Journal of Machine Learning Research 21 (2020) 1-44 Submitted 9/18; Revised 7/19; Published 1/20

Graph-Dependent Implicit Regularisation for Distributed Stochastic Subgradient Descent

Dominic Richards DOMINIC.RICHARDS@SPC.OX.AC.UK
Department of Statistics
University of Oxford
24-29 St Giles', Oxford, OX1 3LB

Patrick Rebeschini PATRICK.REBESCHINI@STATS.OX.AC.UK
Department of Statistics
University of Oxford
24-29 St Giles', Oxford, OX1 3LB

Editor: Sathiya Keerthi

Abstract

We propose graph-dependent implicit regularisation strategies for synchronised distributed stochastic subgradient descent (Distributed SGD) for convex problems in multi-agent learning. Under the standard assumptions of convexity, Lipschitz continuity, and smoothness, we establish statistical learning rates that retain, up to logarithmic terms, single-machine serial statistical guarantees through implicit regularisation (step size tuning and early stopping) with appropriate dependence on the graph topology. Our approach avoids the need for explicit regularisation in decentralised learning problems, such as adding constraints to the empirical risk minimisation rule. Particularly for distributed methods, the use of implicit regularisation allows the algorithm to remain simple, without projections or dual methods. To prove our results, we establish graph-independent generalisation bounds for Distributed SGD that match the single-machine serial SGD setting (using algorithmic stability), and we establish graph-dependent optimisation bounds that are of independent interest. We present numerical experiments to show that the qualitative nature of the upper bounds we derive can be representative of real behaviours.

Keywords: Distributed machine learning, implicit regularisation, generalisation bounds, algorithmic stability, multi-agent optimisation.

1. Introduction

In machine learning, a canonical setting involves assuming that training data is made of independent samples from a certain unknown distribution, and the goal is to construct a model that can perform well on new unseen data from the same distribution (Vapnik, 1995). Given a certain loss function that measures the performance of a model against an individual data point, the classical framework of regularised empirical risk minimisation involves looking for the model that minimises the empirical risk, i.e., the average loss over the training set, under some notions of regularisation, and investigating the performance of this model on the expected risk or Test Risk, i.e., on the expected value of the loss function taken with respect to a new data point.

In the distributed setting, data is stored and processed in different locations by different agents. Each agent is represented by a node in a graph, and synchronised communication is allowed between neighbouring agents in this graph. In the decentralised setting typical of peer-to-peer networks, there is no central authority that can aggregate information from all the nodes and coordinate the distribution of computations. In sensor networks, for instance, data is collected on different sensors and each sensor communicates with nearby sensors by sharing model parameters. In the setting where the distributed data is assumed to be generated from the same unknown distribution, the goal is to design iterative algorithms so that agents can leverage local exchange of information to learn models that have better prediction capabilities as compared to the models they would obtain by only using the data they own.

In recent years, primarily due to the explosion in the size of modern data sets, the decentralised nature in which modern data is collected, and the rise of distributed computing platforms, the setting of distributed machine learning has received increased attention. From an optimisation point of view, problems in decentralised multi-agent learning are typically treated as a particular instance of consensus optimisation, and a variety of techniques have been developed to address this general framework, starting from the early work of Tsitsiklis (1984); Tsitsiklis et al. (1986) to more recent work that relates to the setting that we consider, which includes Johansson et al. (2007); Nedic and Ozdaglar (2009); Nedić et al. (2009); Johansson et al. (2009); Ram et al. (2010); Lobel and Ozdaglar (2011); Matei and Baras (2011); Boyd et al. (2011); Duchi et al. (2012); Shi et al. (2015); Mokhtari and Ribeiro (2016). From a statistical point of view, however, as emphasised in Shamir and Srebro (2014), distributed learning problems have more structure than general consensus problems, due to the possible statistical similarities in the data owned by different agents, for instance. Aside from the client-server (star network) setting where a central aggregator can coordinate the exchange of information with every other node so that divide and conquer protocols are admissible (Lin and Cevher, 2018), the literature on statistical guarantees for distributed methods seems to have focused exclusively on the investigation of models with explicit regularisation, i.e., when constraints and/or penalty terms are added to the minimisation of the empirical loss function (Agarwal and Duchi, 2011; Zhang et al., 2012; Shamir et al., 2014; Zhang and Lin, 2015; Bijral et al., 2017). The presence of explicit regularisation typically increases the complexity of both the algorithms and the resulting theoretical analysis, particularly for the distributed setting (Lian et al., 2017). For example, constraints can require the need for projection steps which are potentially costly for low-powered sensors, and deriving error bounds that depend on the graph topology for distributed algorithms in the presence of constraints is known to be challenging (Duchi et al., 2012). We are not aware of any result that investigates the performance of distributed and decentralised algorithms (i.e., not divide and conquer methods) on the Test Risk in the absence of explicit regularisation. This is in sharp contrast with the single-machine setting, where recent progress has been made giving optimal statistical learning guarantees for algorithms based on unregularised empirical risk minimisation via notions of implicit regularisation, i.e., proper tuning of algorithmic parameters (Ying and Pontil, 2008; Tarres and Yao, 2014; Dieuleveut and Bach, 2016; Lin et al., 2016a; Lin and Rosasco, 2017).

1.1. Contributions

This paper investigates the learning capabilities of a simple synchronised distributed first-order method for multi-agent learning using notions of implicit regularisation that depend on the topology of the underlying communication graph. We consider the unconstrained and unpenalised empirical risk minimisation problem in the setting where n agents have access to m independent data points coming from the same unknown distribution, and where agents can only exchange information with their neighbours. We consider a synchronised distributed version of stochastic subgradient descent

(Distributed SGD), which is a stochastic variant of one of the most widely-studied first-order method in multi-agent optimisation (Nedic and Ozdaglar, 2009). In the implementation that we look at, at every iteration each agent first performs a standard SGD step, where only one data point is uniformly sampled with replacement among those individually-owned to compute the local subgradient, and then performs a classical synchronised consensus step, where a local exchange of information is implemented via an average of the updated iterates across neighbouring agents. We treat Distributed SGD as a *statistical* device, and look at its performance on unseen data by bounding the Test Error, i.e., the expected value of the excess risk defined as the difference between the Test Risk evaluated at the output of the algorithm and the minimal Test Risk. Under different assumptions on the convex loss function (we consider the standard assumptions of Lipschitz and smoothness) we establish upper bounds for the Test Error of Distributed SGD that exhibit explicit dependence on both the algorithmic tuning parameters (the learning rate and the time horizon) and the graph topology (the spectral gap of the communication matrix). Minimising these upper bounds yields implicit regularisation strategies, allowing to recover the single-machine serial statistical rates by proper tuning of the learning rates and of the time horizon (a.k.a. early stopping) as a function of the network topology. In the case of convex, Lipschitz, and smooth losses, we recover, up to logarithmic terms, the optimal rate of $O(1/\sqrt{nm})$ for single-pass constrained single machine serial SGD (Lan, 2012; Xiao, 2010). In the case of convex and Lipschitz losses, we recover, up to logarithmic terms, the best-known rate of $O(1/(nm)^{1/3})$ for single-machine serial SGD with implicit regularisation Lin et al. (2016a,b).[1] We present numerical experiments to show that the qualitative nature of the upper bounds we derive can be representative of real behaviours.

To establish learning rates for Distributed SGD, we follow the general framework pioneered in the single-machine setting by Bousquet and Bottou (2008) and, in particular, by Hardt et al. (2016). We consider, in the distributed setting, a decomposition of the Test Error which involves the Generalisation Error (i.e., the expected value of the difference between the loss incurred on the training data versus the loss incurred on a new data point) and the Optimisation Error (i.e., the expected value of the error on the training set). To bound the Generalisation Error, we use algorithmic stability or sensitivity as initially put forward by Bousquet and Elisseeff (2002) and later applied for single-machine serial stochastic subgradient descent in Hardt et al. (2016). The notion of stability that we use measures how much the output of an algorithm differs when a single observation is resampled. In our case, as the observations are spread throughout the communication graph, we need to consider stability not only with respect to time (i.e., the iteration time of the algorithm), but also with respect to space (i.e., the communication graph). This technology allows us to establish generalisation bounds for Distributed SGD that do not depend on the topology of the communication graph, and we recover the same type of results that hold in the single machine serial setting. This is in contrast to optimisation bounds for distributed subgradient methods, which typically depend on the graph topology, as initially seen in the work of Johansson et al. (2009, 2007); Duchi et al. (2012). To bound the Optimisation Error, we follow the approach pioneered in Nedic and Ozdaglar (2009) and compare the behaviour of Distributed SGD with its network average, and we take inspiration from the analysis of the network term in the work of Duchi et al. (2012) (in the case of dual methods for constrained problems with Lipschitz losses) to derive upper bounds that depend on the graph topology via the inverse of the spectral gap of the communication matrix. In our setting, as we investigate implicit regularisation strategies, we deal with unconstrained problems

[1] Lin et al. (2016b) considers implicit regularisation for gradient descent, although they remark that the analysis can be modified to account for stochastic gradients.

and the evolution of the network-averaged process admits a simple form that facilitates the analysis. This approach avoids the difficulties with the nonlinearity of projection that have been previously challenging in distributed learning models, and that motivated the investigation of dual methods such as in Duchi et al. (2012). The bounds that we establish for the Optimisation Error of Distributed SGD seem novel and are of independent interest.

Finally, our results show that one can also think of the graph itself as a regularisation parameter. To give an example, agents can achieve the same statistical guarantees by trading off communication against iterations: they can choose to communicate by using a low-energy sparse communication protocol per iteration (for instance, communicating using a grid-like protocol even if the underlying topology is that of a complete graph and all agents are connected with each others), but would need to communicate for a longer time horizon in order to be guaranteed to reach the same level of statistical accuracy.

The main contributions of this work are here summarised.

1. **Graph-dependent implicit regularisation.** We propose graph-dependent implicit regularisation strategies for problems in distributed machine learning, specifically, step size tuning and early stopping as a function of the spectral gap of the communication matrix. Our results also show that the graph itself can be interpreted as a regularisation parameter.

2. **Optimal statistical rates using a simple algorithm.** Using implicit regularisation, we show how a simple, primal, unconstrained, first-order method (Distributed SGD) recovers, up to logarithmic terms, centralised statistical rates, in particular matching the optimal rates in the case of smooth loss functions for constrained single-pass serial SGD.

3. To establish statistical rates and control the Test Error of Distributed SGD, we use a distributed version of the error decomposition proposed in Hardt et al. (2016). We establish error bounds on the Generalisation Error and Optimisation Error, respectively.

   (a) **Distributed generalisation bounds.** We establish graph-independent Generalisation Error bounds for Distributed SGD that match those within Hardt et al. (2016) for the single-machine serial case. In the case of convex losses that are Lipschitz and smooth, we prove upper bounds that grow linearly with the number of iterations and step size.

   (b) **Distributed optimisation bounds.** We establish graph-dependent Optimisation Error bounds for Distributed SGD. In the case of convex and Lipschitz loss functions, our analysis is inspired by Nedic and Ozdaglar (2009); Duchi et al. (2012). When smoothness is considered, our analysis is inspired by Bubeck et al. (2015); Dekel et al. (2012).

The remainder of the work is laid out as follows. Section 2 introduces the framework of multi-agent learning. Section 3 introduces the Distributed SGD algorithm. Section 4 presents the main results of this work, Test Error bounds for Distributed SGD with convex, Lipschitz, and either smooth or non-smooth losses. Section 5 presents the specific Generalisation and Optimisation Error bounds, as well as the notion of stability that we use. Section 6 gives a simulation study for the case of smooth losses. Section 7 contains the conclusion. Appendix A provides proofs for all Generalisation and Test Error bounds. Appendix B gives proofs for Optimisation Error bounds under a general first-order stochastic oracle model.

2. Multi-Agent Learning

In this section we introduce the framework of distributed and decentralised machine learning that we consider. We address the case in which agents or nodes in a network are given their own independent data sets and they want to cooperate, by iteratively exchanging information with their neighbours, to develop a good learning model for new unseen data.

Let (V, E) be a simple undirected graph with n nodes, $V = \{1, \dots, n\} \equiv [n]$ being the vertex set and $E \subseteq V \times V$ being the edge set. Let $\mathcal{Z}$ be the space of observations, and to each $v \in V$ let $\mathcal{D}_v := \{Z_{v,1}, \dots, Z_{v,m}\} \in \mathcal{Z}^m$ denote the training set associated to node v, which consists of m i.i.d. data points sampled from a certain unknown distribution supported on $\mathcal{Z}$. Let $\mathcal{D} := \cup_{v \in V} \mathcal{D}_v$ denote the collection of all data points, that is, the entire/global training data set. Let $d > 0$ be a given positive integer, and define $\mathcal{X} = \mathbb{R}^d$. Each agent wants to find a model $x^\star \in \mathcal{X}$ that minimises of the Test Risk r, which is defined as

$$r(x) := \mathbf{E}\, \ell(x, Z).$$

Here, the function $\ell : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is a given loss function, and $\ell(x, Z)$ represents the loss of the model x on the random sample Z, which represents a new (unseen, independent) data point from the same distribution. We assume that the minimum of r can be achieved. As the distribution of the data is unknown, the expected risk r can not be computed, and a popular approach in machine learning is to consider the empirical risk as a proxy. In the distributed setting, the global empirical risk R is defined as

$$R(x) := \frac{1}{nm} \sum_{v \in V} \sum_{i=1}^{m} \ell(x, Z_{v,i}) = \frac{1}{n} \sum_{v \in V} R_v(x).$$

Here, we have further defined the local empirical risk $R_v(x) := \frac{1}{m} \sum_{i=1}^{m} \ell(x, Z_{v,i})$, for any $v \in V$. Let us denote by $X^\star \in \operatorname{argmin}_{x \in \mathcal{X}} R(x)$ a minimiser of the global empirical risk. In the decentralised setting that we consider, each agent $v \in V$ iteratively exchanges information with their neighbours for a certain amount of time steps t to construct a model $X_v^t \in \mathcal{X}$ that can be a good proxy for the minimiser of the expected risk, i.e., for $x^\star \in \operatorname{argmin}_{x \in \mathcal{X}} r(x)$. A way to assess the quality of a model X_v^t is to consider the Test Error, which we define as the expected value of the excess risk $r(X_v^t) - r(x^\star)$, namely,

$$\mathbf{E}\, r(X_v^t) - r(x^\star).$$

In the next section we introduce the specific distributed algorithm that we consider to generate the models' estimates X_v^t's, and we then present the main results on the bounds for the Test Error. The general paradigm that we adopt to bound the Test Error is given by a generalisation to the distributed setting of the error decomposition given in Hardt et al. (2016) for the single-machine setting. This decomposition allows to bound the Test Error of a model into the sum of two errors: the *Generalisation Error*, which controls the difference between the performance of the model on a new data point and the performance of the model on the training data in $\mathcal{D}$; and the *Optimisation Error*, which controls how well the model optimises the empirical risk.

Proposition 1 (Hardt et al. (2016)) *For each $v \in V$, $t \geq 1$ we have*

$$\underbrace{\mathbf{E}\, r(X_v^t) - r(x^\star)}_{\textit{Test Error}} \leq \underbrace{\mathbf{E}[r(X_v^t) - R(X_v^t)]}_{\textit{Generalisation Error}} + \underbrace{\mathbf{E}[R(X_v^t) - R(X^\star)]}_{\textit{Optimisation Error}}.$$

Proof For completeness, the proof from Hardt et al. (2016) is given in Appendix A.1. ■

By using the error decomposition in Proposition 1, we are able to consider the unregularised empirical risk minimisation problem introduced above and develop implicit regularisation strategies for a simple iterative algorithm, which we introduce next.

Remark 2 (Statistical optimisation) *From the statistical point of view, the distributed setting where each agent is given a subset of the data has received a lot of attention in the literature (see introduction), though most of the literature on statistical optimisation has focused on the client-server (also known as master-slave) architecture typical of data centers, where a central aggregator in the network (the server) can communicate with every other nodes (the clients) and can thus coordinate the processing and exchange of information. This amounts to a star network topology that can be used to model shared-memory protocols. This type of architecture makes divide-and-conquer strategies possible, and most of the literature on statistical optimisation has focused on investigating statistical rates on the Test Error for one-shot-averaging techniques. In this work, we focus on the decentralised setting where all nodes iteratively perform the same type of computations and communications with respect to the underlying graph structure, without the presence of any special node. We are not aware of any prior work that directly investigates the statistical performance of decentralised methods on the Test Error. Most of the literature on decentralised methods seem to have focused on bounding the Optimisation Error on the training data, as we explain in Remark 3.*

Remark 3 (Consensus optimisation) *From the optimisation point of view, the literature on multi-agent learning has largely focused on the investigation of the Optimisation Error via consensus methods in the presence of explicit regularisation, typically in the form of a convex constraint set $\mathcal{R}$ (see literature review in the introduction). Statistically, this approach is justified, for instance, by the distributed version of the classical error decomposition given in Bousquet and Bottou (2008):*

$$\underbrace{\mathbf{E}\, r(X_v^t) - r(x^\star)}_{\textit{Test Error}} \leq 2 \underbrace{\mathbf{E} \sup_{x \in \mathcal{R}} |r(x) - R(x)|}_{\textit{Uniform Generalisation Error}} + \underbrace{\mathbf{E}[R(X_v^t) - R(X_{\mathcal{R}}^\star)]}_{\textit{Regularised Optimisation Error}} + \underbrace{r(x_{\mathcal{R}}^\star) - r(x^\star)}_{\textit{Approximation Error}},$$

with $x_{\mathcal{R}}^\star \in \operatorname{argmin}_{x \in \mathcal{R}} r(x)$ and $X_{\mathcal{R}}^\star \in \operatorname{argmin}_{x \in \mathcal{R}} R(x)$. In this setting, consensus optimisation deals with algorithms that minimise the quantity $R(X_v^t) - R(X_{\mathcal{R}}^\star)$, where $R(x) = \frac{1}{n}\sum_{v \in V} R_v(x)$. Bounds on the Regularised Optimisation Error can then be combined with bounds on the Uniform Generalisation Error using notions of complexity for the constraint set $\mathcal{R}$ (e.g., VC dimension, Rademacher complexity, etc.). As highlighted in Shamir and Srebro (2014), and as we mentioned in the introduction, however, distributed learning problems have more structure than general consensus problems, as the local functions R_v are random and have a specific design. In this work, we analyse a stochastic algorithm that is tailor-made for distributed learning problems (not for general consensus problems), and use the error decomposition in Proposition 1 to develop implicit regularisation strategies for the unregularised empirical risk minimisation problem.

3. Distributed Stochastic Subgradient Descent

The algorithm that we consider to generate the model estimates X_v^t's assumes that each node $v \in V$ can query subgradients $\partial \ell$ of the loss function ℓ with respect to the first parameter, evaluated at points in the local data set $\mathcal{D}_v$. We consider the stochastic setting where at each time step agent v does not evaluate the full subgradient of the local empirical risk R_v, but instead only a subgradient $\partial\ell$ at a single randomly chosen sample in the locally-owned data set $\mathcal{D}_v$. This is well tailored to situations where m is large, as this reduces the per-iteration complexity to a constant factor.

The algorithm is defined as follows. Let $\partial\ell(x, Z_{v,k})$ represent an element of the subgradient of $\ell(\,\cdot\,, Z_{v,k})$ at x, with $k \in \{1, \dots, m\} \equiv [m]$. Let $P \in \mathbb{R}^{n\times n}$ be a doubly stochastic matrix supported on the graph (V, E), that is, $P_{ij} \neq 0$ only if $\{i, j\} \in E$. Distributed stochastic subgradient descent (Distributed SGD) generates a collections of vectors $\{X_v^s\}_{v\in V, s\geq 1}$ in $\mathcal{X}$ as follows. Given initial vectors $\{X_v^1\}_{v\in V}$, possibly random, for $s \geq 1$,

$$X_v^{s+1} = \sum_{w\in V} P_{vw}(X_w^s - \eta\partial\ell(X_w^s, Z_{w,K_w^{s+1}})), \tag{1}$$

where for each $v \in V$, $\{K_v^2, K_v^3, \dots\}$ is a collection of i.i.d. random variables uniform in $[m]$, and $\eta > 0$ is the step size. The above algorithm can be described as performing two steps: a stochastic gradient update $Y_w^{s+1} = X_w^s - \eta\partial\ell(X_w^s, Z_{w,K_w^{s+1}})$, and a synchronised consensus step $\sum_{w\in V} P_{vw} Y_w^{s+1}$. This framework for decentralised optimisation (albeit for a slightly different protocol, see remark 4) has been largely explored with the early works of Nedic and Ozdaglar (2009); Ram et al. (2009); Lobel and Ozdaglar (2011); Duchi et al. (2012). The fact that we consider implicit regularisation strategies allows us to focus on the unconstrained risk minimisation problem. In turn, this allows us to consider an algorithm that is much simpler to analyse than the ones previously considered in the literature, avoiding projections or dual approaches (see introduction for the relevant literature review). We also highlight the randomised sampling mechanism in algorithm (1), which is tailor-made for the machine learning problem at hand and not for generic consensus problems.

Remark 4 *In the stochastic setting, the protocol put forward by Nedic and Ozdaglar (2009) updates the iterates as $X_v^{s+1} = \sum_{w\in V} P_{vw} X_w^s - \eta\partial\ell(X_v^s, Z_{v,K_v^{s+1}})$, which is slightly different from the protocol that we consider where also the gradients are averaged across neighbours. The two main motivations for the original protocol are that it is fully decentralised, in that nodes are only required to communicate locally, and that it reduces to a consensus protocol to solve network averaging problems when $\ell = 0$. The protocol* (1) *that we consider preserves these properties and it makes the error analyses simpler. The difference between these two protocols in a general setting has been investigated in the literature, see Sayed (2014) for instance.*

In the next section we present results on the performance of Distributed SGD under various assumptions on the loss function ℓ.

4. Results

This section presents the main results of this work: Test Error bounds for Distributed SGD with smooth and non-smooth losses, Section 4.1 and Section 4.2, respectively.

Henceforth, let $\|\cdot\|$ be the ℓ_2 norm. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be L-Lipschitz, with $L > 0$, if $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$, and β-smooth, with $\beta > 0$, if $\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$ for all $x, y \in \mathbb{R}^d$. Let $\sigma_2(P)$ be the second largest eigenvalue in absolute value for the matrix P. Unless stated otherwise, we use the big-O notation $O(\,\cdot\,)$ to denote order of magnitudes up to constants in n and m, and the notation $\widetilde{O}(\,\cdot\,)$ to denote order of magnitudes up to both constants and logarithmic terms in n and m. Equality modulo constants and logarithmic terms is denoted by $\simeq$.

4.1. Smooth Losses

We analyse the statistical rates for smooth losses. First, we present the Test Error bound in its full form. Then, we present a corollary that summarises the order of magnitudes of the bounds obtained under different choices of implicit regularisation, tuning the step size and the stopping time as a function of the graph topology. Full details are given in Appendix A.

For smooth losses, we present a bound that depends on both the variance of the gradient estimates and the statistical deviations between the local empirical losses $\{R_v\}_{v\in V}$. Let $\sigma, \kappa > 0$ be such that the following holds for any $v \in V$ and $s \geq 1$,

$$\mathbf{E}\big[\|\nabla \ell(X_v^s, Z_{v,K_v^{s+1}}) - \nabla R_v(X_v^s)\|^2\big] \leq \sigma^2, \tag{2}$$

$$\mathbf{E}\big[\|\nabla \ell(X_v^s, Z_{v,K_v^{s+1}}) - \frac{1}{n}\sum_{w\in V} \nabla R_w(X_w^s)\|^2\big] \leq \kappa^2. \tag{3}$$

The quantity σ^2 in (2) yields a uniform control on the variance of the stochastic gradients, while the quantity κ^2 in (3) yields a uniform control on both the variance of the gradients and the deviation between local objectives. Note that if $\ell(\,\cdot\,, z)$ is L-Lipschitz for any $z \in \mathcal{Z}$, then both σ^2 and κ^2 are bounded by $4L^2$ by the triangle inequality. A detailed discussion of these assumptions is given in Appendix B in the more general context of stochastic optimisation.

Theorem 5 (Test Error bounds for convex, Lipschitz, and smooth losses) *Assume that for any $z \in \mathcal{Z}$ the function $\ell(\,\cdot\,, z)$ is convex, L-Lipschitz, β-smooth and satisfies* (2) *and* (3). *Let $X_v^1 = 0$ for all $v \in V$, $\|X^\star\| \leq G$. Then, Distributed SGD with $\eta = 1/(\beta + 1/\rho)$, $\rho > 0$, and $\eta\beta \leq 2$, yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) \leq \underbrace{\frac{L^2}{nm(\beta+1/\rho)}(t+1)}_{\textit{Generalisation Error bound}}$$
$$+ \underbrace{\frac{\rho}{2}\sigma^2 + \frac{(\beta+1/\rho)G^2}{2t} + \frac{3\kappa}{\beta+1/\rho}\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big(L + \frac{3}{2}\frac{\beta(3+\beta\rho)\kappa}{\beta+1/\rho}\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)}_{\textit{Optimisation Error bound}}.$$

Proof See Appendix A.5. ■

We highlight that the Generalisation Error bound is independent of the graph topology, while the Optimisation Error bound naturally depends upon inverse of the spectral gap of the communication matrix: $1/(1-\sigma_2(P))$. The following corollary gives the order of magnitudes for the Test Error bounds obtained with three different choices of step size and corresponding early stopping. The different choices for the parameter $\rho > 0$ correspond to the following (modulo the simplifications used to perform the minimisations, as explained in detail in Section A.6):

- $\rho^\star$ is the choice for serial SGD, see for instance Dekel et al. (2012); Bubeck et al. (2015);

- $\rho^\star_{\text{Opt}}$ is the choice that minimises the Optimisation Error bound in Theorem 5;
- $\rho^\star_{\text{Test}}$ is the choice that minimises the Test Error bound in Theorem 5.

Corollary 6 (Implicit regularisation for convex, Lipschitz, and smooth losses) *In the setting of Theorem 5, the following holds for different choices of ρ, function of the time horizon t:*

ρ	Size	Test Error at ρ, t	Test Error at $\rho, t^\star(\rho)$
$\rho^\star$	$O\Big(\frac{1}{\sqrt{t}}\Big)$	$\widetilde{O}\Big(\frac{1}{(1-\sigma_2(P))\sqrt{t}} + \frac{\sqrt{t}}{nm}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{nm(1-\sigma_2(P))}}\Big)$
$\rho^\star_{\text{Opt}}$	$\widetilde{O}\Big(\sqrt{\frac{1-\sigma_2(P)}{t}}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{t(1-\sigma_2(P))}} + \frac{\sqrt{t(1-\sigma_2(P))}}{nm}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{nm}}\Big)$
$\rho^\star_{\text{Test}}$	$\widetilde{O}\Big(\frac{1}{\sqrt{t}}\frac{1}{\sqrt{\frac{1}{1-\sigma_2(P)}+\frac{t}{nm}}}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{t(1-\sigma_2(P))}} + \frac{1}{\sqrt{nm}}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{nm}}\Big)$

Table 1: $t^\star(\rho^\star) \simeq t^\star(\rho^\star_{\text{Opt}}) \simeq t^\star(\rho^\star_{\text{Test}}) \simeq nm/(1-\sigma_2(P))$.

Proof See Appendix A.6. ■

We note that the Test Error bound given by the choice $\rho^\star_{\text{Test}}$ is the only one that is guaranteed to converge as the number of iterations t goes to infinity. With this choice, $t^\star(\rho^\star_{\text{Test}}) \simeq nm/(1-\sigma_2(P))$ iterations are guaranteed to reach the rate $\widetilde{O}(1/\sqrt{nm})$. Minimising (approximately) with respect to time the Test Error bounds that are obtained with the choices $\rho^\star$ and $\rho^\star_{\text{Opt}}$ gives early stopping rules with the same order of iterations, i.e., $t^\star(\rho^\star) \simeq t^\star(\rho^\star_{\text{Opt}}) \simeq nm/(1-\sigma_2(P))$. The choices $\rho^\star_{\text{Test}}$ and $\rho^\star_{\text{Opt}}$ with early stopping yield, up to logarithmic terms, the optimal rate $O(1/\sqrt{nm})$ for single-pass constrained serial SGD (Lan, 2012; Xiao, 2010). On the other hand, the choice $\rho^\star$ that aligns with serial SGD, with no dependence on the graph topology, yields a suboptimal statistical guarantee with a rate $\widetilde{O}(1/\sqrt{nm(1-\sigma_2(P))})$.

Remark 7 (Knowledge of Network Spectrum) *Algorithmic parameter choices in Table 1 depend on the network through the spectral gap of the communication matrix $1-\sigma_2(P)$. While outside the scope of this work, this quantity can be estimated in a decentralised manner, see for instance (Yang et al., 2010; Yang and Tang, 2015) and references therein.*

Remark 8 (Early Stopping with a Constant Step Size) *When performing early stopping a step size constant in the number of iterations is commonly chosen so a single instance of single-machine serial SGD is required. Theorem 5 demonstrates optimal statistical rates up to logarithmic factors can be achieved for Distributed SGD when choosing the step size $\rho = O((1-\sigma_2(P)))/\sqrt{nm})$ and iterations $t = O(nm/(1-\sigma_2(P)))$. For the calculation of this fact see Appendix A.8.*

4.2. Non-Smooth Losses

We now analyse the statistical rates for non-smooth losses. Before presenting the results, we introduce and motivate the technical assumptions that we need.

Assumptions 1

(a) There exist constants $C \leq B$ such that for any $z \in \mathcal{Z}$ the loss function $\ell(\,\cdot\,, z)$ is bounded from above at zero, i.e., $\ell(0, z) \leq B$, and is uniformly bounded from below, i.e., $C \leq \ell(x, z)$ for any $x \in \mathbb{R}^d$;

(b) There exists a constant $D \geq 0$ such that for any $z_1, \ldots, z_{nm} \in \mathcal{Z}$ and any $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ we have

$$\mathbf{E} \sup_{x \in \widetilde{\mathcal{X}}} \frac{1}{nm} \sum_{i=1}^{nm} \sigma_i \ell(x, z_i) \leq D \frac{\sup_{x \in \widetilde{\mathcal{X}}} \|x\|}{\sqrt{nm}},$$

where $\{\sigma_i\}_{i \in [nm]}$ is a collection of independent Rademacher random variables, namely, $\mathbf{P}(\sigma_i = 1) = \mathbf{P}(\sigma_i = -1) = 1/2$.

Assumption *(a)* is a common boundedness assumption for controlling the norm of the iterates of gradient descent algorithms through a centring argument. Assumption *(b)* represents a control on the Rademacher complexity of the function class $\{\ell(x, \,\cdot\,) : x \in \mathcal{X}\}$ with respect to the ℓ_2 norm. These assumptions are satisfied, for instance, in the setting of supervised learning with linear predictors, bounded data, and hinge loss (with is convex, Lipschitz, and non-smooth). See Remark 11 below.

First, we present the Test Error bound for non-smooth losses under Assumptions 1. Then, we present a corollary that summarises the order of magnitudes of the bounds obtained under different choices of implicit regularisation, tuning the step size and the stopping time as a function of the graph topology. Full details are given in Appendix A.

Theorem 9 (Test Error bounds for convex and Lipschitz losses) *Assume that for any $z \in \mathcal{Z}$ the loss function $\ell(\,\cdot\,, z)$ is convex and L-Lipschitz. Consider Assumptions 1. Let $X_v^1 = 0$ for all $v \in V$, $\|X^\star\| \leq G$. Then, Distributed SGD with $\eta > 0$ yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^s\Big) - r(x^\star) \leq \underbrace{2D\sqrt{\frac{(t-1)(\eta^2 L^2 + 2\eta(B-C))}{nm}}}_{\textit{Generalisation Error bound}} + \underbrace{\frac{19}{2}\frac{\eta L^2 \log(t\sqrt{n})}{1-\sigma_2(P)} + \frac{G^2}{2\eta t}}_{\textit{Optimisation Error bound}}.$$

Proof See Appendix A.5. ■

The following corollary gives the order of magnitudes for the Test Error bound obtained with three different choices of step size and corresponding early stopping. The different choices for the step size $\eta > 0$ correspond to the following (modulo the simplifications used to perform the minimisations, as explained in detail in Section A.7):

- $\eta^\star$ is the choice for serial SGD, see for instance Bubeck et al. (2015);
- $\eta^\star_{\text{Opt}}$ is the choice that minimises the Optimisation Error bound in Theorem 9;
- $\eta^\star_{\text{Test}}$ is the choice that minimises the Test Error bound in Theorem 9.

Corollary 10 (Implicit regularisation for convex and Lipschitz losses) *In the setting of Theorem 9, the following holds for different choices of η, function of the time horizon t:*

η	Size	Test Error at η, t	Test Error at $\eta, t^\star(\eta)$
$\eta^\star$	$O\Big(\frac{1}{\sqrt{t}}\Big)$	$\widetilde{O}\Big(\frac{1}{(1-\sigma_2(P))\sqrt{t}} + \sqrt{\frac{\sqrt{t}}{nm}}\Big)$	$\widetilde{O}\Big(\frac{1}{(nm(1-\sigma_2(P)))^{1/3}}\Big)$
$\eta^\star_{\text{Opt}}$	$\widetilde{O}\Big(\sqrt{\frac{1-\sigma_2(P)}{t}}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{t(1-\sigma_2(P))}} + \sqrt{\frac{\sqrt{t(1-\sigma_2(P))}}{nm}}\Big)$	$\widetilde{O}\Big(\frac{1}{(nm)^{1/3}}\Big)$
$\eta^\star_{\text{Test}}$	$\widetilde{O}\Big(\frac{1}{\sqrt{t}}\frac{1}{\sqrt{\frac{1}{1-\sigma_2(P)}+\frac{t}{(nm)^{2/3}}}}\Big)$	$\widetilde{O}\Big(\frac{1}{\sqrt{t(1-\sigma_2(P))}} + \frac{1}{(nm)^{1/3}}\Big)$	$\widetilde{O}\Big(\frac{1}{(nm)^{1/3}}\Big)$

Table 2: $t^\star(\eta^\star) \simeq (nm)^{2/3}/(1-\sigma_2(P))^{4/3}$ and $t^\star(\eta^\star_{\text{Opt}}) \simeq t^\star(\eta^\star_{\text{Test}}) \simeq (nm)^{2/3}/(1-\sigma_2(P))$.

Proof See Appendix A.7. ■

Corollary 10 shows asymptotic behaviours for the Test Error bounds (as a function of time t upon different choices of the step size) that are analogous to the ones established in Corollary 6 in the case of smooth losses. In particular, as in Corollary 6, the step sizes accounting for the graph topology, i.e., $\eta^\star_{\text{Test}}$ and $\eta^\star_{\text{Opt}}$, give improved statistical rates over the step size independent of the graph topology $\eta^\star$.

The statistical rate obtained by both $\eta^\star_{\text{Test}}$ and $\eta^\star_{\text{Opt}}$, upon performing early stopping, matches, up to logarithmic terms, the best-known rate of $O(1/(nm)^{1/3})$ obtained by serial SGD with implicit regularisation (Lin et al., 2016a). Differing from the smooth case, additional iterations with respect to the graph topology are required for the step size independent of the graph topology $\eta^\star$ to achieve its best statistical rates (as prescribed by our upper bounds), when compared to step sizes accounting for the topology $\eta^\star_{\text{Test}}$ and $\eta^\star_{\text{Opt}}$. As highlighted in (Lin et al., 2016a), we note that these rates are not sharp, leaving it to future work to obtain better bounds.

Remark 11 *Assumptions 1 is satisfied in the setting of supervised learning with bounded data, linear predictors, and hinge loss, for instance. In this setting, each observation $z \in \mathcal{Z}$ decomposes into a d-dimensional feature vector and a real-valued response, i.e., $z = \{w, y\}$ with $w \in \mathcal{W} \subset \mathbb{R}^d$ and $y \in \mathcal{Y} \subset \mathbb{R}$ such that $\|w\| \le D_{\mathcal{W}} < \infty$, and $|y| \le D_{\mathcal{Y}} < \infty$. The linear predictors are parametrised by $x \in \widetilde{\mathcal{X}} \subseteq \mathcal{X} = \mathbb{R}^d$, i.e., $w \to w^\top x$, and the loss function is of the form $\ell(x, z) = \tilde{\ell}(w^\top x, y)$ with the function $\tilde{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ measuring the discrepancy between the predicted response $w^\top x$ and the observed response y. For the hinge loss, $\tilde{\ell}(\tilde{y}, y) = \max(0, 1 - \tilde{y}y)$. Assumption 1 (a) is satisfied with $B = 1$ and $C = 0$. By Talagrand's contraction lemma and standard results on the Rademacher complexity of linear predictors, assumption (b) is satisfied with $D = D_{\mathcal{Y}} D_{\mathcal{W}}$, as the hinge loss $\tilde{\ell}(\,\cdot\,, y)$ is $|y|$-Lipschitz. Also the Lipschitz constant in Theorem 9 reads $L = D$, as $|\ell(x_1, z) - \ell(x_2, z)| \le D_{\mathcal{Y}} |(x_1 - x_2)^\top w| \le D_{\mathcal{Y}} D_{\mathcal{W}} \|x_1 - x_2\|$ by the Cauchy-Schwarz's inequality.*

5. Generalisation and Optimisation Error Bounds

In this section we present the Generalisation and Optimisation Error bounds that yield the Test Error bounds presented within Section 4. Section 5.1 begins with the stability analysis used to derive the Generalisation Error bounds for smooth losses. This is followed by the Generalisation Error bound for non-smooth losses in Section 5.2. Finally, Section 5.3 presents Optimisation Error bounds for both classes of losses.

5.1. Generalisation Error Bound for Smooth Losses through Stability

To bound the Generalisation Error for smooth losses we utilise its link with stability. This has previously been investigated in Rogers and Wagner (1978); Kearns and Ron (1999); Bousquet and Elisseeff (2002); Mukherjee et al. (2006); Shalev-Shwartz et al. (2010), with Bousquet and Elisseeff (2002) and Hardt et al. (2016) providing the work upon which we rely. Specifically, Hardt et al. (2016) investigated the Generalisation Error of serial SGD in the multi-pass setting, giving, in the case of convex, Lipschitz, and smooth losses, upper bounds that grow linearly with the number of iterations and step size. The method used is algorithmic stability (or sensitivity) as introduced in Bousquet and Elisseeff (2002). This method investigates the deviation of an algorithm when a single data point in the data set $\mathcal{D}$ is resampled. By iterating through all of the observations, accounting for the deviation in each case, the Generalisation Error is then equal to the average deviation, as we see next. In our case the observations are spread throughout a graph, so the deviations of the algorithm depends on the location of the observation that is resampled.

For each $w \in V$ and $k \in [m]$, let $\widetilde{Z}_{w,k}$ be a resampled (independent) observation coming from the same data distribution. Let $\widetilde{X}(w,k)_v^t$ denote the output of Distributed SGD at node v after t iterations when the algorithm is run on the perturbed data set $\{\mathcal{D}\backslash Z_{w,k}\} \cup \widetilde{Z}_{w,k}$ in which the k-th observation for node w, i.e., $Z_{w,k}$, is replaced by $\widetilde{Z}_{w,k}$. The Generalisation Error is then equal to the average mean deviance of the loss function evaluated at the perturbed outputs.

Proposition 12 *For any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}[r(X_v^t) - R(X_v^t)] = \frac{1}{nm} \sum_{w \in V} \sum_{k=1}^{m} \mathbf{E}[\ell(X_v^t, \widetilde{Z}_{w,k}) - \ell(\widetilde{X}(w,k)_v^t, \widetilde{Z}_{w,k})].$$

Proof The proof is given in Appendix A.2. ■

The identity in Proposition 12 involves a double sum over the mean deviations of the algorithm applied to locally perturbed data sets: one sum relates to the graph location where the perturbation is supported (w), and the other sum relates to the index of the perturbed data point at that location (k). Each *individual* deviation depends on the graph topology via the location of the resampled observation w relative to the node of reference v. This dependence is captured by the bound that we give in Proposition 18 in Appendix A.3.2, where we show that the non-expansive property of the gradient descent update in the smooth case controls the spatial propagation of the deviation across the network via the term $\sum_{s=1}^{t-1}(P^s)_{vw}$. Proposition 12 involves the *average* across all deviations, and once the summation over $w \in V$ is considered, we get a final bound that increases linearly with time but does not depend on the graph topology, as we state next.

Lemma 13 (Generalisation Error bound for convex, Lipschitz, and smooth losses) *Assume that for any $z \in \mathcal{Z}$ the function $\ell(\,\cdot\,, z)$ is convex, L-Lipschitz, and β-smooth. Let $X_v^1 = 0$ for all $v \in V$. Then, Distributed SGD with $\eta\beta \leq 2$ yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}[r(X_v^t) - R(X_v^t)] \leq \frac{2\eta L^2}{nm}(t-1).$$

Proof See Appendix A.3. ■

For completeness, and to fully establish in the decentralised case the results derived in Hardt et al.

(2016) in the single machine case, we include in Appendix A.3 also the time-uniform Generalisation Error bound for the constrained and strongly-convex case. In this case, the *contraction* property of the gradient descent update controls the spatial propagation of the deviation across the network via the term $\sum_{s=1}^{t-1} \iota^s (P^s)_{vw}$, for a given $\iota < 1$. Once the summation over $w \in V$ in Proposition 12 is taken, we get a final bound that does not depend on time, nor on the graph topology. The bounds that we give are identical to the ones in Hardt et al. (2016) for a single agent with nm observations.

5.2. Generalisation Error Bound for Non-Smooth Losses through Rademacher Complexity

In the case of non-smooth losses we follow the approach used in Lin et al. (2016a) for the single-machine case that involves controlling the Generalisation Error by using standard Rademacher complexity's arguments through Assumption 1 *(b)* and bounding the norm of the iterates $\|X_v^t\|$ as a function of the parameters of the algorithm.

Lemma 14 (Generalisation Error for convex and Lipschitz losses) *Assume that for any $z \in \mathcal{Z}$ the loss function $\ell(\,\cdot\,, z)$ is convex and L-Lipschitz. Consider Assumptions 1. Let $X_v^1 = 0$ for all $v \in V$. Then, Distributed SGD yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}[r(X_v^t) - R(X_v^t)] \leq 2D\sqrt{\frac{(t-1)(\eta^2 L^2 + 2\eta(B-C))}{nm}}.$$

Proof See Appendix A.4. ■

We now go on to give Optimisation Error bounds which, once combined the Generalisation Error bounds in Section 5.1 and 5.2, give the Test Error bounds presented within Section 4.

5.3. Optimisation Error Bounds

In this section we present Optimisation Error bounds for Distributed SGD with convex, Lipschitz, and either smooth or non-smooth losses. These results follow from theorems proved within Appendix B under the more general setting of the first-order stochastic oracle model. We note that constants within these bounds have not been optimised.

The bounds that we derive are proved using the techniques developed in Nedić et al. (2009) and, in particular, in Duchi et al. (2012), where the evolution of the algorithm X_v^s is compared against the evolution of its network average $\overline{X}^s := \frac{1}{n}\sum_{v\in V} X_v^s$ to derive graph-dependent error bounds. Appendix B contains the full scheme of the proof, along with the error decomposition into a network term, an optimisation term, and a gradient noise term (only in the smooth case). As previously emphasised, the fact that we investigate implicit regularisation strategies allows us to deal with unconstrained problems, and in this case the evolution of the network-averaged process $\overline{X}^s$ admits a simple form that facilitates the analysis. This approach avoids the difficulties with the nonlinearity of projection that have been previously challenging in distributed learning models, and that motivated the investigation of dual methods such as in Duchi et al. (2012).

We start with the case of Lipschitz and smooth losses. The proof for this case is inspired from the proof for serial SGD applied to smooth objectives, specifically, Theorem 6.3 in Bubeck et al. (2015), itself extracted from Dekel et al. (2012). The bound that we present depends upon both the quantity σ and the quantity κ defined, respectively, in (2) and (3).

Lemma 15 (Optimisation Error bound for convex, Lipschitz, and smooth losses) *Assume that for any $z \in \mathcal{Z}$ the function $\ell(\,\cdot\,, z)$ is convex, L-Lipschitz, β-smooth and satisfies* (2) *and* (3). *Let $X^1_v = 0$ for all $v \in V$, $\|X^\star\| \leq G$. Then, Distributed SGD with $\eta = 1/(\beta + 1/\rho)$ and $\rho > 0$, yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}\Big[R\Big(\frac{1}{t}\sum_{s=1}^{t} X^{s+1}_v\Big) - R(X^\star)\Big]$$
$$\leq \frac{\rho}{2}\sigma^2 + \frac{(\beta+1/\rho)G^2}{2t} + \frac{3\kappa}{\beta+1/\rho}\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big(L + \frac{3}{2}\frac{\beta(3+\beta\rho)\kappa}{\beta+1/\rho}\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big).$$

Proof The result follows from Corollary 27 in Appendix B and from Section B.4. ■

Next is the Optimisation Error bound for non-smooth losses, inspired from Duchi et al. (2012).

Lemma 16 (Optimisation Error bound for convex and Lipschitz losses) *Assume that for any $z \in \mathcal{Z}$ the function $\ell(\cdot, z)$ is convex and L-Lipschitz. Let $X^1_v = 0$ for all $v \in V$, $\|X^\star\| \leq G$. Then, Distributed SGD yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}\Big[R\Big(\frac{1}{t}\sum_{s=1}^{t} X^{s}_v\Big) - R(X^\star)\Big] \leq \frac{\eta L^2}{2}\Big(19\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}\Big) + \frac{G^2}{2\eta t}.$$

Proof The result follows from Corollary 25 in Appendix B and from Section B.4. ■

When optimising either of these bounds with respect to ρ or η, a rate no better than $O(1/\sqrt{t})$ can be obtained, matching the rate of stochastic gradient descent in the single-machine case. From the bound in Lemma 15, however, we note that if $\sigma = \kappa = 0$ then the accelerated rate of $O(1/t)$ can be obtained, matching the rate of full-gradient descent in the single-machine case. For a general discussion on these lines, we refer to Appendix B and to Remark 22 in particular.

6. Numerical Experiments

In this section we provide a simulation study to investigate if the previously proven bounds can be representative of real behaviours. Specifically, we investigate the Test Error bounds given in Corollary 6 for convex, Lipschitz, and smooth losses. We start by introducing the notation and quantities of interest in Section 6.1, then we present the results of the experiments in Section 6.2.

6.1. Setup

As we want to minimise the expected risk $r(x) = \mathbf{E}\,\ell(x, Z)$ but a closed form expression is typically not available, we use a Monte Carlo approximation constructed from an independent out of sample data set $\{Z_j\}_{j\in[\widehat{N}]}$, namely, $\hat{r}(x) := \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \ell(x, Z_j)$. Given t iterations of the Distributed SGD algorithm, we denote the ergodic average of the iterates by $\widehat{X}^t_v := \frac{1}{t}\sum_{s=1}^{t} X^s_v$, for $v \in V$. We investigate the Out of Sample Risk defined as $\max_{v\in V} \hat{r}(\widehat{X}^t_v)$, which is set to be a proxy for the Test Risk for Distributed SGD, as defined in Section 2. We recall that the Test Error is defined as the expectation of the Test Risk minus the minimum expected risk $r(x^\star)$, which is a constant. Therefore, modulo a constant shift, Out of Sample Risk is also a proxy for the Test Error.

Given a graph (V, E) with $n = |V|$ nodes, let $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix defined as $A_{vw} := 1$ if $\{v, w\} \in E$ and $A_{vw} := 0$ otherwise. For each $v \in V$, let $d_v = \sum_{w \in V} A_{vw}$ denote the degree of node v, $d_{\text{max}} = \max_{v \in V} d_v$ the maximum degree, and $D = \text{diag}(d_1, \dots, d_n)$ the diagonal degree matrix. We consider the doubly stochastic matrix $P = I - \frac{1}{d_{\text{max}}+1}(D - A)$. This choice is standard in distributed optimisation (see Shah (2009), for instance). In this case, the spectral gap is known to be of the following orders (see Duchi et al. (2012), for instance):

$$O\left(\frac{1}{\sqrt{1-\sigma_2(P)}}\right) = \begin{cases} n & \text{Cycle} \\ \sqrt{n} & \text{Grid} \\ 1 & \text{Complete Graph} \end{cases}$$

We adopt the following parametrisation: $O(1/\sqrt{1-\sigma_2(P)}) = O(n^{\alpha})$, for $\alpha \geq 0$. These topologies are typical of those used in decentralised networks (Shah, 2009; Dimakis et al., 2010).

We consider an instance of logistic regression in supervised learning, where for a given positive integer d, we have $Z = \{W, Y\}$ with the feature vector $W \in \mathbb{R}^d$ and the label $Y \in \{-1, 1\}$, and the parameter of interest is $X \in \mathbb{R}^d$. The loss function in this case is given by

$$\ell(X, Z) = \log(1 + e^{-Y \times \langle X, W \rangle}),$$

where $\langle X, W \rangle = X^\top W = \sum_{i=1}^{d} X_i W_i$. Given the node count n and m locally-owned data points, a simulated data set with a total of $N = mn$ observations $\{Z_i\}_{i \in [N]}$ are sampled following the experiment within Duchi et al. (2012). Specifically, a true parameter $X^{\star\star}$ is sampled from a standard d-dimensional Gaussian $\mathcal{N}(0, I)$, the feature vectors W_i's are sampled uniformly from the unit sphere $\{w \in \mathbb{R}^d : \|w\| \leq 1\}$, and the responses are set as $Y_i = \text{sign}(\langle W_i, X^{\star\star} \rangle)$ where $\text{sign}(a) = 1$ if $a \geq 0$ and -1 if $a < 0$. The data set is then randomly spread across the graph with each node getting m samples. It can easily be seen that the Lipschitz parameter is $L = 1$ and the smoothness parameter is $\beta = 1/4$. Parameters depending upon the gradient noise were upper bounded by distribution-independent quantities and set to $\sigma^2 \to 4L^2$ and $\kappa \to L$ (see Proposition 23 in Appendix B for the interplay between L and κ as far as bounding the network term is concerned). A solution $X^\star$ to the empirical risk minimisation rule is calculated with tolerance 10^{-15} using the `lbfgs` solver within the `LogisticRegression` function of the python library `scikit` (Pedregosa et al., 2011). We set $G = \|X^\star\|$. Dimension and Monte Carlo estimate size are set to $d = 100$ and $\widehat{N} = 1000$, respectively. We investigate the performance of Distributed SGD in two sample size regimes $m = 2$ and $m = 100$, which we now go on to describe in more detail.

6.2. Experimental Results - Small Sample Regime

This setting explores the small sample size regime where by agent receive $m = 2$ samples each. Distributed SGD is run for 15 different time horizons t, between 10^2 and either 10^7 or $10^{6.5}$ for graph sizes $n = 3^2$ or $n = 10^2$, respectively. All runs are initialised from $X_v^1 = 0$ for all $v \in V$. Comparisons are made for three choices of the step size, as prescribed in Corollary 6, and for three choices of the graph topology: complete graph ($\alpha = 0$), grid ($\alpha = 1/2$), and cycle ($\alpha = 1$). Referring to the *upper* bounds in Corollary 6, we outline what we expect to see plotting the Test Error against the time horizon t, with $\log - \log$ scales, across the three different step sizes:

- $\rho^\star$ - For small t, linear decrease with graph-dependent intercept; for large t, linear increase with intercept independent of the graph topology. Minimum attained is graph-dependent;

- $\rho^\star_{\text{Opt}}$ - For small and large t, respectively, linear decrease and increase with graph-dependent intercept. Minimum attained is independent of graph topology;
- $\rho^\star_{\text{Test}}$ - Linear decrease with graph-dependent intercept up to a threshold independent of the graph topology.

Figure 1 presents $\log - \log$ plots of the Out of Sample Risk against the time horizon t, using the step sizes stated in Corollary 6. All of the behaviours described above, as suggested by our upper bounds, are observed. In particular, recall that our bounds suggest the sub-optimality of the sample rate achieved by the step size aligned with serial SGD ($\rho^\star$), as opposed to the other two choices ($\rho^\star_{\text{Opt}}$ and $\rho^\star_{\text{Test}}$) that depend on the graph topology. Corollary 6 states that the Test Error for $\rho^\star$ yields the rate $\widetilde{O}(n^\alpha/\sqrt{nm})$, as opposed to the rate $\widetilde{O}(1/\sqrt{nm})$ achieved by the other two choices. The former rate is worse (i.e., larger) than the latter for the cycle ($\alpha = 1$) and the grid ($\alpha = 1/2$), while it is of the same order for the complete graph ($\alpha = 0$). Evidence of this behaviour is observed in Figure 1 for $n = 100$, where the Out of Sample Risk related to the cycle and grid is seen to achieve a lower minimum when the step sizes that account for the graph topology are used.

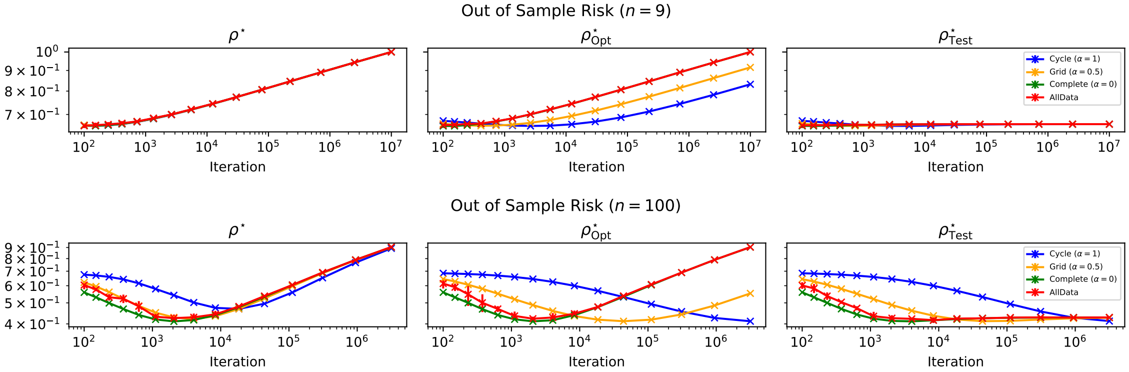

Figure 1: Out of Sample Risk against time horizon for different choices of step size: $\rho^\star$, $\rho^\star_{\text{Opt}}$, and $\rho^\star_{\text{Test}}$. Scales are $\log - \log$. Graph size $n = 9$ (*top*), 100 (*bottom*). Simulations run for 15 values of t from 10^2 to 10^7 (*top*) or $10^{6.5}$ (*bottom*). Each point is an average over 10 (*top*) or 4 (*bottom*) replications with error bars representing 2 standard deviations before taking the log scale (error bars are not visible for large t due to the small variance between repeated runs). *AllData*: serial SGD run on the full data set of 18 (*top*) or 200 (*bottom*) samples with $\rho^\star = \rho^\star_{\text{Opt}} = O(1/\sqrt{t})$ and $\rho^\star_{\text{Test}} = O(1/(\sqrt{t}\sqrt{1 + t/m})$. The behaviour of serial SGD is seen to correspond to the behaviour of Distributed SGD on the complete graph, as expected.

6.3. Experimental Results - Large Sample Regime

In this section a larger sample regime ($n = 100$, $m = 25$) is investigated. Due to the number of iterations scaling with the total number of data points i.e. stopping time being of the order $t \sim nm/(1 - \sigma_2(P))$, following Remark 8, a fixed step size is used to save running multiple instances of Distributed SGD and save on computational cost. Specifically, the two fixed step size choices considered are: $\rho^\star_{\text{Const}} = O(1/\sqrt{nm})$, to align with serial single-machine SGD; and $\rho^\star_{\text{ConstNet}} = O((1 - \sigma_2(P))/\sqrt{nm})$, the step size suggested by Theorem 5 Remark 8 that adjusts for the network topology. Furthermore, the true underlying optimal parameter $X^{\star\star}$ has its first $\sqrt{d}$ co-ordinates fixed to zero in order to simulate an over parameterised setting. The resulting Out of Sample Risks have been presented within Figure 2.

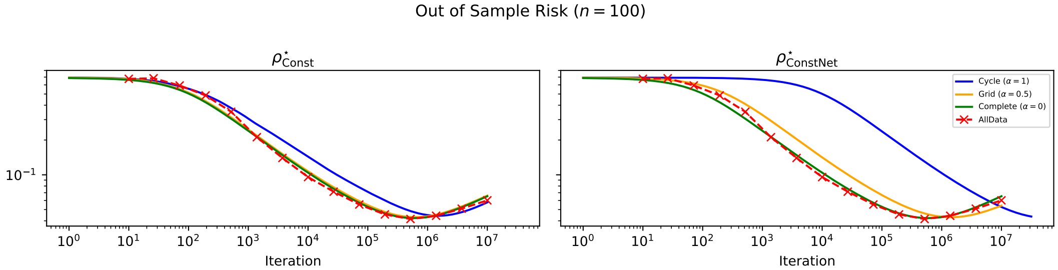

Figure 2: Out of Sample Risk for Distributed SGD with step sizes $\rho^{\star}_{\text{Const}}$ (*Left*) and $\rho^{\star}_{\text{ConstNet}}$ (*Right*) for graph topologies Cycle, Grid and Complete. Each run for 10^7 iterations, while Distributed SGD on cycle topology with $\rho^{\star}_{\text{ConstNet}}$ run for $10^{7.5}$ iterations. Quantity plotted is for a single instance of Distributed SGD. *AllData*: single-machine serial SGD run for 15 different iterations t between 10 and 10^7 with decreasing step size $\rho = O(1/\sqrt{t})$. Both x-axis and y-axis are logarithmic scales.

Firstly, observe that the minimum Out of Sample Risk achieved by Distributed SGD with $\rho^{\star}_{\text{ConstNet}}$ matches the minimum achieved by Serial Single-Machine SGD with decreasing step size (Dashed Red line with markers). Secondly, aligning with the small sample regime in Section 6.2, the minimum out of sample risk (0.0442) for a cycle topology with a constant single-machine serial step size $\rho^{\star}_{\text{Const}}$ is higher than the minimum out of sample risk (0.0436) attained with the constant step size adjusted for the network topology $\rho^{\star}_{\text{ConstNet}}$. We note the simulation for the cycle topology with $\rho^{\star}_{\text{ConstNet}}$ were stopped early at $10^{7.5}$ iterations due to computational cost.

7. Conclusion

We have proposed and investigated graph-dependent implicit regularisation strategies for synchronised Distributed SGD for convex problems in multi-agent learning. Specifically, we have shown how Distributed SGD can retain single-machine serial statistical guarantees by proper tuning of the algorithmic parameters as a function of the graph topology. For convex, Lipschitz, and smooth losses, we showed that Distributed SGD recovers, up to logarithmic terms, the optimal rate of $O(1/\sqrt{nm})$ for single-pass constrained serial SGD (Lan, 2012; Xiao, 2010). For convex and Lipschitz losses, we showed that Distributed SGD recovers, up to logarithmic terms, the best-known rate of $O(1/(nm)^{1/3})$ for single-machine serial SGD with implicit regularisation (Lin et al., 2016b). To obtain these results we: proved Generalisation Error bounds that do not depend on the graph topology and match the bounds in the single-machine serial setting; and derived Optimisation Error bounds that depend on the graph topology. We provided numerical simulations showing that our bounds can be representative of real behaviours.

Our work motivates further investigation of graph-dependent implicit regularisation strategies for decentralised protocols. Since synchronisation and communication are often a dominant bottleneck in distributed computations, further research is needed to investigate the improvement on the communicational and computational complexity that can be obtained by exploiting the interplay between the statistical regularities of the local objective functions and schemes involving minibatching, acceleration, and graph sparsification. The latter relates to Gossip protocols where only a random subset of nodes communicate at each iteration (Dimakis et al., 2010). Another direction for future investigation lies in the analysis of adaptive schemes that can contemplate time-dependent step sizes and that can automatically infer the algorithmic parameters of interests, in primis the spectral gap of the communication matrix.

Appendix A. Proofs of Generalisation and Test Error Bounds

This appendix provides the proofs for the Generalisation and Test Error bounds presented within the main body of this paper. First, for completeness, we include the proofs of Proposition 1 and Proposition 12 in Section A.1 and Section A.2, respectively. These results generalise to the distributed setting the Test Error decomposition and the Generalisation Error decomposition via stability used in the single-machine setting, and the proofs follow the exact same arguments as in the single-machine case. Second, we present the proofs of the Generalisation Error bounds for smooth and non-smooth losses in Section A.3 and Section A.4, respectively. For completeness, Section A.3 also includes the proof of stability for the strongly convex case with constraints, which is not covered in the main body but is here presented as it fully generalises the results in Hardt et al. (2016) for Distributed SGD. Third, in Section A.5 we present the proofs of Test Error bounds for smooth and non-smooth losses, referring to Theorem 5 and Theorem 9 within the main body of the work. Finally, in Section A.6 and Section A.7 we give the calculations deriving the rates presented in Corollary 6 and Corollary 10 for smooth and non-smooth losses, respectively. Throughout, we use the notations $\lesssim, \simeq, \gtrsim$, to indicate $\leq, =, \geq$ modulo constants and $\log$ terms.

A.1. Proof of Proposition 1

The proof is analogous to the one given in Hardt et al. (2016) for the single-machine case.

Proof [Proposition 1] We have $r(X_v^t)-r(x^\star) = r(X_v^t)-R(X_v^t)+R(X_v^t)-R(X^\star)+R(X^\star)-r(x^\star)$. Note that $\mathbf{E}R(X^\star) \leq r(x^\star)$, as for any x we have $R(X^\star) \leq R(x)$ so that $\mathbf{E}R(X^\star) \leq \mathbf{E}R(x) = r(x)$, which holds for $x = x^\star$. Thus, $\mathbf{E}\, r(X_v^t)-r(x^\star) \leq \mathbf{E}[r(X_v^t)-R(X_v^t)]+\mathbf{E}[R(X_v^t)-R(X^\star)]$. ■

A.2. Proof of Proposition 12

The proof follows the ideas in Bousquet and Elisseeff (2002) and Hardt et al. (2016) for the single-machine case.

Proof [Proposition 12] As the resampled observation $\widetilde{Z}_{w,k}$ has the same distribution than Z and is independent of both X_v^t and $\mathcal{D}$, we have $\mathbf{E}\, r(X_v^t) = \frac{1}{nm}\sum_{w\in V}\sum_{k=1}^m \mathbf{E}\,\ell(X_v^t, \widetilde{Z}_{w,k})$. As the pair $(X_v^t, Z_{w,k})$ has the same distribution as the pair $(\widetilde{X}(w,k)_v^t, \widetilde{Z}_{w,k})$, the expectation of the empirical risk can be written as $\mathbf{E}R(X_v^t) = \frac{1}{nm}\sum_{w\in V}\sum_{k=1}^m \mathbf{E}\,\ell(\widetilde{X}(w,k)_v^t, \widetilde{Z}_{w,k})$. Thus, $\mathbf{E}[r(X_v^t) - R(X_v^t)] = \frac{1}{nm}\sum_{w\in V}\sum_{k=1}^m \mathbf{E}[\ell(X_v^t, \widetilde{Z}_{w,k}) - \ell(\widetilde{X}(w,k)_v^t, \widetilde{Z}_{w,k})]$. ■

A.3. Proof of Generalisation Error Bounds for Smooth Losses

In this section we prove the Generalisation Error bound presented in Lemma 13 for smooth losses, and we establish a Generalisation Error bound for strongly convex functions. The proof that we present follows the spirit of the proof in Hardt et al. (2016) for the single-machine setting, using algorithmic stability. Specifically, deviations of the algorithm are studied when a single data point in the entire data set is resampled. In the distributed setting that we consider, the training data is spread throughout the communication graph, and we need to consider stability not only with respect to time (i.e., the iteration time of the algorithm), but also with respect to space (i.e., the communication graph). As established in Proposition 12, the Generalisation Error is the average of these deviations. Intermediate steps show that the individual deviations have a dependence on the graph topology, as encoded by the communication matrix P. However, once the average over all deviations is taken, we get results that do not depend on the graph topology.

First, in Section A.3.1 we describe the setup for the stability analysis. Then, in Section A.3.2 we present the proof for the case of convex, Lipschitz, and smooth losses. Finally, in Section A.3.3 we present the case of Lipschitz, smooth, and strongly-convex losses with constraints.

A.3.1. SETUP

For any $w \in V$ and $k \in [m]$, let $\widetilde{\mathcal{D}}(w,k) := \{\mathcal{D}\backslash Z_{w,k}\} \cup \widetilde{Z}_{w,k}$ be the data set in which node w has the k-th observation resampled. Recall that $\widetilde{X}(w,k)^t_v$ denotes the output at node v and time step t of Distributed SGD (1) run with respect to the data set $\widetilde{\mathcal{D}}(w,k)$. From Proposition 12, the link between the Generalisation Error and the ℓ_2 deviation

$$\delta(w,k)^t_v := \|\widetilde{X}(w,k)^t_v - X^t_v\|$$

can be made explicit when the loss function ℓ is L-Lipschitz in the first coordinate (uniformly in the second). Specifically, each term in the double sum $\sum_{k=1}^m \sum_{w\in V}$ in Proposition 12 is bounded by

$$\ell(X^t_v, \widetilde{Z}_{w,k}) - \ell(\widetilde{X}(w,k)^t_v, \widetilde{Z}_{w,k}) \leq L\delta(w,k)^t_v.$$

The results that we derive directly bound the deviation $\delta(w,k)^t_v$. Henceforth, for a given matrix $M \in \mathbb{R}^{n\times n}$ we use the notation M^s_{vw} to represent the quantity $(M^s)_{vw}$, where M^s is the s-th power of M, and the notation M_v to represent the v-th row of M. Hence, for a given vector x, we write $M_v x$ to indicate $\sum_{w\in V} M_{vw}x_w$. For any $x,y \in \mathbb{R}^d$, we let $\langle x,y\rangle = x^\top y = \sum_{i=1}^d x_i y_i$.

Before proceeding to the main proofs we require some standard results relating to the expansive properties of gradient descent updates with smooth and either convex or strongly convex functions. Specifically, for a sufficiently small step size, a result showing that gradient descent updates with smooth and convex function are non-expansive. Meanwhile, for additionally strongly-convex functions, a result showing that gradient descent updates are contractive. The proof can be found in Appendix A of Hardt et al. (2016) and it utilises the co-coercivity of gradients for smooth and convex functions (Nesterov, 2013).

Lemma 17 *Let f be a β-smooth function, convex, and $\eta\beta \leq 2$ with $\eta > 0$. Then, for any $x, y \in \mathbb{R}$,*

$$\|x - y - \eta(\nabla f(x) - \nabla f(y))\| \leq \|x-y\|.$$

Let f be a β-smooth function, γ-strongly convex, and $\eta \leq 2/(\beta+\gamma)$. Then, for any $x, y \in \mathbb{R}$,

$$\|x - y - \eta(\nabla f(x) - \nabla f(y))\| \leq \Big(1 - \frac{\eta\beta\gamma}{\beta+\gamma}\Big)\|x-y\|.$$

A.3.2. CONVEX, LIPSCHITZ, AND SMOOTH LOSSES

We start by stating Proposition 18 that establishes a bound on the deviation $\delta(w,k)^t_v$ that explicitly depends on the graph topology. This is followed by the proof of Lemma 13.

Proposition 18 (Stability for convex, Lipschitz, and smooth losses) *Assume the setting of Lemma 13. Then, for any $v, w \in V, k \in [m]$ and $t \geq 1$,*

$$\mathbf{E}\,\delta(w,k)^t_v = \mathbf{E}\|\widetilde{X}(w,k)^t_v - X^t_v\| \leq \frac{2\eta L}{m}\sum_{s=1}^{t-1} P^s_{vw}.$$

Proof [Proposition 18] Let $\mathcal{F}_1$ be the σ-algebra generated by $\mathcal{D}$ and $\widetilde{\mathcal{D}} := \{\widetilde{\mathcal{D}}(w,k)\}_{w\in V, k\in[m]}$. For any $t \geq 2$, let $\mathcal{F}_t$ be the σ-algebra generated by the data sets $\mathcal{D}$ and $\widetilde{\mathcal{D}}$, and by the collection of uniform random variables $\{K^2_v, \dots, K^t_v\}_{v\in V}$. Plugging the algorithm updates (1) into $\delta(w,k)^t_v$, applying the triangle inequality and using the fact that $\{X^{t-1}_v\}_{v\in V}$, $\{\widetilde{X}(w,k)^{t-1}_v\}_{v\in V}$, $\mathcal{D}$, and $\widetilde{\mathcal{D}}$ are measurable with respect to $\mathcal{F}_{t-1}$, we get

$$\begin{aligned}
&\mathbf{E}[\delta(w,k)^t_v|\mathcal{F}_{t-1}] \\
&\leq \sum_{l\neq w} P_{vl}\mathbf{E}\Big[\Big\|\widetilde{X}(w,k)^{t-1}_l - X^{t-1}_l - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_l, Z_{l,K^t_l}) - \nabla\ell(X^{t-1}_l, Z_{l,K^t_l})\Big)\Big\|\Big|\mathcal{F}_{t-1}\Big] \quad (4)\\
&\quad + \frac{P_{vw}}{m}\sum_{i\neq k}\Big\|\widetilde{X}(w,k)^{t-1}_w - X^{t-1}_w - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_w, Z_{w,i}) - \nabla\ell(X^{t-1}_w, Z_{w,i})\Big)\Big\| \quad (5)\\
&\quad + \frac{P_{vw}}{m}\Big\|\widetilde{X}(w,k)^{t-1}_w - X^{t-1}_w - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_w, \widetilde{Z}_{w,k}) - \nabla\ell(X^{t-1}_w, Z_{w,k})\Big)\Big\|. \quad (6)
\end{aligned}$$

The above decomposition is in three parts: (4), the terms aligning with agents who do not have a resample datapoint $\forall \ell, \ell \neq w$; (5), the terms at w conditioned on not picking the resample datapoint; and (6), the term at w when picking the resample datapoint. In particular (6) is the only one to involve the difference of two gradients evaluated at different data points ($\widetilde{Z}_{w,k}$ and $Z_{w,k}$). To bound this term, we use the Lipschitz property, $\|\nabla\ell(\,\cdot\,,z)\| \leq L$ for all $z \in \mathcal{Z}$, and get

$$(6) \leq \Big(\delta(w,k)^{t-1}_w + 2\eta L\Big)\frac{P_{vw}}{m}.$$

To bound terms (4) and (5), we use the non-expansive property of the gradient updates arising from the convexity and smoothness of $\ell(\,\cdot\,,z)$, specifically, the inequality $\|x - y - \eta(\nabla\ell(x,z) - \nabla\ell(y,z))\| \leq \|x - y\|$ for $x, y \in \mathbb{R}^d, z \in \mathcal{Z}$ in Lemma 17. In particular we have

$$\begin{aligned}
(4) &\leq \sum_{\ell\neq w} P_{v\ell}\delta(w,k)^{t-1}_\ell \\
(5) &\leq \frac{P_{vw}}{m}\sum_{i\neq k}\delta(w,k)^{t-1}_w = \frac{P_{vw}}{m}(m-1)\delta(w,k)^{t-1}_w.
\end{aligned}$$

This yields

$$\begin{aligned}
\mathbf{E}[\delta(w,k)^t_v|\mathcal{F}_{t-1}] &\leq \sum_{l\neq w} P_{vl}\delta(w,k)^{t-1}_l + \Big(1 - \frac{1}{m}\Big)P_{vw}\delta(w,k)^{t-1}_w + \Big(\delta(w,k)^{t-1}_w + 2\eta L\Big)\frac{P_{vw}}{m} \\
&= \sum_{l\in V} P_{vl}\delta(w,k)^{t-1}_l + \frac{2\eta L}{m}P_{vw}.
\end{aligned}$$

Let $e_v \in \mathbb{R}^n$ be the vector with 1 in the coordinate aligning with node v and 0 everywhere else. Recursively applying the bound above in vector form with $\delta(w,k)^t = \{\delta(w,k)^t_v\}_{v\in V} \in \mathbb{R}^n$ yields (the inequality is meant coordinate-wise)

$$\mathbf{E}\,\delta(w,k)^t = \mathbf{E}[\mathbf{E}[\delta(w,k)^t|\mathcal{F}_{t-1}]] \leq P\,\mathbf{E}\,\delta(w,k)^{t-1} + \frac{2\eta L}{m} P e_w \leq \frac{2\eta L}{m}\sum_{s=1}^{t-1} P^s e_w,$$

where we used $\delta(w,k)^1_l = \|\widetilde{X}(w,k)^1_l - X^1_l\| = 0$ for all $l \in V$. Recall $(P^s e_w)_v = P^s_{vw}$. ■

The bound in Proposition 18 shows that the expected deviation between the algorithms remains zero until the number of iterations exceeds the natural distance in the graph between node v and node w. This bound naturally reflects the graph topology and captures the propagation of the deviation due to resampling a data point in a specific location of the graph. When combined with the summation over $w \in V$ in Proposition 12, this bound yields a Generalisation Error bound that does not depend on the graph topology: Lemma 13.

Proof [Lemma 13] Plugging the bound from Proposition 18 into the identity from Proposition 12, using that the rows of the matrix P sum to 1, we get

$$\mathbf{E}[r(X^t_v) - R(X^t_v)] \leq \frac{L}{nm}\sum_{w\in V}\sum_{k=1}^{m}\mathbf{E}\,\delta(w,k)^t_v \leq \frac{2\eta L^2}{nm}\sum_{s=1}^{t-1}\sum_{w\in V} P^s_{vw} = \frac{2\eta L^2}{nm}(t-1).$$

■

A.3.3. STRONGLY CONVEX, LIPSCHITZ, AND SMOOTH LOSSES

This section presents a Generalisation Error bound for Distributed SGD when the loss function is strongly convex, smooth, and Lipschitz continuous, generalising the results in Hardt et al. (2016) to the distributed setting. Recall that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is γ-strongly convex, with $\gamma > 0$, if $f(x) - f(y) \geq \nabla f(y)^\top (x-y) + \gamma\|x-y\|^2/2$ for all $x, y \in \mathbb{R}^d$. As strongly convex functions have unbounded gradients on $\mathbb{R}^d$, we consider the setting where parameters are constrained to be on a compact convex set $\mathcal{X} \subset \mathbb{R}^d$. Let $x \to \Pi(x) = \arg\min_{y\in\mathcal{X}} \|x-y\|$ be the Euclidean projection on $\mathcal{X}$. Then, iteration (1) becomes, for $s \geq 1$,

$$X^{s+1}_v = \Pi\Big(\sum_{w\in V} P_{vw}(X^s_w - \eta\nabla\ell(X^s_w, Z_{w,K^{s+1}_w}))\Big). \tag{7}$$

We refer to this variant as Distributed Projected SGD.

To motivate these assumptions, consider the specific case of Tikhonov regularisation, as done in Hardt et al. (2016). If the loss function ℓ is convex, β-smooth, and L-Lipschitz, then the penalised loss function $x \to \ell(x,z) + \frac{\gamma}{2}\|x\|^2$ is γ-strongly convex, $(\beta+\gamma)$-smooth, and $(L+\gamma r)$-Lipschitz when the constraint set is contained in a ball of radius r, i.e., $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\| \leq r\}$. The next result is the analogue of Lemma 13 with the additional assumption of strong convexity.

Lemma 19 (Generalisation Error bound for strongly-convex, Lipschitz, and smooth losses) *Assume that for any $z \in \mathcal{Z}$ the function $\ell(\,\cdot\,, z)$ is γ-strongly convex, L-Lipschitz, and β-smooth. Let $X^1_v = 0$ for all $v \in V$. Then, Distributed Projected SGD run on a compact, convex set $\mathcal{X}$ with $\eta \leq 2/(\beta+\gamma)$ yields, for any $v \in V$ and $t \geq 1$,*

$$\mathbf{E}[r(X^t_v) - R(X^t_v)] \leq \frac{2L^2}{mn}\frac{\beta+\gamma}{\beta\gamma}.$$

Observe that, for a sufficiently small step size $\eta \leq 2/(\beta+\gamma)$, the bound obtained is independent of the step size η and number of iterations t. As for the convex and smooth case of Lemma 13, also this bound aligns with the one given in Hardt et al. (2016) for a single agent with nm observations.

The next result is the analogue of Proposition 18.

Proposition 20 (Stability for strongly-convex, Lipschitz, and smooth losses) *Assume the setting of Lemma 19. Then, for any $v, w \in V$, $k \in [m]$ and $t \geq 1$,*

$$\mathbf{E}\,\delta(w,k)^t_v = \mathbf{E}\,\|\widetilde{X}(w,k)^t_v - X^t_v\| \leq \frac{2\eta L}{m}\sum_{s=1}^{t-1}\Big(1 - \frac{\eta\beta\gamma}{\beta+\gamma}\Big)^{s-1} P^s_{vw}.$$

Proof [Proposition 20]

The proof follows the same outline for the proof of Proposition 18. Consider the same setup and notation there defined. Using the non-expansive property of the Euclidean projection, the triangle inequality, and the fact that $\{X^{t-1}_v\}_{v\in V}$, $\{\widetilde{X}(w,k)^{t-1}_v\}_{v\in V}$, $\mathcal{D}$, and $\widetilde{\mathcal{D}}$ are measurable with respect to $\mathcal{F}_{t-1}$, we get

$$\begin{aligned}
&\mathbf{E}[\delta(w,k)^t_v|\mathcal{F}_{t-1}] \leq \mathbf{E}\|\widetilde{X}(w,k)^t_v - X^t_v\|\\
&\leq \sum_{l\neq w} P_{vl}\mathbf{E}\Big[\Big\|\widetilde{X}(w,k)^{t-1}_l - X^{t-1}_l - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_l, Z_{l,K^t_l}) - \nabla\ell(X^{t-1}_l, Z_{l,K^t_l})\Big)\Big\|\Big|\mathcal{F}_{t-1}\Big] \qquad (8)\\
&\quad + \frac{P_{vw}}{m}\sum_{i\neq k}\Big\|\widetilde{X}(w,k)^{t-1}_w - X^{t-1}_w - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_w, Z_{w,i}) - \nabla\ell(X^{t-1}_w, Z_{w,i})\Big)\Big\| \qquad (9)\\
&\quad + \frac{P_{vw}}{m}\Big\|\widetilde{X}(w,k)^{t-1}_w - X^{t-1}_w - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_w, \widetilde{Z}_{w,k}) - \nabla\ell(X^{t-1}_w, Z_{w,k})\Big)\Big\|. \qquad (10)
\end{aligned}$$

Term (10) is the only one to involve the difference of two gradients evaluated at different data points ($\widetilde{Z}_{w,k}$ and $Z_{w,k}$). To use the contraction property arising from strong convexity, add and subtract the term $\eta\nabla\ell(\widetilde{X}(w,k)^{t-1}_w, Z_{w,k})$ inside the norm, and use the Lipschitz property to get

$$(10) \leq \frac{P_{vw}}{m}\Big\|\widetilde{X}(w,k)^{t-1}_w - X^{t-1}_w - \eta\Big(\nabla\ell(\widetilde{X}(w,k)^{t-1}_w, Z_{w,k}) - \nabla\ell(X^{t-1}_w, Z_{w,k})\Big)\Big\| + \frac{2\eta L}{m}P_{vw}.$$

To bound terms (8) and (9), as well as the bound above for (10), we use the contraction property of the gradient updates from Lemma 17, specifically, the inequality $\|x - y - \eta(\nabla\ell(x,z) - \nabla\ell(y,z))\| \leq$

$(1-\frac{\eta\beta\gamma}{\beta+\gamma})\|x-y\|$ for $x, y \in \mathbb{R}^d$, $z \in \mathcal{Z}$, and $\eta \leq \frac{2}{\beta+\gamma}$. In particular,

$$
\begin{aligned}
(8) &\leq \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)\sum_{\ell\neq w} P_{v\ell}\delta(w,k)^{t-1}_{\ell} \\
(9) &\leq \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)\frac{P_{vw}}{m}\sum_{i\neq k}\delta(w,k)^{t-1}_{w} = \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)\frac{P_{vw}}{m}(m-1)\delta(w,k)^{t-1}_{w} \\
(10) &\leq \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)\frac{P_{vw}}{m}\delta(w,k)^{t-1}_{w} + \frac{2\eta L}{m}P_{vw}
\end{aligned}
$$

This yields

$$
\begin{aligned}
&\mathbf{E}[\delta(w,k)^t_v|\mathcal{F}_{t-1}] \\
&\leq \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)\Bigg[\sum_{l\neq w} P_{vl}\delta(w,k)^{t-1}_{l} + \Big(1-\frac{1}{m}\Big)P_{vw}\delta(w,k)^{t-1}_{w} + \frac{1}{m}P_{vw}\delta(w,k)^{t-1}_{w}\Bigg] + \frac{2\eta L}{m}P_{vw} \\
&= \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)\sum_{l\in V} P_{vl}\delta(w,k)^{t-1}_{l} + \frac{2\eta L}{m}P_{vw}.
\end{aligned}
$$

In vector notation, the above reads

$$
\mathbf{E}\,\delta(w,k)^t \leq \Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)P\,\mathbf{E}\,\delta(w,k)^{t-1} + \frac{2\eta L}{m}Pe_w \leq \frac{2\eta L}{m}\sum_{s=1}^{t-1}\Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)^{s-1}P^s e_w
$$

where we used $\delta(w,k)^1_l = \|\widetilde{X}(w,k)^1_l - X^1_l\| = 0$ for all $l \in V$ and recursively applied the above bound to $\mathbf{E}[\delta(w,k)^t]$. ■

With Proposition 20 in hand, we prove Lemma 19.

Proof [Lemma 19] Plugging the bound from Proposition 20 into the identity from Proposition 12, using that the rows of the matrix P sum to 1, we get

$$
\mathbf{E}[r(X^t_v) - R(X^t_v)] \leq \frac{L}{nm}\sum_{w\in V}\sum_{k=1}^{m}\mathbf{E}\,\delta(w,k)^t_v \leq \frac{2\eta L^2}{mn}\sum_{s=1}^{t-1}\Big(1-\frac{\eta\beta\gamma}{\beta+\gamma}\Big)^{s-1},
$$

and the proof is concluded by summing the geometric projection for t going to infinity, using that the assumption $\eta \leq \frac{2}{\beta+\gamma}$ implies that $\frac{\eta\beta\gamma}{\beta+\gamma} < 1$. ■

A.4. Proof of Generalisation Error Bound for Non-Smooth Losses

This section presents Generalisation Error bounds for Distributed SGD when losses are assumed to be non-smooth, aligning with Lemma 14 within the main body of the text. In this case we follow the approach in (Lin et al., 2016a, Appendix B) that involves controlling the Generalisation Error by using standard Rademacher complexity's arguments through Assumption 1 *(b)* and bounding the norm of the iterates through Assumption 1 *(a)*. We start by presenting Lemma 21 which bounds the iterates produced by the Distributed SGD. This is followed by the proof for the Generalisation Error bound for non-smooth losses Lemma 14.

Lemma 21 *Assume there exist $C \leq B$ such that for each $z \in \mathcal{Z}$ the function $\ell(\,\cdot\,, z)$ is convex, L-Lipschitz, bounded above at zero $\ell(0, z) \leq B$, and bound uniformly from below $\ell(x, z) \geq C$ for $x \in \mathbb{R}^d$. Let $X_v^1 = 0$ for all $v \in V$. Then, Distributed SGD yields, for any $v \in V$ and $t \geq 1$,*

$$\|X_v^t\| \leq \sqrt{(t-1)(\eta^2 L^2 + 2\eta(B-C))}.$$

Proof Let $x \in \mathbb{R}^d$. By the Distributed SGD update (1) we get

$$\|X_v^t - x\| \leq \sum_{w \in V} P_{vw} \|X_w^{t-1} - \eta \partial \ell(X_w^{t-1}, Z_{w,K_w^t}) - x\|. \tag{11}$$

The convexity of $\ell(\,\cdot\,, z)$ yields

$$\langle \partial \ell(X_w^{t-1}, Z_{w,K_w^t}), x - X_w^{t-1} \rangle \leq \ell(x, Z_{w,K_w^t}) - \ell(X_w^{t-1}, Z_{w,K_w^t}),$$

and the Lipschitz continuity of $\ell(\,\cdot\,, z)$ yields $\|\partial \ell(X_w^{t-1}, Z_{w,K_w^t})\| \leq L$. Thus,

$$\begin{aligned}
&\|X_w^{t-1} - \eta \partial \ell(X_w^{t-1}, Z_{w,K_w^t}) - x\|^2 \\
&= \|X_w^{t-1} - x\|^2 + \eta^2 \|\partial \ell(X_w^{t-1}, Z_{w,K_w^t})\|^2 + 2\eta \langle \partial \ell(X_w^{t-1}, Z_{w,K_w^t}), x - X_w^{t-1} \rangle \\
&\leq \|X_w^{t-1} - x\|^2 + \eta^2 L^2 + 2\eta(\ell(x, Z_{w,K_w^t}) - \ell(X_w^{t-1}, Z_{w,K_w^t})).
\end{aligned}$$

Setting $x = 0$, using that $\ell(X_w^{t-1}, Z_{w,K_w^t}) \geq C$ as well as the assumption $\ell(0, Z_{w,K_w^t}) \leq B$, we get

$$\|X_w^{t-1} - \eta \partial \ell(X_w^{t-1}, Z_{w,K_w^t})\|^2 \leq \|X_w^{t-1}\|^2 + \eta^2 L^2 + 2\eta(B-C).$$

Using that the matrix P is doubly stochastic, the bound (11) yields the recursion

$$\max_{v \in V} \|X_v^t\|^2 \leq \max_{w \in V} \|X_w^{t-1} - \eta \partial \ell(X_w^{t-1}, Z_{w,K_w^t})\|^2 \leq \max_{v \in V} \|X_v^{t-1}\|^2 + \eta^2 L^2 + 2\eta(B-C),$$

so recursively applying the above bound and taking square root gives

$$\|X_v^t\| \leq \max_{v \in V} \|X_v^t\| \leq \sqrt{(t-1)(\eta^2 L^2 + 2\eta(B-C))}.$$

■

Proof [Lemma 14] Standard Rademacher complexity's arguments utilising the symmetrisation technique and Assumption 1 *(b)* yield, for any $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$,

$$\mathbf{E} \sup_{x \in \widetilde{\mathcal{X}}} (r(x) - R(x)) \leq 2\mathbf{E} \sup_{x \in \widetilde{\mathcal{X}}} \frac{1}{nm} \sum_{i=1}^{nm} \sigma_i \ell(x, z_i) \leq 2D \frac{\sup_{x \in \widetilde{\mathcal{X}}} \|x\|}{\sqrt{nm}}.$$

By Lemma 21 we know that the iterates are contained in the ball $\widetilde{\mathcal{X}} = \{x \in \mathbb{R}^d : \|x\| \leq \sqrt{(t-1)(\eta^2 L^2 + 2\eta(B-C))}\}$, so that

$$\mathbf{E}[r(X_v^t) - R(X_v^t)] \leq \mathbf{E} \sup_{x \in \widetilde{\mathcal{X}}} (r(x) - R(x)) \leq 2D \sqrt{\frac{(t-1)(\eta^2 L^2 + 2\eta(B-C))}{nm}}.$$

■

A.5. Proof of Test Error Bounds for Smooth and Non-Smooth Losses

This section gives the proofs of the Test Error bounds presented within the main body of the work, namely Theorem 5 for convex, Lipschitz, and smooth losses, and Theorem 9 for convex and Lipschitz losses. This is achieved by using the error decomposition given in Proposition 1, and by bringing together the Generalisation Error bounds and the Optimisation Error bounds in Section 5.

Proof [Theorem 5] By the convexity of the Test Risk r, using Proposition 1, we get

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) \leq \frac{1}{t}\sum_{s=1}^{t}\Big(\underbrace{\mathbf{E}[r(X_v^{s+1}) - R(X_v^{s+1})]}_{\text{Generalisation Error}} + \underbrace{\mathbf{E}[R(X_v^{s+1}) - R(X^\star)]}_{\text{Optimisation Error}}\Big).$$

The proof follows by applying Lemma 13 for the Generalisation Error, which yields

$$\frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[r(X_v^{s+1}) - R(X_v^{s+1})] \leq \frac{2\eta L^2}{nm}\frac{1}{t}\sum_{s=1}^{t} s = \frac{\eta L^2}{nm}(t+1),$$

and by the Optimisation Error bound from Lemma 15. ■

Proof [Theorem 9] By the convexity of the Test Risk r, using Proposition 1, we get

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s}\Big) - r(x^\star) \leq \frac{1}{t}\sum_{s=1}^{t}\Big(\underbrace{\mathbf{E}[r(X_v^{s}) - R(X_v^{s})]}_{\text{Generalisation Error}} + \underbrace{\mathbf{E}[R(X_v^{s}) - R(X^\star)]}_{\text{Optimisation Error}}\Big).$$

The proof follows by applying Lemma 14 for the Generalisation Error, which yields

$$\frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[r(X_v^{s}) - R(X_v^{s})] \leq 2D\sqrt{\frac{(t-1)(\eta^2 L^2 + 2\eta(B-C))}{nm}},$$

and by the Optimisation Error bound from Lemma 16. ■

A.6. Calculations for Corollary 6 (Convex, Lipschitz, and Smooth)

This section presents the calculations needed to populate the table of rates in Corollary 6 in the case of convex, Lipschitz, and smooth losses. The simplification $1/(\beta + 1/\rho) \leq \rho$ is used. Additionally, minimisations are performed up to first-order terms in ρ, possibly neglecting logarithmic terms. This section is divided into four parts:

- **Optimisation Error** calculates the step size $\rho^\star_{\text{Opt}}$ minimising the Optimisation Error bound;
- **Test Error** calculates the step size $\rho^\star_{\text{Test}}$ that minimises the Test Error bound;
- **Early Stopping Optimisation** calculates the number of iterations that minimises the Test Error bound when the step size $\rho^\star_{\text{Opt}}$ is used;
- **Early Stopping Single-Machine Serial** calculates the number of iterations that minimises the Test Error bound when the step size $\rho^\star = O(1/\sqrt{t})$ is used.

Optimisation Error. Optimising over first-order terms in ρ in the Optimisation Error bound of Lemma 15 with $1/(\beta+1/\rho)\leq\rho$ we get

$$\rho^{\star}_{\text{Opt}}=\text{argmin}_{\rho}\Big\{\frac{\rho}{2}\sigma^2+\frac{G^2}{2t\rho}+3L\kappa\rho\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big\}=\frac{G}{\sqrt{t}}\frac{1}{\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2}},$$

which yields with $\frac{3+\beta\rho}{\beta+1/\rho}\leq 4\rho$ from $3/(\beta+1/\rho)\leq 3\rho$ and $\beta/(\beta+1/\rho)\leq\rho$ the Optimisation Error bound

$$\begin{aligned}
&\mathbf{E}\Big[R\Big(\frac{1}{t}\sum_{s=1}^{t}X_v^{s+1}\Big)-R(X^{\star})\Big]\\
&\leq\frac{G}{\sqrt{t}}\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2}+\frac{\beta G^2}{2t}+18\kappa^2\beta\rho^2_{\text{Opt}}\Big(\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)^2\\
&\leq\frac{G}{\sqrt{t}}\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2}+\frac{\beta G^2}{2t}\Big[1+\frac{6\kappa}{L}\frac{\Big(\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)^2}{\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\frac{\sigma^2}{6L\kappa}}\Big].
\end{aligned}$$

This bound is $\widetilde{O}(1/\sqrt{(1-\sigma_2(P))t})$ as the second term is $\widetilde{O}(1/((1-\sigma_2(P))t))$.

Test Error. Consider the Test Error bound in Theorem 5 with $1/(\beta+1/\rho)\leq\rho$. Optimising over first-order terms in ρ we get

$$\begin{aligned}
\rho^{\star}_{\text{Test}}&=\text{argmin}_{\rho}\Big\{\frac{\rho}{2}\sigma^2+\frac{G^2}{2t\rho}+3L\kappa\rho\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\frac{\rho L^2}{nm}(t+1)\Big\}\\
&=\frac{G}{\sqrt{t}}\frac{1}{\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2+\frac{2L^2(t+1)}{nm}}},
\end{aligned}$$

which yields with with $\frac{3+\beta\rho}{\beta+1/\rho}\leq 4\rho$ from $3/(\beta+1/\rho)\leq 3\rho$ and $\beta/(\beta+1/\rho)\leq\rho$ the Test Error bound

$$\begin{aligned}
&\mathbf{E}\,r\Big(\frac{1}{t}\sum_{s=1}^{t}X_v^{s+1}\Big)-r(x^{\star})\\
&\leq\frac{G}{\sqrt{t}}\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2+\frac{2L^2}{nm}(t+1)}+\frac{\beta G^2}{2t}+18\kappa^2\beta\rho^2_{\text{Test}}\Big(\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)^2\\
&\leq\frac{G}{\sqrt{t}}\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2+\frac{2L^2}{nm}(t+1)}\\
&\qquad\qquad+\frac{\beta G^2}{2t}\Big[1+\frac{6\kappa}{L}\frac{\Big(\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)^2}{\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\frac{1}{6L\kappa}(\sigma^2+\frac{2L^2}{nm}(t+1))}\Big].
\end{aligned}$$

This bound is $\widetilde{O}\Big(\sqrt{\frac{1}{t(1-\sigma_2(P))}+\frac{1}{nm}}\Big)$ as the second term is $\widetilde{O}(1/((1-\sigma_2(P))t))$. This is $\widetilde{O}(\frac{1}{\sqrt{nm}})$ when $t\gtrsim nm/(1-\sigma_2(P))$.

Early Stopping Optimisation. Considering the Test Error bound from Theorem 5 with step size $\rho = \rho^\star_{\text{Opt}}$ and $1/(\beta + 1/\rho) \le \rho$ we get

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) \le G\left[\frac{1}{\sqrt{t}}\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2} + \frac{2L^2\sqrt{t}}{nm}\sqrt{\frac{1}{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2}}\right]$$
$$+ \frac{\beta G^2}{2t}\Big[1+\frac{6\kappa}{L}\frac{\Big(\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)^2}{\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\frac{\sigma^2}{6L\kappa}}\Big],$$

where $(t+1)/\sqrt{t} \le 2\sqrt{t}$ was used. The first term is dominant and $O\Big(\sqrt{\frac{\log(t\sqrt{n})}{t(1-\sigma_2(P))}}+\frac{1}{nm}\sqrt{\frac{t(1-\sigma_2(P))}{\log(t\sqrt{n})}}\Big)$ while the second term is $\widetilde{O}(1/(1-\sigma_2(P)t))$. To minimise the first term with respect to t, consider the more tractable form

$$\frac{1}{\sqrt{t}}\sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2} + \frac{2L^2\sqrt{t}}{nm}\sqrt{\frac{1}{6L\kappa\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\sigma^2}}$$
$$\le \frac{\sigma}{\sqrt{t}} + \sqrt{6L\kappa\frac{\log((t+1)\sqrt{n})}{t(1-\sigma_2(P))}} + \frac{2L^2}{nm}\sqrt{\frac{t(1-\sigma_2(P))}{6L\kappa\log((t+1)\sqrt{n})}}.$$

An approximate minimiser in t neglecting the $\log((t+1)\sqrt{n})$ in the denominator is given by

$$\frac{t}{\log((t+1)\sqrt{n})} = \text{argmin}_{c\ge 0}\left\{\sqrt{\frac{6L\kappa}{c(1-\sigma_2(P))}} + \frac{2L^2}{nm}\sqrt{\frac{c(1-\sigma_2(P))}{6L\kappa}}\right\} = 3\frac{\kappa}{L}\frac{nm}{1-\sigma_2(P)}.$$

This choice yields the Test Error bound

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) \le \frac{G}{\sqrt{nm}}\Big[\sigma\sqrt{\frac{L(1-\sigma_2(P))}{3\kappa}} + 2\sqrt{2}L\Big]$$
$$+ \frac{\beta G^2 L(1-\sigma_2(P))}{6\kappa nm}\Big[1+\frac{6\kappa}{L}\frac{\Big(\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)^2}{\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}+\frac{\sigma^2}{6L\kappa}}\Big]$$

which is a $\widetilde{O}(\frac{1}{\sqrt{nm}})$ Test Error bound obtained with $t \simeq nm/(1-\sigma_2(P))$ iterations.

Early Stopping Single-Machine Serial. Considering the Test Error bound of Theorem 5 with $1/\beta + 1/\rho \le \rho$ and $\rho = \rho^\star = \frac{G}{Lc\sqrt{t}}$ for some constant $c > 0$. Plugging in we get

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) \le \frac{G}{c}\Big[3\kappa\frac{\log((t+1)\sqrt{n})}{(1-\sigma_2(P))\sqrt{t}} + \frac{2L}{nm}\sqrt{t}\Big]$$
$$+ \frac{G}{2\sqrt{t}}\Big(\frac{\sigma^2}{cL} + cL\Big) + \frac{\beta G^2}{2t}\Big[1 + \frac{9(3+\beta\rho)\kappa^2}{c^2L^2}\frac{\log^2((t+1)\sqrt{n})}{(1-\sigma_2(P))^2}\Big],$$

where $(t+1)/\sqrt{t} \leq 2\sqrt{t}$ for $t \geq 1$ was used on the Generalisation Error bound. The above bound is dominated by the first term which is $\widetilde{O}\Big(\frac{1}{(1-\sigma_2(P))\sqrt{t}} + \frac{\sqrt{t}}{nm}\Big)$. Minimising up to log terms yields

$$t = \frac{3\kappa}{2L}\frac{nm}{1-\sigma_2(P)}.$$

This choice yields the Test Error bound

$$\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) \leq \frac{G}{c}\sqrt{\frac{6\kappa L}{nm(1-\sigma_2(P))}}\Big[\log((t+1)\sqrt{n}) + 1\Big]$$
$$+ \sqrt{\frac{L(1-\sigma_2(P))}{6\kappa nm}}\Big(\frac{\sigma^2}{cL} + cL\Big) + \frac{L\beta G^2(1-\sigma_2(P))}{3\kappa nm}\Big[1 + \frac{9(3+\beta\rho)\kappa^2}{c^2L^2}\frac{\log^2((t+1)\sqrt{n})}{(1-\sigma_2(P))^2}\Big],$$

which is dominated by the first term that is $\widetilde{O}\Big(\frac{1}{\sqrt{nm(1-\sigma_2(P))}}\Big)$, as the third term is $\widetilde{O}\Big(\frac{1}{nm(1-\sigma_2(P))}\Big)$. Regarding the choice of constant c, note the above is decreasing up to $c = (1-\sigma_2(P))^{-1/2}$, in which case the $O(1/\sqrt{nm})$ rate for $\rho^\star_{\text{Opt}}$ is recovered.

A.7. Calculations for Corollary 10 (Convex and Lipschitz)

This section presents the calculations needed to populate the table of rates in Corollary 10 in the case of convex and Lipschitz losses. This section is divided into four parts:

- **Optimisation Error** calculates the step size $\eta^\star_{\text{Opt}}$ minimising the Optimisation Error bound;
- **Test Error** calculates the step size $\eta^\star_{\text{Test}}$ that minimises the Test Error bound;
- **Early Stopping Optimisation** calculates the number of iterations that minimises the Test Error bound when the step size $\eta^\star_{\text{Opt}}$ is used;
- **Early Stopping Single-Machine Serial** calculates the number of iterations that minimises the Test Error when the step size $\eta^\star = O(1/\sqrt{t})$ is used.

Optimisation Error. Minimising the Optimisation Error bound in Lemma 16 with respect to the step size yields $\eta = \eta^\star_{\text{Opt}} = \frac{G}{L\sqrt{19t}}\sqrt{\frac{1-\sigma_2(P)}{\log(t\sqrt{n})}}$ and

$$\mathbf{E}\Big[R\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^s\Big) - R(X^\star)\Big] \leq \sqrt{19}\frac{GL}{\sqrt{t}}\sqrt{\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}}.$$

Test Error. In this section the step size

$$\eta = \eta^\star_{\text{Test}} = \frac{G}{L\sqrt{t}}\frac{1}{\sqrt{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)} + \frac{t}{(nm)^{2/3}}}}$$

is shown to ensure that the Test Error bound in Theorem 9 converges in a time uniform manner to a quantity of order $\widetilde{O}(1/(nm)^{1/3})$. We consider the Optimisation and Generalisation Error separately.

The Optimisation Error bound with this step size yields

$$\begin{aligned}\frac{19}{2}\frac{\eta^\star_{\text{Test}} L^2 \log(t\sqrt{n})}{1-\sigma_2(P)} + \frac{G^2}{2\eta^\star_{\text{Test}} t} &= \frac{GL}{\sqrt{t}}\sqrt{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)} + \frac{t}{(nm)^{2/3}}}\Big[\frac{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}}{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}+\frac{t}{(nm)^{2/3}}} + \frac{1}{2}\Big] \\ &\leq \frac{3}{2}\frac{GL}{\sqrt{t}}\sqrt{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)} + \frac{t}{(nm)^{2/3}}},\end{aligned}$$

which is $\widetilde{O}(\frac{1}{(nm)^{1/3}})$ when the number of iterations satisfies $t \geq \frac{19}{2}\log(t\sqrt{n})(nm)^{2/3}/(1-\sigma_2(P))$. We split the Generalisation Error bound term into two parts

$$2D\sqrt{\frac{(t-1)(\eta^2L^2+2\eta(B-C))}{nm}} \leq 2\eta DL\sqrt{\frac{t}{nm}} + 2\sqrt{2}D\sqrt{\frac{\eta t(B-C)}{nm}} \qquad (12)$$

and bounded each part separately. The first quantity in (12) with the step size $\eta = \eta^\star_{\text{Test}}$ becomes

$$2\eta^\star_{\text{Test}} DL\sqrt{\frac{t}{nm}} = \frac{GD}{\sqrt{nm}}\frac{1}{\sqrt{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}+\frac{t}{(nm)^{2/3}}}} \leq \frac{GD}{\sqrt{nm}}\sqrt{\frac{2}{19}\frac{1-\sigma_2(P)}{\log(t\sqrt{n})}},$$

which is $O(1/\sqrt{nm})$, and thus $O(1/(nm)^{1/3})$. For the second quantity in (12), its square yields

$$\begin{aligned}8D^2\frac{\eta^\star_{\text{Test}} t(B-C)}{nm} &= 8D^2\frac{(B-C)G}{Lnm}\sqrt{\frac{t}{\frac{19}{2}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}+\frac{t}{(nm)^{2/3}}}} \\ &= 8D^2\frac{(B-C)GL}{L(nm)^{2/3}}\sqrt{\frac{t}{\frac{19}{2}\frac{\log(t\sqrt{n})(nm)^{2/3}}{1-\sigma_2(P)}+t}} \\ &\leq 8D^2\frac{(B-C)G}{L(nm)^{2/3}}.\end{aligned}$$

Therefore, when using step size $\eta^\star_{\text{Test}}$ with $t \gtrsim (nm)^{2/3}/(1-\sigma_2(P))$ the Test Error is bounded by the sum of three quantities each of which are $\widetilde{O}(1/(nm)^{1/3})$.

Early Stopping Optimisation. Setting $\eta = \eta^\star_{\text{Opt}}$ in the Test Error bound in Theorem 9 and using (12) to split the Generalisation Error we get

$$\begin{aligned}\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^s\Big) - r(x^\star) &\leq \sqrt{19}\frac{GL}{\sqrt{t}}\sqrt{\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}} \\ &\quad + \frac{2GD}{\sqrt{19nm}}\sqrt{\frac{1-\sigma_2(P)}{\log(t\sqrt{n})}} + 2\sqrt{2}D\sqrt{\frac{G(B-C)}{Lnm}\sqrt{\frac{t(1-\sigma_2(P))}{19\log(t\sqrt{n})}}}.\end{aligned}$$

This is $O\Big(\sqrt{\frac{\log(t\sqrt{n})}{t(1-\sigma_2(P))}}+\sqrt{\frac{1}{nm}\sqrt{\frac{t(1-\sigma_2(P))}{\log(t\sqrt{n})}}}\Big)$ as the second term is dominated by the first and third. Neglecting the $\log(t\sqrt{n})$ term and approximately minimising in t yields

$$\begin{aligned}\frac{t}{\log(t\sqrt{n})} &= \text{argmin}_{c>0}\Big\{GL\sqrt{\frac{1}{c}\frac{19}{1-\sigma_2(P)}}+2\sqrt{2}D\sqrt{\frac{G(B-C)}{Lnm}\sqrt{c\frac{1-\sigma_2(P)}{19}}}\Big\}\\ &= \frac{19L^2(Gnm)^{2/3}}{(1-\sigma_2(P))(2(B-C))^{2/3}D^{4/3}}\end{aligned}$$

with the final bound

$$\mathbf{E}\,r\Big(\frac{1}{t}\sum_{s=1}^{t}X_v^s\Big)-r(x^\star)\le 2^{1/3}3\frac{G^{2/3}(B-C)^{1/3}D^{2/3}}{(nm)^{1/3}}+\frac{2GD}{\sqrt{19nm}}\sqrt{\frac{1-\sigma_2(P)}{\log(t\sqrt{n})}}.$$

This is a $O(1/(nm)^{1/3})$ Test Error bound obtained with $t\simeq(nm)^{2/3}/(1-\sigma_2(P))$ iterations.

Early Stopping Single-Machine Serial. Setting $\eta=\eta^\star=\frac{G}{L\sqrt{19t}}$ in the Test Error bound in Theorem 9 and using (12) to split the Generalisation Error gives

$$\mathbf{E}\,r\Big(\frac{1}{t}\sum_{s=1}^{t}X_v^s\Big)-r(x^\star)\le GL\sqrt{\frac{19}{t}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}}+\frac{2GD}{\sqrt{19nm}}+2\sqrt{2}D\sqrt{\frac{G(B-C)}{Lnm}\sqrt{\frac{t}{19}}},$$

which is $\widetilde{O}\Big(\frac{1}{(1-\sigma_2(P))\sqrt{t}}+\sqrt{\frac{1}{nm}\sqrt{t}}\Big)$ as the second term is dominated by the first and third. Neglecting the $\log(t\sqrt{n})$ term and approximately minimising the above with respect to the number of iterations t yields

$$\begin{aligned}t &= \text{argmin}_{c>0}\Big\{GL\sqrt{\frac{19}{c}\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}}+2\sqrt{2}D\sqrt{\frac{G(B-C)}{Lnm}\sqrt{\frac{c}{19}}}\Big\}\\ &= \frac{19\log^{4/3}(t\sqrt{n})(Gnm)^{2/3}L^2}{(1-\sigma_2(P)^{4/3}(2(B-C))^{2/3}D^{4/3}}\end{aligned}$$

with the resulting bound

$$\mathbf{E}\,r\Big(\frac{1}{t}\sum_{s=1}^{t}X_v^s\Big)-r(x^\star)\le 2^{1/3}3\Big(\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}\Big)^{1/3}\frac{G^{2/3}(B-C)^{1/3}D^{2/3}}{(nm)^{1/3}}+\frac{2GD}{\sqrt{19nm}}.$$

This is $\widetilde{O}(1/(nm(1-\sigma_2(P))^{1/3})$ and is obtained with $t\simeq(nm)^{2/3}/(1-\sigma_2(P))^{4/3}$ iterations.

A.8. Calculation for Remark 8

In this section it is shown that Distributed SGD with step size choice $\rho=O((1-\sigma_2(P))/\sqrt{nm})$ and iterations $t=O(nm/(1-\sigma_2(P)))$ achieves optimal statistical rates up to logarithmic factors for convex, smooth and Lipschitz losses.

Begin by plugging $\rho = G(1-\sigma_2(P))/(L\sqrt{nm})$ into the Test Error bound of Theorem 5 with $1/(\beta+1/\rho) \leq \rho$ and $(3+\beta\rho)/(\beta+1/\rho) \leq 4\rho$, the latter arising from $3/(\beta+1/\rho) \leq 3\rho$ and $\beta\rho/(\beta+1/\rho) \leq \rho$. This then yields the Test Error bound

$$\begin{aligned}\mathbf{E}\, r\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - r(x^\star) &\leq \frac{2(1-\sigma_2(P))GLt}{(nm)^{3/2}} + \frac{\sigma^2(1-\sigma_2(P))G}{2L\sqrt{nm}} \\ &+ \frac{\beta G^2}{2t} + \frac{GL\sqrt{nm}}{2t(1-\sigma_2(P))} + \frac{3G\kappa\log((t+1)\sqrt{n})}{\sqrt{nm}} + 18\frac{\beta G^2\log^2((t+1)\sqrt{n})}{L^2 nm}.\end{aligned}$$

Choosing $t = (nm)/(1-\sigma_2(P))$ we see that the first and fourth terms become $O(1/\sqrt{nm})$ while the remaining terms are in this case $\widetilde{O}(1/\sqrt{nm})$.

Appendix B. Proofs of Optimisation Error bounds

This appendix presents Optimisation Error bounds for the Distributed Stochastic Subgradient Descent algorithm. Here we consider the general setting of stochastic first-order oracles. The Optimisation Error bounds presented within the main body of this work, specifically Lemma 15 and Lemma 16 for smooth and non-smooth losses, follow from Corollary 27 and Corollary 25 within this appendix.

B.1. Setup

Let (V,E) be a simple undirected graph with n nodes, and let $P \in \mathbb{R}^{n\times n}$ be a doubly stochastic matrix supported on the graph, i.e., $P_{ij} \neq 0$ only if $\{i,j\} \in E$. For each $v \in V$, let $F_v : \mathbb{R}^d \to \mathbb{R}$ be a random convex function. We consider the problem of minimizing the function $x \to \overline{F}(x) := \frac{1}{n}\sum_{v\in V} F_v(x)$ over $x \in \mathbb{R}^d$. Let $X^\star$ be a minimum of $\overline{F}$. Assume that $\mathbf{E}[\|X^\star\|^2] \leq G^2$ for a positive constant G. Given the initial vectors $\{X_v^1 = 0\}_{v\in V}$, throughout this appendix, we consider the following update for $s \geq 1$:

$$X_v^{s+1} = \sum_{w\in V} P_{vw}(X_w^s - \eta G_w^{s+1}), \tag{13}$$

where $\eta > 0$ is the step size, and each $G_v^{s+1} \in \mathbb{R}^d$ is an estimator of the subgradient of F_v evaluated at X_v^s. Specifically, for each $s \geq 1$ let $\mathcal{F}_s$ be the σ-algebra generated by the random functions $\{F_v\}_{v\in V}$ and by the estimators $\{G_v^k\}_{k\in\{2,\dots,s\}}$. We have, for any $s \geq 1$, $v \in V$,

$$\mathbf{E}[G_v^{s+1}|\mathcal{F}_s] \in \partial F_v(X_v^s). \tag{14}$$

Note that both $\{X_v^s\}_{v\in V}$ and $X^\star$ are measurable with respect to $\mathcal{F}_s$. Assume, for any $s \geq 1$, $v \in V$,

$$\mathbf{E}[\|G_v^{s+1}\|^2|\mathcal{F}_s] \leq L^2. \tag{15}$$

Section B.2 presents results for the setting just introduced under the additional assumption that the functions $\{F_v\}_{v\in V}$ are L-Lipschitz. Section B.3 presents results for the case where the functions $\{F_v\}_{v\in V}$ are smooth (Lipschitz continuity is not assumed in this case). Finally, Section B.4 checks that the assumptions of this general setting are satisfied for the specific case of algorithm (1).

The bounds that we derive are proved controlling the deviation of the output of the algorithm X_v^s from its network average $\overline{X}^s := \frac{1}{n}\sum_{v\in V} X_v^s$ on the one hand (*network term*), and bounding the deviation of $\overline{X}^s$ from the solution $X^\star$ on the other end (*optimisation term*). This strategy was originally proposed in Nedić et al. (2009) and used in Duchi et al. (2012) to get bounds that depend on the graph topology for a dual method in constrained optimisation. In the smooth case, we present a bound that also depends on the noise of the gradient (*gradient noise term*).

Remark 22 *The bounds that we derive naturally generalise the bounds in the single-machine setting. If no gradient noise is present and all the functions $\{F_v\}_{v\in V}$ are the same, then the network terms vanish as there is no difference between X_v^s and $\overline{X}^s$ (recall that the initial conditions are the same for each node, i.e., $X_v^1 = 0$ for all $v \in V$) and optimal tuning of the step sizes recovers the same rates as for serial SGD: $O(1/\sqrt{t})$ for the Lipschitz case and $O(1/t)$ for the smooth case.*

As the matrix P is doubly stochastic, the network average $\overline{X}^s$ admits the following simple evolution:

$$\overline{X}^{s+1} = \overline{X}^s - \eta\frac{1}{n}\sum_{v\in V} G_v^{s+1}. \tag{16}$$

In particular, note that by rearranging the previous expression we get

$$\frac{1}{n}\sum_{v\in V} G_v^{s+1} = \frac{1}{\eta}(\overline{X}^s - \overline{X}^{s+1}), \tag{17}$$

which will be used in the proofs in the next sections.

Before moving on to the next sections and presenting the Optimisation Error bounds, we establish bounds on the network terms that hold in the setting introduced so far. The next proposition bounds the deviation of X_v^s from $\overline{X}^s$ as a function of the second largest eigenvalue in magnitude of the matrix P, i.e., $\sigma_2(P)$. We present different bounds, that either depend on the Lipschitz parameter L or on a *Gradient Noise and Function Deviation Term* κ, as defined in (18). If no gradient noise is present and all the functions $\{F_v\}_{v\in V}$ are the same, then $\kappa = 0$, reflecting the comment in Remark 22.

Proposition 23 (Network term) *Consider the assumptions of Section B.1. Let κ^2 be such that, for any $v \in V, s \geq 1$,*

$$\underbrace{\mathbf{E}\Big[\Big\|G_v^{s+1} - \frac{1}{n}\sum_{\ell=1}^{n} \nabla F_\ell(X_\ell^s)\Big\|^2\Big]}_{\text{Gradient Noise and Function Deviation Term}} \leq \kappa^2. \tag{18}$$

For any $v \in V, s \geq 1$, we have

$$\mathbf{E}[\|X_v^s - \overline{X}^s\|^2] \leq \eta^2 \min\{L^2, \kappa^2\}\Big(2\frac{\log(s\sqrt{n})}{1-\sigma_2(P)} + 1\Big)^2.$$

Proof Fix $v \in V, s \geq 1$. By unraveling the updates in (13) and (16), using that $X_v^1 = 0$ for all $v \in V$, we get

$$X_v^s = -\eta\sum_{k=1}^{s-1}\sum_{w\in V} P_{vw}^k G_w^{s-k+1}, \qquad \overline{X}^s = -\eta\sum_{k=1}^{s-1}\sum_{w\in V} (\tfrac{1}{n}11^\top)_{vw} G_w^{s-k+1},$$

where $1 \in \mathbb{R}^n$ is the all-one vector. Using that the rows of the matrix P sum to one, note that for any collection of vectors $\{\nu^k\}_{k=1}^{s-1}$ in $\mathbb{R}^d$ we have

$$X_v^s - \overline{X}^s = \eta \sum_{k=1}^{s-1} \sum_{w \in V} (P^k - \tfrac{1}{n} 1 1^\top)_{vw} (G_w^{s-k+1} - \nu^{s-k}).$$

We have

$$\mathbf{E}[\|X_v^s - \overline{X}^s\|^2] = \mathbf{E}\langle X_v^s - \overline{X}^s, X_v^s - \overline{X}^s \rangle$$
$$\leq \eta^2 \sum_{k,k'=1}^{s-1} \sum_{w,w' \in V} |(P^k - \tfrac{1}{n} 1 1^\top)_{vw}| |(P^{k'} - \tfrac{1}{n} 1 1^\top)_{vw'}| \, \mathbf{E}|\langle G_w^{s-k+1} - \nu^{s-k}, G_{w'}^{s-k'+1} - \nu^{s-k'} \rangle|.$$

By Cauchy-Schwarz's inequality and Hölder's inequality,

$$\mathbf{E}|\langle G_w^{s-k+1} - \nu^{s-k}, G_{w'}^{s-k'+1} - \nu^{s-k'} \rangle| \leq \sqrt{\mathbf{E}[\|G_w^{s-k+1} - \nu^{s-k}\|^2]} \sqrt{\mathbf{E}[\|G_{w'}^{s-k'+1} - \nu^{s-k'}\|^2]},$$

and the above yields

$$\mathbf{E}[\|X_v^s - \overline{X}^s\|^2] \leq \left(\eta \sum_{k=1}^{s-1} \sum_{w \in V} |(P^k - \tfrac{1}{n} 1 1^\top)_{vw}| \sqrt{\mathbf{E}[\|G_w^{s-k+1} - \nu^{s-k}\|^2]} \right)^2.$$

By choosing $\nu^k = 0$ and using (15), we get

$$\mathbf{E}[\|X_v^s - \overline{X}^s\|^2] \leq \eta^2 L^2 \left(\sum_{k=1}^{s-1} \sum_{w \in V} |(P^k - \tfrac{1}{n} 1 1^\top)_{vw}| \right)^2.$$

By choosing $\nu^k = \frac{1}{n} \sum_{\ell=1}^n \nabla F_\ell(X_\ell^k)$ and using the assumption of the proposition, we get

$$\mathbf{E}[\|X_v^s - \overline{X}^s\|^2] \leq \eta^2 \kappa^2 \left(\sum_{k=1}^{s-1} \sum_{w \in V} |(P^k - \tfrac{1}{n} 1 1^\top)_{vw}| \right)^2.$$

Note that $\sum_{k=1}^{s-1} \sum_{w \in V} |(P^k - \frac{1}{n} 1 1^\top)_{vw}| = \sum_{k=1}^{s-1} \|e_v^\top P^k - \frac{1}{n} 1^\top\|_1$, where $e_v \in \mathbb{R}^n$ is the vector with 1 in the coordinate aligning with node v and 0 everywhere else, and $\|\cdot\|_1$ denotes the ℓ_1 norm. Standard results from Perron-Frobenius theory yield, for any $k \geq 1$,

$$\|e_v^\top P^k - \tfrac{1}{n} 1^\top\|_1 \leq \sqrt{n} \|e_v^\top P^k - \tfrac{1}{n} 1^\top\| \leq \sqrt{n} \sigma_2(P)^k.$$

To bound the quantity $\sum_{k=1}^{s-1} \|e_v^\top P^k - \frac{1}{n} 1^\top\|_1$, break the sum and bound each part separately. For the first half use $\|e_v^\top P^k - \frac{1}{n} 1^\top\|_1 \leq \|e_v^\top P^k\|_1 + \|\frac{1}{n} 1^\top\|_1 = 2$ so

$$\sum_{k=1}^{s-1} \|e_v^\top P^k - \tfrac{1}{n} 1^\top\|_1 = \sum_{k=1}^{\tilde{s}} \|e_v^\top P^k - \tfrac{1}{n} 1^\top\|_1 + \sum_{k=\tilde{s}+1}^{s-1} \|e_v^\top P^k - \tfrac{1}{n} 1^\top\|_1 \leq 2\tilde{s} + \sqrt{n} \sum_{k=\tilde{s}+1}^{s-1} \sigma_2(P)^k.$$

Requiring $\sigma_2(P)^k \leq \frac{1}{s\sqrt{n}}$ for k between $\tilde{s}+1$ and $s-1$, set $\tilde{s} = \lfloor \frac{\log(s\sqrt{n})}{\log(\sigma_2(P)^{-1})} \rfloor$. As there are no more than s terms in the sum, using that $\log(x^{-1}) \geq 1 - x$, we get

$$\sum_{k=1}^{s-1} \|e_v^\top P^k - \tfrac{1}{n} 1^\top\|_1 \leq 2\tilde{s} + 1 \leq 2 \frac{\log(s\sqrt{n})}{1 - \sigma_2(P)} + 1.$$

■

B.2. Convex and Lipschitz

The following result controls the evolution of algorithm (13) in the setting defined in Section B.1, under the additional assumption of Lipschitz continuity. The proof is inspired from the analysis in Duchi et al. (2012),

Theorem 24 (Optimisation bound for convex and Lipschitz objectives) *Consider the setting of Section B.1. Let the functions $\{F_v\}_{v\in V}$ be L-Lipschitz. Then, Distributed SGD yields, for any $v \in V$ and $t \geq 1$,*

$$\begin{aligned}\mathbf{E}\Big[\overline{F}\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^s\Big) - \overline{F}(X^\star)\Big] &\leq \frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[\overline{F}(X_v^s) - \overline{F}(X^\star)] \\ &\leq \underbrace{\frac{3L}{t}\max_{w\in V}\sum_{s=1}^{t}\mathbf{E}\|X_w^s - \overline{X}^s\|}_{\textit{Network Term}} + \underbrace{\frac{1}{t}\sum_{s=1}^{t}\frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1}, \overline{X}^s - X^\star\rangle}_{\textit{Optimisation Term}}.\end{aligned}$$

and the Optimisation Term is upper bounded by $\frac{G^2}{2\eta t} + \frac{\eta L^2}{2}$.

Proof For any $s \geq 1$ and $v \in V$, adding and subtracting the term $\frac{1}{n}\sum_{w\in V} F_w(X_w^s)$, we find

$$\begin{aligned}\mathbf{E}[\overline{F}(X_v^s) - \overline{F}(X^\star)] &= \frac{1}{n}\sum_{w\in V}\mathbf{E}[F_w(X_v^s) - F_w(X_w^s)] + \frac{1}{n}\sum_{w\in V}\mathbf{E}[F_w(X_w^s) - F_w(X^\star)] \\ &\leq \frac{1}{n}\sum_{w\in V} L\mathbf{E}\|X_v^s - X_w^s\| + \frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1}, X_w^s - X^\star\rangle,\end{aligned}$$

where for the first summand we used the Lipschitz property, and for the second summand we used convexity, assumption (14), and that both $\{X_v^s\}_{v\in V}$ and $X^\star$ are measurable with respect to $\mathcal{F}_s$. In fact, for any $w \in V$, we have

$$F_w(X_w^s) - F_w(X^\star) \leq \langle \partial F_w(X_w^s), X_w^s - X^\star\rangle = \langle \mathbf{E}[G_w^{s+1}|\mathcal{F}_s], X_w^s - X^\star\rangle = \mathbf{E}[\langle G_w^{s+1}, X_w^s - X^\star\rangle|\mathcal{F}_s],$$

so that $\mathbf{E}[F_w(X_w^s) - F_w(X^\star)] \leq \mathbf{E}\langle G_w^{s+1}, X_w^s - X^\star\rangle$ by the tower property of conditional expectations. By adding and subtracting $\overline{X}^s$ and applying the Cauchy-Schwarz's inequality, we have

$$\mathbf{E}\langle G_w^{s+1}, X_w^s - X^\star\rangle \leq \mathbf{E}[\|G_w^{s+1}\|\|X_w^s - \overline{X}^s\|] + \mathbf{E}\langle G_w^{s+1}, \overline{X}^s - X^\star\rangle,$$

and the first term on the right-hand side is further bounded by using Jensen's inequality and the fact that $(X_w^s - \overline{X}^s)$ is $\mathcal{F}_s$-measurable, along with assumption (15), giving

$$\mathbf{E}[\|G_w^{s+1}\|\|X_w^s - \overline{X}^s\|] \leq \mathbf{E}[(\mathbf{E}[\|G_w^{s+1}\|^2|\mathcal{F}_s])^{1/2}\|X_w^s - \overline{X}^s\|] \leq L\mathbf{E}\|X_w^s - \overline{X}^s\|.$$

All together with $\|X_v^t - X_w^s\| \leq 2\max_{w'\in V}\|X_{w'}^s - \overline{X}^s\|$ we have

$$\frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[\overline{F}(X_v^s) - \overline{F}(X^\star)] \leq \frac{3L}{t}\max_{w\in V}\sum_{s=1}^{t}\mathbf{E}\|X_w^s - \overline{X}^s\| + \frac{1}{t}\sum_{s=1}^{t}\frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1}, \overline{X}^s - X^\star\rangle.$$

To bound the Optimisation Term we proceed as follows. Using (17) and that $\langle a,b\rangle = (\|a\|^2 + \|b\|^2 - \|a-b\|^2)/2$ we obtain

$$\begin{aligned}\frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1},\overline{X}^s - X^\star\rangle &= \frac{1}{\eta}\mathbf{E}\langle \overline{X}^s - \overline{X}^{s+1},\overline{X}^s - X^\star\rangle\\ &= \frac{1}{2\eta}(\mathbf{E}[\|\overline{X}^{s+1} - \overline{X}^s\|^2] + \mathbf{E}[\|\overline{X}^s - X^\star\|^2] - \mathbf{E}[\|\overline{X}^{s+1} - X^\star\|^2])\\ &\le \frac{1}{2\eta}\Big(\mathbf{E}[\|\overline{X}^s - X^\star\|^2] - \mathbf{E}[\|\overline{X}^{s+1} - X^\star\|^2] + \eta^2\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}G_w^{s+1}\Big\|^2\Big]\Big)\\ &\le \frac{1}{2\eta}(\mathbf{E}[\|\overline{X}^s - X^\star\|^2] - \mathbf{E}[\|\overline{X}^{s+1} - X^\star\|^2] + \eta^2 L^2),\end{aligned}$$

where we used Cauchy-Schwarz's and Hölder's inequalities, along with assumption (15), to get

$$\begin{aligned}\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}G_w^{s+1}\Big\|^2\Big] &= \frac{1}{n^2}\sum_{u,w\in V}\mathbf{E}\langle G_u^{s+1},G_w^{s+1}\rangle \le \frac{1}{n^2}\sum_{u,w\in V}\mathbf{E}[\|G_u^{s+1}\|\|G_w^{s+1}\|]\\ &\le \frac{1}{n^2}\sum_{u,w\in V}\sqrt{\mathbf{E}[\|G_u^{s+1}\|^2]}\sqrt{\mathbf{E}[\|G_w^{s+1}\|^2]} \le L^2. \qquad (19)\end{aligned}$$

Summing over s, using that $X_v^1 = 0$ for all $v \in V$ and that $\mathbf{E}[\|X^\star\|^2] \le G^2$, we get the following bound for the Optimisation Term

$$\frac{1}{t}\sum_{s=1}^{t}\frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1},\overline{X}^s - X^\star\rangle \le \frac{1}{2\eta t}\mathbf{E}[\|\overline{X}^1 - X^\star\|^2] + \frac{\eta L^2}{2} \le \frac{G^2}{2\eta t} + \frac{\eta L^2}{2}.$$

■

Corollary 25 *Consider the assumptions of Section B.1. Let the functions $\{F_v\}_{v\in V}$ be L-Lipschitz. Then, Distributed SGD yields, for any $v \in V$ and $t \ge 1$,*

$$\mathbf{E}\Big[\overline{F}\Big(\frac{1}{t}\sum_{s=1}^{t}X_v^s\Big) - \overline{F}(X^\star)\Big] \le \frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[\overline{F}(X_v^s) - \overline{F}(X^\star)] \le \frac{\eta L^2}{2}\Big(19\frac{\log(t\sqrt{n})}{1-\sigma_2(P)}\Big) + \frac{G^2}{2\eta t}.$$

Proof It follows from Theorem 24 and Proposition 23, as $\mathbf{E}\|X_v^s - \overline{X}^s\| \le \sqrt{\mathbf{E}[\|X_v^s - \overline{X}^s\|^2]}$ by Jensen's inequality. ■

B.3. Convex and Smooth

The following result controls the evolution of algorithm (13) in the setting defined in Section B.1, under the additional assumption of smoothness. The proof is inspired by the one given Dekel et al. (2012) for single-machine serial SGD applied to smooth losses, the specific exposition of which more closely follows Bubeck et al. (2015). The bound that we give is made of three components: the Optimisation Term that decays like $1/t$; the Gradient Noise Term that captures the average noise of the gradient across the graph; the Network Term that controls the deviation of the algorithm from its network average.

Theorem 26 (Optimisation bound for convex and smooth objectives) *Consider the Assumptions of Section B.1. Let the functions $\{F_v\}_{v\in V}$ be β-smooth. Then, Distributed SGD with $\eta = 1/(\beta + 1/\rho)$ and $\rho \ge 0$, yields, for any $v \in V$ and $t \ge 1$,*

$$\mathbf{E}\Big[\overline{F}\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - \overline{F}(X^\star)\Big] \le \frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[\overline{F}(X_v^{s+1}) - \overline{F}(X^\star)]$$

$$\le \underbrace{\frac{1}{t}\sum_{s=1}^{t}\Big(L\mathbf{E}\|X_v^{s+1}-\overline{X}^{s+1}\| + \beta\max_{w\in V}\mathbf{E}[\|X_w^{s+1}-\overline{X}^{s+1}\|^2] + \frac{\beta}{2}\Big(1+\beta\rho\Big)\max_{w\in V}\mathbf{E}[\|X_w^{s}-\overline{X}^{s}\|^2]\Big)}_{\textit{Network Term}}$$

$$+\underbrace{\frac{\rho}{2}\frac{1}{t}\sum_{s=1}^{t}\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(G_w^{s+1}-\nabla F_w(X_w^s))\Big\|^2\Big]}_{\textit{Gradient Noise Term}}$$

$$+\underbrace{\frac{1}{t}\sum_{s=1}^{t}\Big(\frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1},\overline{X}^{s+1}-X^\star\rangle + \frac{1}{2}\Big(\beta+\frac{1}{\rho}\Big)\mathbf{E}[\|\overline{X}^{s+1}-\overline{X}^{s}\|^2]\Big)}_{\textit{Optimisation Term}},$$

and the Optimisation Term is upper bounded by $\frac{1}{2}(\beta+\frac{1}{\rho})\frac{G^2}{t}$.

Proof Recall that if a function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth then for any $x, y \in \mathbb{R}^d$ we have (Nesterov, 2013) $f(x) - f(y) \le \langle \nabla f(y), x - y\rangle + \frac{\beta}{2}\|x-y\|^2$. Fix $s \ge 1$, $v \in V$. Consider the following decomposition.

$$\overline{F}(X_v^{s+1}) - \overline{F}(X^\star) = \underbrace{\overline{F}(X_v^{s+1}) - \overline{F}(\overline{X}^{s+1})}_{\text{Term (a)}} + \underbrace{\overline{F}(\overline{X}^{s+1}) - \overline{F}(X^\star)}_{\text{Term (b)}}. \tag{20}$$

Term (a). To bound Term (a), we use smoothness and convexity to get

$$\overline{F}(X_v^{s+1}) - \overline{F}(\overline{X}^{s+1}) = \frac{1}{n}\sum_{w\in V}\Big(F_w(X_v^{s+1}) - F_w(X_w^{s+1}) + F_w(X_w^{s+1}) - F_w(\overline{X}^{s+1})\Big)$$

$$\le \frac{1}{n}\sum_{w\in V}\Big(\langle\nabla F_w(X_w^{s+1}), X_v^{s+1}-X_w^{s+1}\rangle + \frac{\beta}{2}\|X_v^{s+1}-X_w^{s+1}\|^2 + \langle\nabla F_w(X_w^{s+1}), X_w^{s+1}-\overline{X}^{s+1}\rangle\Big)$$

$$= \frac{1}{n}\sum_{w\in V}\Big(\langle\nabla F_w(X_w^{s+1}), X_v^{s+1}-\overline{X}^{s+1}\rangle + \frac{\beta}{2}\|X_v^{s+1}-X_w^{s+1}\|^2\Big).$$

As $\nabla F_w(X_w^{s+1}) = \mathbf{E}[G_w^{s+2}|\mathcal{F}_{s+1}]$ and $\{X_w^{s+1}\}_{w\in V}$ are $\mathcal{F}_{s+1}$-measurable, we get

$$\begin{aligned}\langle\nabla F_w(X_w^{s+1}), X_v^{s+1}-\overline{X}^{s+1}\rangle &= \mathbf{E}[\langle G_w^{s+2}, X_v^{s+1}-\overline{X}^{s+1}\rangle|\mathcal{F}_{s+1}]\\ &\le \mathbf{E}[\|G_w^{s+2}\|\|X_v^{s+1}-\overline{X}^{s+1}\||\mathcal{F}_{s+1}]\\ &\le \sqrt{\mathbf{E}[\|G_w^{s+2}\|^2|\mathcal{F}_{s+1}]}\|X_v^{s+1}-\overline{X}^{s+1}\|\\ &\le L\|X_v^{s+1}-\overline{X}^{s+1}\|,\end{aligned}$$

where we used Cauchy-Schwarz's inequality, Jensen's inequality, and $\mathbf{E}[\|G_w^{s+2}\|^2|\mathcal{F}_{s+1}] \leq L^2$. Thus,

$$\mathbf{E}[\overline{F}(X_v^{s+1}) - \overline{F}(\overline{X}^{s+1})] \leq L\mathbf{E}\|X_v^{s+1} - \overline{X}^{s+1}\| + \beta \max_{w\in V} \mathbf{E}[\|X_w^{s+1} - \overline{X}^{s+1}\|^2]. \tag{21}$$

Term (b). To bound Term (b), we use smoothness to find a bound that involves a telescoping sum whose terms cancel out when we take the summation over time s. Using smoothness, adding and subtracting $\langle G_w^{s+1}, \overline{X}^{s+1}, \overline{X}^s\rangle = \langle G_w^{s+1}, \overline{X}^{s+1} - X^\star\rangle + \langle G_w^{s+1}, X^\star - \overline{X}^{s+1}\rangle$ and using Cauchy-Schwarz's inequality ($2\langle a,b\rangle \leq \rho\|a\|^2 + \|b\|^2/\rho$ for $\rho \geq 0$) we get

$$\begin{aligned}\overline{F}(\overline{X}^{s+1}) - \overline{F}(\overline{X}^s) &\leq \frac{1}{n}\sum_{w\in V}\langle \nabla F_w(\overline{X}^s), \overline{X}^{s+1} - \overline{X}^s\rangle + \frac{\beta}{2}\|\overline{X}^{s+1} - \overline{X}^s\|^2 \\ &= \Big\langle \frac{1}{n}\sum_{w\in V}(\nabla F_w(\overline{X}^s) - G_w^{s+1}), \overline{X}^{s+1} - \overline{X}^s\Big\rangle + \frac{1}{n}\sum_{w\in V}\langle G_w^{s+1}, \overline{X}^{s+1} - X^\star\rangle \\ &\quad + \frac{1}{n}\sum_{w\in V}\langle G_w^{s+1}, X^\star - \overline{X}^s\rangle + \frac{\beta}{2}\|\overline{X}^{s+1} - \overline{X}^s\|^2 \\ &\leq \frac{\rho}{2}\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(\overline{X}^s) - G_w^{s+1})\Big\|^2 + \frac{1}{n}\sum_{w\in V}\langle G_w^{s+1}, \overline{X}^{s+1} - X^\star\rangle \\ &\quad + \frac{1}{n}\sum_{w\in V}\langle G_w^{s+1}, X^\star - \overline{X}^s\rangle + \frac{1}{2}\Big(\beta + \frac{1}{\rho}\Big)\|\overline{X}^{s+1} - \overline{X}^s\|^2. \end{aligned} \tag{22}$$

Adding $\overline{F}(\overline{X}^s)$ to both sides, taking expectation, using that $\{X_w^s\}_{w\in V}$ and $X^\star$ are $\mathcal{F}_s$-measurable, and that $\mathbf{E}\langle[G_w^{s+1}, X^\star - \overline{X}^s\rangle|\mathcal{F}_s] = \langle \nabla F_w(X_w^s), X^\star - \overline{X}^s\rangle$, we get

$$\begin{aligned}\mathbf{E}[\overline{F}(\overline{X}^{s+1}) - \overline{F}(X^\star)] \leq \mathbf{E}[\overline{F}(\overline{X}^s) - \overline{F}(X^\star)] &+ \frac{\rho}{2}\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(\overline{X}^s) - G_w^{s+1})\Big\|^2\Big] \\ &+ \frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1}, \overline{X}^{s+1} - X^\star\rangle + \frac{1}{2}\Big(\beta + \frac{1}{\rho}\Big)\mathbf{E}[\|\overline{X}^{s+1} - \overline{X}^s\|^2] \\ &+ \frac{1}{n}\sum_{w\in V}\mathbf{E}\langle \nabla F_w(X_w^s), X^\star - \overline{X}^s\rangle. \end{aligned} \tag{23}$$

To bound the first term on the right-hand side of bound (23) and cancel the dependence on $X^\star$ from the term $\langle \nabla F_w(X_w^s), X^\star - \overline{X}^s\rangle$, note that by convexity and smoothness we get

$$\begin{aligned}\mathbf{E}[\overline{F}(\overline{X}^s) - \overline{F}(X^\star)] &= \frac{1}{n}\sum_{w\in V}\mathbf{E}[F_w(\overline{X}^s) - F_w(X_w^s) + F_w(X_w^s) - F_w(X^\star)] \\ &= \frac{1}{n}\sum_{w\in V}\mathbf{E}[F_w(\overline{X}^s) - F_w(X_w^s) + \langle \nabla F_w(X_w^s), X_w^s - \overline{X}^s\rangle + \langle \nabla F_w(X_w^s), \overline{X}^s - X^\star\rangle] \\ &\leq \frac{\beta}{2}\max_{w\in V}\mathbf{E}\|X_w^s - \overline{X}^s\|^2 + \frac{1}{n}\sum_{w\in V}\mathbf{E}\langle \nabla F_w(X_w^s), \overline{X}^s - X^\star\rangle. \end{aligned} \tag{24}$$

To bound the second term on the right-hand side of bound (23), note that

$$\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(\overline{X}^s)-G_w^{s+1})\Big\|^2\Big]=\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}\Big(\nabla F_w(\overline{X}^s)-\nabla F_w(X_w^s)+\nabla F_w(X_w^s)-G_w^{s+1}\Big)\Big\|^2\Big]$$
$$=\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(\overline{X}^s)-\nabla F_w(X_w^s))\Big\|^2\Big]+\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(X_w^s)-G_w^{s+1})\Big\|^2\Big], \qquad (25)$$

where we used that the cross terms are zero as $\mathbf{E}[G_w^{s+1}|\mathcal{F}_s]=\nabla F_w(X_w^s)$ and both $\{F_w\}_{w\in V}$ and $\{X_w^s\}_{w\in V}$ are $\mathcal{F}_s$-measurable. The first term in (25) can be bounded as follows:

$$\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(X_w^s)-\nabla F_w(\overline{X}^s))\Big\|^2\Big]$$
$$=\frac{1}{n^2}\sum_{w,l\in V}\mathbf{E}\langle\nabla F_w(X_w^s)-\nabla F_w(\overline{X}^s),\nabla F_l(X_l^s)-\nabla F_l(\overline{X}^s)\rangle$$
$$\leq\frac{1}{n^2}\sum_{w,l\in V}\mathbf{E}\big[\|\nabla F_w(X_w^s)-\nabla F_w(\overline{X}^s)\|\|\nabla F_l(X_l^s)-\nabla F_l(\overline{X}^s)\|\big]$$
$$\leq\frac{\beta^2}{n^2}\sum_{w,l\in V}\mathbf{E}\big[\|X_w^s-\overline{X}^s\|\|X_l^s-\overline{X}^s\|\big]$$
$$\leq\frac{\beta^2}{n^2}\sum_{w,l\in V}\sqrt{\mathbf{E}\big[\|X_w^s-\overline{X}^s\|^2\big]}\sqrt{\mathbf{E}\big[\|X_l^s-\overline{X}^s\|^2\big]}$$
$$\leq\beta^2\max_{w\in V}\mathbf{E}\big[\|X_w^s-\overline{X}^s\|^2\big], \qquad (26)$$

where applied Cauchy-Schwarz's inequality, smoothness, and Hölder's inequality. Plugging (24), (25), and (26) into (23) we get the following bound for the expected value of term (b):

$$\mathbf{E}[\overline{F}(\overline{X}^{s+1})-\overline{F}(X^\star)]\leq\frac{\beta}{2}\Big(1+\beta\rho\Big)\max_{w\in V}\mathbf{E}[\|X_w^s-\overline{X}^s\|^2]+\frac{\rho}{2}\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(\nabla F_w(X_w^s)-G_w^{s+1})\Big\|^2\Big]$$
$$+\frac{1}{n}\sum_{w\in V}\mathbf{E}\langle G_w^{s+1},\overline{X}^{s+1}-X^\star\rangle+\frac{1}{2}\Big(\beta+\frac{1}{\rho}\Big)\mathbf{E}[\|\overline{X}^{s+1}-\overline{X}^s\|^2]. \qquad (27)$$

Term (a) + Term (b). The main result in the theorem follows by using bounds (21) and (27) to bound Term (a) and Term (b) in (20), taking the summation over time from $s=1$ to $s=t$.

To bound the Optimisation Term, use (17) and that $2\langle a,b\rangle=\|a\|^2+\|b\|^2-\|a-b\|^2$ so that

$$\frac{1}{n}\sum_{w\in V}\langle G_w^{s+1},\overline{X}^{s+1}-X^\star\rangle=\frac{1}{\eta}\langle\overline{X}^s-\overline{X}^{s+1},\overline{X}^{s+1}-X^\star\rangle$$
$$=-\frac{1}{\eta}\langle\overline{X}^{s+1}-\overline{X}^s,\overline{X}^{s+1}-X^\star\rangle$$
$$=\frac{1}{2\eta}\Big(-\|\overline{X}^{s+1}-\overline{X}^s\|^2-\|\overline{X}^{s+1}-X^\star\|^2+\|\overline{X}^s-X^\star\|^2\Big).$$

The choice $\eta=\frac{1}{\beta+1/\rho}$ leads to the cancellation of the quantity $\|\overline{X}^{s+1}-\overline{X}^s\|^2$ in the Optimisation Term. The telescoping sum over time, using that $X_w^1=0$ for all $w\in V$ and the assumption

$\mathbf{E}[\|X^\star\|^2] \leq G^2$, yields the final result. ■

As for single-machine serial SGD (Dekel et al., 2012), the error bound that we give in Theorem 26 for the smooth case exhibits explicit dependence on the gradient noise, which in our setting is averaged out across the network. As far as the following corollary is concerned, we assume a time-uniform control on the gradient noise, namely,

$$\mathbf{E}\Big[\Big\|\frac{1}{n}\sum_{w\in V}(G_w^{s+1} - \nabla F_w(X_w^s))\Big\|^2\Big] \leq \sigma^2 \tag{28}$$

for any $s \geq 1$.

Corollary 27 *Consider the Assumptions of Section B.1. Let the functions* $\{F_v\}_{v\in V}$ *be* β*-smooth and satisfy both* (18) *and* (28)*. Then, Distributed SGD with* $\eta = 1/(\beta + 1/\rho)$ *and* $\rho \geq 0$*, yields, for any* $v \in V$ *and* $t \geq 1$*,*

$$\begin{aligned}
&\mathbf{E}\Big[\overline{F}\Big(\frac{1}{t}\sum_{s=1}^{t} X_v^{s+1}\Big) - \overline{F}(X^\star)\Big] \leq \frac{1}{t}\sum_{s=1}^{t}\mathbf{E}[\overline{F}(X_v^{s+1}) - \overline{F}(X^\star)] \\
&\leq \frac{\rho}{2}\sigma^2 + \frac{(\beta+1/\rho)G^2}{2t} + \frac{3\kappa}{\beta+1/\rho}\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big(L + \frac{3}{2}\frac{\beta(3+\beta\rho)\kappa}{\beta+1/\rho}\frac{\log((t+1)\sqrt{n})}{1-\sigma_2(P)}\Big)
\end{aligned}$$

Proof It follows from Theorem 26 and Proposition 23. ■

B.4. Assumptions for Distributed SGD (1)

This section verifies that the more general assumptions considered in this Appendix for Distributed SGD (13) are satisfied within the context of the main body of this work, that is, for Distributed SGD (1) as described within Section 3. This is performed by placing Distributed SGD (1) into the context Distributed SGD (13) as follows. Let the random objective functions be $F_v(x) = R_v(x) = \frac{1}{m}\sum_{k=1}^m \ell(x, Z_{v,k})$ for $v \in V$, which yields the network average $\overline{F}(x) = R(x)$. Consider the following stochastic gradients, for $v \in V$ and $s \geq 1$,

$$G_v^{s+1} = \partial\ell(X_v^s, Z_{v,K_v^{s+1}}),$$

where K_v^s is a uniform random variable on $[m]$. Let $\mathcal{F}_1$ be the σ-algebra generated by the data sets $\mathcal{D}$. For any $s \geq 2$, let $\mathcal{F}_s$ contain the σ-algebra generated by the data sets $\mathcal{D}$ and the uniform random variables up to time s $\{K_v^2, \ldots, K_v^s\}_{v\in V}$. The random functions $\{F_v\}_{v\in V}$ and their optimal value $X^\star$ are $\mathcal{F}_s$-measurable, as $\mathcal{F}_s$ contains the σ-algebra generated by $\mathcal{D}$. The iterates $\{X_v^k\}_{k\leq s, v\in V}$ are also $\mathcal{F}_s$-measurable, as $\mathcal{F}_s$ contains the σ-algebra generated by $\{K_v^2, \ldots, K_v^s\}_{v\in V}$. We now check that assumption (14) and assumption (15) are satisfied. The following hold for any $s \geq 1$.

- Assumption (14) on the unbiasedness of the subgradient estimators is satisfied as for any $v \in V$ we have

$$\mathbf{E}[G_v^{s+1}|\mathcal{F}_s] = \mathbf{E}[\partial\ell(X_v^s, Z_{v,K_v^{s+1}})\big|\mathcal{F}_s] = \frac{1}{m}\sum_{k=1}^m \partial\ell(X_v^s, Z_{v,k}) \in \partial F_v(X_v^s),$$

  where have used that the sum of subgradients belong to the subgradient of sums.

- Assumption (15) on the boundedness of the second moment of the subgradients is satisfied as for any $v \in V$ we have

  $$\mathbf{E}[\|G_v^{s+1}\|^2|\mathcal{F}_s] = \mathbf{E}[\|\partial\ell(X_v^s, Z_{v,K_v^{s+1}})\|^2\big|\mathcal{F}_s] = \frac{1}{m}\sum_{k=1}^{m}\|\partial\ell(X_v^s, Z_{v,k})\|^2 \leq L^2,$$

  where we have used that the function $\ell(\,\cdot\,, z)$ is L-Lipschitz for all $z \in Z$.

References

Alekh Agarwal and John C. Duchi. Distributed Delayed Stochastic Optimization. In *Advances in Neural Information Processing Systems*, pages 873–881, 2011.

Avleen S. Bijral, Anand D. Sarwate, and Nathan Srebro. Data-Dependent Convergence for Consensus Stochastic Optimization. *IEEE Transactions on Automatic Control*, 62(9):4483–4498, 2017.

Olivier Bousquet and Léon Bottou. The Tradeoffs of Large Scale Learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

Olivier Bousquet and André Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

Sébastien Bubeck et al. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal Distributed Online Prediction using Mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

Aymeric Dieuleveut and Francis Bach. Nonparametric Stochastic Approximation with Large Stepsizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Alexandros G. Dimakis, Soummya Kar, José M.F. Moura, Michael G. Rabbat, and Anna Scaglione. Gossip Algorithms for Distributed Signal Processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train Faster, Generalize Better: Stability of Stochastic Gradient Descent. In *International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1225–1234, 2016.

Bjorn Johansson, Maben Rabi, and Mikael Johansson. A Simple Peer-to-Peer Algorithm for Distributed Optimization in Sensor Networks. In *Decision and Control, 2007 46th IEEE Conference on*, pages 4705–4710. IEEE, 2007.

Björn Johansson, Maben Rabi, and Mikael Johansson. A Randomized Incremental Subgradient Method for Distributed Optimization in Networked Systems. *SIAM Journal on Optimization*, 20 (3):1157–1170, 2009.

Michael Kearns and Dana Ron. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural computation*, 11(6):1427–1453, 1999.

Guanghui Lan. An Optimal Method for Stochastic Composite Optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

Junhong Lin and Volkan Cevher. Optimal Distributed Learning with Multi-pass Stochastic Gradient Methods. *Proceedings of the 35th International Conference on Machine Learning*, page 27, 2018.

Junhong Lin and Lorenzo Rosasco. Optimal Rates for Multi-Pass Stochastic Gradient Methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization Properties and Implicit Regularization for Multiple Passes SGM. In *International Conference on Machine Learning*, pages 2340–2348, 2016a.

Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative Regularization for Learning with Convex Loss Functions. *Journal of Machine Learning Research*, 17(1):2718–2755, 2016b.

Ilan Lobel and Asuman Ozdaglar. Distributed Subgradient Methods for Convex Optimization over Random Networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306, 2011.

Ion Matei and John S. Baras. Performance Evaluation of the Consensus-Based Distributed Subgradient Method Under Random Communication Topologies. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):754–771, 2011.

Aryan Mokhtari and Alejandro Ribeiro. DSA: Decentralized Double Stochastic Averaging Gradient Algorithm. *Journal of Machine Learning Research*, 17(61):1–35, 2016.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning Theory: Stability is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Angelia Nedic and Asuman Ozdaglar. Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedić, Alex Olshevsky, Asuman Ozdaglar, and John N. Tsitsiklis. On Distributed Averaging Algorithms and Quantization Effects. *IEEE Transactions on Automatic Control*, 54(11): 2506–2517, 2009.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Aalexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

S. Sundhar Ram, Angelia Nedic, and Venugopal V. Veeravalli. Distributed Subgradient Projection Algorithm for Convex Optimization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3653–3656. IEEE, 2009.

Srinivasan Sundhar Ram, Angelia Nedić, and Venugopal V. Veeravalli. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545, 2010.

William H. Rogers and Terry J. Wagner. A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules. *The Annals of Statistics*, pages 506–514, 1978.

Ali H. Sayed. Adaptive Networks. *Proceedings of the IEEE*, 102(4):460–497, 2014.

Devavrat Shah. Gossip algorithms. *Foundations and Trends® in Networking*, 3(1):1–125, 2009.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, Stability and Uniform Convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

Ohad Shamir and Nathan Srebro. Distributed Stochastic Optimization and Learning. In *Communication, Control, and Computing (Allerton), 2014*, pages 850–857. IEEE, 2014.

Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-Efficient Distributed Optimization using an Approximate Newton-Type Method. In *International Conference on Machine Learning*, pages 1000–1008, 2014.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An Exact First-Order Algorithm for Decentralized Consensus Optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Pierre Tarres and Yuan Yao. Online Learning as Stochastic Approximation of Regularization Paths: Optimality and Almost-Sure Convergence. *IEEE Transactions on Information Theory*, 60(9): 5716–5735, 2014.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms. *IEEE transactions on automatic control*, 31 (9):803–812, 1986.

John Nikolas Tsitsiklis. Problems in Decentralized Decision Making and Computation. Technical report, Massachusetts Inst Of Tech Cambridge Lab For Information And Decision Systems, 1984.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

Lin Xiao. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

Mu Yang and Choon Yik Tang. Distributed Estimation of Graph Spectrum. In *2015 American Control Conference (ACC)*, pages 2703–2708. IEEE, 2015.

Peng Yang, Randy A Freeman, Geoffrey J Gordon, Kevin M Lynch, Siddhartha S Srinivasa, and Rahul Sukthankar. Decentralized Estimation and Control of Graph Connectivity for Mobile Sensor Networks. *Automatica*, 46(2):390–396, 2010.

Yiming Ying and Massimiliano Pontil. Online Gradient Descent Learning Algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Yuchen Zhang and Xiao Lin. DiSCO: Distributed Optimization for Self-concordant Empirical Loss. In *International Conference on Machine Learning*, pages 362–370, 2015.

Yuchen Zhang, Martin J. Wainwright, and John C. Duchi. Communication-Efficient Algorithms for Statistical Optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.