A Unified Theory of Decentralized SGD with Changing Topology and Local Updates

Anastasia Koloskova [*1] Nicolas Loizou [2] Sadra Boreiri [1] Martin Jaggi [1] Sebastian U. Stich [*1]

Abstract

Decentralized stochastic optimization methods have gained a lot of attention recently, mainly because of their cheap per iteration cost, data locality, and their communication-efficiency. In this paper we introduce a unified convergence analysis that covers a large variety of decentralized SGD methods which so far have required different intuitions, have different applications, and which have been developed separately in various communities.

Our algorithmic framework covers local SGD updates and synchronous and pairwise gossip updates on adaptive network topology. We derive universal convergence rates for smooth (convex and non-convex) problems and the rates interpolate between the heterogeneous (non-identically distributed data) and iid-data settings, recovering linear convergence rates in many special cases, for instance for over-parametrized models. Our proofs rely on weak assumptions (typically improving over prior work in several aspects) and recover (and improve) the best known complexity results for a host of important scenarios, such as for instance coorperative SGD and federated averaging (local SGD).

1. Introduction

Training machine learning models in a non-centralized fashion can offer many advantages over traditional centralized approaches in core aspects such as data ownership, privacy, fault tolerance and scalability. In efforts to depart from the traditional parameter server paradigm (Dean et al., 2012), federated learning (Konečnỳ et al., 2016; McMahan et al., 2016; 2017; Kairouz et al., 2019) has emerged, but also fully decentralized approaches have been suggested recently—though yet still at a smaller scale than federated learning (Lian et al., 2017; Assran et al., 2019; Koloskova et al., 2020). However, the community has identified a host of challenges that come along with decentralized training: notably, high communication cost (Tang et al., 2018a; Wang et al., 2019; Koloskova et al., 2019), a need for time-varying topologies (Nedić & Olshevsky, 2014; Assran et al., 2019) and data-heterogeneity (Li et al., 2018; Karimireddy et al., 2019; Li et al., 2020a;b). It is imperative to have a good theoretical understanding of decentralized stochastic gradient descent (SGD) to predict the training performance of SGD in these scenarios and to assist the design of optimal decentralized training schemes for machine learning tasks.

In contrast to the centralized setting, where the convergence of SGD is well understood (Bach & Moulines, 2011; Rakhlin et al., 2012; Dekel et al., 2012), the analyses of SGD in non-centralized settings are often application specific and have been historically developed separately in different communities, besides some recent efforts towards a unified theory. Notably, Wang & Joshi (2018) propose a framework for decentralized optimization with non-heterogeneous data and Li et al. (2019) study decentralized SGD for non-convex heterogeneous settings. We here propose a significantly extended framework that covers these previously proposed ones as special cases.

We provide tight convergence rates for a large family of decentralized SGD variants. Proving convergence rates in a unified framework is much more powerful than studying individual special cases on their own: We are not only able to recover many existing analyses and results, we can also often show improved rates under more general setting. Remarkably, for instance for local SGD (Zinkevich et al., 2010; Stich, 2019b; Patel & Dieuleveut, 2019) we show improved rates for the convex and strongly-convex case and recover the best known rates for the non-convex case under weaker assumptions than assumed in prior work (highlighted in Table 1).

1.1. Contributions

- We present a unified framework for gossip based decentralized SGD methods that captures local updates and time-varying, randomly sampled, mixing distributions.

[*] Equal contribution [1] EPFL, Lausanne, Switzerland [2] Mila and DIRO, Université de Montréal, Canada. Correspondence to: Anastasia Koloskova <anastasia.koloskova@epfl.ch>, Sebastian U. Stich <sebastian.stich@epfl.ch>.

Preprint. Copyright 2020 by the author(s).

Our framework covers a rich class of methods that previously needed individual convergence analyses.

- Our theoretical results rely on weak assumptions that measure the strength of the noise and the dissimilarity of the functions between workers and a novel assumption on the expected mixing rate of the gossip algorithm. This provides us with great flexibility on how to select the topology of the network and the mixing weights.

- We demonstrate the effectiveness and tightness of our results by exemplary showing that our framework gives the best convergence rates for local SGD for both, heterogeneous and iid. data settings, improving over all previous analyses on convex functions.

- We provide a lower bound that confirms that our convergence rates are tight on strongly convex functions.

- We empirically verify the tightness of our theoretical results on strongly convex functions and explain the impact of noise and data diversity on the convergence.

2. Related Work

The study of decentralized optimization algorithms can be tracked back at least to (Tsitsiklis, 1984). For the problem of computing aggregates (finding consensus) among clients, various gossip-based protocols have been proposed. For instance the push-sum algorithm (Kempe et al., 2003), based on the intuition of mixing in Markov chains and allowing for asymmetric communication, or the symmetric randomized gossip protocol for averaging over arbirary graphs (Xiao & Boyd, 2004; Boyd et al., 2006) that we follow closely in this work. For general optimization problems, the most common algorithms are either combinations of standard gradient based methods with gossip-type averaging step (Nedić & Ozdaglar, 2009; Johansson et al., 2010), or specifically designed methods relying on problem structure, such as alternating direction method of multipliers (ADMM) (Wei & Ozdaglar, 2012; Iutzeler et al., 2013), dual averaging (Duchi et al., 2012; Nedić et al., 2015; Rabbat, 2015), primal-dual methods (Alghunaim & Sayed, 2020), or block-coordinate methods for generalized linear models (He et al., 2018). There is a rich literature in the control community that discusses various special cases—motivated by particular applications—such as for instance asynchronity (Boyd et al., 2006) or time-varying graphs (Nedić & Olshevsky, 2014; Nedić & Olshevsky, 2016), see also (Nedić et al., 2018) for an overview.

For the deterministic (non-stochastic) descentralized optimization a recent line of work developed optimal algorithms based on acceleration (Jakovetić et al., 2014; Scaman et al., 2017; 2018; Uribe et al., 2018). In the machine learning context, decentralized implementations of stochastic gradient descent have gained a lot of attention recently (Lian et al., 2017; Tang et al., 2018b; Assran et al., 2019; Koloskova et al., 2020), especially for the particular (but not fully decentralized) case of a star-shaped network topology, the federated learning setting (Konečnỳ et al., 2016; McMahan et al., 2016; 2017; Kairouz et al., 2019). Rates for the stochastic optimization are derived in (Shamir & Srebro, 2014; Rabbat, 2015), under the assumption that the distributions on all nodes are equal. However, this is a very strong assumption for practical problems.

It has been noted quite early that decentralized gradient based methods in heterogenous data setting suffer from a 'client-drift', i.e. the diversity in the functions on each node leads to a drift on each client towards the minima of f_i—potentially far away from the global minima of f. This phenomena has been discussed (and sometimes been adressed by modifing the SGD updates) for example in (Shi et al., 2015; Lee et al., 2015; Nedić et al., 2016) and been rediscovered frequently in the context of stochastic optimization (Zhao et al., 2018; Karimireddy et al., 2019). It is important to note that in analyses based on the bounded gradient assumption—which was traditionally assumend for analyzing SGD (Lacoste-Julien et al., 2012; Rakhlin et al., 2012)—the diversity in the data distribution on each worker sometimes can be hidden in this generous upper bound and the analyses cannot distinguish between iid. and non-iid. data cases, such as e.g. in (Koloskova et al., 2019; Nadiradze et al., 2019; Li et al., 2020b). In this work, we use much weaker assumptions and we show how the convergence rate depends on the similarity between the functions (by providing matching lower and upper bounds). Our results show that in overparametrized settings no drift effects occur and linear convergence can be achieved similar as to the centralized setting (Schmidt & Roux, 2013; Needell et al., 2016; Ma et al., 2018).

For reducing communication cost, various techniques have been proposed. In this work we do not consider gradient compression techniques (Alistarh et al., 2017; Stich et al., 2018; Tang et al., 2018a; 2019; Stich & Karimireddy, 2019)—but such orthogonal techniques could be added on top of our scheme—and instead only focus on local updates steps which are often efficient in practice but challenging to handle in the theoretical analysis (McMahan et al., 2017; Stich, 2019b; Yu et al., 2019; Lin et al., 2020).

3. Setup

We study the distributed stochastic optimization problem

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

where the components $f_i \colon \mathbb{R}^d \to \mathbb{R}$ are distributed among n nodes and are given in stochastic form:

$$f_i(\mathbf{x}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} F_i(\mathbf{x}, \xi_i), \tag{2}$$

where $\mathcal{D}_i$ denotes the distribution of ξ_i over parameter space Ω_i on node i. Standard empirical risk minimization is an important special case of this problem, when each $\mathcal{D}_i$ presents a finite number m_i of elements $\{\xi_i^1, \dots, \xi_i^{m_i}\}$. Then f_i can be rewritten as $f_i(\mathbf{x}) = \frac{1}{m_i} \sum_{j=1}^{m_i} F_i(\mathbf{x}, \xi_i^j)$. In the special case of $m_i = 1$, for each $i \in [n]$, we further recover the deterministic distributed optimization problem.

It is important to note that we do not make any assumptions on the distributions $\mathcal{D}_i$. This means that we especially cover hard heterogeneous machine learning problems where data is only available locally to each worker $i \in [n] := \{1, \dots, n\}$ and the *local minima* $\mathbf{x}_i^\star := \arg\min_{\mathbf{x} \in \mathbb{R}^d} f_i(\mathbf{x})$, can be far away from the global minimizer of (1). This covers a host of practically relevant problems over decentralized training data, as in federated learning (motivated by privacy), or large datasets stored across datacenters or devices (motivated by scalability). We will discuss several important examples in Section 3.2 below.

3.1. Assumptions on the objective function f

For all our theoretical results we assume that f is smooth.

Assumption 1a (L-smoothness)**.** *Each function $F_i(\mathbf{x}, \xi) \colon \mathbb{R}^d \times \Omega_i \to \mathbb{R}$, $i \in [n]$ is differentiable for each $\xi \in \operatorname{supp}(\mathcal{D}_i)$ and there exists a constant $L \geq 0$ such that for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \operatorname{supp}(\mathcal{D}_i)$:*

$$\|\nabla F_i(\mathbf{y}, \xi) - \nabla F_i(\mathbf{x}, \xi)\| \leq L \|\mathbf{x} - \mathbf{y}\| \,. \tag{3}$$

Sometimes it will be enough to just assume smoothness of f_i instead.

Assumption 1b (L-smoothness)**.** *Each function $f_i(\mathbf{x}) \colon \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$ is differentiable and there exists a constant $L \geq 0$ such that for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:*

$$\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\| \leq L \|\mathbf{x} - \mathbf{y}\| \,. \tag{4}$$

Remark 1. *Clearly, Assumption 1b is more general than Assumption 1a. Moreover, for convex $F(\mathbf{y}, \xi)$ Assumption 1a implies Assumption 1b (Nesterov, 2004).*

Assumption 1b is quite common in the literature (e.g. (Lian et al., 2017; Wang & Joshi, 2018)) but sometimes also the stronger Assumption 1a is assumed (Nguyen et al., 2018). We here use this version in the convex case only, to allow for a more general assumption on the noise instead (see Section 3.2 below).

For some of the derived results we need in addition convexity. Specifically, μ-convexity for a parameter $\mu \geq 0$.

Assumption 2 (μ-convexity)**.** *Each function $f_i \colon \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$ is μ-(strongly) convex for constant $\mu \geq 0$. That is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:*

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \leq \langle \nabla f_i(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle \,. \tag{5}$$

3.2. Assumptions on the noise

We now formulate our conditions on the noise. For the convergence analysis of SGD on smooth convex functions it is typically enough to assume a bound on the noise at the optimum only (Needell et al., 2016; Bottou et al., 2018; Gower et al., 2019; Stich, 2019a). Similarly, to express the diversity of the functions f_i in the convex case it is sufficient to measure it only at the optimal point $\mathbf{x}^\star$ (such a point always exists for strongly convex functions).

Assumption 3a (Bounded noise at the optimum)**.** *Let $\mathbf{x}^\star = \arg\min f(\mathbf{x})$ and define*

$$\zeta_i^2 := \|\nabla f_i(\mathbf{x}^\star)\|_2^2 \,, \qquad \bar{\zeta}^2 := \tfrac{1}{n} \textstyle\sum_{i=1}^n \zeta_i^2 \,. \tag{6}$$

Further, define

$$\sigma_i^2 := \mathbb{E}_{\xi_i} \|\nabla F_i(\mathbf{x}^\star, \xi_i) - \nabla f_i(\mathbf{x}^\star)\|_2^2 \tag{7}$$

and similarly as above, $\bar{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n \sigma_i^2$. We assume that $\bar{\sigma}^2$ and $\bar{\zeta}^2$ are bounded.

Here, $\bar{\sigma}^2$ measures the noise level, and $\bar{\zeta}^2$ the diversity of the functions f_i. If all functions are identical, $f_i = f_j$, for all i, j, then $\bar{\zeta}^2 = 0$. Many prior work in the context of stochastic decentralized optimization often assumed bounded diversity and bounded noise *everywhere* (such as e.g. (Lian et al., 2017; Tang et al., 2018b)), whereas we here only need to assume this bound locally at $\mathbf{x}^\star$.

For the non-convex case—where a unique $\mathbf{x}^\star$ does not necessarily exist—we generalize Assumption 3a to:

Assumption 3b (Bounded noise)**.** *We assume that there exists constants P, $\hat{\zeta}$ such that $\forall \mathbf{x} \in \mathbb{R}^d$,*

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla f_i(\mathbf{x})\|_2^2 \leq \hat{\zeta}^2 + P \|\nabla f(\mathbf{x})\|_2^2 \tag{8}$$

and constants M, $\hat{\sigma}$ such that $\forall \mathbf{x}_1, \dots \mathbf{x}_n \in \mathbb{R}^d$

$$\Psi \leq \hat{\sigma}^2 + \tfrac{M}{n} \textstyle\sum_{i=1}^n \|\nabla f(\mathbf{x}_i)\|_2^2 \tag{9}$$

where $\Psi := \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\xi_i} \|\nabla F_i(\mathbf{x}_i, \xi_i) - \nabla f_i(\mathbf{x}_i)\|_2^2$.

We see that Assumption 3a is weaker than Assumption 3b as it only needs ho hold for $\mathbf{x}_i = \mathbf{x}^\star$. Further, it is important to note that we do not assume a uniform bound on the variance (as many prior work, such as Li et al., 2019; Tang et al., 2018b; Lian et al., 2017; Assran et al., 2019) but instead allow the bound on the noise and the diversity to grow with the gradient norm (similar assumptions are common in the convex setting (Bottou et al., 2018)).

Discussion. We now show that the Assumption 3b is weaker than assuming a uniform upper bound on the noise. The uniform variance bound is given as

$$\mathbb{E}\left\|\nabla F_i(\mathbf{x},\xi_i)-\nabla f_i(\mathbf{x})\right\|_2^2\leq\sigma_{\text{unif}}^2\,,\qquad\forall\mathbf{x}\in\mathbb{R}^d$$

similarly for the similarity of functions between nodes

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n\mathbb{E}\left\|\nabla f_i(\mathbf{x})-\nabla f(\mathbf{x})\right\|_2^2\leq\bar{\zeta}_{\text{unif}}^2\,,\quad\forall\mathbf{x}\in\mathbb{R}^d\,.$$

By recalling the inequality $\|\mathbf{a}+\mathbf{b}\|^2\leq 2\|\mathbf{a}\|^2+2\|\mathbf{b}\|^2$ for $\mathbf{a},\mathbf{b}\in\mathbb{R}^d$, it is easy to check that these two bounds imply Assumption 3b with $B=2$, $M=0$, $\hat{\sigma}^2=\sigma_{\text{unif}}^2$ and $\hat{\zeta}^2=2\bar{\zeta}_{\text{unif}}^2$. Thus, our assumptions are weaker and $\hat{\zeta}$ and $\hat{\sigma}$ can be much smaller than $\bar{\zeta}_{\text{unif}}^2,\sigma_{\text{unif}}^2$ in general.

A second common assumption is to assume that the (stochastic) gradients are uniformly bounded (e.g. Koloskova et al., 2019; Li et al., 2020b), that is

$$\mathbb{E}\left\|\nabla F_i(\mathbf{x},\xi_i)\right\|_2^2\leq G^2\,,$$

for a constant G. Under the bounded gradient assumption, Assumption 3b is clearly satisfied, as all terms on the left hand side of (8) and (9) can be upper bounded by $2G^2$.

3.3. Notation

We use the notation $\mathbf{x}_i^{(t)}$ to denote the iterates on node i at time step t. We further define the average

$$\bar{\mathbf{x}}^{(t)}:=\tfrac{1}{n}\textstyle\sum_{i=1}^n\mathbf{x}_i^{(t)}\,. \tag{10}$$

We use both vector and matrix notation whenever it is more convenient, and define

$$X^{(t)}:=\left[\mathbf{x}_1^{(t)},\dots,\mathbf{x}_n^{(t)}\right]\in\mathbb{R}^{d\times n} \tag{11}$$

and likewise define $\bar{X}^{(t)}:=\left[\bar{\mathbf{x}}^{(t)},\dots,\bar{\mathbf{x}}^{(t)}\right]\equiv X^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

4. Decentralized (Gossip) SGD

We now present the generalized decentralized SGD framework. Similar to existing works (Lian et al., 2017; Wang & Joshi, 2018; Li et al., 2019) our proposed method allows only decentralized communications. That is, the exchange of information (through *gossip* averaging) can only occur between connected nodes (neighbors). The algorithm (outlined in Algorithm 1) consists of two phases: (i) stochastic gradient updates, performed locally on each worker (lines 4–5), followed by a (ii) consensus operation, where nodes average their values with their neighbors (line 6).

The gossip averaging protocol can be compactly written in matrix notation, with $\mathcal{N}_i^{(t)}:=\{j\colon w_{ij}^{(t)}>0\}$ denoting the neighbors of node i at iteration t:

$$X^{(t+1)}=X^{(t)}W^{(t)}\quad\Leftrightarrow\quad\mathbf{x}_i^{(t+1)}=\textstyle\sum_{j\in\mathcal{N}_i^{(t)}}w_{ij}^{(t)}\mathbf{x}_j^{(t)}\,,$$

where the mixing matrix $W^{(t)}\in[0,1]^{n\times n}$ encodes the network structure at time t and the averaging weights (nodes i and j are connected if $w_{ij}^{(t)}>0$).

Our scheme shows great flexibility as the mixing matrices can change over iterations and moreover can be selected from (changing) distributions.

Definition 1 (Mixing matrix). *A symmetric ($W=W^\top$) doubly stochastic ($W\mathbf{1}=\mathbf{1}$, $\mathbf{1}^\top W=\mathbf{1}^\top$) matrix $W\in[0,1]^{n\times n}$.*

4.1. Algorithm

Algorithm 1 DECENTRALIZED SGD

```
input  for each node $i\in[n]$ initialize $\mathbf{x}_i^{(0)}\in\mathbb{R}^d$,
       stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations $T$,
       mixing matrix distributions $\mathcal{W}^{(t)}$ for $t\in[0,T]$
1: for $t$ in $0\dots T$ do
2:    Sample $W^{(t)}\sim\mathcal{W}^{(t)}$
3:    In parallel (task for worker $i$, $i\in[n]$)
4:    Sample $\xi_i^{(t)}$, compute $\mathbf{g}_i^{(t)}:=\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})$
5:    $\mathbf{x}_i^{(t+\frac{1}{2})}=\mathbf{x}_i^{(t)}-\eta_t\mathbf{g}_i^{(t)}$    ▷ stochastic gradient updates
6:    $\mathbf{x}_i^{(t+1)}:=\sum_{j\in\mathcal{N}_i^t}w_{ij}^{(t)}\mathbf{x}_j^{(t+\frac{1}{2})}$    ▷ gossip averaging
7: end for
```

In each iteration in Algorithm 1 a new mixing matrix $W^{(t)}$ is sampled from a possibly time-varying distribution $\mathcal{W}^{(t)}$, $t\in\{0,\dots,T\}$ (we will show below that also degenerate mixing matrices, for instance $W^{(t)}=\mathbf{I}_n$ which implies *no communication* in round t, are possible choices). We will discuss several important instances below, but first we now state our assumption on the quality of the mixing matrices. This assumption is novel in the literature to the best of our knowledge and a natural generalization of earlier versions.

4.2. New assumption on mixing matrices

We recall that for randomized gossip averaging with a randomly sampled mixing matrix $W\sim\mathcal{W}$ it holds

$$\mathbb{E}_W\left\|XW-\bar{X}\right\|_F^2\leq(1-p)\left\|X-\bar{X}\right\|_F^2\,, \tag{12}$$

for a value $p\geq 0$ (related to the spectrum of $\mathbb{E}\,W^\top W$), that is, the averaging step brings the values in the columns of $X\in\mathbb{R}^{d\times n}$ closer to their row-wise average $\bar{X}:=X\cdot\frac{1}{n}\mathbf{1}\mathbf{1}^\top$ in expectation (see e.g. Boyd et al., 2006).

In our analysis it will be enough to *assume* that a property similar to (12) holds for the *composition* of mixing matrixes, and does not necessarily hold for every single step.

Assumption 4 (Expected Consensus Rate). *We assume that there exists two constants $p \in (0, 1]$ and integer $\tau \geq 1$ such that for all matrices $X \in \mathbb{R}^{d \times n}$ and all integers $\ell \in \{0, \dots, T/\tau\}$,*

$$\mathbb{E}_W \left\|XW_{\ell,\tau} - \bar{X}\right\|_F^2 \leq (1-p) \left\|X - \bar{X}\right\|_F^2 , \qquad (13)$$

where $W_{\ell,\tau} = W^{((\ell+1)\tau-1)} \dots W^{(\ell\tau)}$ and $\bar{X} := X \frac{\mathbf{1}\mathbf{1}^\top}{n}$ and $\mathbb{E}$ is taken over the distributions $W^{(t)} \sim \mathcal{W}^{(t)}$ and indices $t \in \{\ell\tau, \dots, (\ell+1)\tau - 1\}$.

It is crucial to observe that this assumption does not require every realization W to satisfy a decrease property as for the standard analysis, it is enough if it holds over the concatenation of τ mixing steps. This assumption differs from the connectivity assumptions sometimes used in the control community. For example Nedić & Olshevsky (2014) require strong connectivity of the graph after every τ steps, whereas we here do not require this (for example, even sampling one single random edge leads to a positive decrease in expectation, whereas to ensure connectivity one would need to perform $\Omega(n)$ pairwise communications). This means that our bounds are typically much tighter that bounds derived on the strong connectivity assumption. However, as we require W to be symmetric, our setting is less general than the one considered in (Nedić et al., 2017; Xi & Khan, 2017; Saadatniaki et al., 2018; Assran & Rabbat, 2018; Scutari & Sun, 2019; Assran et al., 2019).

Commonly used weights are for instance the Metropolis-Hastings weights $w_{ij} = w_{ji} = \min\{\frac{1}{\deg(i)+1}, \frac{1}{\deg(j)+1}\}$ for $(i, j) \in E$, see also (Xiao & Boyd, 2004; Boyd et al., 2006) for further guidelines. With these weights, the values of p for commonly used graphs are $p = 1$ for the complete graph, $p = \Theta\big(\frac{1}{n}\big)$ for 2-d torus on n nodes, and $p = \Theta\big(\frac{1}{n^2}\big)$ for a cycle on n nodes. Intuitively, $p^{-1/2}$ correlates with the diameter of the graph and is related to the mixing time of Markov chains. A commonly studied randomized scheme is the pairwise random gossip algorithm (Boyd et al., 2006; Loizou & Richtárik, 2019), where one edge at a time is sampled from an underlying graph $\mathcal{G} = ([n], E)$, i.e. the a random mixing matrix $\mathbf{Z}_{i,j} := \mathbf{I}_n - \frac{1}{2}(e_i - e_j)(e_i - e_j)^\top$, for all edges in the graph $(i, j) \in E$, where $e_i \in \mathbb{R}^n$ is the i^{th} coordinate vector. In this case $p = \rho(\mathcal{G})/|E|$, where $\rho(\mathcal{G})$ denotes the algebraic connectivity of the network (Fiedler, 1973; Boyd et al., 2006; Loizou & Richtárik, 2016). For example, with the complete graph as base graph, pairwise gossip attains $p = \Theta\big(\frac{1}{n^2}\big)$, i.e. enjoys equally fast mixing as averaging over a (fixed) cycle (which requires n pairwise communications per round).

5. Examples Covered in the Framework

Our framework is very general and covers many special cases previously introduced in the literature.

5.1. Fixed Sampling Distribution ($\tau = 1$, $\mathcal{W}^{(t)} \equiv \mathcal{W}$)

The simplest instances of Algorithm 1 arise when the mixing matrix W is kept constant over the iterations. By choosing the fully connected matrix $W = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ we recover • **centralized mini-batch SGD** (Dekel et al., 2012) and by choosing an arbitrary connected W, we recover • **decentralized SGD** (Lian et al., 2017).

To reduce communication overheads, it has been proposed to choose sparse (not necessarily connected) subgraphs of the network topology. For instance in • **MATCHA** (Wang et al., 2019) it is proposed to sample edges from a matching decomposition of the underlying network topology, therefore allowing for pairwise communications between nodes. Whilst no explicit values of p were given for this approach, for the simpler instance of • **pairwise randomized gossip** (Boyd et al., 2006; Ram et al., 2010; Lee & Nedić, 2015; Loizou & Richtárik, 2019) we have $p = \Theta\big(\frac{1}{n^2}\big)$, thus by sampling a linear number of (independent) edges—not necessarily a matching—we approximately have $p = \Theta\big(\frac{1}{n}\big)$ for this • **repeated pairwise randomized gossip** variant (and expect roughly the same parameter for matchings). This approach can be generalized to • **randomized subgraph gossip**, where a subgraph of the base topology is selected for averaging. A special case of this is • **clique gossip** (Liu et al., 2019), or an alternative variant is to • **sample from a fixed set of communication topologies** (known to all decentralized) workers.

One noteably instance of this type is • **loopless local decentralized SGD** where the mixing matrix is (a fixed) W with probability $\frac{1}{\tau}$, and $\mathbf{I}_n$ with probability $1 - \frac{1}{\tau}$, for a parameter $\tau \geq 1$. This algorithm mimicks the behavior of the local SGD (see subsection below), commonly analyzed for $W = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ only, but the loopless variant is much easier to analyze (with p decreased by a factor of τ).

5.2. Periodic Sampling ($\tau > 1$, $\mathcal{W}^{(t)} \equiv \mathcal{W}^{(t+\tau)}$)

Our analysis covers the empirical (finite-sample) versions of the aforementioned algorithms, for instance • **alternating decentralized SGD** that sweeps through τ fixed mixing matrices. A special algorithm of this type is • **local SGD** (Coppola, 2015; Zhou & Cong, 2018; Stich, 2019b) where averaging on the complete graph is performed every τ iterations and only local steps are performed otherwise (mixing matrix $\mathbf{I}_n$ for $\tau - 1$ steps).

Our analysis covers also natural extensions such as • **decentralized local SGD** where mixing is performed with an arbitrary matrix W, and • **random decentralized local SGD** where the mixing matrix is sampled from a distribution. More generally, our framework also allows to combine local steps with all of the examples described in the previous section.

5.3. Non-Periodic Sampling

It is not necessary to have a periodic structure, it is sufficient that the composition of every τ consecutive mixing matrixes satisfies Assumption 4. For instance as in • **distriributed SGD over time-varying graphs** (Nedić & Olshevsky, 2014).

5.4. Other Frameworks

In contrast to many prior works, we here allow the topology and the averaging weights to change between iterations. Our framework covers • **Cooperative SGD** (Wang & Joshi, 2018) which considers only the IID data case ($f_i = f_j$) with local updates and a fixed mixing matrix W, and the recently proposed • **periodic decentralized SGD** (Li et al., 2019) that allows for multiple local update and multiple mixing steps (for fixed W) in a periodic manner. None of these work considered sampling of the mixing matrix and do only provide rates for non-convex functions.

6. Convergence Result

In this section we present the convergence results for decentralized SGD variants that fit the template of Algorithm 1.

6.1. Complexity Estimates (Upper Bounds)

Theorem 2. *For schemes as in Algorithm 1 with mixing matrices such as in Assumption 4 and any target accuracy $\epsilon > 0$ there exists a (constant) stepsize (potentially depending on ϵ) such that the accuracy can be reached after at most the following number of iterations T:*
Non-Convex: *Under Assumption 1b and 3b, it holds $\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \leq \epsilon$ after*

$$\mathcal{O}\left(\frac{\hat\sigma^2}{n\epsilon^2} + \frac{\hat\zeta\tau\sqrt{M+1} + \hat\sigma\sqrt{p\tau}}{p\epsilon^{3/2}} + \frac{\tau\sqrt{(P+1)(M+1)}}{p\epsilon}\right)\cdot LF_0$$

iterations. If we in addition assume convexity,
Convex: *Under Assumption 1a, 3a and 2 for $\mu \geq 0$, the error $\frac{1}{(T+1)}\sum_{t=0}^{T}(\mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - f^\star) \leq \epsilon$ after*

$$\mathcal{O}\left(\frac{\bar\sigma^2}{n\epsilon^2} + \frac{\sqrt{L}(\bar\zeta\tau + \bar\sigma\sqrt{p\tau})}{p\epsilon^{3/2}} + \frac{L\tau}{p\epsilon}\right)\cdot R_0^2$$

iterations, and if $\mu > 0$,
Strongly-Convex: *then $\sum_{t=0}^{T}\frac{w_t}{W_T}(\mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - f^\star) + \mu\,\mathbb{E}\|\bar{\mathbf{x}}^{(T+1)} - \mathbf{x}^\star\|^2 \leq \epsilon$ for*[1]

$$\tilde{\mathcal{O}}\left(\frac{\bar\sigma^2}{\mu n\epsilon} + \frac{\sqrt{L}(\bar\zeta\tau + \bar\sigma\sqrt{p\tau})}{\mu p\sqrt{\epsilon}} + \frac{L\tau}{\mu p}\right)$$

iterations, for positive weights w_t and $F_0 := f(\mathbf{x}_0) - f^\star$ and $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|$ denote the initial errors.

6.2. Lower Bound

We now show that the terms depending on $\bar\zeta$ are necessary for the strongly convex setting and cannot be removed by an improved analysis.

Theorem 3. *For $n > 1$ there exists strongly convex and smooth functions $f_i \colon \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$ with $L = \mu = 1$ and without stochastic noise ($\bar\sigma^2 = 0$), such that Algorithm 1 for every constant mixing matrix $W^{(t)} \equiv W$ with $p < 1$ (see Assumption 4), hence with $\tau = 1$, requires*

$$T = \tilde\Omega\left(\frac{\bar\zeta}{\sqrt{\epsilon p}}\right)$$

iterations to converge to accuracy ϵ.

6.3. Discussion

Exemplary, we focus in our discussion on the strongly convex case only. For strongly convex functions we prove that the expected function value suboptimality decreases as

$$\tilde{\mathcal{O}}\left(\frac{\bar\sigma^2}{n\mu T} + \frac{L(\tau^2\bar\zeta^2 + \tau p\bar\sigma^2)}{\mu^2 p^2 T^2} + \frac{L\tau R_0^2}{p}\exp\left[-\frac{\mu T p}{\tau L}\right]\right)$$

where T denotes the iteration counter. We now argue that this rate is optimal up to acceleration.

Stochastic Terms. If $\bar\sigma^2 > 0$ the convergence rate is asymptotically dominated by the first term, which cannot be further improved for stochastic methods (Nemirovsky & Yudin, 1983). We observe that the dominating first term indicates a linear speedup in the number of workers n, and no dependence on the number of local steps τ, the mixing parameter p or the dissimilarity parameter $\bar\zeta^2$. This means that decentralized SGD methods are ideal for the optimization in the high-noise regime even when network connectivity is low and number of local steps is large (see also (Chaturapruek et al., 2015) and recent work (Pu et al., 2019)). In our rates the variance $\bar\sigma^2$ does also show up in the second term, but affects the convergence only mildly (for $T = \Omega(\tau n/p)$ this second term gets dominated by the first one).

Optimization Terms. Even when $\bar\sigma^2 = 0$, the convergence of decentralized SGD only *sublinear* when $\bar\zeta^2 > 0$:[2]

$$\mathcal{O}\left(\frac{L\tau^2\bar\zeta^2}{\mu^2 p^2 T^2} + \frac{L\tau R_0}{\mu p}\log\left(\frac{1}{\epsilon}\right)\right).$$

The dependence on the dissimilarity $\bar\zeta^2$ cannot be removed in general as we show in Theorem 3. These results show

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

[2] Except for the special case when $p = 1$ (fully connected graph, such as for mini-batch SGD). In this case the rate does not depend on $\bar\zeta^2$. We detail this (known result) in the appendix.

Table 1. Comparison of convergence rates for Local SGD in non-iid settings, most recent results. We improve over the convex results, and recover the non-convex rate of Li et al. (2019).

Reference		convergence to ϵ-accuracy	
	strongly convex	convex	non-convex
Li et al. (2020b)	$\mathcal{O}\left(\frac{\bar{\sigma}^2}{n\mu^2\epsilon} + \frac{\tau^2\bar{\zeta}^2}{\mu^2\epsilon}\right)^a$	-	-
Khaled et al. (2020)	-	$\mathcal{O}\left(\frac{\bar{\sigma}^2+\bar{\zeta}^2}{n\epsilon^2} + \frac{\sqrt{L}\tau(\bar{\zeta}+\bar{\sigma})}{\epsilon^{3/2}} + \frac{L\tau}{\epsilon}\right)$	-
Li et al. (2019)	-	-	$\mathcal{O}\left(\frac{L\bar{\sigma}^2}{n\epsilon^2} + \frac{L(\tau\bar{\zeta}+\sqrt{\tau}\bar{\sigma})}{\epsilon^{3/2}} + \frac{L\tau}{\epsilon}\right)$
this work	$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{n\mu\epsilon} + \frac{\sqrt{L}(\tau\bar{\zeta}+\sqrt{\tau}\bar{\sigma})}{\mu\sqrt{\epsilon}} + \kappa\tau\right)$	$\mathcal{O}\left(\frac{\bar{\sigma}^2}{n\epsilon^2} + \frac{\sqrt{L}(\tau\bar{\zeta}+\sqrt{\tau}\bar{\sigma})}{\epsilon^{3/2}} + \frac{L\tau}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L\hat{\sigma}^2}{n\epsilon^2} + \frac{L(\tau\hat{\zeta}+\sqrt{\tau}\hat{\sigma})}{\epsilon^{3/2}} + \frac{L\tau}{\epsilon}\right)$

[a] The paper relies on slightly different assumptions (bounded gradients and different measure of dissimilarity). For better comparison of the rates we write here $\bar{\zeta}^2$ instead (which is strictly smaller than their parameter).

that decentralized SGD methods without additional modifications (see also Shi et al., 2015; Karimireddy et al., 2019) cannot converge linearly.

We can further observe see that the rates only depend on the ratio p/τ, but not on p or τ individually. This also means that the rates for local variants of decentralized SGD are the same as for their loopless variants (when the mixing is performed with probability $\frac{1}{\tau}$ only). The error term depending on R_0^2 vanishes exponentially fast, as expected for SGD methods (Bach & Moulines, 2011). The linear dependence on $\frac{L}{\mu p}$ is expected here, as we use non-accelerated first order schemes and standard gossip. This term could potentially be improved to $\left(\frac{L}{\mu p}\right)^{1/2}$ with acceleration techniques, such as in (Scaman et al., 2017). The linear dependence on τ cannot further be improved in general. This follows from the lower bound for the communication complexity of distributed convex optimization (Arjevani & Shamir, 2015), as the number of communication rounds is at most $\frac{T}{\tau}$ (no communication happens during the local steps). However, when $\bar{\zeta}^2 = 0$ (as for instance the case for identical functions f_i on each worker), this lower bound becomes vacuous and improvement of the dependence on τ might be possible (which we cannot not exploit here).

Linear Convergence for Overparametrized Settings. In overparametrized problems, there exists always $\mathbf{x}^\star$ s.t. $\|\nabla f_i(\mathbf{x}^\star)\|^2 = 0$, that is $\bar{\sigma}^2 = 0$ and $\bar{\zeta}^2 = 0$. We prove here that decentralized SGD converges linearly in this case, similarly to mini-batch SGD (Bach & Moulines, 2011; Schmidt & Roux, 2013; Needell et al., 2016; Ma et al., 2018; Gower et al., 2019; Loizou et al., 2020).

7. Special Cases: Highlights

Our rates apply to all the examples discussed in Section 5 and of course we could design even more variants and combinations of these schemes. This gives great flexibility in designing new schemes and algorithms for future applications. We leave the exploration of the trade-offs in these approaches for future work, and highlight here only a few special cases that could be of particular interest.

7.1. Best Rates for Local SGD

Local SGD is a simplified version of the federated averaging algorithm (McMahan et al., 2016; 2017) and has recently attracted the attention of the theoretical community in the seek of the best convergence rates (Stich, 2019b; Wang & Joshi, 2018; Yu et al., 2019; Basu et al., 2019; Patel & Dieuleveut, 2019; Stich & Karimireddy, 2019; Li et al., 2019; Khaled et al., 2020). Our work extends this chain and improves previous best results for convex settings and recovers the results of Li et al. (2019) in the non-convex case as we highlight in Table 1. We point out that all these rates are still dominated by large-batch SGD and do not match the lower bounds established in (Woodworth et al., 2018) (for the iid. case $\bar{\zeta}^2 = 0$). See also recent parallel work in (Woodworth et al., 2020). Whilst these previous analysis were often specifically tailored and only applicable to the mixing structure in local SGD, our analysis is much more general and tighter at the same time.

7.2. Comparison to Recent Frameworks

We mentioned major differences to other frameworks in Section 5.4 above already. Our results for the non-convex case recover the best results from (Wang & Joshi, 2018) for the iid. case[3] ($\hat{\zeta}^2 = 0$) and the non-iid. case from (Li et al., 2019) for their specific settings. We point out that our results also cover the convex setting and deterministic setting.

[3] These results can be recovered by optimizing the stepsize in (Wang & Joshi, 2018, Theorem 1) directly, instead of resorting to the worse rate stated in (Wang & Joshi, 2018, Corollary 1).

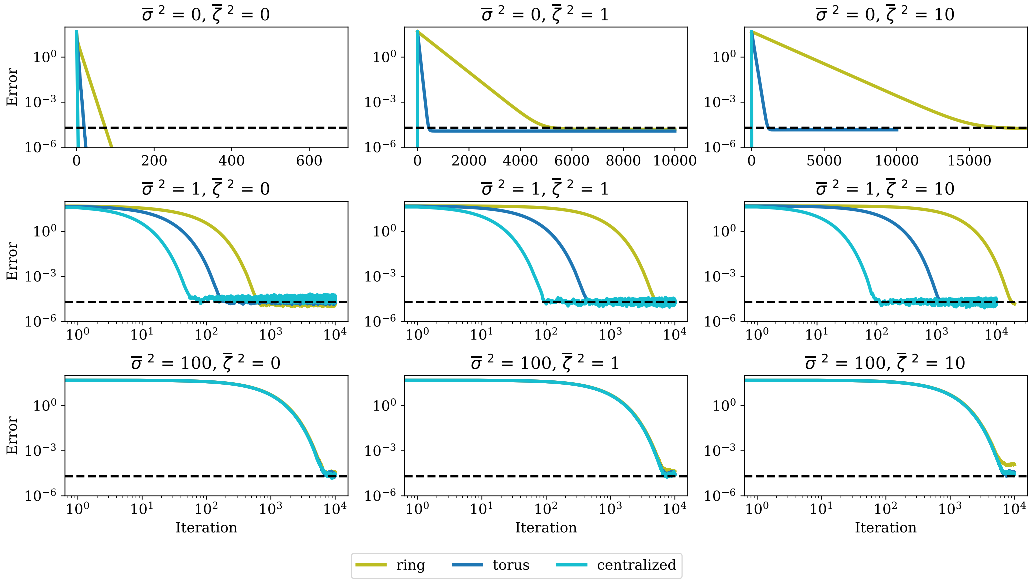

Figure 1. Convergence of $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\right\|_2^2$ to target accuracy $\epsilon = 10^{-5}$ for different problem difficulty ($\bar\sigma^2$ increasing to the bottom, $\bar\zeta^2$ increasing to the right), and different topologies on $n = 25$ nodes, $d = 50$. Stepsizes were tuned for each experiment individually to reach target accuracy in as few iterations as possible.

7.3. Best Rates for Decentralized SGD

We improve best known rates of Decentralized SGD (Olshevsky et al., 2019; Koloskova et al., 2019) for strongly convex objectives and recover the best rates in the non-convex case (Lian et al., 2017).

8. Experiments

Complementing prior work that established the effectiveness of decentralized training methods (Lian et al., 2017; Assran et al., 2019) we here focus on verifying whether the numerical performance of decentralized stochastic optimization algorithms coincides with the rates predicted by theory, focusing on the strongly convex case for now.

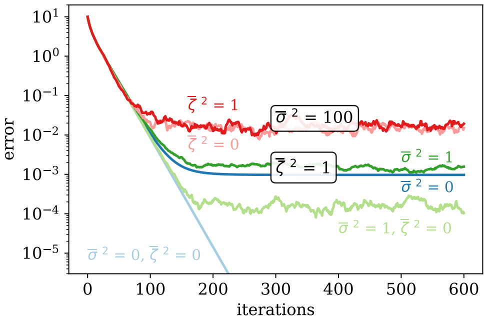

Figure 2. Problem setup. Parameters $\bar\sigma^2$ and $\bar\zeta^2$ change the noise level and the difficulty of the problem. (Here we depict $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\right\|_2^2$ on the ring with $n = 25$ nodes, $d = 10$, using fixed stepsize $\eta = 10^{-2}$ for illustration.

We consider a distributed least squares objective with $f_i(\mathbf{x}) := \frac{1}{2}\left\|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\right\|_2^2$, for fixed Hessian $\mathbf{A}_i^2 = \frac{i^2}{n} \cdot \mathbf{I}_d$ and sample each $\mathbf{b}_i \sim \mathcal{N}(0, \bar\zeta^2/i^2\mathbf{I}_d)$ for a parameter $\bar\zeta^2$, which controls the similarity of the functions (and coincides with the parameter in Assumption 3a). We control the stochastic noise $\bar\sigma^2$ by adding Gaussian noise to every stochastic gradient. We depict the effect of these parameters in Figure 2.

Setup. We consider three common network topologies, *ring*, 2-*d torus* and *fully-connected* graph and use the Metropolis-Hasting mixing matrix W, i.e. $w_{ij} = w_{ji} = \frac{1}{deg(i)+1} = \frac{1}{deg(j)+1}$ for $\{i, j\} \in E$. For all algorithms we tune the stepsize to reach a desired target accuracy ϵ with the fewest number of iterations.

Discussion of Results. In Figure 1 we depict the results. We observe that in the high noise regime (bottom row) the graph topology and the functions similarity $\bar\zeta^2$ do not impact the number of iterations needed to reach the target accuracy (the $\frac{\bar\sigma^2}{T}$ term is dominating in this regime. We also see linear rates when $\bar\sigma^2 = \bar\zeta^2 = 0$ as predicted. When increasing $\bar\zeta^2$ (in the case of $\bar\sigma^2 = 0$) we see that on the ring and torus topology the linear rate changes to a sublinear rate: even thought the curves look like straight lines, they stop converging when reaching the target accuracy (the stepsize must be further decreased to achieve higher accuracy). By comparing two top right plots, we see that for fixed topology the number of iterations increases approximately by a factor of $\sqrt{10}$ when increasing $\bar{\zeta}^2$ by a factor of 10, as one would expect from the term $\frac{\bar{\zeta}^2}{p^2T^2}$ in the convergence rate (see also Figure 3 in the appendix). The difference in number of iterations on the torus vs. ring scales approximately linear in the ratio of their mixing parameters p, (that is, $\Theta(n)$ as mentioned in Section 4.2).

9. Extensions

We presented a unifying framework for the analysis of decentralized SGD methods and provide the best known convergence guarantees. Our results show that when the noise is high, decentralized SGD methods can achieve linear speedup in the number of workers n and the convergence rate does only weakly depend on the graph topology, the number of local steps or the data heterogeneity. This shows that such methods are perfectly suited to solve stochastic optimization problems in a decentralized way. However, our results also reveal that when the noise is small (for e.g. when using large mini-batches), the effect of those parameters become more pronounced and especially function diversity can hamper the convergence of decentralized SGD methods.

Our framework can be further extended by considering gradient compression techniques (Koloskova et al., 2019) or overlapping communication steps (Assran et al., 2019; Wang et al., 2020) to additionally speedup the distributed training.

Acknowledgements

We acknowledge funding from SNSF grant 200021_175796, as well as a Google Focused Research Award. Nicolas Loizou acknowledges support by the IVADO Postdoctoral Funding Program.

References

Alghunaim, S. A. and Sayed, A. H. Linear convergence of primal-dual gradient methods and their performance in distributed optimization. *arXiv preprint arXiv:1904.01196v2*, 2020. (v2 shows improved convergence rate).

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-efficient SGD via gradient quantization and encoding. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *NIPS - Advances in Neural Information Processing Systems 30*, pp. 1709–1720. Curran Associates, Inc., 2017.

Arjevani, Y. and Shamir, O. Communication complexity of distributed convex learning and optimization. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *NIPS - Advances in Neural Information Processing Systems 28*, pp. 1756–1764. Curran Associates, Inc., 2015.

Assran, M. and Rabbat, M. Asynchronous subgradient-push. *arXiv preprint arXiv:1803.08950*, 2018.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. *ICML - Proceedings of the 36th International Conference on Machine Learning*, 2019.

Bach, F. R. and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS - Advances in Neural Information Processing Systems 24*, pp. 451–459. Curran Associates, Inc., 2011.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-SGD: Distributed SGD with quantization, sparsification, and local computations. *arXiv preprint arXiv:1906.02367*, 2019.

Bottou, L., Curtis, F., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE/ACM Trans. Netw.*, 14(SI):2508–2530, June 2006. ISSN 1063-6692. doi: 10.1109/TIT.2006.874516.

Chaturapruek, S., Duchi, J. C., and Ré, C. Asynchronous stochastic convex optimization: the noise is in the noise and SGD don't care. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *NIPS - Advances in Neural Information Processing Systems 28*, pp. 1531–1539. Curran Associates, Inc., 2015.

Coppola, G. *Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing*. PhD thesis, The University of Edinburgh, 2015.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., Le, Q. V., and Ng, A. Y. Large scale distributed deep networks. In *NIPS - Advances in Neural Information Processing Systems*, pp. 1223–1231, 2012.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 13(1):165–202, January 2012. ISSN 1532-4435.

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

Fiedler, M. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In *ICML - International Conference on Machine Learning*, pp. 5200–5209, 2019.

He, L., Bian, A., and Jaggi, M. Cola: Decentralized linear learning. In *NeurIPS - Advances in Neural Information Processing Systems 31*, pp. 4541–4551. 2018.

Iutzeler, F., Bianchi, P., Ciblat, P., and Hachem, W. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *Proceedings of the 52nd IEEE Conference on Decision and Control, CDC 2013, December 10-13, 2013, Firenze, Italy*, pp. 3671–3676. IEEE, 2013. doi: 10.1109/CDC.2013.6760448.

Jakovetić, D., Xavier, J., and Moura, J. M. F. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146, May 2014. ISSN 0018-9286. doi: 10.1109/TAC.2014.2298712.

Johansson, B., Rabi, M., and Johansson, M. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3): 1157–1170, 2010. doi: 10.1137/08073038X.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascn, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konen, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., zgr, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Kempe, D., Dobra, A., and Gehrke, J. Gossip-based computation of aggregate information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '03, pp. 482–, Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-2040-5.

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local SGD on identical and heterogeneous data. *arXiv preprint arXiv:1909.04746v2*, 2020.

Koloskova, A., Stich, S., and Jaggi, M. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML - Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3478–3487. PMLR, 2019.

Koloskova, A., Lin, T., Stich, S. U., and Jaggi, M. Decentralized deep learning with arbitrary communication compression. *ICLR - International Conference on Learning Representations*, 2020.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Lacoste-Julien, S., Schmidt, M. W., and Bach, F. R. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Lee, J. D., Lin, Q., Ma, T., and Yang, T. Distributed stochastic variance reduced gradient methods and a lower bound for communication complexity. *arXiv preprint arXiv:1507.07595*, 2015.

Lee, S. and Nedić, A. Asynchronous gossip-based random projection algorithms over networks. *IEEE Transactions on Automatic Control*, 61(4):953–968, 2015.

Li, T., Sahu, A. K., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Feddane: A federated newton-type method. *arXiv preprint arXiv:2001.01920*, 2020a.

Li, X., Yang, W., Wang, S., and Zhang, Z. Communication efficient decentralized training with multiple local updates. *arXiv preprint arXiv:1910.09126*, 2019.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of FedAvg on non-IID data. *ICLR - International Conference on Learning Representations*, openreview, 2020b.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *NIPS - Advances in Neural Information Processing Systems 30*, pp. 5330–5340. Curran Associates, Inc., 2017.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. Don't use large mini-batches, use local SGD. *ICLR - International Conference on Learning Representations*, 2020.

Liu, Y., Li, B., Anderson, B. D., and Shi, G. Clique gossiping. *IEEE/ACM Transactions on Networking*, 27(6):2418–2431, 2019.

Loizou, N. and Richtárik, P. A new perspective on randomized gossip algorithms. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 440–444. IEEE, 2016.

Loizou, N. and Richtárik, P. Revisiting randomized gossip algorithms: General framework, convergence rates and novel block and accelerated protocols. *arXiv preprint arXiv:1905.08645*, 2019.

Loizou, N., Vaswani, S., Laradji, I., and Lacoste-Julien, S. Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence. *arXiv preprint arXiv:2002.10542*, 2020.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern overparametrized learning. In *ICML - Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3325–3334. PMLR, 2018.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS - Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

Nadiradze, G., Sabour, A., Sharma, A., Markov, I., Aksenov, V., and Alistarh, D. PopSGD: Decentralized Stochastic Gradient Descent in the Population Model. *arXiv e-prints*, art. arXiv:1910.12308, Oct 2019.

Nedić, A. and Olshevsky, A. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

Nedić, A. and Olshevsky, A. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016. doi: 10.1109/TAC.2016.2529285.

Nedić, A. and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009. doi: 10.1109/TAC.2008.2009515.

Nedić, A., Lee, S., and Raginsky, M. Decentralized online optimization with global objectives and local communication. In *2015 American Control Conference (ACC)*, pp. 4497–4503, 2015. doi: 10.1109/ACC.2015.7172037.

Nedić, A., Olshevsky, A., and Shi, W. A geometrically convergent method for distributed optimization over time-varying graphs. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 1023–1029, 2016. doi: 10.1109/CDC.2016.7798402.

Nedić, A., Olshevsky, A., and Shi, W. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Nedić, A., Olshevsky, A., and Rabbat, M. G. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018. doi: 10.1109/JPROC.2018.2817461.

Needell, D., Srebro, N., and Ward, R. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Mathematical Programming*, 155(1):549–573, 2016. doi: 10.1007/s10107-015-0864-7.

Nemirovsky, A. S. and Yudin, D. B. *Problem complexity and method efficiency in optimization.* Wiley, 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization*, volume 87 of *Springer Science & Business Media*. Springer US, Boston, MA, 2004.

Nguyen, L. M., Nguyen, P. H., Richtárik, P., Scheinberg, K., Takáč, M., and van Dijk, M. New convergence aspects of stochastic gradient algorithms. *arXiv preprint arXiv:1811.12403*, 2018.

Olshevsky, A., Paschalidis, I. C., and Pu, S. A non-asymptotic analysis of network independence for distributed stochastic gradient descent. *arXiv preprint arXiv:1906.02702*, art. arXiv:1906.02702, Jun 2019.

Patel, K. K. and Dieuleveut, A. Communication trade-offs for synchronized distributed SGD with large step size. *arXiv preprint arXiv:1904.11325*, 2019.

Pu, S., Olshevsky, A., and Paschalidis, I. C. A sharp estimate on the transient time of distributed stochastic gradient descent. *arXiv preprint arXiv:1906.02702*, 2019.

Rabbat, M. Multi-agent mirror descent for decentralized stochastic optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 517–520, Dec 2015. doi: 10.1109/CAMSAP.2015.7383850.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML - Proceedings of the 29th International Conference on Machine Learning*, pp. 1571–1578. Omnipress, 2012.

Ram, S. S., Nedić, A., and Veeravalli, V. V. Asynchronous gossip algorithm for stochastic optimization: Constant stepsize analysis. In *Recent Advances in Optimization and its Applications in Engineering*, pp. 51–60. Springer, 2010.

Saadatniaki, F., Xin, R., and Khan, U. A. Optimization over time-varying directed graphs with row and column-stochastic matrices. *arXiv preprint arXiv:1810.07393*, 2018.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In Precup, D. and Teh, Y. W. (eds.), *ICML - Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3027–3036, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Scaman, K., Bach, F., Bubeck, S., Massoulié, L., and Lee, Y. T. Optimal algorithms for non-smooth distributed optimization in networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NeurIPS - Advances in Neural Information Processing Systems 31*, pp. 2745–2754. Curran Associates, Inc., 2018.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Scutari, G. and Sun, Y. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176(1-2):497–544, 2019.

Shamir, O. and Srebro, N. Distributed stochastic optimization and learning. *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 850–857, 2014.

Shi, W., Ling, Q., Wu, G., and Yin, W. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Stich, S. U. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019a.

Stich, S. U. Local SGD converges fast and communicates little. *ICLR - International Conference on Learning Representations*, art. arXiv:1805.09767, 2019b.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified SGD with memory. In *NeurIPS - Advances in Neural Information Processing Systems 31*, pp. 4452–4463. 2018.

Tang, H., Gan, S., Zhang, C., Zhang, T., and Liu, J. Communication compression for decentralized training. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NeurIPS - Advances in Neural Information Processing Systems 31*, pp. 7663–7673. Curran Associates, Inc., 2018a.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. D^2: Decentralized training over decentralized data. In Dy, J. and Krause, A. (eds.), *ICML - Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4848–4856, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018b. PMLR.

Tang, H., Lian, X., Qiu, S., Yuan, L., Zhang, C., Zhang, T., and Liu, J. Deepsqueeze: Decentralization meets error-compensated compression. *arXiv preprint arXiv:1907.07346*, 2019.

Tsitsiklis, J. N. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

Uribe, C. A., Lee, S., and Gasnikov, A. A dual approach for optimal algorithms in distributed optimization over networks. *arXiv preprint arXiv:1809.00710*, September 2018.

Wang, J. and Joshi, G. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Wang, J., Sahu, A. K., Yang, Z., Joshi, G., and Kar, S. An exact quantized decentralized gradient descent algorithm. *arXiv preprint arXiv:1905.09435*, 2019.

Wang, J., Liang, H., and Joshi, G. Overlap local-SGD: An algorithmic approach to hide communication delays in distributed SGD. *manuscript*, 2020.

Wei, E. and Ozdaglar, A. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 5445–5450, 2012. doi: 10.1109/CDC.2012.6425904.

Woodworth, B., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., and Srebro, N. Is local SGD better than minibatch SGD? *arXiv preprint arXiv:2002.07839*, 2020.

Woodworth, B. E., Wang, J., Smith, A., McMahan, H. B., and Srebro, N. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *NIPS - Advances in Neural Information Processing Systems*, pp. 8496–8506, 2018.

Xi, C. and Khan, U. A. Dextra: A fast algorithm for optimization over directed graphs. *IEEE Transactions on Automatic Control*, 62(10):4980–4993, 2017.

Xiao, L. and Boyd, S. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2004.02.022.

Yu, H., Yang, S., and Zhu, S. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zhou, F. and Cong, G. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3219–3227. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/447.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. J. Parallelized stochastic gradient descent. In *NIPS - Advances in Neural Information Processing Systems*, pp. 2595–2603, 2010.

APPENDIX
A Unified Theory of Decentralized SGD with Changing Topology and Local Updates

The appendix is organized as follows: In Section A, we rewrite Algorithm 1 equivalently in matrix notation as Algorithm 2 and give a sketch of the proof using this new notation. In Section B we state a few auxiliary technical lemmas, before giving all details for the proof of the theorem in Sections C and D. We conclude the appendix in Section F by presenting additional numerical results that confirm the tightness of our theoretical analysis in the strongly convex case.

A. Proof of Theorem 2

A.1. Decentralized SGD in Matrix Notation

We can rewrite Algorithm 1 using the following matrix notation, extending the definition used in the main text:

$$\begin{aligned} X^{(t)} &:= \left[\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_n^{(t)}\right] \in \mathbb{R}^{d \times n}, \\ \bar{X}^{(t)} &:= \left[\bar{\mathbf{x}}^{(t)}, \dots, \bar{\mathbf{x}}^{(t)}\right] \in \mathbb{R}^{d \times n}, \\ \partial F(X^{(t)}, \xi^{(t)}) &:= \left[\nabla F_1(\mathbf{x}_1^{(t)}, \xi_1^{(t)}), \dots, \nabla F_n(\mathbf{x}_n^{(t)}, \xi_n^{(t)})\right] \in \mathbb{R}^{d \times n}. \end{aligned} \tag{14}$$

Algorithm 2 DECENTRALIZED SGD (MATRIX NOTATION)

```
input : $X^{(0)}$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations $T$, mixing matrix distributions $\mathcal{W}^{(t)}$ for $t \in [0, T]$
for $t$ in $0 \dots T$ do
    Sample $W^{(t)} \sim \mathcal{W}^{(t)}$
    $X^{(t+\frac{1}{2})} = X^{(t)} - \eta_t \partial F_i(X^{(t)}, \xi_i^{(t)})$        ▷ stochastic gradient updates
    $X^{(t+1)} = X^{(t+\frac{1}{2})} W^{(t)}$        ▷ gossip averaging
end for
```

A.2. Proof Sketch—Combining Consensus Progress (Gossip) and Optimization Progress (SGD)

In this section we sketch of the proof for Theorem 2. As a first step in the proof, we will derive an upper bound on the expected progress, measured as distance to the optimum, $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$ for the convex cases, and function suboptimality $r_t = \mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - f^\star$ in the non-convex case. These bounds will have the following form:

$$r_{t+1} \leq (1 - a\eta_t) r_t - b\eta_t e_t + c\eta_t^2 + \eta_t B \Xi_t \,, \tag{15}$$

with $\Xi_t = \frac{1}{n} \mathbb{E}_t \sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ and

- for both convex cases $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$, $e_t = f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)$, $a = \frac{\mu}{2}$, $b = 1$, $c = \frac{\bar{\sigma}^2}{n}$, $B = 3L$ (Lemma 8);
- for the non-convex case $r_t = \mathbb{E} f(\bar{\mathbf{x}}^{(t)}) - f^\star$, $e_t = \left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2$, $a = 0$, $b = \frac{1}{4}$, $c = \frac{L\hat{\sigma}^2}{n}$, $B = L^2$ (Lemma 10).

We will then bound the consensus distance Ξ_t as detailed in Section C; Lemmas 9 and11 by a recursion of the form

$$\Xi_t \leq \left(1 - \frac{p}{2}\right) \Xi_{m\tau} + \frac{p}{16\tau} \sum_{j=m\tau}^{t-1} \Xi_j + D \sum_{j=m\tau}^{t-1} \eta_j^2 e_j + A \sum_{j=m\tau}^{t-1} \eta_j^2, \tag{16}$$

with $A = \bar{\sigma}^2 + \frac{18\tau}{p}\bar{\zeta}^2$, $D = 36L\frac{\tau}{p}$ for convex cases (Lemma 9) and $A = \hat{\sigma}^2 + 2\left(\frac{6\tau}{p} + M\right)\hat{\zeta}^2$, $D = 2P\left(\frac{6\tau}{p} + M\right)$ for non-convex case (Lemma 11).

Next, we simplify this recursive equation (16) using Lemma 12 and some positive weights $\{w_t\}_{t\geq 0}$ (see Lemma 12 for the definition of the weights w_t) to

$$B \cdot \sum_{t=0}^{T} w_t \Xi_t \;\leq\; \frac{b}{2} \cdot \sum_{t=0}^{T} w_t e_t + 64AB \cdot \sum_{t=0}^{T} w_t \eta_t^2 \,, \tag{17}$$

where again $\Xi_t = \frac{1}{n}\,\mathbb{E}_t \sum_{i=1}^{n} \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$.

Then we combine (15) and (17). Firstly rearranging (15), multiplying by w_t and dividing by η_t, we get

$$b w_t e_t \leq \frac{(1 - a\eta_t)}{\eta_t} w_t r_t - \frac{w_t}{\eta_t} r_{t+1} + c w_t \eta_t + B w_t \Xi_t \,,$$

Now summing up and dividing by $W_T = \sum_{t=0}^{T} w_t$,

$$\begin{aligned} \frac{1}{W_T}\sum_{t=0}^{T} b w_t e_t &\leq \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{(1-a\eta_t)w_t}{\eta_t} r_t - \frac{w_t}{\eta_t} r_{t+1}\right) + \frac{c}{W_T}\sum_{t=0}^{T} w_t\eta_t + \frac{1}{W_T} B \sum_{t=0}^{T} w_t \Xi_t \\ &\overset{(17)}{\leq} \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{(1-a\eta_t)w_t}{\eta_t} r_t - \frac{w_t}{\eta_t} r_{t+1}\right) + \frac{c}{W_T}\sum_{t=0}^{T} w_t\eta_t + \frac{1}{2W_T}\sum_{t=0}^{T} w_t e_t + \frac{64BA}{W_T}\sum_{t=0}^{T} w_t \eta_t^2 \,, \end{aligned}$$

Therefore,

$$\frac{1}{2W_T}\sum_{t=0}^{T} b w_t e_t \leq \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{(1-a\eta_t)w_t}{\eta_t} r_t - \frac{w_t}{\eta_t} r_{t+1}\right) + \frac{c}{W_T}\sum_{t=0}^{T} w_t\eta_t + \frac{64BA}{W_T}\sum_{t=0}^{T} w_t \eta_t^2 \tag{18}$$

Finally, to solve this main recursion (18) and obtain the final convergence rates of Theorem 2, we will use the following Lemmas, which will be presented in Section D:

- Lemma 13 for strongly convex case when $a > 0$.
- Lemmas 14 and 15 for both weakly convex and non-convex cases as their common feature is that $a = 0$.

A.3. How the Proof of Theorem 2 Follows

In this section we summarize how the proof of Theorem 2 follows from the results that we establish in Sections C and D below. Note that for convex cases we require both f_i and F_i to be convex as in Lemma 9.

Proof of Theorem 2, strongly convex case. The proof follows by applying the result of Lemma 13 to the equation (18) (obtained with Lemmas 8, 9, 12) with $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$, $e_t = f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)$, $a = \frac{\mu}{2}$, $b = 1$, $c = \frac{\bar{\sigma}^2}{n}$, $d = \frac{96\sqrt{3}\tau L}{p}$, $A = \bar{\sigma}^2 + \frac{18\tau}{p}\bar{\zeta}^2$, $B = 3L$. It is only left to show that chosen weights w_t stepsizes η_t in Lemma 13 satisfy conditions of Lemmas 8, 9, 12, which is shown in Proposition 4. □

Proof of Theorem 2, weakly convex case. The proof follows by applying the result of Lemma 14 to the equation (18) (obtained with Lemmas 8, 9, 12) with $r_t = \mathbb{E}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2$, $e_t = f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)$, $a = 0$, $b = 1$, $c = \frac{\bar{\sigma}^2}{n}$, $d = \frac{96\sqrt{3}\tau L}{p}$, $A = \bar{\sigma}^2 + \frac{18\tau}{p}\bar{\zeta}^2$, $B = 3L$. Weights w_t stepsizes η_t chosen in Lemma 14 satisfy conditions of Lemmas 8, 9, 12, as shown in Proposition 4. □

Proof of Theorem 2, non-convex case. applying the result of Lemma 14 to the equation (18) (obtained with Lemmas 10, 11, 12) with $r_t = \mathbb{E}\, f(\bar{\mathbf{x}}^{(t)}) - f^\star$, $e_t = \left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2$, $a = 0$, $b = \frac{1}{4}$, $c = \frac{L\hat{\sigma}^2}{n}$, $d = 64L\sqrt{2P\left(\frac{6\tau}{p} + M\right)\frac{\tau}{p}}$, $A = \hat{\sigma}^2 + 2\left(\frac{6\tau}{p} + M\right)\hat{\zeta}^2$, $B = L^2$. Weights w_t stepsizes η_t chosen in Lemma 14 satisfy conditions of Lemmas 10, 11, 12, as shown in Proposition 4. □

A.4. Improved rate when $\tau = 1$ (recovering mini-batch SGD convergence results)

In the special case when $\tau = 1$ the proof can be simplified and the rate can be improved: there will be an additional $(1-p)$ factor appearing in the middle term, e.g in strongly convex case the improved rate reads as

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{n\mu T} + \frac{L(\bar{\zeta}^2 + p\bar{\sigma}^2)(1-p)}{\mu^2 p^2 T^2} + \frac{LR_0^2}{p}\exp\left[-\frac{\mu T p}{L}\right]\right) .$$

The main difference to the general result stated in Theorem 2 (for $\tau \geq 1$) is that the second term is multiplied with $(1-p)$, allowing to recover the rate of mini-batch SGD in the case of fully-connected graph when $p = 1$. This improvement also holds for the weakly-convex and non-convex case.

In order to do so, one has to observe that the consensus distance Lemmas 9 and 11 can be improved when $\tau = 1$. In the first lines of both these proofs we multiply with $(1-p)$ not only the first term $\left\|X^{(t)} - \bar{X}^{(t)}\right\|_2^2$ but also the second term with the gradient as during the 1-step averaging both $\mathbf{x}^{(t)}$ and $\eta_t \partial F_i(X^{(t)}, \xi_i^{(t)})$ are averaged with mixing matrix $W^{(t)}$ (line 4 of Algorithm 2). We omit the full derivations for this special case, as they can easily be obtained by following the current proofs.

B. Technical Preliminaries

B.1. Implications of the assumptions

Proposition 1. *One step of gossip averaging with the mixing matrix W (def. 1) preserves the average of the iterates, i.e.*

$$XW\frac{\mathbf{1}\mathbf{1}^\top}{n} = X\frac{\mathbf{1}\mathbf{1}^\top}{n} .$$

Proposition 2 (Implications of the smoothness Assumption 1a)**.** *If for functions $F_i(\mathbf{x}, \xi)$ Assumption 1a holds, then it also holds that*

$$F_i(\mathbf{x}, \xi) \leq F_i(\mathbf{y}, \xi) + \langle \nabla F_i(\mathbf{y}, \xi), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\left\|\mathbf{x} - \mathbf{y}\right\|_2^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \Omega_i \tag{19}$$

If functions $f_i(\mathbf{x}) = \mathbb{E}_\xi F_i(\mathbf{x}, \xi)$, then

$$f_i(\mathbf{x}) \leq f_i(\mathbf{y}) + \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\left\|\mathbf{x} - \mathbf{y}\right\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{20}$$

Moreover, if in addition F_i are convex functions, then

$$\left\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\right\|_2 \leq L\left\|\mathbf{x} - \mathbf{y}\right\|_2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{21}$$

$$\left\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\right\|_2^2 \leq 2L\left(g(\mathbf{x}) - g(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla g(\mathbf{y}) \rangle\right), \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{22}$$

where $g(\mathbf{x})$ is either F_i or f_i.

Proposition 3 (Implications of the smoothness Assumption 1b)**.** *From Assumption 1b it follows that*

$$f_i(\mathbf{x}) \leq f_i(\mathbf{y}) + \langle \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\left\|\mathbf{x} - \mathbf{y}\right\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d . \tag{23}$$

B.2. Useful Inequalities

Lemma 4. *For arbitrary set of n vectors $\{\mathbf{a}_i\}_{i=1}^n$, $\mathbf{a}_i \in \mathbb{R}^d$*

$$\left\|\sum_{i=1}^n \mathbf{a}_i\right\|^2 \leq n\sum_{i=1}^n \left\|\mathbf{a}_i\right\|^2 . \tag{24}$$

Lemma 5. *For given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$*

$$2\langle \mathbf{a}, \mathbf{b} \rangle \leq \gamma\left\|\mathbf{a}\right\|^2 + \gamma^{-1}\left\|\mathbf{b}\right\|^2, \qquad \forall \gamma > 0 . \tag{25}$$

Lemma 6. *For given two vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$

$$\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\alpha)\|\mathbf{a}\|^2 + (1+\alpha^{-1})\|\mathbf{b}\|^2, \qquad \forall \alpha > 0. \tag{26}$$

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

Remark 7. *For* $A \in \mathbb{R}^{d\times n}$, $B \in \mathbb{R}^{n\times n}$

$$\|AB\|_F \leq \|A\|_F \|B\|_2 . \tag{27}$$

B.3. τ-slow Sequences

Definition 2 (τ-slow sequences (Stich & Karimireddy, 2019))**.** *The sequence* $\{a_t\}_{t\geq 0}$ *of positive values is* τ-slow decreasing *for parameter* $\tau > 0$ *if*

$$a_{t+1} \leq a_t, \quad \forall t \geq 0 \qquad \textit{and,} \qquad a_{t+1}\left(1+\frac{1}{2\tau}\right) \geq a_t, \quad \forall t \geq 0 .$$

The sequence $\{a_t\}_{t\geq 0}$ *is* τ-slow increasing *if* $\{a_t^{-1}\}_{t\geq 0}$ *is* τ*-slow decreasing.*

Proposition 4 (Examples)**.**

1. *The sequence* $\{\eta_t^2\}_{t\geq 0}$ *with* $\eta_t = \frac{a}{b+t}$, $b \geq \frac{32}{p}$ *is* $\frac{4}{p}$*-slow decreasing.*
2. *The sequence of constant stepsizes* $\{\eta_t^2\}_{t\geq 0}$ *with* $\eta_t = \eta$ *is* τ*-slow decreasing for any* τ*.*
3. *The sequence* $\{w_t\}_{t\geq 0}$ *with* $w_t = (b+t)^2$, $b \geq \frac{84}{p}$ *is* $\frac{8}{p}$*-slow increasing.*
4. *The sequence of constant weights* $\{w_t\}_{t\geq 0}$ *with* $w_t = 1$ *is* τ*-slow increasing for any* τ*.*

C. Descent Lemmas and Consensus Recursions

In this section, according to our proof sketch we derive descent (15) and consensus recursions (17) for both convex and also non-convex cases.

C.1. Convex Cases

We require both f_i and F_i to be convex. We do not need Assumption 2 to hold for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and we could weaken it to hold only for $\mathbf{x} = \mathbf{x}^\star$ and for all $\mathbf{y} \in \mathbb{R}^d$.

Proposition 5 (Mini-batch variance)**.** *Let functions* $F_i(\mathbf{x}, \xi)$, $i \in [n]$ *be L-smooth (Assumption 1a) with bounded noise at the optimum (Assumption 3a). Then for any* $\mathbf{x}_i \in \mathbb{R}^d$, $i \in [n]$ *and* $\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ *it holds*

$$\mathbb{E}_{\xi_1,\dots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i,\xi_i))\right\|^2 \leq \frac{3L^2}{n^2}\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + \frac{6L}{n}\left(f(\bar{\mathbf{x}}) - f(\mathbf{x}^\star)\right) + \frac{3\bar{\sigma}^2}{n} .$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\xi_1,\dots,\xi_n}\ \left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i,\xi_i))\right\|^2 &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{\xi_i}\ \|\nabla F_i(\mathbf{x}_i,\xi_i) - \nabla f_i(\mathbf{x}_i)\|^2 \\
&\leq \frac{3}{n^2}\sum_{i=1}^n \mathbb{E}_{\xi_i}\ \Big(\|\nabla F_i(\mathbf{x}_i,\xi_i) - \nabla F_i(\bar{\mathbf{x}},\xi_i) - \nabla f_i(\mathbf{x}_i) + \nabla f_i(\bar{\mathbf{x}})\|^2 \\
&\qquad + \left\|\nabla F_i(\bar{\mathbf{x}},\xi_i) - \nabla F_i(\mathbf{x}^\star,\xi_i) - \nabla f_i(\bar{\mathbf{x}}^{(t)}) + \nabla f_i(\mathbf{x}^\star)\right\|^2 + \|\nabla F_i(\mathbf{x}^\star,\xi_i) - \nabla f_i(\mathbf{x}^\star)\|^2 \Big) \\
&\leq \frac{3}{n^2}\sum_{i=1}^n \mathbb{E}_{\xi_i}\ \left(\left\|\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)}) - \nabla F_i(\bar{\mathbf{x}},\xi_i)\right\|^2 + \|\nabla F_i(\bar{\mathbf{x}},\xi_i) - \nabla F_i(\mathbf{x}^\star,\xi_i)\|^2 + \|\nabla F_i(\mathbf{x}^\star,\xi_i) - \nabla f_i(\mathbf{x}^\star)\|^2\right) \\
&\leq \frac{3}{n^2}\sum_{i=1}^n \left(L^2\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}\right\|^2 + 2L\left(f_i(\bar{\mathbf{x}}^{(t)}) - f_i(\mathbf{x}^\star)\right) + \sigma_i^2\right),
\end{aligned}$$

where we used that $\mathbb{E}\left\|Y-a\right\|^2 = \mathbb{E}\left\|Y\right\|^2 - \left\|a\right\|^2 \leq \mathbb{E}\left\|Y\right\|^2$ if $a = \mathbb{E}Y$. □

Lemma 8 (Descent lemma for convex cases)**.** *Under Assumptions 1a, 2, 3a and 4, the averages* $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ *of the iterates of Algorithm 1 with the stepsize* $\eta_t \leq \frac{1}{12L}$ *satisfy*

$$\mathbb{E}_{\boldsymbol{\xi}_1^{(t)},\dots,\boldsymbol{\xi}_n^{(t)}} \left\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^\star\right\|^2 \leq \left(1 - \frac{\eta_t \mu}{2}\right)\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2 + \frac{\eta_t^2 \bar{\sigma}^2}{n} - \eta_t\left(f(\bar{\mathbf{x}}^{(t)}) - f^\star\right) + \eta_t \frac{3L}{n}\sum_{i=1}^n \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2, \tag{28}$$

where $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$*.*

Proof. Because all mixing matrixes preserve the average (Proposition 1), we have

$$\begin{aligned}
\left\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^\star\right\|^2 &= \left\|\bar{\mathbf{x}}^{(t)} - \frac{\eta_t}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \mathbf{x}^\star\right\|^2 \\
&= \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star - \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}) + \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}) - \frac{\eta_t}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2 \\
&= \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star - \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 + \eta_t^2\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2 + \\
&\quad + \frac{2\eta_t}{n}\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star - \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}), \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}) - \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\rangle .
\end{aligned}$$

The last term is zero in expectation, as $\mathbb{E}_{\xi_i^{(t)}} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) = \nabla f_i(\mathbf{x}_i^{(t)})$. The second term is estimated using Proposition 5. The first term can be written as:

$$\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star - \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 = \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2 + \eta_t^2 \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2}_{=:T_1} \underbrace{- 2\eta_t \left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star, \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle}_{=:T_2} .$$

We can estimate

$$\begin{aligned}
T_1 &= \left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)}) + \nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}^\star))\right\|^2 \\
&\overset{(24)}{\leq} \frac{2}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\right\|^2 + 2\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\bar{\mathbf{x}}^{(t)}) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^\star)\right\|^2 \\
&\overset{(21),(22)}{\leq} \frac{2L^2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + \frac{4L}{n}\sum_{i=1}^n \left(f_i(\bar{\mathbf{x}}^{(t)}) - f_i(\mathbf{x}^\star)\right) \\
&= \frac{2L^2}{n}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 + 4L\left(f(\bar{\mathbf{x}}^{(t)}) - f^\star\right) .
\end{aligned}$$

And for the remaining T_2 term:

$$\begin{aligned}
-\frac{1}{\eta_t}T_2 &= -\frac{2}{n}\sum_{i=1}^{n}\left[\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}, \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle + \left\langle \mathbf{x}_i^{(t)} - \mathbf{x}^\star, \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle\right] \\
&\overset{(20),(5)}{\leq} -\frac{2}{n}\sum_{i=1}^{n}\left[f_i(\bar{\mathbf{x}}^{(t)}) - f_i(\mathbf{x}_i^{(t)}) - \frac{L}{2}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 + f_i(\mathbf{x}_i^{(t)}) - f_i(\mathbf{x}^\star) + \frac{\mu}{2}\left\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\right\|^2\right] \\
&\overset{(24)}{\leq} -2\left(f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^\star)\right) + \frac{L+\mu}{n}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 - \frac{\mu}{2}\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2,
\end{aligned}$$

Where at the last step (24) was applied to $\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^\star\right\|^2 \leq 2\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|^2 + 2\left\|\mathbf{x}_i^{(t)} - \mathbf{x}^\star\right\|^2$. Putting everything together and using that $\eta_t \leq \frac{1}{12L}$ we are getting statement of the lemma. □

Lemma 9 (Recursion for consensus distance)**.** *Under Assumptions 1a, 2, 3a and 4, if in addition functions F_i are convex and if stepsizes $\eta_t \leq \frac{p}{96\sqrt{3}\tau L}$, then*

$$\Xi_t \leq \left(1 - \frac{p}{2}\right)\Xi_{m\tau} + \frac{p}{16\tau}\sum_{j=m\tau}^{t-1}\Xi_j + 36L\frac{\tau}{p}\sum_{j=m\tau}^{t-1}\eta_j^2\left(f(\bar{\mathbf{x}}^{(j)}) - f(\mathbf{x}^\star)\right) + \left(\bar{\sigma}^2 + \frac{18\tau}{p}\bar{\zeta}^2\right)\sum_{j=m\tau}^{t-1}\eta_j^2,$$

where $\Xi_t = \frac{1}{n}\mathbb{E}_t\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ is a consensus distance, $m = \lfloor t/\tau \rfloor - 1$.

Proof. Using matrix notation (14), for $t \geq \tau$

$$n\Xi_t = \mathbb{E}\left\|X^{(t)} - \bar{X}^{(t)}\right\|_F^2 = \mathbb{E}\left\|X^{(t)} - \bar{X}^{(m\tau+\frac{1}{2})} - \left(\bar{X}^{(t)} - \bar{X}^{(m\tau+\frac{1}{2})}\right)\right\|_F^2 \leq \mathbb{E}\left\|X^{(t)} - \bar{X}^{(m\tau+\frac{1}{2})}\right\|_F^2,$$

where we used that $\left\|A - \bar{A}\right\|_F^2 = \sum_{i=1}^n\left\|\mathbf{a}_i - \bar{\mathbf{a}}\right\|_2^2 \leq \sum_{i=1}^n\left\|\mathbf{a}_i\right\|_2^2 = \left\|A\right\|_F^2$. Unrolling $X^{(t)}$ up to $X^{(m\tau)}$ using lines 3–4 of the Algorithm 2,

$$\begin{aligned}
n\Xi_t &\leq \mathbb{E}\left\|X^{(m\tau)}\prod_{i=t-1}^{m\tau}W^{(i)} - \bar{X}^{(m\tau+\frac{1}{2})} + \sum_{j=m\tau}^{t-1}\eta_j\partial F(X^{(j)},\xi^{(j)})\prod_{i=t-1}^{j}W^{(i)}\right\|_F^2 \\
&\leq \mathbb{E}\left\|X^{(m\tau)}\prod_{i=t-1}^{m\tau}W^{(i)} - \bar{X}^{(m\tau+\frac{1}{2})} + \sum_{j=m\tau}^{t-1}\eta_j\left(\partial F(X^{(j)},\xi^{(j)}) - \partial F(X^\star,\xi^{(j)}) + \partial f(X^\star)\right)\prod_{i=t-1}^{j}W^{(i)}\right\|_F^2 \\
&\quad + \mathbb{E}\left\|\sum_{j=m\tau}^{t-1}\eta_j\left(\partial F(X^\star,\xi^{(j)}) - \partial f(X^\star)\right)\prod_{i=t-1}^{j}W^{(i)}\right\|_F^2
\end{aligned}$$

where we used that $\mathbb{E}\,\partial F(X^\star, \xi^{(j)}) = \partial f(X^\star)$. Using that all of the $\xi^{(j)}$ are independent for different j,

$$\begin{aligned}
n\Xi_t &\overset{(26)}{\le} (1+\alpha)\,\mathbb{E}\left\|X^{(m\tau)}\prod_{i=t-1}^{m\tau} W^{(i)} - \bar{X}^{(m\tau+\frac{1}{2})}\right\|_F^2 \\
&\quad + (1+\alpha^{-1})\,\mathbb{E}\left\|\sum_{j=m\tau}^{t-1}\eta_j\left(\partial F(X^{(j)},\xi^{(j)}) - \partial F(X^\star,\xi^{(j)}) + \partial f(X^\star)\right)\prod_{i=t-1}^{j} W^{(i)}\right\|_F^2 \\
&\quad + \sum_{j=m\tau}^{t-1}\eta_j^2\,\mathbb{E}\left\|\left(\partial F(X^\star,\xi^{(j)}) - \partial f(X^\star)\right)\prod_{i=t-1}^{j} W^{(i)}\right\|_F^2 \\
&\overset{(13),(24),(27)}{\le} (1+\alpha)(1-p)\,\mathbb{E}\left\|X^{(m\tau)} - \bar{X}^{(m\tau)}\right\|_F^2 + (1+\alpha^{-1})2\tau\sum_{j=m\tau}^{t-1}\eta_j^2\,\mathbb{E}\left\|\partial F(X^{(j)},\xi^{(j)}) - \partial F(X^\star,\xi^{(j)}) + \partial f(X^\star)\right\|_F^2 \\
&\quad + \sum_{j=m\tau}^{t-1}\eta_j^2\,\mathbb{E}\left\|\partial F(X^\star,\xi^{(j)}) - \partial f(X^\star)\right\|_F^2 \\
&\overset{\alpha=\frac{p}{2},(7)}{\le} (1-\frac{p}{2})\,\mathbb{E}\left\|X^{(m\tau)} - \bar{X}^{(m\tau)}\right\|_F^2 + \frac{6\tau}{p}\sum_{j=m\tau}^{t-1}\eta_j^2\underbrace{\mathbb{E}\left\|\partial F(X^{(j)},\xi^{(j)}) - \partial F(X^\star,\xi^{(j)}) + \partial f(X^\star)\right\|_F^2}_{=:T} + \sum_{j=m\tau}^{t-1}\eta_j^2 n\bar{\sigma}^2,
\end{aligned}$$

We estimate the second term as

$$\begin{aligned}
T &= \mathbb{E}\left\|\partial F(X^{(j)},\xi^{(j)}) - \partial F(\bar{X}^{(j)},\xi^{(j)}) + \partial F(\bar{X}^{(j)},\xi^{(j)}) - \partial F(X^\star,\xi^{(j)}) + \partial f(X^\star)\right\|_F^2 \\
&\overset{(24)}{\le} 3\left\|\partial F(X^{(j)},\xi^{(j)}) - \partial F(\bar{X}^{(j)},\xi^{(j)})\right\|_F^2 + 3\left\|\partial F(\bar{X},\xi^{(j)}) - \partial F(X^\star,\xi^{(j)})\right\|_F^2 + 3\left\|\partial f(X^\star)\right\|_F^2 \\
&\overset{(3),(22),(6)}{\le} 3\left(L^2\left\|X^{(j)} - \bar{X}^{(j)}\right\|_F^2 + 2Ln(f(\bar{\mathbf{x}}^{(j)}) - f(\mathbf{x}^\star)) + n\bar{\zeta}^2\right)
\end{aligned}$$

Putting back estimate for T and using that $\eta_t \le \frac{p}{12\sqrt{2}\tau L}$ we arrive to the statement of the lemma. □

C.2. Non-convex Case

Here we derive descent recursive equation (15) and recursion for consensus distance (16) for the non-convex case.

Proposition 6 (Mini-batch variance)**.** *Let functions $F_i(\mathbf{x},\xi)$, $i \in [n]$ be L-smooth (Assumption 1a) with bounded noise as in Assumption 3b. Then for any $\mathbf{x}_i \in \mathbb{R}^d, i \in [n]$ and $\bar{\mathbf{x}} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ it holds*

$$\mathbb{E}_{\xi_1,\dots,\xi_n}\left\|\frac{1}{n}\sum_{i=1}^n\left(\nabla f_i(\mathbf{x}_i) - \nabla F_i(\mathbf{x}_i,\xi_i)\right)\right\|^2 \le \frac{\hat{\sigma}^2}{n} + \frac{M}{n^2}\sum_{i=1}^n\|\nabla f(\mathbf{x}_i)\|^2 \tag{29}$$

Lemma 10 (Descent lemma for non-convex case)**.** *Under Assumptions 1b, 3b and 4, the averages $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ of the iterates of Algorithm 1 with the constant stepsize $\eta < \frac{1}{4L(M+1)}$ satisfy*

$$\mathbb{E}_{t+1}\, f(\bar{\mathbf{x}}^{(t+1)}) \le f(\bar{\mathbf{x}}^{(t)}) - \frac{\eta}{4}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{\eta L^2}{n}\sum_{i=1}^n\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|_2^2 + \frac{L}{n}\eta^2\hat{\sigma}^2. \tag{30}$$

Proof. Because all mixing matrixes preserve the average (Proposition 1) and function f is L-smooth, we have

$$\begin{aligned}\mathbb{E}_{t+1}\, f(\bar{\mathbf{x}}^{(t+1)}) &= \mathbb{E}_{t+1}\, f\left(\bar{\mathbf{x}}^{(t)} - \frac{\eta}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right) \\ &\leq f(\bar{\mathbf{x}}^{(t)}) \underbrace{- \mathbb{E}_{t+1}\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{\eta}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\rangle}_{:=T_1} + \mathbb{E}_{t+1}\, \frac{L}{2}\eta^2 \underbrace{\left\| \frac{1}{n}\sum_{j=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|_2^2}_{:=T_2}\end{aligned}$$

To estimate the second term, we add and subtract $\nabla f(\bar{\mathbf{x}}^{(t)})$

$$\begin{aligned}T_1 &= -\eta\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta}{n}\sum_{i=1}^n \left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle \\ &\overset{(25),\gamma=1;(24)}{\leq} -\frac{\eta}{2}\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta}{2n}\sum_{i=1}^n \left\|\nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2\end{aligned}$$

For the last term,

$$\begin{aligned}T_2 &= \mathbb{E}_{t+1}\left\|\frac{1}{n}\sum_{j=1}^n\left(\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right)\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\| \\ &\overset{(29)}{\leq} \frac{\hat{\sigma}^2}{n} + \frac{M}{n^2}\sum_{i=1}^n \left\|\nabla f(\mathbf{x}_i^{(t)}) \pm \nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}) \pm \nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &\overset{(26)}{\leq} \frac{\hat{\sigma}^2}{n} + \frac{2M}{n^2}\sum_{i=1}^n \left\|\nabla f(\mathbf{x}_i^{(t)}) - \nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2 + (2{}^M\!/\!{}_n + 2)\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{2}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\right\|_2^2\end{aligned}$$

Combining this together and using L-smoothness to estimate $\left\|\nabla f_i(\bar{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\|_2^2$ and $\left\|\nabla f(\bar{\mathbf{x}}^{(t)}) - \nabla f(\mathbf{x}_i^{(t)})\right\|_2^2$,

$$\mathbb{E}_{t+1}\, f(\bar{\mathbf{x}}^{(t+1)}) \leq f(\bar{\mathbf{x}}^{(t)}) - \eta\left(\frac{1}{2} - L\eta(M+1)\right)\left\|\nabla f(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \left(\frac{\eta L^2}{2n} + \frac{L^3\eta^2(M+1)}{n}\right)\sum_{i=1}^n \left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\right\|_2^2 + \frac{L}{n}\eta^2\hat{\sigma}^2.$$

Applying $\eta < \frac{1}{4L(M+1)}$ we get statement of the lemma. □

Lemma 11 (Recursion for consensus distance)**.** *Under Assumptions 1b, 3b and 4, if the stepsize* $\eta_t \leq \frac{p}{96\sqrt{3}\tau L}$*, then*

$$\Xi_t \leq \left(1 - \frac{p}{2}\right)\Xi_{m\tau} + \frac{p}{16\tau}\sum_{j=m\tau}^{t-1}\Xi_j + 2P\left(\frac{6\tau}{p} + M\right)\sum_{j=m\tau}^{t-1}\eta_j^2\left\|\nabla f(\bar{\mathbf{x}}^{(j)})\right\|_2^2 + \left(\hat{\sigma}^2 + 2\left(\frac{6\tau}{p} + M\right)\hat{\zeta}^2\right)\sum_{j=m\tau}^{t-1}\eta_j^2$$

where $\Xi_t = \frac{1}{n}\mathbb{E}_t\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ *is a consensus distance,* $m = \lfloor t/\tau \rfloor - 1$.

Proof. We start exactly the same way as in the convex proof in Lemma 9 Defining $\Xi_t = \frac{1}{n}\mathbb{E}_t\sum_{i=1}^n\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$, $m = \lfloor t/\tau \rfloor - 1$ and using matrix notation (11), for $t \geq \tau$ (and therefore $m \geq 0$)

$$n\Xi_t = \mathbb{E}\left\|X^{(t)} - \bar{X}^{(t)}\right\|_F^2 = \mathbb{E}\left\|X^{(t)} - \bar{X}^{(m\tau+\frac{1}{2})} - \left(\bar{X}^{(t)} - \bar{X}^{(m\tau+\frac{1}{2})}\right)\right\|_F^2 \leq \mathbb{E}\left\|X^{(t)} - \bar{X}^{(m\tau+\frac{1}{2})}\right\|_F^2,$$

where we used that $\left\|A - \bar{A}\right\|_F^2 = \sum_{i=1}^n \|\mathbf{a}_i - \bar{\mathbf{a}}\| \leq \sum_{i=1}^n \|\mathbf{a}_i\|_F^2 = \|A\|_F^2$. Unrolling $X^{(t)}$ up to $X^{(m\tau)}$ using lines 3-4 of the Algorithm 2,

$$\begin{aligned}
n\Xi_t &\leq \mathbb{E}\left\|X^{(m\tau)}\prod_{i=t-1}^{m\tau} W^{(i)} - \bar{X}^{(m\tau)} + \sum_{j=m\tau}^{t-1}\eta_j \partial F(X^{(j)}, \xi^{(j)}) \prod_{i=t-1}^{j} W^{(i)}\right\|_F^2 \\
&\leq \mathbb{E}\left\|X^{(m\tau)}\prod_{i=t-1}^{m\tau} W^{(i)} - \bar{X}^{(m\tau)} + \sum_{j=m\tau}^{t-1}\eta_j \partial f(X^{(j)}) \prod_{i=t-1}^{j} W^{(i)}\right\|_F^2 \\
&\quad + \mathbb{E}\left\|\sum_{j=m\tau}^{t-1}\eta_j \left(\partial F(X^{(j)}, \xi^{(j)}) - \partial f(X^{(j)})\right) \prod_{i=t-1}^{j} W^{(i)}\right\|_F^2
\end{aligned}$$

where we used that $\mathbb{E}\,\partial F(X^{(j)}, \xi^{(j)}) = \partial f(X^{(j)})$ and that all of the $\xi^{(j)}$ are independent for different j.

$$\begin{aligned}
n\Xi_t &\overset{(26)}{\leq} (1+\alpha)\,\mathbb{E}\left\|X^{(m\tau)}\prod_{i=t-1}^{m\tau} W^{(i)} - \bar{X}^{(m\tau)}\right\|_F^2 + (1+\alpha^{-1})\,\mathbb{E}\left\|\sum_{j=m\tau}^{t-1}\eta_j \partial f(X^{(j)}) \prod_{i=t-1}^{j} W^{(i)}\right\|_F^2 \\
&\quad + \sum_{j=m\tau}^{t-1}\eta_j^2\,\mathbb{E}\left\|\left(\partial F(X^{(j)}, \xi^{(j)}) - \partial f(X^{(j)})\right)\prod_{i=t-1}^{j} W^{(i)}\right\|_F^2 \\
&\overset{(13),(24),(27)}{\leq} (1+\alpha)(1-p)\,\mathbb{E}\left\|X^{(m\tau)} - \bar{X}^{(m\tau)}\right\|_F^2 + (1+\alpha^{-1})2\tau\sum_{j=m\tau}^{t-1}\eta_j^2\,\mathbb{E}\left\|\partial f(X^{(j)})\right\|_F^2 \\
&\quad + \sum_{j=m\tau}^{t-1}\eta_j^2\,\mathbb{E}\left\|\partial F(X^{(j)}, \xi^{(j)}) - \partial f(X^{(j)})\right\|_F^2 \\
&\overset{\alpha=\frac{p}{2},(9)}{\leq} \left(1-\frac{p}{2}\right)\mathbb{E}\left\|X^{(m\tau)} - \bar{X}^{(m\tau)}\right\|_F^2 + \left(\frac{6\tau}{p} + M\right)\sum_{j=m\tau}^{t-1}\eta_j^2\left\|\partial f(X^{(j)})\right\|_F^2 + \sum_{j=m\tau}^{t-1}\eta_j^2 n\hat{\sigma}^2 \\
&\overset{(26)}{\leq} \left(1-\frac{p}{2}\right)\mathbb{E}\left\|X^{(m\tau)} - \bar{X}^{(m\tau)}\right\|_F^2 + \left(\frac{6\tau}{p} + M\right)2\sum_{j=m\tau}^{t-1}\eta_j^2\left(\left\|\partial f(X^{(j)}) - \partial f(\bar{X}^{(j)})\right\|_F^2 + \left\|\partial f(\bar{X}^{(j)})\right\|_F^2\right) \\
&\quad + \sum_{j=m\tau}^{t-1}\eta_j^2 n\hat{\sigma}^2 \\
&\overset{(4),(8)}{\leq} \left(1-\frac{p}{2}\right)\mathbb{E}\left\|X^{(m\tau)} - \bar{X}^{(m\tau)}\right\|_F^2 + \left(\frac{6\tau}{p} + M\right)2\sum_{j=m\tau}^{t-1}\eta_j^2\left(L^2\left\|X^{(j)} - \bar{X}^{(j)}\right\|_F^2 + n\hat{\zeta}^2 + Pn\left\|\nabla f(\bar{\mathbf{x}}^{(j)})\right\|_2^2\right) \\
&\quad + \sum_{j=m\tau}^{t-1}\eta_j^2 n\hat{\sigma}^2
\end{aligned}$$

Where $\partial\bar{f}(\bar{X}^{(j)}) = \partial f(\bar{X}^{(j)})\frac{\mathbf{1}\mathbf{1}^\top}{n}$. Using that $\eta_t \leq \frac{p}{4L\sqrt{2\tau(6\tau+pM)}}$

$$n\Xi_t \leq \left(1-\frac{p}{2}\right)n\Xi_{m\tau} + \sum_{j=m\tau}^{t-1}\frac{p}{16\tau}n\Xi_j + 2Pn\left(\frac{6\tau}{p} + M\right)\sum_{j=m\tau}^{t-1}\eta_j^2\left\|\nabla f(\bar{\mathbf{x}}^{(j)})\right\|_2^2 + \sum_{j=m\tau}^{t-1}\eta_j^2 n\left[\hat{\sigma}^2 + 2\left(\frac{6\tau}{p} + M\right)\hat{\zeta}^2\right]$$

□

C.3. Simplifying Consensus Recursion

In Lemmas 9, 11 we obtained the consensus recursive equation (16) for both convex and non-convex cases. In this section we simplify it to be able to easily combine it later with (15).

Lemma 12. *If non-negative sequences $\{\Xi_t\}_{t\geq 0}$, $\{e_t\}_{t\geq 0}$ and $\{\eta_t\}_{t\geq 0}$ satisfy (16) for some constants $0 < p \leq 1, \tau \geq 1, A, D \geq 0$, moreover if the stepsizes $\{\eta_t^2\}_{t\geq 0}$ is $\frac{8\tau}{p}$-slow decreasing sequence (Definition 2), and if $\{w_t\}_{t\geq 0}$ is $\frac{16\tau}{p}$-slow increasing non-negative sequence of weights, then it holds that*

$$E\sum_{t=0}^{T} w_t \Xi_t \leq \frac{b}{2}\sum_{t=0}^{T} w_t e_t + 64BA\frac{\tau}{p}\sum_{t=0}^{T} w_t \eta_t^2,$$

for some constant $E > 0$ with the constraint that stepsizes $\eta_t \leq \frac{1}{16}\sqrt{\frac{pb}{DB\tau}}$.

Proof. Recursively substituting every Ξ_j in the second term of (16) we get

$$\Xi_t \leq \left(1 - \frac{p}{2}\right)\Xi_{m\tau}\left(1 + \frac{p}{16\tau}\right)^{2\tau} + D\sum_{j=m\tau}^{t-1}\left(1 + \frac{p}{16\tau}\right)^{t-1-j}\eta_j^2 e_j + A\sum_{j=m\tau}^{t-1}\left(1 + \frac{p}{16\tau}\right)^{t-1-j}\eta_j^2,$$

Using that $\left(1 + \frac{p}{16\tau}\right)^{2\tau} \leq \exp\left(\frac{p}{8}\right) \leq 1 + \frac{p}{4}$ for $p \leq 1$ and also that $(1 + \frac{p}{16\tau})^{t-1-j} \leq \left(1 + \frac{p}{16\tau}\right)^{2\tau} \leq 1 + \frac{p}{4} \leq 2$

$$\Xi_t \leq \left(1 - \frac{p}{4}\right)\Xi_{m\tau} + 2D\sum_{j=m\tau}^{t-1}\eta_j^2 e_j + 2A\sum_{j=m\tau}^{t-1}\eta_j^2,$$

Unrolling $\Xi_{m\tau}$ recursively up to 0 we get,

$$\Xi_t \leq 2D\sum_{j=0}^{t-1}\left(1 - \frac{p}{4}\right)^{\lfloor (t-j)/\tau \rfloor}\eta_j^2 e_j + 2A\sum_{j=0}^{t-1}\left(1 - \frac{p}{4}\right)^{\lfloor (t-j)/\tau \rfloor}\eta_j^2,$$

For the first term estimating $\left(1 - \frac{p}{4}\right)^{1/\tau} \leq \exp(-\frac{p}{4\tau}) \leq 1 - \frac{p}{8\tau}$ and that $\left(1 - \frac{p}{8\tau}\right)^{\tau\lfloor (t-j)/\tau \rfloor} \leq \left(1 - \frac{p}{8\tau}\right)^{t-j}\left(1 - \frac{p}{8\tau}\right)^{-\tau}$. For the last term, $\left(1 - \frac{p}{8\tau}\right)^{-\tau} \leq \left(\frac{1}{1 - \frac{p}{8\tau}}\right)^{\tau} \leq (1 + \frac{p}{4\tau})^{\tau}$ because $\frac{p}{8\tau} \leq \frac{1}{2}$ and finally $\left(1 + \frac{p}{4\tau}\right)^{\tau} \leq \exp(\frac{p}{4}) < 2$,

$$\Xi_t \leq 4D\sum_{j=0}^{t-1}\left(1 - \frac{p}{8\tau}\right)^{t-j}\eta_j^2 e_j + 4A\sum_{j=0}^{t-1}\left(1 - \frac{p}{8\tau}\right)^{t-j}\eta_j^2,$$

Now using that η_t^2 is $\frac{8\tau}{p}$-slow decreasing, i.e. $\eta_j^2 \leq \eta_t^2\left(1 + \frac{p}{16\tau}\right)^{t-j}$ and using that $(1 - \frac{p}{8\tau})(1 + \frac{p}{16\tau}) \leq (1 - \frac{p}{16\tau})$

$$\Xi_t \leq 4D\eta_t^2\sum_{j=0}^{t-1}\left(1 - \frac{p}{16\tau}\right)^{t-j}e_j + 4A\eta_t^2\sum_{j=0}^{t-1}\left(1 - \frac{p}{16\tau}\right)^{t-j} \leq 4D\eta_t^2\sum_{j=0}^{t-1}\left(1 - \frac{p}{16\tau}\right)^{t-j}e_j + 64A\frac{\tau}{p}\eta_t^2$$

Now averaging Ξ_t with weights w_t and using that w_t is $\frac{16\tau}{p}$-slow increasing sequence, i.e. $w_t \leq w_j\left(1 + \frac{p}{32\tau}\right)^{t-j}$, and also using that $\eta_t \leq \frac{1}{16}\sqrt{\frac{pb}{DB\tau}}$

$$\begin{aligned} B\sum_{t=0}^{T} w_t \Xi_t &\leq 4DB\sum_{t=0}^{T}\eta_t^2\sum_{j=0}^{t-1} w_j\left(1 - \frac{p}{32\tau}\right)^{t-j}e_j + 64AB\frac{\tau}{p}\sum_{t=0}^{T} w_t\eta_t^2 \\ &\leq \underbrace{\frac{pb}{64\tau}\sum_{t=0}^{T}\sum_{j=0}^{t-1} w_j\left(1 - \frac{p}{32\tau}\right)^{t-j}e_j}_{:=T_1} + 64AB\frac{\tau}{p}\sum_{t=0}^{T} w_t\eta_t^2 \end{aligned}$$

And finally,

$$T_1 = \frac{pb}{64\tau}\sum_{j=0}^{T} w_j e_j\sum_{t=j+1}^{T}\left(1 - \frac{p}{32\tau}\right)^{t-j} \leq \frac{pb}{64\tau}\sum_{j=0}^{T} w_j e_j\sum_{t=0}^{\infty}\left(1 - \frac{p}{32\tau}\right)^{t-j} \leq \frac{b}{2}\sum_{t=0}^{T} w_t e_t. \qquad \square$$

D. Solving the Main Recursion (18)

D.1. $a > 0$ (strongly convex case)

Lemma 13. *If non-negative sequences $\{r_t\}_{t\geq 0}, \{e_t\}_{t\geq 0}$ satisfy (18) for some constants $a, b > 0, c, A, B \geq 0$, then there exists a constant stepsize $\eta_t = \eta < \frac{1}{d}$ such that for weights $w_t = (1-a\eta)^{-(t+1)}$ and $W_T := \sum_{t=0}^T w_t$ it holds:*

$$\frac{1}{2W_T}\sum_{t=0}^T b e_t w_t + a r_{T+1} \leq \tilde{\mathcal{O}}\left(r_0 d \exp\left[-\frac{a(T+1)}{d}\right] + \frac{c}{aT} + \frac{BA}{a^2T^2}\right),$$

where $\tilde{\mathcal{O}}$ hides polylogarithmic factors.

Proof. Starting from (18) and using that $\eta_t = \eta$ and that $\frac{w_t(1-a\eta)}{\eta} = \frac{w_{t-1}}{\eta}$ we obtain a telescoping sum,

$$\frac{1}{2W_T}\sum_{t=0}^T b w_t e_t \leq \frac{1}{W_T\eta}\left((1-a\eta)w_0 r_0 - w_T r_{T+1}\right) + c\eta + 64BA\eta^2\,,$$

And hence,

$$\frac{1}{2W_T}\sum_{t=0}^T b w_t e_t + \frac{w_T r_{T+1}}{W_T\eta} \leq \frac{r_0}{W_T\eta} + c\eta + 64BA\eta^2\,,$$

Using that $W_T \leq \frac{w_T}{a\eta}$ and $W_T \geq w_T = (1-a\gamma)^{-(T+1)}$ we can simplify

$$\frac{1}{2W_T}\sum_{t=0}^T b w_t e_t + a r_{T+1} \leq (1-a\eta)^{T+1}\frac{r_0}{\eta} + c\eta + 64BA\eta^2 \leq \frac{r_0}{\eta}\exp\left[-a\eta(T+1)\right] + c\eta + 64BA\eta^2\,,$$

Now lemma follows by tuning η the same way as in (Stich, 2019a).

- If $\frac{1}{d} \geq \frac{\ln(\max\{2, a^2 r_0 T^2/c\})}{aT}$ then we choose $\eta = \frac{\ln(\max\{2, a^2 r_0 T^2/c\})}{aT}$ and get that

$$\tilde{\mathcal{O}}\left(a r_0 T \exp\left[-\ln(\max\{2, a^2 r_0 T^2/c\})\right]\right) + \tilde{\mathcal{O}}\left(\frac{c}{aT}\right) + \tilde{\mathcal{O}}\left(\frac{BA}{a^2T^2}\right) = \tilde{\mathcal{O}}\left(\frac{c}{aT}\right) + \tilde{\mathcal{O}}\left(\frac{BA}{a^2T^2}\right),$$

- Otherwise $\frac{1}{d} \leq \frac{\ln(\max\{2, a^2 r_0 T^2/c\})}{aT}$ we pick $\eta = \frac{1}{d}$ and get that

$$\tilde{\mathcal{O}}\left(r_0 d \exp\left[-\frac{a(T+1)}{d}\right] + \frac{c}{d} + \frac{BA}{d^2}\right) \leq \tilde{\mathcal{O}}\left(r_0 d \exp\left[-\frac{a(T+1)}{d}\right] + \frac{c}{aT} + \frac{BA}{a^2T^2}\right). \qquad \square$$

D.2. $a = 0$ (weakly convex and non-convex cases)

Now we assume that in Assumption 2 $\mu = 0$, which means that $a = 0$ in (18).

Lemma 14. *If non-negative sequences $\{r_t\}_{t\geq 0}, \{e_t\}_{t\geq 0}$ satisfy (18) with $a = 0, b > 0, c, A, B \geq 0$, then there exists a constant stepsize $\eta_t = \eta < \frac{1}{d}$ such that for weights $\{w_t = 1\}_{t\geq 0}$ it holds that:*

$$\frac{1}{(T+1)}\sum_{t=0}^T e_t \leq \mathcal{O}\left(2\left(\frac{cr_0}{T+1}\right)^{\frac{1}{2}} + 2(BA)^{1/3}\left(\frac{r_0}{T+1}\right)^{\frac{2}{3}} + \frac{dr_0}{T+1}\right).$$

Proof. With $a = 0$, constant stepsizes $\eta_t = \eta$ and weights $\{w_t = 1\}_{t\geq 0}$ (18) is equivalent to

$$\frac{1}{2(T+1)}\sum_{t=0}^T e_t \leq \frac{1}{(T+1)\eta}\sum_{t=0}^T (r_t - r_{t+1}) + c\eta + 64BA\eta^2 \leq \frac{r_0}{(T+1)\eta} + c\eta + 64BA\eta^2.$$

To conclude the proof we tune the stepsize using Lemma 15. □

Lemma 15 (Tuning the stepsize)**.** *For any parameters* $r_0 \geq 0, b \geq 0, e \geq 0, d \geq 0$ *there exists constant stepsize* $\eta \leq \frac{1}{d}$ *such that*

$$\Psi_T := \frac{r_0}{\eta(T+1)} + b\eta + e\eta^2 \leq 2\left(\frac{br_0}{T+1}\right)^{\frac{1}{2}} + 2e^{1/3}\left(\frac{r_0}{T+1}\right)^{\frac{2}{3}} + \frac{dr_0}{T+1}$$

Proof. Choosing $\eta = \min\left\{\left(\frac{r_0}{b(T+1)}\right)^{\frac{1}{2}}, \left(\frac{r_0}{e(T+1)}\right)^{\frac{1}{3}}, \frac{1}{d}\right\} \leq \frac{1}{d}$ we have three cases

- $\eta = \frac{1}{d}$ and is smaller than both $\left(\frac{r_0}{b(T+1)}\right)^{\frac{1}{2}}$ and $\left(\frac{r_0}{e(T+1)}\right)^{\frac{1}{3}}$, then

$$\Psi_T \leq \frac{dr_0}{T+1} + \frac{b}{d} + \frac{e}{d^2} \leq \left(\frac{br_0}{T+1}\right)^{\frac{1}{2}} + \frac{dr_0}{T+1} + e^{1/3}\left(\frac{r_0}{T+1}\right)^{\frac{2}{3}}$$

- $\eta = \left(\frac{r_0}{b(T+1)}\right)^{\frac{1}{2}} < \left(\frac{r_0}{e(T+1)}\right)^{\frac{1}{3}}$, then

$$\Psi_T \leq 2\left(\frac{r_0 b}{T+1}\right)^{\frac{1}{2}} + e\left(\frac{r_0}{b(T+1)}\right) \leq 2\left(\frac{r_0 b}{T+1}\right)^{\frac{1}{2}} + e^{\frac{1}{3}}\left(\frac{r_0}{(T+1)}\right)^{\frac{2}{3}},$$

- The last case, $\eta = \left(\frac{r_0}{e(T+1)}\right)^{\frac{1}{3}} < \left(\frac{r_0}{b(T+1)}\right)^{\frac{1}{2}}$

$$\Psi_T \leq 2e^{\frac{1}{3}}\left(\frac{r_0}{(T+1)}\right)^{\frac{2}{3}} + b\left(\frac{r_0}{e(T+1)}\right)^{\frac{1}{3}} \leq 2e^{\frac{1}{3}}\left(\frac{r_0}{(T+1)}\right)^{\frac{2}{3}} + \left(\frac{br_0}{T+1}\right)^{\frac{1}{2}}.$$

□

E. Lower Bound

Proof of Theorem 3. We consider minimization problem of the form (1) with $f_i(x) = \frac{1}{2}(x - y_i)^2$, $x, y_i \in \mathbb{R}$ which has the solution $x^\star = \frac{1}{n}\sum_{i=1}^n y_i$, $L = \mu = 1$. We denote $\mathbf{x} = (x_1, \ldots, x_n)^\top$ and $\nabla f(\mathbf{x}) = (\nabla f_1(x_1), \ldots, \nabla f_n(x_n))^\top$.

We assume that the starting point $\mathbf{x}^{(0)}$ is an eigenvector of W, corresponding to the second largest eigenvalue, i.e. $W\mathbf{x}^{(0)} = (1-p)\mathbf{x}^{(0)}$ and we set y_i such that $\mathbf{y} = \mathbf{1} + \mathbf{x}^{(0)}$. With this choice of $\mathbf{y}$, $\bar{\zeta}^2 = \left\|\mathbf{x}^{(0)}\right\|_2^2$. It will be also useful to note that the average $\bar{\mathbf{x}}^{(0)} = \mathbf{0}$ since it is orthogonal to $\mathbf{1}$, the eigenvector of W corresponding to the largest eigenvalue. We use the notation $\bar{\mathbf{z}} := \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{z}$.

We start the proof by decomposing the error $\left\|\mathbf{x}^{(t)} - \bar{\mathbf{y}}\right\|_2^2$ on consensus and optimization terms

$$\left\|\mathbf{x}^{(t)} - \bar{\mathbf{y}}\right\|_2^2 = \left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} + \bar{\mathbf{x}}^{(t)} - \bar{\mathbf{y}}\right\|_2^2 = \left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \left\|\bar{\mathbf{x}}^{(t)} - \bar{\mathbf{y}}\right\|_2^2.$$

Using that for our chosen functions $\nabla f(\mathbf{x}) = \mathbf{x} - \mathbf{y}$, we can estimate the **optimization term** as

$$\left\|\bar{\mathbf{x}}^{(t)} - \bar{\mathbf{y}}\right\|_2^2 = \left\|(1-\eta)\bar{\mathbf{x}}^{(t-1)} + \eta\bar{\mathbf{y}} - \bar{\mathbf{y}}\right\|_2^2 = (1-\eta)^2\left\|\bar{\mathbf{x}}^{(t-1)} - \bar{\mathbf{y}}\right\|_2^2 = (1-\eta)^{2t}\left\|\bar{\mathbf{x}}^{(0)} - \bar{\mathbf{y}}\right\|_2^2 = (1-\eta)^{2t}\left(\bar{\zeta}^2 + n\right).$$

For the **consensus term**,

$$\begin{aligned}
\left\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 &= \left\|\left(W - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\left(\mathbf{x}^{(t+\frac{1}{2})} - \bar{\mathbf{x}}^{(t+\frac{1}{2})}\right)\right\|_2^2 = \left\|\left(W - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\left((1-\eta)\left(\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\right) + \eta(\mathbf{y} - \bar{\mathbf{y}})\right)\right\|_2^2 = \\
&= \left\|\left(W - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)^t (1-\eta)^t \mathbf{x}^{(0)} + \eta \sum_{\tau=0}^{t-1} (1-\eta)^\tau \left(W - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)^\tau (\mathbf{y} - \bar{\mathbf{y}})\right\|_2^2 = \\
&= \left\|(1-p)^t(1-\eta)^t \mathbf{x}^{(0)} + \eta \sum_{\tau=0}^{t-1}(1-\eta)^\tau (1-p)^\tau \mathbf{x}^{(0)}\right\|_2^2 \\
&= \left((1-p)^t(1-\eta)^t + \eta \sum_{\tau=0}^{t-1}(1-\eta)^\tau(1-p)^\tau\right)^2 \left\|\mathbf{x}^{(0)}\right\|_2^2 \\
&\geq \left((1-p)^{2t}(1-\eta)^{2t} + \eta^2\left(\sum_{\tau=0}^{t-1}(1-\eta)^\tau(1-p)^\tau\right)^2\right)\bar{\zeta}^2
\end{aligned}$$

In order to guarantee error less than ϵ, it is necessary to have simultaneously both optimization and consensus terms less than ϵ, therefore it is required that

$$(1-\eta)^{2t} \leq \frac{\epsilon}{n} \tag{31}$$

$$(1-\eta)^{2t}(1-p)^{2t} \leq \frac{\epsilon}{\bar{\zeta}^2} \tag{32}$$

$$\eta\left(\sum_{\tau=0}^{t-1}(1-\eta)^\tau(1-p)^\tau\right) = \eta\frac{1-(1-\eta)^t(1-p)^t}{1-(1-\eta)(1-p)} \leq \sqrt{\frac{\epsilon}{\bar{\zeta}^2}} \tag{33}$$

Equations (32), (33) imply

$$\eta \leq \sqrt{\frac{\epsilon}{\bar{\zeta}^2}}\frac{1-(1-\eta)(1-p)}{1-\sqrt{\epsilon/\bar{\zeta}^2}} \leq \sqrt{\frac{\epsilon}{\bar{\zeta}^2}}\frac{p+\eta}{1-\sqrt{\epsilon/\bar{\zeta}^2}}$$

And therefore for $\epsilon \leq \frac{\bar{\zeta}^2}{16}$,

$$\eta \leq \frac{\sqrt{\epsilon/\bar{\zeta}^2}\,p}{1-2\sqrt{\epsilon/\bar{\zeta}^2}} \leq 2\sqrt{\epsilon/\bar{\zeta}^2}\,p$$

With this upper bound on η, the inequality (31) gives a lower bound on t:

$$t \geq \frac{\log\frac{n}{\epsilon}}{-2\log(1-\eta)} \geq \frac{\log\frac{1}{\epsilon}}{2\eta} \geq \frac{\bar{\zeta}\log\frac{1}{\epsilon}}{4\sqrt{\epsilon}p}, \tag{34}$$

here we used that $n \geq 1$ and that $\log(1-\eta) \geq -\eta$ for $\eta \leq \frac{4}{5}$. □

F. Additional Experiments to Verify the $\mathcal{O}\left(\frac{1}{T^2}\right)$ Term

In Theorem 2 we proved an upper bound and in Theorem 3 we proved a lower bound, that indicates that in the noiseless ($\bar{\sigma}^2 = 0$) strongly convex case the convergence is not linear when $\bar{\zeta}^2 > 0$. In this section we verify numerically that this rate indeed reflects tightly the convergence behavior of decentralized SGD.

We consider the same setting as in Section 8 before, with $\bar{\sigma}^2 = 0$, $\bar{\zeta}^2 = 10$, $n = 25$, and $d = 10$.

For both ring and 2-d torus (grid), we vary the target accuracy (ϵ) and tune the stepsize to find the smallest number of iterations required (T_ϵ) to achieve this target accuracy. In Figure 3 we depict the results, where x-axis is $\frac{1}{\sqrt{\epsilon}}$ and y-axis is T_ϵ. Based on the Theorem 2 for strongly convex case, ideally each of them should be a line, as we observe in the plots. Moreover, the ratio of the slopes of these lines is $30.2/2.3 = 13.13$ which matches the ratio of the spectral gap of these graphs ($p_{\text{grid}}/p_{\text{ring}} = 0.276/0.021 = 13.142$), as it is shown in Theorems 2 and 3.

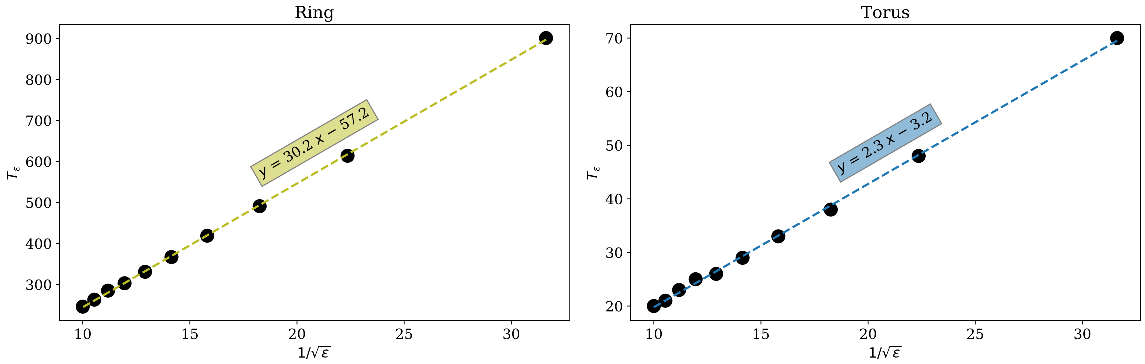

Figure 3. Verifying the $\mathcal{O}\left(\frac{\bar{\zeta}^2}{p^2T^2}\right)$ convergence for the strongly convex noiseless ($\hat{\sigma}^2 = 0$) case. Number of iterations to converge to target accuracy ϵ on ring (left) and 2-d torus (right).