

---

# Optimal Convergence Rate for a Unified Framework of Decentralized SGD

---

Arnaud Watusadisi Mavakala<sup>1</sup>, Anastasia Koloskova<sup>2</sup>, Sebastian U. Stich<sup>2</sup>, Martin Jaggi<sup>2</sup>

<sup>1</sup>African Masters in Machine Intelligence (AMMI), AIMS Senegal

<sup>2</sup>Machine Learning and Optimization Lab (MLO), EPFL, Lausanne, Switzerland.

amavakala@aimsammi.org, anastasia.koloskova@epfl.ch, sebastian.stich@epfl.ch, martin.jaggi@epfl.ch

## Abstract

Decentralized optimization problems are situations where nodes in a communication network privately own a local objective function and communicate with each other using gossip-based methods to minimize the average of these objectives per node. These problems have recently attracted a lot of interest due to the fact that decentralized stochastic optimization methods achieve low cost per iteration, communication efficiency, and data proximity. In this paper, we present an improvement to the unified convergence analysis. This framework comes to cover a wide variety of decentralized stochastic optimization methods because, in the past, these methods were developed separately, had different applications and required different intuitions. The idea being to show an interpolation of the rates between non-identically distributed (heterogeneous) data and i.i.d. data through the parameter  $\bar{\xi}^2$ , we specify in this paper that it is the i.i.d. data that are treated by taking the parameter  $\bar{\xi}^2 = 0$ . We show that the convergence rate in the particular case where  $\bar{\xi}^2 = 0$  which is  $\mathcal{O}\left(\frac{L}{\mu p} \log \frac{1}{\epsilon}\right)$ , can be improved and become  $\mathcal{O}\left(\left(\frac{L}{\mu} + \frac{1}{p}\right) \log \frac{1}{\epsilon}\right)$  while eliminating the noise thanks to the parameter  $\bar{\sigma}^2 = 0$ . Our assertion is based on a simple quadratic objective function and we provide simulation results that show a close agreement between our derived theoretical convergence rate and the numerical results.

## 1 Introduction

The distribution of learning data across many client devices is strongly considered in the large-scale machine learning scenario, such as geo-distributed data centers. The accuracy of the machine learning model trained by decentralized learning methods can be the same as in the case where all the data is aggregated on a single server [1, 16]. Key aspects such as data ownership, privacy, fault tolerance, and scalability show how decentralized machine learning model training can offer many advantages over traditional centralized approaches. It is within this framework that federated learning has emerged in which training is orchestrated by a single entity that communicates with users. Fully decentralized approaches, although on a smaller scale than federated learning, have also been suggested recently. It should be noted that it is important to understand the theory of decentralized stochastic gradient descent (SGD) in order to predict the training performance of SGD and to help design optimal decentralized training schemes for machine learning tasks, as such decentralized training comes with several challenges, namely: high cost on communication [11, 33, 37], a need related to time-varying topologies [1, 22] and data heterogeneity [8, 13–15].

The difference in SGD analyses between decentralized and centralized frameworks is that for decentralized frameworks, SGD analyses are sometimes application specific, unlike the centralized framework, where SGD convergence is well understood [3, 21, 27]. Some works like, [36] had to propose a decentralized framework for decentralized optimization with non-heterogeneous data, but

also [14] which had them study the decentralized SGD for non-convex heterogeneous parameters and [10] which on the other hand had to propose a unified framework which covers these particular cases previously proposed. We propose here an improvement of the convergence rates defined in [10] for a unified framework. Indeed, the study of convergence rates for a unified framework is much more interesting and powerful than when this study of its particular cases is done in isolation. Not only can we recover many analyses and results from previous work, but we can also show improved rates in a more general framework. Taking into account what has been proposed by [10], we prove here the improvement of the unified framework for the strongly convex case and in the presence of i.i.d. data.

### Main contributions.

- We present a general improvement of the convergence rate of strongly convex functions which becomes  $\mathcal{O}\left(r_0 L \exp\left[-\min\left\{\frac{\mu}{L}, p\right\}T\right] + \frac{\bar{\sigma}^2}{n\mu T} + \frac{L(p\bar{\sigma}^2 + \bar{\xi}_1^2)}{\mu^2 p^2 T^2}\right)$  for the unified framework presented by [10] for decentralized gossip-based SGD methods accounting for randomly sampled time-varying mixing distributions.
- We demonstrate the efficiency of our results through the specific case of a simple quadratic objective function for which we derived its rate of convergence without noise  $\bar{\sigma}^2 = 0$  which gave us  $\mathcal{O}\left(r_0 \frac{L}{p} \exp\left[\frac{\mu p T}{L}\right] + \frac{L\bar{\xi}_2^2}{\mu^2 p^2 T^2}\right)$  and we then tried to improve this rate, we thus obtained the improvement which is  $\mathcal{O}\left(r_0 L \exp\left[-\min\left\{\frac{\mu}{L}, p\right\}T\right] + \frac{L\bar{\xi}_2^2}{\mu^2 p^2 T^2}\right)$ .
- We verify by experiments the robustness of our theoretical results on strongly convex functions by neglecting the impact of noise  $\bar{\sigma}^2 = 0$  and dissimilarity  $\bar{\xi}^2 = 0$  (with iid data) on convergence. Thus, we obtain linear convergence rates for decentralized SGD methods of the number of workers  $n$  for five graph topologies.
- We assert through our numerical results by removing the noise  $\bar{\sigma}^2 = 0$  and with iid data ( $\bar{\xi}^2 = 0$ ) that for D-SGD, the convergence rate case  $\mathcal{O}\left(\left(\frac{L}{\mu} + \frac{1}{p}\right) \log \frac{1}{\epsilon}\right)$  is better than the case  $\mathcal{O}\left(\frac{L}{\mu p} \log \frac{1}{\epsilon}\right)$ .

## 2 Related Work

The beginning of the study of decentralized optimization algorithms dates back to at least the 1980s [34]. Decentralized machine learning is indeed more advantageous than traditional centralized approaches in terms of data ownership, confidentiality, fault tolerance and scalability. Based on the fact that information is not disseminated by a central entity but rather by propagation, decentralized algorithms are therefore sometimes called gossip algorithms [2, 9, 39]. The most popular algorithms are based on decentralized gradient descent [24] which is a most classical decentralized algorithm [24, 40] and the alternating direction multiplier method (ADMM) [5, 38]. In terms of convergence, the decentralized gradient descent algorithm converges slowly for obtaining the optimal solution. Therefore, algorithms based on primal and dual formulations or gradient tracking have been proposed to eliminate the convergence bias in DGD-type algorithms and thus improve the convergence, such as D-ADMM [20], DLM [18], NIDS [15],  $D^2$  [33], Exact Diffusion [41], DIGing [23], GSGT [25], etc.

Recently, the decentralized stochastic gradient descent that has received much attention for its fast convergence [1, 10, 16, 17]. As in [11] where it has been proved that while ignoring higher order terms that its convergence is at rate  $\mathcal{O}(1/(nT))$  on strongly convex functions with  $T$  as the number of iterations. In contrast to the centralized setting, where the convergence of SGD is well understood [3, 21, 27], analyses of SGD in decentralized settings are often application specific and developed in separate ways, with the exception of some recent efforts towards a unified theory at the example of [10] that have considered special cases from [36] and [14] that have studied decentralized SGD for heterogeneous non-convex settings.

In the deterministic (non-stochastic) convex version, some advanced of the optimal algorithms have been developed based on acceleration [6, 29, 35] but also the study of quantization in this

framework has been developed by [28]. In [4], they were only able to have a sublinear rate for smooth and strongly convex objectives. Instead, they considered non-smooth objectives and provided sublinear rates that correspond to the optimal rates up to the logarithmic factor [30]. In the stochastic framework, the rates were derived in [26, 31], with the assumption that the distributions over all nodes are equal. This i.i.d. assumption is nothing but a strong restriction that prohibits most applications of distributed machine learning, e.g. also the federated learning framework [19] and thus an algorithm was developed in [11], where the i.i.d. assumption has been removed.

### 3 Setup

In this section we discuss our different settings and assumptions. Gossip averaging consists of information exchange between connected nodes (neighbors) [14, 16, 36]. Stochastic gradient updates are performed locally on each worker, followed by a consensus operation, where nodes average their values with their neighbors.

**Problem Formulation.** We consider the following decentralized optimization:

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

where each  $f_i$  is a  $L_i$ -smooth function. The components  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$  are distributed among  $n$  nodes and are given in stochastic form:

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

where  $\mathcal{D}_i$  denotes the local data distribution on node  $i \in [n]$ . Indeed, when each  $\mathcal{D}_i$  has a finite number  $m_i$  of elements  $(\xi_i^1, \dots, \xi_i^{m_i})$ . Then  $f_i$  can be rewritten as  $f_i(\mathbf{x}) = \frac{1}{m_i} \sum_{j=1}^{m_i} F_i(\mathbf{x}, \xi_i^j)$ . We thus enter a deterministic distributed optimization problem when  $m_i = 1$ , for each  $i \in [n]$ . Decentralized Gradient descent can be written (in matrix notation) as

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

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  $w_{ij}^{(t)} > 0$ ). For now, let's assume  $W$  is fixed and does not change over time.

#### 3.1 Mixing Matrix

**Definition 3.1 (Mixing matrix)** A symmetric ( $W = W^T$ ) doubly stochastic ( $W\mathbf{1} = \mathbf{1}, \mathbf{1}^T W = \mathbf{1}^T$ ) matrix  $W \in [0, 1]^{n \times n}$  with eigenvalues  $1 = |\lambda_1(W)| > |\lambda_2(W)| \geq \dots \geq |\lambda_n(W)|$  and spectral gap

$$\rho := 1 - |\lambda_2(W)| \in [0, 1].$$

We start by stating the assumption about the quality of the mixing matrices. Indeed, there are several of them but we take a generalization of the previous versions.

**Assumption 1 (Mixing matrix)** Every sample of the mixing matrix  $W = \{w_{ij}\} \in \mathbb{R}^{n \times n}$  is doubly stochastic and there exists a parameter  $\rho$  such that :

$$\mathbb{E} \|XW - \bar{X}\|_F^2 \leq (1 - \rho) \|X - \bar{X}\|_F^2, \quad (2)$$

for all  $X$ . This assumption covers a variety of D-SGD parameters with a fixed mixing matrix with spectral gap  $\rho$ , with parameter  $p = 1 - (1 - \rho)^2$  [10]. We define  $\bar{\mathbf{x}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  and  $X := [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ , and  $\bar{X} := [\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}] \equiv X \frac{\mathbf{1}\mathbf{1}^T}{n}$ .

**Proof:** To prove, we use the preservation of the mean over all iterations of the algorithm defined by the identity  $\bar{X} = X \frac{11^T}{n}$ . More explicitly, we have:

$$XW \frac{11^T}{n} = X \frac{11^T}{n} = \bar{X} \frac{11^T}{n} = \bar{X}. \quad (3)$$

Hence we have:

$$\begin{aligned} \|XW - \bar{X}\|_F^2 &= \|XW - \bar{X} - \bar{X} + \bar{X}\|_F^2 \\ &= \left\| XW - X \frac{11^T}{n} - X \frac{11^T}{n} + \bar{X} \frac{11^T}{n} \right\|_F^2 \\ &\leq (1 - \rho)^2 \left\| \left(1 - \frac{11^T}{n}\right) X \right\|_F^2 \\ &\leq (1 - \rho)^2 \|X - \bar{X}\|_F^2 \end{aligned}$$

□

Decentralized topologies converge quickly only when the iterations of individual nodes remain sufficiently close, and to compute this distance, we use the consensus distance which is defined by :

$$\Xi_t := \frac{1}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\|^2$$

### 3.2 Assumptions on $\mathbf{f}$

In this subsection, we list the main assumptions on the optimization problem (1). In order to simplify and facilitate the presentation, we will base ourselves on the most common standard assumptions, but the analyses could be strengthened for many particular cases, following techniques developed in other works. We know that every  $f_i$  is a  $L_i$ -smooth function. So  $f$  is smooth but also sometimes, we will additionally assume that the objective function is convex.

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

$$f_i(\mathbf{x}) - f_i(\mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \leq \langle \nabla f_i(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle. \quad (4)$$

**Assumption 3** [ $L$ -smoothness]. Each function  $F_i(\mathbf{x}, \xi) : \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}, i \in [n]$  is differentiable for each  $\xi \in \text{supp}(\mathbf{D}_i)$  and there exists a constant  $L \geq 0$  such that for each  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \xi \in \text{supp}(\mathbf{D}_i)$ :

$$\|\nabla F_i(\mathbf{y}, \xi) - \nabla F_i(\mathbf{x}, \xi)\| \leq L_i \|\mathbf{x} - \mathbf{y}\|. \quad (5)$$

**Assumption 4** [ $L$ -smoothness]. Each function  $f_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}, i \in [n]$  is differentiable for each  $\xi \in \text{supp}(\mathbf{D}_i)$  and there exists a constant  $L \geq 0$  such that for each  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ :

$$\|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x})\| \leq L_i \|\mathbf{x} - \mathbf{y}\|. \quad (6)$$

Hence, assumption 4 is more general than assumption 3.

**Assumption 5** [Bounded noise  $\bar{\sigma}^2$  and diversity  $\bar{\xi}^2$ ]. There exists constant  $\bar{\sigma}^2, \bar{\xi}^2$  such that  $\forall \mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_\xi \|\nabla F_i(\mathbf{x}_i, \xi_i) - \nabla f_i(\mathbf{x}_i)\|_2^2 \leq \bar{\sigma}^2. \quad (7)$$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_\xi \|\nabla f_i(\mathbf{x}_i) - \nabla f(\mathbf{x}_i)\|_2^2 \leq \bar{\xi}^2. \quad (8)$$

### 3.3 Convergence bounds

In this section, we present convergence results obtained in [10] for decentralized variants of SGD. Decentralized optimization problems are settings where nodes in a communication network privately own a local objective function and communicate with each other using gossip-based methods to minimize the average of these objectives per node. In centralized topologies (corresponding to a star graph), we find a significant bottleneck for the node, which is not the case in decentralized topologies where they are avoided while offering a significantly improved scalability potential. The convergence rate of D-SGD related to the network topology proposed by [10], was defined as follows:

#### 3.3.1 Upper bounds

**Theorem 1** [10]. *There exists a step size (constant) potentially depending on  $\epsilon$  such that the accuracy can be reached after at most the number of iterations following  $T$ , for the D-SGD algorithm with mixing matrices with assumption 1 and any target accuracy  $\epsilon > 0$ ,*

**Strongly-Convex:** *If  $\mu > 0$  then  $\sum_{t=0}^T \frac{w_t}{W_T} \mathbb{E}(f(\mathbf{x}_t) - f^*) + \mu \mathbb{E}\|\mathbf{x}_{T+1} - \mathbf{x}^*\|^2 \leq \epsilon$  after*

$$\mathcal{O}\left(\frac{\hat{\sigma}^2}{\mu n \epsilon^2} + \frac{\sqrt{L}(\hat{\xi}\tau + \hat{\sigma}\sqrt{p\tau})}{\mu p \sqrt{\epsilon}} + \frac{L\tau}{\mu p} \log \frac{1}{\epsilon}\right)$$

iterations, for positive weights  $w_t$ .

#### 3.3.2 Lower bounds

Here was shown that  $\xi$  dependent terms are necessary for the strongly convex framework and cannot be removed by an improved analysis.

**Theorem 2** [10]. *For  $n > 1$ , there exists strongly convex and smooth functions  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $i \in [n]$  with  $L = \mu = 1$  and without stochastic noise  $\sigma^2$ , such that D-SGD Algorithm for every constant mixing matrix  $W^{(t)} = W$  with  $p < 1$  (Assumption 1) for  $T=1$ , requires*

$$T = \Omega\left(\frac{\hat{\xi}(1-p)}{\sqrt{\epsilon p}}\right)$$

## 4 Background

In this section, we do not talk about the decentralized framework but rather about the convergence of the gradient descent at a node, i.e., for a single loss function, in order to allow us to better see how the convergence at a node looks like, and therefore developed further for the decentralized framework in the sections below. We also discuss how the convergence rate of the average gossip is defined. One of the popular generalizations of strong convexity in the literature is the Polyak-Lojasiewicz condition [7, 12]. We first define this condition and then establish the convergence of SGD for a loss function satisfying it. We assume that the function  $f$  has a minimizer and denote it  $f^* = \min_{\mathbf{x}} f(\mathbf{x})$ .

**Lemma 3** *If  $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is  $\mu$ -strongly convex then it also satisfies the Polyak-Lojasiewicz condition, that :*

$$\|\nabla f(\mathbf{x})\|_2^2 \geq 2\mu(f(\mathbf{x}) - f^*), \forall \mathbf{x} \in \mathbb{R}^d,$$

where  $f^* = \min_{\mathbf{x}} f(\mathbf{x})$ .

### 4.1 Convergence of Gradient Descent on a single function

Gradient descent generally assumes that the loss function is differentiable and convex. If the loss function is not convex, the solution may get stuck in a local optimum and not find the global optimum. The stochastic gradient descent algorithm is as follows:

---

**Algorithm 1** CLASSIC SGD

---

Input  $\mathbf{x}_0$ , stepsizes  $\{\eta_t\}_{t=0}^{T-1}$ , number of iterations  $T$ For  $t \in 0 \dots T$ :

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \nabla F(\mathbf{x}^{(t)}, \xi_t)$$

\*stochastic gradient descent

End For

---

**Lemma 4** Let  $f_1 : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  is  $\mu$ -strongly convex with coordinate-wise  $L_1$ -Lipschitz continuous gradient. Let  $\mathbf{x}_t, \mathbf{x}_{t+1} \in \mathbb{R}^n$  denote two successive iterates generated by the Algorithm 1. Then

$$f_1(\mathbf{x}_{t+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^t (f_1(\mathbf{x}_0) - f^*).$$

**4.2 Convergence rate of gossip averaging**

We still know that for the Gossip algorithm (Assumption 1) and applying this rule recursively on the initial point  $\mathbf{x}_0$ , we have:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq (1 - p)^t \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

**5 Decentralized optimization with Gradient Descent and Gossip Averaging**

In this section, we derive the convergence rate of D-SGD in a general case. We simplify the proof of the descent lemma for convex cases presented by [10] step by step for better understanding. We prove the recursion lemma for the consensus distance for strongly convex functions which is an important lemma and is a simplified version of the one presented by [10]. Then, the key is to be able to combine the two lemmas (descent lemma for convex cases and the recursion lemma for the consensus distance for strongly convex functions) for obtaining the desired improved convergence rate. We analyze the following algorithm :

---

**Algorithm 2** DECENTRALIZED SGD (MATRIX NOTATION)

---

Input  $X^{(0)}$ , stepsizes  $\{\eta_t\}_{t=0}^{T-1}$ , number of iterations  $T$ , mixing matrix distribution  $\mathcal{W}^{(t)}$  for  $t \in [T]$ For  $t \in 0 \dots T$ :**Sample**  $W^{(t)} \sim \mathcal{W}^{(t)}$ 

$$X^{(t+1)} = X^{(t)} - \eta_t \nabla F(X^{(t)}, \xi_t)$$

\*stochastic gradient descent

$$X^{(t+1)} = X^{(t+\frac{1}{2})} W^{(t)}$$

\*gossip averaging

End For

---

we use matrix notation:  $X^{(t)} = [\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}, \dots, \mathbf{x}_n^{(t)}]$  (eq. (11) from [10]) where we collect the states of  $n$  workers in one matrix and  $\nabla f(X) = [\nabla f_1(\mathbf{x}_1), \nabla f_2(\mathbf{x}_2), \dots, \nabla f_n(\mathbf{x}_n)]$ . The matrix notation in the Algorithm 2 for the gossip averaging is equivalent to :

$$\mathbf{x}_i^{(t+1)} = \sum_{j \in \mathcal{I}^{(t)}} w_{ij}^{(t)} \left( \mathbf{x}_i^{(t)} - \eta_t \nabla f_i(\mathbf{x}_i^{(t)}) \right), \quad (9)$$

for each node  $i \in [n]$ .

The idea is to combine gradient descent (optimization progress) and gossip averaging. To do so, we will try to derive an upper bound on the expected progress  $r_t = \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2$ , measured as the distance to the optimum in the strongly convex case ( $\mu > 0$ ). This work refers to a fixed sampling distribution, which means that the mixing matrix  $W$  is constant over the iterations ( $\tau = 1$  and  $\mathcal{W}^{(t)} \equiv \mathcal{W}$ ).

**Lemma 5** (Descent Lemma for convex cases [10]). Under Assumptions 2, 4 and 7, the average  $\bar{\mathbf{x}}^{(t)} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(t)}$  of the iterates of Algorithm 2 with the stepsize  $\eta_t \leq \frac{1}{12L}$  satisfy

$$R_{t+1} \leq (1 - m\eta_t)R_t - b\eta_t e_t + c\eta_t^2 + Q\eta_t \Xi_t; \quad (10)$$

where,  $\Xi_t := \frac{1}{n} \mathbb{E}_t \sum_{i=1}^n \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2$ ,  $R_{t+1} = \mathbb{E}_{\xi_1^t, \dots, \xi_n^t} \|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2$ ,  $R_t = \mathbb{E} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2$ ,  $e_t = f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)$ ,  $m = \frac{\mu}{2}$ ,  $b = 1$ ,  $c = \frac{\sigma^2}{n}$  and  $Q = 3L$ .

**Lemma 6** (Recursion for consensus distance). *Under Assumptions 1 and 7, if in addition  $F_i$  are convex and  $\bar{\xi}_1^2 \geq \mathbb{E} \|\nabla f_i(X^{(t)}) - \nabla f(X^{(t)})\|_2^2$ , we have :*

$$\Xi_{t+1} \leq \left(1 - \frac{p}{2}\right) \Xi_t + \frac{3\eta_t^2}{p} \bar{\xi}_1^2 + \eta_t^2 \bar{\sigma}^2, \quad (11)$$

From the combination between (10) and (11) as  $\psi_{t+1} := R_{t+1} + a\Xi_{t+1}$ , we can deduce the following lemma:

**Lemma 7** *Let  $(r_t)_{t \geq 0}$  and  $(e_t)_{t \geq 0}$  be sequences of positive numbers satisfy in*

$$r_{t+1} \leq (1 - \min\{m\eta_t, k\})r_t - B\eta_t e_t + N\eta_t^2 + A\eta_t^3 \quad (12)$$

*For some constants  $B, m > 0$ ,  $N, A, T \geq 0$ , then there exists a constant stepsize  $\eta_t = \eta < \frac{1}{F}$  such that for weights  $w_t = (1 - \max\{m\eta_t, k\})^{-(t+1)}$  and  $W_T := \sum_{t=0}^T w_t$  it holds that*

$$\frac{1}{2W_T} \sum_{t=0}^T B e_t w_t + a r_{T+1} \leq \left( r_0 F \exp \left[ -\min\left\{\frac{m}{F}, k\right\}(T+1) \right] + \frac{N}{mT} + \frac{A}{m^2 T^2} \right)$$

**Proof:** Multiplying (12) by  $w_t$  and dividing by  $\eta_t$ , we get

$$\begin{aligned} \frac{w_t}{\eta_t} r_{t+1} &\leq \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - B \frac{w_t}{\eta_t} \eta_t e_t + N \frac{w_t}{\eta_t} \eta_t^2 - \frac{w_t}{\eta_t} A \eta_t^3 \\ &\leq \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - B w_t e_t + N w_t \eta_t + A w_t \eta_t^2 \end{aligned}$$

We have,

$$B w_t e_t \leq \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - \frac{w_t}{\eta_t} r_{t+1} + N w_t \eta_t + A w_t \eta_t^2$$

Summing up and dividing by  $W_T = \sum_{t=0}^T w_t$ , we have:

$$\frac{1}{W_T} \sum_{t=0}^T B w_t e_t \leq \frac{1}{W_T} \sum_{t=0}^T \left( \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - \frac{w_t}{\eta_t} r_{t+1} \right) + \frac{N}{W_T} \sum_{t=0}^T w_t \eta_t + \frac{A}{W_T} \sum_{t=0}^T w_t \eta_t^2$$

and hence,

$$\frac{1}{2W_T} \sum_{t=0}^T B w_t e_t + \frac{w_T r_{T+1}}{W_T \eta} \leq \frac{r_0}{W_T \eta} + N \eta + A \eta^2$$

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

$$\begin{aligned} \frac{1}{2W_T} \sum_{t=0}^T B w_t e_t + a r_{T+1} &\leq (1 - \min\{m\eta, k\})^{-(T+1)} \frac{r_0}{\eta} + N \eta + A \eta^2 \\ &\leq \frac{r_0}{\eta} \exp[-\min\{m\eta, k\}(T+1)] + N \eta + A \eta^2 \end{aligned}$$

Lemma follows by tuning  $\eta$  the same way as in [32], if  $\frac{1}{F} \leq \frac{\ln(\min\{2, m^2 r_0 T^2 / N\})}{mT}$ , we choose  $\eta = \frac{1}{F}$

$$\mathcal{O} \left( r_0 F \exp \left[ -\min\left\{\frac{m}{F}, k\right\}(T+1) \right] + \frac{N}{F} + \frac{A}{F^2} \right) \leq \mathcal{O} \left( r_0 F \exp \left[ -\min\left\{\frac{m}{F}, k\right\}(T+1) \right] + \frac{N}{mT} + \frac{A}{m^2 T^2} \right)$$

We have for  $m = \frac{\mu}{2}$ ,  $k = \frac{p}{4}$ ,  $N = \frac{\bar{\sigma}^2}{2}$  and  $A = \left( \frac{36\bar{\xi}_1^2}{p^2} + \frac{12\bar{\sigma}^2}{p} \right) L$ ,

$$\mathcal{O} \left( r_0 L \exp \left[ -\min \left\{ \frac{\mu}{L}, p \right\} T \right] + \frac{\bar{\sigma}^2}{n\mu T} + \frac{L(p\bar{\sigma}^2 + \bar{\xi}_1^2)}{\mu^2 p^2 T^2} \right)$$

□

**Corollary 8** *If  $\bar{\sigma}^2 = 0$  and  $\bar{\xi}^2 = 0$ , then :*

$$\mathcal{O} \left( r_0 L \exp \left[ -\min \left\{ \frac{\mu}{L}, p \right\} T \right] \right)$$

## 6 Simple Quadratics

In this section, we consider a simple quadratic objective function  $f_i(\mathbf{x}) = \frac{1}{2}(a_i \mathbf{x} - b_i)^2$  where  $a_i \in \mathbb{R}$  and  $b_i \equiv 0$ . We seek to obtain the rate of convergence of D-SGD defined by [10] with zero stochastic noise ( $\bar{\sigma} = 0$ ) for this function presented in Theorem 11 and, we then seek to obtain the desired improvement of this rate presented in Theorem 12 while using the advanced theories above.

**Remark 9** *Let each  $f_i$  is  $L_i$ -smooth and  $\mu_i$ -strongly convex, then  $L = \max_i L_i$ ,  $\mu = \min_i \mu_i$  and  $L_{avg} = \frac{1}{n} \sum_{i=1}^n L_i$ .*

**Lemma 10** *Let  $f_i$  be  $L_i$ -smooth and  $L_i = \mu_i = a_i^2$ , then*

$$L_{avg} \leq L_{\max}$$

Let's take the example where  $L_1 = \dots L_{n-1} = 1$ ,  $L_n = n$ . We obtain  $L_{avg} = \frac{1}{n} \sum_{i=1}^n a_i^2$  and  $\sum_{i=1}^n a_i^2 = \sum_{i=1}^{n-1} a_i^2 + a_n^2 = n - 1 + n = 2n - 1$  which implies  $L_{avg} = \frac{1}{n}(2n - 1) = 2 - \frac{1}{n}$  and  $L = \max_i a_i^2 = n$ .

**Theorem 11** *Under global Lipschitz-continuity and strong convexity ( $\mu > 0$ ), there exists a step size  $\eta_t \leq \frac{p}{2\sqrt{6}L}$  such that the accuracy can be reached after at most the number of iterations following  $T$ , for the algorithm 2 with mixing matrices with assumption 1 and any target accuracy  $\epsilon > 0$ . Then  $\sum_{t=0}^T \frac{w_t}{W_T} \mathbb{E}(f(\mathbf{x}_t) - f^*) + \mu \mathbb{E}\|\mathbf{x}_{T+1} - \mathbf{x}^*\|^2 \leq \epsilon$  after*

$$\mathcal{O} \left( r_0 \frac{L}{p} \exp \left[ \frac{\mu p T}{L} \right] + \frac{L\bar{\xi}_2^2}{\mu^2 p^2 T^2} \right)$$

where  $\mathbb{E} \left\| \nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f(\mathbf{x}_i^{(t)}) \right\|_2^2 \leq \bar{\xi}_2^2$

We see through Theorem 11 that for the simple quadratic objective function, we could find its rate of convergence of the D-SGD. Indeed, we found that the rate of convergence obtained is the one that has been proved and demonstrated by [10] with  $\bar{\sigma} = 0$ . We then tried to find an improvement of the result of the convergence rate of the theorem 11 by using the idea of obtaining lemma 7. This improvement is presented in the following theorem.

**Theorem 12** *Under global Lipschitz-continuity and strong convexity ( $\mu > 0$ ), there exists a step size  $\eta_t \leq \frac{1}{4L}$  such that the accuracy can be reached after at most the number of iterations following  $T$ , for the algorithm 2 with mixing matrices with Assumption 1 and any target accuracy  $\epsilon > 0$ . Then  $\sum_{t=0}^T \frac{w_t}{W_T} \mathbb{E}(f(\mathbf{x}_t) - f^*) + \mu \mathbb{E}\|\mathbf{x}_{T+1} - \mathbf{x}^*\|^2 \leq \epsilon$  after*

$$\mathcal{O} \left( r_0 L \exp \left[ -\min \left\{ \frac{\mu}{L}, p \right\} T \right] + \frac{L\bar{\xi}_2^2}{\mu^2 p^2 T^2} \right)$$

where  $\mathbb{E} \left\| \nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f(\mathbf{x}_i^{(t)}) \right\|_2^2 \leq \bar{\xi}_2^2$

We can see that by considering the special case where  $\bar{\xi}_2^2 = 0$  for this specific function it is really possible to improve the convergence rate  $\mathcal{O} \left( \frac{L}{\mu p} \log \frac{1}{\epsilon} \right)$  in  $\mathcal{O} \left( \left( \frac{L}{\mu} + \frac{1}{p} \right) \log \frac{1}{\epsilon} \right)$ . It should be noted that  $\bar{\xi}_1^2 \neq \bar{\xi}_2^2$  because  $\bar{\xi}_1^2$  has been defined for the general case.



## 7 Experiments results

In this section, we present experiments related to the comparison between our proven optimal convergence on a strong monotonic problem for the i.i.d. ( $\xi^2 = 0$ ) data case of the gossip algorithm and the one found in [10], using its theoretical Convergence to validate our performance.

**Setup.** To validate our theory on the optimal Convergence, we consider a distributed least squares objective with  $f_i(\mathbf{x}) := \frac{1}{2} \|A_i \mathbf{x} - b_i\|^2$  where  $A_i^2 = \frac{i^2}{n} I_d$  and  $b_i \sim \mathcal{N}(0, \frac{\xi^2}{i^2} I_d)$  with  $\bar{\xi}^2$  which allows us to control the similarity of the functions, we treat only the case i.i.d. ( $\xi^2 = 0$ ) and  $\bar{\sigma}^2$  which on the other hand allows to control the stochastic noise by adding a Gaussian noise to each stochastic gradient. We consider five network topologies, namely the ring, the centralized graph, the chain graph, the tree graph and the fully connected graph. All connections between nodes in our topologies are of the form  $w_{ij} = w_{ji} = \frac{1}{\deg(i)+1} = \frac{1}{\deg(j)+1}$  with  $\{i, j\} \in E$ . We were able to evaluate our experiment by changing the number of nodes and the dimension many times, by fixing the step sizes for the cases of the theoretical convergences and by setting the step sizes individually in each experiment for the gossip algorithm.

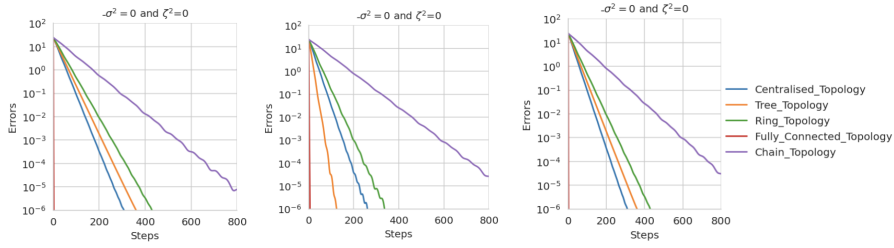


Figure 1: Convergence of  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2$  for different topology on  $n = 28$  nodes and  $d = 24$ , at target accuracy  $\epsilon = 10^{-3}$  (left) and  $\epsilon = 10^{-5}$  (middle and right). Stepsizes were tuned for each experiment individually to reach target accuracy in as few iterations as possible.

We observe in Figure 1 linear rates for all graph topologies while noting that the graph topology has no impact on the number of iterations required. We had to start with the same initial learning rate  $\frac{1}{n}$ , where  $n$  is the number of workers (middle) and with a different initial learning rate for each topology (left and right) to then perform the tuning and thus obtain the best learning rate to reach our target. We found that for the case where the initial learning rates are the same, the tree topology requires less number of iterations in convergence than for the centralizes topology and for the case where the initial learning rates are different, it is the opposite.

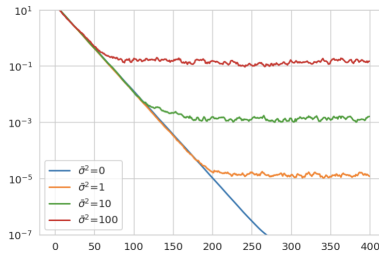


Figure 2: Impact of stochastic noise for the ring topology with  $n = 26$  nodes and  $d = 22$ . The stepsize was set to  $10^{-2}$ .

**Convergence Behavior.** In Figure 2, we show the convergence of the algorithm on the ring topology with a fixed stepsize. We observe a linear convergence at the noise parameter. The convergence is thus slightly affected by the noise, but interestingly, we see a linear convergence with small oscillations and without oscillations when there is no noise. Instead, in Figure 3, we see that increasing

the number of workers for the Ring topology only increases the number of iterations to reach the target.

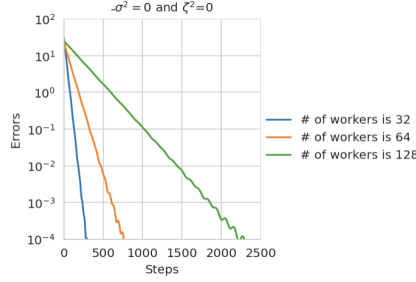


Figure 3: Convergence of  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2$  for the ring topology on  $n = [32, 64, 128]$  and  $d = 25$ , at target accuracy  $\epsilon = 10^{-3}$ . Stepsizes were tuned for each experiment individually to reach target accuracy in as few iterations as possible.

**Comparison with the Theoretical Convergence Rate defined in [10].** Figure 4 shows the comparison between the gossip algorithm and the theoretical Convergence rate that we have defined as follows  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2 / \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \leq (1 - \frac{\mu p}{L})^t$  where  $\mu$  is constant and  $L = n$ , scales in  $n$  on the ring topology while setting the stepsizes in the case of the gossip algorithm and fixing it for the theoretical Convergence rate. We have seen that the gossip algorithm converges linearly (left), faster (right) and requires fewer iterations to converge than the proposed theoretical convergence rate. In order to better notice this optimal Convergence, we have presented in Figure 5, the Convergence of the gossip algorithm and the rate of theoretical Convergence according to the minimum errors obtained for different numbers of workers. Indeed, we see that in the Convergence of the gossip algorithm (left), the mean square error (consensus distance) committed between the average gossip and the optimal value for each iteration, increases with the number of workers. Note that the smaller the consensus distance, the higher the prediction accuracy. Thus, we notice from the number of nodes 25 to the number of nodes 50, a multiplied growth of  $10^{-2}$ , from the number of nodes 50 to the number of nodes 75, a multiplied growth of  $10^{-4}$ , from the number of nodes 75 to the number of nodes 125, a multiplied growth of  $10^{-1}$  before noticing a kind of stability for the number of remaining nodes. Comparing with the theoretical convergence rate (on the right), we can see that the theoretical convergence rate gives large values in terms of average error and therefore not optimal and that its convergence is higher than that of the gossip algorithm.

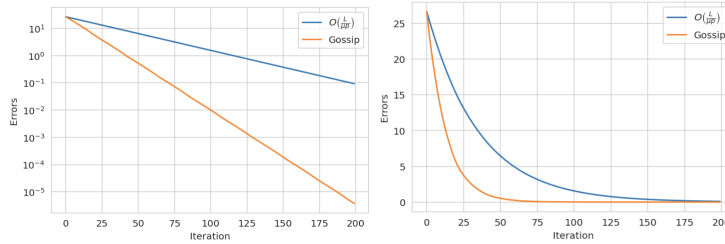


Figure 4: Convergence of  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2$  for the ring topology in  $n = 25$  nodes,  $d = 24$ . The stepsizes were set individually for each number of nodes for the Gossip algorithm and  $\eta = \frac{p}{L}$  for the theoretical convergence.

We can see that the optimal convergence rate of D-SGD is  $\mathcal{O}\left(\left(\frac{L}{\mu} + \frac{1}{p}\right) \log \frac{1}{\epsilon}\right)$  by taking  $\bar{\sigma}^2 = 0$  and  $\bar{\sigma}^2 = 0$ . This can be seen on figure 5 (right) and in figure 6, we added the theoretical convergence rate defined by  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2 / \|\mathbf{x}^0 - \mathbf{x}^*\|^2 \leq (1 - \frac{\mu}{L})^t$  to make the comparison even better and to confirm the lemma 6 in the general case and in the particular case the theorem 12 developed in our work. Indeed, we find that the theoretical convergence rate with step size  $\frac{1}{L}$  obtains lower

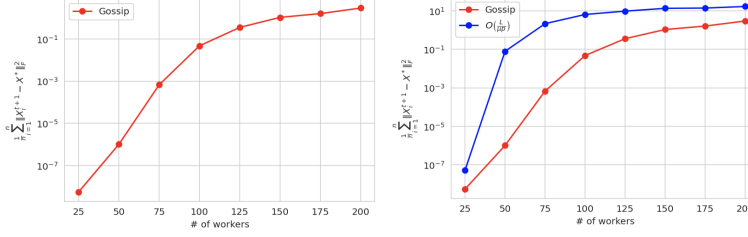


Figure 5: Convergence Rate compared to the minimum value of  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2$  for the ring topology as a function of different number of nodes, at the target accuracy  $\epsilon = 10^{-5}$  and  $d = 24$ . The step sizes were tuned individually for each number of nodes for the Gossip algorithm and  $\eta = \frac{p}{L}$  for the theoretical convergence.

error values than the others because it is the convergence rate of a single node.

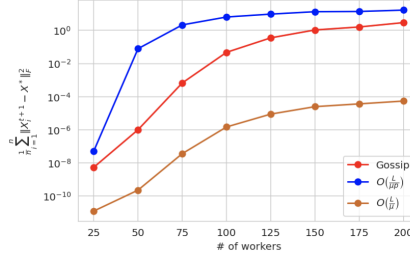


Figure 6: Convergence rate compared to the minimum value of  $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(t+1)} - \mathbf{x}^*\|^2$  for the ring topology for each numbers of nodes,  $d = 24$ . The step sizes were set individually for each number of nodes for the Gossip algorithm and for the theoretical convergence, we fix two stepsizes which are  $\eta = \frac{p}{L}$  and  $\eta = \frac{1}{L}$ .


## 8 Conclusion

We presented the convergence rate improvement of the unified framework for analyzing decentralized SGD methods proposed by [10] while presenting the best known convergence guarantees. We had to use a method to combine the two important lemmas, namely the descent lemma for the convex case and the recursion lemma for the consensus distance, in order to obtain this desired improvement in the convergence rate. To assert this improvement, we performed the test on a simple quadratic objective function and not only recovered the rate of convergence that [10] presented on the function but also managed to improve it. Our experimental results show that when noise and dissimilarity are zero, the decentralized SGD methods have linear rates of the number of workers  $n$  for all graph topologies that have been presented. As future work, we could improve the convergence rate of the unified framework using gradient compression techniques or propose a unified framework and its rate improvement in the case of asynchronous communication.

## References

- [1] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pages 344–353. PMLR, 2019.
- [2] Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.
- [3] Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

- [4] Thinh T Doan, Joseph Lubars, Carolyn L Beck, and R Srikant. Convergence rate of distributed random projections. *IFAC-PapersOnLine*, 51(23):373–378, 2018.
- [5] Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE conference on decision and control*, pages 3671–3676. IEEE, 2013.
- [6] Dušan Jakovetić, Joao Xavier, and José MF Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146, 2014.
- [7] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.
- [8] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. 2019.
- [9] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 482–491. IEEE, 2003.
- [10] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.
- [11] Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019.
- [12] Yunwen Lei, Ting Hu, Guiying Li, and Ke Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE transactions on neural networks and learning systems*, 31(10):4394–4400, 2019.
- [13] Gen Li and Irina Gaynanova. A general framework for association analysis of heterogeneous data. *The Annals of Applied Statistics*, 12(3):1700–1726, 2018.
- [14] Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication-efficient local decentralized sgd methods. *arXiv preprint arXiv:1910.09126*, 2019.
- [15] Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506, 2019.
- [16] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *arXiv preprint arXiv:1705.09056*, 2017.
- [17] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pages 3043–3052. PMLR, 2018.
- [18] Qing Ling, Wei Shi, Gang Wu, and Alejandro Ribeiro. Dlm: Decentralized linearized alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(15):4051–4064, 2015.
- [19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [20] Joao FC Mota, Joao MF Xavier, Pedro MQ Aguiar, and Markus Püschel. D-admm: A communication-efficient distributed algorithm for separable optimization. *IEEE Transactions on Signal processing*, 61(10):2718–2723, 2013.
- [21] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24:451–459, 2011.
- [22] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

- [23] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.
- [24] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- [25] Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.
- [26] Michael Rabbat. Multi-agent mirror descent for decentralized stochastic optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 517–520. IEEE, 2015.
- [27] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and randomize: From value to algorithms. 2012.
- [28] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. Quantized decentralized consensus optimization. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5838–5843. IEEE, 2018.
- [29] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pages 3027–3036. PMLR, 2017.
- [30] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for non-smooth distributed optimization in networks. *arXiv preprint arXiv:1806.00291*, 2018.
- [31] Ohad Shamir and Nathan Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 850–857. IEEE, 2014.
- [32] Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.
- [33] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu.  Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018.
- [34] John Nikolas Tsitsiklis. Problems in decentralized decision making and computation. Technical report, Massachusetts Inst of Tech Cambridge Lab for Information and Decision Systems, 1984.
- [35] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–37. IEEE, 2020.
- [36] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.
- [37] Jianyu Wang and Gauri Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in local-update sgd. *Proceedings of Machine Learning and Systems*, 1:212–229, 2019.
- [38] Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450. IEEE, 2012.
- [39] Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.
- [40] Bo Yang and Mikael Johansson. Distributed optimization and games: A tutorial overview. *Networked Control Systems*, pages 109–148, 2010.
- [41] Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H Sayed. Exact diffusion for distributed optimization and learning—part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2018.

## Appendix

### Lemma 4. .

**Proof:** From the second-order Taylor expansion, we have that:

$$f_1(\mathbf{y}) = f_1(\mathbf{x}) + \langle \nabla f_1(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T H_f(z)(\mathbf{y} - \mathbf{x})$$

Strong convexity implies there exists a constant  $L > 0$  such that:

$$f_1(\mathbf{y}) \leq f_1(\mathbf{x}) + \langle \nabla f_1(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

For  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t g_t$ , where  $g_t = \nabla f_1(\mathbf{x}_t)$  and  $\mathbf{x} = \mathbf{x}_t$  we get:

$$\begin{aligned} f_1(\mathbf{x}_t - \eta_t g_t) &\leq f_1(\mathbf{x}_t) + \langle \nabla f_1(\mathbf{x}_t), \mathbf{x}_t - \eta_t g_t - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_t - \eta_t g_t - \mathbf{x}_t\|_2^2 \\ &= f_1(\mathbf{x}_t) + \langle \nabla f_1(\mathbf{x}_t), -\eta_t g_t \rangle + \frac{L}{2} \|\eta_t g_t\|_2^2 \\ &= f_1(\mathbf{x}_t) - \eta_t \|g_t\|_2^2 + \frac{L}{2} \eta_t^2 \|g_t\|_2^2 \end{aligned}$$

When we take  $\eta_t = \frac{1}{L}$ , subtracting  $f^*$  from both sides and applying Lemma 3, we get :

Subtracting  $f^*$  from both sides and applying Lemma 4, we get :

$$\begin{aligned} f_1(\mathbf{x}_{t+1}) - f^* &\leq f_1(\mathbf{x}_t) - f^* - \frac{1}{2L} \|g_t\|_2^2 \\ &= f_1(\mathbf{x}_t) - f^* - \frac{\mu}{L} (f_1(\mathbf{x}_t) - f^*) \\ &= \left(1 - \frac{\mu}{L}\right) (f_1(\mathbf{x}_t) - f^*) \end{aligned}$$

Applying this rule recursively on the initial point  $\mathbf{x}_0$ , we get:

$$f_1(\mathbf{x}_{t+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^t (f_1(\mathbf{x}_0) - f^*)$$

Thus,  $f_1(\mathbf{x}_{t+1}) - f^* \leq \epsilon$  when

$$t \geq \frac{L}{\mu} \log \left( \frac{1}{\epsilon} \right)$$

So the convergence rate for a single descent gradient is  $\mathcal{O} \left( \frac{L}{\mu} \log \frac{1}{\epsilon} \right)$ . □

### Lemma 5 (Descent Lemma for convex cases [10]).

**Proof:** Thus, recall Assumption 1 where it was signified that a gossip averaging step with the mixing matrix  $W$  preserves the iteration mean. We need therefore solve  $\mathbb{E}_{\xi_1^t, \dots, \xi_n^t} \|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2$  to obtain the form of this bound.

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

We know that  $f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ , where  $f_i(\mathbf{x}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} F_i(\mathbf{x}, \xi_i)$

$$XW \frac{11^T}{n} = X \frac{11^T}{n} = \hat{X} \frac{11^T}{n} = \hat{X}. \quad (13)$$

Let's add  $\pm \frac{\eta_t}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)$  and we know that  $\|a + b\|^2 = \|a\|^2 + 2\langle a, b \rangle + \|b\|^2$ , we have:

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

1.  $i_1 = \left\| \bar{\mathbf{x}}^t - \mathbf{x}^* - \frac{\eta_t}{n} \sum_{i=1}^n \nabla f_i(\bar{\mathbf{x}}_i^t) \right\|^2$  and we use  $\|a + b\|^2 = \|a\|^2 + 2\langle a, b \rangle + \|b\|^2$ , we have :

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

Let us first estimate  $T_1 = \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) \right\|^2$ . Add  $\pm \nabla f_i(\bar{\mathbf{x}})$

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

And for  $T_2 = -2\eta_t \left\langle \bar{\mathbf{x}}^t - \mathbf{x}^*, \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) \right\rangle = \frac{-2\eta_t}{n} \sum_{i=1}^n (\bar{\mathbf{x}}^t \nabla f_i(\mathbf{x}_i^t) - \mathbf{x}^* \nabla f_i(\mathbf{x}_i^t))$ .  
Let's add  $\pm \mathbf{x}^t \nabla f_i(\mathbf{x}_i^t)$ , we have :

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

Let's use

$$\|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2 \leq 2\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 + 2\|\mathbf{x}_i^t - \mathbf{x}^*\|^2 \Rightarrow -2\|\mathbf{x}_i^t - \mathbf{x}^*\|^2 \leq 2\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 - \|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2$$

We replace

$$\begin{aligned}
\frac{-1}{\eta_t} T_2 &\leq \frac{-2}{n} \sum_{i=1}^n \left[ f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}^*) - \frac{L}{2} \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 + \frac{\mu}{2} \|\mathbf{x}_i^t - \mathbf{x}^*\| \right] \\
&\leq \frac{-2}{n} \sum_{i=1}^n (f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}^*)) + \frac{L}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 - \frac{\mu}{n} \sum_{i=1}^n \|\mathbf{x}_i^t - \mathbf{x}^*\| \\
&\leq \frac{-2}{n} \sum_{i=1}^n (f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}^*)) + \frac{L}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 + \frac{\mu}{2n} \sum_{i=1}^n (2\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 - \|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2) \\
&\leq \frac{-2}{n} \sum_{i=1}^n (f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}^*)) + \frac{L}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 + \frac{\mu}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 - \frac{\mu}{2n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2 \\
&\leq \frac{-2}{n} n (f(\bar{\mathbf{x}}^t) - f(\mathbf{x}^*)) + \frac{L}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 + \frac{\mu}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 - \frac{\mu}{2n} n \|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2 \\
&\leq -2 (f(\bar{\mathbf{x}}^t) - f(\mathbf{x}^*)) + \frac{L + \mu}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 - \frac{\mu}{2} \|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2
\end{aligned}$$

2.  $i_2 = \mathbb{E}_{\xi_1^t, \dots, \xi_n^t} \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) - \frac{1}{n} \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t, \xi_i^t) \right\|^2$ . For this case, we add  $\pm \nabla f_i(\bar{\mathbf{x}}_i^t)$ ,  $\pm \nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t)$ ,  $\pm \nabla f_i(\mathbf{x}^*)$  and  $\pm \nabla F_i(\mathbf{x}^*, \xi_i^t)$ , we have

$$\begin{aligned}
i_2 &= \frac{1}{n^2} \mathbb{E}_{\xi_1^t, \dots, \xi_n^t} \left\| \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) - \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t, \xi_i^t) \right\|^2 \\
&= \frac{1}{n^2} \mathbb{E}_{\xi_1^t, \dots, \xi_n^t} \left\| \sum_{i=1}^n (\nabla f_i(\mathbf{x}_i^t) - \nabla F_i(\mathbf{x}_i^t, \xi_i^t)) \right\|^2 \\
&\leq \frac{3}{n^2} \sum_{i=1}^n \mathbb{E}_{\xi_i^t} \|\bar{\mathbf{x}} \nabla F_i(\mathbf{x}_i^t, \xi_i^t) - \nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t) - \nabla f_i(\mathbf{x}_i^t) + \nabla f_i(\bar{\mathbf{x}}^t)\|^2 + \|\nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t) - \nabla F_i(\mathbf{x}^*, \xi_i^t) - \nabla f_i(\bar{\mathbf{x}}^t) + \nabla f_i(\mathbf{x}^*)\|^2 \\
&\quad + \|\nabla F_i(\mathbf{x}^*, \xi_i^t) - \nabla f_i(\mathbf{x}^*)\|^2
\end{aligned}$$



We can use  $\mathbb{E}\|Y - a\|^2 = \mathbb{E}\|Y\|^2 - \|a\|^2 = \mathbb{E}\|Y\|^2$  if  $a = \mathbb{E}Y$ , we have

$$\begin{aligned}
i_2 &\leq \frac{3}{n^2} \sum_{i=1}^n \mathbb{E}_{\xi_i^t} (\|\nabla F_i(\mathbf{x}_i^t, \xi_i^t) - \nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t) - \nabla f_i(\mathbf{x}_i^t) + \nabla f_i(\bar{\mathbf{x}}^t)\|^2 + \|\nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t) - \nabla F_i(\mathbf{x}^*, \xi_i^t) - \nabla f_i(\bar{\mathbf{x}}^t) + \nabla f_i(\mathbf{x}^*, \xi_i^t) - \nabla f_i(\mathbf{x}^*)\|^2) \\
&\leq \frac{3}{n^2} \sum_{i=1}^n \mathbb{E}_{\xi_i^t} (\|\nabla F_i(\mathbf{x}_i^t, \xi_i^t) - \nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t)\|^2 + \|\nabla F_i(\bar{\mathbf{x}}^t, \xi_i^t) - \nabla F_i(\mathbf{x}^*, \xi_i^t)\|^2 + \|\nabla F_i(\mathbf{x}^*, \xi_i^t) - \nabla f_i(\mathbf{x}^*)\|^2) \\
&\leq \frac{3}{n^2} \sum_{i=1}^n \mathbb{E}_{\xi_i^t} (L^2 \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 + 2L(f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}^*)) + \sigma_i^2) \\
&\leq \frac{3L^2}{n^2} \sum_{i=1}^n \mathbb{E}_{\xi_i^t} \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 + \frac{6L}{n} (f(\bar{\mathbf{x}}^t) - f(\mathbf{x}^*)) + \frac{3\bar{\sigma}^2}{n}
\end{aligned}$$

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

$$3. \ i_3 = \left\langle \bar{\mathbf{x}}^t - \mathbf{x}^* - \frac{\eta_t}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t), \frac{\eta_t}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) - \frac{\eta_t}{n} \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t, \xi_i^t) \right\rangle.$$

We know that  $f_i(\mathbf{x}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} F_i(\mathbf{x}, \xi_i) \Rightarrow \nabla f_i(\mathbf{x}) := \mathbb{E}_{\xi_i \sim \mathcal{D}_i} \nabla F_i(\mathbf{x}, \xi_i)$ , so

$$i_3 = \left\langle \bar{\mathbf{x}}^t - \mathbf{x}^* - \frac{\eta_t}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t), \frac{\eta_t}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) - \frac{\eta_t}{n} \sum_{i=1}^n \nabla F_i(\mathbf{x}_i^t, \xi_i^t) \right\rangle = 0$$

by summing all the results obtained and taking  $\eta_t \leq \frac{1}{12L}$ , we have:

$$\mathbb{E}_{\xi_1^t, \dots, \xi_n^t} \|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq (1 - \frac{\mu\eta_t}{2}) \mathbb{E} \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 - \eta_t (f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)) + \frac{\bar{\sigma}^2 \eta_t^2}{n} + 3L \frac{\eta_t}{n} \mathbb{E}_t \sum_{i=1}^n \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2$$

□

**Lemma 6 (Recursion for consensus distance).**

**Proof:** Using matrix notation, we have :

$$\begin{aligned}
n\Xi_t &= \mathbb{E} \left\| X^{(t)} - \bar{X}^{(t)} \right\|_F^2 \\
&= \mathbb{E} \left\| (X^{(t-1)} - \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)})) W^{(t-1)} - X^{(t-1)} \frac{1}{n} 11^T + \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) \frac{1}{n} 11^T \right\|_F^2 \\
&\leq (1-p) \left\| X^{(t-1)} - \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) - X^{(t-1)} \frac{1}{n} 11^T + \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) \frac{1}{n} 11^T \right\|_F^2 \\
&\leq (1-p) \left\| X^{(t-1)} - \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) - X^{(t-1)} \frac{1}{n} 11^T + \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) \frac{1}{n} 11^T \pm \eta_{t-1} \nabla f_i(X^{(t-1)}) \right\|_F^2 \\
&\leq (1-p) \left\| X^{(t-1)} - X^{(t-1)} \frac{1}{n} 11^T + \eta_{t-1} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) \frac{1}{n} 11^T - \eta_{t-1} \nabla f_i(X^{(t-1)}) \right\|_F^2 + (1-p) \eta_{t-1}^2 \\
&\quad \left\| \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) - \nabla f_i(X^{(t-1)}) \right\|_F^2
\end{aligned}$$

We know that  $\mathbb{E} \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) = \nabla f_i(X^{(t-1)})$ ,  $\left\| \nabla F_i(X^{(t-1)}, \xi^{(t-1)}) - \nabla f_i(X^{(t-1)}) \right\|_F^2 \leq n\bar{\sigma}^2$  and the average of  $\nabla f_i(X^{(t-1)})$  is represented by  $\nabla f(X^{(t-1)}) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(X^{(t-1)})$ . Let's

use the propriety  $\|a + b\|^2 \leq (1 + \alpha)\|a\|^2 + (1 + \alpha^{-1})\|b\|^2$  and take  $\alpha = \frac{p}{2}$ , we have :

$$\begin{aligned}
n\Xi_t &\leq (1-p) \left\| X^{(t-1)} - X^{(t-1)} \frac{1}{n} 11^T + \eta_{t-1} \nabla f(X^{(t-1)}) - \eta_{t-1} \nabla f_i(X^{(t-1)}) \right\|_F^2 + (1-p) \eta_{t-1}^2 n \bar{\sigma}^2 \\
&\leq (1-p)(1+\alpha) \left\| X^{(t-1)} - \bar{X}^{(t-1)} \right\|_F^2 + (1-p)(1+\alpha^{-1}) \eta_{t-1}^2 \left\| \nabla f_i(X^{(t-1)}) - \nabla f(X^{(t-1)}) \right\|_F^2 + (1-p) \eta_{t-1}^2 n \bar{\sigma}^2 \\
&\leq (1-p) \left( 1 + \frac{p}{2} \right) \left\| X^{(t-1)} - \bar{X}^{(t-1)} \right\|_F^2 + (1-p) \left( 1 + \frac{2}{p} \right) \eta_{t-1}^2 \left\| \nabla f_i(X^{(t-1)}) - \nabla f(X^{(t-1)}) \right\|_F^2 + (1-p) \eta_{t-1}^2 n \bar{\sigma}^2 \\
&\leq \left( 1 - \frac{p}{2} \right) \left\| X^{(t-1)} - \bar{X}^{(t-1)} \right\|_F^2 + \frac{3(1-p)}{p} \eta_{t-1}^2 \left\| \nabla f_i(X^{(t-1)}) - \nabla f(X^{(t-1)}) \right\|_F^2 + (1-p) \eta_{t-1}^2 n \bar{\sigma}^2
\end{aligned}$$

We know that :

$$\sum_{i=1}^n \left\| \nabla f_i(X^{(t-1)}) - \nabla f(X^{(t-1)}) \right\|_2^2 \leq n \bar{\xi}_1^2$$

We have :

$$\begin{aligned}
n\Xi_t &\leq \left( 1 - \frac{p}{2} \right) n\Xi_{t-1} + \frac{3(1-p)}{p} \eta_{t-1}^2 n \bar{\xi}_1^2 + (1-p) \eta_{t-1}^2 n \bar{\sigma}^2 \\
\Xi_t &\leq \left( 1 - \frac{p}{2} \right) \Xi_{t-1} + \frac{3(1-p)}{p} \eta_{t-1}^2 \bar{\xi}_1^2 + (1-p) \eta_{t-1}^2 \bar{\sigma}^2
\end{aligned}$$

Let's take  $1 - p \leq 1$ , We get :

$$\Xi_{t+1} \leq \left( 1 - \frac{p}{2} \right) \Xi_t + \frac{3}{p} \eta_t^2 \bar{\xi}_1^2 + \eta_t^2 \bar{\sigma}^2$$

□

Let's combine (10) and (11) as  $\psi_{t+1} := R_{t+1} + a\Xi_{t+1}$ , we have :

$$\begin{aligned}
R_{t+1} + a\Xi_{t+1} &\leq \left( 1 - \frac{\mu\eta_t}{2} \right) R_t - \eta_t e_t + \frac{\bar{\sigma}^2}{2} \eta_t^2 + 3L\eta_t \Xi_t + a \left( 1 - \frac{p}{2} \right) \Xi_t + \frac{3a}{p} \eta_t^2 \bar{\xi}_1^2 + a\eta_t^2 \bar{\sigma}^2 \\
&\leq \left( 1 - \frac{\mu\eta_t}{2} \right) R_t + \left( 3L\eta_t + a \left( 1 - \frac{p}{2} \right) \right) \Xi_t - \eta_t e_t + \frac{3a}{p} \eta_t^2 \bar{\xi}_1^2 + \frac{\bar{\sigma}^2}{2} \eta_t^2 + a\eta_t^2 \bar{\sigma}^2
\end{aligned}$$

Let's take :

$$\frac{3L\eta_t}{a} = \frac{p}{4} \implies a = \frac{12L\eta_t}{p}$$

We get :

$$R_{t+1} + a\Xi_{t+1} \leq \left( 1 - \frac{\mu\eta_t}{2} \right) R_t + a \left( 1 - \frac{p}{4} \right) \Xi_t - \eta_t e_t + \frac{\bar{\sigma}^2}{2} \eta_t^2 + \frac{3a}{p} \eta_t^2 \bar{\xi}_1^2 + a\eta_t^2 \bar{\sigma}^2$$

We know that  $a = \frac{12L\eta_t}{p}$ , we get :

$$\psi_{t+1} \leq (1-h)\psi_t - \eta_t e_t + \frac{\bar{\sigma}^2}{2} \eta_t^2 + \left( \frac{36\bar{\xi}_1^2}{p^2} + \frac{12\bar{\sigma}^2}{p} \right) L\eta_t^3,$$

where  $h = \min\{\frac{\mu\eta_t}{2}, \frac{p}{4}\}$ .

It will be useful to define a virtual sequence for every  $i \in [n]$  and  $t \in [T]$  such that :

$$\begin{aligned}
X^{t+1} &= X^t - \eta_t \nabla f_i(X^t) & \bar{X}^{t+1} &= \bar{X}^t - \frac{\eta_t}{n} \sum_{i=0}^n \nabla f_i(X^t) \\
\nabla f(\mathbf{x}_i^t) &= \frac{1}{n} \sum_{i=0}^n a_i^2 \bar{\mathbf{x}}_i^t & \nabla f(\bar{\mathbf{x}}) &= \frac{1}{n} \sum_{i=0}^n a_i^2 \bar{\mathbf{x}} & \nabla f_i(\mathbf{x}_i^t) &= a_i^2 \mathbf{x}_i^t
\end{aligned}$$

**Theorem 11.**

**Proof:** Using the idea of Lemma 6, we have

$$n\Xi_{t+1} = \mathbb{E} \left\| \mathbf{x}^{(t+1)} - \bar{\mathbf{x}}^{(t+1)} \right\|_F^2 = \mathbb{E} \left\| \left( \mathbf{x}^{(t)} - \eta_t a_i^2 \mathbf{x}_i^{(t)} \right) W - \bar{\mathbf{x}}^{(t)} + \frac{\eta_t}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2$$

Let's use the Assumption 1 and the propriety  $\|a + b\|^2 \leq (1 + \alpha)\|a\|^2 + (1 + \alpha^{-1})\|b\|^2$  and  $\left\| a_i^2 \mathbf{x}_i^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2 \leq n\bar{\xi}_2^2$ , we have:

$$\begin{aligned} n\Xi_{t+1} &\leq (1-p) \left\| \mathbf{x}^{(t)} - \eta_t a_i^2 \mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)} + \frac{\eta_t}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2 \\ &\leq \left(1 - \frac{p}{2}\right) \|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\|_F^2 + \frac{3(1-p)}{p} \eta_t^2 \left\| a_i^2 \mathbf{x}_i^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2 \\ &\leq \left(1 - \frac{p}{2}\right) \|\mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)}\|_F^2 + \frac{3(1-p)}{p} \eta_t^2 \|a_i^2 \mathbf{x}_i^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \bar{\mathbf{x}}^{(t)}\|_F^2 + \frac{1}{n} \sum_{i=1}^n a_i^2 \bar{\mathbf{x}}^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \\ &\quad + \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \|_F^2 \\ &\leq \left(1 - \frac{p}{2}\right) n\Xi_t + \frac{3(1-p)}{p} \eta_t^2 n\bar{\xi}_2^2 + \frac{6(1-p)}{p} \eta_t^2 L^2 n\Xi_t \end{aligned}$$

Let's take  $(1-p) \leq 1$  and choosing  $\eta_t$  by

$$\frac{6}{p} \eta_t^2 L^2 \leq \frac{p}{4} \Rightarrow \eta_t^2 \leq \frac{p^2}{24L^2}$$

We have,

$$\Xi_{t+1} \leq \left(1 - \frac{p}{4}\right) \Xi_t + K\eta_t^2, \quad (14)$$

with  $K = \frac{3\bar{\xi}_2^2}{p}$ . Unrolling  $\Xi_t$  recursively up to 0 we get,

$$\Xi_t \leq K \sum_{j=0}^{t-1} \left(1 - \frac{p}{4}\right)^{t-j} \eta_j^2$$

From [10] we have  $1 - \frac{p}{4} \leq 1 - \frac{p}{8}$ ,  $\eta_t^2$  is  $\frac{8}{p}$ -slow decreasing, i.e.  $\eta_j^2 \leq \left(1 + \frac{p}{16}\right)^{t-j} \eta_t^2$  and  $\left(1 - \frac{p}{8}\right) \left(1 + \frac{p}{16}\right) \leq \left(1 - \frac{p}{16}\right)$ . We multiply the both sides by  $B \sum_{t=0}^T w_t$ , we have:

$$B \sum_{t=0}^T w_t \Xi_t \leq \frac{16KB}{p} \sum_{t=0}^T w_t \eta_t^2$$

From lemma 5, let's rearrange the equation (10) with  $\bar{\sigma} = 0$ . Multiplying by  $w_t$  and dividing by  $\eta_t$  and summing up and dividing by  $W_T = \sum_{t=0}^T w_t$ , we get

$$\frac{1}{W_T} \sum_{t=0}^T w_t e_t \leq \frac{1}{W_T} \sum_{t=0}^T \left( \frac{(1 - \eta_t m) w_t}{\eta_t} R_t - \frac{w_t}{\eta_t} R_{t+1} \right) + \frac{16KB}{p} \eta^2$$

Using that  $\eta_t = \eta$  and that  $\frac{(1 - \eta_t m) w_t}{\eta} = \frac{w_{t-1}}{\eta}$  we obtain a telescoping sum,

$$\frac{1}{W_T} \sum_{t=0}^T w_t e_t + \frac{w_T}{W_T \eta} R_{T+1} \leq \frac{R_0}{W_T \eta} + \frac{16KB}{p} \eta^2$$

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

$$\begin{aligned} \frac{1}{2W_T} \sum_{t=0}^T Bw_t e_t + a r_{T+1} &\leq (1 - \eta m)^{-(T+1)} \frac{r_0}{\eta} + \frac{16KB}{p} \eta^2 \\ &\leq \frac{r_0}{\eta} \exp[-\eta m(T+1)] + \frac{16KB}{p} \eta^2 \end{aligned}$$

Lemma follows by tuning  $\eta$  the same way as in [32], if  $\frac{1}{d} \leq \frac{\ln(\min\{2, m^2 r_0 T^2/c\})}{mT}$ , we choose  $\eta = \frac{1}{d}$

$$\mathcal{O} \left( r_0 d \exp \left[ \frac{-m(T+1)}{d} \right] + \frac{KB}{d^2 p} \right) \leq \mathcal{O} \left( r_0 d \exp \left[ \frac{-m(T+1)}{d} \right] + \frac{KB}{pm^2 T^2} \right)$$

We have for  $m = \frac{\mu}{2}$ ,  $K = \frac{3\xi_2^2}{p}$ ,  $B = 3L$  and  $d = \frac{L}{p}$ ,

$$\mathcal{O} \left( r_0 \frac{L}{p} \exp \left[ \frac{\mu p T}{L} \right] + \frac{L \xi_2^2}{\mu^2 p^2 T^2} \right)$$

□

The improvement of the convergence rate result of Theorem 11 is therefore presented in the theorem below. The improvement of the convergence rate result of Theorem 11 is therefore presented in the theorem below. The improvement of the convergence rate result of Theorem 11 is therefore presented in the theorem below. The improvement of the convergence rate result of Theorem 11 is therefore presented in the theorem below.

### Theorem 12.

**Proof:** Let's review Lemma 6, we have :

$$n\Xi_{t+1} = \mathbb{E} \left\| \mathbf{x}^{(t+1)} - \bar{\mathbf{x}}^{(t+1)} \right\|_F^2 = \mathbb{E} \left\| \left( \mathbf{x}^{(t)} - \eta_t a_i^2 \mathbf{x}_i^{(t)} \right) W - \bar{\mathbf{x}}^{(t)} + \frac{\eta_t}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2$$

Let's use the Assumption 1 and the propriety  $\|a + b\|^2 \leq (1 + \alpha)\|a\|^2 + (1 + \alpha^{-1})\|b\|^2$  and  $\left\| a_i^2 \mathbf{x}_i^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2 \leq n\bar{\xi}_2^2$ , we have:

$$\begin{aligned} n\Xi_{t+1} &\leq (1-p) \left\| \mathbf{x}^{(t)} - \eta_t a_i^2 \mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)} + \frac{\eta_t}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2 \\ &\leq \left(1 - \frac{p}{2}\right) \left\| \mathbf{x}^{(t)} - \bar{\mathbf{x}}^{(t)} \right\|_F^2 + \frac{3(1-p)}{p} \eta_t^2 \left\| a_i^2 \mathbf{x}_i^{(t)} - \frac{1}{n} \sum_{i=1}^n a_i^2 \mathbf{x}_i^{(t)} \right\|_F^2 \\ &\leq \left(1 - \frac{p}{2}\right) n\Xi_t + \frac{3(1-p)}{p} \eta_t^2 n\bar{\xi}_2^2 \end{aligned}$$

And  $(1-p) \leq 1$ , we obtain,

$$\Xi_{t+1} \leq \left(1 - \frac{p}{2}\right) \Xi_t + \frac{3}{p} \eta_t^2 \bar{\xi}_2^2 \quad (15)$$

From the lemma 5, let us take the equation (10) with  $\bar{\sigma} = 0$  and combine this equation with (15) as  $\Phi_{t+1} = R_{t+1} + a\Xi_{t+1}$ , we have:

$$\begin{aligned} R_{t+1} + a\Xi_{t+1} &\leq (1 - \eta_t \mu) R_t - \eta_t e_t + 3L\eta_t \Xi_t + a \left(1 - \frac{p}{2}\right) \Xi_t + \frac{3a}{p} \eta_t^2 \bar{\xi}_2^2 \\ &\leq (1 - \eta_t \mu) R_t + a \left( \frac{3L\eta_t}{a} + \left(1 - \frac{p}{2}\right) \right) \Xi_t - \eta_t e_t + \frac{3a}{p} \eta_t^2 \bar{\xi}_2^2 \end{aligned}$$

Let's take ,

$$\frac{3L\eta_t}{a} = \frac{p}{4} \Rightarrow a = \frac{36L\eta_t}{p}$$

We have,

$$\Phi_{t+1} \leq (1-h)\Phi_{t+1} - \eta_t e_t + \frac{24L}{p^2} \eta_t^3 \bar{\xi}_2^2$$

where  $h = \min\{\eta_t \mu, \frac{p}{4}\}$ . Multiplying by  $w_t$  and dividing by  $\eta_t$ , we get

$$\begin{aligned} \frac{w_t}{\eta_t} r_{t+1} &\leq \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - B \frac{w_t}{\eta_t} \eta_t e_t + \frac{w_t}{\eta_t} A \eta_t^3 \\ &\leq \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - B w_t e_t + A w_t \eta_t^2 \end{aligned}$$

We have,

$$B w_t e_t \leq \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - \frac{w_t}{\eta_t} r_{t+1} + A w_t \eta_t^2$$

Summing up and dividing by  $W_T = \sum_{t=0}^T w_t$ , we have:

$$\frac{1}{W_T} \sum_{t=0}^T B w_t e_t \leq \frac{1}{W_T} \sum_{t=0}^T \left( \frac{w_t}{\eta_t} (1 - \min\{m\eta_t, k\}) r_t - \frac{w_t}{\eta_t} r_{t+1} \right) + \frac{A}{W_T} \sum_{t=0}^T w_t \eta_t^2$$

and hence,

$$\frac{1}{2W_T} \sum_{t=0}^T B w_t e_t + \frac{w_T r_{T+1}}{W_T \eta} \leq \frac{r_0}{W_T \eta} + A \eta^2$$

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

$$\begin{aligned} \frac{1}{2W_T} \sum_{t=0}^T B w_t e_t + a r_{T+1} &\leq (1 - \min\{m\eta, k\})^{-(T+1)} \frac{r_0}{\eta} + A \eta^2 \\ &\leq \frac{r_0}{\eta} \exp[-\min\{m\eta, k\}(T+1)] + A \eta^2 \end{aligned}$$

Lemma follows by tuning  $\eta$  the same way as in [32], if  $\frac{1}{F} \leq \frac{\ln(\min\{2, m^2 r_0 T^2 / N\})}{mT}$ , we choose  $\eta = \frac{1}{F}$

$$\mathcal{O} \left( r_0 F \exp \left[ -\min \left\{ \frac{m}{F}, k \right\} (T+1) \right] + \frac{A}{F^2} \right) \leq \mathcal{O} \left( r_0 F \exp \left[ -\min \left\{ \frac{m}{F}, k \right\} (T+1) \right] + \frac{A}{m^2 T^2} \right)$$

We have for  $m = \frac{\mu}{2}$ ,  $A = \frac{36L}{p^2} \bar{\xi}_2^2$ ,  $k = \frac{p}{4}$  and  $F = \frac{p}{4}$ ,

$$\mathcal{O} \left( r_0 L \exp \left[ -\min \left\{ \frac{\mu}{L}, p \right\} T \right] + \frac{L \bar{\xi}_2^2}{\mu^2 p^2 T^2} \right)$$

□