Improved Stability and Generalization Guarantees of the Decentralized SGD Algorithm

Batiste Le Bars [1] **Aurélien Bellet** [2] **Marc Tommasi** [3] **Kevin Scaman** [1] **Giovanni Neglia** [4]

Abstract

This paper presents a new generalization error analysis for Decentralized Stochastic Gradient Descent (D-SGD) based on algorithmic stability. The obtained results overhaul a series of recent works that suggested an increased instability due to decentralization and a detrimental impact of poorly-connected communication graphs on generalization. On the contrary, we show, for convex, strongly convex and non-convex functions, that D-SGD can always recover generalization bounds analogous to those of classical SGD, suggesting that the choice of graph does not matter. We then argue that this result is coming from a worst-case analysis, and we provide a refined optimization-dependent generalization bound for general convex functions. This new bound reveals that the choice of graph can in fact improve the worst-case bound in certain regimes, and that surprisingly, a poorly-connected graph can even be beneficial for generalization.

1. Introduction

Studying the ability of machine learning models to generalize to unseen data is a fundamental and long-standing objective. Among the several approaches that have been proposed to bound generalization errors, the most prominent ones are based on the complexity of the hypothesis class like the Vapnik-Chervonenkis dimension or Rademacher complexity (Bousquet et al., 2004), algorithmic stability (Bousquet & Elisseeff, 2002), PAC-Bayesian bounds (Shawe-Taylor & Williamson, 1997; McAllester, 1998; Catoni, 2007; Alquier et al., 2024), or more recently information-theoretic generalization bounds (Xu & Raginsky, 2017).

[1]Inria Paris - Ecole Normale Supérieure, PSL Research University [2]Inria, Université de Montpellier [3]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189, CRIStAL, F-59000 Lille [4]Inria, Université Côte d'Azur. Correspondence to: Batiste Le Bars <batiste.le-bars@inria.fr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Over the last few years, a substantial amount of work has been dedicated to the study of the generalization properties of *optimization algorithms*, more specifically gradient-based methods (Lin et al., 2016; London, 2017; Zhou et al., 2018; Amir et al., 2021; Neu et al., 2021; Scaman et al., 2024). In particular, since the seminal work of Hardt et al. (2016), approaches based on *algorithmic stability* have encountered a large success as they allow to shed light on the implicit regularization brought by (stochastic) gradient methods (Kuzborskij & Lampert, 2018; Bassily et al., 2020; Lei & Ying, 2020; Schliserman & Koren, 2022). However, this large amount of work is mostly focusing on *centralized* gradient-based algorithms.

Decentralized learning algorithms, such as the celebrated Decentralized Stochastic Gradient Descent (D-SGD) algorithm (Nedic & Ozdaglar, 2009), allow several agents to train models on their local data by exchanging model updates rather than the data itself. In D-SGD, agents solve an empirical risk minimization task by alternating between computing local gradient steps and averaging model parameters with their neighbors in a communication graph. A sparser graph thus reduces the per-iteration communication cost but tends to increase the number of iterations needed to converge. Most theoretical analyses of D-SGD and its variants focus on understanding the *optimization error* by characterizing the convergence rate to the empirical risk minimizer. They notably highlight the impact of the communication graph and data heterogeneity across agents (Koloskova et al., 2020; Neglia et al., 2020; Ying et al., 2021; Le Bars et al., 2023). In contrast, the *generalization error* of decentralized learning algorithms is far less understood.

In the work of Richards & Rebeschini (2020), the authors focus on a specific variant of D-SGD (thereafter referred to as Variant A, see Algorithm 1) where the agents perform a local stochastic gradient update *before* averaging their parameters with their neighbors. With an analysis based on algorithmic stability, they come to the conclusion that for convex functions the decentralization does not have any impact, recovering the same generalization bounds as those obtained by Hardt et al. (2016) for centralized SGD. In contrast, a more recent line of work (Sun et al., 2021; Zhu et al., 2022; Taheri & Thrampoulidis, 2023) has investigated this question for a more practically relevant variant of D-SGD (thereafter referred to as Variant B, see Algorithm 1), where this time the averaging step and the local gradient computation are done *in parallel*. The existing generalization results for this variant, which is known to be more difficult to analyze (see Richards & Rebeschini, 2020, Remark 4 therein), mainly differ in their technical assumptions. Sun et al. (2021) consider Lipschitz and smooth loss functions, Zhu et al. (2022) focus on smooth and convex losses, while Taheri & Thrampoulidis (2023) investigate the overparameterized regime, for convex and Polyak-Lojasiewicz functions. Despite these differences, these three studies all come to the same conclusion: "decentralization has a negative impact on generalization." Specifically, their generalization bounds get larger as the graph gets sparser, and eventually become vacuous for non-connected graphs. In summary, the current literature exhibits strikingly contrasting generalization properties for these two common D-SGD variants, despite their similar optimization performance. While questioning this gap is the main motivation for our work, we also wonder how generalization upper bounds like those obtained by Richards & Rebeschini (2020) can be completely independent of the communication graph. This result is indeed rather counter-intuitive, as we know how important the choice of communication graph can be for optimization (Neglia et al., 2020).

1.1. Contributions

In this work, we focus on the more complex Variant B and prove that the dichotomy between the two variants of D-SGD is only apparent. We show that they are in fact equivalent in terms of stability and generalization performance, improving upon the recent conclusions of Sun et al. (2021); Zhu et al. (2022) and Taheri & Thrampoulidis (2023). Our contributions, summarized in Table 1, are the following:

(1) In Section 3, we first consider convex and strongly convex loss functions and show that we can recover, for Variant B of D-SGD, the exact same generalization upper-bounds than those of Variant A (Richards & Rebeschini, 2020) and standard SGD (Hardt et al., 2016). This leads to the conclusion that, contrary to the optimization error, the choice of graph and in particular **poorly-connected graphs do not have a detrimental impact on generalization**.

(2) We then consider in Section 4 the less studied case of non-convex functions, which, contrary to convex ones, were not considered by Richards & Rebeschini (2020) for Variant A of D-SGD. Again, we show that, for both variants, it is possible to recover almost identical generalization upper-bounds as those obtained by Hardt et al. (2016) for SGD, leading to the same conclusions as for convex cases regarding the lack of impact of decentralization on generalization.

(3) We finally argue in Section 5 that our analysis, as well as the one of Richards & Rebeschini (2020), characterize

Table 1. Simplified generalization bounds for (D)-SGD with Lipschitz and smooth loss functions. [H] indicates the results of Hardt et al. (2016), [R] those of Richards & Rebeschini (2020), and [S] those of Sun et al. (2021). For simplicity, we omit constant factors. T is the number of iterations, m the number of agents, n the number of local data points, and $\rho \in [0, 1]$ the spectral gap of the communication graph. We also have $a \in (0, 1)$ a constant that depends on the model parameters and C_ρ a constant that depends on ρ. For centralized SGD, we consider that the algorithm is run over mn data points. We refer to Section 2.2 for the definitions of Variants A and B of D-SGD.

	SGD	D-SGD	
		Variant A	Variant B
Convex	$\frac{T}{mn}$ [H]	$\frac{T}{mn}$ [R]	$\frac{T}{mn} + \frac{T}{\rho}$ [S] $\frac{T}{mn}$ **[ours]**
μ-Strongly convex	$\frac{1}{\mu mn}$ [H]	$\frac{1}{\mu mn}$ [R]	$\frac{1}{\mu mn} + \frac{1}{\mu \rho}$ [S] $\frac{1}{\mu mn}$ **[ours]**
Non-convex	$\frac{T^a}{mn}$ [H]	$\frac{T^a}{m^{1-a}n}$ **[ours]**	$\frac{T^a}{n} + C_\rho T^a$ [S][1] $\frac{T^a}{m^{1-a}n}$ **[ours]**

"worst-case" generalization guarantees across many possible losses and data distributions. One may then wonder if the communication graph can play a role under more specific losses or distributions. To address this point, we propose a refined analysis for convex functions, inspired by optimization-dependent generalization bounds for classical SGD (Kuzborskij & Lampert, 2018; Lei & Ying, 2020), which confirms this is indeed the case. Quite surprisingly, our new bound not only shows that in low-noise regimes specific choice of graphs can improve the worst-case bound, but also that **poorly-connected graphs can even be beneficial to generalization**.

Before moving to our main contributions, the following section provides relevant background on the relationship between algorithmic stability and generalization, presents D-SGD, and discusses the main assumptions considered throughout the paper.

2. Background

2.1. Stability and Generalization in Decentralized Learning

We consider the general setting of statistical learning, adapted to a decentralized framework with m agents. We consider that agent k observes examples drawn from a local data distribution $\mathcal{D}_k$ with support $\mathcal{Z}$. The objective is to find a global model $\theta \in \mathbb{R}^d$ minimizing the *population risk*

[1]In the original paper (Sun et al., 2021), the authors give a bound in $\mathcal{O}(\frac{T^a}{mn} + C_\rho T^a)$. In Appendix E.3, we reveal that there is a mistake in their proof, corrected in Table 1.

defined by:

$$R(\theta) \triangleq \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}_{Z\sim\mathcal{D}_k}[\ell(\theta;Z)] \ ,$$

where ℓ is some loss function. We denote by $\theta^\star$ a global minimizer of the population risk, i.e., $\theta^\star \in \arg\min_\theta R(\theta)$.

Although we cannot evaluate the population risk $R(\theta)$, we have access to an empirical counterpart, computed over m local datasets $S \triangleq (S_1,\dots,S_m)$ where $S_k = \{Z_{1k},\dots,Z_{nk}\}$ is the dataset of agent k with $Z_{ik}\sim\mathcal{D}_k$. Note that for simplicity we consider that all local datasets are of same size n, but our analysis can be extended to the heterogeneous case. The resulting *empirical risk* is given by:

$$R_S(\theta) \triangleq \frac{1}{m}\sum_{k=1}^{m}R_{S_k}(\theta) \triangleq \frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n}\ell(\theta;Z_{ik}) \ .$$

One of the most famous and studied estimators is the empirical risk minimizer, denoted by $\widehat{\theta}_{\text{ERM}} \triangleq \arg\min_\theta R_S(\theta)$. However, in most situations, this estimator cannot be directly computed. Instead, one relies on a potentially random *decentralized optimization* algorithm A, taking as input the full dataset S, and returning an approximate minimizer $A(S)\in\mathbb{R}^d$ of the empirical risk $R_S(\theta)$.

In this setting, we can upper-bound the expected *excess risk* $R(A(S)) - R(\theta^\star)$ by the sum of the (expected) *generalization* error (ϵ_{gen}), and the (expected) *optimization error* (ϵ_{opt}):

$$\mathbb{E}_{A,S}[R(A(S)) - R(\theta^\star)] \leq \epsilon_{\text{gen}} + \epsilon_{\text{opt}} \ ,$$

where $\epsilon_{\text{gen}} \triangleq \mathbb{E}_{A,S}[R(A(S)) - R_S(A(S))]$ and $\epsilon_{\text{opt}} \triangleq \mathbb{E}_{A,S}[R_S(A(S)) - R_S(\widehat{\theta}_{\text{ERM}})]$. The present work focuses on the control of the expected generalization error ϵ_{gen}, for which a popular approach is based on the stability analysis of the algorithm A.[2]

Contrary to a large body of works using the well-known *uniform stability* (Bousquet & Elisseeff, 2002; Shalev-Shwartz et al., 2010), our analysis relies on the notion of *on-average model stability* (Lei & Ying, 2020) which has the advantage to give tighter bounds in our analysis. Below, we recall this notion, with a slight adaptation to the decentralized setting.

Definition 2.1. *(On-average model stability).* Let $S = (S_1,\dots,S_m)$ with $S_k = \{Z_{1k},\dots,Z_{nk}\}$ and $\tilde{S} = (\tilde{S}_1,\dots,\tilde{S}_m)$ with $\tilde{S}_k = \{\tilde{Z}_{1k},\dots,\tilde{Z}_{nk}\}$ be two independent copies such that $Z_{ik}\sim\mathcal{D}_k$ and $\tilde{Z}_{ik}\sim\mathcal{D}_k$. For any $i\in\{1,\dots,n\}$ and $j\in\{1,\dots,m\}$, let us denote by $S^{(ij)} = (S_1,\dots,S_{j-1},S_j^{(i)},S_{j-1},\dots,S_m)$, with $S_j^{(i)} = \{Z_{1j},\dots,Z_{i-1j},\tilde{Z}_{ij},Z_{i+1j},\dots,Z_{nj}\}$, the dataset formed from S by replacing the i-th element of the j-th agent's dataset by $\tilde{Z}_{ij}$. A randomized algorithm A is said to be *on-average model ε-stable* if

$$\mathbb{E}_{S,\tilde{S},A}\Big[\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}||A(S) - A(S^{(ij)})||_2\Big] \leq \varepsilon \ . \quad (1)$$

A key aspect of on-average model stability is that it can directly be linked to the generalization error, as shown in the following lemma.

Lemma 2.2. (Generalization via on-average model stability (Lei & Ying, 2020)). *Let A be on-average model ε-stable. Then, if $\ell(\cdot;z)$ is L-Lipschitz for all $z\in\mathcal{Z}$ (see Assumption 2.4), we have $|\mathbb{E}_{A,S}[R(A(S)) - R_S(A(S))]| \leq L\varepsilon$.*

Thanks to this lemma, it suffices to control the on-average model stability of the decentralized algorithm A, in order to get the desired generalization bound.

2.2. Decentralized SGD

Throughout this paper, we focus on the popular Decentralized Stochastic Gradient Descent (D-SGD) algorithm (Nedic & Ozdaglar, 2009; Lian et al., 2017), which aims to find minimizers (or saddle points) of the empirical risk $R_S(\theta)$ in a fully decentralized fashion. This algorithm is based on peer-to-peer communications between agents, where a graph is used to encode which pairs of agents (also referred to as nodes) can interact together. More specifically, this *communication graph* is represented by a weight matrix $W\in[0,1]^{m\times m}$, where $W_{jk}>0$ gives the weight that agent j gives to messages received from agent k, while $W_{jk}=0$ (no edge) means that j does not receive messages from k.

Algorithm 1 Decentralized SGD (Lian et al., 2017)

Input: Initialize $\forall k$, $\theta_k^{(0)} = \theta^{(0)}\in\mathbb{R}^d$, iterations T, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, weight matrix W.
for $t = 0,\dots,T-1$ **do**
  for each node $k = 1,\dots,m$ **do**
    Sample $I_k^t \sim \mathcal{U}\{1,\dots,n\}$
    Variant A:
      $\theta_k^{(t+1)} \leftarrow \sum_{l=1}^{m} W_{kl}\left(\theta_l^{(t)} - \eta_t\nabla\ell(\theta_l^{(t)};Z_{I_l^t l})\right)$
    Variant B:
      $\theta_k^{(t+1)} \leftarrow \sum_{l=1}^{m} W_{kl}\theta_l^{(t)} - \eta_t\nabla\ell(\theta_k^{(t)};Z_{I_k^t k})$
  end for
end for

D-SGD is summarized in Algorithm 1. As mentioned in the introduction, there exists two main variants of this algorithm. In Variant A, each agent k first performs a stochastic

[2] While we focus here on the *expected* version of the generalization error, some of these tools are also well-suited to provide *high-probability* generalization bounds (Feldman & Vondrak, 2019).

gradient update based on $\nabla \ell(\theta_k^{(t)}; Z_{I_k^t k})$, i.e., the stochastic gradient of ℓ evaluated at $\theta_k^{(t)}$ with $I_k^t \sim \mathcal{U}\{1, \ldots, n\}$ the index of the data point uniformly selected by agent k from its local dataset S_k at iteration t. Then, it aggregates its updated parameter vector with its neighbors according to the weight matrix W. Variant B somehow reverses the two steps: each agent first aggregates its parameter vector with its neighbors and then performs a stochastic gradient update based on the parameter vector it had before the averaging step.

Variant B is often seen as more efficient because the aggregation step and the stochastic gradient calculation can be done in *parallel*, while in Variant A, the aggregation step cannot be performed before all agents have finished their stochastic gradient update. From an optimization perspective however, both variants are guaranteed to converge at the same rate (Lian et al., 2017).

The state of the art currently suggests that these two variants diverge mainly in terms of generalization error. While Richards & Rebeschini (2020) focus on Variant A and show that we can recover the same generalization bounds as centralized SGD, Sun et al. (2021); Zhu et al. (2022) and Taheri & Thrampoulidis (2023) focus on Variant B and show an increased instability due to decentralization. One of our main contributions is to fill this gap in the current theory, by showing that Variant B can also reach the same generalization error as centralized SGD, making it equivalent to Variant A.

Remark 2.3. Due to its fully decentralized nature, D-SGD (both variants) outputs m different parameters $A_1(S) \triangleq \theta_1^{(T)}, \ldots, A_m(S) \triangleq \theta_m^{(T)}$ at the end of the optimization process (one per agent). For this reason, the stability and generalization analysis of the next sections will not be made with respect to a single output $A(S)$ as described in Section 2.1, but rather with respect to (one of) these different outputs.

2.3. Main Assumptions

We focus on the classic setup of Hardt et al. (2016), also considered by Richards & Rebeschini (2020) and Sun et al. (2021) in prior work on the generalization analysis of D-SGD. These works rely on the standard assumptions of L-Lipschitzness and β-smoothness of the loss function.

Assumption 2.4. *(L-Lipschitzness).* We assume that the loss function ℓ is differentiable w.r.t. θ and uniformly Lipschitz, i.e., $\exists L > 0$ such that $\forall \theta, \theta' \in \mathbb{R}^d, z \in \mathcal{Z}$, $|\ell(\theta; z) - \ell(\theta'; z)| \leq L\|\theta - \theta'\|_2$, or equivalently, $\|\nabla \ell(\theta; z)\|_2 \leq L$.

Assumption 2.5. *(β-smoothness).* The loss function ℓ is β-smooth i.e. $\exists \beta > 0$ such that $\forall \theta, \theta' \in \mathbb{R}^d, z \in \mathcal{Z}$, $\|\nabla \ell(\theta; z) - \nabla \ell(\theta'; z)\|_2 \leq \beta \|\theta - \theta'\|_2$.

Remark 2.6. By considering Lipschitz and smooth loss functions, our results will be directly comparable to those of Hardt et al. (2016); Richards & Rebeschini (2020) and Sun et al. (2021). Nevertheless, we expect the conclusions of this paper to be the same for the analyses with relaxed hypotheses (Zhu et al., 2022; Taheri & Thrampoulidis, 2023). We leave this study for future research.

Our last assumption concerns the weight matrix W. It is again very standard and used extensively in the literature of decentralized optimization (see e.g., Lian et al., 2017; Koloskova et al., 2020).

Assumption 2.7. *(Mixing matrix).* W is doubly stochastic, i.e., $W^T\mathbf{1} = W\mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is the vector (of size m) that contains only ones.

Note that contrary to what is usually considered in the literature, we do not assume the communication graph W to be connected. As an example, we allow W to be the identity matrix, which would reduce D-SGD to m independent local SGD algorithms.

3. Generalization Error for Convex Loss Functions

This section presents our first main contribution. Focusing on convex and strongly convex functions, we first demonstrate that we can recover the exact same generalization upper bounds for Variant A (obtained by Richards & Rebeschini, 2020) and Variant B of D-SGD. Hence, our bounds unify these two variants and contradict (and greatly improve upon) the recent results of Sun et al. (2021); Zhu et al. (2022); Taheri & Thrampoulidis (2023). These authors suggested that the generalization error in D-SGD (Variant B) was adversely affected by sparse communication graphs. Crucially, our analysis demonstrates that the generalization error of D-SGD, regardless of the variant, is in fact not impacted by the choice of communication graph, or by decentralization at all, as our bounds align closely with those established by Hardt et al. (2016) for *centralized* SGD.

3.1. General Convexity

Theorem 3.1. *Assume that the loss function $\ell(\cdot; z)$ is convex, L-lipschitz (Assumption 2.4) and β-smooth (Assumption 2.5). Let $A_1(S) = \theta_1^{(T)}, \ldots, A_m(S) = \theta_m^{(T)}$ be the m final iterates of D-SGD (Variant B) run for T iterations, with mixing matrix W satisfying Assumption 2.7 and with $\eta_t \leq \frac{2\min_k\{W_{kk}\}}{\beta}$. Then, $\forall k = 1, \ldots, m$, $A_k(S)$ has a bounded expected generalization error:*

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{2L^2 \sum_{t=0}^{T-1} \eta_t}{mn} . \quad (2)$$

Sketch of proof (see Appendix B.1 for details). Prior results (Sun et al., 2021; Taheri & Thrampoulidis, 2023) are suboptimal because they try to mimic state-of-the-art *optimization error* analyses (Kong et al., 2021) which require to control a *consensus distance* term $\sum_k \|\theta_k^{(t)} - \bar{\theta}^{(t)}\|^2$, where

$\bar{\theta}^{(t)} = (1/m)\sum_k \theta_k^{(t)}$. This term, important to ensure the minimization of the empirical risk, is small when all local parameters are close to one another, which is the case only if the communication graph is sufficiently connected. Our proof relies on a tighter analysis, which does not require the control of such consensus distance term.

Denote by $A_k(S) = \theta_k^{(T)}$ and $A_k(S^{(ij)}) = \tilde{\theta}_k^{(T)}$, the final iterates of agent k for D-SGD (Variant B) run over two data sets S and $S^{(ij)}$ that differ only in the i-th sample of agent j (see Def. 2.1 for notations). The objective is to control the on-average model stability $\frac{1}{mn}\sum_{i,j}\mathbb{E}[\delta_k^{(T)}(i,j)]$, with $\delta_k^{(T)}(i,j) = \|\theta_k^{(T)} - \tilde{\theta}_k^{(T)}\|_2$, and then apply Lemma 2.2 to conclude.

The crux of the proof is to recognize, in the updates of Variant B, the gradient updates of a classical SGD with step-size η_t/W_{kk}, and then use its 1-expansivity property (Lemma A.3 in Appendix A) when $\eta_t \leq 2\min_k\{W_{kk}\}/\beta$ to obtain the recursion

$$\begin{aligned}\mathbb{E}[\delta_k^{(t+1)}(i,j)] &\leq \sum_{l=1}^m W_{kl}\mathbb{E}[\delta_l^{(t)}(i,j)] \\ &\quad + \frac{2\eta_t}{n}\mathbb{E}[\|\nabla\ell(\theta_k^{(t)};Z_{ij})\|_2]\mathbb{1}_{\{k=j\}} \\ &\leq \sum_{l=1}^m W_{kl}\mathbb{E}[\delta_l^{(t)}(i,j)] + \frac{2L\eta_t}{n}\mathbb{1}_{\{k=j\}}\ , \end{aligned} \quad (3)$$

where the second inequality is obtained by bounding the gradient norm by L (Assumption 2.4). Having established Eq. (3), the remainder of our proof proceeds along a path similar to that taken by Richards & Rebeschini (2020) in their proof for Variant A. We can recursively apply Eq. (3) until $t = 0$ and use the fact that $\delta_k^{(0)}(i,j) = 0$ for all k to get $\mathbb{E}[\delta_k^{(T)}(i,j)] \leq \frac{2L}{n}\sum_{t=0}^{T-1}(W^{T-t-1})_{kj}\eta_t$. Averaging over i and j and using the fact that any power of W is also doubly stochastic, we obtain that the on-average model stability is upper bounded by $\frac{2L\sum_{t=0}^{T-1}\eta_t}{mn}$, which concludes the proof with a direct application of Lemma 2.2. □

Richards & Rebeschini (2020, Lemma 13) and Theorem 3.1 demonstrate that, for convex functions, the generalization bounds for both Variants A and B of D-SGD are identical—including constant factors—to those obtained by Hardt et al. (2016) for centralized SGD over mn data points. Moreover, this bound is optimal in the centralized setting (Zhang et al., 2022). The only difference resides in the fact that, for Variant B, we need to take smaller stepsizes (below $2\min_k\{W_{kk}\}/\beta$ in Theorem 3.1 compared to $2/\beta$ for the others). This difference stems from the greater difficulty in linking the iterates of Variant B to those of a standard gradient descent, but is rather mild, as the assumptions to ensure convergence of the associated optimization problem are usually stronger (see Appendix E.2 for more details).

Overall, our theorem strictly improves upon the recent result of the closest work (Sun et al., 2021), which obtained, for Variant B of D-SGD, an upper-bound with an extra additive term: $\frac{2L^2\sum_{t=0}^{T-1}\eta_t}{mn} + \mathcal{O}(\frac{T}{\rho})$, where $\rho \in [0,1]$ is the spectral gap of W. This earlier result suggested that the generalization error is significantly influenced by the connectivity of W, specifically suggesting that the error would diverge as the graph becomes sparser ($\rho \to 0$). However, our findings contradict this claim. Remarkably, the bound by Sun et al. (2021) does not exhibit consistency, as it fails to approach zero even as n increases.

Remark 3.2. In Sun et al. (2021), but also Zhu et al. (2022) and Taheri & Thrampoulidis (2023), the authors control the generalization error of the averaged final models $\bar{\theta}^{(T)} = (1/m)\sum_{k=1}^m \theta_k^{(T)}$, but not of individual final models. In this sense, our result is stronger, for two key reasons. Firstly, Theorem 3.1 and all results in this paper can be directly extended to the average model $\bar{\theta}^{(T)}$, as detailed in Appendix E.1. Secondly, this average model is not computed at any stage of the D-SGD process (except where the communication graph is complete). Therefore, examining the generalization properties of this parameter introduces a certain conflict with the fully decentralized context.

3.2. Strong Convexity

We now consider strongly convex functions. As such functions cannot be Lipschitz (Assumption 2.4) over $\mathbb{R}^d$, we restrict our analysis to the optimization over a convex compact set Θ as done by Hardt et al. (2016). Denoting by $\Pi_\Theta(\tilde{\theta}) = \arg\min_{\theta\in\Theta}\|\tilde{\theta} - \theta\|$ the Euclidean projection onto Θ, we consider the *projected* extension of the D-SGD algorithm, which replaces the updates from Algorithm 1 by:

$$\theta_j^{(t+1)} \leftarrow \begin{cases} \sum_{k=1}^m W_{jk}\Pi_\Theta\left(\theta_k^{(t)} - \eta_t\nabla\ell(\theta_k^{(t)};Z_{I_k^t k})\right) & \text{(A)} \\ \Pi_\Theta\left(\sum_{k=1}^m W_{jk}\theta_k^{(t)} - \eta_t\nabla\ell(\theta_j^{(t)};Z_{I_j^t j})\right) & \text{(B)} \end{cases}$$

Theorem 3.3. *Assume that the loss function $\ell(\cdot;z)$ is μ-strongly convex, L-Lipschitz over Θ (Assumption 2.4) and β-smooth (Assumption 2.5). Let $A_1(S) = \theta_1^{(T)}, \ldots, A_m(S) = \theta_m^{(T)}$ be the m final iterates of the projected D-SGD (Variant B) run for T iterations, with mixing matrix W satisfying Assumption 2.7 and with constant stepsize $\eta \leq \min_k\{W_{kk}\}/\beta$. Then, $\forall k = 1,\ldots,m$, $A_k(S)$ has a bounded expected generalization error:*

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{4L^2}{\mu mn}\ . \quad (4)$$

The proof of Theorem 3.3, provided in Appendix B.2, essentially follows the same scheme as the one derived above for convex functions. Once again, the bound matches the optimal one obtained for centralized SGD with strongly convex functions in Hardt et al. (2016), and the one obtained by Richards & Rebeschini (2020) for Variant A of D-SGD. Notice that, contrary to the general convex case, the generalization bound for strongly convex functions is independent of the number of iterations T, which makes these problems more stable and less likely to overfit.

Sun et al. (2021) derive a similar generalization bound but with an extra additive error term in $\mathcal{O}(\frac{1}{\mu\rho})$. Their bound is therefore strictly weaker: it is *not* converging to 0 as the number of samples increases and is vacuous when the communication graph is not connected ($\rho = 0$). This again illustrates the suboptimality of these previous results and the major improvement brought by ours.

3.3. Deriving excess risk bounds

Recall from Section 2.1 that the main objective of statistical learning is to control the excess risk $\epsilon_{\text{excess}} \triangleq \mathbb{E}_{A,S}[R(A(S)) - R(\theta^\star)]$, which can be upper-bounded by the sum of the generalization error (ϵ_{gen}) and the optimization error (ϵ_{opt}).

Our work is centered on the control of the generalization error. However, with the rather abundant literature on the control of optimization errors for D-SGD (Koloskova et al., 2020; Neglia et al., 2020; Ying et al., 2021), one can combine the results of these papers with ours to obtain bounds on the excess risk, as explained below. Note that most bounds on the optimization error from the literature are given for the averaged parameter $A(S) = \bar{\theta}^{(T)}$ or $A(S) = \frac{1}{T}\sum_{t=1}^{T}\bar{\theta}^{(t)}$. Since our generalization bounds are also valid for these averaged parameters (see Section E.1, proofs for the time-average parameter are analogous), the following discussions are made with respect to them.

For **convex functions**, one can adapt the optimization error bound from Neglia et al. (2020, Proposition 3.1) to our notations and obtain $\epsilon_{\text{opt}} = \mathcal{O}\left(\frac{1}{\eta T} + \frac{\eta L^2}{\rho}\right)$, where some constant factors have been omitted for simplicity. Combining this with our generalization bound of order $\mathcal{O}(\frac{T\eta L^2}{mn})$, one can take $T = \Theta(\frac{\sqrt{mn}}{\eta L})$ with $\eta \leq \min\{\frac{\rho}{L\sqrt{mn}}; \frac{2\min_k W_{kk}}{\beta}\}$, and recover the classical rate of order $\mathcal{O}(\frac{L}{\sqrt{mn}})$ for ϵ_{excess}, a rate that can be found for instance in Hardt et al. (2016) or Lei & Ying (2020) for centralized SGD.

For **μ-strongly convex functions**, our result from Theorem 3.3 exhibits a generalization bound independent of the algorithm parameters η, W and T (as soon as η satisfies the constraint of our theorem). Moreover, we know (see e.g. Koloskova et al., 2020; Neglia et al., 2020) that the optimization error ϵ_{opt} can be set arbitrary small (in particular smaller than ϵ_{gen}), as soon as the graph W is connected, the number of iterations T is sufficiently large and the stepsize η is sufficiently small (in particular satisfying our constraint). Hence, as soon as W is connected, there exists an instance of parameters η, and T of D-SGD that gives an excess risk bound ϵ_{excess} with a "fast" rate of order $\epsilon_{\text{gen}} = \mathcal{O}(\frac{L^2}{\mu mn})$, a rate that can be found for instance in (Hardt et al., 2016) for centralized SGD.

We emphasize that the convergence of the above excess risk bounds actually depend on the communication graph. Indeed, the convergence is possible only if the graph is connected (i.e. $\rho < 1$, a necessary condition to control the optimization error), and the number of iterations T needed to make the optimization error small depends on ρ.

4. Generalization Error for Non-Convex Loss Functions

The case of non-convex (but bounded) loss functions was only investigated by Sun et al. (2021), for Variant B of D-SGD. Similar to their findings in convex scenarios, they established a generalization upper bound akin to that of Hardt et al. (2016). However, their bound again includes an additional term that does not diminish with increasing sample size. This raises the following question: can a finer-grained analysis than that of Sun et al. (2021) recover, *for both variants of D-SGD*, a result analogous to that of Hardt et al. (2016, Theorem 3.12) for centralized SGD with bounded non-convex loss functions?

To answer this question, we adopt the set of hypotheses of Hardt et al. (2016) and seek to extend their proof technique to the decentralized framework. Hereafter, we provide our generalization bound for bounded non-convex functions.

Theorem 4.1. *Assume that $\ell(\cdot; z) \in [0, 1]$ is an L-Lipschitz (Assumption 2.4) and β-smooth (Assumption 2.5) loss function for every z. Let $A_1(S), \dots, A_m(S)$ be the m final iterates of D-SGD (Variant A and B) run for T iterations, with mixing matrix W satisfying Assumption 2.7, such that $\min_k\{W_{kk}\} > 0$, and with monotonically non-increasing step sizes $\eta_t \leq \frac{c}{t+1}$, $c > 0$. Then, we have:*

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq (1 + \frac{1}{\beta c})(2cL^2)^{\frac{1}{\beta c+1}} \frac{T^{\frac{\beta c}{\beta c+1}}}{nm^{\frac{1}{\beta c+1}}} . \quad (5)$$

Sketch of proof (see Appendix C.1 for details). The crux of the proof is to condition the analysis on the time t_0 at which the swapped sample is first selected, then it is possible to minimize the generalization upper bound with respect to t_0 and obtain a tighter bound. Hardt et al. (2016) employed this same technique in conjunction with uniform stability arguments; our approach diverges by integrating it with an on-average model stability argument. This distinction is critical, as it is precisely what eliminates the impact of the graph in our analysis. Indeed, the discussion after Equation (3) shows that, before averaging over i and j, model deviations $\delta_k(i, j)$ depend on the powers of the communication graph W. The appropriate conditioning on t_0 combined with on-average model stability is given in the key Lemma C.1 and ensures that $\forall t_0 \in \mathbb{N}$ we have

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{t_0}{n} + L\Delta_k^{(T)} ,$$

where $\Delta_k^{(t)} = \frac{1}{mn}\sum_{i,j}\mathbb{E}[\delta_k^{(t)}(i,j)\big|\delta^{(t_0)}(i,j) = \mathbf{0}]$ and δ is the vector containings $\delta_1, \ldots, \delta_m$ (see the proof sketch of Theorem 3.1 for other notations). Then, using the $(1+\eta_t\beta)$-expansivity property of our updates, combined with the fact that $\Delta_k^{(t_0)} = 0$, we prove that:

$$\Delta_k^{(T)} \leq \frac{2L}{\beta mn}\Big(\frac{T}{t_0}\Big)^{c\beta} .$$

It then suffices to plug this equation into the first one and to minimize it with respect to t_0 to complete the proof. □

For the clarity of the discussion, below we omit constant factors in β, c and L, but we stress that our bound has the exact same constant factors as those in Hardt et al. (2016). Our generalization bound is of order $\mathcal{O}(T^{\frac{\beta c}{\beta c+1}}/nm^{\frac{1}{\beta c+1}})$ and several comments can be made. First, contrary to the convex cases, our bound does not exactly match the one of Hardt et al. (2016). Indeed, when centralized SGD is run over mn data points, they obtain a bound of order $\mathcal{O}(T^{\frac{\beta c}{\beta c+1}}/nm)$ which is strictly better than our bound. This comes from the fact that the proof technique relies on characterizing the number of steps that occur before the algorithm picks the data point that differs in S and $S^{(ij)}$. In centralized SGD, the probability to pick this point is $1/mn$ at each iteration, while it is only $1/n$ for D-SGD. Importantly, this means that the weaker bound is not directly due to decentralization, but rather to the fact that D-SGD selects m samples at each iteration (instead of only one for SGD). A fairer comparisonthus be to compare D-SGD to centralized SGD with batch size m.

Importantly, our generalization bound is valid for the two variants of D-SGD, is still independent of the choice of communication graph, and tends towards 0 as n and m increase. This significantly improves the results obtained for D-SGD in the prior work of Sun et al. (2021) for Variant B only, where the obtained bound has an extra additive term of order $\mathcal{O}(T^{\frac{\beta c}{\beta c+1}}C_\rho)$, where C_ρ depends on the spectral gap ρ of W and can be arbitrarily large. Note that, as in convex cases (Remark 3.2), their result is given for the average of final iterates, for which our result is also valid (see Proposition E.2 in Appendix E.1).

Remark 4.2. As pointed out in Table 1, there is a mistake in the original proof of the upper bound of Sun et al. (2021, Theorem 3). We explain this in Appendix E.3.

Finally, notice that for non-convex functions, the excess population risk ϵ_{excess} cannot be directly upper-bounded and a discussion analogue to the one of Section 3.3 cannot be made. Indeed, in the non-convex case, optimization errors usually control an upper bound on the gradient norm of the objective function and not on function values, which is required by our definition of ϵ_{opt}.

5. Towards Optimization-Dependent Generalization Bounds

Based on the results of the previous sections, one could conclude that decentralization and the choice of communication graph do not have an impact on the generalization of D-SGD. In this section, we suggest that this rather counter-intuitive result comes from the fact that the previous analyses are "worst-case", thereby hiding the true influence of the graph on certain types of loss functions and data distributions. In order to highlight such potential effects, we propose to investigate a certain type of generalization bounds referred to as *optimization-dependent*. This type of refined bounds, also referred as "data-dependent" in Kuzborskij & Lampert (2018), has been widely investigated in the generalization error analysis of *centralized* gradient methods (Kuzborskij & Lampert, 2018; Lei & Ying, 2020). Also based on algorithmic stability arguments, they reveal that a good optimization of the empirical risk can be beneficial for generalization (hence the term "optimization-dependent"). Since it is well known that the choice of graph affects the specific trajectories of the optimization algorithm and have an impact on the optimization error (Koloskova et al., 2020), we can expect this type of analysis to be appropriate to reveal the impact of the graph's connectivity on generalization.

Let us start with an additional assumption.

Assumption 5.1. *(Bounded empirical variance).* For all agents $k = 1, \ldots, m$, training dataset $S \in \mathcal{Z}^{mn}$ and model parameter $\theta \in \mathbb{R}^d$, there exists $\sigma^2 > 0$ such that $\frac{1}{n}\sum_{i=1}^n \|\nabla\ell(\theta; Z_{ik}) - \nabla R_{S_k}(\theta)\|_2^2 \leq \sigma^2$.

This assumption is rather standard in the control of the optimization error of stochastic gradient methods. Here, it is necessary to reveal different regimes in our new optimization-dependent bound, but notice that Assumption 5.1 is always satisfied under Assumption 2.4, with $\sigma^2 = L^2$.

The following lemma first links generalization errors to empirical risk minimization errors. Here, we focus on averaged generalization errors, instead of looking at the one of any fixed agent. This makes the analysis more tractable and reveals interesting links with optimization errors.

Lemma 5.2. (Link with local optimization errors). *Under the same hypotheses as in Theorem 3.1 and additional*

Assumption 5.1, we have:

$$\frac{1}{m}\sum_{k=1}^{m}|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{2L\sigma}{mn}\sum_{t=0}^{T-1}\eta_t + \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]$$

Lemma 5.2, and all the results of this section, are proved in Appendix D.

From Lemma 5.2, we notice that we can bound σ and the gradient norms $\|\nabla R_{S_j}(\theta_j^{(t)})\|_2$ by L and recover (up to a factor 2), the worst-case upper bound from Theorem 3.1. This illustrates well the "worst-case" notion mentioned above, but also the fact that the upper bound of Lemma 5.2 can be better than the one of Theorem 3.1 in some regimes. This will notably be the case in "low noise" regimes, when $\sigma \ll L$, and when the expected gradient norms $\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]$ reach small values. Interestingly, these gradient norms are linked with the optimization error of local empirical risks: $\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]$ will get smaller as the parameter $\theta_j^{(t)}$ of the j-th agent minimizes the associated local empirical risk $R_{S_j}(\theta)$. In other words, the more rapidly each agent optimizes its own local empirical risk, the smaller (and the better) the bound in Lemma 5.2.

The fact that the agents should minimize their *local* empirical risks may seem surprising at first. Indeed, as opposed to the minimization of the *full* empirical risk, it suggests that local SGD (D-SGD with identity graph) should be preferred to D-SGD with any other communication graph. This quite counter-intuitive observation comes from the averaging of the node-wise generalization errors, which increases stability. Using tools from convex optimization, we can now provide a more explicit upper bound, given in the following theorem.

Theorem 5.3. (optimization-dependent generalization bound) *Consider the same setting as in Theorem 3.1, with a constant stepsize η and additional Assumption 5.1. Assume further that W is symmetric. Then, there exists a graph-dependent constant $C_W < \infty$ such that:*

$$\frac{1}{m}\sum_{k=1}^{m}|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{2\sqrt{2}L\sqrt{T\eta}}{mn}\sqrt{\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[R_{S_j}(\theta^{(0)}) - R_{S_j}(\theta^*_{S_j})]} + \frac{2L\sigma\eta T}{mn} + \frac{2L\sqrt{\beta}\sigma\eta^{\frac{3}{2}}T}{mn} + \frac{2L^2T\eta C_W}{mn},$$

where $\theta^*_{S_j}$ is the (local) empirical risk minimizer of R_{S_j}.

Here, C_W corresponds to an upper bound on the series $C_W^{(t)} \triangleq \sum_{s=0}^{t-1}\|W^s - W^{s+1}\|_2$. Its existence is guaranteed (see Lemma D.4 in Appendix D), but unfortunately in most cases $C_W^{(t)}$ and C_W do not have a closed form expression. However, it can be shown for instance that $C_W = 0$ for $W = I$ (local SGD) or $C_W = 1$ for $W = \frac{1}{m}\mathbf{1}\mathbf{1}^T$ (complete graph with uniform weights). More generally, a condition for C_W to be small is to be close to the identity graph.

Upper bound analysis. The first term of the upper bound is of order $\mathcal{O}(\frac{\sqrt{T}}{mn})$ and depends on the averaged optimization error at initialization. This illustrates that if we are good at the initial point θ^0, few steps of D-SGD are going to be necessary before reaching a stable point. The other terms are of order $\mathcal{O}(\frac{T}{mn})$, the last one being graph-dependent with the constant C_W, while the other two depend on the variance σ^2. From the worst case point of view, we therefore recover the rate $\mathcal{O}(\frac{T}{mn})$ provided in Theorem 3.1. However, this new bound is more informative as it showcases other regimes. For instance, when σ and C_W are sufficiently small, the first term becomes dominant, and the bound becomes of order $\mathcal{O}(\frac{\sqrt{T}}{mn})$, which strictly improves the worst-case upper bound. Last but not least, if $W = I$ and $\eta \leq \frac{1}{\sqrt{T}}$, we have $C_W = 0$ and we obtain a bound of order

$$\mathcal{O}\left(\frac{\sqrt[4]{T}}{mn}\sqrt{\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[R_{S_j}(\theta^{(0)}) - R_{S_j}(\theta^*_{S_j})]} + \frac{\sigma\sqrt{T}}{mn}\right),$$

which is the same as the optimization-dependent bound obtained in Kuzborskij & Lampert (2018, Theorem 3) for centralized SGD with mn data points. This illustrates that our results generalize those obtained for centralized SGD in past studies.

In Figure 1, we represent the generalization errors observed empirically for different communication graphs (see Appendix E.4 for experimental details). In the low noise regime (left plot), we observe that the generalization error is strongly impacted by the choice of graph, the best one being the identity. This is in line with our analysis, as in the low noise regime the graph-dependent term of our bound becomes dominant. On the contrary, in the high noise regime (right plot), the choice of communication graph is less significant as we essentially recover the worst-case behavior, in which the choice of communication graph does not matter. Interestingly, we observe that in the firsts iterations of the low-noise regime, all curves have the same slope. This suggests that during this phase, generalization errors evolve linearly and do not depend on the graph, exhibiting the worst-case behavior described by Eq. (2). Then, as the algorithm continues, the optimization progresses (depending on the graph), making the algorithm more stable and the worst-case bound too conservative.

In the end, after refuting results which claimed that a poorly connected graph was detrimental to generalization, our bound and our empirical results show that such a graph

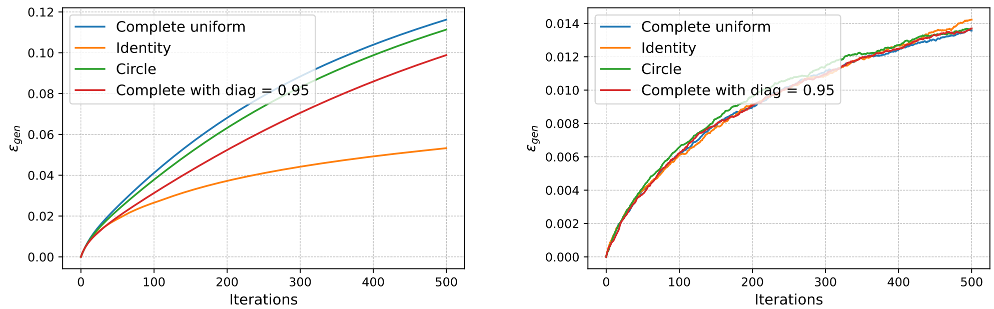

Figure 1. Empirical generalization error, as a function of the number of iterations T, and for different communication graphs. Constant stepsize $\eta = 0.03$. (Left) Low-noise regime with $\sigma \simeq 0$. (Right) Noisy regime with $\sigma > 0$. See Appendix E.4 for experimental details.

can, on the contrary, help generalization in certain regimes. This can be contrasted, however, with the fact that the optimization error ϵ_{opt} of the *full* empirical risk must also be controlled, which can only be done with a connected graph. Overall, our analysis paves the way for the future development of optimization-dependent generalization bounds, whose ability to characterize the practical impact of decentralization and choice of graph is well illustrated by our results.

6. Conclusion

In this paper, we showed that previous generalization error analyses of Variant B of D-SGD were very loose and led to incorrect conclusions regarding the impact of decentralization on generalization. On the contrary, we show that Variants A and B recover upper bounds analogous to those obtained in the centralized setting, suggesting that decentralization and the choice of graph do not have an impact on generalization. We then argue that this result is coming from a worst-case analysis and propose a refined bound revealing that the choice of graph can in fact improve the worst-case bound in certain regimes, and that a poorly-connected graph can even be beneficial for generalization.

All our generalization results, however, should not be completely dissociated from the optimization error. As seen in Section 3.3, if we want to recover the optimal excess risk bounds from the centralized setting, the optimization error must be sufficiently small. Contrary to the generalization error, this means that the graph should be connected and the number of iterations sufficiently large (depending on the connectivity of the communication graph). Future work could therefore include a better understanding of the generalization-optimization trade-off, notably with respect to the minimum number of iterations needed to reach the optimal bounds. In an other vein, future investigations could relax the Lipschitz condition, by considering an analysis similar to those proposed by Lei & Ying (2020); Schliserman & Koren (2022; 2023), or develop more refined optimization-dependent generalization bounds that would be able to capture, for instance, the impact of data heterogeneity between agents.

Acknowledgments

This work was supported by the French government managed by the Agence Nationale de la Recherche (ANR) through grant ANR-20-CE23-0015 (Project PRIDE), through the 3IA Côte d'Azur Investments in the Future project with the reference number ANR-19-P3IA-0002 and under France 2030 program with the reference ANR-23-PEIA-005 (REDEEM project). It was also funded in part by the European Network of Excellence dAIEDGE under Grant Agreement Nr. 101120726 and by the Groupe La Poste, sponsor of the Inria Foundation, in the framework of the FedMalin Inria Challenge. Batiste Le Bars was supported by an Inria-EPFL fellowship.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alquier, P. et al. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Amir, I., Koren, T., and Livni, R. SGD generalizes better than GD (and regularization doesn't help). In *Conference on Learning Theory*, pp. 63–92. PMLR, 2021.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33:4381–4391, 2020.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia*, pp. 169–207, 2004.

Bubeck, S. Convex optimization: Algorithms and complexity. 2015.

Catoni, O. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279. PMLR, 2019.

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods, 2023.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

Kong, L., Lin, T., Koloskova, A., Jaggi, M., and Stich, S. Consensus control for decentralized deep learning. In *International Conference on Machine Learning*, pp. 5686–5696. PMLR, 2021.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

Le Bars, B., Bellet, A., Tommasi, M., Lavoie, E., and Kermarrec, A.-M. Refined convergence and topology learning for decentralized sgd with heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1672–1702. PMLR, 2023.

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 2017.

Lin, J., Camoriano, R., and Rosasco, L. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pp. 2340–2348. PMLR, 2016.

London, B. A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

McAllester, D. A. Some PAC-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 230–234, 1998.

Meyer, C. D. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

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

Neglia, G., Xu, C., Towsley, D., and Calbi, G. Decentralized gradient methods: does topology matter? In *International Conference on Artificial Intelligence and Statistics*, pp. 2348–2358. PMLR, 2020.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*, pp. 3526–3545. PMLR, 2021.

Richards, D. and Rebeschini, P. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34): 1–44, 2020.

Scaman, K., Even, M., Le Bars, B., and Massoulie, L. Minimax excess risk of first-order methods for statistical learning with data-dependent oracles. In *International Conference on Artificial Intelligence and Statistics*, pp. 3709–3717. PMLR, 2024.

Schliserman, M. and Koren, T. Stability vs implicit bias of gradient methods on separable data and beyond. In *Conference on Learning Theory*, pp. 3380–3394. PMLR, 2022.

Schliserman, M. and Koren, T. Tight risk bounds for gradient descent on separable data. *Advances in Neural Information Processing Systems*, 2023.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Shawe-Taylor, J. and Williamson, R. C. A PAC analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pp. 2–9, 1997.

Sun, T., Li, D., and Wang, B. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9756–9764, 2021.

Taheri, H. and Thrampoulidis, C. On generalization of decentralized learning with separable data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4917–4945. PMLR, 2023.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Ying, B., Yuan, K., Chen, Y., Hu, H., Pan, P., and Yin, W. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhang, Y., Zhang, W., Bald, S., Pingali, V., Chen, C., and Goswami, M. Stability of SGD: Tightness analysis and improved bounds. In *Uncertainty in Artificial Intelligence*, pp. 2364–2373. PMLR, 2022.

Zhou, Y., Liang, Y., and Zhang, H. Generalization error bounds with probabilistic guarantee for SGD in nonconvex optimization. *arXiv preprint arXiv:1802.06903*, 2018.

Zhu, T., He, F., Zhang, L., Niu, Z., Song, M., and Tao, D. Topology-aware generalization of decentralized SGD. In *International Conference on Machine Learning*, pp. 27479–27503. PMLR, 2022.

Appendix

A. Technical lemmas

Below, we provide important definitions and lemmas that are going to be useful in our analysis. All proofs can be found in Hardt et al. (2016).

Definition A.1. The (stochastic) gradient update rule with $\eta > 0$, $z \in \mathcal{Z}$ and loss function ℓ is given by

$$G_{\eta,z}(\theta) = \theta - \eta \nabla \ell(\theta; z).$$

Definition A.2. An update rule $G(\theta)$ is said to be ν-expansive if:

$$\sup_{\theta,\theta'} \frac{\|G(\theta) - G(\theta')\|_2}{\|\theta - \theta'\|_2} \leq \nu.$$

Lemma A.3. (Expansivity of $G_{\eta,z}$). *If ℓ is β-smooth (Assumption 2.5), we have:*

1. *$G_{\eta,z}(\theta)$ is $(1+\eta\beta)$-expansive;*
2. *Assume in addition that $\ell(\cdot; z)$ is convex and $\eta < 2/\beta$. Then $G_{\eta,z}(\theta)$ is 1-expansive;*
3. *Assume in addition that $\ell(\cdot; z)$ is μ-strongly convex and $\eta < \frac{2}{\beta+\mu}$. Then $G_{\eta,z}(\theta)$ is $(1 - \frac{\eta\beta\mu}{\beta+\mu})$-expansive.*

Lemma A.4. (Growth recursion) *Fix an arbitrary sequence of gradient update rule $G_{\eta_1,z_1}, \ldots, G_{\eta_T,z_T}$ and another sequence $G_{\eta_1,z_1'}, \ldots, G_{\eta_T,z_T'}$ with same loss function ℓ (Def. A.1). Let $\theta_0 = \theta_0'$ be a starting point in $\mathbb{R}^d$ and define $\delta_t = \|\theta_t - \theta_t'\|$ where θ_t, θ_t' are defined recursively through*

$$\theta_{t+1} = G_{\eta_t,z_t}(\theta_t), \theta_{t+1}' = G_{\eta_t',z_t'}(\theta_t').$$

Then, we have the recurrence relation

$$\delta_0 = 0$$

$$\delta_{t+1} \leq \begin{cases} \nu\delta_t & \text{if } G_{\eta_t,z_t} = G_{\eta_t,z_t'} \text{ is } \nu\text{-expansive} \\ \min\{1,\nu\}\delta_t + 2\eta_t L & \text{if } \ell \text{ is } L\text{-Lipschitz and } G_{\eta_t,z_t} \text{ is } \nu\text{-expansive} \end{cases}$$

B. Proofs of Section 3

B.1. Theorem 3.1

First, notice that if $\min_k\{W_{kk}\} = 0$, by assumption we have $\eta_t \leq \frac{2\min_k\{W_{kk}\}}{\beta} = 0$. Hence our algorithm is perfectly stable and the bound is trivially obtained as $A(S) = \theta^{(0)}$ is data-independent. In the following, we focus on the case where $\min_k\{W_{kk}\} > 0$.

Thanks to Lemma 2.2, we simply need to prove that $A(S)$ is on average ε-stable with $\varepsilon \leq \frac{2L\sum_{t=0}^{T-1}\eta_t}{mn}$. Taking the notations of Def. 2.1, we denote by $A_k(S) = \theta_k^{(T)}$, and $A_k(S^{(ij)}) = \tilde{\theta}_k^{(T)}$, the outputs of agent k for D-SGD (Variant B) at round T, run over S and $S^{(ij)}$ respectively. More generally, $\{\theta_k^{(t)}\}_{t=0}^T$ (respectively $\{\tilde{\theta}_k^{(t)}\}_{t=0}^T$), refer to the iterates of agent k run over S (respectively $S^{(ij)}$). We also denote by $\{Z'_{vk}\}_{vk}$ the elements of the data set $S^{(ij)}$, i.e. $Z'_{vk} = Z_{vk}$ for $(v,k) \neq (i,j)$ and $Z'_{ij} = \tilde{Z}_{ij} \neq Z_{ij}$.

For all $k = 1, \ldots, m$ and $t \geq 1$, we have

$$\begin{aligned}
\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2 &= \Big\| \sum_{l=1}^m W_{kl}\theta_l^{(t)} - \eta_t \nabla\ell(\theta_k^{(t)}; Z_{I_k^t k}) - \sum_{l=1}^m W_{kl}\tilde{\theta}_l^{(t)} + \eta_t\nabla\ell(\tilde{\theta}_k^{(t)}; Z'_{I_k^t k}) \Big\|_2 \\
&= \Big\| W_{kk}\Big(\theta_k^{(t)} - \frac{\eta_t}{W_{kk}}\nabla\ell(\theta_k^{(t)}; Z_{I_k^t k}) - \tilde{\theta}_k^{(t)} + \frac{\eta_t}{W_{kk}}\nabla\ell(\tilde{\theta}_k^{(t)}; Z'_{I_k^t k})\Big) + \sum_{l\neq k}^m W_{kl}(\theta_l^{(t)} - \tilde{\theta}_l^{(t)}) \Big\|_2 \\
&\leq W_{kk}\Big\|\theta_k^{(t)} - \frac{\eta_t}{W_{kk}}\nabla\ell(\theta_k^{(t)}; Z_{I_k^t k}) - \tilde{\theta}_k^{(t)} + \frac{\eta_t}{W_{kk}}\nabla\ell(\tilde{\theta}_k^{(t)}; Z'_{I_k^t k})\Big\|_2 + \sum_{l\neq k}^m W_{kl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2
\end{aligned} \tag{6}$$

Thanks to Lemma A.3 (part. 2) in Appendix A and the fact that, by assumption, $\forall k, \frac{\eta_t}{W_{kk}} \leq \frac{2}{\beta}$, the update rules $G_{\eta_t, Z_{I_k^t k}}(\theta) = \theta - \frac{\eta_t}{W_{kk}} \nabla \ell(\theta; Z_{I_k^t k})$ and $G_{\eta_t, Z'_{I_k^t k}}(\theta) = \theta - \frac{\eta_t}{W_{kk}} \nabla \ell(\theta; Z'_{I_k^t k})$ are 1-expansive. Hence:

If $k \neq j$, we have $Z_{I_k^t k} = Z'_{I_k^t k}$, which gives from Eq. (6) and Lemma A.4 that:

$$\begin{aligned} \|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2 &\leq W_{kk} \Big\| \theta_k^{(t)} - \frac{\eta_t}{W_{kk}} \nabla \ell(\theta_k^{(t)}; Z_{I_k^t k}) - \tilde{\theta}_k^{(t)} + \frac{\eta_t}{W_{kk}} \nabla \ell(\tilde{\theta}_k^{(t)}; Z_{I_k^t k}) \Big\|_2 + \sum_{l \neq k}^{m} W_{kl} \Big\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} \Big\|_2 \\ &\leq W_{kk} \Big\| \theta_k^{(t)} - \tilde{\theta}_k^{(t)} \Big\|_2 + \sum_{l \neq k}^{m} W_{kl} \Big\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} \Big\|_2 \\ &\leq \sum_{l=1}^{m} W_{kl} \Big\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} \Big\|_2 \end{aligned} \tag{7}$$

If $k = j$:

With probability $1 - \frac{1}{n}$, $I_j^t \neq i$ so $Z_{I_j^t j} = Z'_{I_j^t j}$ and we therefore have again the relation of Equation (7).

With probability $\frac{1}{n}$ however, $I_j^t = i$ and in that case $Z_{I_j^t j} = Z_{ij} \neq \tilde{Z}_{ij} = Z'_{I_j^t j}$. With probability $\frac{1}{n}$, we therefore have:

$$\begin{aligned} \|\theta_j^{(t+1)} - \tilde{\theta}_j^{(t+1)}\|_2 &\leq W_{jj} \Big\| \theta_j^{(t)} - \frac{\eta_t}{W_{jj}} \nabla \ell(\theta_j^{(t)}; Z_{ij}) - \tilde{\theta}_j^{(t)} + \frac{\eta_t}{W_{jj}} \nabla \ell(\tilde{\theta}_j^{(t)}; \tilde{Z}_{ij}) \Big\|_2 + \sum_{l \neq j}^{m} W_{jl} \Big\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} \Big\|_2 \qquad (8) \\ &\overset{Lem. A.4}{\leq} W_{jj} \Big(\|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\|_2 + \frac{2\eta_t L}{W_{jj}} \Big) + \sum_{l \neq j}^{m} W_{jl} \Big\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} \Big\|_2 \\ &= \sum_{l=1}^{m} W_{jl} \Big\| \theta_l^{(t)} - \tilde{\theta}_l^{(t)} \Big\|_2 + 2\eta_t L \end{aligned}$$

Denoting by $\delta_k^{(t)}(i,j) \triangleq \|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\|_2$ and combining previous results, we get for all $k = 1, \ldots, m$, the recursion:

$$\mathbb{E}[\delta_k^{(T)}(i,j)] \leq \sum_{l=1}^{m} W_{kl} \mathbb{E}[\delta_l^{(T-1)}(i,j)] + \frac{2\eta_t L}{n} \mathbb{1}_{\{k=j\}} \, .$$

In vector format, this writes $\mathbb{E}[\delta^{(T)}(i,j)] \leq W \mathbb{E}[\delta^{(T-1)}(i,j)] + \frac{2\eta_t L}{n} e_j$ (the inequality is meant coordinate-wise), where $\delta^{(t)}(i,j)$ contains the values of $\delta_k^{(t)}(i,j), \forall k$ and e_j is the j-th element of the canonical basis. Unrolling this recursion until $t = 0$, and noticing that $\delta_k^{(0)}(i,j) = 0$, we get:

$$\mathbb{E}[\delta^{(T)}(i,j)] \leq \frac{2L}{n} \sum_{t=0}^{T-1} W^{T-t-1} e_j \eta_t \Longrightarrow \mathbb{E}[\delta_k^{(T)}(i,j)] \leq \frac{2L}{n} \sum_{t=0}^{T-1} (W^{T-t-1})_{kj} \eta_t \tag{9}$$

Averaging over i and j and using the fact that the power of W is also doubly stochastic, i.e. $\sum_j (W^{T-t-1})_{kj} = 1$, we obtain that the on-average model stability is upper bounded by $\frac{2L \sum_{t=0}^{T-1} \eta_t}{mn}$, which concludes the proof with a direct application of Lemma 2.2.

□

B.2. Theorem 3.3

Like for convex functions, if $\min_k\{W_{kk}\} = 0$ we have $\eta_t = 0$, and the bound is trivially obtained. In the following, we therefore focus on the case where $\min_k\{W_{kk}\} > 0$.

The proof is analogous to the one obtained for the general convex case (Theorem 3.1). We keep the same notations, where $\mathbb{E}\delta_k^{(T)}(i,j) = \mathbb{E}\|\theta_k^{(T)} - \tilde{\theta}_k^{(T)}\|_2$ is the quantity we wish to control, on average over i and j. Using the fact that the Euclidean projection Π_Θ is 1-expansive (see e.g. Lemma 4.6 in Hardt et al. (2016)), we can directly obtain Equation (6) using the same arguments.

Thanks to Lemma A.3 (part. 3), we notice that for all k, the update rules $G_{\eta, Z_{I_k^t k}}(\theta) = \theta - \frac{\eta}{W_{kk}}\nabla\ell(\theta; Z_{I_k^t k})$ and $G_{\eta, Z'_{I_k^t k}}(\theta) = \theta - \frac{\eta}{W_{kk}}\nabla\ell(\theta; Z'_{I_k^t k})$ are $(1 - \frac{\eta\mu}{2W_{kk}})$-expansive. Indeed, since we always have $\mu \leq \beta$ and by assumption $\frac{\eta}{W_{kk}} \leq \frac{1}{\beta} \leq \frac{2}{\beta+\mu}$, we can apply the lemma and then use the fact that $\frac{\eta\beta\mu}{W_{kk}(\beta+\mu)} \geq \frac{\eta\beta\mu}{2W_{kk}\beta} = \frac{\eta\mu}{2W_{kk}}$. We now follow the proof of the convex case by splitting the analysis similarly.

If $k \neq j$, we have $Z_{I_k^t k} = Z'_{I_k^t k}$, which gives from Eq. (6) and Lemma A.4 with the $(1 - \frac{\eta\mu}{2W_{kk}})$-expansivity, that:

$$
\begin{aligned}
\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2 &\leq W_{kk}\Big(1 - \frac{\eta\mu}{2W_{kk}}\Big)\Big\|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\Big\|_2 + \sum_{l\neq j}^{m} W_{kl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 \\
&= \sum_{l=1}^{m} W_{kl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 - \frac{\eta\mu}{2}\Big\|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\Big\|_2 \qquad (10)
\end{aligned}
$$

If $k = j$:

With probability $1 - \frac{1}{n}$, $I_j^t \neq i$ so $Z_{I_j^t j} = Z'_{I_j^t j}$ and we therefore have again the relation of Eq. (10).

With probability $\frac{1}{n}$ however, $I_j^t = i$ and in that case $Z_{I_j^t j} = Z_{ij} \neq \tilde{Z}_{ij} = Z'_{I_j^t j}$. With probability $\frac{1}{n}$, we therefore have:

$$
\begin{aligned}
\|\theta_j^{(t+1)} - \tilde{\theta}_j^{(t+1)}\|_2 &\leq W_{jj}\Big\|\theta_j^{(t)} - \frac{\eta}{W_{jj}}\nabla\ell(\theta_j^{(t)}; Z_{ij}) - \tilde{\theta}_j^{(t)} + \frac{\eta}{W_{jj}}\nabla\ell(\tilde{\theta}_j^{(t)}; \tilde{Z}_{ij})\Big\|_2 + \sum_{l\neq j}^{m} W_{jl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 \\
&\overset{Lem. A.4}{\leq} W_{jj}\Big(\big(1 - \frac{\eta\mu}{2W_{jj}}\big)\|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\|_2 + \frac{2\eta L}{W_{jj}}\Big) + \sum_{l\neq j}^{m} W_{jl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 \\
&= \sum_{l=1}^{m} W_{jl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 - \frac{\eta\mu}{2}\Big\|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\Big\|_2 + 2\eta L
\end{aligned}
$$

Combining previous results, we get for all $k = 1, \dots, m$, the recursion:

$$
\mathbb{E}[\delta_k^{(T)}(i,j)] \leq \sum_{l=1}^{m} W_{kl}\mathbb{E}[\delta_l^{(T-1)}(i,j)] - \frac{\eta\mu}{2}\mathbb{E}[\delta_k^{(T-1)}(i,j)] + \frac{2\eta L}{n}\mathbb{1}_{\{k=j\}} \;.
$$

In vector format, this writes (the inequality is meant coordinate-wise)

$$
\mathbb{E}[\delta^{(T)}(i,j)] \leq \Big(W - \frac{\eta\mu}{2}I\Big)\mathbb{E}[\delta^{(T-1)}(i,j)] + \frac{2\eta_t L}{n}e_j \;,
$$

where $\delta^{(T)}(i,j)$ contains the values of $\delta_k^{(T)}(i,j)$, $\forall k$ and e_j is the j-th element of the canonical basis. Unrolling this recursion until $t = 0$, and noticing that $\delta_k^{(0)}(i,j) = 0$, we get:

$$
\mathbb{E}[\delta^{(T)}(i,j)] \leq \frac{2\eta L}{n}\sum_{t=0}^{T-1}\Big(W - \frac{\eta\mu}{2}I\Big)^t e_j \;.
$$

Averaging the previous equation over i and j and using the fact that $\sum_{j=1}^{m} e_j = \mathbf{1}$, we have

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\delta^{(T)}(i,j)] \leq \frac{2\eta L}{mn}\sum_{t=0}^{T-1}\Big(W-\frac{\eta\mu}{2}I\Big)^t\mathbf{1}\ .$$

Since $(W-\frac{\eta\mu}{2}I)\mathbf{1}=(1-\frac{\eta\mu}{2})\mathbf{1}$, by induction we have $(W-\frac{\eta\mu}{2}I)^t\mathbf{1}=(1-\frac{\eta\mu}{2})^t\mathbf{1}$ and we can finally get $\forall k$:

$$\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\delta_k^{(T)}(i,j)] \leq \frac{2\eta L}{mn}\sum_{t=0}^{T-1}\Big(1-\frac{\eta\mu}{2}\Big)^t \leq \frac{4L}{\mu mn}\ , \tag{11}$$

which makes A_k on average ε-stable (Def. 2.1) with $\varepsilon=\frac{4L}{\mu mn}$. Like in the convex case, a direct application of Lemma 2.2 completes the proof.

□

C. Proofs of Section 4

C.1. Theorem 4.1

Our analysis for the non-convex case relies on on-average model stability and leverages the fact that D-SGD can make several steps before using the one example that has been swapped. This idea is summarized in the following lemma.

Lemma C.1. *Assume that the loss function $\ell(\cdot,z)$ is nonnegative and L-Lipschitz for all z. For all $i=1,\ldots,n$ and $j=1,\ldots,m$, let $\{\theta_k^{(t)}\}_{t=0}^T$ and $\{\tilde{\theta}_k^{(t)}(i,j)\}_{t=0}^T$, the iterates of agent $k=1,\ldots,m$ for D-SGD (Variant A and B) run on S and $S^{(ij)}$ respectively. Then, for every $t_0\in\{0,1,\ldots,T\}$ we have:*

$$|\mathbb{E}_{A,S}[R(A_k(S))-R_S(A_k(S))]| \leq \frac{t_0}{n}\sup_{\theta,z}\ell(\theta;z)+\frac{L}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\delta_k^{(T)}(i,j)\big|\delta^{(t_0)}(i,j)=\mathbf{0}] \tag{12}$$

where $\delta^{(t)}(i,j)$ is the vector containing $\forall k=1,\ldots,m$, $\delta_k^{(t)}(i,j)=\|\theta_k^{(t)}-\tilde{\theta}_k^{(t)}(i,j)\|_2$.

Proof. Consider the notation of Def. 2.1 and notice that

$$R(A_k(S))=\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{Z\sim\mathcal{D}_j}[\ell(A_k(S);Z)]=\frac{1}{mn}\sum_{j=1}^{m}\sum_{j=1}^{n}\mathbb{E}_{\tilde{S}}[\ell(A_k(S);\tilde{Z}_{ij})].$$

Then, for all $k=1,\ldots,m$, by linearity of expectation we have

$$\begin{aligned}\mathbb{E}_{A,S}[R(A_k(S))-R_S(A_k(S))] &= \mathbb{E}_{A,S,\tilde{S}}\left[\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\Big(\ell(A_k(S);\tilde{Z}_{ij})-\ell(A_k(S);Z_{ij})\Big)\right]\\ &= \mathbb{E}_{A,S,\tilde{S}}\left[\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\Big(\ell(A_k(S^{(ij)});Z_{ij})-\ell(A_k(S);Z_{ij})\Big)\right].\end{aligned}$$

Hence,

$$\begin{aligned}|\mathbb{E}_{A,S}[R(A_k(S))-R_S(A_k(S))]| &\leq \mathbb{E}_{A,S,\tilde{S}}\left[\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\Big|\ell(A_k(S^{(ij)});Z_{ij})-\ell(A_k(S);Z_{ij})\Big|\right]\\ &= \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{A,S,\tilde{S}}\left[\Big|\ell(A_k(S^{(ij)});Z_{ij})-\ell(A_k(S);Z_{ij})\Big|\right]\end{aligned}$$

Let the event $\mathcal{E}(i,j)=\{\delta^{(t_0)}(i,j)=\mathbf{0}\}$, we have $\forall i,j$:

$$
\begin{aligned}
&\mathbb{E}_{A,S,\tilde{S}}\Big[\big|\ell(A_k(S^{(ij)});Z_{ij}) - \ell(A_k(S);Z_{ij})\big|\Big] \\
&\quad= \mathbb{P}(\mathcal{E}(i,j))\mathbb{E}[|\ell(A_k(S^{(ij)});Z_{ij}) - \ell(A_k(S);Z_{ij})|\big|\mathcal{E}(i,j)] \\
&\qquad\qquad\qquad\qquad + \mathbb{P}(\mathcal{E}(i,j)^c)\mathbb{E}[|\ell(A_k(S^{(ij)});Z_{ij}) - \ell(A_k(S);Z_{ij})|\big|\mathcal{E}(i,j)^c] \\
&\quad\leq \mathbb{E}[|\ell(A_k(S^{(ij)});Z_{ij}) - \ell(A_k(S);Z_{ij})|\big|\mathcal{E}(i,j)] + \mathbb{P}(\mathcal{E}(i,j)^c)\cdot\sup_{\theta,z}\ell(\theta;z) \\
&\quad\leq L\mathbb{E}[\|A_k(S) - A_k(S^{(ij)})\|\big|\mathcal{E}(i,j)] + \mathbb{P}(\mathcal{E}(i,j)^c)\cdot\sup_{\theta,z}\ell(\theta;z) \\
&\quad= L\mathbb{E}[\delta_k^{(T)}(i,j)\big|\mathcal{E}(i,j)] + \mathbb{P}(\mathcal{E}(i,j)^c)\cdot\sup_{\theta,z}\ell(\theta;z)
\end{aligned}
$$

It remains to bound $\mathbb{P}(\mathcal{E}(i,j)^c)$. Let T_0 be the random variable of the first time step D-SGD uses the swapped example. Since we necessarily have $\{T_0 > t_0\} \subset \mathcal{E}(i,j)$, we have $\mathcal{E}(i,j)^c \subset \{T_0 \leq t_0\}$ and therefore $\mathbb{P}(\mathcal{E}(i,j)^c) \leq \mathbb{P}(T_0 \leq t_0) = \sum_{t=1}^{t_0}\mathbb{P}(T_0 = t) \leq \sum_{t=1}^{t_0}\frac{1}{n} = \frac{t_0}{n}$. Averaging over i and j completes the proof.

□

We can now move on to the proof of the main theorem. We first apply Lemma C.1 and the fact that, by assumption, $\ell \in [0,1]$, so that for any $t_0 \in \{0,1,\dots,T\}$ and any $k = 1,\dots,m$, we have:

$$
|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{t_0}{n} + \frac{L}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\delta_k^{(T)}(i,j)\big|\delta^{(t_0)}(i,j) = \mathbf{0}] \tag{13}
$$

It remains to control the right-hand term of Equation (13). We start with the proof for the variant B of D-SGD. The proof for Variant A will follow.

Variant B:

For a fixed couple (i,j), we are first going to control the vector $\Delta^{(t)}(i,j) \triangleq \mathbb{E}[\delta^{(t)}(i,j)|\delta^{(t_0)}(i,j) = \mathbf{0}]$. When it is clear from context, we simply write $\tilde{\theta}_k^{(t)}(i,j) = \tilde{\theta}_k^{(t)}$. The proof is analogous to the one obtained for convex cases (Theorem 3.1 and 3.3) and we can start directly from Equation (6) using the same arguments.

Thanks to Lemma A.3 (part. 3), we notice that for all k, the update rules $G_{\eta_t, Z_{I_k^t k}}(\theta) = \theta - \frac{\eta_t}{W_{kk}}\nabla\ell(\theta; Z_{I_k^t k})$ and $G_{\eta_t, Z'_{I_k^t k}}(\theta) = \theta - \frac{\eta_t}{W_{kk}}\nabla\ell(\theta; Z'_{I_k^t k})$ are $(1 + \frac{\eta_t\beta}{W_{kk}})$-expansive. Following the proof of the convex cases, we can split the analysis similarly.

If $k \neq j$, we have $Z_{I_k^t k} = Z'_{I_k^t k}$, which gives from Eq. (6) and Lemma A.4 with the $(1 + \frac{\eta_t\beta}{W_{kk}})$-expansivity, that:

$$
\begin{aligned}
\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2 &\leq W_{kk}\Big(1 + \frac{\eta_t\beta}{W_{kk}}\Big)\Big\|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\Big\|_2 + \sum_{l\neq j}^{m} W_{kl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 \\
&= \sum_{l=1}^{m} W_{kl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 + \eta_t\beta\Big\|\theta_k^{(t)} - \tilde{\theta}_k^{(t)}\Big\|_2
\end{aligned} \tag{14}
$$

If $k = j$:

With probability $1 - \frac{1}{n}$, $I_j^t \neq i$ so $Z_{I_j^t j} = Z'_{I_j^t j}$ and we therefore have again the relation of Eq. (14).

With probability $\frac{1}{n}$ however, $I_j^t = i$ and in that case $Z_{I_j^t j} = Z_{ij} \neq \tilde{Z}_{ij} = Z'_{I_j^t j}$. With probability $\frac{1}{n}$, we therefore have:

$$
\begin{aligned}
\|\theta_j^{(t+1)} - \tilde{\theta}_j^{(t+1)}\|_2 &\le W_{jj}\Big\|\theta_j^{(t)} - \frac{\eta_t}{W_{jj}}\nabla\ell(\theta_j^{(t)}; Z_{ij}) - \tilde{\theta}_j^{(t)} + \frac{\eta_t}{W_{jj}}\nabla\ell(\tilde{\theta}_j^{(t)}; \tilde{Z}_{ij})\Big\|_2 + \sum_{l\neq j}^{m} W_{jl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 \\
&\overset{Lem. A.4}{\le} W_{jj}\Big((1 + \frac{\eta_t\beta}{W_{jj}})\|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\|_2 + \frac{2\eta_t L}{W_{jj}}\Big) + \sum_{l\neq j}^{m} W_{jl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 \\
&= \sum_{l=1}^{m} W_{jl}\Big\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\Big\|_2 + \eta_t\beta\Big\|\theta_j^{(t)} - \tilde{\theta}_j^{(t)}\Big\|_2 + 2\eta_t L
\end{aligned}
$$

From the previous equations, we get that $\Delta^{(t+1)}(i,j) \le (W + \eta_t\beta I)\Delta^{(t)}(i,j) + \frac{2\eta_t L}{n} e_j$ (the inequality, and the following ones are meant coordinate-wise). Let $\Delta^{(t)} = \frac{1}{mn}\sum_{i,j}\Delta^{(t)}(i,j)$, then using the fact that $\eta_t \le \frac{c}{t+1}$, $c > 0$, we have $\forall t \ge t_0$:

$$
\begin{aligned}
\Delta^{(t+1)} &\le (W + \eta_t\beta I)\Delta^{(t)} + \frac{2\eta_t L}{mn}\mathbf{1} \\
&\le (W + \frac{c\beta}{t+1} I)\Delta^{(t)} + \frac{2cL}{mn(t+1)}\mathbf{1}
\end{aligned}
$$

Since $\Delta^{(t_0)} = \mathbf{0}$, we can unroll the previous recursion from T to $t_0 + 1$ and get:

$$
\begin{aligned}
\Delta^{(T)} &\le \frac{2cL}{Tmn}\mathbf{1} + \sum_{t=t_0+1}^{T-1}\Big\{\prod_{\tau=t+1}^{T}\Big(W + \frac{c\beta}{\tau}I\Big)\Big\}\frac{2cL}{tmn}\mathbf{1} \\
&= \frac{2cL}{Tmn}\mathbf{1} + \sum_{t=t_0+1}^{T-1}\Big\{\prod_{\tau=t+1}^{T}\Big(1 + \frac{c\beta}{\tau}\Big)\Big\}\frac{2cL}{tmn}\mathbf{1}\,,
\end{aligned}
$$

where in the last equality we used the fact that $(W + \frac{c\beta}{\tau}I)\mathbf{1} = (1 + \frac{c\beta}{\tau})\mathbf{1}$, which by induction gives $\prod_\tau \Big(W + \frac{c\beta}{\tau}I\Big)\mathbf{1} = \prod_\tau \Big(1 + \frac{c\beta}{\tau}\Big)\mathbf{1}$. Then, we focus on the coordinate of interest k and using the fact that $1 + x \le \exp(x)$, we have:

$$
\begin{aligned}
\Delta_k^{(T)} &\le \frac{2cL}{Tmn} + \sum_{t=t_0+1}^{T-1}\Big\{\prod_{\tau=t+1}^{T}\exp\Big(\frac{c\beta}{\tau}\Big)\Big\}\frac{2cL}{tmn} \\
&= \frac{2cL}{Tmn} + \sum_{t=t_0+1}^{T-1}\exp\Big(c\beta\sum_{\tau=t+1}^{T}\frac{1}{\tau}\Big)\frac{2cL}{tmn} \\
&\le \frac{2cL}{Tmn} + \sum_{t=t_0+1}^{T-1}\exp\Big(c\beta\log\big(\frac{T}{t}\big)\Big)\frac{2cL}{tmn} \\
&= \frac{2cL}{Tmn} + \sum_{t=t_0+1}^{T-1}\Big(\frac{T}{t}\Big)^{c\beta}\frac{2cL}{tmn} \\
&= \frac{2cLT^{c\beta}}{T^{c\beta+1}mn} + \frac{2cLT^{c\beta}}{mn}\sum_{t=t_0+1}^{T-1} t^{-c\beta-1} \\
&= \frac{2cLT^{c\beta}}{mn}\sum_{t=t_0+1}^{T} t^{-c\beta-1} \\
&\le \frac{2cLT^{c\beta}}{mn}\frac{t_0^{-\beta c}}{c\beta} = \frac{2L}{\beta mn}\Big(\frac{T}{t_0}\Big)^{c\beta}, \qquad (15)
\end{aligned}
$$

where the last inequality is obtained using bounds over (partial) harmonic series.

Plugging this result into (13), we obtain

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{t_0}{n} + \frac{2L^2}{\beta mn}\Big(\frac{T}{t_0}\Big)^{c\beta}.$$

Then, taking $t_0 = \left(\frac{2L^2c}{m}\right)^{\frac{1}{c\beta+1}} T^{\frac{c\beta}{c\beta+1}}$ (approximate minimizer of the right-hand term above), we have

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq (1 + \frac{1}{\beta c})(2cL^2)^{\frac{1}{\beta c+1}} \frac{T^{\frac{\beta c}{\beta c+1}}}{m^{\frac{1}{\beta c+1}} n} ,$$

which concludes the proof for Variant B.

Remark C.2. Note that the inequality (15) is not optimal in the specific case $t_0 = 0$ (it diverges) and somehow prevents us from taking $t_0 = 0$ in the proof. However, taking this value could be optimal in some regimes. Hence, the analysis and our final bound could be slightly improved by looking at the minimum between the cases where $t_0 = 0$ or not. When taking $t_0 = 0$, the bound in Eq. (15) can be improved to $\frac{2cLT^{c\beta}}{mn}(1 + \frac{1}{c\beta})$.

Variant A:

The proof for the variant A is essentially the same, where instead of Equation (6), we have:

$$\begin{aligned}
\|\theta_k^{(t+1)} - \tilde{\theta}_k^{(t+1)}\|_2 &= \Big\| \sum_{l=1}^{m} W_{kl}\Big(\theta_l^{(t)} - \eta_t \nabla\ell(\theta_l^{(t)}; Z_{I_l^t l})\Big) - \sum_{l=1}^{m} W_{kl}\Big(\tilde{\theta}_l^{(t)} + \eta_t \nabla\ell(\tilde{\theta}_l^{(t)}; Z'_{I_l^t l})\Big)\Big\|_2 \\
&= \sum_{l=1}^{m} W_{kl}\Big\|\theta_l^{(t)} - \eta_t \nabla\ell(\theta_l^{(t)}; Z_{I_l^t l}) + \tilde{\theta}_l^{(t)} + \eta_t \nabla\ell(\tilde{\theta}_l^{(t)}; Z'_{I_l^t l})\Big\|_2 \qquad (16)
\end{aligned}$$

Since, thanks to Lemma A.3 (part. 1), the update rules $G_{\eta_t, Z_{I_l^t l}}(\theta) = \theta - \eta_t \nabla\ell(\theta; Z_{I_l^t l})$ and $G_{\eta_t, Z'_{I_l^t l}}(\theta) = \theta - \eta_t \nabla\ell(\theta; Z'_{I_l^t l})$ are $(1 + \eta_t\beta)$-expansive, we can use the same arguments as before to show that

$$\mathbb{E}[\|\theta_l^{(t)} - \eta_t \nabla\ell(\theta_l^{(t)}; Z_{I_l^t l}) + \tilde{\theta}_l^{(t)} + \eta_t \nabla\ell(\tilde{\theta}_l^{(t)}; Z'_{I_l^t l})\| |\mathcal{F}_t] \leq (1 + \eta_t\beta)\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\| + \frac{2L\eta_t}{n}\mathbb{1}_{\{l=j\}} ,$$

where $\mathcal{F}_t$ is the natural filtration at time t.

Combining previous equations and the notation of the proof of Variant B, we get the vector format relation:

$$\begin{aligned}
\Delta^{(t+1)} &\leq (1 + \eta_t\beta) W\Delta^{(t)} + \frac{2\eta_t L}{mn}\mathbf{1} \\
&\leq (1 + \frac{c\beta}{t+1}) W\Delta^{(t)} + \frac{2cL}{mn(t+1)}\mathbf{1}
\end{aligned}$$

Since $\Delta^{(t_0)} = \mathbf{0}$, we can unroll the previous recursion from T to $t_0 + 1$ and get:

$$\begin{aligned}
\Delta^{(T)} &\leq \frac{2cL}{Tmn}\mathbf{1} + \sum_{t=t_0+1}^{T-1} \Big\{ \prod_{\tau=t+1}^{T} \Big(1 + \frac{c\beta}{\tau}\Big) W \Big\} \frac{2cL}{tmn}\mathbf{1} \\
&= \frac{2cL}{Tmn}\mathbf{1} + \sum_{t=t_0+1}^{T-1} \Big\{ \prod_{\tau=t+1}^{T} \Big(1 + \frac{c\beta}{\tau}\Big) \Big\} \frac{2cL}{tmn}\mathbf{1} ,
\end{aligned}$$

where in the last equality we used the fact that $(1 + \frac{c\beta}{\tau})W\mathbf{1} = (1 + \frac{c\beta}{\tau})\mathbf{1}$. From this point, the proof is the same as the one of Variant B, starting from the beginning of the derivation of Equation (15).

□

D. Proofs of Section 5

Like in the proof of Theorem 3.1, we simply need to consider the case where $\min_k\{W_{kk}\} > 0$, the case $\min_k\{W_{kk}\} = 0$ being trivial since it implies that $\eta_t = 0$.

D.1. Lemma 5.2

We start the proof with the following lemma.

Lemma D.1. (Link with point-wise gradient norms). *Under the same hypothesis as Theorem 3.1:*

$$|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| \leq \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\sum_{j=1}^{m}(W^{T-t-1})_{kj}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\|_2]$$

Proof. Until Equation (8), the proof of Lemma D.1 is exactly the same as the one of Theorem 3.1. Let's start the proof from this point:

If $k = j$, then with probability $\frac{1}{n}$ we have:

$$\begin{aligned}\|\theta_j^{(t+1)} - \tilde{\theta}_j^{(t+1)}\|_2 &\leq W_{jj}\left\|\theta_j^{(t)} - \frac{\eta_t}{W_{jj}}\nabla\ell(\theta_j^{(t)}; Z_{ij}) - \tilde{\theta}_j^{(t)} + \frac{\eta_t}{W_{jj}}\nabla\ell(\tilde{\theta}_j^{(t)}; \tilde{Z}_{ij})\right\|_2 + \sum_{l\neq j}^{m}W_{jl}\left\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\right\|_2 \\ &\leq \sum_{l=1}^{m}W_{jl}\left\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\right\|_2 + \eta_t\left\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\right\|_2 + \eta_t\left\|\nabla\ell(\tilde{\theta}_j^{(t)}; \tilde{Z}_{ij})\right\|_2\end{aligned}$$

Since the gradient norms $\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\|_2$ and $\|\nabla\ell(\tilde{\theta}_j^{(t)}; \tilde{Z}_{ij})\|_2$ have same law, they have the same expectation and we have:

$$\mathbb{E}[\|\theta_j^{(t+1)} - \tilde{\theta}_j^{(t+1)}\|_2] \leq \sum_{l=1}^{m}W_{jl}\mathbb{E}[\|\theta_l^{(t)} - \tilde{\theta}_l^{(t)}\|_2] + \frac{2\eta_t}{n}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\|_2]$$

Combining with the result for $k \neq j$ in the proof of Theorem 3.1, we have the following relation in vector format:

$$\begin{aligned}\mathbb{E}[\delta^{(T)}(i,j)] &\leq W\mathbb{E}[\delta^{(T-1)}(i,j)] + \frac{2\eta_t}{n}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\|_2]\cdot e_j \\ &\leq \frac{2}{n}\sum_{t=0}^{T-1}W^{T-t-1}\eta_t\mathbb{E}[\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\|_2]\cdot e_j\end{aligned}$$

where the second inequality is obtained by unrolling the recursion until $t = 0$. For any agent $k = 1, \ldots, n$, we therefore have

$$\mathbb{E}[\delta_k^{(T)}(i,j)] \leq \frac{2}{n}\sum_{t=0}^{T-1}(W^{T-t-1})_{kj}\eta_t\mathbb{E}[\|\nabla\ell(\theta_j^{(t)}; Z_{ij})\|_2] \tag{17}$$

Averaging over i and j and using Lemma 2.2 gives the final result.

□

Going back to the proof of Lemma 5.2, we can now average over k the equation from Lemma D.1 and use the double stochasticity of W^{T-t-1} to get

$$
\begin{aligned}
&\frac{1}{m}\sum_{k=1}^{m}|\mathbb{E}_{A,S}[R(A_k(S))-R_S(A_k(S))]| \leq \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)};Z_{ij})\|_2] \qquad (18)\\
&\leq \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)};Z_{ij})-\nabla R_{S_j}(\theta_j^{(t)})\|_2] + \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]\\
&= \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}\Big[\sqrt{\|\nabla\ell(\theta_j^{(t)};Z_{ij})-\nabla R_{S_j}(\theta_j^{(t)})\|_2^2}\Big] + \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]\\
&\leq \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\|\nabla\ell(\theta_j^{(t)};Z_{ij})-\nabla R_{S_j}(\theta_j^{(t)})\|_2^2\Big]} + \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]\ ,
\end{aligned}
$$

where we used Jensen inequality in the last step. Using the fact that, by Assumption 5.1, we have $\frac{1}{n}\sum_{i=1}^{n}\|\nabla\ell(\theta_j^{(t)};Z_{ij})-\nabla R_{S_j}(\theta_j^{(t)})\|_2^2 \leq \sigma^2$ finishes the proof.

□

D.2. Theorem 5.3

We start proving that under, the hypothesis of Theorem 5.3, all the eigenvalues of W belong to $(-1,1]$.

Lemma D.2. *Let W be an $n\times n$ symmetric, stochastic matrix with positive elements on the diagonal, then all the eigenvalues of W are in* $(-1,1]$.

Proof. As W is symmetric and stochastic, then the module of its largest eigenvalue is equal to 1 and all eigenvalues are in $[-1,1]$. We want now to prove that no eigenvalue can be equal to -1. By an opportune permutation of the nodes, we can write the matrix W as follows

$$
W = \begin{pmatrix} W_1 & 0_{n_1\times n_2} & \cdots & 0_{n_1\times n_C}\\ 0_{n_2\times n_1} & W_2 & \cdots & 0_{n_2\times n_C}\\ \cdots & \cdots & \cdots & \cdots\\ 0_{n_C\times n_1} & 0_{n_C\times n_2} & \cdots & W_C \end{pmatrix},
$$

where $0_{n\times m}$ denotes an $n\times m$ matrix with 0 elements, $\sum_{c=1}^{C} n_c = n$, and each matrix W_c has size $n_c\times n_c$ and is irreducible (Meyer, 2001)[p. 671]. Each matrix corresponds to a connected component of the communication graph.

The eigenvalues of W (taken with their multiplicity) are the eigenvalues of the different matrices W_c, for $c=1,\dots,C$. It is then sufficient to prove that the eigenvalues of each W_c are in $(-1,1]$. This result follows immediately from the fact that W_c is irreducible with non-negative elements on the diagonal, then it is primitive (Meyer, 2001)[Example 8.3.3], i.e., 1 is the only eigenvalue on the unit circle. □

The objective of the proof is to control the term $\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2]$ in Lemma 5.2. It starts with the following descent lemma.

Lemma D.3. (Descent Lemma). *Let the same setting as Theorem 3.1, with a constant stepsize $\eta>0$ and additional Assumption 5.1. We have:*

$$
\frac{\eta}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2^2] \leq \frac{2}{m}\sum_{j=0}^{m}\mathbb{E}\Big[R_{S_j}(\theta_j^{(t)})-R_{S_j}(\theta_j^{(t+1)})\Big] + \beta\sigma^2\eta^2 + \frac{1}{m\eta}\sum_{j=1}^{m}\mathbb{E}\|\sum_{l=1}^{m}W_{jl}\theta_l^{(t)}-\theta_j^{(t)}\|_2^2
$$

Proof. For all $j = 1, \dots, m$, the convexity and β-smoothness of R_{S_j} gives:

$$
\begin{aligned}
R_{S_j}(\theta_j^{(t+1)}) - R_{S_j}(\theta_j^{(t)}) &\le \langle \nabla R_{S_j}(\theta_j^{(t)}), \theta_j^{(t+1)} - \theta_j^{(t)} \rangle + \frac{\beta}{2} \|\theta_j^{(t+1)} - \theta_j^{(t)}\|_2^2 \\
&\le \Big\langle \nabla R_{S_j}(\theta_j^{(t)}), \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \eta \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) - \theta_j^{(t)} \Big\rangle + \frac{\beta}{2} \Big\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \eta \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) - \theta_j^{(t)} \Big\|_2^2 \\
&\le -\eta \Big\langle \nabla R_{S_j}(\theta_j^{(t)}), \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) \Big\rangle + \Big\langle \nabla R_{S_j}(\theta_j^{(t)}), \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\rangle \\
&\quad + \frac{\beta}{2} \Big(\Big\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\|_2^2 - 2\eta \Big\langle \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}), \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\rangle + \eta^2 \Big\| \nabla \ell(\theta_j^{(t)}; Z_{I_j^t j}) \Big\|_2^2 \Big)
\end{aligned} \tag{19}
$$

Taking the conditional expectation of (19) given $\mathcal{F}_t$, the natural filtration at time t, and the dataset S gives:

$$
\begin{aligned}
\mathbb{E}[R_{S_j}(\theta_j^{(t+1)}) - R_{S_j}(\theta_j^{(t)}) | \mathcal{F}_t, S] &\le -\eta \Big\| \nabla R_{S_j}(\theta_j^{(t)}) \Big\|_2^2 + (1 - \beta\eta) \Big\langle \nabla R_{S_j}(\theta_j^{(t)}), \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\rangle \\
&\quad + \frac{\beta}{2} \Big\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\|_2^2 + \frac{\beta\eta^2}{2n} \sum_{i=1}^n \Big\| \nabla \ell(\theta_j^{(t)}; Z_{ij}) \Big\|_2^2 \\
&\le \Big(\frac{\beta\eta^2}{2} - \eta \Big) \Big\| \nabla R_{S_j}(\theta_j^{(t)}) \Big\|_2^2 + \frac{\beta}{2} \Big\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\|_2^2 + \frac{\beta\sigma^2\eta^2}{2} \\
&\quad + (1 - \beta\eta) \Big\langle \nabla R_{S_j}(\theta_j^{(t)}), \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\rangle
\end{aligned} \tag{20}
$$

Applying the inequality $\langle a, b \rangle \le \frac{\alpha}{2}\|a\|_2^2 + \frac{\alpha^{-1}}{2}\|b\|_2^2$, which is true $\forall \alpha > 0$, to the last term in Equation 20 gives, with $\alpha = \eta$:

$$
\mathbb{E}[R_{S_j}(\theta_j^{(t+1)}) - R_{S_j}(\theta_j^{(t)}) | \mathcal{F}_t, S] \le -\frac{\eta}{2} \Big\| \nabla R_{S_j}(\theta_j^{(t)}) \Big\|_2^2 + \frac{1}{2\eta} \Big\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \Big\|_2^2 + \frac{\beta\sigma^2\eta^2}{2}
$$

Passing to the expectation and rearranging the terms gives:

$$
\eta \mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2^2] \le 2\mathbb{E}[R_{S_j}(\theta_j^{(t+1)}) - R_{S_j}(\theta_j^{(t)})] + \beta\sigma^2\eta^2 + \frac{1}{\eta} \mathbb{E}[\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \|_2^2] \; .
$$

Summing over $j = 1, \dots, m$ and dividing by m provides the desired result. ∎

The following lemma controls the last term in Lemma D.3, under the setting of Theorem 5.3.

Lemma D.4. (Decentralization error control). *Let the same setting as Theorem 3.1, with a constant stepsize η and additional Assumption 5.1. Assume further that W is symmetric. Then, there exist a graph-dependent constant $C_W < \infty$ such that:*

$$
\frac{1}{m\eta} \sum_{j=1}^m \mathbb{E}[\| \sum_{l=1}^m W_{jl}\theta_l^{(t)} - \theta_j^{(t)} \|_2^2] \le \eta L^2 (C_W^{(t)})^2 \; ,
$$

where $C_W^{(t)} \triangleq \sum_{s=0}^{t-1} \|W^s - W^{s+1}\|_2 \le C_W$ and $\|\cdot\|_2$ is the ℓ_2-operator norm.

Proof. Let us rewrite the desired quantity in matrix form. Let $\Theta^{(t)} \in \mathbb{R}^{m\times p}$ be the matrix that contains the parameters $\theta_1^{(t)}, \dots, \theta_m^{(t)}$ row-wise. In other word, the j-th row of $\Theta^{(t)}$ is $\theta_j^{(t)T}$. Similarly, let $\nabla L(\Theta^{(t)}; Z_{I^t}) \in \mathbb{R}^{m\times p}$ be the matrix that contains the stochastic gradients $\nabla \ell(\theta_j^{(t)}; Z_{I_j^t j})$, $j = 1, \dots, m$, also row-wise. When it is clear from context, we simply write $\nabla L(\Theta^{(t)})$. In matrix form, the quantity of interest is equal to $\frac{1}{m\eta}\mathbb{E}\|W\Theta^{(t)} - \Theta^{(t)}\|_F^2$ and the D-SGD (variant B) updates are:

$$\begin{aligned}\Theta^{(t)} &= W\Theta^{(t-1)} - \eta\nabla L(\Theta^{(t-1)}; Z_{I^{t-1}}) \\ &= W^t\Theta^{(0)} - \eta\sum_{s=0}^{t-1} W^s \nabla L(\Theta^{(t-s-1)}) \ .\end{aligned}$$

Hence,

$$\begin{aligned}\frac{1}{m\eta}\mathbb{E}\|W\Theta^{(t)} - \Theta^{(t)}\|_F^2 &= \frac{1}{m\eta}\mathbb{E}\|(I-W)\Theta^{(t)}\|_F^2 \\ &= \frac{1}{m\eta}\mathbb{E}\Big\|(I-W)\Big(W^t\Theta^{(0)} - \eta\sum_{s=0}^{t-1} W^s\nabla L(\Theta^{(t-s-1)})\Big)\Big\|_F^2 \\ &= \frac{1}{m\eta}\mathbb{E}\Big\|(W^t - W^{t+1})\Theta^{(0)} - \eta\sum_{s=0}^{t-1}(W^s - W^{s+1})\nabla L(\Theta^{(t-s-1)})\Big\|_F^2 \\ &= \frac{\eta}{m}\mathbb{E}\Big\|\sum_{s=0}^{t-1}(W^s - W^{s+1})\nabla L(\Theta^{(t-s-1)})\Big\|_F^2 \ , \end{aligned} \tag{21}$$

where we used the fact that $(W^t - W^{t+1})\Theta^{(0)} = (W^t - W^{t+1})(\Theta^{(0)} - \frac{\mathbf{1}\mathbf{1}^T}{m}\Theta^{(0)}) = \mathbf{0}$, since all agents start from the same initialization point $\theta^{(0)}$.

Let's now control the quantity of interest without the square over the norm:

$$\begin{aligned}\Big\|\sum_{s=0}^{t-1}(W^s - W^{s+1})\nabla L(\Theta^{(t-s-1)})\Big\|_F &\leq \sum_{s=0}^{t-1}\Big\|(W^s - W^{s+1})\nabla L(\Theta^{(t-s-1)})\Big\|_F \\ &\leq \sum_{s=0}^{t-1}\|W^s - W^{s+1}\|_2\|\nabla L(\Theta^{(t-s-1)})\|_F \\ &\overset{A.2.4}{\leq} \sqrt{m}L\sum_{s=0}^{t-1}\|W^s - W^{s+1}\|_2 = \sqrt{m}LC_W^{(t)}\end{aligned}$$

Raising the last quantity to the square and plugging it into Equation (21) gives the main result of Lemma D.4. It remains to prove that $\exists C_W < \infty$ such that $C_W^{(t)} \leq C_W$. To this aim, we are going to show the sufficient condition that the series $C_W^{(\infty)} = \sum_{s=0}^{\infty}\|W^s - W^{s+1}\|_2$ converges to some limit $C_W < \infty$.

Let $a_s \triangleq \|W^s - W^{s+1}\|_2$ be any term of the series and denote by $\lambda_1, \dots, \lambda_m$, the eigenvalues of W, which are in $(-1, 1]$ by Lemma D.2. We therefore have $a_s = \sup_k\{|\lambda_k^s - \lambda_k^{s+1}|\} = \sup_k\{|\lambda_k|^s|1-\lambda_k|\}$. We note that if $\lambda_k = 1$, $|\lambda_k|^s|1-\lambda_k| = 0$, so we can omit the eigenvalues equal to 1 is the computation of a_s:

$$a_s = \sup_{k:\lambda_k\neq 1}\{|\lambda_k|^s|1-\lambda_k|\} \ .$$

If all eigenvalues are equal to 1, $a_s = 0$ for all $s \in \mathbb{N}$ and we directly have the convergence. Otherwise, we are going to show that the series converge by using the Cauchy root test, which states that a series converges if $\limsup_{s\to\infty}|a_s|^{\frac{1}{s}} = r < 1$. As a matter of fact, we have

$$
\begin{aligned}
|a_s|^{\frac{1}{s}} &= \sup_{k:\lambda_k \neq 1} \{|\lambda_k|^s |1-\lambda_k|\}^{\frac{1}{s}} \\
&= \sup_{k:\lambda_k \neq 1} \{|\lambda_k| |1-\lambda_k|^{\frac{1}{s}}\} \underset{s\to\infty}{\longrightarrow} \sup_{k:\lambda_k \neq 1} \{|\lambda_k|\} < 1 \,,
\end{aligned}
$$

which allows to conclude that the series converges and that there exists $C_W < \infty$ such that $C_W^{(t)} \leq C_W$.

□

We can now prove Theorem 5.3.Combining Lemma D.4 with Lemma D.3, we have:

$$
\begin{aligned}
\frac{\eta}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2^2] &\leq \frac{2}{m}\sum_{j=0}^{m}\mathbb{E}\Big[R_{S_j}(\theta_j^{(t)}) - R_{S_j}(\theta_j^{(t+1)})\Big] + \beta\sigma^2\eta^2 + \eta L^2 (C_W^{(t)})^2 \\
&\leq \frac{2}{m}\sum_{j=0}^{m}\mathbb{E}\Big[R_{S_j}(\theta_j^{(t)}) - R_{S_j}(\theta_j^{(t+1)})\Big] + \beta\sigma^2\eta^2 + \eta L^2 C_W^2
\end{aligned} \tag{22}
$$

Moreover, from Lemma 5.2 and with multiple use of Jensen inequality, we have

$$
\begin{aligned}
\frac{1}{m}\sum_{k=1}^{m}|\mathbb{E}_{A,S}[R(A_k(S)) - R_S(A_k(S))]| &\leq \frac{2L\sigma\eta T}{mn} + \frac{2L}{mn}\sum_{t=0}^{T-1}\eta\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2] \\
&\leq \frac{2L\sigma\eta T}{mn} + \frac{2L\sqrt{T\eta}}{mn}\sqrt{\sum_{t=0}^{T-1}\eta\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2^2]}
\end{aligned} \tag{23}
$$

Summing Equation (22) over $t = 0\ldots, T-1$ gives:

$$
\begin{aligned}
\sum_{t=0}^{T-1}\eta\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2^2] &\leq \frac{2}{m}\sum_{j=0}^{m}\mathbb{E}\Big[R_{S_j}(\theta_j^{(0)}) - R_{S_j}(\theta_j^{(T)})\Big] + T\beta\sigma^2\eta^2 + T\eta L^2 C_W^2 \\
&\leq \frac{2}{m}\sum_{j=0}^{m}\mathbb{E}\Big[R_{S_j}(\theta_j^{(0)}) - R_{S_j}(\theta_{S_j}^{*})\Big] + T\beta\sigma^2\eta^2 + T\eta L^2 C_W^2
\end{aligned}
$$

Plugging this last equation into (23) gives the final result.

□

E. Additional results and discussions

E.1. On the generalization of $A(S) = \bar{\theta}^{(T)}$

In Remarks 3.2, we claimed that our generalization bounds are all also valid for the average of final iterates $A(S) = \bar{\theta}^{(T)}$. This is ensured by the following propositions.

Proposition E.1. *Let $A(S) = \bar{\theta}^{(T)}$. Under the same set of hypotheses, the upper-bounds derived in Theorem 3.1, Theorem 3.3 and Theorem 5.3 are also valid upper-bounds on $|\mathbb{E}_{A,S}[R(A(S)) - R_S(A(S))]|$.*

Proof. Extension of Theorem 3.1: Like in the proof of the latter theorem, we are going to show that $A(S)$ is on-average ε-stable with $\varepsilon \leq \frac{2L\sum_{t=0}^{T-1}\eta_t}{mn}$.

$$\frac{1}{mn}\sum_{i,j}\mathbb{E}[\|A(S)-A(S^{(i,j)})\|_2] \leq \frac{1}{mn}\sum_{i,j}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}[\|A_k(S)-A_k(S^{(i,j)})\|_2]$$
$$\leq \frac{1}{mn}\sum_{i,j}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}[\delta_k^{(T)}(i,j)] \ ,$$

where we took back the notation of the proof in B.1. Then based on Equation (9) and the double stochasticity of W and its powers, we get the desired result.

Extension of Theorem 3.3: In the same way, using this time Equation (11) we have:

$$\frac{1}{mn}\sum_{i,j}\mathbb{E}[\|A(S)-A(S^{(i,j)})\|_2] \leq \frac{1}{mn}\sum_{i,j}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}[\delta_k^{(T)}(i,j)] \leq \frac{4L}{\mu mn} \ ,$$

which shows that $A(S)$ with μ-strongly convex functions is on-average $\frac{4L}{\mu mn}$-stable.

Extension of Theorem 5.3: It suffices to show that Lemma 5.2 is also valid for $|\mathbb{E}_{A,S}[R(A(S))-R_S(A(S))]|$. Again we are going to use the link between generalization and on-average stability. This time, using Equation (17) , we have:

$$\frac{1}{mn}\sum_{i,j}\mathbb{E}[\|A(S)-A(S^{(i,j)})\|_2] \leq \frac{1}{mn}\sum_{i,j}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}[\delta_k^{(T)}(i,j)]$$
$$\leq \frac{1}{mn}\sum_{i,j}\frac{1}{m}\sum_{k=1}^{m}\frac{2}{n}\sum_{t=0}^{T-1}(W^{T-t-1})_{kj}\eta_t\mathbb{E}[\|\nabla\ell(\theta_j^{(t)};Z_{ij})\|_2]$$
$$= \frac{2}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)};Z_{ij})\|_2]$$

Using Lemma 2.2, we therefore have

$$|\mathbb{E}_{A,S}[R(A(S))-R_S(A(S))]| \leq \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}[\|\nabla\ell(\theta_j^{(t)};Z_{ij})\|_2] \ .$$

We recognize the right-hand term of Equation (18) and using the same arguments, we finally show that

$$|\mathbb{E}_{A,S}[R(A(S))-R_S(A(S))]| \leq \frac{2L\sigma}{mn}\sum_{t=0}^{T-1}\eta_t + \frac{2L}{mn}\sum_{t=0}^{T-1}\eta_t\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\|\nabla R_{S_j}(\theta_j^{(t)})\|_2] \ .$$

□

Proposition E.2. *Let $A(S)=\bar{\theta}^{(T)}$. Under the same set of hypotheses, the upper-bound derived in Theorem 4.1 is also a valid upper-bound on $|\mathbb{E}_{A,S}[R(A(S))-R_S(A(S))]|$.*

Proof. Replacing A_k by A in the proof of Lemma C.1, and using the fact that $\ell\in[0,1]$ we get:

$$
\begin{aligned}
|\mathbb{E}_{A,S}[R(A(S)) - R_S(A(S))]| &\leq \frac{L}{mn}\sum_{i,j}\mathbb{E}[\|A(S) - A(S^{(ij)})\| \big| \mathcal{E}(i,j)] + \frac{t_0}{n} \\
&\leq \frac{L}{mn}\sum_{i,j}\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}[\|\theta_k^{(T)} - \tilde{\theta}_k^{(T)}\| \big| \mathcal{E}(i,j)] + \frac{t_0}{n} \\
&= \frac{1}{m}\sum_{k=1}^{m}\Delta_k^{(T)} + \frac{t_0}{n}
\end{aligned}
$$

where $\Delta_k^{(T)}$ is defined in the proof of Theorem 4.1 and can be controlled in the exact same way (Eq. (15)), leading to the same final result.

□

E.2. On the stepsize assumption in Theorem 3.1

In this section, we show that the assumption $\eta \leq \frac{2\min_k W_{kk}}{\beta}$ is rather mild, and automatically verified for typical choices of stepsize that ensure the convergence of D-SGD. More specifically, note that, when W is symmetric, the iterates of Variant B are precisely the (stochastic) gradient steps for the optimization of the objective function

$$
F(\Theta) = \sum_{k=1}^{m}\mathbb{E}_{Z\sim\mathcal{D}_k}[\ell(\theta_k; Z)] + \frac{1}{2\eta}\Theta^\top(I - W)\Theta\,,
$$

where $\Theta \in \mathbb{R}^{m\times d}$ is the concatenation of all local parameters. As this objective function is smooth and convex, typical convex optimization analysis requires the stepsize to be smaller than $1/\beta_F$, where $\beta_F > 0$ is the smoothness constant of F (see e.g. Bubeck, 2015 or Garrigos & Gower, 2023, Theorem 3.4). However, a simple calculation shows that $\beta_F \leq \beta + \frac{1-\lambda_m}{\eta}$ (and this bound is tight as we have equality for the loss function $\ell(\theta, z) = \theta^2/2$), and the condition on the stepsize under our assumptions is thus $\eta\left(\beta + \frac{1-\lambda_m}{\eta}\right) \leq 1$, which gives

$$
\eta \leq \frac{\lambda_m}{\beta}\,.
$$

Finally, we conclude by noting that $\min_k W_{kk} = \min_k e_k^\top W e_k \geq \min_{u\,:\,\|u\|=1} u^\top W u = \lambda_m$, and thus the condition $\eta \leq \frac{\lambda_m}{\beta}$ directly implies $\eta \leq \frac{\min_k W_{kk}}{\beta}$ and the assumption of Theorem 3.1.

E.3. Mistake in Sun et al. (2021)

In Table 1 and Section 4, we claim that there is unfortunately a mistake in the proof of Theorem 3 in Sun et al. (2021), where the authors provide their generalization upper-bound for non-convex functions. In the paper, they provide an upper bound of order $\mathcal{O}\big(T^{\frac{\beta c}{\beta c+1}}(\frac{1}{mn} + C_\rho)\big)$, however, here we suggest that they should have a bound of order $\mathcal{O}\big(T^{\frac{\beta c}{\beta c+1}}(\frac{1}{n} + C_\rho)\big)$ instead.

Let's start by determining which part of the proof is wrong. To do this, we take the most recent version of the article on Arxiv as a reference, which can be found at: `https://arxiv.org/pdf/2102.01302.pdf`. The proof of Theorem 3 can be found on page 15, it relies on Lemma 7 which can be found on page 11. This Lemma is key in their proof and analogue to our Lemma C.1 with a uniform rather than an on-average model stability argument. With our notation, it states that $\forall z \in Z$, S and S' that differ in a single data point, they have:

$$
\mathbb{E}[|\ell(\bar{\theta}^{(T)}; z) - \ell(\bar{\tilde{\theta}}^{(T)}; z)|] \leq \frac{t_0}{n}\sup_{\theta,z}\ell(\theta; z) + L\mathbb{E}\Big[\|\bar{\theta}_k^{(T)} - \bar{\tilde{\theta}}^{(T)}\|\Big|\|\bar{\theta}^{(t_0)} - \bar{\tilde{\theta}}^{(t_0)}\| = 0\Big]\,.
$$

At this point, we observe that their Lemma is very similar to ours and notably that the first term, like us, is divided by n. The rest of the proof essentially consists of controlling the second term in the upper bound above. They show (page 16) that:

$$
\mathbb{E}\Big[\|\bar{\theta}_k^{(T)} - \bar{\tilde{\theta}}^{(T)}\|\Big|\|\bar{\theta}^{(t_0)} - \bar{\tilde{\theta}}^{(t_0)}\| = 0\Big] \leq \Big(\frac{2Lc}{mn} + 4(1 + cL)LC_\rho\Big)c\beta\left(\frac{T}{t_0}\right)^{c\beta}\,.
$$

However, when they plug this last equation into the one coming from Lemma 7 above, they claim that they have:

$$\mathbb{E}[|\ell(\bar{\theta}^{(T)};z) - \ell(\bar{\tilde{\theta}}^{(T)};z)|] \leq \frac{t_0}{mn} + L\Big(\frac{2Lc}{mn} + 4(1+cL)LC_\rho\Big)c\beta\left(\frac{T}{t_0}\right)^{c\beta} . \tag{24}$$

Here, we can see that for no specific reason, t_0 **is now divided by** mn **instead of** n **only**, which is a mistake. This obviously increases stability, and taking $t_0 = c^{\frac{1}{c\beta+1}}T^{\frac{c\beta}{c\beta+1}}$ leads to their final result. However, by taking this value for t_0 and dividing the first term by n only would give a rate of order $\mathcal{O}\big(T^{\frac{\beta c}{\beta c+1}}(\frac{1}{n} + C_\rho)\big)$ instead. We could imagine to take a better value for t_0 in the corrected version of Eq. (24). However, this is out of the scope of this paper and and in any case, the resulting bound would be worse than ours due to the additional term in C_ρ.

E.4. Experimental setup

We consider a logistic regression problem with two classes. Each data point (X,Y) is i.i.d. and drawn as follows. With probability 0.5, the point is first associated to a class $Y=0$ or 1. If $Y=0$ then X follows a bivariate random Gaussian variable with vector mean $(1,-1)$ and isotropic covariance I. If $Y=1$, then the vector mean is $(-1,1)$. To make the problem slightly more complicated and avoid separability, Y is then flipped with probability 0.1. We take the loss ℓ to be:

$$\ell(\theta;(x,y)) = -y\log\left(\frac{1}{1+\exp(-x^T\theta)}\right) - (1-y)\log\left(\frac{\exp(-x^T\theta)}{1+\exp(-x^T\theta)}\right) .$$

For the training, we have $m=20$ agents. To simulate the low noise regime, we take $n=1$ local data point (i.e. full batch: $\sigma^2=0$), while we take $n=10$ local data points in the higher noise regime. We then run D-SGD (Variant B) for $T=500$ iterations, with constant step size $\eta=0.03$ and initial point $\theta^{(0)}=\mathbf{0}$. We consider four communication graphs: (i) Complete graph with uniform weights $1/m$, (ii) Identity graph I (local SGD), (iii) Circle graph with self-edges and uniform weighs $1/3$, and (iv) Complete graph with diagonal elements equal to 0.95 and remaining elements uniformly equal to $0.05/(m-1)$.

At each iteration $t=1,\dots,T$, we compute a test loss (empirical population risk) using 500 i.i.d. data points, evaluated at all parameters $\theta_1^{(t)},\dots,\theta_m^{(t)}$, and compute the difference with the associated training loss (full empirical risk). Theses differences empirically correspond to local generalization errors, we then average them to match with Theorem 5.3.

We repeat the experiment over 50 different training-test data sets and for each data set, we run the algorithm 3 times, which leads to 150 different runs of D-SGD for each communication graphs. At the end, we average all these 150 runs and take the absolute value.