Decentralized gradient methods: does topology matter?

Giovanni Neglia, Chuan Xu, Don Towsley, Gianmarco Calbi

Decentralized gradient methods: does topology matter?

Giovanni Neglia
Inria

Chuan Xu
Inria

Don Towsley
UMass

Gianmarco Calbi
Inria

Abstract

Consensus-based distributed optimization methods have recently been advocated as alternatives to parameter server and ring all-reduce paradigms for large scale training of machine learning models. In this case, each worker maintains a local estimate of the optimal parameter vector and iteratively updates it by averaging the estimates obtained from its neighbors, and applying a correction on the basis of its local dataset. While theoretical results suggest that worker communication topology should have strong impact on the number of epochs needed to converge, previous experiments have shown the opposite conclusion. This paper sheds lights on this apparent contradiction and show how sparse topologies can lead to faster convergence even in the absence of communication delays.

1 INTRODUCTION

In 2014, Google's Sybil machine learning (ML) platform was processing hundreds of terabytes through thousands of cores to train models with hundreds of billions of parameters (Canini et al., 2014). At this scale, no single machine can solve these problems in a timely manner, and, as time goes on, the need for efficient distributed solutions becomes even more urgent. For example, recent experiments in (Young et al., 2017) rely on more than 10^4 computing nodes to iteratively improve the (hyper)parameters of a deep neural network.

The example in (Young et al., 2017) is typical of a large class of iterative ML distributed algorithms. Such algorithms begin with a guess of an optimal vector of parameters and proceed through multiple iterations over the input data to improve the solution. The process evolves in a data-parallel manner: input data is divided among worker threads. Currently, two communication paradigms are commonly used to coordinate the different workers (Google I/O, 2018): parameter server and ring all-reduce. Both paradigms are natively supported by TensorFlow (Abadi et al., 2016).

In the first case, a stateful parameter server (PS) (Smola and Narayanamurthy, 2010) maintains the current version of the model parameters. Workers use locally available versions of the model to compute "delta" updates of the parameters (e.g. through a gradient descent step). Updates are then aggregated by the parameter server and combined with its current state to produce a new estimate of the optimal parameter vector.

As an alternative, it is possible to remove the PS, by letting each worker aggregate the inputs of all other workers through the ring all-reduce algorithm (Gibiansky, 2017). With M workers, each aggregation phase requires $2(M-1)$ communication steps with $\mathcal{O}(1)$ data transmitted per worker. There are many efficient low level implementations of ring all-reduce, e.g. in NVIDIA's library NCCL.

We observe that both the PS and the ring all-reduce paradigms 1) maintain a unique candidate parameter vector at any given time and 2) rely *logically* on an all-to-all communication scheme.[1] Recently Lian et al. (2017, 2018) have promoted an alternative approach in the ML research community, where each worker 1) keeps updating a local version of the parameters and 2) broadcasts its updates only to a subset of nodes (its neighbors). This family of algorithms became originally popular in the control community, starting from the seminal work of Tsitsiklis et al. (1986) on distributed gradient methods. They are often referred to as *consensus-based distributed optimization methods*. Experimental results in (Lian et al., 2017, 2018;

Preliminary work. Under review by AISTATS 2020. Do not distribute.

[1] Each node needs to receive the aggregate of all other nodes' updates to move to the next iteration. Aggregation is performed by the PS or along the ring through multiple rounds.

Luo et al., 2019) show that

1. in terms of number of epochs, the convergence speed is almost the same when the communication topology is a ring or a clique, *contradicting* theoretical findings that predict convergence to be faster on a clique;

2. in terms of wall-clock time, convergence is faster for sparser topologies, an effect attributed in (Lian et al., 2017) to smaller worker communication costs.

In particular, Luo et al. (2019) summarize their findings as follows "*in theory, the bigger the spectral gap,* [i.e. the more connected the topology] *the fewer iterations it takes to converge. However, our experiments do not show a significant difference in the convergence rate w.r.t. iterations, even when spectral gaps are very dissimilar.*"

In this paper we contribute to a better understanding of the potential advantages of consensus-based gradient methods. In particular,

1. we present a refined convergence analysis that helps to explain the apparent contradiction among theoretical results and empirical observations,

2. we show that sparse topologies can speed-up wall-clock time convergence even when communication costs are negligible, because they intrinsically mitigate the straggler problem,

3. our experiments indicate that, under a realistic distribution of computation times, sparse topologies like rings and 3-degree expander graphs are the best practical choices.

The paper is organized as follows. Section 2 provides required background. Our theoretical analysis of the effect of communication topology is in Sect. 3. Experiment results in Sect. 4 confirm our findings. Section 5 concludes the paper.

2 NOTATION AND BACKGROUND

The goal of supervised learning is to learn a function that maps an input to an output using S examples from a training dataset $\mathbb{S} = \{(\boldsymbol{x}^{(l)}, y^{(l)}), l = 1, \dots S\}$. Each example $(\boldsymbol{x}^{(l)}, y^{(l)})$ is a pair consisting of an input object $\boldsymbol{x}^{(l)}$ and an associated target value $y^{(l)}$. In order to find the best statistical model, ML techniques often find the set of parameters $\boldsymbol{w} \in \mathbb{R}^n$ that solves the following optimization problem:

$$\underset{\boldsymbol{w}}{\text{minimize}} \sum_{l=1}^{S} f(\boldsymbol{w}, \boldsymbol{x}^{(l)}, y^{(l)}) \tag{1}$$

where function $f(\boldsymbol{w}, \boldsymbol{x}^{(l)}, y^{(l)})$ represents the error the model commits on the l-th element of the dataset $\mathbb{S}$ when parameter vector $\boldsymbol{w}$ is used. The objective function may also include a regularization term that enforces some "simplicity" (e.g. sparseness) on $\boldsymbol{w}$; such a term is easily taken into account in our analysis.

Due to increases in available data and statistical model complexity, distributed solutions are often required to determine the parameter vector in a reasonable time. The dataset in this case is divided among M workers ($\mathbb{S} = \cup_{j=1}^{M} \mathbb{S}_j$), possibly with some overlap. For simplicity, we consider that all local datasets $\mathbb{S}_i$ have the same size. Problem (1) can be restated in an equivalent form as minimization of the sum of functions local to each node:

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad F(\boldsymbol{w}) = \sum_{j=1}^{M} F_j(\boldsymbol{w}), \tag{2}$$

where $F_j(\boldsymbol{w}) = \frac{1}{|\mathbb{S}_j|} \sum_{(\boldsymbol{x}^{(l)}, y^{(l)}) \in \mathbb{S}_j} f(\boldsymbol{w}, \boldsymbol{x}^{(l)}, y^{(l)})$.

The distributed system can be represented by a directed *dataflow graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \dots M\}$ is the set of nodes (the workers) and an edge $(i, j) \in \mathcal{E}$ indicates that, at each iteration, node j waits for updates from node i for the previous iteration. We assume the graph is strongly connected. Let $N_j = \{i | (i, j) \in \mathcal{E}\}$ denote the in-neighborhood of node j, i.e. the set of predecessors of node j in $\mathcal{G}$. Each node j maintains a local estimate of the parameter vector $\boldsymbol{w}_j(k)$ and broadcasts it to its successors. The local estimate is updated as follows:

$$\boldsymbol{w}_j(k+1) = \sum_{i \in N_j \cup \{j\}} \boldsymbol{w}_i(k) A_{i,j} - \eta(k) \boldsymbol{g}_j(\boldsymbol{w}_j(k)). \tag{3}$$

The node computes a weighted average (consensus/gossip component) of the estimates of its neighbors and itself, and then corrects it taking into account a stochastic subgradient[2] $\boldsymbol{g}_j(\boldsymbol{w}_j(k))$ of its local function, i.e.

$$\boldsymbol{g}_j(\boldsymbol{w}_j(k)) = \frac{1}{B} \sum_{(\boldsymbol{x}^{(l)}, y^{(l)}) \in \xi_j(k)} \partial f(\boldsymbol{w}_j(k), \boldsymbol{x}^{(l)}, y^{(l)}),$$

where $\partial f(\boldsymbol{w}, \boldsymbol{x}^{(l)}, y^{(l)})$ denotes a subgradient of f with respect to $\boldsymbol{w}$, and $\xi_j(k)$ is a random minibatch of size B drawn from $\mathbb{S}_j$. Parameter $\eta(k) > 0$ is the (potentially time-varying) learning rate. $\boldsymbol{A} = (A_{i,j})$ is an

[2] With some abuse of notation, we indicate *a* subgradient in $\boldsymbol{w}$ as $\boldsymbol{g}(\boldsymbol{w})$, even if $\boldsymbol{g}$ is not a function.

$M \times M$ matrix of non-negative weights. We call $\boldsymbol{A}$ the *consensus matrix*.[3]

The operation of a synchronous PS or ring all-reduce is captured by (3) when the underlying graph $\mathcal{G}$ is a clique, $\boldsymbol{A} = \mathbf{1}\mathbf{1}^\intercal/M$, where $\mathbf{1}$ is the $M \times 1$ vector consisting of all ones, and $\boldsymbol{w}_i(0) = \boldsymbol{w}_j(0), \forall i, j \in \mathcal{V}$.

Under some standard technical conditions, (Nedić et al., 2018, Thm. 8) and (Duchi et al., 2012, Thm. 2) conclude, respectively for Distributed Subgradient Method (DSM) and for the Dual Averaging Distributed method, that the number of iterations K_ϵ needed to approximate the minimum objective function value by the desired error ϵ is

$$K_\epsilon \in \mathcal{O}\left(\frac{1}{\epsilon^2 \gamma(\boldsymbol{A})}\right), \tag{4}$$

where $0 \leq \gamma(\boldsymbol{A}) \leq 1$ is the spectral gap of the matrix $\boldsymbol{A}$, i.e. the difference between the moduli of the two largest eigenvalues of $\boldsymbol{A}$. The spectral gap quantifies how information flows in the network. In particular, the spectral gap is maximal for a clique with weights $A_{i,j} = 1/M$. Motivated by these convergence results, existing theoretically-oriented literature has concluded that a more connected network topology leads to faster convergence (Nedić et al., 2018; Duchi et al., 2012). But some recent experimental results report that consensus-based gradient methods achieve similar performance after the same number of iterations/epochs on topologies as different as rings and cliques. For example (Lian et al., 2017, Fig. 3) shows almost overlapping training losses for different ResNet architectures trained on CIFAR-10 with up to one hundred workers. (Luo et al., 2019, Fig. 20) and our experimental results in Sect. 4 confirm these findings.

Lian et al. (2017) provide a partial explanation for this insensitivity in their Corollary 2, showing that the convergence rate is topology-independent 1) after a large number of iterations ($\mathcal{O}(M^5/\gamma(A)^2)$), 2) for a vanishing learning rate, and 3) when the functions F_j are differentiable with Lipschitzian gradients. Under the additional hypothesis of strong convexity, Olshevsky et al. (2019) prove that topology insensitivity should manifest after $\mathcal{O}(M/\gamma(A)^2)$ iterations. These results do not explain why insensitivity is often observed in practice (as shown in (Lian et al., 2017, 2018; Luo et al., 2019)) 1) since the beginning of the training phase, 2) with constant learning rates, and 3) for non-differentiable machine learning models (e.g. neural networks). In the following section, we present a refined convergence analysis that explains when and why the effect of topology on the number of iterations needed to converge is weaker than what previously predicted.

[3] We are describing a synchronous DSM. The consistency model could be weaker, allowing node i to use older estimates from its neighbors (Li et al., 2014).

3 ANALYSIS

A less connected topology requires more iterations to achieve a given precision as indicated by (4). Our detailed analysis below shows that, when consensus-based optimization methods are used for ML training, the increase in the number of iterations is much less pronounced than previous studies predict. This is due to two different effects. First, consensus is affected only by variability in initial estimates and subgradients across nodes, and not by their absolute values. Second, certain configurations of initial estimates and subgradients are more difficult to achieve a consensus over, and would make the training highly dependent on the topology, but they are unlikely to be obtained by randomly partitioning the dataset.

Let n be the number of parameters of the model, and $\boldsymbol{W}(k)$ and $\boldsymbol{G}(k)$ be $n \times M$ matrices, whose columns are, respectively, node estimates $\boldsymbol{w}_1(k), \ldots, \boldsymbol{w}_M(k)$ and subgradients $\boldsymbol{g}_1(\boldsymbol{w}_1(k)), \ldots, \boldsymbol{g}_M(\boldsymbol{w}_M(k))$ at the completion of iteration k. Equation (3) can be rewritten in the form $\boldsymbol{W}(k+1) = \boldsymbol{W}(k)\boldsymbol{A} - \eta(k)\boldsymbol{G}(k)$, from which we obtain iteratively

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(0)\boldsymbol{A}^{k+1} - \sum_{h=0}^{k} \eta(h)\boldsymbol{G}(h)\boldsymbol{A}^{k-h}. \tag{5}$$

We make the following assumptions:[4]

A1 all functions F_i are convex,

A2 the set of (global) minimizers $\mathbb{W}^*$ is non-empty,

A3 graph $\mathcal{G}$ is strongly connected,

A4 matrix $\boldsymbol{A}$ is normal (i.e. $\boldsymbol{A}^\intercal\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\intercal$) and doubly stochastic,

A5 the squared Frobenius norm of subgradient matrix $\boldsymbol{G}(k)$ is bounded in expectation over the vector $\boldsymbol{\xi}$ of minibatches randomly drawn at nodes, i.e. there exists E, such that $\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{G}(k)\|_F^2\right] \leq E$.

Assumptions A1-A4 are standard ones in the related literature, see for example (Nedić and Ozdaglar, 2009; Duchi et al., 2012; Nedić et al., 2018). Assumption A5 imposes a bound on the (expected) *energy* of the subgradients, because $\|\boldsymbol{G}(k)\|_F^2 = \sum_j \|\boldsymbol{g}_j(\boldsymbol{w}_j(k))\|_2^2$. In the literature it is often replaced by the stronger

[4] Experiments in Sect. 4 show that our conclusions hold also when these assumptions are not satisfied.

requirement that the norm-2 of the subgradients $\boldsymbol{g}_j(\boldsymbol{w}_j(k))$ is bounded. Let $\Delta\boldsymbol{G}(k)$ denote the matrix $\boldsymbol{G}(k)-\boldsymbol{G}(k)\mathbf{1}\mathbf{1}^\intercal/M$, whose column j is the difference between subgradient $\boldsymbol{g}_j(\boldsymbol{w}_j(k))$ and the average of subgradients $\sum_{j=1}^M \boldsymbol{g}_j(\boldsymbol{w}_j(k))/M$. $\|\Delta\boldsymbol{G}(k)\|_F^2$ captures the variability in the subgradients. Assumption A5 also implies that there exist two constants $E_{\text{sp}} \leq E$ and $H \leq \sqrt{E}$ such that

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta\boldsymbol{G}(k)\|_F^2\right] \leq E_{\text{sp}}, \quad \|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}(k)]\|_F \leq H.$$

Similarly, let R denote the energy of the initial parameter vectors (or an upper bound for it), i.e. $R \triangleq \|\boldsymbol{W}(0)\|_F^2$. We also denote by R_{sp} the energy for the difference matrix $\Delta\boldsymbol{W}(0) \triangleq \boldsymbol{W}(0) - \boldsymbol{W}(0)\mathbf{1}\mathbf{1}^\intercal/M$, i.e. $R_{\text{sp}} \triangleq \|\Delta\boldsymbol{W}(0)\|_F^2$. R_{sp} captures the variability in initial estimates. It holds $R_{\text{sp}} \leq R$.

Because of Assumption A4, the consensus matrix has a spectral decomposition with orthogonal projectors $\boldsymbol{A} = \sum_{q=1}^Q \lambda_q \boldsymbol{P}_q$, where $\lambda_1, \ldots, \lambda_Q$ are the $Q \leq M$ distinct eigenvalues of $\boldsymbol{A}$, $\boldsymbol{P}_q$ is the orthogonal projector onto the nullspace of $\boldsymbol{A}-\lambda_q\boldsymbol{I}$ along the range of $\boldsymbol{A}-\lambda_q\boldsymbol{I}$. We assume that the eigenvalues are ordered so that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_Q|$. Assumptions A3 and A4 imply that $\lambda_1 = 1$, and $|\lambda_2| < 1$ (Authors, 2019, App. B). Finally, we define

$$\alpha \triangleq \begin{cases} \sqrt{\sum_{q=2}^Q e_q \left|\frac{\lambda_q}{\lambda_2}\right|^2}, & \text{if } \lambda_2 \neq 0, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

where e_q is an upper-bound for the normalized fraction of energy $\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta\boldsymbol{G}(k)\|_F^2\right]$ in the subspace defined by projector $\boldsymbol{P}_q$ (Authors, 2019, App. C). The quantity α can be interpreted as an effective bound for the fraction of the energy E_{sp} that falls in the subspace relative to the second largest eigenvalue λ_2.

We are now ready to introduce our main convergence result. We state it for the average model over nodes and time, i.e. for $\hat{\overline{\boldsymbol{w}}}(K-1) \triangleq \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{M}\sum_{i=1}^M \boldsymbol{w}_i(k)$. We have also derived a similar bound for the local time-average model at each node, i.e. for $\hat{\boldsymbol{w}}_i(K-1) \triangleq \frac{1}{K}\sum_{k=0}^{K-1}\boldsymbol{w}_i(k)$ (Authors, 2019, App. C.3).[5]

Proposition 3.1. *Under assumptions A1-A5 and that a constant learning rate $\eta(k) = \eta$ is used, an upper bound for the objective value at the end of the $(K-1)$th iteration is given by:*

$$\begin{aligned}\mathbb{E}\left[F\left(\hat{\overline{\boldsymbol{w}}}(K-1)\right)\right] - F^* \leq{}& \frac{M}{2\eta K}\text{dist}\left(\hat{\overline{\boldsymbol{w}}}(0), \mathbb{W}^*\right)^2 + \frac{\eta E}{2} \\ &+ 2H\sqrt{R_{sp}}\frac{\sqrt{M}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\ &+ 2\eta H\sqrt{E_{sp}}\left((1-\alpha)\frac{K-1}{K}\right. \\ &\left.+\frac{\alpha}{1-|\lambda_2|}\left(1-\frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right)\right).\end{aligned} \tag{7}$$

Here, $\text{dist}(\boldsymbol{x}, \mathbb{W}^*)$ denotes the Euclidean distance between vector $\boldsymbol{x}$ and set of global minimizers $\mathbb{W}^*$. The proof is in (Authors, 2019, App. C.1). The first two terms on the right hand side of (7) also appear when studying the convergence of centralized subgradient methods. The last two terms appear because of the distributed consensus component of the algorithm and depend on $|\lambda_2| < 1$. We observe that $1-|\lambda_2|$ is the spectral gap of $\boldsymbol{A}$. It measures how well connected the graph is. In particular, the larger the spectral gap (the smaller λ_2), the better the connectivity and the smaller the bound in (7).

From Proposition 3.1, we can derive a looser bound analogous to the bound for DSM in (Nedić and Ozdaglar, 2009). In fact, observing that $R_{\text{sp}} \leq R$, $E_{\text{sp}} \leq E$, $H \leq \sqrt{E}$, and $0 \leq \alpha \leq 1$, we can prove (Authors, 2019, App. C.2):

Corollary 3.2. *Under assumptions A1-A5 and that constant learning rate $\eta(k) = \eta$ is used, an upper bound for the objective value at the end of the $(K-1)$th iteration is given by:*

$$\begin{aligned}\mathbb{E}\left[F\left(\hat{\overline{\boldsymbol{w}}}(K-1)\right)\right] - F^* \leq{}& \frac{M}{2\eta K}\text{dist}\left(\hat{\overline{\boldsymbol{w}}}(0), \mathbb{W}^*\right)^2 + \frac{\eta E}{2} \\ &+ 2\sqrt{E}\sqrt{R}\frac{\sqrt{M}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\ &+ 2\eta E\frac{1}{1-|\lambda_2|}\left(1-\frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right).\end{aligned} \tag{8}$$

In particular, if workers compute full-batch subgradients and the 2-norm of subgradients of functions F_i is bounded by a constant L, we obtain:

$$\begin{aligned}F\left(\hat{\overline{\boldsymbol{w}}}(K-1)\right) - F^* \leq{}& \frac{M}{2\eta K}\text{dist}\left(\hat{\overline{\boldsymbol{w}}}(0), \mathbb{W}^*\right)^2 + \frac{\eta M L^2}{2} \\ &+ 2L\sqrt{R}\frac{M}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\ &+ 2\eta L^2\frac{M}{1-|\lambda_2|}\left(1-\frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right).\end{aligned} \tag{9}$$

When K is large enough, the fourth term in (8) and (9) is dominant, so that the error is essentially propor-

[5] For subgradient methods, convergence results are usually for the time-average model.

tional to $1/(1-|\lambda_2|)$. Note that the constant multiplying $1/(1-|\lambda_2|)$ in (8) is larger than the corresponding one in (7) by a factor

$$\beta \triangleq \frac{1}{\alpha} \times \frac{E}{\sqrt{E_{\text{sp}}}H} \tag{10}$$

The value β roughly indicates how much looser bound (8) is in comparison to bound (7).

Existing theoretical works like (Nedić and Ozdaglar, 2009; Duchi et al., 2012; Nedić et al., 2018) derived bounds similar to (9) and concluded then that one should select the learning rate proportional to $\sqrt{1-|\lambda_2|}$ to reduce the effect of topology. In particular, one obtains (4) when $\eta = \eta_0\sqrt{(1-|\lambda_2|)/K}$. Our bound (7) improves bound (8) by replacing R in the third terms of (8) by the smaller value R_{sp}, and $\sqrt{E}$ in the third and fourth terms by the smaller values H and $\sqrt{E_{\text{sp}}}$, and introducing the new coefficient α. We qualitatively describe the effect of these constants.

$\mathbf{R_{sp}}$ Bound (7) shows that the norm of the initial estimates (R) does not really matter, but rather variability among workers does. In particular, for ML computation we can make $\boldsymbol{w}_i(0) = \boldsymbol{w}_j(0)$ for each i and j, and then $R_{\text{sp}} = 0$, so that the third term in the RHS of (7) vanishes.

$\mathbf{E_{sp}}$, $\mathbf{H}$ For E_{sp}, considerations similar to those applying to R_{sp} hold. What matters is the variability of the subgradients. Assume that the dataset is replicated at each node and each node computes the subgradient over the full batch ($B = S$). In this case all subgradients would be equal, and $\|\Delta\boldsymbol{G}(k)\| = 0$, $E_{\text{sp}} = 0$, and the fourth term would also vanish. This corresponds to the fact that, when initial parameter vectors, as well as local functions, are the same, the parameter vectors are equal at any iteration k and the system evolves exactly as it would under a centralized subgradient method. In general, local subgradients can be expected to be close (and $E \gg E_{\text{sp}}$), if 1) local datasets are representative of the entire dataset (the dataset has been randomly split and $|\mathbb{S}_j| \gg M$), and 2) large batch sizes are used. On the other hand, when batch sizes are very small, one expects stochastic subgradients to be very noisy, and as a consequence the energy of the matrix $\boldsymbol{G}$ to be much larger than the energy of $\mathbb{E}_\xi[\boldsymbol{G}]$, so that $\sqrt{E} \gg H$. In both cases, we expect $E/(\sqrt{E_{\text{sp}}}H)$ to be large (in the first case because $\sqrt{E} \gg \sqrt{E_{\text{sp}}}$, and in the second because $\sqrt{E} \gg H$). We quantify these effects below.

$\boldsymbol{\alpha}$ From (5) we see that the effect of the subgradients is modulated by $\boldsymbol{A}^{k-h}$, that equals $\sum_{q=1}^{Q}\lambda_q^{k-h}\boldsymbol{P}_q$. The energy of the subgradients is spread across the different subspaces defined by the eigenvectors of $\boldsymbol{A}$. The classic bound (8) implicitly assumes that all energy falls in the subspace corresponding to λ_2 (this occurs if the row of the matrices $\boldsymbol{G}(k)$ are aligned with the second eigenvector). In reality, on average each subspace will only get $1/Q$-th of the total energy ($e_q \approx 1/Q$), and the energy in other subspaces will be averaged faster than what happens for the subspace corresponding to λ_2. $\alpha \leq 1$ quantifies this effect.

A toy example in (Authors, 2019, App. F) illustrates qualitatively these effects. Here we provide estimates for the expected values of E, E_{sp}, and H over all possible ways to distribute the dataset $\mathbb{S}$ randomly across the nodes. We reason as follows. For a given parameter vector $\boldsymbol{w}$, consider the set of subgradients at all dataset points , i.e. $\cup_{(\boldsymbol{x}^{(l)},y^{(l)})\in\mathbb{S}}\{\partial f(\boldsymbol{w},\boldsymbol{x}^{(l)},y^{(l)})\}$. The average subgradient over all datapoints is $\partial F(\boldsymbol{w})$. Let $\sigma^2(\boldsymbol{w})$ denote the trace of the covariance matrix of all subgradients. $\sigma^2(\boldsymbol{w})$ then equals the sum of the variances of all the components of the subgradients. We now distribute the dataset across M nodes replicating each point C times (where $C \in \{1, 2, \dots M\}$). We denote by $\mathbb{S}_C$ the expanded dataset. Finally, we let each node select a random minibatch from its local dataset and we denote by $\boldsymbol{G}$ the corresponding subgradient matrix.

Proposition 3.3. *Consider a uniform random permutation π of $\mathbb{S}_C$ with the constraint that C copies of the same point are placed at C different nodes. The following holds*

$$\begin{aligned}
&\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{G}\|_F^2\right]\right] = M\left(\|\partial F\|_2^2 + \frac{S-B}{B(S-1)}\sigma^2\right),\\
&\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta\boldsymbol{G}\|_F^2\right]\right] = \sigma^2\frac{MC(S-B)-CS+MB}{CB(S-1)},\\
&\mathbb{E}_\pi\left[\|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}]\|_F\right]\\
&\quad\in\left[\sqrt{M}\|\partial F\|_2, \sqrt{M}\sqrt{\|\partial F\|_2^2 + \frac{M-C}{C(S-1)}\sigma^2}\right].
\end{aligned} \tag{11}$$

We can use (11) to study how E, E_{sp}, and H vary with dataset size, batch size, and number of replicas, using the following approximations:

$$\begin{aligned}
\widehat{E} &= \mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{G}\|_F^2\right]\right],\\
\widehat{E}_{\text{sp}} &= \mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta\boldsymbol{G}\|_F^2\right]\right],\\
\widehat{H} &= \sqrt{M}\sqrt{\|\partial F\|_2^2 + \frac{M-C}{C(S-1)}\sigma^2}.
\end{aligned} \tag{12}$$

Figure 1 illustrates the ratio $\widehat{E}/(\sqrt{\widehat{E}_{\text{sp}}}\widehat{H})(=\beta\alpha)$ for a particular setting. It also highlights the two regimes discussed above: $\beta \approx 1/\alpha \times \sqrt{E/E_{\text{sp}}}$ for large batch

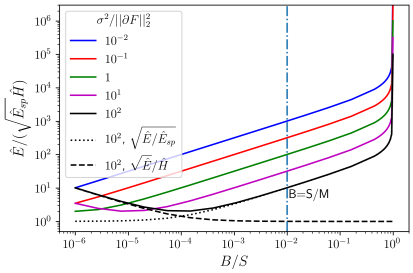


Figure 1: Estimate of $E/(\sqrt{E_{\text{sp}}}H)$ versus the relative batch size B/S for $M = 100$, $S = 10^6$, $C = M$, and different level of heterogeneity ($\sigma^2/\|\partial F\|_2^2$) of the dataset. Curves for $C = 1$ (in (Authors, 2019, App. E)) are very similar, but for the fact that the batch size can scale only up to S/M.

sizes and $\beta \approx 1/\alpha \times \sqrt{E}/H$ for small ones. As β indicates how much looser bound (8) is in comparison to bound (7), and $\beta > E/(\sqrt{E_{\text{sp}}}H)$, the figure shows that (8) may indeed overestimate the effect of the topology by many orders of magnitudes.

4 EXPERIMENTS

With our experiments we want to 1) evaluate the effect of topology on the number of epochs to converge, and in particular quantify E, E_{sp}, H, and α in practical ML problems, 2) evaluate the effect of topology on the convergence *time*. We considered three different optimization problems:

1. Minimization of mean squared error (MSE) for linear regression on the dataset "Relative location of CT slices on axial axis" from (uci; Graf et al., 2011). Convexity holds, but gradients are potentially unbounded.

2. Minimization of cross-entropy loss through a neural network with two convolutional layers on MNIST dataset (Lecun et al., 1998). Neither convexity, nor subgradient boundness hold.

3. Minimization of cross-entropy loss through ResNet18 neural network (He et al., 2016) on CIFAR-10 dataset (Krizhevsky, 2009). Neither convexity, nor subgradient boundness hold. Moreover, we employ local subgradients with classical momentum (Sutskever et al., 2013) (with coefficient 0.9).

We have developed an ad-hoc Python simulator that allows us to test clusters with a large number of nodes, as well as a distributed application using PyTorch MPI backend to run experiments on a real GPU cluster platform.[6] In general, datasets have been randomly split across the different workers without any replication ($C = 1$). For MNIST we have also considered a scenario with $M = 10$ workers, where each worker has received only images for a specific digit. A constant learning rate has been set using the configuration rule from (Smith, 2017) described in (Authors, 2019, App. G). Interestingly, for a given ML problem, when the dataset is split randomly, this procedure has led to choose the learning rate independently of the average node degree. The values selected are indicated in Table 1. Each node starts from the same model parameters ($R_{\text{sp}} = 0$) that have been initialized through PyTorch default functions. We report here a subset of all results, the others can be found in (Authors, 2019, App. G).

Table 1 shows values of $\sqrt{E/E_{\text{sp}}}$, $\sqrt{E}/H$, $1/\alpha$, and their product β for different problems and different settings.[7] E, E_{sp}, and H have been evaluated through empirical averages using the random mini-batches drawn at the first iteration. α is computed for an undirected ring topology. Remember that the value β (defined in (10)) indicates how much tighter the new bound (7) is in comparison to the classic one (8). We also use (12) to provide an estimate of β as follows $\widehat{\beta} = \widehat{E}/(\sqrt{\widehat{E}_{\text{sp}}}\widehat{H})$. The approximation is very accurate when the dataset is split randomly across the nodes. On the other hand, for MNIST, when all images for a given digit are assigned to the same node, local datasets are very different and approximations (12) are too crude (but our bound (7) still holds). Interestingly, β is dominated by different effects for the three datasets; similarity of local datasets for CT ($\sqrt{E/E_{\text{sp}}}$ dominates), energy spread over different eigenspaces for MNIST ($1/\alpha$ dominates), and very noisy stochastic subgradients for CIFAR ($\sqrt{E}/H$ dominates). This can be explained considering that, even if local datasets have similar sizes, they are statistically more different the more complex the model to train, i.e. the larger n.

From (7) and (8), we can also compute at which iteration the two bounds predict that the effect of the topology becomes significant, by identifying when the training loss difference between the clique and the ring accounts for a given percentage of the loss decrease over the entire training period. Figure 3 qualitatively illustrates the procedure.[8] These predictions are indi-

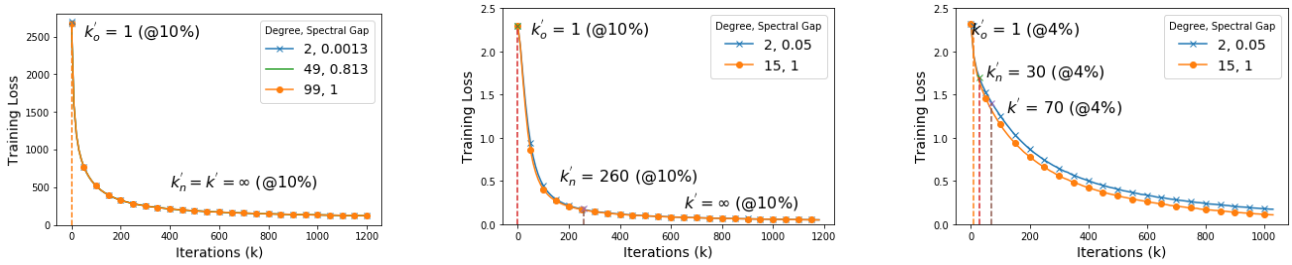
[6] The platform is composed of various types of GPUs, e.g., GeForce GTX 1080 Ti, GeForce GTX Titan X and Nvidia Tesla V100.

[7] Some additional experiments in (Authors, 2019, App. G) show that R_{sp} and R have a smaller effect on the bounds. This is due to the fact third term in (7) and in (8) converging to 0 when K diverges.

[8] In order to be able to compare the upper bounds (8) and (7) with the actual loss curves, we rescale them by

Table 1: Empirical estimation of E, E_{sp}, H, α on different ML problems and comparison of their joint effect (β) with the value $\widehat{\beta}$ predicted through (12). Number of iterations by which training losses for the ring and the clique differ by 4%, 10%, as predicted by the old bound (8), k'_o, by the new one (7), k'_n, and as measured in the experiment, k'. When a value exceeds the total number of iterations we ran (respectively 1200 for CT, 1190 for MNIST, and 1040 for CIFAR-10), we simply indicate it as ∞.

Dataset	Model	M	B	η	$\sqrt{E/E_{sp}}$	$\sqrt{E}/H$	$\frac{1}{\alpha}$	β	$\widehat{\beta}$	@4%			@10%		
										k'_o	k'_n	k'	k'_o	k'_n	k'
CT (S=52000)	Linear regr. n=384	16	128	0.0003	7.92	1.01	1.53	12.23	12.31	1	∞	∞	1	∞	∞
			3250		38.45	1.00	1.64	62.86	60.97	1	∞	∞	1	∞	∞
		100	128		7.75	1.01	1.54	12.05	11.56	1	10	∞	1	∞	∞
			520		15.58	1.00	1.51	23.60	22.96	1	17	∞	1	∞	∞
MNIST (S=60000)	2-conv layers n=431080	16	128	0.1	1.45	1.42	1.49	3.07	2.92	1	16	∞	1	72	∞
			500		2.15	1.14	1.53	3.75	3.71	1	22	40	1	260	∞
		64	128		1.41	1.42	1.51	3.02	3.03	1	10	∞	1	24	∞
split by digit		10	500	0.01	1.01	1.00	1.42	1.42	3.62	1	3	60	1	7	100
CIFAR-10 (S= 50000)	ResNet18 n=11173962	16	128	0.05	1.07	2.85	1.46	4.45	3.05	1	7	70	1	20	∞
			500		1.19	1.83	1.47	5.11	4.14	1	30	70	1	∞	∞



(a) CT: M=100, B = 128 (b) MNIST: M=16, B = 500 (c) CIFAR-10: M=16, B = 500

Figure 2: Effect of network connectivity (degree d) on the iterations to convergence.

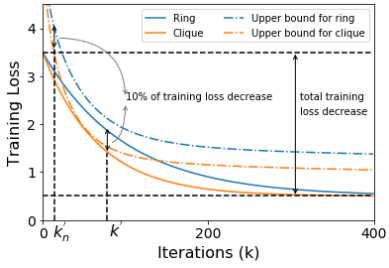


Figure 3: How to determine the number of iteration at which training loss for the clique and for the ring differs significantly.

cated in the last columns of Table 1 and are compared with the values observed in the experiments (k').

a factor determined so that the upper bound curve and the experimental one are tangent for the clique topology (Fig. 3). Moreover, once determined at which iteration rings and cliques should differ, we update the upper-bounds with new estimates for E, E_{sp}, H, R, and R_{sp} computed at this iteration, and check if they now predict a larger number of iterations.

We note that forecasts are very different, despite the fact that, in some settings, our bound is only 3 times tighter than the classic one. Bound (8) predicts that the training loss curves should differ by more than 10% since the first iteration ($k'_o = 1$). The new bound (7) correctly identifies that the topology's effect becomes evident later, sometimes beyond the total number of iterations performed in the experiment (in this case we indicate $k'_n = \infty$).

Figure 2 shows the training loss evolution $F(\hat{\boldsymbol{w}}(k))$ for specific settings (one for each ML problem) and different topologies, when the dataset is split randomly across the nodes. The behaviour is qualitatively similar to what observed in previous works (Lian et al., 2017, 2018; Luo et al., 2019); despite the remarkable difference in the level of connectivity (quantified also by the spectral gap), the curves are very close, sometimes indistinguishable.

Figure 4 shows the same plot for the case when MNIST images for the same digit have been assigned to the same node. In this case the local datasets are very different and $\sqrt{E/E_{\text{sp}}} \approx 1$; the topology has a re-

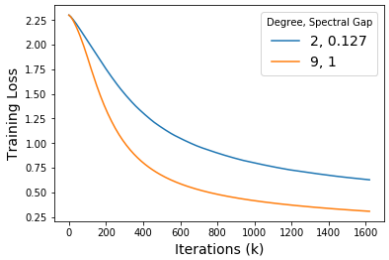


Figure 4: MNIST: 10 workers, B = 500

markable effect! This plot warns against extending the empirical finding in (Lian et al., 2017, 2018; Luo et al., 2019) to settings where local datasets can be highly different as it can be for example in the case of federated learning (Konecný et al., 2015).

The experiments above confirm that the communication topology has little influence on the number of *epochs* needed to converge (when local datasets are statistically similar). Our analysis reconciles (at least in part) theory and experiments by pushing farther the training epoch at which the effect of the topology should be evident.

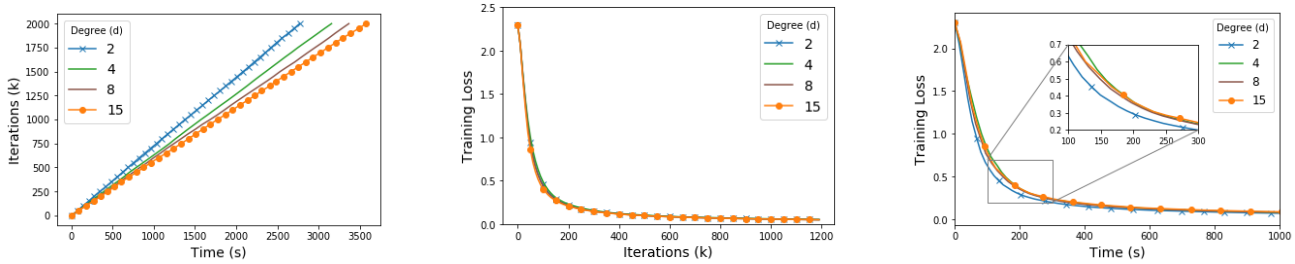
The conclusion about the role of the topology is radically different if one considers the *time* to converge. For example, Karakus et al. (2017); Luo et al. (2019) observe experimentally that sparse topologies can effectively reduce the convergence wall-clock time. A possible explanation is that each iteration is faster because less time is spent in the communication phase: the less connected the graph, the smaller the communication load at each node. Lian et al. (2017, 2018) advance this explanation to justify why DSM on ring-like topologies can converge faster than the centralized PS.

Here, we show that sparse topologies can speed-up wall-clock time convergence even when communication costs are negligible, because they intrinsically mitigate the effect of stragglers, i.e. tasks whose completion time can be occasionally much longer than its typical value. Transient slowdowns are common in computing systems (especially in shared ones) and have many causes, such as resource contention, background OS activities, garbage collection, and (for ML tasks) stopping criteria calculations. Stragglers can significantly reduce computation speed in a multi-machine setting (Ananthanarayanan et al., 2013; Karakus et al., 2017; Li et al., 2018). For consensus-based method, once can hope that, when the topology is sparse, a temporary straggler only slows down a limited number of nodes (its out-neighbors in $\mathcal{G}$), so that the system can still maintain a high throughput.

Neglia et al. (2019) have proposed approximate formulas to evaluate the throughput of distributed ML systems for some specific random distribution of the computation time (uniform, exponential, and Pareto). Here, we take a more practical approach. Our PyTorch-based distributed application allow us to simulate systems with arbitrary distributions of the computation times and communication delays. We have carried out experiments with zero communication delays (an ideal network) and two different empirical distributions for the computation time. One was obtained by running stochastic gradient descent on a production Spark cluster with sixteen servers, each with two 8-core Intel E5-2630 CPUs running at 2.40GHz. The other was extracted from ASCI-Q super-computer traces (Petrini et al., 2003, Fig. 4). Figure 5 shows the effect of topology connectivity on the convergence time for a MNIST experiment with Spark computation distribution. The number of iterations completed per node grows faster the less connected the topology (Fig. 5 (a)). As the training loss is almost independent of the topology (Fig. 5 (b)), the ring achieves the shortest convergence time (Fig. 5 (c)), even if there is no communication delay. Qualitatively similar results for other ML problems and time distributions are in (Authors, 2019, App. G).

5 CONCLUSIONS

We have explained, both through analysis and experiments, when and why the communication topology does not affect the number of epochs consensus-based optimization methods need to converge, an effect recently observed in many papers, but not thoroughly investigated. We have also shown that, as a consequence of this invariance, a less connected topology achieves a shorter convergence time, not necessarily because it incurs a smaller communication load, but because it mitigates the stragglers' problem. The distributed operation of consensus-based approaches appears particularly suited for federated and multi-agent learning. Our study points out that further research is required for these applications, because the benefits observed until now are dependent on the statistical similarity of the local datasets, an assumption that is not satisfied in federated learning.



(a) Throughput (b) Error vs iterations (c) Error vs time

Figure 5: Effect of network connectivity (degree d) on the convergence for dataset MNIST with computation times from a Spark cluster. M = 16, B = 500.

References

UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association. ISBN 978-1-931971-33-1.

Ganesh Ananthanarayanan, Ali Ghodsi, Scott Shenker, and Ion Stoica. Effective straggler mitigation: Attack of the clones. In *Proc. of the 10th USENIX Conf. NSDI*, 2013.

Authors. Submitted supplementary material file, 2019.

Kevin Canini, Tushar Chandra, Eugene Ie, Jim McFadden, Ken Goldman, Mike Gunter, Jeremiah Harmsen, Kristen LeFevre, Dmitry Lepikhin, Tomas Lloret Llinares, Indraneel Mukherjee, Fernando Pereira, Josh Redstone, Tal Shaked, and Yoram Singer. Sibyl: A system for large scale supervised machine learning, 2014. Technical talk.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Trans. on Automatic Control*, 57(3):592–606, 2012. ISSN 0018-9286. doi: 10.1109/TAC.2011.2161027.

Andrew Gibiansky. Bringing hpc techniques to deep learning. online, http://research.baidu.com/bringing-hpc-techniques-deep-learning, 2017.

Google I/O. Distributed tensorflow training. online, https://www.youtube.com/watch?v=bRMGoPqsn20, 2018.

Franz Graf, Hans-Peter Kriegel, Matthias Schubert, Sebastian Pölsterl, and Alexander Cavallaro. 2D Image Registration in CT Images Using Radial Image Descriptors. In *Proc. of MICCAI*, pages 607–614, Berlin, Heidelberg, 2011. Springer. ISBN 978-3-642-23629-7.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Can Karakus, Yifan Sun, Suhas Diggavi, and Wotao Yin. Straggler mitigation in distributed optimization through data encoding. In *Proc. of NIPS)*, pages 5434–5442. 2017.

Jakub Konecný, Brendan McMahan, and Daniel Ramage. Federated Optimization: Distributed Optimization Beyond the Datacenter. In *Neural Information Processing Systems (workshop)*, 2015.

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

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. doi: 10.1109/5.726791.

Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 19–27. Curran Associates, Inc., 2014.

Songze Li, Seyed Mohammadreza Mousavi Kalan, A. Salman Avestimehr, and Mahdi Soltanolkotabi. Near-Optimal Straggler Mitigation for Distributed Gradient Methods. In *Proc. of the 7th Intl. Workshop ParLearning*, May 2018.

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 *NIPS*, 2017.

Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *ICML*, 2018.

Qinyi Luo, Jinkun Lin, Youwei Zhuo, and Xuehai Qian. Hop: Heterogeneity-aware decentralized training. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '19, pages 893–907, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6240-5.

Brendan D. McKay. The expected eigenvalue distribution of a large regular graph. *Linear Algebra and its Applications*, 40:203 – 216, 1981. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(81)90150-6.

Brendan D. McKay and Nicholas C. Wormald. Uniform generation of random regular graphs of moderate degree. *J. Algorithms*, 11(1):52–67, February 1990. ISSN 0196-6774. doi: 10.1016/0196-6774(90)90029-E.

Carl D. Meyer, editor. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, PA, USA, 2000. ISBN 0-89871-454-0.

Angelia Nedić and Asuman E. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Trans. Automat. Contr.*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, and Michael G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proc. of the IEEE*, 106(5):953–976, May 2018. ISSN 0018-9219. doi: 10.1109/JPROC.2018.2817461.

Giovanni Neglia, Gianmarco Calbi, Don Towsley, and Gayane Vardoyan. The Role of Network Topology for Distributed Machine Learning. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2019.

Alex Olshevsky, Ioannis Ch. Paschalidis, and Shi Pu. A non-asymptotic analysis of network independence for distributed stochastic gradient descent, 2019. arXiv preprint arXiv:1906.02702.

Fabrizio Petrini, Darren J. Kerbyson, and Scott Pakin. The case of the missing supercomputer performance: Achieving optimal performance on the 8,192 processors of asci q. In *Proceedings of the 2003 ACM/IEEE Conference on Supercomputing*, SC '03, pages 55–, New York, NY, USA, 2003. ACM. ISBN 1-58113-695-1. doi: 10.1145/1048935.1050204.

Leslie N Smith. Cyclical learning rates for training neural networks. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 464–472. IEEE, 2017.

Alexander Smola and Shravan Narayanamurthy. An architecture for parallel topic models. *Proc. VLDB Endow.*, 3(1-2):703–710, September 2010. ISSN 2150-8097. doi: 10.14778/1920841.1920931.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, September 1986. ISSN 0018-9286. doi: 10.1109/TAC.1986.1104412.

Steven R. Young, Derek C. Rose, Travis Johnston, William T. Heller, Thomas P. Karnowski, Thomas E. Potok, Robert M. Patton, Gabriel Perdue, and Jonathan Miller. Evolving deep networks using hpc. In *Proceedings of the Machine Learning on HPC Environments*, MLHPC'17, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5137-9. doi: 10.1145/3146347.3146355.

A Notation

We use an overline to denote an average over all the nodes and a "hat" to denote the time-average. For example

$$\overline{\boldsymbol{w}}(k) = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{w}_i(k), \qquad \hat{\boldsymbol{w}}_i(k) = \frac{1}{k+1}\sum_{h=0}^{k} \boldsymbol{w}_i(h). \tag{13}$$

For a matrix, e.g. $\boldsymbol{W}(k) = (\boldsymbol{w}_1(k), \dots \boldsymbol{w}_M(k))$, $\overline{\boldsymbol{W}}(k)$ denotes the matrix whose column i is $\overline{\boldsymbol{w}}(k)$, i.e.

$$\overline{\boldsymbol{W}}(k) = \boldsymbol{W}(k)\boldsymbol{P}(\mathbf{1}),$$

where $\boldsymbol{P}(\mathbf{1}) = \frac{\mathbf{1}\mathbf{1}^\intercal}{M}$ is the orthogonal projector on the subspace generated by $\mathbf{1}$. $\Delta\boldsymbol{W}(k)$ is used to denote the difference $\boldsymbol{W}(k) - \overline{\boldsymbol{W}}(k)$.

Given a matrix $\boldsymbol{A}$, $\boldsymbol{A}_{i,:}$ and $\boldsymbol{A}_{:,j}$ denote the i-th row and the j-th column, respectively.

We use different standard matrix norms, whose definitions are reported here for completeness. Let $\boldsymbol{A}$ be a $I \times J$ matrix:

$$\|\boldsymbol{A}\|_2 = \sigma_{\max}(\boldsymbol{A}), \tag{14}$$

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{1\le i\le I, 1\le j\le J} |A_{i,j}|^2} = \sqrt{\sum_{i=1}^{\min(I,J)} \sigma_i^2(\boldsymbol{A})}, \tag{15}$$

where $\{\sigma_i(\boldsymbol{A})\}$ are the singular values of the matrix $\boldsymbol{A}$ and $\sigma_{\max}(\boldsymbol{A})$ is the largest one.

We will also consider the Frobenius inner product between matrices defined as follows

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle_F \triangleq \sum_{i,j} \boldsymbol{A}_{i,j}\boldsymbol{B}_{i,j} = \mathrm{Tr}\left(\boldsymbol{A}^\intercal\boldsymbol{B}\right) \tag{16}$$

All the results in Appendix C assume that the matrix $\boldsymbol{A}$ is irreducible, primitive, doubly stochastic, non-negative, and normal.

B Linear algebra reminders

B.1 Irreducible primitive doubly stochastic non-negative matrices

We remind some results from Perron-Frobenius theory (Meyer, 2000, Ch. 8). Because our communication graph is strongly connected, the $M \times M$ consensus matrix $\boldsymbol{A}$ is irreducible. Moreover, the consensus matrix has non-null diagonal elements and then it is also primitive. The spectral radius $\rho(\boldsymbol{A}) \triangleq \max_i |\lambda_i|$ is then itself a simple eigenvalue. Because $\boldsymbol{A}$ is also stochastic, its eigenvalue $\lambda_1 = 1$ coincides with the spectral radius ($1 \le \max|\lambda_i| \le \|\boldsymbol{A}\|_1 = 1$).

We also observe that, for any non-negative matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 = 1$ if and only if the matrix is doubly stochastic. In fact, if $\boldsymbol{A}$ is doubly stochastic, so is $\boldsymbol{A}^\intercal\boldsymbol{A}$ and hence ($\|A\|_2^2 \triangleq$)$\rho(\boldsymbol{A}^\intercal\boldsymbol{A}) = 1$. For the opposite direction, assume that $\mathbf{1}$ is not a left eigenvector, then the vector $\mathbf{1}^\intercal\boldsymbol{A}$ is not aligned with $\mathbf{1}$ and it follows from Cauchy-Schwarz inequality:

$$M = \mathbf{1}^\intercal\boldsymbol{A}\mathbf{1} < \|\boldsymbol{A}^\intercal\mathbf{1}\|_2 \times \|\mathbf{1}\|_2 = \|\mathbf{1}^\intercal\boldsymbol{A}\|_2\sqrt{M}.$$

Hence $\|\boldsymbol{A}^\intercal\mathbf{1}\|_2 > \sqrt{M} = \|\mathbf{1}\|_2$, from which it follows $\|\boldsymbol{A}^\intercal\|_2 > 1$, contradicting the hypothesis $\|\boldsymbol{A}\|_2 = 1$.

Let $\boldsymbol{P}(\mathbf{1}) = \frac{\mathbf{1}\mathbf{1}^\intercal}{M}$ be the orthogonal projector on the subspace generated by the unit vector $\mathbf{1}$.

The non-zero singular values of $\boldsymbol{A}$ are the positive square roots of the non-zero eigenvalues of $\boldsymbol{A}^\intercal\boldsymbol{A}$. We observe that $(\boldsymbol{A} - \boldsymbol{P}(\mathbf{1}))^\intercal(\boldsymbol{A} - \boldsymbol{P}(\mathbf{1})) = \boldsymbol{A}^\intercal\boldsymbol{A} - \boldsymbol{P}(\mathbf{1})$. The spectrum of $\boldsymbol{A}^\intercal\boldsymbol{A} - \boldsymbol{P}(\mathbf{1})$ is equal to the spectrum of $\boldsymbol{A}^\intercal\boldsymbol{A}$ but for one eigenvalue 1 that is replaced by an eigenvalue 0. It follows that $\sigma_1(\boldsymbol{A} - \boldsymbol{P}(\mathbf{1})) = \sigma_2(\boldsymbol{A})$.

B.2 Normal matrices

An $M \times M$ matrix $\boldsymbol{A}$ is *normal* if $\boldsymbol{A}^\intercal \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\intercal$. A matrix $\boldsymbol{A}$ is normal if and only if it is unitarily diagonalizable (Meyer, 2000, p. 547), i.e. it exists a complete orthonormal set of eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \dots \boldsymbol{u}_M$ such that $\boldsymbol{U}^\intercal \boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}$, where $\boldsymbol{D}$ is the diagonal matrix containing the eigenvalues and $\boldsymbol{U}$ has the eigenvectors as columns.

Normal matrices have a spectral decomposition with orthogonal projectors (Meyer, 2000, p. 517), i.e.

$$\boldsymbol{A} = \sum_{q=1}^{Q} \lambda_q \boldsymbol{P}_q,$$

where $\lambda_1, \dots \lambda_Q$ are the $Q \leq M$ eigenvalues,, $\boldsymbol{P}_i$ is the orthogonal projector onto the nullspace of $\boldsymbol{A} - \lambda_i \boldsymbol{I}$ along the range of $\boldsymbol{A} - \lambda_i \boldsymbol{I}$, $\boldsymbol{P}_i \boldsymbol{P}_j = \boldsymbol{0}$ for $i \neq j$, and $\sum_{i=1}^{M} \boldsymbol{P}_i = \boldsymbol{I}$. Because the projectors $\boldsymbol{P}_i$ are orthogonal and non null it holds $\boldsymbol{P}_i^\intercal = \boldsymbol{P}_i$ and $||\boldsymbol{P}_i||_2 = 1$ (Meyer, 2000, p. 433). Moreover, for any vector $\boldsymbol{x}$ and $h \geq 0$, it holds:

$$\left\| \boldsymbol{A}^h \boldsymbol{x} \right\|_2^2 = \sum_{q=1}^{Q} |\lambda_q|^{2h} \left\| \boldsymbol{P}_q \boldsymbol{x} \right\|_2^2 . \tag{17}$$

Symmetric matrices as well as circulant matrices are normal. In fact, a circulant matrix is always diagonalizable by the Fourier matrix and then it is normal.

The non-zero singular values of a normal matrix $\boldsymbol{A}$ are the modules of its eigenvalues, i.e. $\sigma_q(\boldsymbol{A}) = |\lambda_q(\boldsymbol{A})|$. If the matrix $\boldsymbol{A}$ is also imprimitive, irreducible, doubly stochastic and non-negative, it holds

$$\sigma_1(\boldsymbol{A} - \boldsymbol{P}(\boldsymbol{1})) = \sigma_2(\boldsymbol{A}) = |\lambda_2| < 1. \tag{18}$$

Moreover, observe that in this case $\boldsymbol{P}_1 = \boldsymbol{P}(\boldsymbol{1})$.

C Convergence results

All the following results assume that the matrix $\boldsymbol{A}$ is irreducible, primitive, doubly stochastic, non-negative, and normal.

Lemma C.1. *The following inequality holds.*

$$\mathbb{E}_{\boldsymbol{\xi}}[||\Delta \boldsymbol{W}(k+1)||_F] \leq \sqrt{M}\|\Delta \boldsymbol{W}(0)\|_F |\lambda_2|^{k+1} + \sum_{h=0}^{k} \eta(h) \sqrt{\sum_{l=2}^{Q} |\lambda_l|^{2(k-h)}\, \mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta \boldsymbol{G}(h)\boldsymbol{P}_l\|_F^2\right]}. \tag{19}$$

Proof.

$$\|\Delta \boldsymbol{W}(k+1)\|_F = \left\| \boldsymbol{W}(0)\left(\boldsymbol{A}^{k+1} - \boldsymbol{P}(\boldsymbol{1})\right) - \sum_{h=0}^{k} \eta(h)\boldsymbol{G}(h)\left(\boldsymbol{A}^{k-h} - \boldsymbol{P}(\boldsymbol{1})\right)\right\|_F \tag{20}$$

$$= \left\| \Delta\boldsymbol{W}(0)\left(\boldsymbol{A}^{k+1} - \boldsymbol{P}(\boldsymbol{1})\right) - \sum_{h=0}^{k} \eta(h)\Delta\boldsymbol{G}(h)\left(\boldsymbol{A}^{k-h} - \boldsymbol{P}(\boldsymbol{1})\right)\right\|_F \tag{21}$$

$$\leq \left\| \Delta\boldsymbol{W}(0)\left(\boldsymbol{A}^{k+1} - \boldsymbol{P}(\boldsymbol{1})\right)\right\|_F + \sum_{h=0}^{k} \eta(h) \left\|\Delta\boldsymbol{G}(h)\left(\boldsymbol{A}^{k-h} - \boldsymbol{P}(\boldsymbol{1})\right)\right\|_F \tag{22}$$

$$= \left\| \Delta\boldsymbol{W}(0)\left(\boldsymbol{A} - \boldsymbol{P}(\boldsymbol{1})\right)^{k+1}\right\|_F + \sum_{h=0}^{k} \eta(h) \left\|\Delta\boldsymbol{G}(h)\left(\boldsymbol{A} - \boldsymbol{P}(\boldsymbol{1})\right)^{k-h}\right\|_F \tag{23}$$

$$= \left\| \Delta\boldsymbol{W}(0)(\Delta\boldsymbol{A})^{k+1}\right\|_F + \sum_{h=0}^{k} \eta(h) \left\|\Delta\boldsymbol{G}(h)(\Delta\boldsymbol{A})^{k-h}\right\|_F. \tag{24}$$

The first equality (20) follows from (5), $\overline{\boldsymbol{W}}(k+1) = \boldsymbol{W}(k+1)\boldsymbol{P}(\boldsymbol{1})$, and $\boldsymbol{A}\boldsymbol{P}(\boldsymbol{1}) = \boldsymbol{P}(\boldsymbol{1})$, because $\boldsymbol{A}$ is row stochastic. Equation (21) follows from the fact that, for any matrix $\boldsymbol{B}$, $\overline{\boldsymbol{B}} = \boldsymbol{B}\boldsymbol{P}(\boldsymbol{1})$ and then $\overline{\boldsymbol{B}}(\boldsymbol{A}^h - \boldsymbol{P}(\boldsymbol{1})) = \boldsymbol{B}\boldsymbol{P}(\boldsymbol{1})\boldsymbol{A}^h - \boldsymbol{B}\boldsymbol{P}(\boldsymbol{1})^2 = \boldsymbol{B}\boldsymbol{P}(\boldsymbol{1}) - \boldsymbol{B}\boldsymbol{P}(\boldsymbol{1}) = \boldsymbol{0}$, because $\boldsymbol{A}$ is column stochastic and $\boldsymbol{P}(\boldsymbol{1})$ is a projector. For (23) expand $(\boldsymbol{A} - \boldsymbol{P}(\boldsymbol{1}))^h$ taking into account again that $\boldsymbol{A}$ is row stochastic.

Let us now bound separately the two terms on the right hand side of inequality (23). For the first one it holds

$$\begin{aligned}\left\| \Delta\boldsymbol{W}(0)(\Delta\boldsymbol{A})^{k+1}\right\|_F &\leq \|\Delta\boldsymbol{W}(0)\|_F \left\|(\Delta\boldsymbol{A})^{k+1}\right\|_F \\ &= \|\Delta\boldsymbol{W}(0)\|_F \sqrt{\sum_{l=2}^{M} |\lambda_l|^{2(k+1)}} \\ &\leq \sqrt{M}\|\Delta\boldsymbol{W}(0)\|_F |\lambda_2|^{k+1},\end{aligned} \tag{25}$$

where the first inequality follows from the sub-multiplicative property of Frobenius norm. For the second term on the right hand side of (23), we carry out a more careful analysis.

$$\begin{aligned}\left\| \Delta\boldsymbol{G}(h)(\Delta\boldsymbol{A})^{k-h}\right\|_F^2 &= \sum_{i=1}^{n} \left\|\Delta\boldsymbol{G}_{i,:}(h)(\Delta\boldsymbol{A})^{k-h}\right\|_2^2 \\ &= \sum_{i=1}^{n} \left\| \Delta\boldsymbol{G}_{i,:}(h) \sum_{l=2}^{Q} |\lambda_l|^{k-h}\boldsymbol{P}_l\right\|_2^2 \\ &= \sum_{i=1}^{n}\sum_{l=2}^{Q} |\lambda_l|^{2(k-h)} \|\Delta\boldsymbol{G}_{i,:}(h)\boldsymbol{P}_l\|_2^2 \\ &= \sum_{l=2}^{Q} |\lambda_l|^{2(k-h)} \sum_{i=1}^{n} \|\Delta\boldsymbol{G}_{i,:}(h)\boldsymbol{P}_l\|_2^2 \\ &= \sum_{l=2}^{Q} |\lambda_l|^{2(k-h)} \|\Delta\boldsymbol{G}(h)\boldsymbol{P}_l\|_F^2\end{aligned} \tag{26}$$

From (23), (25), (26), and Jensen's inequality, it follows

$$\mathbb{E}_{\boldsymbol{\xi}}[||\Delta \boldsymbol{W}(k+1)||_F] \leq \sqrt{M}\|\Delta \boldsymbol{W}(0)\|_F|\lambda_2|^{k+1} + \sum_{h=0}^{k}\eta(h)\sqrt{\sum_{l=2}^{Q}|\lambda_l|^{2(k-h)}\,\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta \boldsymbol{G}(h)\boldsymbol{P}_l\|_F^2\right]}$$

□

Let R be a bound on the energy of the initial parameter vector across the different nodes, i.e. $\|\boldsymbol{W}(0)\|_F^2 \leq R$, and R_{sp} be the corresponding bound for the spread of the parameter vectors around their averages, i.e. $\left\|\boldsymbol{W}(0) - \overline{\boldsymbol{W}}(0)\right\|_F^2 \leq R_{\text{sp}}$. Similarly, we define E and E_{sp} as bounds for the subgradient matrix ΔG for any time h: $\sup_{h\geq 0}\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{G}(h)\|_F^2\right] \leq E$, $\sup_{h\geq 0}\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta \boldsymbol{G}(h)\|_F^2\right] \leq E_{\text{sp}}$. We observe that $R_{\text{sp}} \leq R$ and $E_{\text{sp}} \leq E$.

Moreover, for a normal matrix $\boldsymbol{A}$, we define for $l = 1, \dots M$:

$$E_{\text{sp},l} \triangleq \sup_{h\geq 0}\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\Delta \boldsymbol{G}_{i,:}(h)\sum_{l'=2}^{l}\boldsymbol{P}_{l'}\right\|_F^2\right],$$

with the usual convention that $\sum_i^j \cdot = 0$ if $j < 1$ and then $E_{\text{sp},1} = 0$. We observe that $E_{\text{sp},l}$ represents the maximum expected energy $\Delta \boldsymbol{G}(h)$ in the projection subspace defined by the first l projectors. In particular it holds

$$E_{\text{sp},M} = \sup_{h\geq 0}\left\|\Delta \boldsymbol{G}(h)\sum_{l'=2}^{M}\boldsymbol{P}_{l'}\right\|_F^2 = \sup_{h\geq 0}\|\Delta \boldsymbol{G}(h)(\boldsymbol{I} - \boldsymbol{P}_1)\|_F^2 \tag{27}$$

$$= \sup_{h\geq 0}\|\Delta \boldsymbol{G}(h)\|_F^2 \leq E_{\text{sp}}. \tag{28}$$

Let us now consider the normalized fraction of energy in each subspace, defined as follows:

$$e_l \triangleq \frac{E_{\text{sp},l} - E_{\text{sp},l-1}}{E_{\text{sp},M}}, \tag{29}$$

so that $\sum_l e_l = 1$. Finally, let

$$\alpha(h) \triangleq \sqrt{\sum_{l=2}^{M} e_l \left|\frac{\lambda_l}{\lambda_2}\right|^{2h}}, \tag{30}$$

and we denote $\alpha(1)$ simply as α. We observe that $|\lambda_l/\lambda_2| \leq 1$, then $\alpha(h)$ is decreasing in h.

$$\sqrt{e_2} = \sqrt{e_2\left|\frac{\lambda_2}{\lambda_2}\right|^{2h}} \leq \sqrt{\sum_{l=2}^{M} e_l\left|\frac{\lambda_l}{\lambda_2}\right|^{2h}} \leq \sqrt{\sum_{l=2}^{M} e_l} = 1$$

$\alpha(h)$ can be considered a bound for the effective energy contribution of the vector $\Delta \boldsymbol{G}(h)$ in the projection subspace defined by $\boldsymbol{P}_2$.

Corollary C.2. *Considering the definition of R_{sp}, E_{sp}, and $\alpha(l)$, the following inequality holds for a constant learning rate η:*

$$||\Delta \boldsymbol{W}(k)||_F \leq \sqrt{M}\sqrt{R_{sp}}|\lambda_2|^k + \eta\sqrt{E_{sp}}\left((1-\alpha)\mathbb{1}_{k\geq 1} + \alpha\frac{1-|\lambda_2|^k}{1-|\lambda_2|}\right). \tag{31}$$

Proof. The first term on the right hand side bounds $\sqrt{M}\|\boldsymbol{W}(0)\|_F$ by definition of R_{sp}.

Observe that

$$\sqrt{\sum_{l=2}^{M} |\lambda_l|^{2(k-h)} \mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta \boldsymbol{G}(h)\boldsymbol{P}_l\|_F^2\right]} \leq |\lambda_2|^{k-h}\sqrt{E_{\text{sp}}}\sqrt{\sum_{l=2}^{M} \frac{\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta \boldsymbol{G}(h)\boldsymbol{P}_l\|_F^2\right]}{E_{\text{sp}}}\left|\frac{\lambda_l}{\lambda_2}\right|^{2(k-h)}} \tag{32}$$

$$\leq |\lambda_2|^{k-h}\sqrt{E_{\text{sp}}}\sqrt{\sum_{l=2}^{M} e_l\left|\frac{\lambda_l}{\lambda_2}\right|^{2(k-h)}} \tag{33}$$

$$\leq |\lambda_2|^{k-h}\sqrt{E_{\text{sp}}}\alpha(k-h). \tag{34}$$

Using this bound, we obtain

$$\sum_{h=0}^{k}\sqrt{\sum_{l=2}^{M} |\lambda_l|^{2(k-h)} \mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta \boldsymbol{G}(h)\boldsymbol{P}_l\|_2^2\right]} \leq \sqrt{E_{\text{sp}}}\sum_{h=0}^{k}\alpha(h)|\lambda_2|^h \tag{35}$$

$$= \sqrt{E_{\text{sp}}}\left(1+\sum_{h=1}^{k}\alpha(h)|\lambda_2|^h\right) \tag{36}$$

$$\leq \sqrt{E_{\text{sp}}}\left(1+\alpha(1)\sum_{h=1}^{k}|\lambda_2|^h\right) \tag{37}$$

$$= \sqrt{E_{\text{sp}}}\left(1+\alpha\sum_{h=1}^{k}|\lambda_2|^h\right) \tag{38}$$

$$= \sqrt{E_{\text{sp}}}\left((1-\alpha)+\alpha\sum_{h=0}^{k}|\lambda_2|^h\right) \tag{39}$$

$$= \sqrt{E_{\text{sp}}}\left((1-\alpha)+\alpha\frac{1-|\lambda_2|^{k+1}}{1-|\lambda_2|}\right) \tag{40}$$

From this last bound and (19), an inequality similar to (31), but without the indicator function, follows immediately. The indicator function can be introduced because, for $k=0$, it is simply $||\Delta \boldsymbol{W}(k)||_F \leq \sqrt{R_{\text{sp}}}$.

□

Lemma C.3. *Let $\mathbb{W}^*$ the (non-empty) optimal solution set. It holds:*

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\text{dist}(\overline{\boldsymbol{w}}(k+1),\mathbb{W}^*)^2\right] \leq \mathbb{E}_{\boldsymbol{\xi}}\left[\text{dist}(\overline{\boldsymbol{w}}(k),\mathbb{W}^*)^2\right] + \frac{\eta^2 E}{M} + \frac{4\eta H}{M}\|\Delta \boldsymbol{W}(k)\|_F - \frac{2\eta}{M}\left(\mathbb{E}_{\boldsymbol{\xi}}[F(\overline{\boldsymbol{w}}(k))]-F^*\right). \tag{41}$$

where $\text{dist}(\boldsymbol{x},\mathbb{X})$ *denotes the distance between a vector* $\boldsymbol{x}$ *and the set* $\mathbb{X}$.

Proof. The proof follows closely the proof in (Nedić and Ozdaglar, 2009, Lemma 5), replacing the Euclidean norm with the Frobenius norm used in this paper.

We denote the subgradients of F_j in $\boldsymbol{w}_j(k)$ and $\overline{\boldsymbol{w}}(k)$ simply as $\boldsymbol{g}_j(k)$ and $\tilde{\boldsymbol{g}}_j(k)$, respectively, *i.e.* $\boldsymbol{g}_j(k) = \boldsymbol{g}_j(\boldsymbol{w}_j(k))$ and $\tilde{\boldsymbol{g}}_j(k) = \boldsymbol{g}_j(\overline{\boldsymbol{w}}(k))$. Let $\boldsymbol{G}(k)$ and $\tilde{\boldsymbol{G}}(k)$ be respectively the matrices whose columns are $\boldsymbol{g}_j(k)$ and $\tilde{\boldsymbol{g}}_j(k)$. Let $\boldsymbol{x}$ be a generic vector in $\mathbb{R}^n$.

$$\|\overline{\boldsymbol{w}}(k+1)-\boldsymbol{x}\|_2^2 = \left\|\overline{\boldsymbol{w}}(k)-\boldsymbol{x}-\frac{\eta}{M}\sum_{j=1}^{M}\boldsymbol{g}_j(\boldsymbol{w}_j(k))\right\|_2^2$$

$$= \|\overline{\boldsymbol{w}}(k)-\boldsymbol{x}\|_2^2 + \frac{\eta^2}{M^2}\left\|\sum_{j=1}^{M}\boldsymbol{g}_j(k)\right\|_2^2 - \frac{2\eta}{M}\sum_{j=1}^{M}\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{x})$$

$$\leq \|\overline{\boldsymbol{w}}(k)-\boldsymbol{x}\|_2^2 + \frac{\eta^2}{M}\|\boldsymbol{G}(k)\|_F^2 - \frac{2\eta}{M}\sum_{j=1}^{M}\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{x}). \tag{42}$$

Let us bound the scalar product $\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{x})$:

$$\begin{aligned}\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{x}) &= \boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{w}_j(k)) + \boldsymbol{g}_j(k)^\intercal(\boldsymbol{w}_j(k)-\boldsymbol{x}) \\ &\geq \boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{w}_j(k)) + F_j(\boldsymbol{w}_j(k)) - F_j(\boldsymbol{x}) \qquad (43)\\ &= \boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{w}_j(k)) + F_j(\boldsymbol{w}_j(k)) - F_j(\overline{\boldsymbol{w}}(k)) + F_j(\overline{\boldsymbol{w}}(k)) - F_j(\boldsymbol{x}) \\ &\geq \boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{w}_j(k)) + \tilde{\boldsymbol{g}}_j(k)^\intercal(\boldsymbol{w}_j(k)-\overline{\boldsymbol{w}}(k)) + F_j(\overline{\boldsymbol{w}}(k)) - F_j(\boldsymbol{x}), \qquad (44)\end{aligned}$$

where (43) follows from $\boldsymbol{g}_j(k)$ being a subgradient of F_j in $\boldsymbol{w}_j(k)$ and (44) from $\tilde{\boldsymbol{g}}_j(k)$ being a subgradient of F_j in $\overline{\boldsymbol{w}}(k)$. Summing over j the LHS and RHS of the above inequality, we obtain:

$$\begin{aligned}\sum_{j=1}^{M}\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{x}) &\geq \sum_{j=1}^{M}\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{w}_j(k)) + \sum_{j=1}^{M}\tilde{\boldsymbol{g}}_j(k)^\intercal(\boldsymbol{w}_j(k)-\overline{\boldsymbol{w}}(k)) + \sum_{j=1}^{M}F_j(\overline{\boldsymbol{w}}(k)) - \sum_{j=1}^{M}F_j(\boldsymbol{x}) \\ &= -\langle \boldsymbol{G}(k), \Delta\boldsymbol{W}(k)\rangle_F + \langle \tilde{\boldsymbol{G}}(k), \Delta\boldsymbol{W}(k)\rangle_F + F(\overline{\boldsymbol{w}}(k)) - F(\boldsymbol{x}),\end{aligned}$$

where we have used the definition of the Frobenius inner product (16). By computing the expected value we obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\left[\sum_{j=1}^{M}\boldsymbol{g}_j(k)^\intercal(\overline{\boldsymbol{w}}(k)-\boldsymbol{x})\right] &\geq \left(\langle \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}(k)], \Delta\boldsymbol{W}(k)\rangle_F - \langle \mathbb{E}_{\boldsymbol{\xi}}\left[\tilde{\boldsymbol{G}}(k)\right], \Delta\boldsymbol{W}(k)\rangle_F\right) + \mathbb{E}_{\boldsymbol{\xi}}[F(\overline{\boldsymbol{w}}(k)) - F(\boldsymbol{x})] \\ &\geq -\left(\|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}(k)]\|_F + \left\|\mathbb{E}_{\boldsymbol{\xi}}\left[\tilde{\boldsymbol{G}}(k)\right]\right\|_F\right)\|\Delta\boldsymbol{W}(k)\|_F + \mathbb{E}_{\boldsymbol{\xi}}[F(\overline{\boldsymbol{w}}(k)) - F(\boldsymbol{x})] \\ &\geq -2H\|\Delta\boldsymbol{W}(k)\|_F + \mathbb{E}_{\boldsymbol{\xi}}[F(\overline{\boldsymbol{w}}(k)) - F(\boldsymbol{x})] \qquad (45)\end{aligned}$$

From (42) and (45), we obtain:

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\overline{\boldsymbol{w}}(k+1)-\boldsymbol{x}\|_2^2\right] \leq \mathbb{E}_{\boldsymbol{\xi}}\left[\|\overline{\boldsymbol{w}}(k)-\boldsymbol{x}\|_2^2\right] + \frac{\eta^2 E}{M} + \frac{4\eta H}{M}\|\Delta\boldsymbol{W}(k)\|_F - \frac{2\eta}{M}\,\mathbb{E}_{\boldsymbol{\xi}}[F(\overline{\boldsymbol{w}}(k)) - F(\boldsymbol{x})]. \qquad (46)$$

Then the thesis follows from considering $\boldsymbol{x}$ a generic point in the optimal set $\mathbb{W}^*$. □

C.1 Proof of Proposition 3.1

Proof. We start computing a bound for the time average of $\|\Delta\boldsymbol{W}(k)\|_F$ using Corollary C.2:

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\|\Delta\boldsymbol{W}(k)\|_F \leq& \frac{\sqrt{M}\sqrt{R_{\text{sp}}}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} + \eta\sqrt{E_{\text{sp}}}(1-\alpha)\frac{K-1}{K} \\ &+ \frac{\eta\sqrt{E_{\text{sp}}}\alpha}{1-|\lambda_2|}\left(1-\frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right) \qquad (47)\end{aligned}$$

If we take into account convexity of F, Lemma C.3, and (47), we obtain:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\widehat{\overline{\boldsymbol{w}}}(k)\right)\right] - F^* &= \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\frac{1}{K}\sum_{k=0}^{K-1}\overline{\boldsymbol{w}}(k)\right)\right] - F^* \\
&\underset{\text{convexity}}{\leq} \frac{1}{K}\sum_{k=0}^{K-1}\left(\mathbb{E}_{\boldsymbol{\xi}}[F(\overline{\boldsymbol{w}}(k))] - F^*\right) \\
&\underset{\text{Lem C.3}}{\leq} \frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{M}{2\eta}\left(\operatorname{dist}(\overline{\boldsymbol{w}}(k),\mathbb{W}^*)^2 - \operatorname{dist}(\overline{\boldsymbol{w}}(k+1),\mathbb{W}^*)^2\right) + \frac{\eta E}{2} + 2H\|\Delta\boldsymbol{W}(k)\|_F\right) \\
&= \frac{M}{2\eta K}\left(\operatorname{dist}(\overline{\boldsymbol{w}}(0),\mathbb{W}^*)^2 - \operatorname{dist}(\overline{\boldsymbol{w}}(K),\mathbb{W}^*)^2\right) + \frac{\eta E}{2} + 2H\frac{1}{K}\sum_{k=0}^{K-1}\|\Delta\boldsymbol{W}(k)\|_F \\
&\leq \frac{M}{2\eta K}\operatorname{dist}(\overline{\boldsymbol{w}}(0),\mathbb{W}^*)^2 + \frac{\eta E}{2} + 2H\frac{1}{K}\sum_{k=0}^{K-1}\|\Delta\boldsymbol{W}(k)\|_F \\
&\underset{(47)}{\leq} \frac{M}{2\eta K}\operatorname{dist}(\overline{\boldsymbol{w}}(0),\mathbb{W}^*)^2 + \frac{\eta E}{2} + 2H\frac{\sqrt{M}\sqrt{R_{\text{sp}}}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\
&\quad + 2H\eta\sqrt{E_{\text{sp}}}\left((1-\alpha)\frac{K-1}{K} + \frac{\alpha}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right)\right)
\end{aligned} \tag{48}$$

□

C.2 Proof of Corollary 3.2

Proof. Because of the relations $R_{\text{sp}} \leq R$ and $E_{\text{sp}} \leq E$, the only step to prove is that

$$(1-\alpha)\frac{K-1}{K} + \alpha\frac{1}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right) \leq \frac{1}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right). \tag{49}$$

The two sides can be rewritten as the sums indicated below:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left((1-\alpha)\mathbb{1}_{k\geq 1} + \alpha\frac{1-|\lambda_2|^k}{1-|\lambda_2|}\right) \leq \frac{1}{K}\sum_{k=0}^{K-1}\frac{1-|\lambda_2|^k}{1-|\lambda_2|}. \tag{50}$$

It is then sufficient to prove that for each $k \geq 0$

$$(1-\alpha)\mathbb{1}_{k\geq 1} + \alpha\frac{1-|\lambda_2|^k}{1-|\lambda_2|} \leq \frac{1-|\lambda_2|^k}{1-|\lambda_2|}. \tag{51}$$

This relation is obviously satisfied for $k = 0$ ($\alpha < 1$). For any $k > 0$, it follows from

$$\frac{1-|\lambda_2|^k}{1-|\lambda_2|} = \sum_{h=0}^{k-1}|\lambda_2|^h \geq 1. \tag{52}$$

□

C.3 Convergence of local estimates

Proposition C.4. *Under assumptions A1-A5 and that a constant learning rate $\eta(k) = \eta$ is used, an upper bound for the objective value at the end of the $(K-1)$th iteration is given, for each i, by:*

$$\begin{aligned}
\mathbb{E}[F(\hat{\boldsymbol{w}}_i(K-1))] - F^* \leq& \frac{M}{2\eta K}\operatorname{dist}(\overline{\boldsymbol{w}}(0),\mathbb{W}^*)^2 + \frac{\eta E}{2} + H\frac{3M\sqrt{R_{sp}}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\
&+ 3\eta\sqrt{M}H\sqrt{E_{sp}}\left((1-\alpha)\frac{K-1}{K} + \frac{\alpha}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right)\right).
\end{aligned} \tag{53}$$

Proof. We consider the local model at node i ($\boldsymbol{w}_i(k)$). Using convexity of the local functions and the definition of subgradients, we obtain:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}}[F(\hat{\boldsymbol{w}}_i(k))] &= \sum_{j=1}^{M} \mathbb{E}_{\boldsymbol{\xi}}[F_j(\hat{\boldsymbol{w}}_i(k))] \underset{\text{convexity}}{\leq} \sum_{j=1}^{M} \left(\mathbb{E}_{\boldsymbol{\xi}}\left[F_j\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_j(\hat{\boldsymbol{w}}_i(k))]^\intercal(\hat{\boldsymbol{w}}_i(k) - \hat{\overline{\boldsymbol{w}}}(k))\right) \\
&= \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sum_{j=1}^{M} \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_j(\hat{\boldsymbol{w}}_i(k))]^\intercal(\hat{\boldsymbol{w}}_i(k) - \hat{\overline{\boldsymbol{w}}}(k)) \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sum_{j=1}^{M} \|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}_j(\hat{\boldsymbol{w}}_i(k))]\|_2 \|\hat{\boldsymbol{w}}_i(k) - \hat{\overline{\boldsymbol{w}}}(k)\|_2 \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sqrt{M}\|\mathbb{E}_{\boldsymbol{\xi}}[G(\hat{\boldsymbol{w}}_i(k))]\|_F \left\|\hat{\boldsymbol{w}}_i(k) - \hat{\overline{\boldsymbol{w}}}(k)\right\|_2 \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sqrt{M}H\left\|\hat{\boldsymbol{w}}_i(k) - \hat{\overline{\boldsymbol{w}}}(k)\right\|_2 \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sqrt{M}H\frac{1}{K}\sum_{k=0}^{K-1}\|\boldsymbol{w}_i(k) - \overline{\boldsymbol{w}}(k)\|_2 \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sqrt{M}H\frac{1}{K}\sum_{k=0}^{K-1}\left\|\boldsymbol{W}(k) - \overline{\boldsymbol{W}}(k)\right\|_F \\
&\leq \mathbb{E}_{\boldsymbol{\xi}}\left[F\left(\hat{\overline{\boldsymbol{w}}}(k)\right)\right] + \sqrt{M}H\frac{\sqrt{M}\sqrt{R_{\text{sp}}}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\
&\quad + \sqrt{M}H\eta\sqrt{E_{\text{sp}}}\left((1-\alpha)\frac{K-1}{K} + \frac{\alpha}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right)\right)
\end{aligned} \tag{54}$$

Putting together (48) and (54), we obtain:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}}[F(\hat{\boldsymbol{w}}_i(k))] - F^* &\leq \frac{M}{2\eta K}\text{dist}(\overline{\boldsymbol{w}}(0), \mathbb{W}^*)^2 + \frac{\eta E}{2} + (2+\sqrt{M})H\frac{\sqrt{M}\sqrt{R_{\text{sp}}}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\
&\quad + (2+\sqrt{M})H\eta\sqrt{E_{\text{sp}}}\left((1-\alpha)\frac{K-1}{K} + \frac{\alpha}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right)\right) \\
&\leq \frac{M}{2\eta K}\text{dist}(\overline{\boldsymbol{w}}(0), \mathbb{W}^*)^2 + \frac{\eta E}{2} + H\frac{3M\sqrt{R_{\text{sp}}}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} \\
&\quad + 3\sqrt{M}H\eta\sqrt{E_{\text{sp}}}\left((1-\alpha)\frac{K-1}{K} + \frac{\alpha}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right)\right),
\end{aligned} \tag{55}$$

where the last inequality is simply to slightly compact the formula. □

Corollary C.5. *Under assumptions A1-A5 and that constant learning rate $\eta(k) = \eta$ is used, an upper bound for the objective value at the end of the $(K-1)$th iteration is given, for each i, by:*

$$\mathbb{E}[F(\hat{\boldsymbol{w}}_i(K-1))] - F^* \leq \frac{M}{2\eta K}\text{dist}(\overline{\boldsymbol{w}}(0), \mathbb{W}^*)^2 + \frac{\eta E}{2} + \sqrt{E}\frac{3M\sqrt{R}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} + \frac{3\eta\sqrt{M}E}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right). \tag{56}$$

In particular, if workers compute full-batch subgradients and the 2-norm of subgradients of functions F_i is bounded by a constant L, we obtain:

$$F(\hat{\boldsymbol{w}}_i(K-1)) - F^* \leq \frac{M}{2\eta K}\text{dist}(\overline{\boldsymbol{w}}(0), \mathbb{W}^*)^2 + \frac{\eta M L^2}{2} + L\frac{3M^{3/2}\sqrt{R}}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|} + \frac{3\eta M^{3/2}L^2}{1-|\lambda_2|}\left(1 - \frac{1}{K}\frac{1-|\lambda_2|^K}{1-|\lambda_2|}\right). \tag{57}$$

Proof. The result follows immediately from Proposition C.4 and (49). □

D Proof of Proposition 3.3

Proof. Our first remark is that selecting a uniform permutation and then a uniform random batch of size B without resampling is equivalent to selecting uniformly B elements from the dataset $\mathbb{S}$ without resampling. We denote this sampling with the random variable $\xi_{\mathbb{S}}$. This is true independently from the presence of data replication, as far as data point replicas are located at different nodes. From this observation and the formula for the variance of the average of a sample drawn without resampling, it follows:

$$\begin{aligned}\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{G}\|_F^2\right]\right] &= \sum_{i,j}\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[G_{i,j}^2\right]\right] = \sum_{i,j}\mathbb{E}_{\xi_{\mathbb{S}}}\left[G_{i,j}^2\right] = \sum_{i,j}\left(\left(\mathbb{E}_{\xi_{\mathbb{S}}}[G_{i,j}]\right)^2 + \mathrm{Var}_{\xi_{\mathbb{S}}}[G_{i,j}]\right)\\ &= \sum_{i,j}\left((\partial F)_i^2 + \frac{S}{S-1}\frac{\sigma_i^2}{B}\left(1-\frac{B}{S}\right)\right) = \sum_{i,j}\left((\partial F)_i^2 + \frac{S-B}{S-1}\frac{\sigma_i^2}{B}\right)\\ &= M\left(\|\partial F\|_2^2 + \frac{S-B}{S-1}\frac{\sigma^2}{B}\right) \end{aligned} \tag{58}$$

From $\sum_{j=1}^M x_j^2 = \sum_{j=1}^M (x_j-\bar{x})^2 + M\bar{x}^2$, where $\bar{x}$ denote the mean of the M values x_j, it follows

$$\|\Delta\boldsymbol{G}\|_F^2 = \|\boldsymbol{G}\|_F^2 - \left\|\boldsymbol{G}\frac{\mathbf{1}\mathbf{1}^\intercal}{M}\right\|_F^2 \tag{59}$$

In order to compute $\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\Delta\boldsymbol{G}\|_F^2\right]\right]$, we will then compute $\mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{G}\mathbf{1}\mathbf{1}^\intercal/M\|_F^2\right]\right]$ and then use (58) and (59). We denote the double expectation over ξ and π simply as $\mathbb{E}[]$.

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{M}\sum_{j=1}^M G_{i,j}\right)^2\right] &= \frac{1}{M^2}\left(\sum_j \mathbb{E}\left[G_{i,j}^2\right] + \sum_{j\neq j'}^M \mathbb{E}[G_{i,j}G_{i,j'}]\right)\\ &= \frac{1}{M}\left(\mathbb{E}\left[G_{i,j}^2\right] + (M-1)\,\mathbb{E}[G_{i,j}G_{i,j'}]\right), \end{aligned} \tag{60}$$

where j and j' are arbitrary values in $\{1,\dots M\}$ with $j'\neq j$.

$$\mathbb{E}\left[G_{i,j}^2\right] = \mathbb{E}[G_{i,j}]^2 + \mathrm{Var}\left[G_{i,j}\right] = (\partial F)_i^2 + \frac{S-B}{S-1}\frac{\sigma_i^2}{B}. \tag{61}$$

Let $\mathbb{1}_{s,s'}$ be the indicator function denoting if the datapoint $(x^{(s)},y^{(s)})$ has been selected in the minibatch at node j and the datapoint $(x^{(s')},y^{(s')})$ has been selected in the minibatch at node j'. In order to keep the notation compact we also denote by $\partial f_{i,s}$ and $\partial f_{i,s'}$ respectively $(\partial f(\boldsymbol{w},(x^{(s)},y^{(s)})))_i$ and $(\partial f(\boldsymbol{w},(x^{(s')},y^{(s')})))_i$.

$$\mathbb{E}[G_{i,j}G_{i,j'}] = \frac{1}{B^2}\sum_{s=1}^S\sum_{s'=1}^S \partial f_{i,s}\partial f_{i,s'}\,\mathbb{E}[\mathbb{1}_{s,s'}]. \tag{62}$$

The probability that the point s is in the local dataset at node i and the point is selected in the minibatch is $\frac{CS/M}{S}\times\frac{B}{CS/M} = \frac{B}{S}$. Let us consider first the case $s'=s$. Given that s is in j, the probability that one of the other $C-1$ copies is stored in j' is $\frac{C-1}{M-1}$ and this copy has a probability $\frac{B}{CS/M}$ to be selected. Then the total probability of the event that a copy of s is selected in the minibatch at j and another copy is selected in the minibatch at j' is

$$\frac{B}{S}\times\frac{C-1}{M-1}\times\frac{B}{CS/M} = \frac{B^2}{S^2}\frac{C-1}{C}\frac{M}{M-1}.$$

If $s'\neq s$, then s' may be present in j or not. The first event occurs with probability $\frac{\frac{CS}{M}-1}{S-1}$ (because s has already been located in j) and in this case the probability that s' is also located in j' is $\frac{C-1}{M-1}$. s' is not present in j with

probability $1 - \frac{\frac{CS}{M}-1}{S-1}$ and in this case it is located in j' with probability $\frac{C}{M-1}$. Then the total probability of the event that a copy of s' is located at j' given that a copy of s is located at j is

$$\frac{\frac{CS}{M}-1}{S-1}\frac{C-1}{M-1} + \left(1 - \frac{\frac{CS}{M}-1}{S-1}\right)\frac{C}{M-1} = \frac{C}{M} + \frac{M-C}{M(M-1)(S-1)}$$

and the probability that a copy of s is selected in the minibatch at j and a copy of $s' \neq s$ is located at j' is

$$\frac{B}{S} \times \frac{C}{M} + \frac{M-C}{M(M-1)(S-1)} \times \frac{B}{\frac{CS}{M}} = \frac{B^2}{S^2}\left(1 + \frac{M-C}{C(M-1)(S-1)}\right)$$

In conclusion, we have just proved that:

$$\mathbb{E}[\mathbb{1}_{s,s'}] = \begin{cases} \frac{B^2}{S^2}\frac{C-1}{C}\frac{M}{M-1}, & \text{for } s' = s \\ \frac{B^2}{S^2}\left(1 + \frac{M-C}{C(M-1)(S-1)}\right), & \text{for } s' \neq s \end{cases} \tag{63}$$

We observe that

$$\frac{1}{S}\sum_s (\partial f_{i,s})^2 = (\partial F)_i^2 + \sigma_i^2, \tag{64}$$

$$(\partial F)_i^2 = \left(\frac{1}{S}\sum_s \partial f_{i,s}\right)^2 = \frac{1}{S^2}\left(\sum_s (\partial f_{i,s})^2 + \sum_{s \neq s'} \partial f_{i,s}\partial f_{i,s'}\right). \tag{65}$$

Using these two relations and (63) in (62) we obtain

$$\mathbb{E}[G_{i,j}G_{i,j'}] = (\partial F)_i^2 + \sigma_i^2 \frac{C-M}{C(M-1)(S-1)}. \tag{66}$$

This equality together with (60) and (61) leads to:

$$\begin{aligned} \mathbb{E}\left[\|\boldsymbol{G}\mathbf{1}\mathbf{1}^\intercal/M\|_F^2\right] &= \sum_i \mathbb{E}\left[G_{i,j}^2\right] + (M-1)\,\mathbb{E}[G_{i,j}G_{i,j'}] \\ &= \|\partial F\|_2^2 + \frac{S-B}{S-1}\frac{\sigma^2}{B} + (M-1)\left(\|\partial F\|_2^2 + \sigma^2\frac{C-M}{C(M-1)(S-1)}\right) \\ &= M\|\partial F\|_2^2 + \sigma_2\frac{CS-MB}{CB(S-1)}. \end{aligned} \tag{67}$$

Finally,

$$\mathbb{E}\left[\|\Delta\boldsymbol{G}\|_F^2\right] = \mathbb{E}\left[\|\boldsymbol{G}\|_F^2\right] - \mathbb{E}\left[\|\boldsymbol{G}\mathbf{1}\mathbf{1}^\intercal/M\|_F^2\right] \tag{68}$$

$$= M\sigma^2\left(\frac{S-B}{B(S-1)} - \frac{CS-MB}{MCB(S-1)}\right). \tag{69}$$

We now move to bound $\mathbb{E}_\pi\left[\|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}]\|_F\right]$. The lower bound follows immediately from the fact that any norm is a convex function, so that

$$\mathbb{E}_\pi\left[\|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}]\|_F\right] \geq \|\mathbb{E}_\pi[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}]]\|_F = \sqrt{M}\|\partial F\|_2.$$

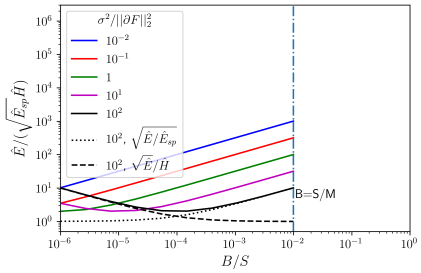


Figure 6: Estimate of $E/(\sqrt{E_{\text{sp}}}H)$ versus the relative batch size B/S for $M = 100$, $S = 10^6$, $C = 1$, and different level of heterogeneity ($\sigma^2/\|\partial F\|_2^2$) of the dataset.

For the upper bound:

$$\begin{aligned}
\mathbb{E}_\pi\left[\|\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{G}]\|_F\right] &= \mathbb{E}_\pi\left[\sqrt{\sum_{i,j}\left(\mathbb{E}_{\boldsymbol{\xi}}[G_{i,j}]\right)^2}\right] \\
&\leq \sqrt{\sum_{i,j}\mathbb{E}_\pi\left[\left(\mathbb{E}_{\boldsymbol{\xi}}[G_{i,j}]\right)^2\right]} && (70)\\
&= \sqrt{\sum_{i,j}\left(\mathbb{E}_\pi[\mathbb{E}_{\boldsymbol{\xi}}[G_{i,j}]]\right)^2 + \mathrm{Var}_\pi\left[\mathbb{E}_{\boldsymbol{\xi}}[G_{i,j}]\right]} && (71)\\
&= \sqrt{\sum_{i,j}(\partial F)_i^2 + \sigma_i^2\frac{M-C}{C(S-1)}} && (72)\\
&= \sqrt{M\left(\|\partial F\|_2^2 + \frac{M-C}{C(S-1)}\sigma^2\right)}, && (73)
\end{aligned}$$

where the first inequality follows from the concavity of square root function. Observe that $\mathbb{E}_{\boldsymbol{\xi}}[G_{:,j}]$ is equal to the full-batch gradient computed on the local dataset at node j. As such, its expected value over the permutations π is equal to ∂F, and its variance corresponds to the variance of the sample average when we draw CS/M samples from a dataset of S elements without resampling. □

E Effect of the replication factor C on β

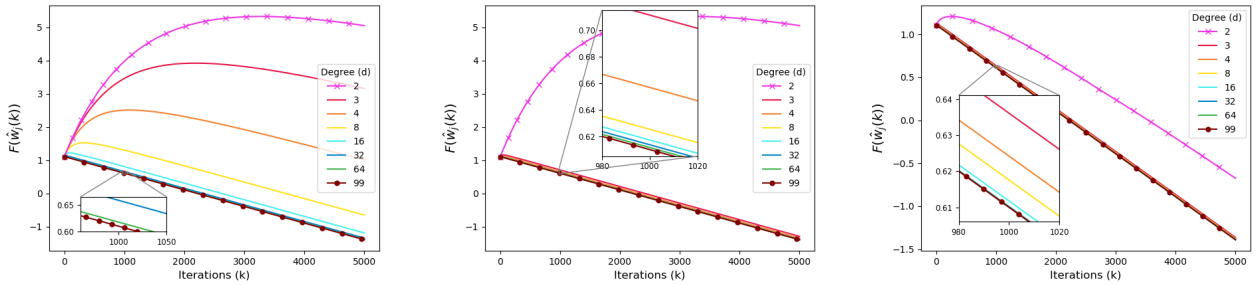
The comparison of Fig. 6 and Fig. 1 shows that $\frac{E}{\sqrt{E_{\text{sp}}}H}$ (and then β), does not depend much on the replication factor C, but for the fact that the batch size can scale up to CS/M.

F Toy example

In this section we consider a toy example to illustrate the implications of the findings in the previous section. We want to solve the following problem:

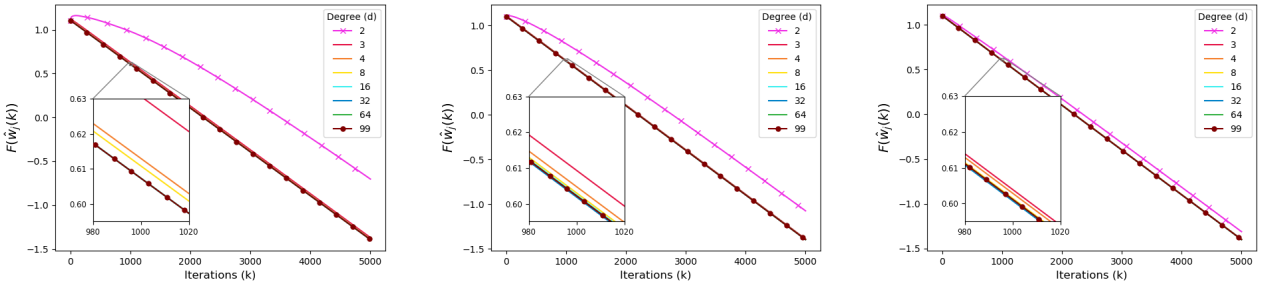
$$\begin{aligned}
\underset{w}{\text{minimize}} \quad & F(w) = \frac{1}{M}\sum_{l=1}^{M} 1 - y^{(l)}x^{(l)}w && (74)\\
\text{subject to} \quad & w \in \mathbb{W}
\end{aligned}$$

where $x^{(l)}$ is a data feature and $y^{(l)} \in \{-1, 1\}$ identifies the class label. Note that there is only one parameter $w \in \mathbb{R}$ ($n = 1$) and, for the moment, the number of nodes equals the size of the dataset ($M = m$). $\Delta\boldsymbol{G}(k) = \boldsymbol{G}(k) - \boldsymbol{G}(k)\frac{\mathbf{1}\mathbf{1}^\intercal}{M}$ is in this case a constant row vector.



(a) Rings with aligned gradient (b) Expanders with aligned gradient (c) Expanders with generic gradient

Figure 7: Objective function (74) evaluated for the worst model $\boldsymbol{w}_j(k)$ versus number of iterations: effect of the topology and its relation with the gradients.



(a) 2X dataset (b) 20X dataset (c) 100X dataset

Figure 8: Objective function (74) evaluated for the worst model $\boldsymbol{w}_j(k)$ versus number of iterations: effect of dataset size. Expanders with generic gradient.

We consider an undirected d-regular graph with $M = 100$ nodes and homogeneous non-negative weights, i.e. $A_{i,j} = 1/(d+1)$ for $(i,j) \in \mathcal{E}$ and $A_{i,j} = 0$ otherwise. Matrix $\boldsymbol{A}$ is then symmetric.

Figure 7(a) presents results for undirected d-regular rings, where each node i is connected to nodes $i-1$ and $i+1$ (hence forming a cycle) as well as the $d-2$ closest nodes on the cycle. Among graphs with degree d, these rings are poorly connected, with a diameter that is order of M/d and low spectral gap $1-|\lambda_2|$. The curves show the evolution of the objective function evaluated at the worst estimate, i.e. $\max_i F(\hat{w}_i(k))$ for different values of the degree d. For each degree, the dataset is chosen selecting a vector $\boldsymbol{u}$ orthogonal to $\mathbf{1}$ and perturbing each of its components by a small amount ζ, so that $\boldsymbol{G} = \boldsymbol{u}^\intercal + \zeta \mathbf{1}^\intercal$ and $\Delta \boldsymbol{G}(k) = \boldsymbol{u}^\intercal$. Because $\mathbf{1}$ and $\boldsymbol{u}$ are orthogonal, it follows that

$$E = \|\boldsymbol{u}^\intercal + \zeta \mathbf{1}^\intercal\|_F^2 = \|\boldsymbol{u}\|_2^2 + \zeta^2 \|\mathbf{1}\|_2^2 = \|\boldsymbol{u}\|_2^2 + M\zeta^2 = E_{\text{sp}} + M\zeta^2.$$

Then, by selecting ζ small enough we can make E_{sp} and E arbitrarily close. This also a full-batch setting, so that $H = \sqrt{E}$. We can select $\boldsymbol{u}$ equal to a left eigenvector of $\boldsymbol{A}$ relative to λ_2, because $\boldsymbol{A}$ has orthogonal eigenvectors among which there is $\mathbf{1}$. In this case, we say that gradients are *aligned with the topology*,[9] If $\boldsymbol{u}$ is a left eigenvector of $\boldsymbol{A}$, all the energy of $\Delta \boldsymbol{G}(k)$ is in the subspace defined by $\boldsymbol{P}_2$, then $\alpha = \sqrt{e_2} = 1$. Details about how the dataset is built are in the sections below, together with calculations showing that the value of the objective function is

$$\max_i (F(\hat{w}_i(k-1))) = 1 + \zeta + \frac{\eta\zeta}{1-\lambda_2}\left(1 - \frac{1-\lambda_2^k}{k(1-\lambda_2)}\right) - \eta\zeta^2\frac{k}{2}. \tag{75}$$

Equation (75) exactly matches the plots in Fig. 7(a). Comparing (75) with (8) and (7), we recognize the same dependence on the second largest eigenvalue.[10] Because $E_{\text{sp}} \approx E$, $E = \sqrt{E}$, and $\alpha(1) = 1$, the two bounds (8)

[9] It corresponds to the slowest convergence scenario for distributed dual averaging, see (Duchi et al., 2012, Prop. 1).

[10] Equation (75) indicates a much faster convergence than those bounds, because, for simplicity, we considered a differentiable linear objective function (74).

and (7) are almost equivalent, but for the fact that (7) correctly takes into account that there is no dependence on the initial estimates ($R_{\text{sp}} = 0$).

Figure 7(a) shows clearly a high variability of the performance of the optimization algorithm across different topologies. The number of iterations required to achieve a given approximation of the optimum is orders of magnitude larger for the cycle ($d = 2$) than for the clique ($d = 99$).

In Fig. 7(b) the same experiments are executed on d-regular connected random graphs. These graphs are known to be Ramanujan graphs with high-probability, i.e. they have the smallest $|\lambda_2|$ among all graphs with the same degree (McKay, 1981). The curves still follow (75), but because λ_2 is smaller, objective function $F(\cdot)$ takes on smaller values. Note that the curves for $d = 2$ and $d = M - 1$ are unchanged, because the corresponding graphs are always the same (the cycle and the clique, respectively). We also observe that the relative performance differences for graphs with $d \geq 3$ are much smaller: the marginal benefit to increase connectivity is much less for random graphs.

Until now, we have assumed that row matrix $\Delta \boldsymbol{G}$ is aligned with a left eigenvector of $\boldsymbol{A}$, but there is no particular reason to think this should be the case. Figure 7(c) shows numerical results for the case when the same datasets as in Fig. 7(b) are used, but the graph is a new independently generated d-regular random graph.[11] In the figure, we see that connectivity is even less important and that the curve for $d = 2$ starts approaching the others. This experiment shows the effect of α. $\boldsymbol{u}$ is the most difficult configuration to average for the consensus matrix $\boldsymbol{A}$. Now vector $\Delta \boldsymbol{G}$ and the vector $\boldsymbol{u}$ are arbitrarily oriented, so that on average only $(1/M)$th of the energy of $\Delta \boldsymbol{G}$ falls in the direction of $\boldsymbol{u}$.

We now look at how gradient variability affects convergence. We increase the dataset size and have each node compute its local function on the basis of more data samples. The additional data is built as above from the second eigenvectors of new independently generated d-regular random graphs. The new data have then the same statistical properties. Figure 8 shows what happens when each node stores respectively 2, 20, and 100 data points. Because the local datasets become more and more similar as dataset size increases, E_{sp} reduces and the curves get closer and closer. For a 100x dataset, the convergence rate of a cycle ($d = 2$) differs little from that of a fully connected graph.

F.1 Datasets' generation

The datasets are generated as follows. We start from a vector $\boldsymbol{u}$ of values in $[-1, 1]$ which sum up to zero. We describe later how the vector $\boldsymbol{u}$ is selected in the different experiments. Features $x^{(l)}$ are defined as $x^{(l)} = |u_l + \zeta|$, where ζ is a small positive constant. The labels are defined as $y^{(l)} = -\operatorname{sign}(u_l + \zeta)$. We select $\mathbb{W} = [-30, 1]$. Observe that $F(w) = 1 - \sum_l y^{(l)} x^{(l)}/M = 1 + \zeta w$, and consequently the minimizer of problem (74) is $w^* = -30$. We select $\zeta = 1/10$, $\eta(k) = 1/10$ and all nodes start with the same initial estimate $w_i(0) = 1$.

We say that gradients are *aligned with the topology*, when vector $\boldsymbol{u}$ is a left eigenvector of $\boldsymbol{A}$ relative to λ_2, normalized so that $\|\boldsymbol{u}\|_\infty = 1$ and $\min_i u_i = -1$. In this case, $\Delta \boldsymbol{G}(k) = \boldsymbol{G}(k) - \boldsymbol{G}(k)\frac{\mathbf{1}\mathbf{1}^\intercal}{M} = \boldsymbol{u}^\intercal$.

F.2 Proof of (75)

In the toy example, the estimates matrix (a $1 \times M$ vector in this case) evolves as

$$\begin{aligned}
\boldsymbol{W}(k) &= \boldsymbol{W}(0)\boldsymbol{A}^k - \sum_{h=0}^{k-1} \eta \boldsymbol{G} \boldsymbol{A}^{k-1-h} \\
&= \mathbf{1}^\intercal - \eta \sum_{h=0}^{k-1} \left(\boldsymbol{u}^\intercal \boldsymbol{A}^{k-1-h} + \zeta \mathbf{1}^\intercal \boldsymbol{A}^{k-1-h} \right) = \mathbf{1}^\intercal - \eta \sum_{h=0}^{k-1} \left(\lambda_2^{k-1-h} \boldsymbol{u}^\intercal + \zeta \mathbf{1}^\intercal \right) \\
&= \mathbf{1}^\intercal - \eta \boldsymbol{u}^\intercal \frac{1 - \lambda_2^k}{1 - \lambda_2} - \eta \zeta k \mathbf{1}^\intercal.
\end{aligned}$$

If $\lambda_2 > 0$, and ζ is small enough, the worst local model is the one of the node, call it j, that stores the data pair

[11] For the case $d = 2$, the original dataset is randomly distributed across the nodes.

$x^{(j)} = |-1+\zeta|$ for which $u_j = -1$. Its local model evolves as:

$$w_j(k) = 1 + \eta\frac{1-\lambda_2^k}{1-\lambda_2} - \eta\zeta k,$$

For node j the time-average model is

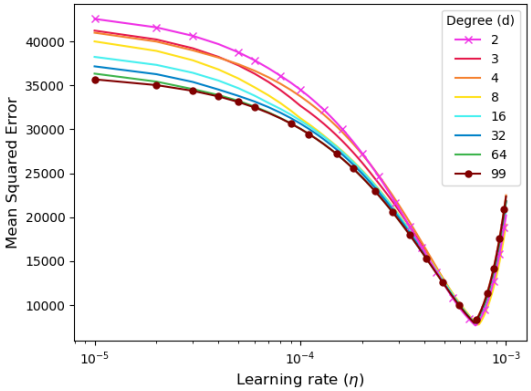
$$\hat{w}_j(k-1) = 1 + \frac{\eta}{1-\lambda_2}\left(1 - \frac{1-\lambda_2^k}{k(1-\lambda_2)}\right) - \eta\zeta\frac{k}{2}.$$

Finally, the objective function is

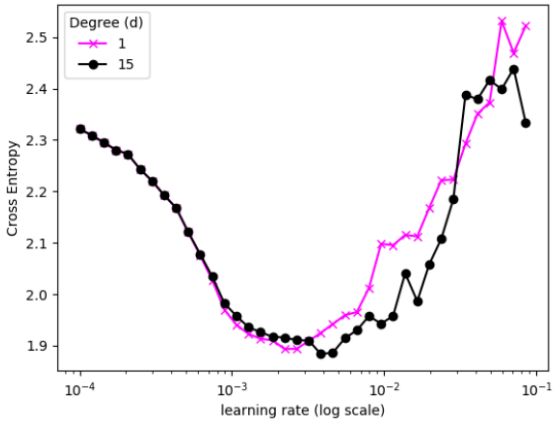
$$\begin{aligned}\max_i(F(\hat{w}_i(k-1))) &= F(\hat{w}_j(k-1)) \\ &= 1 + \zeta + \frac{\eta\zeta}{1-\lambda_2}\left(1 - \frac{1-\lambda_2^k}{k(1-\lambda_2)}\right) - \eta\zeta^2\frac{k}{2}. \end{aligned} \tag{75}$$

Equation (75) exactly matches the plots in Fig. 7(a). This implicitly confirms that $\lambda_2 > 0$.

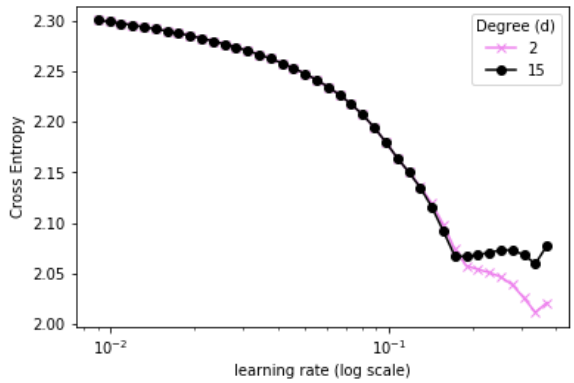
Because $\Delta\boldsymbol{G}(k) = \boldsymbol{u}^\intercal$ is a left eigenvector of $\boldsymbol{A}$ corresponding to λ_2, all the energy of $\Delta\boldsymbol{G}(k)$ is in the subspace defined by $\boldsymbol{P}_2$, then $\alpha = \sqrt{e_2} = 1$.



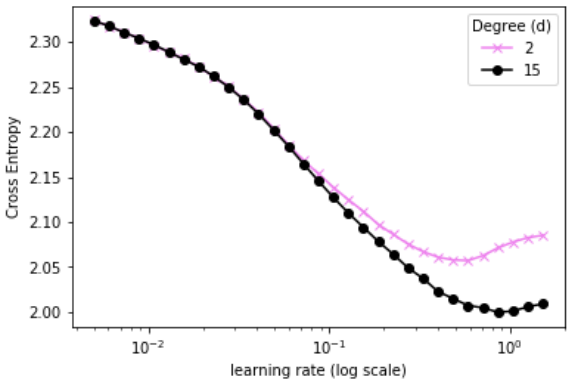
(a) Linear regression on CT dataset, M=100



(b) ResNet18 on MNIST dataset, M=16



(c) 2-conv layers on MNIST, M=16



(d) ResNet18 on CIFAR-10, M=16

Figure 9: Error after one iteration vs learning rate.

G Experiments

We consider two families of regular graphs that we call directed ring lattices, and undirected expanders. In a directed regular ring lattice with degree d, each node i is connected to nodes $(i+1)\%M$, $(i+2)\%M$, ..., $(i+d)\%M$. An undirected regular expander is obtained by generating a large number (200) of regular random graphs (using the NetworkX implementation of the algorithm in (McKay and Wormald, 1990)), and selecting the one with the largest spectral gap. All the experiments presented in the main paper are performed on undirected regular expanders.

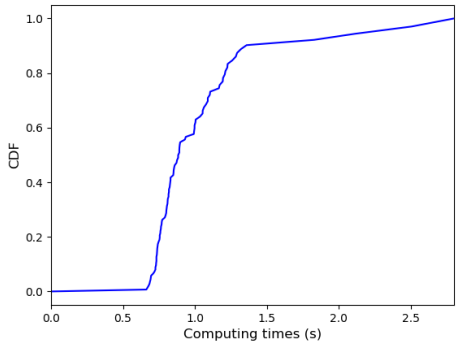
For each experiment, the learning rate has been set using the configuration rule in (Smith, 2017). We increase geometrically the learning rate and evaluate the training loss after one iteration. We then determine two "knees" in the loss versus learning rates plots: the learning rate value for which the loss starts decreasing significantly and the value for which it starts increasing again. We set the learning rate to the geometric average of these two values. As Fig. 9 shows, for a given experiment, this procedure leads to select the same learning rate independently of the degree of the topology, respectively equal to $\eta = 3 \times 10^{-4}$ for the linear regression (Fig. 9(a)), $\eta = 6 \times 10^{-4}$ for ResNet18 on MNIST dataset (Fig. 9(b)), $\eta = 0.1$ for 2-conv layers on MNIST dataset (Fig. 9(c)) and $\eta = 0.05$ for ResNet18 on CIFAR-10 dataset (Fig. 9(d)).

In comparison to Table 1, Table 2 also shows the predictions for an "intermediate" bound, obtained by replacing in (8) R with R_{sp}. These predictions are denoted by k''_o. Comparing k''_o with k'_o and k'_n reveals that the difference between k'_o and k'_n is mainly due to the effect of E_{sp}, H, and α, while R_{sp} plays a less important role.

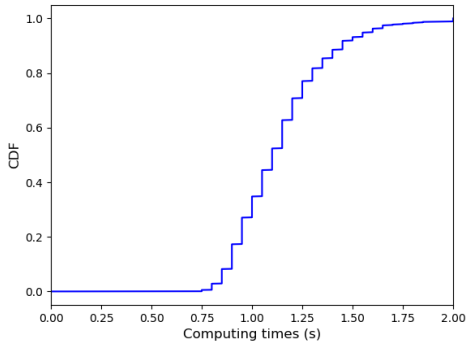
Figures 10(a) and 10(b) show the Cumulative Distribution Functions of computing times for the Spark cluster and for ASCI Q super-computer, respectively.

Table 2: Empirical estimation of E, E_{sp}, H, α on different ML problems and comparison of their joint effect (β) with the value $\widehat{\beta}$ predicted through (12). Number of iterations by which training losses for the ring and the clique differ by 4%, 10%, as predicted by the old bound (8), k'_o, by the old bound (8) where R is replaced by R_{sp}, k''_o, by the new one (7), k'_n, and as measured in the experiment, k'. When their values exceeds the total number of iterations we ran (respectively 1200 for CT, 1190 for MNIST, and 1040 for CIFAR-10), we simply indicate it as ∞.

Dataset	Model	M	B	$\sqrt{E/E_{sp}}$	$\sqrt{E}/H$	$\frac{1}{\alpha}$	β	$\widehat{\beta}$	@4%				@10%			
									k'_o	k''_o	k'_n	k'	k'_o	k''_o	k'_n	k'
CT (S=52000)	Linear regr. n=384	16	128	7.92	1.01	1.53	12.23	12.31	1	2	∞	∞	1	5	∞	∞
			3250	38.45	1.00	1.64	62.86	60.97	1	2	∞	∞	1	5	∞	∞
		100	128	7.75	1.01	1.54	12.05	11.56	1	2	10	∞	1	4	∞	∞
			520	15.58	1.00	1.51	23.60	22.96	1	2	17	∞	1	4	∞	∞
MNIST (S=60000)	2-conv layers n=431080	16	128	1.45	1.42	1.49	3.07	2.92	1	5	16	∞	1	11	72	∞
			500	2.15	1.14	1.53	3.75	3.71	1	6	22	40	1	14	260	∞
		64	128	1.41	1.42	1.51	3.02	3.03	1	5	10	∞	1	11	24	∞
split by digit		10	500	1.01	1.00	1.42	1.42	3.62	1	2	3	60	1	3	7	100
CIFAR-10 (S= 50000)	ResNet18 n=11173962	16	128	1.07	2.85	1.46	4.45	3.05	1	2	7	70	1	4	20	∞
			500	1.19	1.83	1.47	5.11	4.14	1	10	30	70	1	20	∞	∞



(a) Spark cluster

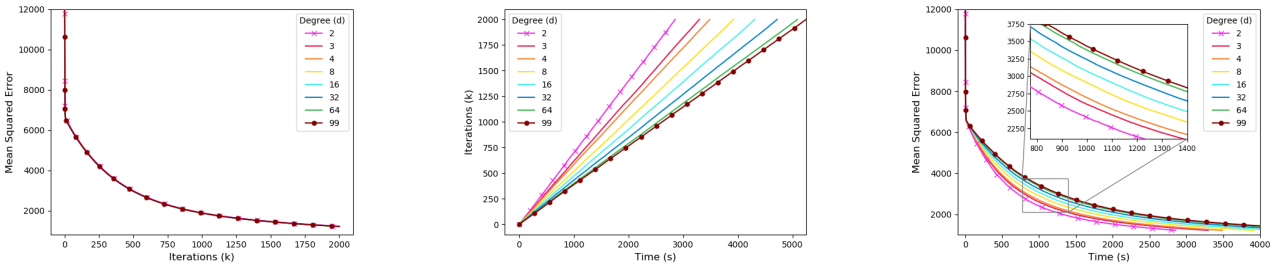


(b) ASCI Q

Figure 10: Empirical distribution of the computation times.

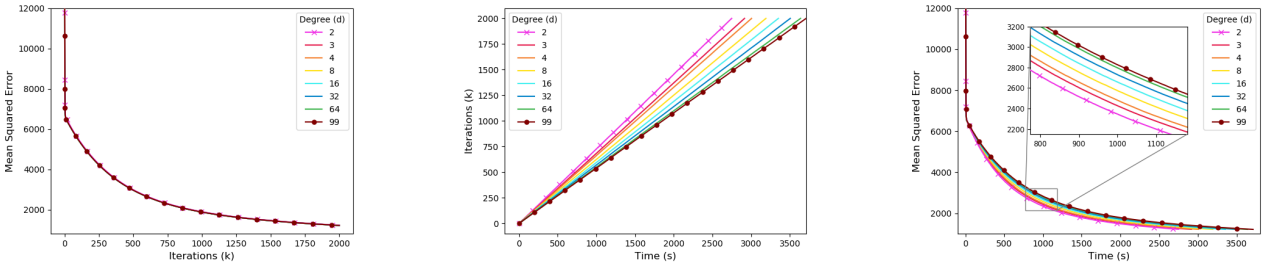
The following experiments have been carried out:

1. Linear regression on CT dataset on undirected expanders with computation time distribution from the Spark cluster (Fig. 11) and from ASCI Q super-computer (Fig. 12).
2. ResNet18 on MNIST dataset on directed ring lattices with computation time distribution from the Spark cluster (Fig. 13).
3. 2-conv layers on MNIST dataset on undirected expanders with computation time distribution from ASCI Q super-computer (Fig. 14)
4. ResNet18 on CIFAR-10 dataset on undirected expanders with computation time distribution from the Spark cluster (Fig. 15) and from ASCI Q super-computer (Fig. 16).



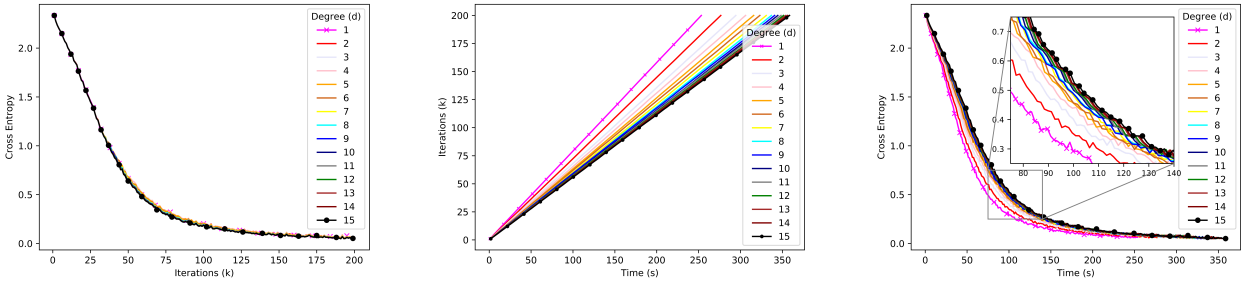
(a) Error vs iterations (b) Throughput (c) Error vs time

Figure 11: Effect of network connectivity (degree d) on the convergence for linear regression on dataset CT with computation times from a Spark cluster. M = 100, B = 128.



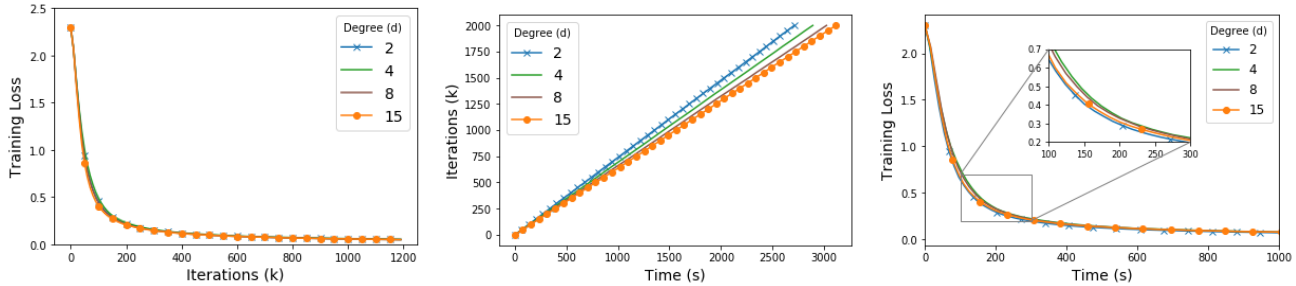
(a) Error vs iterations (b) Throughput (c) Error vs time

Figure 12: Effect of network connectivity (degree d) on the convergence for linear regression on dataset CT with computation times from ASCI-Q super-computer. M = 100, B = 128.



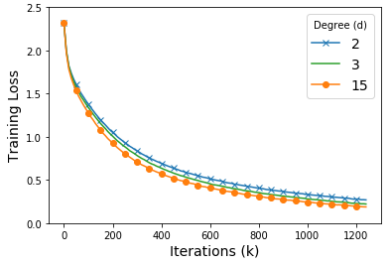
(a) Error vs iterations (b) Throughput (c) Error vs time

Figure 13: Effect of network connectivity (degree d) on the convergence for ResNet18 on dataset MNIST with computation times from a Spark cluster. M = 16, B = 500.

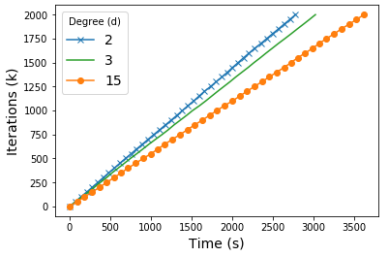


(a) Error vs iterations (b) Throughput (c) Error vs time

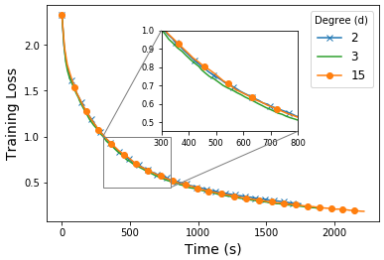
Figure 14: Effect of network connectivity (degree d) on the convergence for 2-conv layers on dataset MNIST with computation times from ASCI-Q super-computer. M = 16, B = 500.



(a) Error vs iterations

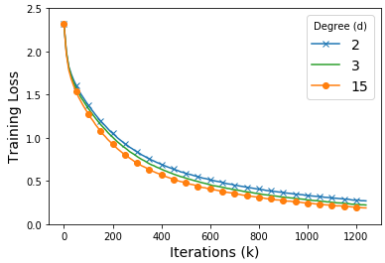


(b) Throughput

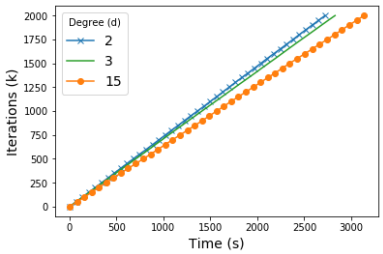


(c) Error vs time

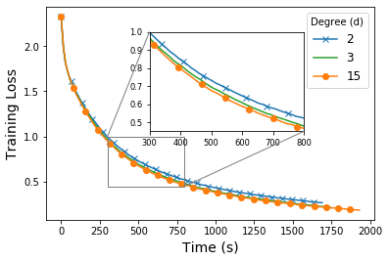
Figure 15: Effect of network connectivity (degree d) on the convergence for CIFAR-10 with computation times from a spark cluster. M = 16, B = 128.



(a) Error vs iterations



(b) Throughput



(c) Error vs time

Figure 16: Effect of network connectivity (degree d) on the convergence for CIFAR-10 with computation times from ASCI-Q super computer. M = 16, B = 128.