Beyond spectral gap (extended): The role of the topology in decentralized learning

Thijs Vogels, Hadrien Hendrikx, Martin Jaggi

Beyond spectral gap (extended): The role of the topology in decentralized learning

Thijs Vogels* THIJS.VOGELS@EPFL.CH
Hadrien Hendrikx* HADRIEN.HENDRIKX@EPFL.CH
Martin Jaggi MARTIN.JAGGI@EPFL.CH

Machine Learning and Optimization Laboratory
EPFL
Lausanne, Switzerland

Abstract

In data-parallel optimization of machine learning models, workers collaborate to improve their estimates of the model: more accurate gradients allow them to use larger learning rates and optimize faster. In the decentralized setting, in which workers communicate over a sparse graph, current theory fails to capture important aspects of real-world behavior. First, the 'spectral gap' of the communication graph is not predictive of its empirical performance in (deep) learning. Second, current theory does not explain that collaboration enables *larger* learning rates than training alone. In fact, it prescribes *smaller* learning rates, which further decrease as graphs become larger, failing to explain convergence dynamics in infinite graphs. This paper aims to paint an accurate picture of sparsely-connected distributed optimization. We quantify how the graph topology influences convergence in a quadratic toy problem and provide theoretical results for general smooth and (strongly) convex objectives. Our theory matches empirical observations in deep learning, and accurately describes the relative merits of different graph topologies. This paper is an extension of the conference paper by Vogels et al. (2022). Code: github.com/epfml/topology-in-decentralized-learning.

Keywords: Decentralized Learning, Convex Optimization, Stochastic Gradient Descent, Gossip Algorithms, Spectral Gap

1. Introduction

Distributed data-parallel optimization algorithms help us tackle the increasing complexity of machine learning models and of the data on which they are trained. We can classify those training algorithms as either *centralized* or *decentralized*, and we often consider those settings to have different benefits over training 'alone'. In the *centralized* setting, workers compute gradients on independent mini-batches of data, and they average those gradients between all workers. The resulting lower variance in the updates enables larger learning rates and faster training. In the *decentralized* setting, workers average their models with only a sparse set of 'neighbors' in a graph instead of all-to-all, and they may have private datasets sampled from different distributions. As the benefit of decentralized learning, we usually focus only on the (indirect) access to other worker's datasets, and not of faster training.

Homogeneous (i.i.d.) setting. While decentralized learning is typically studied with heterogeneous datasets across workers, sparse (decentralized) averaging between them is also useful when worker's data is identically distributed (i.i.d.) (Lu and Sa, 2021). As an example,

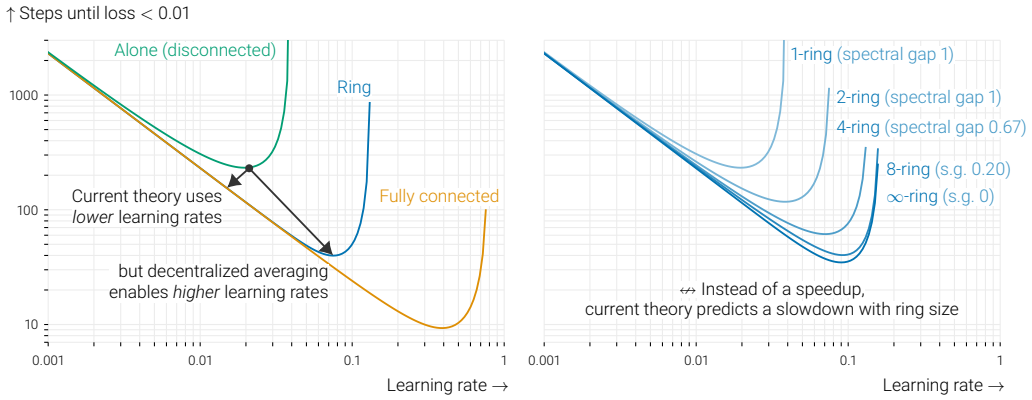

Figure 1: 'Time to target' for D-SGD (Lian et al., 2017) with constant learning rates on an i.i.d. isotropic quadratic dataset (Section 3.1). The noise disappears at the optimum. Compared to optimizing alone, 32 workers in a ring (*left*) are faster for any learning rate, but the largest improvement comes from being able to use a large learning rate. This benefit is not captured by current theory, which prescribes a smaller learning rate than training alone. On the *right*, we see that rings of increasing size enable larger learning rates and faster optimization. Because a ring's spectral gap goes to zero with the size of the ring, this cannot be explained by current theory.

sparse averaging is used in data centers to mitigate communication bottlenecks (Assran et al., 2019). When the data is i.i.d. (or heterogeneity is mild), the goal of sparse averaging is to optimize faster, just like in centralized (all-to-all) graphs. Yet, current decentralized learning theory poorly explains this speed-up. Analyses typically show that, for *small enough* learning rates, training with sparse averaging behaves the same as with all-to-all averaging (Lian et al., 2017; Koloskova et al., 2020) and so it reduces the gradient variance by the number of workers compared to training alone with the *same small learning rate*. In practice, however, such small learning rates would never be used. In fact, a reduction in variance should allow us to use a *larger* learning rate than training alone, rather than imposing a *smaller* one. Contrary to current theory, we show that (sparse) averaging lowers variance throughout all phases of training (both initially and asymptotically), allowing to take higher learning rates, which directly speeds up convergence. We characterize how much averaging with various communication graphs reduces the variance, and show that centralized performance (variance divided by the number of workers) is not always achieved when using optimal large learning rates. The behavior we explain is illustrated in Figure 1.

Heterogeneous (non-i.i.d.) setting. In standard analyses, heterogeneity affects convergence in a very worst-case manner. Standard guarantees intuitively correspond to the pessimistic case in which the most distant workers have the most different functions. These guarantees are typically loose in the settings where workers have different finite datasets sampled i.i.d. from the same distribution, or if each worker has a lot of diversity in its close neighbors. In this work, we characterize the impact of heterogeneity together with the communication graph,

enabling non-trivial guarantees even for infinite graphs under non-adversarial heterogeneity patterns.

Spectral gap. In both the homogeneous and heterogeneous settings, the graph topology appears in current convergence rates through the *spectral gap* of its averaging (gossip) matrix. The spectral gap poses a conservative lower bound on how much one averaging step brings all worker's models closer together. The larger, the better. If the spectral gap is small, a significantly smaller learning rate is required to make the algorithm behave close to SGD with all-to-all averaging with the same learning rate. Unfortunately, we experimentally observe that, both in deep learning and in convex optimization, the spectral gap of the communication graph is *not predictive* of its performance under tuned learning rates.

The problem with the spectral gap quantity is clearly illustrated in a simple example. Let the communication graph be a ring of varying size. As the size of the ring increases to infinity, its spectral gap goes to zero since it becomes harder and harder to achieve consensus between all the workers. This leads to the optimization progress predicted by current theory to go to zero as well. In some cases, when the worker's objectives are adversarially heterogeneous in a way that requires workers to obtain information from all others, this is indeed what happens. In typical cases, however, this view is overly pessimistic. In particular, this view does not match the empirical behavior with i.i.d. data. With i.i.d. data, as the size of the ring increases, the convergence rate actually *improves* (Figure 1), until it saturates at a point that depends on the problem.

In this work, we aim to accurately describe the behavior of distributed learning algorithms with sparse averaging, both in theory and in practice. We aim to do so both in the high learning rate regime, which was previously studied in the conference version of this paper Vogels et al. (2022), as well as in the small learning rate regime, in which we characterize the interplay between topology and data heterogeneity, as well as stochastic noise.

- We quantify the role of the graph in a quadratic toy problem designed to mimic the initial phase of deep learning (Section 3.1), showing that averaging enables a larger learning rate.

- From these insights, we derive a problem-independent notion of 'effective number of neighbors' in a graph that is consistent with time-varying topologies and infinite graphs, and is predictive of a graph's empirical performance in both convex and deep learning.

- We provide convergence proofs for (strongly) convex objectives that do not depend on the spectral gap of the graph (Section 4), and consider finer spectral quantities instead. Our rates disentangle the homogeneous and heterogeneous settings, and highlight that all problems behave as if they were homogeneous when the iterates are far from the optimum.

At its core, our analysis does not enforce global consensus, but only between workers that are close to each other in the graph. Our theory shows that sparse averaging provably enables larger learning rates and thus speeds up optimization. These insights prove to be relevant in deep learning, where we accurately describe the performance of a variety of topologies, while their spectral gap does not (Section 5).

2. Related work

Decentralized SGD. This paper studies decentralized SGD. Koloskova et al. (2020) obtain the tightest bounds for this algorithm in the general setting where workers optimize heterogeneous objectives. They show that gossip averaging reduces the asymptotic variance suffered by the algorithm at the cost of a degradation (depending on the spectral gap of the gossip matrix) of the initial linear convergence term. This key term does not improve through collaboration and gives rise to a *smaller learning rate* than training alone. Besides, as discussed above, this implies that optimization is not possible in the limit of large graphs, even in the absence of heterogeneity: for instance, the spectral gap of an infinite ring is zero, which would lead to a learning rate of zero as well.

These rates suggest that decentralized averaging speeds up the last part of training (dominated by variance), at the cost of slowing down the initial (linear convergence) phase. Beyond the work of Koloskova et al. (2020), many papers focus on *linear speedup* (in the variance phase) over optimizing alone, and prove similar results in a variety of settings (Lian et al., 2017; Tang et al., 2018; Lian et al., 2018). All these results rely on the following insight: while linear speedup is only achieved for small learning rates, SGD eventually requires such small learning rates anyway (because of, e.g., stochastic noise, or non-smoothness). This observation leads these works to argue that "topology does not matter". This is the case indeed, but only for very small learning rates, as shown in Figure 1. Besides, while linear speedup might be achievable indeed for very small learning rates, some level of variance reduction should be obtained by averaging for *any* learning rate. In practice, averaging speeds up both the initial *and* last part of training and in a possibly non-linear way. This is what we show in this work, both in theory and in practice.

Another line of work studies decentralized SGD under statistical assumptions on the local data. In particular, Richards and Rebeschini (2020) show favorable properties for D-SGD with graph-dependent implicit regularization and attain optimal statistical rates. Their suggested learning rate does depend on the spectral gap of the communication network, and it goes to zero when the spectral gap shrinks. Richards and Rebeschini (2019) also show that larger (constant) learning rates can be used in decentralized GD, but their analysis focuses on decentralized kernel regression. Their analysis relies on statistical concentration of local objectives rather, while the analysis in this paper relies on the notion of local neighborhoods.

Gossiping in infinite graphs. An important feature of our results is that they do not depend on the spectral gap, and so they apply independently of the size of the graph. Instead, our results rely on new quantities that involve a combination of the graph topology and the heterogeneity pattern. These may depend on the spectral gap in extreme cases, but are much better in general. Berthier et al. (2020) study acceleration of gossip averaging in infinite graphs, and obtain the same conclusions as we do: although spectral gap is useful for asymptotics (how long does information take to spread in the whole graph), it fails to accurately describe the transient regime of gossip averaging, *i.e.*, how quickly information spreads over local neighborhoods in the first few gossip rounds. This is especially limiting for optimization (compared to just averaging), as new local updates need to be averaged at every step. The averaging for latest gradient updates always starts in the transient regime, implying that the transient regime of gossip averaging deeply affects the asymptotic regime of decentralized SGD. In this work, we build on tools from Berthier et al. (2020) to show

how the effective number of neighbors, a key quantity we introduce, is related to the graph's spectral dimension.

The impact of the graph topology. Lian et al. (2017) argue that the topology of the graph does not matter. This is only true for asymptotic rates in specific settings, as illustrated in Figure 1. Neglia et al. (2020) investigate the impact of the graph on decentralized optimization, and contradict this claim. Similarly to us, they show that the graph has an impact in the early phases of training. Their analysis of the heterogeneous setting, their analysis depends on how gradient heterogeneity spans the eigenspace of the Laplacian. Their assumptions, however, differ from ours, and they retain an unavoidable dependence on the spectral gap of the graph. Our results are different in nature, and show the benefits of averaging and the impact of the graph through the choice of large learning rates, and a better dependence on the noise and the heterogeneity for a given learning rate. Even et al. (2021) also consider the impact of the graph on decentralized learning. They focus on non-worst-case dependence on heterogeneous delays, and still obtain spectral-gap-like quantities but on a reweighted gossip matrix.

Another line of work studies the interaction of topology with particular patterns of data heterogeneity (Le Bars et al., 2022; Dandi et al., 2022), and how to optimize graphs with this heterogeneity in mind. Our analysis highlights the role of heterogeneity through a different quantity than these works, that we believe is tight. Besides, both works either try to reduce this heterogeneity all along the trajectory, or optimize for both the spectral gap of the graph and the heterogeneity term. Instead, we show that heterogeneity changes the fixed-point of the algorithm but not the global dynamics.

Time-varying topologies. Time-varying topologies are popular for decentralized deep learning in data centers due to their strong mixing (Assran et al., 2019; Wang et al., 2019). The benefit of varying the communication topology over time is not easily explained through standard theory, but requires dedicated analysis (Ying et al., 2021). While our proofs only cover static topologies, the quantities that appear in our analysis can be computed for time-varying schemes, too. With these quantities, we can empirically study static and time-varying schemes in the same framework.

Conference version. This paper is an extension of Vogels et al. (2022), which focused on the homogeneous setting where all workers share the same global optimum. In this extension, we introduce a simpler analysis that strictly improves and generalizes the previous one, extending the results to the important heterogeneous setting. In the conference version, it remained unclear if larger learning rates could only be achieved thanks to homogeneity. We also connect the quantities we introduce to the spectral dimension of a graph, and use this connection to derive explicit formulas for the optimal learning rates based on the spectral dimension. This allows us to accurately compare with previous bounds (for instance Koloskova et al. (2020)) and show that we improve on them in all settings.

3. Measuring collaboration in decentralized learning

Both this paper's analysis of decentralized SGD for general convex objectives and its deep learning experiments revolve around a notion of 'effective number of neighbors' that we would introduce in Section 3.2. The aim of this section is to motivate the quantity based on

a simple toy model for which we can exactly characterize the convergence (Section 3.1). We then connect this quantity to the typical graph metrics such as spectral gap and spectral dimensions in Section 3.3.

3.1 A toy problem: D-SGD on isotropic random quadratics

The aim of this section is to provide intuition while avoiding the complexities of general analysis. To keep this section light, we omit any derivations. The appendix of (Vogels et al., 2022) contains a longer version of this section that includes derivations and proofs.

We consider n workers that jointly optimize an isotropic quadratic $\mathbb{E}_{\mathbf{d}\sim\mathcal{N}^d(0,1)}\frac{1}{2}(\mathbf{d}^\top\mathbf{x})^2 = \frac{1}{2}\|\mathbf{x}\|^2$ with a unique global minimum $\mathbf{x}^\star = \mathbf{0}$. The workers access the quadratic through stochastic gradients of the form $\mathbf{g}(\mathbf{x}) = \mathbf{d}\mathbf{d}^\top\mathbf{x}$, with $\mathbf{d} \sim \mathcal{N}^d(0,1)$. This corresponds to a linear model with infinite data, and where the model can fit the data perfectly, so that stochastic noise goes to zero close to the optimum. We empirically find that this simple model is a meaningful proxy for the initial phase of (over-parameterized) deep learning (Section 5). A benefit of this model is that we can compute exact rates for it. These rates illustrate the behavior that we capture more generally in the theory of Section 4.

The stochasticity in this toy problem can be quantified by the *noise level*

$$\zeta = \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbb{E}_{\mathbf{d}}\|\mathbf{g}(\mathbf{x})\|^2}{\|\mathbf{x}\|^2} = \sup_{\mathbf{x}\in\mathbb{R}^d} \frac{\mathbb{E}_{\mathbf{d}}\|\mathbf{d}\mathbf{d}^\top\mathbf{x}\|^2}{\|\mathbf{x}\|^2}, \tag{1}$$

which is equal to $\zeta = d + 2$, due to the random normal distribution of $\mathbf{d}$.

The workers run the D-SGD algorithm (Lian et al., 2017). Each worker i has its own copy $\mathbf{x}_i \in \mathbb{R}^d$ of the model, and they alternate between local model updates $\mathbf{x}_i \leftarrow \mathbf{x}_i - \eta\mathbf{g}(\mathbf{x}_i)$ and averaging their models with others: $\mathbf{x}_i \leftarrow \sum_{j=1}^n w_{ij}\mathbf{x}_j$. The averaging weights w_{ij} are summarized in the *gossip matrix* $\mathbf{W} \in \mathbb{R}^{n\times n}$. A non-zero weight w_{ij} indicates that i and j are directly connected. In the following, we assume that $\mathbf{W}$ is symmetric and doubly stochastic: $\sum_{j=1}^n w_{ij} = 1\ \forall i$.

On our objective, D-SGD either converges or diverges linearly. Whenever it converges, i.e., when the learning rate is small enough, there is a convergence rate r such that

$$\mathbb{E}\|\mathbf{x}_i^{(t)}\|^2 \leq (1-r)\|\mathbf{x}_i^{(t-1)}\|^2,$$

with equality as $t \to \infty$. When the workers train alone ($\mathbf{W} = \mathbf{I}$), the convergence rate for a given learning rate η reads:

$$r_{\text{alone}} = 1 - (1-\eta)^2 - (\zeta-1)\eta^2. \tag{2}$$

The optimal learning rate $\eta^\star = \frac{1}{\zeta}$ balances the optimization term $(1-\eta)^2$ and the stochastic term $(\zeta-1)\eta^2$. In the centralized (fully connected) setting ($w_{ij} = \frac{1}{n}\ \forall i,j$), the rate is simple as well:

$$r_{\text{centralized}} = 1 - (1-\eta)^2 - \frac{(\zeta-1)\eta^2}{n}. \tag{3}$$

Averaging between n workers reduces the impact of the gradient noise, and the optimal learning rate grows to $\eta^\star = \frac{n}{n+\zeta-1}$. We find that D-SGD with a general gossip matrix $\mathbf{W}$ interpolates those results.

3.2 The effective number of neighbors

To quantify the reduction of the $(\zeta - 1)\eta^2$ term in general, we introduce the problem-independent notion of *effective number of neighbors* $n_{\mathbf{W}}(\gamma)$ of the gossip matrix $\mathbf{W}$ and *decay parameter* γ.

Definition 1 (Effective number of neighbors) *The effective number of neighbors* $n_{\mathbf{W}}(\gamma) = \lim_{t\to\infty} \frac{\sum_{i=1}^n \mathrm{Var}[\mathbf{y}_i^{(t)}]}{\sum_{i=1}^n \mathrm{Var}[\mathbf{z}_i^{(t)}]}$ *measures the ratio of the asymptotic variance of the processes*

$$\mathbf{y}^{(t+1)} = \sqrt{\gamma} \cdot \mathbf{y}^{(t)} + \boldsymbol{\xi}^{(t)}, \quad \text{where } \mathbf{y}^{(t)} \in \mathbb{R}^n \text{ and } \boldsymbol{\xi}^{(t)} \sim \mathcal{N}^n(0, 1) \tag{4}$$

and

$$\mathbf{z}^{(t+1)} = \mathbf{W}(\sqrt{\gamma} \cdot \mathbf{z}^{(t)} + \boldsymbol{\xi}^{(t)}), \quad \text{where } \mathbf{z}^{(t)} \in \mathbb{R}^n \text{ and } \boldsymbol{\xi}^{(t)} \sim \mathcal{N}^n(0, 1). \tag{5}$$

We call $\mathbf{y}$ and $\mathbf{z}$ *random walks* because workers repeatedly add noise to their state, somewhat like SGD's parameter updates. This should not be confused with a 'random walk' over nodes in the graph.

Since averaging with $\mathbf{W}$ decreases the variance of the random walk by at most n, the effective number of neighbors is a number between 1 and n. The decay γ modulates the sensitivity to communication delays. If $\gamma = 0$, workers only benefit from averaging with their direct neighbors. As γ increases, multi-hop connections play an increasingly important role. As γ approaches 1, delayed and undelayed noise contributions become equally weighted, and the reduction tends to n for any connected topology.

Proposition 2 *For regular doubly-stochastic symmetric gossip matrices* $\mathbf{W}$ *with eigenvalues* $\lambda_1, \ldots, \lambda_n$, $n_{\mathbf{W}}(\gamma)$ *has a closed-form expression*

$$n_{\mathbf{W}}(\gamma) = \frac{\frac{1}{1-\gamma}}{\frac{1}{n}\sum_{i=1}^n \frac{{\lambda_i}^2}{1-\lambda_i^2\gamma}}. \tag{6}$$

This follows from unrolling the recursions for $\mathbf{y}$ and $\mathbf{z}$, using the eigendecomposition of $\mathbf{W}$, and the limit $\lim t \to \infty \sum_{k=1}^t x^k = \frac{x}{1-x}$.

While this closed-form expression only covers a restricted set of gossip matrices, the notion of variance reduction in random walks, however, naturally extends to infinite topologies or time-varying averaging schemes. Figure 2 illustrates $n_{\mathbf{W}}$ for various topologies.

In our exact characterization of the convergence of D-SGD on the isotropic quadratic toy problem, we find that the effective number of neighbors appears in place of the number of workers n in the fully-connected rate of Equation 3. The rate r is the unique solution to

$$r = 1 - (1-\eta)^2 - \frac{(\zeta-1)\eta^2}{n_{\mathbf{W}}\left(\frac{(1-\eta)^2}{1-r}\right)}. \tag{7}$$

For fully-connected and disconnected $\mathbf{W}$, $n_{\mathbf{W}}(\gamma) = n$ or 1 respectively, irrespective of γ, and Equation 7 recovers Equations 2 and 3. For other graphs, the effective number of workers depends on the learning rate. Current theory only considers the case where $n_{\mathbf{W}} \approx n$, but

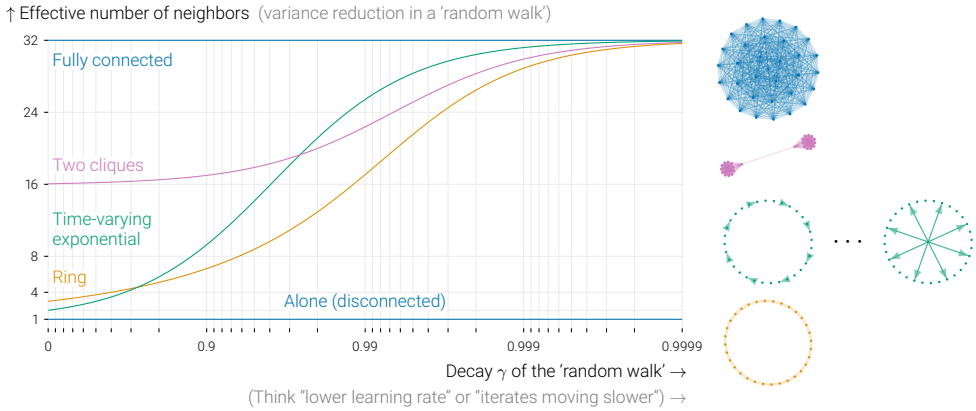

Figure 2: The effective number of neighbors for several topologies measured by their variance reduction in (5). The point γ on the x-axis that matters depends on the learning rate and the task. Which topology is 'best' varies from problem to problem. For large decay rates γ (corresponding small learning rates), all connected topologies achieve variance reduction close to a fully connected graph. For small decay rates (large learning rates), workers only benefit from their direct neighbors (e.g. 3 in a ring). These curves can be computed explicitly for constant topologies, and simulated efficiently for the time-varying exponential scheme (Assran et al., 2019).

the small learning rates this requires can make the term $(1-\eta)^2$ too large, defeating the purpose of collaboration.

Beyond this toy problem, we find that the proposed notion of effective number of neighbors is also meaningful in the analysis of general objectives (Section 4) and in deep learning (Section 5).

3.3 Links between the effective number of neighbors and other graph quantities

In general, the effective number of neighbors function $n_{\mathbf{W}}(\gamma)$ cannot be summarized by a single scalar. Figure 2 demonstrates that the behavior of this function varies from graph to graph. We can, however, bound the effective number of neighbors by known graph quantities such as its spectral gap or spectral dimension.

We aim to create bounds for both finite and infinite graphs. To allow for this, we introduce a generalization of Proposition 2 as an integral over the spectral measure $\mathbf{d}\sigma$ of the gossip matrix, instead of a sum over its eigenvalues:

$$n_{\mathbf{W}}(\gamma)^{-1} = (1-\gamma)\int_0^1 \frac{\lambda^2}{1-\gamma\lambda^2}\mathbf{d}\sigma(\lambda). \tag{8}$$

For finite graphs, $\mathbf{d}\sigma$ is a sum of Dirac deltas of mass $\frac{1}{n}$ at each eigenvalue of matrix $\mathbf{W}$, recovering Equation (6).

3.3.1 Upper and lower bounds

We can use the fact that there all eigenvalues λ are ≤ 1, leading to:

$$n_{\mathbf{W}}(\gamma)^{-1} \leq (1-\gamma)\int_0^1 \frac{1}{1-\gamma}\mathbf{d}\sigma(\lambda) = 1, \tag{9}$$

This lower bound to the 'effective number of neighbors' corresponds to a disconnected graph.

On the other hand, for finite graphs, we can use the fact that $\sigma(\lambda)$ contains a series of n Diracs. The peak at $\lambda = 1$, corresponding to the fully-averaged state, has value $\frac{1}{n}$, while the other peaks have values ≥ 0. Using this bound, we obtain

$$n_{\mathbf{W}}(\gamma)^{-1} \geq \frac{1-\gamma}{1-\gamma}\frac{1}{n} = \frac{1}{n}. \tag{10}$$

This upper bound to the 'effective number of neighbors' is tight for a fully-connected graph.

3.3.2 Bounding by spectral gap

If the graph has a spectral gap α, this means that $\sigma(\lambda)$ contains a Dirac delta with mass $\frac{1}{n}$ at $\lambda = 1$, corresponding to the fully-averaged state. The rest of $\sigma(\lambda)$ has mass $\frac{n-1}{n}$ and is contained in the subdomain $\lambda \in [0, 1-\alpha]$. In this setting, we obtain

$$n_{\mathbf{W}}(\gamma)^{-1} \leq \frac{1}{n} + \frac{n-1}{n}\frac{(1-\gamma)(1-\alpha)^2}{1-\gamma(1-\alpha)^2}. \tag{11}$$

This lower bound to the 'effective number of neighbors' is typically pessimistic, but it is tight for the finite gossip matrix $\mathbf{W} = (1-\alpha)\mathbf{I} + \frac{\alpha}{n}\mathbb{1}\mathbb{1}^\top$.

3.3.3 Bounding by spectral dimension

Next, we will link the notion of 'effective number of neighbors' to the spectral dimension d_s of the graph (Berthier, 2021, e.g. Definition 1.9), which controls the decay of eigenvalues near 1. This notion is usually linked with the spectral measure of the Laplacian of the graph. However, to avoid introducing too many graph-related quantities, we define spectral dimension with respect to the gossip matrix $\mathbf{W}$. Standard definitions using the Laplacian $\mathbf{L_W} = \mathbf{I} - \mathbf{W}$ are equivalent. In the remainder of this paper, the 'graph' will always refer to the communication graph implicitly induced by $\mathbf{W}$ of Laplacian $\mathbf{L_W}$.

Definition 3 (Spectral Dimension) *A gossip matrix has a spectral dimension at least d_s if there exists $c_s > 0$ such that for all $\lambda \in [0,1]$, the density of its eigenvalues is bounded by*

$$\sigma((\lambda, 1)) \leq c_s^{-1}(1-\lambda)^{\frac{d_s}{2}}. \tag{12}$$

The notation $\sigma((\lambda,1))$ here refers to the integral $\int_\lambda^1 \sigma(l)\,\mathbf{d}l$. The spectral dimension of a graph has a natural geometric interpretation. For instance, the line (or ring) are of spectral dimension $d_s = 1$, whereas 2-dimensional grids are of spectral dimension 2. More generally, a d-dimensional torus is of spectral dimension d. Besides, the spectral dimension describes macroscopic topological features and are robust to microscopic changes. For instance, random geometric graphs are of spectral dimension 2.

Note that since finite graphs have a spectral gap, $\sigma((\lambda_2(\mathbf{W}), 1)) = 0$ and so finite graphs verify (12) for any spectral dimension d_s. However, the notion of spectral dimension is still relevant for finite graphs, since the constant c_s blows up when d_s is bigger than the actual spectral dimension of an infinite graph with similar topology. Alternatively, it is sometimes helpful to explicitly take the spectral gap into account in (12), as in Berthier et al. (2020, Section 6).

We now proceed to bounding $n_{\mathbf{W}}(\gamma)$ using the spectral dimension. Since $\lambda \mapsto \lambda^2(1-\gamma\lambda^2)^{-1}$ is a non-negative non-decreasing function on $[0,1]$, we can use Berthier et al. (2020, Lemma C.1) to obtain that:

$$n_{\mathbf{W}}(\gamma)^{-1} \leq \frac{1}{n} + c_s^{-1}(1-\gamma)\int_0^1 \frac{\lambda^2}{1-\gamma\lambda^2}(1-\lambda)^{\frac{d_s}{2}-1}\mathbf{d}\lambda. \tag{13}$$

The term $\frac{1}{n}$ comes from the fact that for finite graphs, the density $\mathbf{d}\sigma$ includes a Dirac delta with mass $\frac{1}{n}$ at eigenvalue 1. This Dirac is not affected by spectral dimension, and is required for consistency, as it ensures that $n_{\mathbf{W}}(\gamma) \leq n$ for any finite graph. To evaluate the integral, we then distinguish three cases.

Case $d_s > 2$. Since $\gamma\lambda < 1$, then $1-\lambda \leq 1-\gamma\lambda^2$. In particular we use integration by parts to get:

$$\begin{aligned} n_{\mathbf{W}}(\gamma)^{-1} - n^{-1} &\leq c_s^{-1}(1-\gamma)\int_0^1 \lambda^2(1-\gamma\lambda^2)^{\frac{d_s}{2}-2}\mathbf{d}\lambda \\ &\leq -\frac{(1-\gamma)c_s^{-1}}{2\gamma(d_s/2-1)}\int_0^1 -2\gamma\lambda(d_s/2-1)(1-\gamma\lambda^2)^{\frac{d_s}{2}-2}\mathbf{d}\lambda \\ &= \frac{(1-\gamma)c_s^{-1}}{\gamma(d_s-2)}\left[1-(1-\gamma)^{\frac{d_s}{2}-1}\right]. \end{aligned}$$

This leads to a scaling of:

$$n_{\mathbf{W}}(\gamma) \geq \left(\frac{1}{n} + \frac{(1-\gamma)}{\gamma(d_s-2)c_s}\right)^{-1}. \tag{14}$$

For large enough n, we obtain the same scaling of $(1-\gamma)^{-1}$ as in the previous section, thus indicating that for networks that are well-enough connected ($d_s > 2$), the spectral dimension only affects the constants, and not the scaling in γ.

Case $d_s = 2$. When $d_s = 2$, only the primitive of the integrand changes, leading to:

$$n_{\mathbf{W}}(\gamma) \geq \left(\frac{1}{n} - \frac{(1-\gamma)\ln(1-\gamma)}{2\gamma c_s}\right)^{-1} \tag{15}$$

Case $d_s < 2$. In this case, we start by splitting the integral as:

$$(1-\gamma)\int_0^1 \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda = (1-\gamma)\int_0^\gamma \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda + (1-\gamma)\int_\gamma^1 \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda$$

For the first term, note that $\gamma\lambda \leq 1$, so $(1-\gamma\lambda^2)^{-1} \leq (1-\lambda)^{-1}$, leading to:

$$(1-\gamma)\int_0^\gamma \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda \leq (1-\gamma)\int_0^\gamma (1-\lambda)^{\frac{d_s}{2}-2}\mathbf{d}\lambda$$
$$= \frac{2(1-\gamma)}{2-d_s}\left[(1-\gamma)^{\frac{d_s}{2}-1}-1\right] \leq \frac{2}{2-d_s}(1-\gamma)^{\frac{d_s}{2}}.$$

For the second term, note that $\lambda^2 \leq 1$, so $(1-\gamma\lambda^2)^{-1} \leq (1-\gamma)^{-1}$, leading to:

$$(1-\gamma)\int_\gamma^1 \frac{\lambda^2(1-\lambda)^{\frac{d_s}{2}-1}}{(1-\gamma\lambda^2)}\mathbf{d}\lambda \leq \int_\gamma^1 (1-\lambda)^{\frac{d_s}{2}-1}\mathbf{d}\lambda = \frac{2}{d_s}(1-\gamma)^{\frac{d_s}{2}}. \tag{16}$$

In the end, we obtain that $n_{\mathbf{W}}(\gamma)^{-1} - \frac{1}{n} \leq \frac{2}{c_s}\left[\frac{1}{2-d_s}+\frac{1}{d_s}\right](1-\gamma)^{\frac{d_s}{2}}$, and so:

$$n_{\mathbf{W}}(\gamma) \geq \left(\frac{1}{n} + \frac{4(1-\gamma)^{\frac{d_s}{2}}}{d_s(2-d_s)c_s}\right)^{-1}. \tag{17}$$

In this case, scaling in γ is impacted by the spectral dimension. Better-connected graphs benefit more from higher γ.

4. Convergence analysis

4.1 Notations and Definitions

In the previous section, we have derived exact rates for a specific function. Now we present convergence rates for general (strongly) convex functions that are consistent with our observations in the previous section. We obtain rates that depend on the level of noise, the hardness of the objective, and the topology of the graph. More formally, we assume that we would like to solve the following problem:

$$\min_{\theta\in\mathbb{R}^d}\sum_{i=1}^n f_i(\theta) = \min_{\mathbf{x}\in\mathbb{R}^{nd},\mathbf{x}_i=\mathbf{x}_j}\sum_{i=1}^n f_i(\mathbf{x}_i). \tag{18}$$

In this case, $\mathbf{x}_i \in \mathbb{R}^d$ represents the local variable of node i, and $\mathbf{x} \in \mathbb{R}^{nd}$ the stacked variables of all nodes. We will assume the following iterations for D-SGD:

$$\text{(D-SGD):} \qquad \mathbf{x}_i^{(t+1)} = \sum_{j=1}^n w_{ij}\mathbf{x}_j^{(t)} - \eta\nabla f_{\xi_i^{(t)}}(\mathbf{x}_i^{(t)}) \tag{19}$$

where $f_{\xi_i^{(t)}}$ represent sampled data points and the gossip weights w_{ij} are elements of $\mathbf{W}$. Denoting $\mathbf{L_W} = \mathbf{I} - \mathbf{W}$, we rewrite this expression in matrix form as:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \left[\eta\nabla F_{\xi^{(t)}}(\mathbf{x}^{(t)}) + \mathbf{L_W}\mathbf{x}^{(t)}\right], \tag{20}$$

where $(\nabla F_{\xi^{(t)}}(\mathbf{x}^{(t)}))_i = \nabla f_{\xi_i^{(t)}}(\mathbf{x}_i^{(t)})$. We abuse notations in the sense that $\mathbf{W} \in \mathbb{R}^{nd\times nd}$ is now the Kronecker product of the standard $n\times n$ gossip matrix and the $d\times d$ identity matrix.

This definition is a slight departure from the conference version of this work (Vogels et al., 2022), which alternated randomly between gossip steps and gradient updates instead of in turns. The analysis of the randomized setting is still possible, but with heterogeneous objectives $\mathbf{x}_i \neq \sum_{j=1}^n w_{ij}\mathbf{x}_j$, even for the fixed points of D-SGD (19), and randomizing the updates adds undesirable variance. Similarly, it is also possible to analyze the popular variant $\mathbf{x}^{(t+1)} = \mathbf{W}[\mathbf{x}^{(t)} - \eta\nabla F_{\xi^{(t)}}(\mathbf{x}^{(t)})]$, which locally averages the stochastic gradients before they are applied. Yet, the D-SGD algorithm in (19) allows communications and computations to be performed in parallel, and leads to a simpler analysis. We analyze this model under the following assumptions, where $D_f(x,y) = f(x) - f(y) - \nabla f(y)^\top(x-y)$ denotes the Bregman divergence of f between points x and y.

Assumption 4 *The stochastic gradients are such that: (I) the sampled data points $\xi_i^{(t)}$ and $\xi_j^{(\ell)}$ are independent across times t, ℓ and nodes $i \neq j$. (II) stochastic gradients are locally unbiased: $\mathbb{E}\left[f_{\xi_i^{(t)}}\right] = f_i$ for all t, i (III) the objectives $f_{\xi_i^{(t)}}$ are convex and ζ_ξ-smooth for all t, i, with $\mathbb{E}\left[\zeta_\xi D_{f_\xi}(x,y)\right] \leq \zeta D_f(x,y)$ for all x, y. (IV) all local objectives f_i are μ-strongly-convex for $\mu \geq 0$ and L-smooth.*

Large learning rates. The smoothness constant ζ of the stochastic functions f_ξ defines the level of noise in the problem (the lower, the better) in the transient regime. The ratio ζ/L compares the difficulty of optimizing with stochastic gradients to the difficulty with the true global gradient before reaching the 'variance region' in which the iterates of D-SGD with a constant learning rate lie almost surely as $t \to \infty$. This ratio is thus especially important in interpolating settings when all $f_{\xi_i^{(t)}}$ have the same minimum, so that the 'variance region' is reduced to the optimum $\mathbf{x}^\star$. Assuming better smoothness for the global average objective than for the local functions is key to showing that averaging between workers allows for larger learning rates. Without communication, convergence to the 'variance region' is ensured for learning rates $\eta \leq 1/\zeta$. If $\zeta \approx L$, there is little noise and cooperation only helps to reduce the final variance, and to get closer to the *global* minimum (instead of just your own). Yet, in noisy regimes ($\zeta \gg L$), such as in Section 3.1 in which $\zeta = d + 2 \gg 1 = L$, averaging enables larger learning rates up to $\min(1/L, n/\zeta)$, greatly speeding up the initial training phase. This is precisely what we will prove in Theorem 6.

If the workers always remain close ($\mathbf{x}_i \approx \frac{1}{n}(\mathbf{x}_1 + \ldots + \mathbf{x}_n)$ $\forall i$, or equivalently $\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{x} \approx \mathbf{x}$), D-SGD behaves the same as SGD on the average parameter $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, and the learning rate depends on $\max(\zeta/n, L)$, showing a reduction of variance by n. Maintaining "$\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{x} \approx \mathbf{x}$", however, requires a small learning rate. This is a common starting point for the analysis of D-SGD, in particular for the proofs in Koloskova et al. (2020). On the other extreme, if we do not assume closeness between workers, "$\mathbf{I}\mathbf{x} \approx \mathbf{x}$" always holds. In this case, there is no variance reduction, but no requirement for a small learning rate either. In Section 3.1, we found that, at the optimal learning rate, workers are *not* close to all other workers, but they *are* close to others that are not too far away in the graph.

We capture the concept of 'local closeness' by defining a neighborhood matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$. It allows us to consider semi-local averaging beyond direct neighbors, but without fully averaging with the whole graph. We ensure that "$\mathbf{M}\mathbf{x} \approx \mathbf{x}$", leading to an improvement in the smoothness somewhere between ζ (achieved alone) and ζ/n (achieved when global consensus

is maintained). Each neighborhood matrix $\mathbf{M}$ implies a requirement on the learning rate, as well as an improvement in smoothness.

While we can conduct our analysis with any $\mathbf{M}$, those matrices that strike a good balance between the learning rate requirement and improved smoothness are most interesting. Based on Section 3.1, we therefore focus on a specific construction of matrices: We choose $\mathbf{M}$ as the covariance of a decay-γ 'random walk process' with the graph, as in (5), meaning that

$$\mathbf{M} = (1-\gamma)\sum_{k=1}^{\infty}\gamma^{k-1}\mathbf{W}^{2k} = (1-\gamma)\mathbf{W}^2(\mathbf{I}-\gamma\mathbf{W}^2)^{-1}. \tag{21}$$

Varying γ induces a spectrum of averaging neighborhoods from $\mathbf{M} = \mathbf{W}^2$ ($\gamma = 0$) to $\mathbf{M} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ ($\gamma = 1$). γ also implies an effective number of neighbors $n_{\mathbf{W}}(\gamma)$: the larger γ, the larger $n_{\mathbf{W}}(\gamma)$. We make the following assumption on the neighborhood matrix $\mathbf{M}$:

Assumption 5 *The neighborhood matrix $\mathbf{M}$ is of the form of (21), and all the diagonal elements have the same value,* i.e., $\mathbf{M}_{ii} = \mathbf{M}_{jj}$ *for all i, j.*

Assumption 5 implies that ${\mathbf{M}_{ii}}^{-1} = n_{\mathbf{W}}(\gamma)$: the effective number of neighbors defined in (6) is equal to the inverse of the self-weights of $\mathbf{M}$. This comes from the fact that the trace of $\mathbf{M}$ is equal to the sum of its eigenvalues. Otherwise, all results that require Assumption 5 hold by replacing $n_{\mathbf{W}}(\gamma)$ with $\min_i {\mathbf{M}_{ii}}^{-1}$. Besides this interesting relationship with the effective number of neighbors $n_{\mathbf{W}}(\gamma)$, we will be interested in another spectral property of $\mathbf{M}$, namely the constant $\beta(\gamma)$ (which only depends on γ through $\mathbf{M}$, but we make this dependence explicit), which is such that:

$$\mathbf{L}_{\mathbf{M}} \preccurlyeq \beta(\gamma)^{-1}\mathbf{L}_{\mathbf{W}}\mathbf{W} \tag{22}$$

This constant can be interpreted as the strong convexity of the semi-norm defined by $\mathbf{L}_{\mathbf{W}}\mathbf{W}$ relatively to the one defined by $\mathbf{L}_{\mathbf{M}}$. Due to the form of $\mathbf{M}$, we have $1 - \lambda_2(\mathbf{W}) \leq \beta(\gamma) \leq 1$, and the lower bound is tight for $\gamma \to 1$. However, the specific form of $\mathbf{M}$ (involving neighborhoods as defined by $\mathbf{W}$) and the use of $\gamma < 1$ ensure a much larger constant $\beta(\gamma)$ in general.

Fixed points of D-(S)GD. In Vogels et al. (2022), we consider a homogeneous setting, in which $\mathbb{E}\, f_{\xi_i^{(t)}} = f$ for all i. We now go beyond this analysis, and consider a setting in which local functions f_i might be different. In this case, constant-learning-rate Decentralized Gradient Descent (the deterministic version of D-SGD) does not converge to the minimizer of the average function but to a different one. Let us now consider this fixed point $\mathbf{x}_\eta^\star$, which verifies:

$$\eta\nabla F(\mathbf{x}_\eta^\star) + \mathbf{L}_{\mathbf{W}}\mathbf{x}_\eta^\star = 0. \tag{23}$$

Note that $\mathbf{x}_\eta^\star$ crucially depends on the learning rate η (which we emphasize in the notation) and that it is generally not at consensus ($\mathbf{L}_{\mathbf{W}}\mathbf{x}_\eta^\star \neq 0$). In the presence of stochastic noise, D-SGD will oscillate in a neighborhood (proportional to the gradients' variance) of this fixed point $\mathbf{x}_\eta^\star$, and so from now on we will refer to $\mathbf{x}_\eta^\star$ as the fixed point of D-SGD.

In the remainder of this section, we show that the results from Vogels et al. (2022) still hold as long as we replace the global minimizer $\mathbf{x}^\star$ (solution of Problem (18)) by this fixed

point $\mathbf{x}_\eta^\star$. More specifically, we measure convergence by ensuring the decrease of the following Lyapunov function:

$$\mathcal{L}_t = \|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{M}}^2 + \omega\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{L_M}}^2 = (1-\omega)\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|_{\mathbf{M}}^2 + \omega\|\mathbf{x}^{(t)} - \mathbf{x}_\eta^\star\|^2, \tag{24}$$

for some parameter $\omega \in [0,1]$, and where $\mathbf{L_M} = \mathbf{I} - \mathbf{M}$. Then, we will show how these results imply convergence to a neighborhood of $\mathbf{x}_\eta^\star$, and that this neighborhood shrinks with smaller learning rates η. More specifically, the section unrolls as follows:

1. Theorem 6 first proves a general convergence result to $\mathbf{x}_\eta^\star$, the fixed point of D-(S)GD.
2. Theorem 9 then bounds the distance to the true optimum for general learning rates.
3. Corollary 10 finally gives a full convergence result with optimized learning rates. Readers interested in quickly comparing our results with state-of-the art ones can skip to this result.

4.2 General convergence result

Theorem 6 provides convergence rates for any choice of the parameter γ that determines the neighborhood matrix $\mathbf{M}$, and for any Lyapunov parameter ω. The best rates are obtained for specific γ and ω that balance the benefit of averaging with the constraint it imposes on closeness between neighbors. We will discuss these choices more in depth in the next section.

Theorem 6 *If Assumptions 4 and 5 hold and if η is such that*

$$\eta \leq \min\left(\frac{\beta(\gamma)\omega}{L}, \frac{1}{4\left(\left[n_{\mathbf{W}}(\gamma)^{-1} + \omega\right]\zeta + L\right)}\right), \tag{25}$$

then the Lyapunov function defined in (24) verifies the following:

$$\mathcal{L}^{(t+1)} \leq (1-\eta\mu)\mathcal{L}^{(t)} + \eta^2\sigma_{\mathbf{M}}^2,$$

where $\sigma_{\mathbf{M}}^2 = 2[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1} + \omega]\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right]$.

This theorem shows convergence (up to a variance region) to the fixed point $\mathbf{x}_\eta^\star$ of D-SGD, regardless of the 'true' minimizer $\mathbf{x}^\star$. Although converging to $\mathbf{x}_\eta^\star$ might not be ideal depending on the use case (but do keep in mind that $\mathbf{x}_\eta^\star \to \mathbf{x}^\star$ as η shrinks), this is what D-SGD does, and so we believe it is important to start by stating this clearly. The homogeneous case did not have this problem since $\mathbf{x}_\eta^\star = \mathbf{x}^\star$ for all η for η that implied convergence.

Parameter $\omega \in [0,1]$ is free, and it is often convenient to choose it as $\omega = \eta L/\beta(\gamma)$ to get rid of the first condition on η. However, we present the result with a free parameter ω since, as we will see in the remainder of this section, setting $\omega = n_{\mathbf{W}}(\gamma)^{-1}$ allows for simple corollaries.

Proof We now detail the proof, which is both a simplification and generalization of Theorem IV from Vogels et al. (2022).

1 - General decomposition We first analyze the first term in the Lyapunov (24), and use the fixed-point conditions of (23) to write:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}_\eta^\star\|_{\mathbf{M}}^2\right] &= \|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{M}}^2 + \|\eta\nabla F_{\xi_t}(\mathbf{x}^{(t)})+\mathbf{L}_{\mathbf{W}}\mathbf{x}^{(t)}\|_{\mathbf{M}}^2 \\ &\quad - 2\eta\left[\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]^\top \mathbf{M}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star) - 2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L}_{\mathbf{W}}\mathbf{M}}^2.\end{aligned} \tag{26}$$

The second term is the same with $\mathbf{M}$ in place of $\mathbf{I}$.

2 - Error terms We start by bounding the error terms, and use the optimality conditions to obtain:

$$\begin{aligned}&\mathbb{E}\left[\|\eta\nabla F_{\xi_t}(\mathbf{x}^{(t)})+\mathbf{L}_{\mathbf{W}}\mathbf{x}^{(t)}\|_{\mathbf{M}}^2\right] \\ &= \mathbb{E}\left[\|\eta\left[\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]+\mathbf{L}_{\mathbf{W}}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right] \\ &= \mathbb{E}\left[\|\eta\left(\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\right)+\left[\eta\left(\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\right)+\mathbf{L}_{\mathbf{W}}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)\right]\|_{\mathbf{M}}^2\right] \\ &\leq 2\eta^2\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right]+2\eta^2\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right]+2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L}_{\mathbf{W}}\mathbf{M}\mathbf{L}_{\mathbf{W}}}^2,\end{aligned}$$

where the last inequality comes from the bias-variance decomposition. The second term corresponds to variance, whereas the first and last one will be canceled by descent terms.

Stochastic gradient noise. To bound the first term, we crucially use that stochastic noises are *independent* for two different nodes, so in particular:

$$\begin{aligned}\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right] &= n_{\mathbf{W}}(\gamma)^{-1}\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)\|^2\right] \\ &\quad + \|\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|_{\mathbf{M}-n_{\mathbf{W}}(\gamma)^{-1}\mathbf{I}}^2 \\ &\leq 2n_{\mathbf{W}}(\gamma)^{-1}\,\mathbb{E}\left[\zeta_{\xi_t}D_{F_{\xi_t}}(\mathbf{x}_\eta^\star,\mathbf{x}^{(t)})\right]+\|\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|^2 \\ &\leq 2\left[n_{\mathbf{W}}(\gamma)^{-1}\zeta+L\right]D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star),\end{aligned}$$

where we used that $\mathbf{M}\preccurlyeq\mathbf{I}$, the L-cocoercivity of F, and the noise assumption, i.e., $\mathbb{E}\left[\zeta_{\xi_t}D_{F_{\xi_t}}\right]\leq\zeta D_F$. The effective number of neighbors $n_{\mathbf{W}}(\gamma)$ kicks in since Assumption 5 implies that the diagonal of $\mathbf{M}$ is equal to $n_{\mathbf{W}}(\gamma)^{-1}\mathbf{I}$. Using independence again, we obtain:

$$\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|_{\mathbf{M}}^2\right] = n_{\mathbf{W}}(\gamma)^{-1}\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|^2\right] \tag{27}$$

Performing the same computations for the $\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|^2\right]$ term and adding consensus error leads to:

$$\begin{aligned}\mathbb{E}\left[\|\eta\nabla F_{\xi_t}(\mathbf{x}^{(t)})+\mathbf{L}_{\mathbf{W}}\mathbf{x}^{(t)}\|_{(1-\omega)\mathbf{M}+\omega\mathbf{I}}^2\right] &\leq 4\left[\left[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1}+\omega\right]\zeta+(1-\omega)L\right]D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star) \\ &\quad + 2\eta^2((1-\omega)n_{\mathbf{W}}(\gamma)^{-1}+\omega)\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|^2\right] \\ &\quad + 2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L}_{\mathbf{W}}[\mathbf{M}+\omega\mathbf{L}_{\mathbf{M}}]\mathbf{L}_{\mathbf{W}}}^2\end{aligned} \tag{28}$$

Here, the first term will be controlled by the descent obtained through the gradient terms, and the second one through communication terms.

3 - Descent terms

Gradient terms We first analyze the effect of all gradient terms. In particular, we use that $(1-\omega)\mathbf{M}+\omega\mathbf{I}=\mathbf{I}-(1-\omega)\mathbf{L_M}$. Then, we use that

$$\left[\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]^\top(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)=D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star)+D_F(\mathbf{x}_\eta^\star,\mathbf{x}^{(t)}),$$

and:

$$\begin{aligned}2\left[\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]^\top\mathbf{L_M}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)&\leq 2\|\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|\|\mathbf{L_M}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)\|\\&\leq\frac{1}{2L}\|\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\|^2+2L\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2}\\&\leq D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star)+2L\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2}.\end{aligned}$$

Overall, the gradient terms sum to:

$$\begin{aligned}&-2\left[\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]^\top(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)+2(1-\omega)\left[\nabla F(\mathbf{x}^{(t)})-\nabla F(\mathbf{x}_\eta^\star)\right]^\top\mathbf{L_M}(\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star)\\&\leq-2D_F(\mathbf{x}_\eta^\star,\mathbf{x}^{(t)})-(1+\omega)D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star)+2(1-\omega)L\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2}\\&\leq-\mu\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|^2-D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star)+2L\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_M}^2}\\&\leq-(1-\omega)\mu\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{M}}^2-\omega\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|^2-D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star)+2\beta(\gamma)^{-1}L\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_M}\mathbf{L_W}\mathbf{W}},\end{aligned}\tag{29}$$

where we used that $\mathbf{L_M}\preccurlyeq\beta(\gamma)^{-1}\mathbf{L_W}\mathbf{W}$.

Gossip terms. We simply recall the gossip terms we use for descent here, which write:

$$-2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_W}\mathbf{M}}^2-2\omega\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_W}\mathbf{L_M}}^2.\tag{30}$$

4 - Putting everything together. We now add all the descent and error terms together. More specifically, using Equations (28), (29) and (30) we obtain:

$$\begin{aligned}\mathcal{L}^{(t+1)}\leq&\,(1-\eta\mu)\mathcal{L}^{(t)}\\&-2\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_W}\mathbf{M}(\mathbf{I}-\mathbf{L_W})}^2\\&-2\omega\left[1-\eta L/(\omega\beta(\gamma))\right]\|\mathbf{x}^{(t)}-\mathbf{x}_\eta^\star\|_{\mathbf{L_W}\mathbf{L_M}\mathbf{W}}^2\\&-\eta\left(1-4\eta\left[\left[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1}+\omega\right]\zeta+(1-\omega)L\right]\right)D_F(\mathbf{x}^{(t)},\mathbf{x}_\eta^\star)\\&+2\eta^2\left[(1-\omega)n_{\mathbf{W}}(\gamma)^{-1}+\omega\right]\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star)-\nabla F(\mathbf{x}_\eta^\star)\|^2\right].\end{aligned}$$

The conditions in the theorem are chosen so that the terms from lines 3 and 4 are positive (which is automatically true for line 2), and using that $1-\omega\leq 1$ (since ω is small anyway). ■

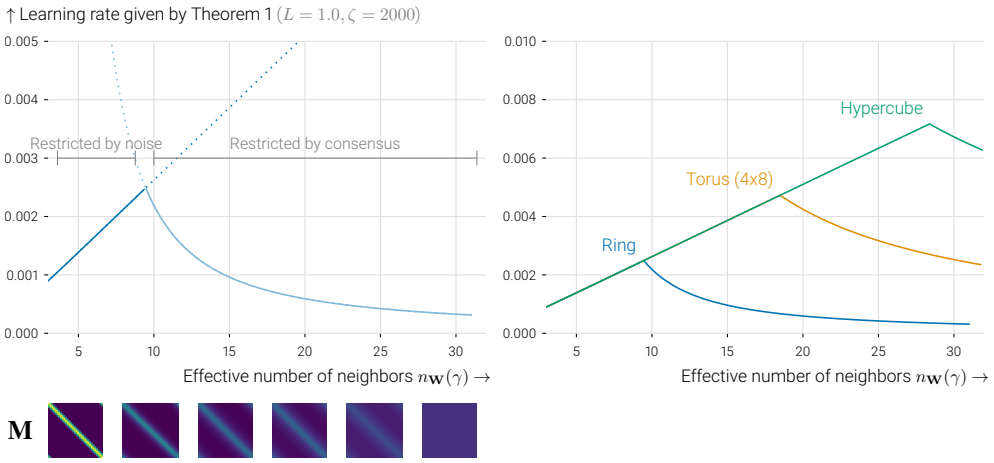

Figure 3: Maximum learning rates prescribed by Theorem 6, varying the parameter γ that implies an effective neighborhood size (x-axis) and an averaging matrix $\mathbf{M}$ (drawn as *heatmaps*). On the *left*, we show the details for a 32-worker ring topology, and on the *right*, we compare it to more connected topologies. Increasing γ (and with it $n_{\mathbf{W}}(\gamma)$) initially leads to larger learning rates thanks to noise reduction. At the optimum, the cost of consensus exceeds the benefit of further reduced noise.

4.3 Main corollaries

4.3.1 Large learning rate: speeding up convergence for large errors

We now investigate Theorem 6 in the case in which both the noise σ^2 and the heterogeneity $\|\nabla F(\mathbf{x}^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}$ are small (compared to $\mathcal{L}^{(0)}$), and so we would like to have the highest possible learning rate in order to ensure fast decrease of the objective (which is consistent with Figure 1). Using (25), we obtain a rate for each parameter γ that controls the local neighborhood size (remember that $\beta(\gamma)$ depends on γ). The task that remains is to find the γ parameter that gives the best convergence guarantees (the largest learning rate). As explained before, one should never reduce the learning rate in order to be close to others, because the goal of collaboration (in this regime in which we are not affected by variance and heterogeneity) is to *increase* the learning rate.

We illustrate this in Figure Figure 3, that we obtain by choosing $\omega = n_{\mathbf{W}}(\gamma)^{-1}$, and evaluating the two terms of (25) for different values of γ. The expression for the linear part of the curve (before consensus dominates) is given in Corollary 7.

Corollary 7 *Consider that Assumptions 4 and 5 hold, then the largest (up to constants) learning rate is obtained as:*

$$\eta = (8\zeta/n_{\mathbf{W}}(\gamma) + 4L)^{-1}\,, \text{ for } \gamma \text{ such that } 4n_{\mathbf{W}}(\gamma)^{-1}\beta(\gamma)(2n_{\mathbf{W}}(\gamma)^{-1}\zeta + L) \leq L \qquad (31)$$

We see that the learning rate scales linearly with the number of effective neighbors in this case (which is equivalent to taking a mini-batch of size linear in $n_{\mathbf{W}}(\gamma)$) until a certain number

of neighbors is reached (condition on the right), or centralized performance is achieved ($\zeta = n_{\mathbf{W}}(\gamma)L$). The condition on γ always has a solution since when $\gamma \approx 0$, both $\beta(\gamma)$ and $n_{\mathbf{W}}(\gamma)^{-1}$ are close to 1, and they both decrease when γ grows. This corollary directly follows from taking $\omega = n_{\mathbf{W}}(\gamma)^{-1}$ in Theorem 6. Note that a slightly tighter choice could be obtained by setting $\omega = \eta\beta(\gamma)/L$.

Investigating $\beta(\gamma)$. We now evaluate $\beta(\gamma)$ in order to obtain more precise bounds. In particular, choosing $\mathbf{M}$ as in (21), the eigenvalues of $\mathbf{L}_{\mathbf{M}}$ are equal to:

$$\lambda_i^{\mathbf{L}_{\mathbf{M}}} = \frac{1-\lambda_i^2}{1-\gamma\lambda_i^2}, \tag{32}$$

where λ_i are the eigenvalues of $\mathbf{W}$. In particular, $\beta(\gamma)\mathbf{L}_{\mathbf{M}} \preccurlyeq \mathbf{W}\mathbf{L}_{\mathbf{W}}$ translates into the fact that for all i such that $\lambda_i \neq 1$ (automatically verified in this case), we want for all i:

$$\beta(\gamma) \leq \frac{1-\gamma\lambda_i^2}{1-\lambda_i^2}(1-\lambda_i)\lambda_i = \frac{\lambda_i(1-\gamma\lambda_i^2)}{1+\lambda_i}. \tag{33}$$

We now make the simplifying assumption that $\lambda_{\min}(\mathbf{W}) \geq \frac{1}{2}$ (which we can always enforce by taking $\mathbf{W}' = (\mathbf{I}+\mathbf{W})/2$), but note that the theory holds regardless. We motivate this simplifying assumption by the fact that the for arbitrarily small spectral gaps, the right side of (33) will always be minimized for $\lambda_2(\mathbf{W})$ assuming γ is large enough, so the actual value of $\lambda_{\min}(\mathbf{W}) < 1$ does not matter. In particular, in this case, *neglecting the effect of the spectral gap*, we can just take:

$$\beta(\gamma) = \frac{1-\gamma\lambda_2(\mathbf{W})}{4} \geq \frac{1-\gamma}{4}, \tag{34}$$

Note that $\beta(\gamma)$ allows for large γ when the spectral gap $1-\lambda_2(\mathbf{W})$ is large, but we allow non-trivial learning rates $\eta > 0$ even when $\lambda_2(\mathbf{W}) = 1$ (infinite graphs) as long as $\gamma < 1$.

Optimal choice of $n_{\mathbf{W}}(\gamma)$. Leveraging the spectral dimension results from Section 3.1, we obtain the following corollary:

Corollary 8 *Under Assumption 4 and 5, and assuming that $\lambda_{\min}(\mathbf{W}) \geq \frac{1}{2}$, that the communication graph has spectral dimension $d_s > 2$, and that $\zeta \gg L$, the highest possible learning rate is*

$$\eta = \frac{1}{8}\left(\frac{c_s(d_s-2)}{\zeta^2 L}\right)^{\frac{1}{3}}, \text{ obtained for } n_{\mathbf{W}}(\gamma) = \left(c_s(d_s-2)\frac{\zeta}{L}\right)^{\frac{1}{3}} \tag{35}$$

This result follows from Corollary 7, which, if $\zeta \gg L$, writes:

$$\frac{L}{\zeta} \geq 8 n_{\mathbf{W}}(\gamma)^{-2}\beta(\gamma) = n_{\mathbf{W}}(\gamma)^{-3} c_s(d_s-2), \tag{36}$$

where the right part is obtained by plugging in the expressions for $\beta(\gamma)$ from (34) into $n_{\mathbf{W}}(\gamma)^{-1} \leq \frac{2(1-\gamma)}{c_s(d_s-2)}$ from (14) (assuming $\gamma \geq 1/2$). Then, one can solve for $1-\gamma$. Assumptions besides Assumption 4 allow to give a simple result in this specific case, but similar expressions can easily be obtained for $d_s \leq 2$ and $\zeta < L n_{\mathbf{W}}(\gamma)$.

4.3.2 Small learning rate: approaching the optimum arbitrarily closely

Theorem 6 gives a convergence result to $\mathbf{x}_\eta^\star$, the fixed point of D-SGD, and we have investigated in the previous section the behavior of D-SGD for large learning rates. In Theorem 9, we focus on small error levels, for which the *variance* and *heterogeneity* terms dominate, and we would like to take small learning rates η. In this setting, we bound the distance between the current iterate and the *true minimizer* $\mathbf{x}^\star$ instead of $\mathbf{x}_\eta^\star$. We also provide a result that gets rid of all dependence on $\mathbf{x}_\eta^\star$, and only explicitly depends on the learning rate η.

Theorem 9 *Under the same assumptions and conditions on the learning rate as Theorem 6 and Corollary 8, we have that:*

$$\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}} \leq 2(1-\eta\mu)^t \mathcal{L}^{(0)} + \frac{2\eta\sigma_{\mathbf{M}}^2}{\mu} + 2\eta^2(1+\kappa)\|\mathbf{L}_{\mathbf{W}}{}^\dagger \nabla F(\mathbf{x}_\eta^\star)\|^2 \tag{37}$$

We can further remove $\mathbf{x}_\eta^\star$ from the bound, and obtain:

$$\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}} \leq 2(1-\eta\mu)^t \mathcal{L}^{(0)} + \frac{6\eta\sigma_{\mathbf{M},\star}^2}{\mu} + 6\eta^2\kappa p^{-1}\Delta_{\mathbf{W}}^2,$$

where $\sigma_{\mathbf{M},\star}^2 = (n_{\mathbf{W}}(\gamma)^{-1} + \omega)\,\mathbb{E}\left[\|\nabla F_\xi(\mathbf{x}^\star) - \nabla F(\mathbf{x}^\star)\|^2\right]$ *and* $p^{-1} = \max_\eta \frac{\|\mathbf{L}_{\mathbf{W}}{}^\dagger \nabla F(\mathbf{x}_\eta^\star)\|^2}{\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}}$, *so that* $p \geq 1 - \lambda_2(\mathbf{W})$, *and* $\Delta_{\mathbf{W}}^2 = \|\nabla F(\mathbf{x}^\star)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}$

The norm $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}}^2$ considers convergence of locally averaged neighborhoods, but $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|_{\mathbf{M}}^2 \geq \|\overline{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2$ since $\mathbb{1}$ is an eigenvector of $\mathbf{M}$ with eigenvalue 1. We now briefly discuss the various terms in this corollary, and then prove it.

Heterogeneity term. The term due to heterogeneity only depends on the distance between the true optimum $\mathbf{x}^\star$ and the fixed point $\mathbf{x}_\eta^\star$, which we then transform into a condition on $\|\nabla F(\mathbf{x}^\star)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}$. In particular, it is not influenced by the choice of $\mathbf{M}$ (and thus of γ).

Constant p. We introduce constant p to get rid of the explicit dependence on $\mathbf{x}_\eta^\star$. Indeed, p^{-1} intuitively denotes how large $\mathbf{L}_{\mathbf{W}}{}^\dagger$ is in the direction of $\nabla F(\mathbf{x}_\eta^\star)$. For instance, if $\nabla F(\mathbf{x}_\eta^\star)$ is an eigenvector of $\mathbf{W}$ associated with eigenvalue λ, then we have $p = 1-\lambda$. In the worst case, we have that $p = 1 - \lambda_2(\mathbf{W})$, but p can be much better in general, when the heterogeneity is spread evenly, instead of having very different functions on distant nodes.

Variance term. In this case, the largest variance reduction (of order n) is obtained by taking ω and $n_{\mathbf{W}}(\gamma)^{-1}$ as small as possible. For learning rates that are too large to imply $n_{\mathbf{W}}(\gamma)^{-1} \approx n^{-1}$, decreasing it decreases the variance term in two ways: (I) directly, through the η term, (II) indirectly, by allowing to take smaller values of $n_{\mathbf{W}}(\gamma)^{-1}$.

For very large (infinite) graphs, we can take $\omega = n_{\mathbf{W}}(\gamma)^{-1}$, and in this case Theorem 6 gives that the smallest $n_{\mathbf{W}}(\gamma)^{-1}$ is given by $n_{\mathbf{W}}(\gamma)^{-1}\beta(\gamma) = \eta L$. Using spectral dimension results (for instance with $d_s > 2$), we obtain (similarly to Corollary 8) that we can take

$\beta(\gamma) = n_{\mathbf{W}}(\gamma)^{-1} c_s(d_s-2)/8$, and so:

$$n_{\mathbf{W}}(\gamma)^{-1} = \sqrt{\frac{8\eta L}{c_s(d_s-2)}}, \tag{38}$$

so the residual variance term for this choice of $n_{\mathbf{W}}(\gamma)^{-1}$ is of order:

$$\mathcal{O}\left(\frac{\eta^{\frac{3}{2}}}{\mu}\sqrt{\frac{L}{c_s(d_s-2)}}\,\mathbb{E}\left[\|\nabla F_\xi(\mathbf{x}^\star) - \nabla F(\mathbf{x}^\star)\|^2\right]\right) \tag{39}$$

In particular, we obtain *super-linear* scaling when reducing the learning rate η thanks to the added benefit of gaining more effective neighbors. Note that again, the cases $d_s \leq 2$ can be treated in the same way.

Proof [Theorem 9] We start by writing:

$$\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2_{\mathbf{M}} \leq 2\|\mathbf{x}^{(t)} - \mathbf{x}^\star_\eta\|^2_{\mathbf{M}} + 2\|\mathbf{x}^\star_\eta - \mathbf{x}^\star\|^2_{\mathbf{M}} \leq 2\mathcal{L}^{(t)} + 2\|\mathbf{x}^\star_\eta - \mathbf{x}^\star\|^2. \tag{40}$$

Theorem 6 ensures that $\mathcal{L}^{(t)}$ becomes small, and so we are left with bounding the distance between $\mathbf{x}^\star_\eta$ and $\mathbf{x}^\star$.

1 - Distance to the global minimizer. We define $\overline{\mathbf{x}^\star_\eta} = \mathbb{1}\mathbb{1}^\top \mathbf{x}^\star_\eta/n$. Using the fact that both $\overline{\mathbf{x}^\star_\eta}$ and $\mathbf{x}^\star$ are at consensus, and $\mathbb{1}^\top \nabla F(\mathbf{x}^\star_\eta) = 0$ (immediate from (23)), we write:

$$\begin{aligned} D_F(\mathbf{x}^\star, \mathbf{x}^\star_\eta) &= F(\mathbf{x}^\star) - F(\mathbf{x}^\star_\eta) - \nabla F(\mathbf{x}^\star_\eta)^\top(\mathbf{x}^\star - \mathbf{x}^\star_\eta) \\ &= F(\overline{\mathbf{x}^\star_\eta}) - F(\mathbf{x}^\star_\eta) - \nabla F(\mathbf{x}^\star_\eta)^\top(\overline{\mathbf{x}^\star_\eta} - \mathbf{x}^\star_\eta) + F(\mathbf{x}^\star) - F(\overline{\mathbf{x}^\star_\eta}) \\ &\leq D_F(\overline{\mathbf{x}^\star_\eta}, \mathbf{x}^\star_\eta), \end{aligned} \tag{41}$$

where the last line comes from the fact that $\mathbf{x}^\star$ is the minimizer of F on the consensus space. Therefore:

$$\begin{aligned} \|\mathbf{x}^\star_\eta - \mathbf{x}^\star\|^2 &= \|\overline{\mathbf{x}^\star_\eta} - \mathbf{x}^\star\|^2 + \|\mathbf{x}^\star_\eta - \overline{\mathbf{x}^\star_\eta}\|^2 \\ &\leq \frac{1}{\mu} D_F(\mathbf{x}^\star, \mathbf{x}^\star_\eta) + \|\mathbf{x}^\star_\eta - \overline{\mathbf{x}^\star_\eta}\|^2 \\ &\leq \frac{1}{\mu} D_F(\overline{\mathbf{x}^\star_\eta}, \mathbf{x}^\star_\eta) + \|\mathbf{x}^\star_\eta - \overline{\mathbf{x}^\star_\eta}\|^2 \\ &\leq \left(1 + \frac{L}{\mu}\right)\|\overline{\mathbf{x}^\star_\eta} - \mathbf{x}^\star_\eta\|^2 = \eta^2\left(1 + \frac{L}{\mu}\right)\|\mathbf{L}_{\mathbf{W}}{}^\dagger \nabla F(\mathbf{x}^\star_\eta)\|^2. \end{aligned}$$

Note that the result depends on the heterogeneity pattern of the gradients at the fixed point, and might be bounded (and even small) even when $\mathbf{W}$ has no spectral gap. However, this quantity is proportional to the squared inverse spectral gap in the worst case.

2 - Monotonicity in η. We now prove that $\|\nabla F(\mathbf{x}^\star_\eta)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}$ decreases when η increases, and so is maximal for $\eta = 0$, corresponding to $\mathbf{x}^\star_\eta = \mathbf{x}^\star$. More specifically:

$$\frac{\mathbf{d}\|\nabla F(\mathbf{x}^\star_\eta)\|^2_{\mathbf{L}_{\mathbf{W}}{}^\dagger}}{\mathbf{d}\eta} = \frac{\mathbf{d}\left[\eta^{-2}\|\mathbf{x}^\star_\eta\|^2_{\mathbf{L}_{\mathbf{W}}}\right]}{\mathbf{d}\eta} = -\frac{2\|\mathbf{x}^\star_\eta\|^2_{\mathbf{L}_{\mathbf{W}}}}{\eta^3} + 2\eta^{-2}(\mathbf{x}^\star_\eta)^\top \mathbf{L}_{\mathbf{W}} \frac{\mathbf{d}\mathbf{x}^\star_\eta}{\mathbf{d}\eta}$$

Differentiating the fixed-point conditions, we obtain that

$$\eta\nabla^2 F(\mathbf{x}_\eta^\star)\frac{\mathbf{d}\mathbf{x}_\eta^\star}{\mathbf{d}\eta} + \nabla F(\mathbf{x}_\eta^\star) + \mathbf{L}_\mathbf{W}\frac{\mathbf{d}\mathbf{x}_\eta^\star}{\mathbf{d}\eta} = 0, \tag{42}$$

so that:

$$\frac{\mathbf{d}\mathbf{x}_\eta^\star}{\mathbf{d}\eta} = -\left(\eta\nabla^2 F(\mathbf{x}_\eta^\star) + \mathbf{L}_\mathbf{W}\right)^{-1}\nabla F(\mathbf{x}_\eta^\star) = \eta^{-1}\left(\eta\nabla^2 F(\mathbf{x}_\eta^\star) + \mathbf{L}_\mathbf{W}\right)^{-1}\mathbf{L}_\mathbf{W}\mathbf{x}_\eta^\star. \tag{43}$$

Plugging this into the previous expression and using that $\nabla^2 F(\mathbf{x}_\eta^\star)$ is positive semi-definite, we obtain:

$$\begin{aligned}\frac{\mathbf{d}\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}}{\mathbf{d}\eta} &= -\frac{2}{\eta^3}(\mathbf{x}_\eta^\star)^\top\left[\mathbf{L}_\mathbf{W} - \mathbf{L}_\mathbf{W}\left(\mathbf{L}_\mathbf{W} + \eta\nabla^2 F(\mathbf{x}_\eta^\star)\right)^{-1}\mathbf{L}_\mathbf{W}\right]\mathbf{x}_\eta^\star \\ &\leq -\frac{2}{\eta^3}(\mathbf{x}_\eta^\star)^\top\left[\mathbf{L}_\mathbf{W} - \mathbf{L}_\mathbf{W}\mathbf{L}_\mathbf{W}^\dagger\mathbf{L}_\mathbf{W}\right]\mathbf{x}_\eta^\star = 0.\end{aligned}$$

3 - Getting rid of $\mathbf{x}_\eta^\star$. By definition of p, we can write:

$$\|\mathbf{L}_\mathbf{W}^\dagger\nabla F(\mathbf{x}_\eta^\star)\|^2 \leq p^{-1}\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger} \leq p^{-1}\|\nabla F(\mathbf{x}^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}. \tag{44}$$

Note that we have to bound this constant p in order to use the monotonicity in η of $\|\nabla F(\mathbf{x}_\eta^\star)\|^2_{\mathbf{L}_\mathbf{W}^\dagger}$ since this result does not hold for $\|\mathbf{L}_\mathbf{W}^\dagger\nabla F(\mathbf{x}_\eta^\star)\|^2$. For the variance, we write that:

$$\begin{aligned}\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right] &\leq 3\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F_{\xi_t}(\mathbf{x}^\star)\|^2\right] \\ &\quad + 3\,\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}^\star) - \nabla F(\mathbf{x}^\star)\|^2\right] + 3\|\nabla F(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}^\star)\|^2 \\ &\leq 3\sigma^2_{\mathbf{M},\star} + 3\left(\zeta + L\right)D_F(\mathbf{x}^\star, \mathbf{x}_\eta^\star).\end{aligned}$$

From here, we use Equation (41) and obtain that:

$$\mathbb{E}\left[\|\nabla F_{\xi_t}(\mathbf{x}_\eta^\star) - \nabla F(\mathbf{x}_\eta^\star)\|^2\right] \leq 3\sigma^2_{\mathbf{M},\star} + 3L\left(\zeta + L\right)\eta^2\|\mathbf{L}_\mathbf{W}^\dagger\nabla F(\mathbf{x}_\eta^\star)\|^2. \tag{45}$$

To obtain the final result, we use that $\eta(n_\mathbf{W}(\gamma)^{-1} + \omega)(\zeta + L) \leq 1/4$ thanks to the conditions on the learning rate.

■

4.3.3 Comparison with existing work.

Expressed in the form of Koloskova et al. (2020), we can summarize the previous corollaries into the following result by taking either η as the largest possible constant (as indicated in Corollary 8) or $\eta = \tilde{O}(1/(\mu T))$. Here, $\tilde{O}$ denotes inequality up to logarithmic factors, and recall that $\|\mathbf{x}^{(t)} - \mathbf{x}^\star\|^2_\mathbf{M} \geq \|\overline{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2$. We recall that L is the smoothness of the global objective f, ζ is the smoothness of the stochastic functions f_ξ, μ is the strong convexity parameter, d_s is the spectral dimension of the gossip matrix $\mathbf{W}$ (and we assume $d_s > 2$) and c_s is the associated constant.

Corollary 10 (Final result.) *Under the same assumptions as Corollary 8, there exists a choice of learning rate (and, equivalently, of decay parameters γ^*_{large} and γ^*_{small}) such that the expected squared distance to the global optimum after T steps of D-SGD $\|\overline{\mathbf{x}}^{(t)} - \mathbf{x}^\star\|^2$ is of order:*

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu^2 T n_{\mathbf{W}}(\gamma^*_{\text{small}})} + \frac{L\Delta^2_{\mathbf{W}}}{\mu^3 p T^2} + \exp\left[-n_{\mathbf{W}}(\gamma^*_{\text{large}})\frac{\mu}{\zeta}T\right]\right), \tag{46}$$

where $\Delta^2_{\mathbf{W}}$ and p are defined in Theorem 9, and $\overline{\mathbf{x}}^{(t)}$ is the average parameter. The optimal effective number of neighbors in respectively the small and large learning rate settings are:

$$n_{\mathbf{W}}(\gamma^*_{\text{small}}) = \min\left(\sqrt{\frac{d_s T}{L c_s}}, n\right) \text{ and } n_{\mathbf{W}}(\gamma^*_{\text{large}}) = \min\left(\left(\frac{c_s d_s \zeta}{L}\right)^{\frac{1}{3}}, n\right). \tag{47}$$

This result can be contrasted with the result from Koloskova et al. (2020), which writes:

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu^2 T}\left[\frac{1}{n} + \frac{L}{\mu(1-\lambda_2(\mathbf{W}))T}\right] + \frac{L\Delta^2}{\mu^3(1-\lambda_2(\mathbf{W}))^2 T^2} + \exp\left[-\frac{\mu}{(1-\lambda_2(\mathbf{W}))\zeta}T\right]\right), \tag{48}$$

We can now make the following observations.

Scheduling the learning rate. Here, the learning rate is either chosen as $\eta_{\text{large}} = n_{\mathbf{W}}(\gamma^*_{\text{large}})/\zeta$, or as $\eta_{\text{small}} = \tilde{O}((\mu T)^{-1})$. In practice, one would start with the large learning rate, and switching to η when training does not improve anymore (heterogeneity/variance terms dominate).

Exponential decrease term. We first show a significant improvement in the exponential decrease term. Indeed, $n_{\mathbf{W}}(\gamma^*_{\text{large}})/(1-\lambda_2(\mathbf{W}))$, the ratio between the largest learning rate permitted in our analysis versus existing ones, is always large since $n_{\mathbf{W}}(\gamma^*_{\text{large}}) \geq 1$ and $1-\lambda_2(\mathbf{W}) \leq 1$. Besides, the exponential decrease term is no longer affected by the spectral gap in our analysis, which only affects how big $n_{\mathbf{W}}(\gamma)$ can be. This improvement holds even when $\zeta = L$ (in this case $n_{\mathbf{W}}(\gamma) = 1$ is enough), and is due to the fact that *heterogeneity only affects lower-order terms*, so that when cooperation brings nothing it doesn't hurt convergence either.

Impact of heterogeneity. The improvement in the heterogeneous case does not depend on some γ, and relies on bounding heterogeneity in a non-worst case fashion. Indeed, $\zeta_{\mathbf{W}}$ and p capture the interplay between how heterogeneity is distributed among nodes, and the actual topology of the graph. Note that this does not contradict the lower bound from Koloskova et al. (2020), since $\Delta^2_{\mathbf{W}}/p = \Delta^2/(1-\lambda_2(\mathbf{W}))^2$ in the worst case. In the worst case, the heterogeneity pattern of $\nabla F(\mathbf{x}^\star)$ is aligned with the smallest eigenvalue of $\mathbf{L}_{\mathbf{W}}$, *i.e.*, very distant nodes have very different objectives. The quantity p, however, gives more fine-grained bounds that depend on the actual heterogeneity pattern in general.

Variance term. One key difference between the analyses is on the variance term that involves σ^2. Both analyses depend on the variance of a single node, $\sigma^2/(\mu T)$, which is

then multiplied by a 'variance reduction' term. In both cases, this term is of the form $n_{\mathbf{W}}(\gamma)^{-1}+\eta L\beta(\gamma)^{-1}$. However, the standard analysis implicitly use $\gamma = 1$, and so $n_{\mathbf{W}}(\gamma) = n$, and $\beta(\gamma) = 1 - \lambda_2(\mathbf{W})$. Then, the form from (48) follows from taking $\eta = \tilde{O}(1/(\mu T))$. Our analysis on the other hands relies on tuning γ such that $n_{\mathbf{W}}(\gamma)^{-1} + \eta L\beta(\gamma)^{-1}$ is the smallest possible, and is therefore strictly better than just considering $\gamma = 1$. Assuming a given spectral dimension $d_s > 2$ for the graph leads to (46), but any assumption that precisely relates $n_{\mathbf{W}}(\gamma)$ and γ would allow getting similar results.

While the $\tilde{O}(T^{-2})$ in the variance term of Koloskova et al. (2020) seems better than our $\tilde{O}(T^{-3/2})$ term, this is misleading because constants are very important in this case. Our rate is optimized by over γ, which accounts for the fact that *if* the $\tilde{O}(T^{-2})$ term dominates, then it is better to just consider a smaller neighborhood. In that case, we would not benefit from n^{-1} variance reduction anyway. Our result optimally balances the two variance terms from (48) instead. Thanks to this balancing, we obtain that in graphs of spectral dimension $d_s > 2$, the variance decreases as $\tilde{O}(T^{-\frac{3}{2}})$ with a learning rate of $\tilde{O}(T^{-1})$ due to the combined effect of a smaller learning rate and adding more effective neighbors. In finite graphs, this effect caps at $n_{\mathbf{W}}(\gamma) = n$.

Finally, note that our analysis and the analysis of Koloskova et al. (2020) allow for different generalizations of the standard framework: our analysis applies to arbitrarily large (infinite) graphs, while Koloskova et al. (2020) can handle time-varying graphs with weak (multi-round) connectivity assumptions.

5. Empirical relevance in deep learning

While the theoretical results in this paper are for convex functions, the initial motivation for this work comes from observations in deep learning. First, it is crucial in deep learning to use a large learning rate in the initial phase of training (Li et al., 2019). Contrary to what current theory prescribes, we do not use smaller learning rates in decentralized optimization than when training alone (even when data is heterogeneous.) And second, we find that the spectral gap of a topology is not predictive of the performance of that topology in deep learning experiments.

In this section, we experiment with a variety of 32-worker topologies on Cifar-10 (Krizhevsky et al.) with a VGG-11 model (Simonyan and Zisserman, 2015). Like other recent works (Lin et al., 2021; Vogels et al., 2021), we opt for this older model, because it does not include BatchNorm (Ioffe and Szegedy, 2015) which forms an orthogonal challenge for decentralized SGD. Please refer to Appendix E of (Vogels et al., 2022) for full details on the experimental setup. Our set of topologies includes regular graphs like rings and toruses, but also irregular graphs such as a binary tree (Vogels et al., 2021) and social network Davis et al. (1930), and a time-varying exponential scheme (Assran et al., 2019). We focus on the initial phase of training, 25k steps in our case, where both train and test loss converge close to linearly. Using a large learning rate in this phase is found to be important for good generalization (Li et al., 2019).

Figure 4 shows the loss reached after the first 2.5k SGD steps for all topologies and for a dense grid of learning rates. The curves have the same global structure as those for isotropic quadratics Figure 1: (sparse) averaging yields a small increase in speed for small learning rates, but a large gain over training alone comes from being able to increase the learning

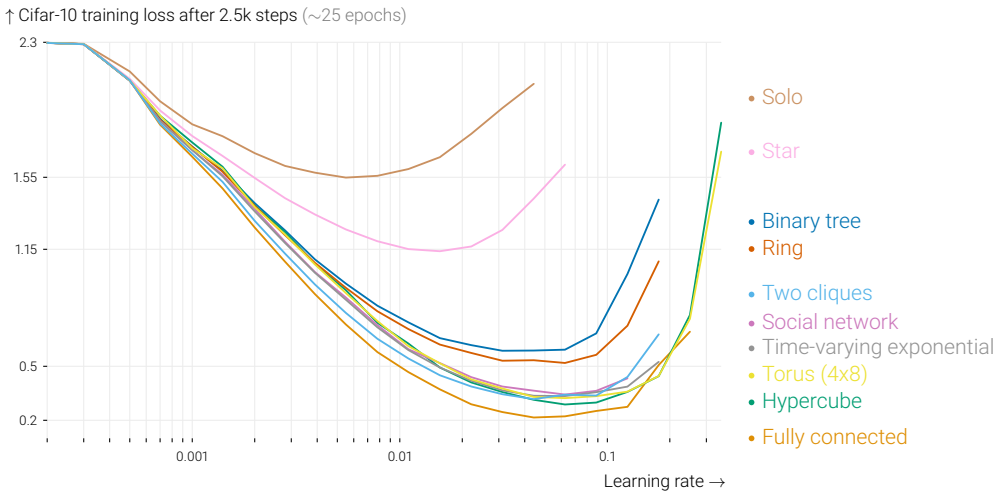

Figure 4: Training loss reached after 2.5k SGD steps with a variety of graph topologies. In all cases, averaging yields a small increase in speed for small learning rates, but a large gain over training alone comes from being able to increase the learning rate. While the star has a better spectral gap (0.031) than the ring (0.013), it performs worse, and does not allow large learning rates. For reference, similar curves for fully-connected graphs of varying sizes are in the appendix of Vogels et al. (2022).

rate. The best schemes support almost the same learning rate as 32 fully-connected workers, and get close in performance.

We also find that the random walks introduced in Section 3.1 are a good model for variance between workers in deep learning. Figure 5 shows the empirical covariance between the workers after 100 SGD steps. Just like for isotropic quadratics, the covariance is accurately modeled by the covariance in the random walk process for a certain decay rate γ.

Finally, we observe that the effective number of neighbors computed by the variance reduction in a random walk (Section 3.1) accurately describes the relative performance under tuned learning rates of graph topologies on our task, including for irregular and time-varying topologies. This is in contrast to the topology's spectral gaps, which we find to be not predictive. We fit a decay rate $\gamma = 0.951$ that seems to capture the specifics of our problem, and show the correlation in Figure 6.

Appendix F of (Vogels et al., 2022) replicates the same experiments in a different setting. There, we use larger graphs (of 64 workers), a different model and data set (an MLP on Fashion MNIST Xiao et al. (2017)), and no momentum or weight decay. The results in this setting are qualitatively comparable to the ones presented above.

6. Conclusion

We have shown that the sparse averaging in decentralized learning allows larger learning rates to be used, and that it speeds up training. With the optimal large learning rate, the workers' models are not guaranteed to remain close to their global average. Enforcing global consensus is often unnecessary and the small learning rates it requires can be counter-productive. Indeed,

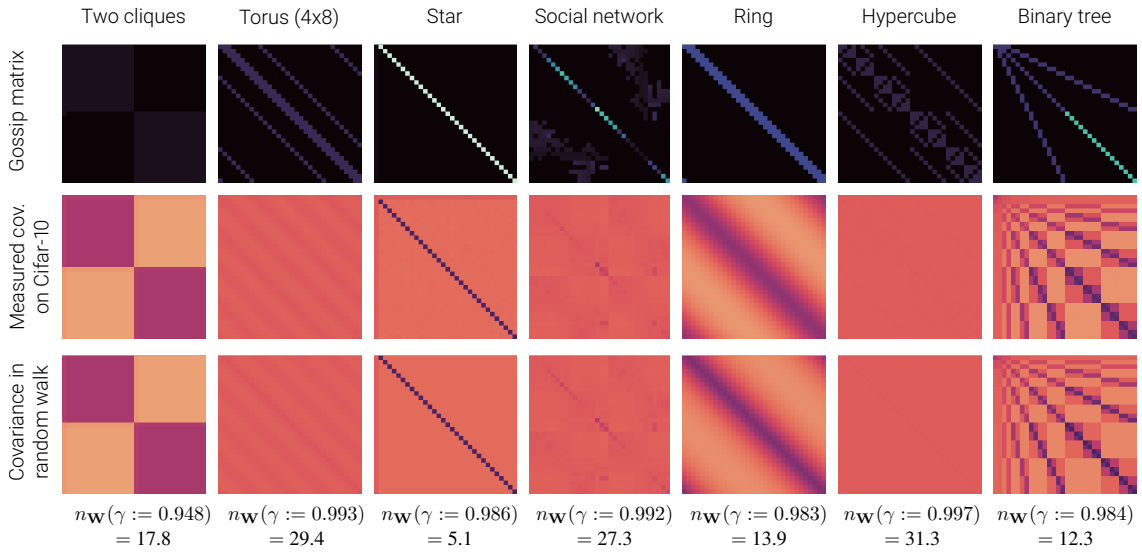

Figure 5: Measured covariance in Cifar-10 (second row) between workers using various graphs (top row). After 10 epochs, we store a checkpoint of the model and train repeatedly for 100 SGD steps, yielding 100 models for 32 workers. We show normalized covariance matrices between the workers. These are very well approximated by the covariance in the random walk process of Section 3.1 (third row). We print the fitted decay parameters and corresponding 'effective number of neighbors'.

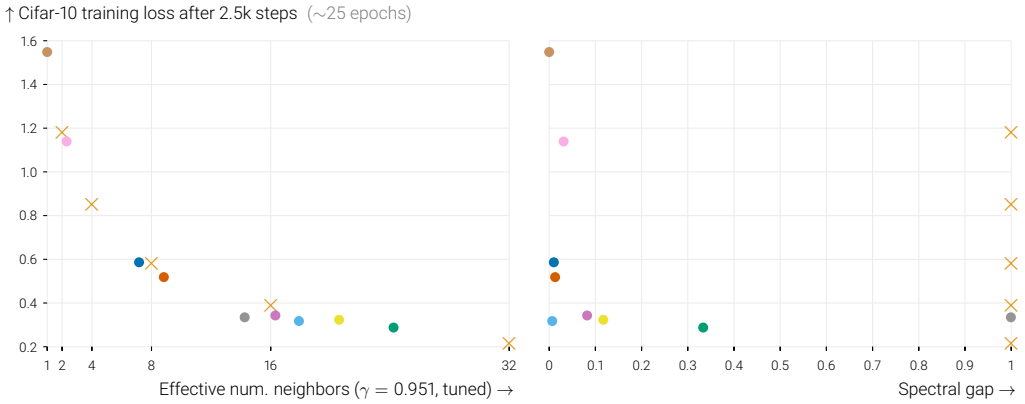

Figure 6: Cifar-10 training loss after 2.5k steps for all studied topologies with their optimal learning rates. Colors match Figure 4, and × indicates fully-connected graphs with varying number of workers. After fitting a decay parameter $\gamma = 0.951$ that captures problem specifics, the effective number of neighbors (left) as measured by variance reduction in a random walk (like in Section 3.1) explains the relative performance of these graphs much better than the spectral gap of these topologies (right).

models *do* remain close to some local average in a weighted neighborhood around them even with high learning rates. The workers benefit from a number of 'effective neighbors', potentially smaller than the whole graph, that allow them to use larger learning rates while retaining sufficient consensus within the 'local neighborhood'.

Similar insights apply when nodes have heterogeneous local functions: there is no need to enforce global averaging over the whole network when heterogeneity is small across local neighborhoods. Besides, there is no need to compensate for heterogeneity in the early phases of training, when models are all far from the global optimum.

Based on our insights, we encourage practitioners of sparse distributed learning algorithms to look beyond the spectral gap of graph topologies, and to investigate the actual 'effective number of neighbors' that is used. We also hope that our insights motivate theoreticians to be mindful of assumptions that artificially limit the learning rate, even though they are tight in worst cases. Indeed, the spectral gap is omnipresent in the decentralized litterature, which sometimes hides some subtle phenomena such as the superlinear decrease of the variance in the learning rate, that we highlight.

We show experimentally that our conclusions hold in deep learning, but extending our theory to the non-convex setting is an important open direction that could reveal interesting new phenomena. Another interesting direction would be to better understand (beyond the worst-case) the effective number of neighbors for irregular graphs.

Acknowledgments and Disclosure of Funding

This project was supported by SNSF grant 200020_200342.

We thank Lie He for valuable conversations and for identifying the discrepancy between a topology's spectral gap and its empirical performance.

We also thank Raphaël Berthier for helpful discussions that allowed us to clarify the links between effective number of neighbors and spectral dimension.

We also thank Aditya Vardhan Varre, Yatin Dandi and Mathieu Even for their feedback on the manuscript.

References

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Michael G. Rabbat. Stochastic gradient push for distributed deep learning. In *Proc. ICML*, volume 97, pages 344–353, 2019.

Raphaël Berthier. Analysis and acceleration of gradient descents and gossip algorithms. *PhD Thesis, Université Paris Sciences & Lettres*, 2021.

Raphaël Berthier, Francis R. Bach, and Pierre Gaillard. Accelerated gossip in networks of given dimension using jacobi polynomial iterations. *SIAM J. Math. Data Sci.*, 2(1):24–47, 2020.

Yatin Dandi, Anastasia Koloskova, Martin Jaggi, and Sebastian U. Stich. Data-heterogeneity-aware mixing for decentralized learning. *CoRR*, abs/2204.06477, 2022.

Allison Davis, Burleigh Bradford Gardner, and Mary R Gardner. *Deep South: A social anthropological study of caste and class*. Univ of South Carolina Press, 1930.

Mathieu Even, Hadrien Hendrikx, and Laurent Massoulie. Decentralized optimization with heterogeneous delays: a continuous-time approach. *arXiv preprint arXiv:2106.03585*, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, volume 37, pages 448–456, 2015.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proc. ICML*, volume 119, pages 5381–5393, 2020.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (Canadian Institute for Advanced Research).

B. Le Bars, Aurélien Bellet, Marc Tommasi, and Anne-Marie Kermarrec. Yes, topology matters in decentralized optimization: Refined convergence and topology learning under heterogeneous data. *CoRR*, abs/2204.04452, 2022.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *NeurIPS*, pages 11669–11680, 2019.

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 *NeurIPS*, pages 5330–5340, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *Proc. ICML*, volume 80, pages 3049–3058, 2018.

Tao Lin, Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Quasi-global momentum: Accelerating decentralized deep learning on heterogeneous data. In *Proc. ICML*, volume 139, pages 6654–6665, 2021.

Yucheng Lu and Christopher De Sa. Optimal complexity in decentralized training. In *Proc. ICML*, volume 139, pages 7111–7123, 2021.

Giovanni Neglia, Chuan Xu, Don Towsley, and Gianmarco Calbi. Decentralized gradient methods: does topology matter? In *AISTATS,*, volume 108, pages 2348–2358, 2020.

Dominic Richards and Patrick Rebeschini. Optimal statistical rates for decentralised non-parametric regression with linear speed-up. In *NeurIPS*, pages 1214–1225, 2019.

Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *J. Mach. Learn. Res.*, 21:34:1–34:44, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. d^2: Decentralized training over decentralized data. In *Proc. ICML*, volume 80, pages 4855–4863, 2018.

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 *NeurIPS*, pages 28004–28015, 2021.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap: the role of topology in decentralized learning. In *NeurIPS*, 2022.

Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. MATCHA: speeding up decentralized SGD via matching decomposition sampling. *CoRR*, abs/1905.09435, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. In *NeurIPS*, pages 13975–13987, 2021.