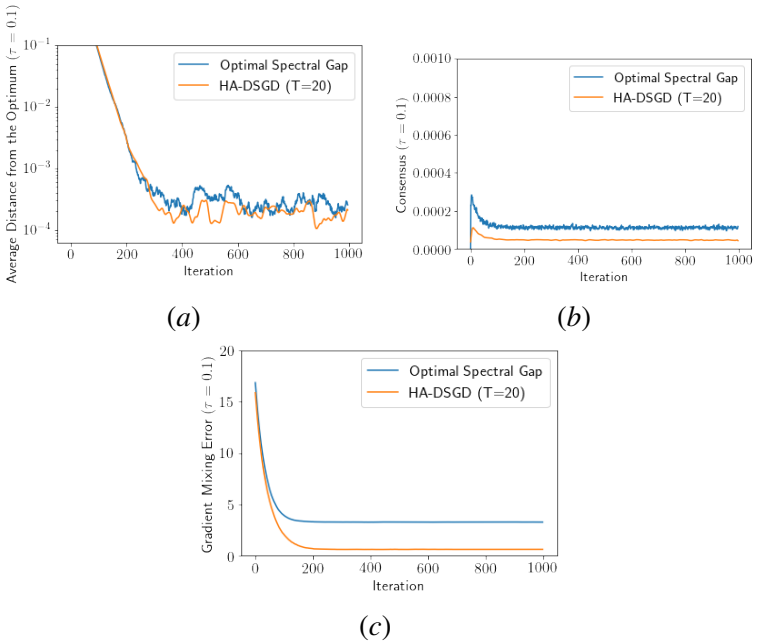


Figure 5: Comparison of HA-DSGD to D-SGD. (a) Average distance from the optimum, (b) consensus distance $\frac{1}{n}\left\|X-\overline{X}\right\|_F^2$, and (c) gradient mixing error $\left\|\partial F(X,\xi)W-\overline{\partial F}(X,\xi)\right\|_F^2$ vs. the number of iterations for quadratic objectives with heterogeneity parameter $\tau = 0.1$ We report an average over a window of 5 iterations of corresponding quantity on each plot.

analysis can also be extended to the case of optimizing parameter averaging and the use of two separate mixing matrices to mix parameters and gradients respectively (similar as in Bellet et al. 6) as discussed in sections I.2 and F respectively.

I.1. Optimizing mixing of updates of arbitrary algorithms

Our approach can be generalized to arbitrary additive updates to the parameters of the form $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \eta \mathbf{u}_i^{(t)}$. Concretely, let $U^{(t)}$ denote the matrix with its i_{th} column being $\mathbf{u}_i^{(t)}$. Then, the updates to X can be decoupled as follows:

$$X^{(t)} - \overline{X}^{(t)} = \left(X^{(t-1)} - \overline{X}^{(t-1)}\right)W^{(t-1)} - \eta_t\left(U^{(t)} - \overline{U^{(t)}}\right)W^{(t-1)}. \tag{29}$$

Therefore, the contribution to the deviation from the mean of the i_{th} node due to the mixing of the updates is again given by $\sum_{j=1}^{n} W_{ji}u_j - \overline{u}$.

For example, replacing the gradients in the Algorithm 1 by the updates of the Adam algorithm [20] results in the minimization of the mixing error involved in decentralized Adam updates. We empirically verify the effectiveness of such an algorithm for an NLP task as discussed in Section J.2. In Appendix I, we discuss other variations and extensions covered by our framework, such as directly optimizing the mixing of parameters.

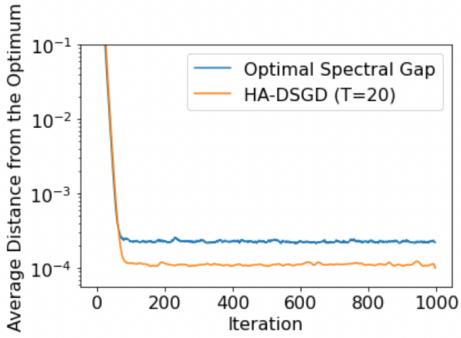


Figure 6: Distance from the optimum vs number of iterations for quadratic objectives on 64 nodes

I.2. Optimizing Mixing of Parameters

An alternate way of simultaneously maximizing the consensus factor p and the gradient mixing error is to directly optimize the mixing error of the parameters i.e. $\left\|(X^{(t)} - \overline{X}^{(t)})W\right\|_F^2$. Our theoretical analysis covers such a choice of mixing matrices as a special case that involves trying to obtain a mixing matrix having both small $(1-p)$ and the gradient mixing error. However, unlike the gradient mixing error that involves changes of the order η^2 as shown by Lemma 1, the distribution of the parameters across nodes can change rapidly due to the mixing. Moreover, we found both approaches to yield similar improvements in practice and focus on the gradient mixing error since it covers a wider range of design choices such as mixing within unbiased cliques.

Appendix J. Additional Experiments and Details

J.1. Quadratic Objectives

Details of the sampling procedure for graphs: We generate a random connected graph of 16 nodes through the following procedure: starting from a fully connected graph having 16 nodes and 120 edges, we repeatedly construct graphs by deleted 60 of the edges chosen uniformly at random until we obtain a connected graph having 60 edges.

Experiments on 64 nodes: Similar to Section 6, we generate a random graph containing 64 nodes with half of the edges randomly removed from a complete graph. Figure 6 shows the results in this setting for quadratic objectives, confirming the effectiveness of HADSGD on large graphs.

J.2. Deep Learning Benchmarks

In all our results, the period denotes the number of updates after which the mixing matrix is re-computed i.e. a period of 100 implies that the communication of the compressed gradients and the computation of the mixing matrix occurs only for a $\frac{1}{100}$ fraction of the updates. Furthermore, in Appendix J.4, we empirically verify that a sketching dimension of $100-1000$ is sufficient for models having millions of parameters. These two factors make our overhead in communication cost negligible compared to the baselines. In the Appendix J.3, we discuss utilizing the periodic global communication to perform averaging using the decoded sketched gradients.

CIFAR10. We evaluate our approach on the CIFAR10 dataset [25] by training the Resnet20 model [41] with Evonorm [29] for 300 epochs for each model. Following Sec. I.1, we consider the extension of our algorithm to the mixing of Nesterov momentum updates, denoted by HA-DSGD (momentum) in Table 1, and compare against the corresponding version of DSGD with momentum. We also compare against the D^2 algorithm [46] and Gradient Tracking [23, 30, 31, 37, 40] for completeness. The results show that our approach generally outperforms the baselines across three topologies, ring ($n = 16$), torus ($n = 16$), as well as the topology defined by the Davis Southern Women dataset as available in the `Networkx` library [14]. Since both the Metropolis-Hastings and the optimal spectral gap mixing schemes lead to similar results, we only compare against the Metropolis-Hastings schemes in the subsequent tasks.

Method	Ring (n=16)	Torus (n=16)	Social Network (n=32)
DSGD	74.71 ± 2.24	76.13 ± 1.65	77.68 ± 1.42
HA-DSGD	78.21 ± 2.19	79.08 ± 2.07	79.54 ± 1.61
HA-DSGD (mn, period=100)	80.75 ± 1.84	82.22 ± 1.87	83.24 ± 1.15
DSGD (mn, Metropolis-Hastings)	77.52 ± 2.78	80.45 ± 2.27	80.71 ± 1.93
DSGD (mn, Optimal Spectral Gap)	79.06 ± 1.82	80.28 ± 2.12	80.91 ± 1.74
DSGD (mn, Gradient Tracking)	78.42 ± 2.71	78.76 ± 2.43	80.14 ± 2.59
D^2	49.68 ± 3.19	51.37 ± 2.68	52.15 ± 2.43

Table 1: Top-1 test accuracy on CIFAR10 under different topologies. The results in the table are averaged over three random seeds.

Method	Ring (n=16)
HA-DSGD(mn, period=1000)	55.14 ± 0.215
DSGD (mn)	53.22 ± 0.25

Table 2: Top-1 Test accuracy on the Imagenet dataset, The results in the table are averaged over three random seeds.

Method	Ring (n=16)	Torus (n=16)
DAdam	87.14 ± 0.71	87.42 ± 0.65
HA-DAdam	89.29 ± 0.48	89.73 ± 0.54

Table 3: Top-1 test accuracy on the AG-News dataset under different topologies. The results in the table are averaged over three random seeds.

Transformer on AG News. We evaluate the extension of our algorithm to the mixing of Adam [20] updates on the NLP task of fine-tuning the `distilbert-base-uncased` model [54] on the AGNews dataset [61]. Table 3 verifies the applicability of our approach to Adam updates.

Imagenet. To evaluate our approach on a large-scale dataset, we consider the task of training a Resnet18 model [41] with evonorm on the Imagenet dataset [11]. We use a larger period of 1000 for the optimization of the mixing matrix to account for the larger number of steps per epoch. We train each model using Nesterov momentum for 90 epochs using a ring topology defined on 16 nodes.

Compression Dimension	Test Accuracy
1	75.66
100	81.55
1000	81.97

Table 4: Effect of the Compression Dimension on the top-1 test accuracy on the CIFAR dataset.

Similar to other settings, our approach as shown in Table 2 outperforms DSGD, demonstrating its effectiveness under large period and dataset sizes.

J.3. Effect of Periodic Averaging

While the use of sketching significantly reduces the communication cost, our algorithm is still not fully decentralized due to the requirement of periodically communicating the sketched gradients to a central server or node. When the full gradient is utilized, or a suitable compression scheme is used, we can further utilize the central communication step to perform an averaging of the parameters.

J.4. Effect of the Compression Dimension

Proposition 3 predicts that a low approximation error in the entries of the Gram matrix can be achieved through compression with dimension independent of the number of parameters and logarithmic in the number of nodes. We empirically verify this for the CIFAR10 dataset using HA-DSGD with Nesterov momemtum and a period of 100 in table 4

Table 5: Experimental settings for Cifar-10

Dataset	Cifar-10
Data augmentation	random horizontal flip and random 32×32 cropping
Architecture	Resnet20 with evonorm
Training objective	cross entropy
Evaluation objective	top-1 accuracy
Number of workers	16, 32
Topology	Ring, Torus, Social Network
Data distribution	Heterogeneous, not shuffled, according to Dirichlet sampling procedure from [28]
Batch size	32 patches per worker
Momentum	0.9 (Nesterov)
Learning rate	0.1 for $\alpha = 0.1$
LR decay	$/10$ at epoch 150 and 180
LR warmup	Step-wise linearly within 5 epochs, starting from 0
# Epochs	300
Weight decay	10^{-4}
Normalization scheme	no normalization layer
Repetitions	3, with varying seeds

Table 6: Experimental settings for finetuning distilBERT

Dataset	AG News
Data augmentation	none
Architecture	DistilBERT
Training objective	cross entropy
Evaluation objective	top-1 accuracy
Number of workers	16
Topology	ring
Data distribution	Heterogeneous, not shuffled, according to Dirichlet sampling procedure from [28]
Batch size	32 patches per worker
Adam β_1	0.9
Adam β_2	0.999
Adam ε	10^{-8}
Learning rate	1e-6
LR decay	constant learning rate
LR warmup	no warmup
# Epochs	10
Weight decay	0
Normalization layer	LayerNorm [4],
Repetitions	3, with varying seeds

Table 7: Experimental settings for ImageNet

Dataset	ImageNet
Data augmentation	random resized crop (224×224), random horizontal flip
Architecture	ResNet-20-EvoNorm [29, 41]
Training objective	cross entropy
Evaluation objective	top-1 accuracy
Number of workers	16
Topology	Ring
Gossip weights	Metropolis-Hastings (1/3 for ring)
Data distribution	Heterogeneous, not shuffled, according to Dirichlet sampling procedure from [28]
Batch size	32 patches per worker
Momentum	0.9 (Nesterov)
Learning rate	$0.1 \times \frac{32*16}{256}$)
LR decay	$/10$ at epoch $30, 60, 80$
LR warmup	Step-wise linearly within 5 epochs, starting from 0.1
# Epochs	90
Weight decay	10^{-4}
Normalization layer	EvoNorm [29]

Appendix K. Limitations

Like other works in the optimization literature, our convergence analysis does not directly explain generalization. However, we empirically validate improvements in generalization performance on several deep learning benchmarks. Our work also assumes a fixed topology and incorporating time-varying and adapting topology is a promising direction for future work.

Appendix L. Societal Impact

The demands for preserving privacy in machine learning training systems have been constantly growing over the past few years [18, 36]. We believe that Decentralized learning can play a major role in meeting such demands. This can improve the trust towards Machine Learning applications as well as maintenance of data-ownership in society.

Additionally, improvements in efficiency of decentralized optimization algorithms can reduce the environmental impact of training large machine learning models. We believe that the focus of our work towards the setting of heterogeneous data makes it especially relevant for practical settings.

Appendix M. Conclusion and Future Work

In this work, we extended the analysis of DSGD to incorporate the interaction between the mixing matrix and the data heterogeneity, leading to a novel technique for dynamically adapting the mixing matrix throughout training. Future work could involve extending our technique to algorithms designed for specific settings such as EXTRA [44] for convex non-stochastic cases, as well as approaches based on row-stochastic, column-stochastic matrices and time-varying topologies. On the theoretical side, promising directions include extending our analysis to the mixing of momentum.