
Asynchronous SGD on Graphs: a Unified Framework for Asynchronous Decentralized and Federated Optimization

Mathieu Even
Inria - ENS Paris

Anastasia Koloskova
EPFL, Switzerland

Laurent Massoulié
Inria - ENS Paris

Abstract

Decentralized and asynchronous communications are two popular techniques to speedup communication complexity of distributed machine learning, by respectively removing the dependency over a central orchestrator and the need for synchronization. Yet, combining these two techniques together still remains a challenge. In this paper, we take a step in this direction and introduce Asynchronous SGD on Graphs (AGRAF SGD) — a general algorithmic framework that covers asynchronous versions of many popular algorithms including SGD, Decentralized SGD, Local SGD, FedBuff, thanks to its relaxed communication and computation assumptions. We provide rates of convergence under much milder assumptions than previous decentralized asynchronous works, while still recovering or even improving over the best know results for all the algorithms covered.

1 Introduction

We consider solving stochastic optimization problems that are distributed amongst n agents (indexed by a set \mathcal{V}) who can compute stochastic gradients in parallel. This includes classical federated setups, such as distributed and federated learning. Depending on the application, agents have access to either same shared data distribution or a different agent-specific distributions. In recent years, such stochastic optimization problems have continued to grow rapidly in size, both in terms of the dimension d of the optimization variable—i.e., the number of model parameters in ma-

chine learning—and in terms of the quantity of data—i.e., the number of data samples m being used over all agents. With d and m regularly reaching the hundreds or thousands of billions [Chowdhery et al., 2022, Tourvion et al., 2023], it is increasingly necessary to use parallel optimization algorithms to handle the large scale.

With *communication cost* being one of the major bottlenecks of parallel optimization algorithms, there are several directions aimed to improve communication efficiency. Amongst the others (such as local update steps [Stich, 2019, Woodworth et al., 2020] and communication compression [Alistarh et al., 2017, Koloskova et al., 2019]), **decentralization** and **asynchrony** are the two popular techniques for reducing the communication time. Decentralization [Koloskova et al., 2020, Lian et al., 2017a] eliminates the dependency on the central server—frequently a major bottleneck in distributed learning—while naturally amplifying privacy guarantees [Cyffers et al., 2022]. Asynchrony Recht et al. [2011], Baudet [1978], Tsitsiklis et al. [1986] shortens the time per computation rounds and allows more updates to be made during the same period of time. It aims to overcome several possible sources of delays: nodes may have *heterogeneous hardware* with different computational throughputs [Kairouz et al., 2019, Horváth et al., 2021], *network latency* can slow the communication of gradients, and nodes may even just *drop out* [Ryabinin et al., 2021]. Moreover, slower “*straggler*” compute nodes can arise in many natural parallel settings, including training ML models using multiple GPUs [Chen et al., 2016] or in the cloud; sensitivity to these stragglers poses a serious problem for synchronous algorithms, that depend on the slowest agent. In decentralized synchronous optimization where communication times between pairs of nodes may be heterogeneous, the algorithm can even be further slowed down by *straggling communication links*.

Combining both decentralization and asynchrony is a challenging problem, and it is only recently that this question has risen a surge of interest [Assran and Rab-

bat, 2021, Bornstein et al., 2023, Luo et al., 2020, Liu et al., 2022, Nadiradze et al., 2021, Even et al., 2021c, Zhang and You, 2021]. These works are however restricted to a given communication protocol and static topologies [Assran and Rabbat, 2021, Lian et al., 2015, Bornstein et al., 2023, Nadiradze et al., 2021, Even et al., 2021c], no communication delays [Lian et al., 2015, Bornstein et al., 2023, Nadiradze et al., 2021], or their analyses rely on an upper-bound on the maximal computation delay [Assran and Rabbat, 2021, Lian et al., 2017b, Bornstein et al., 2023, Luo et al., 2020, Liu et al., 2022, Nadiradze et al., 2021, Zhang and You, 2021, Wu et al., 2023]. In this work we aim to circumvent these shortcomings. We study an asynchronous version of decentralized SGD in a unified framework that relaxes overly strong communication assumptions imposed by prior works. Our framework covers time-varying topologies, arbitrary computation orders and local update steps. We prove an improved rates of convergence under such a weaker communication assumptions, covering and improving asynchronous versions of many common distributed and federated algorithms.

1.1 Contributions

(i) We introduce **AGRAF SGD** (Asynchronous SGD on graphs), a unified formulation of an asynchronous version of the synchronous Decentralized SGD as formulated by Koloskova et al. [2020]. One of the strengths of AGRAP SGD is that it formally takes the form of a simple sequence (Equation (3)), allowing for an effective theoretical analysis, while covering asynchronous versions of many distributed algorithms such as Asynchronous SGD, Decentralized SGD, FedAvg or FedBuff.

(ii) We analyze the AGRAP SGD sequence under various combinations of convexity, non-convexity, smoothness and Lipschitzness assumptions. We use a relaxed communication assumption that only imposes that the different topologies mix in a given window of time, while our computation assumption depends on whether the local functions are homogeneous or heterogeneous. In special cases, our rates recover best known rates of Minibatch SGD, Asynchronous SGD or Decentralized SGD, while for Asynchronous Decentralized SGD, our rates improve the previous works by up to factors of order n^2 , under relaxed assumptions (as summarized in Table 1).

(iii) Finally, we show that AGRAP SGD allows to efficiently handle communication delays in decentralized optimization, by introducing *Decentralized SGD on Loss Networks*. We show that the assumptions required in our analysis are satisfied by this algorithm, giving explicit rates of convergence that depend on the underlying graph topology, pairwise communica-

tion delays, and each device computation time.

1.2 Related works

Asynchronous optimization. Asynchronous optimization has a long history. In the 1970s, Baudet [1978] considered shared-memory asynchronous fixed-point iterations, and an early convergence result for Asynchronous SGD was established by Tsitsiklis et al. [1986]. Recent analysis typically relies on bounded delays [Agarwal and Duchi, 2011, Recht et al., 2011, Lian et al., 2015, Stich and Karimireddy, 2020], while some algorithms try to adapt to the delays [Sra et al., 2016, Zheng et al., 2017, Mishchenko et al., 2018, Koloskova et al., 2022, Mishchenko et al., 2022, Feyzmahdavian and Johansson, 2021], in order to depend only on an average delay. For more examples of stochastic asynchronous algorithms, we refer readers to the surveys by Ben-Nun and Hoefler [2019], Assran et al. [2020]. More closely related to our analysis techniques, Mania et al. [2017] proposed and utilized the analysis tool of **virtual iterates** for Asynchronous SGD under bounded delays, extended by Koloskova et al. [2020], Mishchenko et al. [2022] who proved that Asynchronous SGD performs well under arbitrary delays. We adapt this proof approach to decentralized optimization in order to obtain some robustness towards large delays and introduce a different virtual sequence for the averaged model over all the nodes.

Decentralized SGD and asynchrony. Decentralized SGD [Koloskova et al., 2022, e.g.] consists in iterations where at every time step, all nodes perform local SGD steps, and communicate their local model with their neighbors in a graph (that may vary with time, but that needs to mix in an ergodic way). The closest works to ours Lian et al. [2017b], Bornstein et al. [2023] proposed asynchronous versions of decentralized SGD where at each iteration, **one** node v_k is sampled *independently from the past* (with fixed probabilities), and this node performs a local stochastic gradient step and an averaging operation with its neighbors. We extend their sequence and results to a more general (due to relaxed communication and computation assumptions) asynchronous version of decentralized SGD, that keeps the “unified” point of view of the work of Koloskova et al. [2020]. Assran and Rabbat [2021] considers asymmetric communications (*push sum*) and all the agents performing computations at every iterations in a synchronous way, Nadiradze et al. [2021] considers quantized pairwise communications as in the historical gossip algorithm [Boyd et al., 2006], but no communication nor computation delays, while Luo et al. [2020], Agarwal et al. [2009] do not provide convergence guarantees. Orthogonally, Even et al. [2021c] consider both communication and

computation delays in a *continuized* framework [Even et al., 2021a], allowing more degrees of freedom for the algorithm and the analysis, but their work does not apply to modern ML tasks; still, our Loss Network section relates to this line of work due to the introduction of continuous-time physical delays. Nabli et al. [2023] considers a similar continuized framework for communication acceleration on time-varying topologies with local stochastic gradient steps: while their work seems communication efficient, they do not consider computation or communication delays.

2 AGRAF Algorithmic Framework

In this section we present AGRAF SGD—our algorithmic framework for asynchronous decentralized SGD—and give examples of existing popular algorithms that it can cover.

2.1 Asynchronous SGD on graphs

We consider a connected undirected graph $G = (\mathcal{V}, \mathcal{E})^1$ on a set of nodes $\mathcal{V} = \{1, \dots, n\}$. Let the function $f_v : \mathbb{R}^d \rightarrow \mathbb{R}$ of agent $v \in \mathcal{V}$ be defined as

$$f_v(x) := \mathbb{E}[F_v(x, \xi_v)] \quad \xi_v \sim \mathcal{D}_v, \quad x \in \mathbb{R}^d, \quad (1)$$

where \mathcal{D}_v is some local distribution. Let the global objective function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined as follows, and consider the optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \sum_{v \in \mathcal{V}} q_v f_v(x) \right\}, \quad (2)$$

for some non-negative weights (q_v) that sum to 1. We classically assume that node v in the graph has access to unbiased stochastic gradients of f_v (of the form $F_v(x, \xi_v)$). The standard goal of decentralized optimization is to minimize f using only local computations and communications (only neighboring nodes in the graph can communicate).

Notations. Standard small letters (x, g, y, z , etc) are for vectors in \mathbb{R}^d . Capital letters (mostly W) are for matrices in $\mathbb{R}^{\mathcal{V} \times \mathcal{V}}$. Bold letters $\mathbf{x}, \mathbf{g}, \dots$ are for *concatenated vectors* in $\mathbb{R}^{\mathcal{V} \times d}$, that we write as $\mathbf{x} = (x_v)_{v \in \mathcal{V}}$. For some vector $x \in \mathbb{R}^d$, we denote $\mathbf{x} \in \mathbb{R}^{\mathcal{V} \times d}$ the concatenated vector such that $\mathbf{x}_v = x$ for all $v \in \mathcal{V}$. $\mathbf{1} \in \mathbb{R}^{\mathcal{V}}$ is the vector with all entries equal to 1. For $\mathbf{x} \in \mathbb{R}^{\mathcal{V} \times d}$, we denote $\bar{\mathbf{x}} = \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x}$.

In this paper we study a general scheme for *asynchronous SGD on graphs (AGRAF)* which is summarized in Algorithm 1: workers asynchronously perform local SGD steps (lines 3-4), while an underlying **linear communication algorithm** is running

¹Since we consider varying topologies, this graph should be thought as the union of graphs considered over time.

Algorithm 1 Asynchronous SGD on graph G (AGRAF SGD)

- 1: **Input:** $\bar{x}^0 \in \mathbb{R}^d$, $x_v = \bar{x}^0$ for $v \in \mathcal{V}$ initialized local variables, stepsize $\gamma > 0$
- 2: **for** $v \in \mathcal{V}$, **do**
- 3: Upon finishing computation of a stochastic gradient $\nabla F(\tilde{x}_v, \tilde{\xi}_v)$ at some previous local current state \tilde{x}_v ,

$$x_v \leftarrow x_v - \gamma \nabla F_v(\tilde{x}_v, \tilde{\xi}_v).$$

- 4: Compute $\nabla F_v(x_v, \xi_v)$ for $\xi_v \sim \mathcal{D}_v$ independently from the past, at current state x_v .
 - 5: **end for**
 - 6: **while** procedure still running **do**
 - 7: Run any **linear communication algorithm** on graph G incurring no communication delay.
 - 8: **end while**
-

without incurring communication delays (line 7). A linear communication algorithm on graph G implies that any communication update can be formulated as $\mathbf{x}_+ = W\mathbf{x}_-$ where $\mathbf{x}_+, \mathbf{x}_- \in \mathbb{R}^{\mathcal{V} \times d}$ are respectively the global state after and before the communication update, and $W \in \mathbb{R}^{\mathcal{V} \times \mathcal{V}}$ is a **communication matrix** with $W_{v,w}$ being zero for disconnected nodes v, w , i.e. $W_{v,w} \neq 0$ iff $\{v, w\} \in \mathcal{E}$.

Since every agent asynchronously works at their own speed and communicates in a decentralized way, there is no global state. Keeping track of a global ordering of the iterates involving both computation and communication updates is thus a challenge. In the next subsection we address this challenge and propose a way to effectively cast Algorithm 1 into equations with ordered updates. This reformulation is a key novelty of our work. It allows for an improved theoretical analysis with better rates together with relaxed communication and computations assumptions, allowing AGRAF SGD to cover asynchronous versions of many popular distributed and federated algorithms.

2.2 The sequence studied

We denote by $T_0 = 0$ the initialization time of the algorithm and by $\{0 < T_1 < T_2 < \dots\}$ the times at which the local computation updates are made. Note that these are physical (continuous) times, and that several agents may possibly finish their local computations at the same time T_k . We also assume that computational updates are **atomic**. For some time T , we denote as $T-$ the left limit ($\lim_{t \rightarrow T, t < T}$) and $T+$ the right limit ($\lim_{t \rightarrow T, t > T}$). For time $t \in \mathbb{R}^+$ (physical time), let $x_v(t) \in \mathbb{R}^d$ denote the state of the local variable at time t , and let $\mathbf{x}(t) = (x_v(t))_{v \in \mathcal{V}}$. For $k \geq 0$ and $v \in \mathcal{V}$,

let x_v^k denote the state of the local variable at node v at time T_k+ i.e., $x_v^k = x_v(T_k+) = \lim_{t \rightarrow T_k+, t > T_k} x_v(t)$ and let $\mathbf{x}^k = (x_v^k)_{v \in \mathcal{V}}$.

Communication updates. For $k \geq 0$, none to plenty of communication updates may have happened between the computational update times T_k and T_{k+1} . We encode these communication updates by a *single* matrix W_k : W_k is thus the product of all communication matrices corresponding to communication updates between times T_k and T_{k+1} . Hence, we can write:

$$\mathbf{x}(T_{k+1}-) = W_k \mathbf{x}(T_k+).$$

If no communication happened between two gradients computed, we have $W_k = I_d$. If there are r communications between times T_k+ and $T_{k+1}-$ that happened at times $T_k < T_{k,1} < \dots < T_{k,r} < T_{k+1}$, denoting by $W_{k,r}$ the communication matrix corresponding to communication updates at time $T_{k,r}$, we have $W_k = W_{k,r} \cdot \dots \cdot W_{k,2} \cdot W_{k,1}$. Note that for $r = 0$ this product is taken equal to I_d .

Computation updates. For $k \geq 1$, let $\mathcal{I}_k \subset \mathcal{V}$ be the set of nodes that finish computing stochastic gradients $\nabla F_v(\tilde{x}_v^k, \tilde{\xi}_v^k)$ for $v \in \mathcal{I}_k$ at time T_k- . The computation updates, that are assumed to be **atomic**, then read:

$$x_v(T_k+) = x_v(T_k-) - \gamma \nabla F(\tilde{x}_v^k, \tilde{\xi}_v^k), \quad v \in \mathcal{I}_k,$$

where $\tilde{x}_v^k = x_v^{k-1-\tau(k,v)}$ and $\tilde{\xi}_v^k = \xi_{v_k}^{k-1-\tau(k,v)}$, for $\tau(k,v) \geq 0$ the delay of this update that corresponds to the number of computation updates performed by other nodes during the computation of the local stochastic gradient.

The sequence studied. Combining communication and computation updates, the sequence generated by Algorithm 1 follows the following recursion:

$$\mathbf{x}^{k+1} = W_k \mathbf{x}^k - \gamma \mathbf{g}^k, \quad (3)$$

where $g_w^k = 0$ for $v \notin \mathcal{I}_k$, and $g_v^k = \nabla F_v(x_v^{k-\tau(k,v)}, \xi_v^{k-\tau(k,v)})$.

What is important to keep in mind is that the iterates \mathbf{x}^{k+1} are taken at the time just after computation updates (time T_k+) so that k denotes the number of computation updates. \mathcal{I}_k is the set of nodes that perform computation updates at iteration k , it can be any subset of \mathcal{V} , and $\sum_{k < K} |\mathcal{I}_k|$ denotes the total number of stochastic gradients computed up to iteration K by **all** the agents. The matrix W_k encodes all communications that happened between the k -th and $(k+1)$ -th computation updates (there can be any number such communications, the more there are the more (W_k) will mix).

2.3 AGRAF SGD is the right formulation of Asynchronous Decentralized SGD

Recall that the Decentralized SGD algorithm [Koloskova et al., 2022, e.g.] consists in iterations of the form:

$$x_v^{k+1} = \sum_{w \sim v} W_{\{v,w\}}^{(k)} x_w^k - \gamma \nabla F_v(x_v^k, \xi_v^k), \quad \forall v \in \mathcal{V}, \quad (4)$$

for communication matrices $(W^{(k)})_k$ satisfying Assumption 2. The question thus arises: how can Decentralized SGD be turned into an asynchronous algorithm? Previous works [Lian et al., 2017b, Bornstein et al., 2023] proposed and analyzed schemes that take the following form: at each iteration, **one** node v_k is sampled independently from the past (with fixed or lower bounded probability), and this node performs a local stochastic gradient step together with an averaging operation with its neighbors in the graph. This results in updates of the form of AGRAF SGD, for $\mathcal{I}_k = \{v_k\}$ and W_k a matrix that depends on v_k and that mixes (in mean) independently from the past ($\mathbb{E}[W_k | W_0, \dots, W_{k-1}]$ mixes well).

Leaving the analyses aside, this prior approach is too restrictive: **(i)** *communication assumptions* do not allow varying topologies that may mix but only in the long run, which may particularly be the case for asynchronous algorithms, and **(ii)** *computation assumptions* do not allow for more than one worker to update their value at the same time; having a sampling assumption restricts the type of delays that the algorithm can handle; and nodes that compute should not necessarily be correlated to communicating edges since this forbids the use of several local SGD steps.

AGRAF SGD thus appears as a natural way to make Decentralized SGD asynchronous: nodes are not forced to all perform computations at the same time as in eq. (4), and having the relaxed communication assumption (Assumption 2) allows any communication order, especially when one considers W_k as a concatenation of all communications that may happen between two consecutive computations.

2.4 Some examples covered by AGRAF

We now give a few examples of algorithms (*i.e.* communication and computation schedules) that can be cast as AGRAF SGD. The three first are degenerate cases.

Minibatch SGD and Asynchronous SGD are obtained by setting $W_k = \frac{1}{n} \mathbf{1}\mathbf{1}^\top$, and $\mathcal{I}_k = \mathcal{V}$ and $\mathcal{I}_k = \{v_k\}$ for some node v_k respectively.

Decentralized (local) SGD. Set $\mathcal{I}_k = \mathcal{V}$ and $(W_k)_k$ a sequence of gossip matrices to obtain Decentralized

SGD [Ram et al., 2010]. Note that in that case there are no computation delays, since this algorithm is inherently synchronous and all nodes perform updates at the same time. As done in Koloskova et al. [2020], periodic communications are possible, allowing to recover algorithms with several local gradient steps between each communication round, such as **Local (Decentralized) SGD** [Stich, 2019] or **FedAvg** [McMahan et al., 2017].

Asynchronous Decentralized SGD. As explained in Section 2.3, AGRAF SGD covers Asynchronous Decentralized SGD beyond particular instances previously studied [Lian et al., 2017b, Bornstein et al., 2023]. Furthermore, since we make relaxed communication/computation assumptions, we cover more general decentralized algorithms that allow **local** gradient steps between communications, varying topologies, and arbitrary computations. As such, together with covering an asynchronous version of Decentralized SGD (4), we also cover asynchronous versions of **FedAvg** or **Local SGD**, together with **FedBuff** [Nguyen et al., 2022].

Asynchronous SGD on Loss Networks. If communication latencies are not negligible compared to computational ones, designing an algorithm that is asynchronous and decentralized becomes much more challenging, as the naive implementation might lead to deadlocks. In order to handle non-negligible communication delays, we use *loss networks* [Kelly, 1991] to enforce that the edges adjacent to “busy” nodes are prohibited to be used for communicating². This enables us to design communication/computation schemes that fit in the AGRAF framework, while not violating the physical delay constraints. We introduce these Loss Networks in Section 5.2: we define them more thoroughly, and provide their ergodic mixing properties with explicit constants that depend on the graph topology and local communication and computation delays.

3 Assumptions and Notations

We consider solving the problem (2) under several standard [see, e.g., Bubeck, 2015] combinations of conditions on the objective F . We denote the minimum of f as $f^* := \min_{x \in \mathbb{R}^d} f(x)$, an upper bound on the **initial suboptimality** $\Delta \geq f(\bar{x}^0) - f^*$, and an upper bound on the **initial distance** to the minimizer $D \geq \min \{\|x_0 - x^*\| : x^* \in \operatorname{argmin}_x f(x)\}$ that we assume

²Loss Networks were initially introduced by F. Kelly to model telecommunication networks, where the same mobile phone cannot initiate another phone call while being busy with another call. In our case, phone calls should be thought as communicating with a neighbor.

to exist. $\|\cdot\|$ denotes the Euclidean norm. A function F is **convex** if for each x, y and subgradient $g \in \partial F(x)$, we have $F(y) \geq F(x) + \langle g, y - x \rangle$. When F_v and f_v are convex, we do not necessarily assume they are differentiable, but we abuse notation and use $\nabla f_v(x)$ and $\nabla F_v(x; \xi)$ to denote an arbitrary subgradient at x . The loss F_v is **B -Lipschitz-continuous** if for each x, y and ξ , we have $|F_v(x; \xi) - F_v(y; \xi)| \leq B\|x - y\|$. The objective f_v is **L -smooth** if it is differentiable and its gradient is L -Lipschitz-continuous. We also assume the stochastic gradients have **σ^2 -bounded variance**³.

Assumption 1 (Noise). *There exists σ^2 such that for all x and $v \in \mathcal{V}$, we have $\mathbb{E}[\nabla_x F_v(x, \xi_v)] = \nabla f_v(x)$ and $\mathbb{E}[\|\nabla f_v(x) - \nabla_x F_v(x, \xi_v)\|^2] \leq \sigma^2$, where $\xi_v \sim \mathcal{D}_v$.*

Graph, communications and mixing. We now formulate the communication assumptions we will make. For $k \geq 0$, as opposed to some previous Asynchronous Decentralized SGD analyses [Lian et al., 2017b, Bornstein et al., 2023], we do not want to assume that W_k mixes well in mean (i.e., that the spectral gap of $\mathbb{E}[W_k | W_0, \dots, W_{k-1}]$ is non-null or some other related assumption), since W_k may possibly be the identity matrix. We use the least restrictive assumption under which convergence of (synchronous) decentralized SGD is established [Koloskova et al., 2020], by assuming that if we wait enough communication updates, a consensus will ultimately be achieved.

Assumption 2 (Ergodic mixing). $W_k \in [0, 1]^{\mathcal{V} \times \mathcal{V}}$, $W_k \mathbf{1} = \mathbf{1}$, and there exist $\rho, k_\rho > 0$ such that we have $\forall k \in \mathbb{N}$ and $\forall \mathbf{x} \in \mathbb{R}^{\mathcal{V}}$:

$$\mathbb{E} \left[\left\| W^{(k+k_\rho)} \mathbf{x} - \bar{\mathbf{x}} \right\|^2 \middle| \mathcal{F}_k \right] \leq (1 - \rho)^2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2, \quad (5)$$

where for $k, \ell \geq 0$, $W^{(k, \ell)} = W_{\ell-1} \dots W_{k+1} W_k$, and $\mathcal{F}_k = \sigma(\mathbf{x}^s, \mathbf{g}^{s-1}, W_{s-1}, s \leq k)$ is the filtration up to step k .

This assumption makes it possible to consider any “reasonable” communication scheme. In the rest of the paper, when assuming that Assumption 2 holds for some constants (ρ, k_ρ) , we write $\bar{\rho} = \frac{e-1}{e} \frac{\rho}{k_\rho}$ (with $e = \exp(1)$), and this quantity is used in our main results.

Heterogeneous and homogeneous settings, sampling assumptions. Assuming that the sequence of nodes $(\mathcal{I}_k)_{k \geq 0}$ that iteratively perform local updates is arbitrary makes it possible to encompass all possible computation orderings and cover arbitrary delays. It is much more general than assuming that $\mathcal{I}_k = \mathcal{V}$ for all k (decentralized SGD) or $\mathcal{I}_k = \{v_k\}$ for

³which can easily be generalized to $\mathbb{E}[\|\nabla f_v(x) - \nabla_x F_v(x, \xi_v)\|^2] \leq \sigma^2 + \delta^2 \|\nabla f_v(x)\|^2$.

v_k sampled independently from the past, as assumed in most previous asynchronous decentralized works [Lian et al., 2015, Bornstein et al., 2023]. However, if functions f_v are not all equal and if the sequence v_k is arbitrary, convergence to the global function f cannot be assured (some of the nodes v might simply never appear during training). We therefore need to make some sampling assumption if we assume that local functions can be heterogeneous. We will thus assume either one the two following assumptions: **(i)** the heterogeneous setting where local functions f_v can be different, but where we make some node-sampling assumption for computations, and **(ii)** the homogeneous setting, where computations can be arbitrary, but functions f_v are all the same. Note that it is classical in asynchronous optimization to either assume **(i)** or **(ii)**; for instance, Asynchronous SGD with arbitrary orderings is proved to converge only under such assumptions [Mishchenko et al., 2022, Koloskova et al., 2022]. However, Asynchronous Decentralized works only assume that the sampling assumption **(i)** holds. Formally, we summarize these into the following two assumptions.

Assumption 3 (Heterogeneous setting). *There exists ζ^2 such that the **population variance** satisfies:*

$$\sum_{v \in \mathcal{V}} q_v \|\nabla f_v(x) - \nabla f(x)\|^2 \leq \zeta^2, \quad \forall x \in \mathbb{R}^d. \quad (6)$$

There exists $\mathbf{p} = (p_v)_{v \in \mathcal{V}} \in [0, 1]^{\mathcal{V}}$ such that the sequence $(\mathbb{1}_{v \in \mathcal{I}_k})_{k \geq 0}$ is i.i.d. distributed, with $\mathbb{P}(v \in \mathcal{I}_k) = p_v$ for all $k \geq 0, v \in \mathcal{V}$. We denote $\kappa_{\mathbf{p}} = \frac{p_{\max}}{\bar{p}}$, $p_{\max} = \max_v p_v$ and $\bar{p} = 1/n \sum_{v \in \mathcal{V}} p_v$. Furthermore, we assume that \mathbf{p} is proportional to \mathbf{q} : $\mathbf{p} = \beta \mathbf{q}$, and since $\sum_v q_v = 1$, we thus have $\beta = n\bar{p}$.

Assumption 4 (Homogeneous setting). *All functions f_v satisfy $f_v \equiv f$. No assumption on $(\mathcal{I}_k)_{k \geq 0}$.*

4 General Convergence Analysis

We now turn to our main results: convergence guarantees for AGRAF SGD, under a variety of regularity assumptions and settings. Note that in almost all cases, our rates do not depend on any upper bound on the maximal delays, which is a key feature of our analysis. This is also the case for asynchronous SGD [Koloskova et al., 2022, Mishchenko et al., 2022] or a recent asynchronous decentralized SGD work [Bornstein et al., 2023]. In this section, while presenting the results, we will only compare our results to degenerate baselines such as minibatch SGD, asynchronous SGD or decentralized SGD, in order to give simple arguments to show that our rates have expected order of magnitudes, leaving more complex comparisons and applications to be developed in Section 5. We first start with convex-Lipschitz losses. In this section, all

the rates are obtained for a **constant stepsize** γ (that differs in each different case and is time-horizon dependent), explicited in the proofs in the Appendix.

Theorem 1 (Lipschitz-convex rate). *Assume that f is convex and that for almost all (i.e., with probability 1) $\xi \sim \mathcal{D}$ $F_v(\cdot, \xi)$ is B -Lipschitz for some $B > 0$, let $D^2 \geq \|x_0 - x^*\|^2$, and $F_K = \mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right]$.*

1. In the **homogeneous** setting (Assumption 4),

$$F_K = \mathcal{O} \left(\sqrt{\frac{B^2 D^2 n \bar{\rho}^{-1}}{\sum_{k < K} |\mathcal{I}_k|}} \right).$$

2. In the **heterogeneous** setting (Assumption 3),

$$F_K = \mathcal{O} \left(\sqrt{\frac{B^2 D^2}{\sum_{k < K} |\mathcal{I}_k|}} \times n \sqrt{p_{\max}} (\sqrt{\kappa_{\mathbf{p}}} + \bar{\rho}^{-1}) \right).$$

We thus recover the well-known rate of minibatch SGD for convex-Lipschitz losses, by setting $\bar{\rho} = 1$ and $|\mathcal{I}_k| = n$, leading to the optimal rate $\mathcal{O}(\sqrt{B^2 D^2 / K})$ [Nemirovsky and Yudin, 1983]. Asynchronous SGD has also been studied under such assumptions, with the rate $\mathcal{O}(\sqrt{B^2 D^2 n / K})$ that we recover here ($\bar{\rho} = 1$ and $|\mathcal{I}_k| = 1$) [Mishchenko et al., 2022], that is min-max optimal [Woodworth et al., 2018]. No rates under the given assumptions existed for Decentralized (local) SGD, that thus exhibits a rate of $\mathcal{O}(\sqrt{B^2 D^2 \bar{\rho}^{-1} / K})$. Finally, adding the sampling assumption not only enables to handle heterogeneous functions, but also leads to improved rates: for well balanced weights ($p_v \approx \bar{p}$ and $\kappa_{\mathbf{p}} \approx 1$) we have $n\sqrt{\bar{p}}\bar{\rho}^{-1}$ instead of $n\bar{\rho}^{-1}$, which can improve the rate by a factor $1/\sqrt{n}$ if $\mathcal{O}(1)$ agents compute at the same time, which is usually the case in the asynchronous setting. This phenomenon (better rates under the sampling assumption) appears in all our other rates below.

Theorem 2 (Lipschitz-smooth-convex rate). *Assume that f is convex, for almost all $\xi \sim \mathcal{D}$, $F(\cdot, \xi)$ is B -Lipschitz for some $B > 0$, f_v is L -smooth, Assumption 1 holds, and let $D^2 \geq \|x_0 - x^*\|^2$. In the **homogeneous** setting,*

$$\begin{aligned} & \mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] \\ &= \mathcal{O} \left(\frac{L \bar{\rho}^{-1} n D^2}{\sum_{k < K} |\mathcal{I}_k|} \sqrt{\frac{\sigma^2 B^2}{\sum_{k < K} |\mathcal{I}_k|}} \right. \\ & \quad \left. + \left(\frac{D^2 n \sqrt{L} (B^2 + \bar{\rho}^{-1} \sigma^2)}{\sum_{k < K} |\mathcal{I}_k|} \right)^{\frac{2}{3}} \right) \end{aligned}$$

For Lipschitz-smooth functions, setting $\bar{\rho}^{-1} = 1$ and $|\mathcal{I}_k| = 1$, we recover the exact same rates as Asynchronous SGD under arbitrary delays, recently derived by Mishchenko et al. [2022], Koloskova et al. [2022], and that do not depend on any upper bound on the delays. These rates are thus extended to the more general AGRF SGD algorithm.

Theorem 3 (Smooth-convex). *Assume that f is convex, all f_v are L -smooth, and let $D^2 \geq \|x_0 - x^*\|^2$.*

1. In the **homogeneous** setting,

$$\begin{aligned} & \mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] \\ &= \mathcal{O} \left(\frac{LD^2(n\bar{\rho}^{-1} + \sqrt{n\tau_{\max}})}{\sum_{k < K} |\mathcal{I}_k|} + \sqrt{\frac{D\sigma^2}{\sum_{k < K} |\mathcal{I}_k|}} \right. \\ & \quad \left. + \left[\frac{D^2 \sqrt{L\sigma^2 n^2 \bar{\rho}^{-1}}}{\sum_{k < K} |\mathcal{I}_k|} \right]^{2/3} \right), \end{aligned}$$

where $\tau_{\max} \geq \sup_{k < K, v \in \mathcal{V}} \sum_{\ell=k}^{\tau(k+1, v)} |\mathcal{I}_\ell|$ is an upper bound on the maximal compute delay.

2. In the **heterogeneous** setting,

$$\begin{aligned} & \mathbb{E} \left[f \left(\frac{1}{K} \sum_{k < K} \bar{x}^k \right) - f(x^*) \right] \\ &= \mathcal{O} \left(\frac{LD^2 \sqrt{\kappa_{\mathbf{P}}} \left(\frac{1}{\bar{p}} + (\bar{\rho}\sqrt{\bar{p}})^{-1} \right)}{K} + \sqrt{\frac{D^2(\sigma^2 + \zeta^2)}{n\bar{p}K}} \right. \\ & \quad \left. + \left[\frac{D^2 \sqrt{L\sigma^2 p_{\max} \bar{\rho}^{-1} + L\zeta p_{\max} \bar{\rho}^{-2}}}{\bar{p}K} \right]^{2/3} \right). \end{aligned}$$

Removing the Lipschitz assumption, we are still able to recover and extend the rates of Asynchronous SGD with **constant** stepsizes. Note that under no sampling assumption, this rate depends on $\sqrt{n\tau_{\max}}$ instead of n as in the previous two theorems; however, this dependency is still better than depending on τ_{\max} since we always have $\tau_{\max} \geq n$. We expect to be able to remove this dependency by the use of varying stepsizes as was done for Asynchronous SGD (where stepsizes scale as $1/(L\tau(k))$, inversely proportional to the actual delay). However, such stepsizes cannot be used in a fully decentralized setting, since a given node cannot be aware of the iteration counter k and thus of the delay $\tau(k)$. Note also that in the sampling case, we have $\mathbb{E}[\sum_{k < K} |\mathcal{I}_k|] = n\bar{p}K$, so that the statistical rate is still reached. These comments also applies to the **non-convex and smooth setting** below, for which we fall back to showing that the algorithm will find an approximate first-order stationary point of the

objective. We recover, as in the convex-smooth case just above, the exact same rates as Koloskova et al. [2020] for Decentralized (local) SGD.

Theorem 4 (Non-convex and smooth rates). *Assume that the functions f_v are L -smooth.*

1. In the **homogeneous** setting,

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k < K} |\mathcal{I}_k| \|\nabla f(\bar{x}^k)\|^2 \right] \\ &= \mathcal{O} \left(\frac{LF_0(\sqrt{n\tau_{\max}} + n\bar{\rho}^{-1})}{K} + \left(\frac{L\sigma^2 F_0}{K} \right)^{\frac{1}{2}} \right. \\ & \quad \left. + \left(\frac{L\sigma n F_0}{K\sqrt{\bar{\rho}}} \right)^{\frac{2}{3}} \right). \end{aligned}$$

2. In the **heterogeneous** setting,

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{K} \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 \right] \\ &= \mathcal{O} \left(\frac{LF_0 \sqrt{\kappa_{\mathbf{P}}} \left(\frac{1}{\bar{p}} + (\bar{\rho}\sqrt{\bar{p}})^{-1} \right)}{K} + \left(\frac{L(\sigma^2 + \zeta^2) F_0}{K} \right)^{\frac{1}{2}} \right. \\ & \quad \left. + \left(\frac{LnF_0 \sqrt{\sigma^2 p_{\max} \bar{\rho}^{-1} + \zeta^2 p_{\max} \bar{\rho}^{-2}}}{K} \right)^{\frac{2}{3}} \right). \end{aligned}$$

Remark 1 (Heterogeneous without sampling). *So far, the heterogeneous setting was only considered under a sampling assumption. In fact, generalizing [Mishchenko et al., 2022, Theorem 4] to AGRF SGD, under both **heterogeneous functions** with population variance ζ^2 (as in eq. (6)) and **arbitrary ordering of the updates**, the exact same rate as Theorem 4.1 up to an additional term $\mathcal{O}(\zeta^2)$ could be obtained.*

5 Applications

5.1 Better rates for Asynchronous Decentralized SGD

A first direct application of our theory is a better analysis of Asynchronous Decentralized SGD. Comparing our analysis with those of Lian et al. [2017b], Bornstein et al. [2023], we highlight that our work handles arbitrary computation orders and delays in the homogeneous settings, as opposed to Lian et al. [2017b], Bornstein et al. [2023] that are only valid for $k_\rho = 1$ in Assumption 2 (which means that at any step k , conditionally on the current state, the graph of edges that can be sampled must be connected) and under a sampling assumption. In both homogeneous and heterogeneous cases, our communication assumptions are much less restrictive. Furthermore, under similar computation and regularity assumptions as Lian

Table 1: We compare the number of iterations required to reach the statistical regime $\mathcal{O}(\sigma^2/\sum_{k<K}|\mathcal{I}_k|)$ (if it is reached) of previous Asynchronous Decentralized SGD works [Lian et al., 2017b, Bornstein et al., 2023] with our rates (Theorems 2 and 3). **Strong** communication assumption : Assumption 2 with $k_\rho = 1$ and W_k independent from the past; **Sampling assumption** : $\mathcal{I}_k = \{v_k\}$ with v_k *i.i.d.* sampled or $\mathbb{P}(v \in \mathcal{I}_k) = p_v$ *i.i.d.* sampled. ^(a) Bornstein et al. [2023] reaches $\mathcal{O}(\bar{\rho}^{-2}\sqrt{\sigma^2/K})$ instead, after $\mathcal{O}(n^2\bar{\rho}^{-4})$ iterations.

Reference	Communication Assumption	Computation Assumption	Regularity	# iterations before $\sqrt{\frac{\sigma^2}{K}}$ regime
Lian et al. [2017b]	Strong	Sampling	Smoothness	$\mathcal{O}(\max(n^4\bar{\rho}^{-4}, \tau_{\max}^4))$
Bornstein et al. [2023]	Strong	Sampling	Smoothness	N.A. ^(a)
Theorem 2.1 (convex)	Assumption 2	Arbitrary	Smooth-Lipschitz, Homogeneous	$\mathcal{O}(n^4\bar{\rho}^{-2})$
Theorem 4.1	Assumption 2	Arbitrary	Smooth, Homogeneous	$\mathcal{O}(\max(n^4\bar{\rho}^{-2}, n\tau_{\max}))$
Theorem 4.2	Assumption 2	Sampling	Smoothness	$\mathcal{O}(n^2\bar{\rho}^{-4})$

et al. [2017b], Bornstein et al. [2023] (sampling and smooth losses, see last line of Table 1), our convergence bound (Theorem 3.2) reaches a statistical rate $\sqrt{\sigma^2/\sum_{k<K}|\mathcal{I}_k|}$ after $\sum_{k<K}|\mathcal{I}_k| = \mathcal{O}(n^2\bar{\rho}^{-4})$, while Bornstein et al. [2023] does not reach such a statistical rate and Lian et al. [2017b] reaches this rate after $K = \mathcal{O}(\max(n^4\bar{\rho}^{-4}, \tau_{\max}^4))$ iterations. For the sake of comparison, we take $\bar{\rho}$ of order 1 in our rates.

5.2 Asynchronous Decentralized SGD on Loss Networks

The previous considerations and the AGRAP SGD rates hold as long as there is no communication delay. The following question then arises: given a communication graph $G = (\mathcal{V}, \mathcal{E})$ with communication delays $\tau_{\{v,w\}}$ and computation delays τ_v for $v \in \mathcal{V}$ and $\{v,w\} \in \mathcal{E}$, can we reverse-engineer and build communication/computation schemes that fit in the AGRAP SGD framework and that do not break the communication and computation constraints? Can we analyze such a scheme and prove that it mixes well (in the sense that Assumption 2 holds, for explicit values of ρ, k_ρ) ?

Overview of the Loss Network scheme. Starting with $\mathcal{I}_k = \{v_k, w_k\} \in \mathcal{E}$, and communication matrices W_k corresponding to an averaging along the edge $\{v_k, w_k\}$ as a baseline (*i.e.*, $W_k = I_{\mathcal{V}} - \frac{(e_{v_k} - e_{w_k})(e_{v_k} - e_{w_k})^\top}{2}$ where (e_v) is the canonical basis of $\mathbb{R}^{\mathcal{V}}$) as a baseline, choosing a sequence \mathcal{I}_k such that there is no induced communication delays becomes tricky. While assuming that $\{v_k, w_k\}$ is sampled independently from the past with fixed probability [Lian et al., 2017b] is amenable for the analysis (since then Assumption 2 directly holds for $k_\rho = 1$), this can incur communication delays if for instance the same node is sampled in two consecutive updates.

To alleviate this issue, we remove the independence between sampled edges in the following way: we impose

that nodes that are already involved in a communication are tagged as *busy*, and that busy nodes cannot be involved in new communications. Then once a node finishes a computation, it can then choose a new neighbor (*who is not busy*) to start communicating with. Doing so, the induced communication matrices are no longer independent, as they follow a Markov process. This scheme is inspired by **Loss-Networks**, introduced in [Kelly, 1991] to model telecommunication networks, in which an edge in the graph models a phone communication that can happen; since a phone cannot make several calls in parallel, once involved in a communication with some neighboring node it cannot be called by another neighbor while it is busy; this is exactly the same process we use, phone calls being replaced by model communications.

How to schedule such a process ? If nodes start a new communication right after they finish their last one, the process can end up in deadlock and thus does not mix at all: this is for instance the case on the cycle or line graphs with an even number of nodes [Kelly, 1991]. We thus need to introduce some randomness and some waiting times. We proceed as follows and use exponential random waiting times as in Kelly [1991].

(i) Once a node v finishes a communication, it waits a time $T_v \sim \text{Exp}(p_v)$ (exponential random variable, of intensity p_v).

(ii) If v is still not busy after this waiting time, v samples some neighboring node $w \sim v$ with probability $\frac{p_{\{v,w\}}}{p_v}$ to communicate with, for $\sum_{w \sim v} p_{\{v,w\}} = p_v$.

(iii) If w is busy, this procedure restarts at (i), else both v and w become busy and can communicate. Once they are busy, they cannot communicate with other nodes. The communication between v and w consists in averaging local values by setting x_v, x_w to $(x_v + x_w)/2$. When this is done, they each perform a local (eventually delayed) gradient step, and then

become *non-busy*. Overall, the k^{th} update reads:

$$x_{v_k}^{k+1} = \frac{x_{v_k}^k + x_{w_k}^k}{2} - \gamma \nabla F_{v_k}(x_{v_k}^{k-\tau(v_k,k)}, \xi_{v_k}^{k-\tau(v_k,k)}), \quad (7)$$

and similarly at node w_k . The procedure described just above ((i)-(ii)-(iii)) to sample pairs of nodes that iteratively perform computations and pairwise communications can be instantiated locally, provided nodes know when their neighbors in the graph are busy — this can be relaxed by adding some “busy-checking” operation. However, the key challenge here lies in that the communication matrices $(W_k)_{k \geq 0}$ induced by the updates Equation (7) are not independent, and analyzing some form of ergodic mixing time becomes highly non-trivial. Still, using the randomness introduced in this procedure through the exponential waiting times and the sampling of neighbors, we are able to prove that Assumption 2 holds, for values of ρ, k_ρ that depend on the physical delays.

Assumption 5 (Loss Network assumptions). *There exist $\tau_v, \tau_{\{v,w\}} \in \mathbb{R}_{>0}$ ⁴ for $v \in \mathcal{V}, \{v,w\} \in \mathcal{E}$ > 0 such that a communication between v and w takes a time at most $\tau_{\{v,w\}}$, and computing a stochastic gradient at node v takes a time at most τ_v .*

Theorem 5. *Under Assumption 5, assume that $p_{\{v,w\}} = \min\left(\frac{1}{\max_{u \sim v} \tau'_{\{u,w\}}}, \frac{1}{2(\max(d_v, d_w) - 1)\tau'_{\{v,w\}}}\right)$, where d_v is the degree of node v and $\tau'_{\{v,w\}} = \tau_{\{v,w\}} + \max(\tau_v, \tau_w)$. Let Λ be the spectral gap (smallest non-null eigenvalue of the weighted Laplacian) of the graph G with weights $\lambda_{\{v,w\}} = \frac{\min_{u \sim \{v,w\}} p_{\{v,w\}}}{d \sum_{e \in \mathcal{E}} p_e}$, $\{v,w\} \in \mathcal{E}$, where d is the max degree in the graph. Then, Assumption 2 is verified for $\frac{\rho}{k_\rho} = \tilde{O}(\Lambda)$.*

Given a graph G with physical communication and computation latencies $\{\tau_v, \tau_{\{v,w\}}\}$ (Assumption 5), we are thus able to exhibit a communication scheme that satisfies communication and computation constraints, while still fitting in the framework of AGRAP SGD under the assumptions used in our convergence rates. Crucially, the mixing constant Λ explicitly depends on the graph and the delays, through the smallest non-null eigenvalue of the weighted graph Laplacian, with explicit weights $\lambda_{\{v,w\}}$ on the edges. These weights depend on **local** delays: having straggler nodes or edges do not slow down communication or computations, if there are fast edges/nodes that are dense enough in the graph. To further highlight the importance of having weights $\lambda_{\{v,w\}}$ that only depend on the local delays, this can be put in perspective of Asynchronous SGD, that is proved to depend only on the averaged computation delay $\frac{1}{n} \sum_{v \in \mathcal{V}} \frac{1}{\tau_v}$ rather than the max delay [Koloskova et al., 2022, Mishchenko et al., 2018]. For

decentralized optimization over a given graph, depending on the averaged communication delays wouldn’t make sense since all communication paths need to be taken into account; hence, the counterpart to the mean delay in the graph is a **weighted** Laplacian, with weights on edge $\{v,w\}$ that are function of **local** delays, instead of a max delay which is the asynchronous speedup [Even et al., 2021c]. **Disclaimer.** The proof of Theorem 5 is adapted from that of Even et al. [2021b], an unpublished work by a subset of the authors.

Conclusion

We introduced a unifying framework for studying asynchronous and decentralized algorithms; our analysis recovers and improves over that of previous asynchronous decentralized SGD works, while being much more general. The flexibility of our framework furthermore enables us to leverage an asynchronous speedup under communication and computation delays, by the introduction of Loss Networks and new analysis tools, thus providing a non-trivial sampling scheme that still satisfies the ergodic mixing property introduced by Koloskova et al. [2020].

Acknowledgements. M.E. thanks Konstantin Mishchenko for initiating discussions and suggesting this subject (asynchronous SGD on graphs) and for all the valuable discussions. A.K. and M.E. also thank Martin Jaggi for interesting discussions.

References

- Alekh Agarwal and John C. Duchi. Distributed delayed stochastic optimization. *Advances in Neural Information Processing Systems*, 24, 2011. 2
- Alekh Agarwal, Martin J. Wainwright, Peter L. Bartlett, and Pradeep K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009. 2
- Dan Alistarh, Demjan Grubic, Jerry Z. Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 1707–1718, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. 1
- Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G. Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020. 2

⁴ $\tau_v, \tau_{\{v,w\}}$ are **physical continuous-time** delays.

- Mahmoud S. Assran and Michael G. Rabbat. Asynchronous gradient push. *IEEE Transactions on Automatic Control*, 66(1):168–183, 2021. doi: 10.1109/TAC.2020.2981035. 1, 2
- Gerard M. Baudet. Asynchronous iterative methods for multiprocessors. *Journal of the ACM (JACM)*, 25(2):226–244, 1978. 1, 2
- Tal Ben-Nun and Torsten Hoefer. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Computing Surveys (CSUR)*, 52(4):1–43, 2019. 2
- Marco Bornstein, Tahseen Rabbani, Evan Z Wang, Amrit Bedi, and Furong Huang. SWIFT: Rapid decentralized federated learning via wait-free model communication. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4, 5, 6, 7, 8
- S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516. 2, 16
- Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. 5
- Jianmin Chen, Xinghao Pan, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous SGD. *arXiv preprint arXiv:1604.00981*, 2016. 1
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 1
- Edwige Cyffers, Mathieu Even, Aurélien Bellet, and Laurent Massoulié. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15889–15902. Curran Associates, Inc., 2022. 1
- Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9412–9439. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/even23a.html>. 17
- Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Hadrien Hendriks, Pierre Gaillard, Laurent Massoulié, and Adrien Taylor. Continuiized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 28054–28066. Curran Associates, Inc., 2021a. URL <https://proceedings.neurips.cc/paper/2021/file/ec26fc2eb2b75aece19c70392dc744c2-Paper.pdf>. 3
- Mathieu Even, Hadrien Hendriks, and Laurent Massoulié. Asynchrony and acceleration in gossip algorithms. *arXiv 2011.02379*, 2021b. 9, 15
- Mathieu Even, Hadrien Hendriks, and Laurent Massoulié. Decentralized optimization with heterogeneous delays: a continuous-time approach. *arXiv:2106.03585*, 2021c. 2, 9
- Hamid Feyzmahdavian and Mikael Johansson. Asynchronous iterations in optimization: New sequence results and sharper algorithmic guarantees. In *JMLR*, 2023, 2021. 2
- Samuel Horváth, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos I. Venieris, and Nicholas D. Lane. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34, 2021. 1
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser,

- Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019. 1
- F. P. Kelly. Loss networks. *The Annals of Applied Probability*, 1(3):319–378, 1991. 5, 8, 16
- Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3478–3487. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/koloskova19a.html>. 1
- 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. 1, 2, 5, 7, 9
- Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Sharper convergence guarantees for asynchronous SGD for distributed and federated learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4_oCZgBIVI. 2, 4, 6, 7, 9
- Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for non-convex optimization. *Advances in Neural Information Processing Systems*, 28, 2015. 2, 6
- Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 5336–5346, Red Hook, NY, USA, 2017a. Curran Associates Inc. ISBN 9781510860964. 1
- Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent, 2017b. URL <https://arxiv.org/abs/1710.06952>. 2, 4, 5, 7, 8
- Qi Liu, Bo Yang, Zhaojian Wang, Dafeng Zhu, Xinyi Wang, Kai Ma, and Xinpeng Guan. Asynchronous decentralized federated learning for collaborative fault diagnosis of pv stations. *IEEE Transactions on Network Science and Engineering*, 9(3):1680–1696, 2022. doi: 10.1109/TNSE.2022.3150182. 2
- Qinyi Luo, Jiaao He, Youwei Zhuo, and Xuehai Qian. Prague: High-performance heterogeneity-aware asynchronous decentralized training. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 401–416, 03 2020. 2
- Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017. doi: 10.1137/16M1057000. 2
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 5
- Konstantin Mishchenko, Franck Iutzeler, Jérôme Malick, and Massih-Reza Amini. A delay-tolerant proximal-gradient algorithm for distributed learning. In *International Conference on Machine Learning*, pages 3584–3592, 2018. 2, 9
- Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake Woodworth. Asynchronous sgd beats minibatch sgd under arbitrary delays, 2022. URL <https://arxiv.org/abs/2206.07638>. 2, 6, 7, 14
- Adel Nabli, Eugene Belilovsky, and Edouard Oyallon. $\mathbf{A}^2\mathbf{CiD}$: Accelerating asynchronous communication in decentralized deep learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=YE04aRkeZb>. 3
- Giorgi Nadiradze, Amirmojtaba Sabour, Peter Davies, Shigang Li, and Dan Alistarh. Asynchronous decentralized SGD with quantized and local updates. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=9x10Q5J8e9W>. 2
- Arkadii Semenovitch Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983. 6

- John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Mike Rabbat, Mani Malek, and Dzmitry Huba. Federated learning with buffered asynchronous aggregation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3581–3607. PMLR, 28–30 Mar 2022. URL <https://proceedings.mlr.press/v151/nguyen22b.html>. 5
- S. Sundhar Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545, July 2010. doi: 10.1007/s10957-010-9737-7. URL <https://doi.org/10.1007/s10957-010-9737-7>. 5
- Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in Neural Information Processing Systems*, 24, 2011. 1, 2
- Max Ryabinin, Eduard Gorbunov, Vsevolod Plokhotnyuk, and Gennady Pekhimenko. Moshpit SGD: Communication-efficient decentralized training on heterogeneous unreliable devices. *Advances in Neural Information Processing Systems*, 34, 2021. 1
- Suvrit Sra, Adams Wei Yu, Mu Li, and Alexander J. Smola. Adadelat: Delay adaptive distributed stochastic optimization. In *Artificial Intelligence and Statistics*, pages 957–965. PMLR, 2016. 2
- Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=S1g2JnRcFX>. 1, 5
- Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019. 14
- Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed updates. *Journal of Machine Learning Research*, 21: 1–36, 2020. 2
- Sebastian U. Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication, 2021. 15
- Mike Tanner. *Practical queueing analysis*. IBM McGraw-Hill. McGraw-Hill, London, 1995. URL <https://cds.cern.ch/record/2678155>. 22
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1
- John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 1986. 1, 2
- Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020. 1
- Blake E. Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in Neural Information Processing Systems*, 31, 2018. 6
- Xuyang Wu, Changxin Liu, Sindri Magnússon, and Mikael Johansson. Delay-agnostic asynchronous coordinate update algorithm. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37582–37606. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/wu23n.html>. 2
- Jiaqi Zhang and Keyou You. Fully asynchronous distributed optimization with linear convergence in directed networks, 2021. 2
- Shuxin Zheng, Qi Meng, Taifeng Wang, Wei Chen, Nenghai Yu, Zhi-Ming Ma, and Tie-Yan Liu. Asynchronous stochastic gradient descent with delay compensation. In *International Conference on Machine Learning*, pages 4120–4129, 2017. 2

A Equivalence of two ergodic mixing assumptions

The following assumption is a consequence of Assumption 2: if Assumption 2 holds for some τ, ρ , then Assumption 6 holds for $\bar{\rho} = c \frac{\rho}{\tau}$ where c is some numerical constant. In fact, as we prove in Proposition 1, they are both equivalent, but the following proves to be easier to handle in the analysis.

Assumption 6. $W_k \mathbf{1} = \mathbf{1}$ and there exist $\bar{\rho}$ such that we have $\forall k, \ell \in \mathbb{N}$ and $\forall \mathbf{x} \in \mathbb{R}^{\mathcal{V}}$:

$$\begin{aligned} \mathbb{E} \left[\left\| W^{(k:k+\ell)} \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \right\|^2 \middle| \mathcal{F}_k \right] \\ \leq 2(1 - \bar{\rho})^{2\ell} \left\| \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \right\|^2. \end{aligned} \quad (8)$$

Proposition 1. *Assumptions 2 and 6 are equivalent, in the following sense.*

1. *If Assumption 2 holds for some $\rho \in [0, 1]$ and for some $k_\rho \in \mathbb{N}^*$, then Assumption 6 holds for $\bar{\rho} = c \frac{\rho}{k_\rho}$, for $c > 0$ some numerical constant.*
2. *If Assumption 6 holds for some $\bar{\rho} \in [0, 1]$, then Assumption 2 holds for any $\rho \in (0, 1)$ and $k_\rho = \left\lceil \frac{\frac{1}{2} \ln(2) \ln(1-\rho)}{\ln(1-\bar{\rho})} \right\rceil$ ($\propto \frac{\rho}{\bar{\rho}}$ for $\rho, \bar{\rho}$ small).*

Proof. We first prove 1. Assume that Assumption 2 holds for some ρ, k_ρ . If Assumption 2 holds for ρ it holds for any $\rho' < \rho$, so that we can assume without loss of generality that $\rho \leq 1 - \sqrt{2}$. Let $k, \ell \in \mathbb{N}$ and $\mathbf{x} \in \mathbb{R}^{\mathcal{V}}$. Using Assumption 2 $\lfloor \frac{\ell}{k_\rho} \rfloor$, we have that:

$$\mathbb{E} \left[\left\| W^{(k:k+\ell)} \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \right\|^2 \middle| W_0, \dots, W_k \right] \leq (1 - \rho)^{2 \lfloor \ell/k_\rho \rfloor} \left\| \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \right\|^2.$$

Thus, $(1 - \rho)^{2 \lfloor \ell/k_\rho \rfloor} \leq (1 - \rho)^{2(\ell/k_\rho - 1)} \leq \frac{1}{(1-\rho)^2} (1 - \rho)^{2\ell/k_\rho}$. Then, $\frac{1}{(1-\rho)^2} \leq 2$ and $(1 - \rho)^{2\ell/k_\rho} \leq e^{-2\ell\rho/k_\rho} \leq (1 - c \frac{\rho}{k_\rho})^{2\ell}$ for $c \in (0, 1)$ some numerical constant ($c = \frac{e-1}{e}$), since $\frac{\rho}{k_\rho} \leq 1$.

We now prove 2. Assume that Assumption 6 holds for $\bar{\rho} > 0$, and let $\rho > 0$. We have:

$$\mathbb{E} \left[\left\| W^{(k:k+\ell)} \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \right\|^2 \middle| W_0, \dots, W_k \right] \leq (1 - \rho)^2 \left\| \mathbf{x} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{x} \right\|^2,$$

provided that ℓ satisfies:

$$2(1 - \bar{\rho})^{2\ell} \leq (1 - \rho)^2.$$

This is satisfied for:

$$\ell \geq \frac{\frac{1}{2} \ln(2) \ln(1 - \rho)}{\ln(1 - \bar{\rho})},$$

and thus Assumption 5 holds for ρ and $k_\rho = \left\lceil \frac{\frac{1}{2} \ln(2) \ln(1-\rho)}{\ln(1-\bar{\rho})} \right\rceil$. □

B Preliminaries for our convergence rates

For $k \geq 0$, , and for any $k \geq 0$ and $v \in \mathcal{V}$:

$$\text{next}(k, v) = \inf \{ \ell \geq k, v \in \mathcal{I}_\ell \}, \quad \text{prev}(k, v) = \sup \{ \ell < k, v \in \mathcal{I}_\ell \} \cup \{0\}, \quad \tau(k, v) = k - \text{prev}(k+1, v).$$

In other words, at a given iteration k , $\text{next}(k, v)$ is the iteration at which the node v will finish computing its current gradient, $\text{prev}(k, v)$ is the iteration at which the node v started computing its current gradient, and $\tau(k, v)$ is the current computational delay of node v at time k .

Let also $\bar{x}^k = \frac{1}{n} \sum_{v \in \mathcal{V}} x_v^k \in \mathbb{R}^d$ and $\mathbf{g}^k = (\mathbf{1}_{v \in \mathcal{I}_k} \nabla F_v(x_v^{\text{prev}(k,v)}, \xi_v^{\text{prev}(k,v)}))$, so that $\mathbf{x}^{k+1} = W_k \mathbf{x}^k - \gamma \mathbf{g}^k$.

B.1 Virtual iterate sequence to handle delays

As in [Mishchenko et al. \[2022\]](#), the delay analysis relies on the study of a virtual sequence. Noticing that $\bar{x}^{k+1} = \bar{x}^k - \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} g_v^{t-\tau(k,v)}$ and mimicking the analysis of asynchronous SGD, we introduce the sequence $\{\hat{x}^k, k \geq 1\}$ that lives in \mathbb{R}^d , defined through the following recursion:

$$\hat{x}^{k+1} = \hat{x}^k - \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} g_v^k, \quad \hat{x}_1 = \bar{x}_0 - \frac{\gamma}{n} \sum_{v \in \mathcal{V}} g_v^0.$$

We then have, for all $k \geq 1$:

$$\hat{x}^k - \bar{x}^k = -\frac{\gamma}{n} \sum_{v \in \mathcal{V} \setminus \mathcal{I}_k} g^{\text{prev}(k,v)}.$$

The difference $\|\hat{x}^k - \bar{x}^k\|$ can thus be easily bounded.

Lemma 1 (Virtual iterates control). *If stochastic gradients are bounded by a constant $B > 0$, we have:*

$$\|\hat{x}^k - \bar{x}^k\| \leq \gamma B. \quad (9)$$

In the general case,

$$\mathbb{E} \left[\|\hat{x}^k - \bar{x}^k\|^2 \right] \leq \frac{2\gamma^2}{n} \left(\sigma^2 + \sum_{v \in \mathcal{V}} \mathbb{E} \left[\|\nabla f_v(x_v^{\text{prev}(v,k)})\|^2 \right] \right). \quad (10)$$

Proof. Equation (9) is proved using a triangle inequality, while Equation (10) is a direct application of [\[Stich and Karimireddy, 2019, Lemma 15\]](#). \square

B.2 Consensus control

Lemma 2 (Consensus control). *We have:*

$$\sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq 2\gamma^2 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| + \frac{4\gamma^2}{\bar{\rho}^2} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\|\nabla f_v(x_v^{k-\tau(k,v)})\|^2 \right] \quad (11)$$

$$\leq 2\gamma^2 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| + \frac{4\gamma^2}{\bar{\rho}^2} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\|\nabla f_v(x_v^k)\|^2 \right]. \quad (12)$$

If the stochastic gradients are bounded by some $B > 0$,

$$\sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq \frac{2\gamma^2 B^2}{\bar{\rho}^2} \sum_{k < K} |\mathcal{I}_k|. \quad (13)$$

Proof. Under Assumption 2, we can bound the variations of $\mathbf{x}^k - \bar{\mathbf{x}}^k$ (here, $\bar{\mathbf{x}}^k = \mathbf{1} \mathbf{1}^\top \mathbf{x}^k$). Using Cauchy-Schwarz inequality, for $a_m > 0$ scalars and $b_m \in \mathbb{R}^p$ vectors, we have:

$$\left\| \sum_m b_m \right\|^2 \leq \left(\sum_m a_m^{-1} \right) \left(\sum_m a_m \|b_m\|^2 \right).$$

We now apply this to $\mathbf{x}^k - \bar{\mathbf{x}}^k = -\gamma \sum_{m=0}^k W^{(m:k)}(\tilde{\mathbf{g}}^m - \bar{\tilde{\mathbf{g}}}^m)$ to obtain:

$$\begin{aligned} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 &= \mathbb{E} \left[\left\| \gamma \sum_{m=0}^k W^{(m:k)}(\tilde{\mathbf{g}}^m - \bar{\tilde{\mathbf{g}}}^m) \right\|^2 \right] \\ &\leq \gamma^2 \sum_{m'=0}^k (1-\bar{\rho})^{k-m'} \sum_{m=0}^k (1-\bar{\rho})^{-(k-m)} \mathbb{E} \left[\left\| W^{(m:k)}(\tilde{\mathbf{g}}^m - \bar{\tilde{\mathbf{g}}}^m) \right\|^2 \right] \\ &\leq 2\gamma^2 \frac{1}{\bar{\rho}} \sum_{m=0}^k (1-\bar{\rho})^{k-m} \mathbb{E} \left[\|\tilde{\mathbf{g}}^m\|^2 \right] \end{aligned}$$

leading to, if stochastic gradients are bounded by B :

$$\mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq \frac{2\gamma^2 B^2}{\bar{\rho}} \sum_{\ell < k} (1-\bar{\rho})^{k-\ell} |\mathcal{I}_\ell|,$$

and thus:

$$\sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq \frac{2\gamma^2 B^2}{\bar{\rho}^2} \sum_{k < K} |\mathcal{I}_k|.$$

We also have, using a bias-variance decomposition (not exactly, since the \mathbf{g}^m are not independent, but using the martingale version as in [Stich and Karimireddy, 2021, Lemma 15]):

$$\begin{aligned} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 &= \mathbb{E} \left[\left\| \gamma \sum_{m=0}^k W^{(m:k)}(\tilde{\mathbf{g}}^{m-\tau(m)} - \bar{\tilde{\mathbf{g}}}^m) \right\|^2 \right] \\ &\leq 2\gamma^2 \sigma^2 \sum_{\ell < k} (1-\bar{\rho})^{k-\ell} |\mathcal{I}_\ell| + \frac{4\gamma^2}{\bar{\rho}} \sum_{m=0}^k (1-\bar{\rho})^{k-m} \sum_{v \in \mathcal{I}_m} \mathbb{E} \left[\left\| \nabla f_v(x_v^{(m-\tau(m),v)}) \right\|^2 \right], \end{aligned}$$

so that:

$$\sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \leq 2\gamma^2 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| + \frac{4\gamma^2}{\bar{\rho}^2} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\left\| \nabla f_v(x_v^{k-\tau(k,v)}) \right\|^2 \right].$$

□

C Loss Networks analysis

Disclaimer. This proof is adapted from that of Even et al. [2021b], an unpublished work by a subset of the authors.

In this section, we prove Theorem 5 and provide some more information on loss networks. The updates of *decentralized SGD on loss networks* write as:

$$\begin{cases} x_{v_k}^{k+1} = \frac{x_{v_k}^k + x_{w_k}^k}{2} - \gamma \nabla F_{v_k}(x_{v_k}^{k-\tau(v_k,k)}, \xi_{v_k}^{k-\tau(v_k,k)}) \\ x_{w_k}^{k+1} = \frac{x_{v_k}^k + x_{w_k}^k}{2} - \gamma \nabla F_{w_k}(x_{w_k}^{k-\tau(w_k,k)}, \xi_{w_k}^{k-\tau(w_k,k)}) \end{cases}, \quad (14)$$

leading to $\mathbf{x}^{k+1} = W_k \mathbf{x}^k - \gamma \mathbf{g}^k$, for $W_k = W_{\{v_k, w_k\}} = I_V - \frac{(e_{v_k} - e_{w_k})(e_{v_k} - e_{w_k})^\top}{2}$, and \mathbf{g}^k the corresponding delayed gradients. Note then that this takes the same form as the **AGRAF SGD** sequence.

Definition 1 (Poisson point process (P.p.p.)). *A Poisson point process of intensity $p > 0$ is a random discrete subset \mathcal{P} of $\mathbb{R}_{\geq 0}$ that can be written as $\mathcal{P} = \{T_0 < T_1 < \dots < T_k < \dots\}$, where $(T_k - T_{k-1})_{k \geq 1}$ are i.i.d. exponential random variables of mean $\frac{1}{p}$.*

Boyd et al. [2006] consider a model (without any delay) for gossip algorithms, where updates are that of Equation (14) without the gradient steps, and these updates happen at the times of Poisson point processes (a *P.p.p.* of intensity $p_{\{v,w\}}$ for an update along $\{v,w\}$). Consequently, W_k is independent from the past, and $\mathbb{P}(W_k = W_{\{v,w\}}) \propto p_{\{v,w\}}$.

The *P.p.p. model* considered in Boyd et al. [2006] where the updates are performed at the times of *Poisson point processes* is particularly amenable to analysis, but it assumes that communications and computations are done instantaneously. Thus, actual implementations differ from its underlying assumptions, unless further synchrony is assumed. To alleviate this issue, with pairwise communications ruled by point processes as a baseline, we consider a protocol in which nodes are tagged as *busy* when they are already engaged in an update, and communications between busy nodes are forbidden. Our model is inspired from classical Loss Network models [Kelly, 1991], in which edges are activated following the same procedure as in the *P.p.p. model*, with a *P.p.p.* of intensity $p_{\{v,w\}}$. Note that we do not consider these intensities to be constraints of the problem, but rather parameters of the algorithm, that can be tuned. Each node has an exponential clock of intensity $p_v \frac{1}{2} \sum_{w \sim v} p_{\{v,w\}}$. At each clock-ticking, if v is not busy, it selects a neighbor w with probability $p_{\{v,w\}} / \sum_{u \sim v} p_{\{u,v\}}$. If w is not *busy*, v and w compute and exchange information, becoming busy for a duration $\tau'_{\{v,w\}}$. We can think of this procedure as classical gossip on an underlying random graph that follows a Markov-Chain process. The difference between our communication model on Loss Networks and the *P.p.p.* model lies in that in our case, W_k is not independent on the past. In fact, we have:

$$\mathbb{P}(\{v_k, w_k\} = \{v, w\} | \mathcal{F}_k) = \frac{\mathbb{1}_{\{v,w \text{ not busy at time } T_k\}} p_{\{v,w\}}}{\sum_{\{u,u'\} \in \mathcal{E}} \mathbb{1}_{\{u,u' \text{ not busy at time } T_k\}} p_{\{u,u'\}}},$$

leading to complicated intricacies between the matrices $(W_k)_k$, that we need to handle.

Proving Theorem 5 requires to show that there exist ρ, k_ρ (that need to be computed) such that for any $k \geq 0$, $\mathbf{x} \in \mathbb{R}^V$,

$$\mathbb{E} \left[\left\| W_{\{v_{k+k_\rho-1}, w_{k+k_\rho-1}\}} \cdots W_{\{v_k, w_k\}} (\mathbf{x} - \bar{\mathbf{x}}) \right\|^2 | \mathcal{F}_k \right] \leq (1 - \rho)^2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2.$$

Our proof of Theorem 5 follows three main steps: *i*) Deriving convergence results for more general communication schemes than loss networks, under deterministic assumptions on the activations. *ii*) Adapting Step i) to stochastic assumptions on the delays. *iii*) Deriving high-probability upper-bounds on the delays between two activations in loss networks in order to fall under the assumptions of Step i).

C.1 Descent lemma under deterministic assumptions on the activations

We consider general activation processes $\mathcal{P}_{\{v,w\}}$, where we define $\mathcal{P}_{\{v,w\}}$ as $\mathcal{P}_{\{v,w\}} = \{T_k : \{v_k, w_k\} = \{v, w\}\}$, and these times are called **activation times of edge $\{v,w\}$** . When edge $\{v,w\}$ is activated, the update described in (14) is performed. The delay of an edge is defined as its (random) waiting time between two activations. Two ergodicity-like conditions on the delays are needed: *(i) edges activated regularly enough and (ii) incident edges must not be activated too many times.*

We now formally introduce these assumptions. We consider discrete time in this section: more precisely, $k \in \mathbb{N}$ stands for the k -th edge activation.

Definition 2. Consider a communication scheme with edge-activation point processes $\mathcal{P}_{\{v,w\}}$. Let $k = 0, 1, 2, \dots$ index the consecutive edge activations. Let $\ell \in \mathbb{N}$, $\{v, w\}$ and $\{u, u'\} \in E$. Let $k_{\{v,w\}} < \ell_{\{v,w\}}$ such that $k_{\{v,w\}} \leq k < \ell_{\{v,w\}}$ be consecutive activation times (in discrete time) of $\{v, w\}$. Denote $T_{\{v,w\}}(k) = \ell_{\{v,w\}} - k_{\{v,w\}} - 1$ the total number of edge activations between the two consecutive activations of $\{v, w\}$. Denote $N(\{u, u'\}, \{v, w\}, k)$ the number of activations of edge $\{v, w\}$ in the activations $\{s_{\{v,w\}}, s_{\{v,w\}} + 1, \dots, t_{\{v,w\}} - 1\}$.

Assumption 7 (Delay Assumptions). There exist $T \in \mathbb{N}^*$, $a, b > 0$, and $\ell_{\{v,w\}} > 0$, $\{v, w\} \in E$ such that, for the quantities and the communication scheme in Definition 2:

1. For all $k \in \mathbb{N}$, all edges are activated between iterations k and $k + T - 1$.
2. $\forall k \geq 0, \forall (\{v, w\}) \in E, T_{\{v,w\}}(k) \leq a \ell_{\{v,w\}}$: $(\{v, w\})$ is activated at least every $a \ell_{\{v,w\}}$ activations.
3. $\forall k \geq 0, \forall (\{v, w\}), (\{u, u'\}) \in E$ such that $(\{u, u'\}) \sim (\{v, w\})$, $N(\{u, u'\}, \{v, w\}, k) \leq \lceil \frac{b \ell_{\{v,w\}}}{\ell_{\{u,u'\}}} \rceil$.

Assumption (1) is implied by Assumption (2) if $T = \max_{\{v,w\}} \ell_{\{v,w\}}$. Taking $\ell_{\{v,w\}}$ as a deterministic upper-bound on the delays of edge $(\{v,w\})$ between two activations in continuous time is sufficient to have Assumption (2) and (3), with some normalizing constant a , and b such that $\ell_{\{v,w\}}/b$ is a lower-bound on these delays.

The main technical difficulty lies in the fact that at a defined activation time t , some nodes are not available: at any time $k \geq 0$, $\sum_{(\{v,w\}) \in E \text{ not busy}} W_k$ usually differs from $\sum_{\{v,w\}} p_{\{v,w\}} W_{\{v,w\}}$ (and $\sum_{(\{v,w\}) \in E \text{ not busy}} W_k$ may have a null spectral gap) as in *Markov-Chain Gradient Descent* [Even, 2023], thus making an analysis such as in the *P.p.p. model* impossible. To alleviate this difficulty, in order to make sure that all edges are taken into account when performing the averaging, the Lyapunov function Λ_k that we study considers the value of the objective for T consecutive activation times. It is defined as follows:

$$\forall k \in \mathbb{N}, \Lambda_k(\mathbf{x}) = \frac{1}{T} \sum_{\ell=k}^{k+T-1} \left\| W^{(0,\ell)}(\mathbf{x} - \bar{\mathbf{x}}) \right\|^2, \quad \mathbf{x} \in \mathbb{R}^{\mathcal{V}}.$$

The first step of the proof of Theorem 5 consists in proving the following.

Theorem 6. *Consider a general communication scheme as in Definition 2, that satisfies Assumption 7 for constants $\ell_{\{v,w\}}, a, b > 0$. Let γ be the smallest positive eigenvalue of the Laplacian of the graph G with weights:*

$$\nu_{\{v,w\}} = C \ell_{\{v,w\}}^{-1} \min_{\{u,u'\} \sim \{v,w\}} \frac{\ell_{\{u,u'\}}}{\ell_{\{v,w\}}}, \quad \{v,w\} \in \mathcal{E},$$

where $C = \frac{1}{2a+8d_{\max}^2 ab}$. Then we have, for all $k, \ell \in \mathbb{N}$:

$$\Lambda_{k+\ell}(\mathbf{x}) \leq (1 - \gamma)^\ell \Lambda_k(\mathbf{x}).$$

Proof. We fix $\mathbf{x} \in \mathbb{R}^{\mathcal{V}}, k, \ell$. To prove this intermediate theorem, we need to study every matrix multiplication involved. At iteration k , not every coordinates is available, hence the need to study the impact of T multiplications together.

A gradient step alongside edge $\{v,w\}$ only involves edges in its neighborhood (thanks to the sparsity of the matrix A), a key element that will need to be explicitated. The proof involves three main steps.

Before that, we need to introduce **edge dual variables**. Matrix multiplications by matrices like $W_{\{v,w\}}$ aim at minimizing the function $F(\mathbf{y}) = \frac{1}{2} \sum_{v \in \mathcal{V}} (y_v - x_v)^2$, which is minimized at $\mathbf{y} = \bar{\mathbf{x}}$. A standard way to deal with the constraint $x_1 = \dots = x_n$, is to use a dual formulation, by introducing a dual variable $\lambda \in \mathbb{R}^{\mathcal{E}}$ indexed by the edges. We first introduce a matrix $A \in \mathbb{R}^{\mathcal{V} \times \mathcal{E}}$ such that $\text{Ker}(A^\top) = \text{Vect}(\mathbb{1})$ where $\mathbb{1}$ is the constant vector $(1, \dots, 1)^\top$. A is chosen such that:

$$\forall \{v,w\} \in E, A e_{\{v,w\}} = \mu_{\{v,w\}}(e_v - e_w). \quad (15)$$

for some non-null constants $\mu_{\{v,w\}}$. We define $\nu_{\{v,w\}} = -\mu_{\{v,w\}}$ for this writing to be consistent. This matrix A is a square root of the laplacian of the graph weighted by $\nu_{\{v,w\}} = \mu_{\{v,w\}}^2$. The constraint $x_1 = \dots = x_n$ can then be written $A^\top x = 0$. The dual problem reads as follows:

$$\min_{\mathbf{y} \in \mathbb{R}^{\mathcal{V}}, A^\top \mathbf{y} = 0} F(\mathbf{y}) = \min_{\mathbf{y} \in \mathbb{R}^{\mathcal{V}}} \max_{\lambda \in \mathbb{R}^{\mathcal{E}}} F(\mathbf{y}) - \langle A^\top \mathbf{y}, \lambda \rangle.$$

Let $F_A^*(\lambda) := F^*(A\lambda) = F_A(\lambda)$ for $\lambda \in \mathbb{R}^{E \times d}$ where F^* is the Fenchel conjugate of F . Now, notice that for our particular form of F , we in fact have $F^* = F$. The dual problem reads

$$\min_{\mathbf{y} \in \mathbb{R}^{\mathcal{V}}, y_1 = \dots = y_n} F(\mathbf{y}) = \max_{\lambda \in \mathbb{R}^{\mathcal{E}}} -F_A(\lambda).$$

Thus $F_A^*(\lambda)$ is to be minimized over the dual variable $\lambda \in \mathbb{R}^{\mathcal{E}}$.

We now make a parallel between pairwise operations between adjacent nodes in the network and coordinate gradient steps on F_A^* . As $F_A^*(\lambda) = \max_{\mathbf{y} \in \mathbb{R}^{\mathcal{V}}} -F(\mathbf{y}) + \langle A\lambda, \mathbf{y} \rangle$, to any $\lambda \in \mathbb{R}^{\mathcal{E}}$ a primal variable $\mathbf{y} \in \mathbb{R}^{\mathcal{V}}$ is

uniquely associated through the formula $\nabla F(\mathbf{y}) = A\lambda$. The partial derivative of F_A^* with respect to coordinate $\{v, w\} \in \mathcal{E}$ of λ reads :

$$\nabla_{\{v,w\}} F_A^*(\lambda) = (Ae_{\{v,w\}})^\top \nabla F^*(A\lambda) = \mu_{\{v,w\}} (\nabla g_v^*((A\lambda)_v) - \nabla g_w^*((A\lambda)_w)),$$

where we denote $g_v(y) : \frac{1}{2}(y - x_v)^2$. Consider then the following step of coordinate gradient descent for F_A^* on coordinate $\{v, w\}$ of λ , performed when edge $\{v, w\}$ is activated at iteration k (corresponding to time T_k), and where $U_{\{v,w\}} = e_{\{v,w\}} e_{\{v,w\}}^\top$:

$$\lambda_{k+1} = \lambda_k - \frac{1}{\mu_{\{v,w\}}^2} U_{\{v,w\}} \nabla_{\{v,w\}} F_A^*(\lambda_k). \quad (16)$$

Denoting $\mathbf{y}_k = A\lambda_k \in \mathbb{R}^{\mathcal{V}}$, we obtain the following formula for updating coordinates v and w of \mathbf{y} when $\{v, w\}$ is activated:

$$y_{v,k+1} = y_{v,k} - \frac{1}{2}(y_{v,k} - y_{w,k}) = \frac{1}{2}(y_{v,k} + y_{w,k}) = y_{w,k+1}. \quad (17)$$

Thus, $\mathbf{y}^{k+1} = W_k \mathbf{y}^k$ is equivalent to $\lambda_{k+1} = \lambda_k - \frac{1}{2\mu_{\{v_k, w_k\}}^2} \nabla_{\{v_k, w_k\}} F_A^*(\lambda_k)$, which is easier to study. Also, notice that this is the consensus distance exactly: $F_A^*(\lambda) = F(\mathbf{y})$ for $\mathbf{y} = A\lambda$.

Hence, $\Lambda_k(\mathbf{x}) = F(\mathbf{y}^k) = F_A^*(\lambda_k)$ here $\mathbf{y}^k = A\lambda^k$ is obtained with the recursion $\lambda^{k+1} = \lambda^k - \frac{1}{2\mu_{\{v_k, w_k\}}^2} \nabla_{\{v_k, w_k\}} F_A^*(\lambda^k)$, with initialisation $\mathbf{y}^0 = \mathbf{x}$: we thus study this sequence.

Step 1: First, notice that F_A^* is $\mu_{\{v,w\}}^2$ -smooth along every coordinate $\{v, w\}$, so that using local smoothness, for all $\{v, w\} \in \mathcal{E}$ and $\lambda \in \mathbb{R}^{\mathcal{E}}$, for $\gamma \leq \frac{1}{2\mu_{\{v,w\}}^2}$, we have:

$$F_A^*(\lambda - \gamma \nabla_{\{v,w\}} F_A^*(\lambda)) - F_A^*(\lambda) \leq \frac{1}{4\mu_{\{v,w\}}^2} \|\nabla_{\{v,w\}} F_A^*(\lambda)\|^2. \quad (18)$$

Applying Equation (18), where $\{v_\ell, w_\ell\}$ is the ℓ^{th} activated edge:

$$F_A^*(\lambda^{\ell+1}) - F_A^*(\lambda^\ell) \leq -\frac{1}{4\mu_{\{v_\ell, w_\ell\}}^2} \|\nabla_{\{v_\ell, w_\ell\}} F_A^*(\lambda^\ell)\|^2. \quad (19)$$

Hence, summing:

$$\Lambda_{k+1} \leq \Lambda_k - \frac{1}{T} \sum_{k \leq \ell < k+T} \frac{1}{4\mu_{\{v_\ell, w_\ell\}}^2} \|\nabla_{\{v_\ell, w_\ell\}} F_A^*(\lambda^\ell)\|^2, \quad (20)$$

Notice that:

$$\frac{1}{T} \sum_{k \leq \ell < k+T} \sum_{\{v,w\} \in \mathcal{E}} \|\nabla_{\{v,w\}} F_A^*(\lambda^\ell)\|^2 = \frac{1}{T} \sum_{k \leq \ell < k+T} \|\nabla F_A^*(\lambda^\ell)\|^2 \geq \sigma_A \Lambda_t \quad (21)$$

σ_A is the strong convexity parameter of F_A^* which is equal to lower bounded by $\lambda_{\min}^+(A^T A)$, which itself is exactly the smallest positive non-null eigenvalue of the graph Laplacian with weights $\mu_{\{v,w\}}^2$. Hence, if an inequality of the type

$$\frac{C}{T} \frac{1}{T} \sum_{k \leq \ell < k+T} \sum_{\{v,w\} \in \mathcal{E}} \|\nabla_{\{v,w\}} F_A^*(\lambda^\ell)\|^2 \leq \frac{1}{4\mu_{\{v_\ell, w_\ell\}}^2} \|\nabla_{\{v_\ell, w_\ell\}} F_A^*(\lambda^\ell)\|^2 \quad (22)$$

holds, we have using strong convexity:

$$\Lambda_{k+1} \leq \Lambda_k - \frac{C}{T} \sum_{k \leq \ell < k+T} \|\nabla F_A^*(\lambda^\ell)\|^2 \leq (1 - C\sigma_A) \Lambda_k. \quad (23)$$

We thus need to tune correctly the $\mu_{\{v,w\}}^2$ and C in order to have (22) verified.

Step 2: We are looking for necessary conditions for (22) to hold. In the left term, every coordinate is present at each time ℓ . However, in the right hand side of the inequality, just the activated one is present. We will need

to compensate this with a bigger factor in front of the gradients. In order to compare these quantities, we need to introduce upper bound inequalities on $\|\nabla_{\{v,w\}}F_A^*(\lambda(s))\|^2$, that only make activated coordinates intervene. Let $s \in \{t, \dots, t+T-1\}$, and suppose that there exists $t \leq r \leq s < r+t_{\{v,w\}} \leq t+T-1$ such that $\{v,w\}$ is activated at times r and $r+t_{\{v,w\}}$. Thanks to the assumption on T , either one of these integers exists. If the other one doesn't, replace it with t for r , and by $t+T-1$ for $r+t_{\{v,w\}}$. Thanks to our assumptions, we know that $t_{\{v,w\}} \leq a\ell_{\{v,w\}}$. We have the following basic inequalities:

$$\|\nabla_{\{v,w\}}F_A^*(\lambda(s))\|^2 \leq (\|\nabla_{\{v,w\}}F_A^*(\lambda(r))\| + \|\nabla_{\{v,w\}}F_A^*(\lambda(s)) - \nabla_{\{v,w\}}F_A^*(\lambda(r))\|)^2 \quad (24)$$

$$\leq 2(\|\nabla_{\{v,w\}}F_A^*(\lambda(r))\|^2 + \|\nabla_{\{v,w\}}F_A^*(\lambda(s)) - \nabla_{\{v,w\}}F_A^*(\lambda(r))\|^2). \quad (25)$$

The quantity $\|\nabla_{\{v,w\}}F_A^*(\lambda(s)) - \nabla_{\{v,w\}}F_A^*(\lambda(r))\|^2$ then needs to be controlled. We use the following lemma.

Lemma 3. *For $\lambda, \lambda' \in R^E$, and $\{v,w\} \in E$, we have:*

$$\|\nabla_{\{v,w\}}F_A^*(\lambda) - \nabla_{\{v,w\}}F_A^*(\lambda')\|^2 \leq 8d_{\{v,w\}}\mu_{\{v,w\}}^2 \sum_{(\{u,u'\} \sim \{v,w\})} \mu_{\{u,u'\}}^2 \|\lambda_{\{u,u'\}} - \lambda'_{\{u,u'\}}\|^2. \quad (26)$$

Proof. First, notice that $\nabla_{\{v,w\}}F_A^*(\lambda) = \mu_{\{v,w\}}(\nabla g_i^*((A\lambda)_v) - \nabla g_j^*((A\lambda)_w))$. Then:

$$\begin{aligned} \|\nabla_{f_v^*}((A\lambda)_v) - \nabla_{f_w^*}((A\lambda)_w)\| &= \|(A(\lambda - \lambda'))_v\| \text{ (smoothness)} \\ &= \left\| \sum_{\{u,u'\} \sim \{v,w\}} \mu_{\{u,u'\}}(\lambda - \lambda')_{\{u,u'\}} \right\| \\ &\leq \sum_{\{u,u'\} \sim \{v,w\}} \mu_{\{u,u'\}} \|(x - x')_{\{u,u'\}}\| \end{aligned}$$

Conclude by taking the square and summing for v and w . \square

Using this with $\lambda = \lambda(s)$ and $\lambda' = \lambda(r)$:

$$\|\nabla_{\{v,w\}}F_A^*(\lambda(s))\|^2 \leq 2\|\nabla_{\{v,w\}}F_A^*(\lambda(r))\|^2 \quad (27)$$

$$+ 2d_{\{v,w\}} \sum_{r < k < r+t_{\{v,w\}}} N(\{v_k, w_k\}, \{v, w\}, k) \frac{\mu_{\{v,w\}}^2}{2\mu_{\{v_k, w_k\}}^2} \|\nabla_{\{v_k, w_k\}}F_A^*(\lambda(k))\|^2 \quad (28)$$

$$\leq 2\|\nabla_{\{v,w\}}F_A^*(\lambda(r))\|^2 \quad (29)$$

$$+ 2d_{\{v,w\}} \sum_{r < k < r+t_{\{v,w\}}} \left[b \frac{\ell_{\{v,w\}}}{L_{\{v_k, w_k\}}} \right] \frac{\mu_{\{v,w\}}^2}{\mu_{\{v_k, w_k\}}^2} \|\nabla_{\{v_k, w_k\}}F_A^*(\lambda(k))\|^2 \quad (30)$$

The advantage of this last expression is that only activated quantities are present on the right hand side.

Step 3: The last step of the proof consists in summing the last inequality for $t \leq \ell < t+T$, $\{v,w\} \in E$. When summing, each $\|\nabla_{\{v_k, w_k\}}F_A^*(\lambda(k))\|^2$ appears on the right hand-side of the inequality, with a factor upper-bounded by (here instead of $\{v_k, w_k\}$ we write $(\{v, w\})$):

$$2a\ell_{\{v,w\}} + 2d_{\{v,w\}} \sum_{\{u,u'\} \sim \{v,w\}} a\ell_{\{u,u'\}} \left[\frac{b\ell_{\{u,u'\}}}{\ell_{\{v,w\}}} \right] \frac{\mu_{\{u,u'\}}^2}{\mu_{\{v,w\}}^2}. \quad (31)$$

We want the expression above multiplied by C defined in Step 1 to be upper-bounded by $\frac{1}{4\mu_{\{v,w\}}^2}$, in order for (22) to be verified. This is possible if and only if:

$$C \left(4a\ell_{\{v,w\}}\mu_{\{v,w\}}^2 + 4d_{\{v,w\}} \sum_{\{u,u'\} \sim \{v,w\}} a \left[\frac{b\ell_{\{u,u'\}}}{\ell_{\{v,w\}}} \right] \ell_{\{u,u'\}}\mu_{\{u,u'\}}^2 \right) \leq \frac{1}{2}, \quad (32)$$

where C is defined in step 1 of the proof. This is equivalent to:

$$C \left(a\ell_{\{v,w\}}\mu_{\{v,w\}}^2 + d_{\{v,w\}} \sum_{\{u,u'\} \sim \{v,w\}} a \frac{b\ell_{\{u,u'\}}^2}{\ell_{\{v,w\}}} \mu_{\{u,u'\}}^2 \right) \leq \frac{1}{8}$$

if $\forall \{u, u'\} \sim \{v, w\}, \ell_{\{v,w\}} \leq b\ell_{\{u,u'\}},$

where we bounded $\lceil b \frac{\ell_{\{v,w\}}}{\ell_{\{u,u'\}}} \rceil$ by $2 \frac{b\ell_{\{v,w\}}}{\ell_{\{u,u'\}}}$ here. We here see that in this case, if

$$\mu_{\{v,w\}}^2 = \frac{1}{2\ell_{\{v,w\}}} \times \min_{\{u,u'\} \sim \{v,w\}} \frac{\ell_{\{u,u'\}}}{\ell_{\{v,w\}}} \quad (33)$$

with $8a + 8d_{max}^2 b \leq C^{-1}$, our inequality holds. However, our inequality on the ceil operator seems not to work in the general case. Let's take $\{u, u'\}$ a neighbor of $\{v, w\}$ such that $\ell_{\{v,w\}} > b\ell_{\{u,u'\}}$. As $\ell_{\{v,w\}} > b\ell_{\{u,u'\}}$, we have $\lceil \frac{b\ell_{\{u,u'\}}}{\ell_{\{v,w\}}} \rceil = 1$, leading to $a \lceil \frac{b\ell_{\{u,u'\}}}{\ell_{\{v,w\}}} \rceil \ell_{\{u,u'\}} \mu_{\{u,u'\}}^2 = a\ell_{\{u,u'\}} \mu_{\{u,u'\}}^2 \leq a \leq ab$. Hence, our result still holds.

Conclusion: We have our result for $C = \frac{1}{2a + 8d_{max}^2 ab}$ and a laplacian weighted with local communication constraints: $\mu_{\{v,w\}}^2 = \frac{1}{2\ell_{\{v,w\}}} \times \min_{\{u,u'\} \sim \{v,w\}} \frac{\ell_{\{u,u'\}}}{\ell_{\{v,w\}}}$. The final rate thus depends on the smallest eigenvalue of the laplacian weighted by:

$$\frac{1}{2a + 8d_{max}^2 ab} \frac{1}{L_{max}} \frac{1}{2\ell_{\{v,w\}}} \times \min_{\{u,u'\} \sim \{v,w\}} \frac{\ell_{\{u,u'\}}}{\ell_{\{v,w\}}}. \quad (34)$$

This ends the proof of Theorem 6. \square

C.2 Adding stochasticity

We now prove the following result.

Theorem 7 (Adding Stochasticity). *Assume that, for all $k \in \mathbb{N}$, there exists a \mathcal{F}_{k+T-1} -measurable event A_k , such that $\mathbb{P}(A_k | \mathcal{F}_k) \geq \frac{1}{2}$ almost surely, and that under A_k , Assumption 7 holds for all $k \leq \ell \leq k + T - 1$. Then, we have the following bound on $\Lambda_k(\mathbf{x})$:*

$$\mathbb{E}[\Lambda_k(\mathbf{x})] \leq \left(\frac{1}{4}(1 - \gamma)^{T/3} + \frac{3}{4} \right)^{\lceil \frac{k}{2T} \rceil} \mathbb{E}[\Lambda_0],$$

where γ is defined in Theorem 6.

Proof. Using the same arguments as in the proof of Theorem 6, we obtain:

$$\mathbb{E}[\Lambda_{t+1} - \Lambda_t | \mathcal{F}_t, A_t] \leq -\sigma \Lambda_t. \quad (35)$$

However, this is not enough to conclude. Under A_t^C , we only know that $\Lambda_{t+1} \leq \Lambda_t$ (our local coordinate gradient steps cannot increase distance to the optimum). Hence:

$$\mathbb{E}[\Lambda_{t+1} | \mathcal{F}_t] \leq (1 - \sigma \mathbb{I}_{A_t}) \Lambda_t. \quad (36)$$

And then, by induction:

$$\mathbb{E}[\Lambda_t] \leq \mathbb{E}[P_t \Lambda_0], \text{ where } P_t = \prod_{s=0}^{t-1} (1 - \sigma \mathbb{I}_{A_s}). \quad (37)$$

However, no direct bound on P_t exists. The interdependencies on the events A_t make it impossible for an induction to prove a bound of the form $\leq (1 - \sigma/2)^t$. However, the logarithm of the product seems easier to study:

$$\log(P_t) = \log(1 - \sigma) \sum_{s=0}^{t-1} \mathbb{I}_{A_s}, \quad (38)$$

giving us $\mathbb{E} \log(P_t) \leq \log(1 - \sigma)t/2$, as $\mathbb{P}(A_t) \geq 1/2$. We are thus going to make a study in probability. For $t \in \mathbb{N}$, let $X_t = \frac{1}{T} \sum_{s=t}^{t+T-1} \mathbb{I}_{A_s}$. Using Markov-type inequalities conditionnaly on \mathcal{F}_t gives:

$$\mathbb{P}(X_t \geq 1/3 | \mathcal{F}_t) + 1/3 \mathbb{P}(X_t \leq 1/3 | \mathcal{F}_t) \geq \mathbb{E}[X_t | \mathcal{F}_t] \geq 1/2 \implies \mathbb{P}(X_t \geq 1/3 | \mathcal{F}_t) \geq 1/4. \quad (39)$$

Thus, we have: $\mathbb{E}[\prod_{s=t}^{t+T-1} (1 - \mathbb{I}_{A_s} \sigma) | \mathcal{F}_t] \leq \frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4}$. We then know how to control T consecutive factors of the product P_t . Skipping the next T terms, we have:

$$\mathbb{E} \left[\prod_{s=t}^{t+3T-1} (1 - \mathbb{I}_{A_s} \sigma) \right] = \mathbb{E} \left[\prod_{s=t}^{t+T-1} (1 - \mathbb{I}_{A_s} \sigma) \prod_{s=t+T}^{t+2T-1} (1 - \mathbb{I}_{A_s} \sigma) \prod_{s=t+2T}^{t+3T-1} (1 - \mathbb{I}_{A_s} \sigma) \right] \quad (40)$$

$$\leq \mathbb{E} \left[\prod_{s=t}^{t+T-1} (1 - \mathbb{I}_{A_s} \sigma) \prod_{s=t+2T}^{t+3T-1} (1 - \mathbb{I}_{A_s} \sigma) \right] \quad (41)$$

$$\leq \mathbb{E} \left[\prod_{s=t}^{t+T-1} (1 - \mathbb{I}_{A_s} \sigma) \mathbb{E}^{\mathcal{F}_{t+2T}} \left\{ \prod_{s=t+2T}^{t+3T-1} (1 - \mathbb{I}_{A_s} \sigma) \right\} \right] \quad (42)$$

as in the last right hand side, the first big product is \mathcal{F}_{t+2T} -measurable (our assumption on the A_s states that they are \mathcal{F}_{s+T-1} -measurable). Then, using inequality $\mathbb{E} \left[\prod_{s=t}^{t+T-1} (1 - \mathbb{I}_{A_s} \sigma) | \mathcal{F}_t \right] \leq \frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4}$ twice, with t and $t + 2T$, we get:

$$\begin{aligned} \mathbb{E} \left[\prod_{s=t}^{t+3T-1} (1 - \mathbb{I}_{A_s} \sigma) \right] &\leq \mathbb{E} \left[\prod_{s=t}^{t+T-1} (1 - \mathbb{I}_{A_s} \sigma) \left(\frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4} \right) \right] \\ &\leq \left(\frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4} \right)^2. \end{aligned}$$

Proceeding the same way by induction leads us to:

$$\mathbb{E}[P_t] \leq \left(\frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4} \right)^{\lfloor t/(2T) \rfloor}, \quad (43)$$

which is the desired bound. \square

From the proof, we thus have the following corollary.

Corollary 1. *Assume that, for all $k \in \mathbb{N}$, there exists a \mathcal{F}_{k+T-1} -measurable event A_k , such that $\mathbb{P}(A_k | \mathcal{F}_k) \geq \frac{1}{2}$ almost surely, and that under A_k , Assumption 7 holds for all $k \leq \ell \leq k + T - 1$. Then, we have the following bound on $\Lambda_k(\mathbf{x})$, for any $k \geq 0$:*

$$\mathbb{E}[\Lambda_{k+2T}(\mathbf{x}) | \mathcal{F}_k] \leq \left(\frac{1}{4}(1 - \gamma)^{T/3} + \frac{3}{4} \right) \mathbb{E}[\Lambda_k(\mathbf{x}) | \mathcal{F}_k].$$

where γ is defined in Theorem 6.

C.3 Expliciting the constants in the loss networks model we consider

We now need to compute and tune the constants introduced in Theorem 6 for the assumptions of Theorem 7 to hold in our Loss Network model. We begin by the following lemma, inspired by queuing theory arguments, that upper bound the probability that an edge stays inactivated for a long period of time.

Note that we here come back to continuous time, to study the loss network model. What is important to keep in mind is that an edge cannot be occupied for a time longer than $\tau'_{\{v,w\}}$.

Lemma 4. *Let $\delta \in (0, 1)$. For any $t_0 \geq 0$, $\{v, w\} \in E$, if the Poisson intensities are such that $p_{\{v,w\}} = \frac{1}{2 \max(d_i, d_j) - 1} (\tau'_{\{v,w\}})^{-1}$ and $\tau'_{\max}(\{v, w\}) = \max_{\{u, u'\} \sim \{v, w\}} \tau'_{\{u, u'\}}$, let:*

$$\ell_{\{v,w\}} = \frac{\log(\delta^{-1})}{\log(1 - (1 - e^{-1})e^{-1})} (p_{\{v,w\}}^{-1} + \tau'_{\max}(\{v, w\})).$$

We have:

$$\mathbb{P}(\{v, w\} \text{ not activated in } [t_0, t_0 + \ell_{\{v,w\}}] | \mathcal{F}_{t_0}) \leq \delta. \quad (44)$$

Proof of Lemma 4. Let $\{v, w\} \in E$ and $t_0 \geq 0$ fixed. We use tools from queuing theory [Tanner, 1995, $M/M/\infty/\infty$ queues] in order to compute the probability that edge $\{v, w\}$ is activable at a time t or not. More formally, we define a process $N_{\{v,w\}}(t)$ with values in \mathbb{N} , such that $N_{\{v,w\}}(t_0) = 1$ if $\{v, w\}$ non-available at time t_0 and 0 otherwise. Then, when an edge $\{u, u'\}$ such that $\{u, u'\} \sim \{v, w\}$ is activated, we make an increment of 1 on $N_{\{v,w\}}(t)$ (a *customer* arrives). This customer stays for a time $\tau'_{\{u,u'\}}$ and when he leaves, $N_{\{v,w\}}$ is decreased by 1. Thus $N_{\{v,w\}} \geq 0$ a.s., and if $N_{\{v,w\}} = 0$, then edge $\{v, w\}$ is available. For $t \geq \max_{\{u,u'\} \sim \{v,w\}} \tau'_{\{u,u'\}} + t_0$, $N_{\{v,w\}}(t)$ follows a Poisson law of parameter $\sum_{\{u,u'\} \sim \{v,w\}} p_{\{u,u'\}} \tau'_{\{u,u'\}}$. For any $t \geq \max_{\{u,u'\} \sim \{v,w\}} \tau'_{\{u,u'\}} + t_0$:

$$\mathbb{P}(\{v, w\} \text{ available at time } t | \mathcal{F}_{t_0}) \geq \mathbb{P}(N_i(t) = 0) = \exp\left(- \sum_{\{u,u'\} \sim \{v,w\}} p_{\{u,u'\}} \tau'_{\{u,u'\}}\right).$$

That leads to taking $p_{\{u,u'\}} = \frac{1}{2} \frac{1}{\max(d_k, d_l) - 1} (\tau'_{\{u,u'\}})^{-1}$ for all edges, in order to have

$$\mathbb{P}(\{v, w\} \text{ available at time } t | \mathcal{F}_{t_0}) \geq 1/e.$$

Then, $\mathbb{P}(\{v, w\} \text{ rings in } [t, t + p_{\{v,w\}}^{-1}]) = 1 - e^{-1}$, giving:

$$\begin{aligned} & \mathbb{P}(\{v, w\} \text{ activated in } [t_0, t_0 + \tau'_{\max}(\{v, w\}) + p_{\{v,w\}}^{-1}] | \mathcal{F}_{t_0}) = \mathbb{P}(\{v, w\} \text{ rings in } [t, t + p_{\{v,w\}}^{-1}]) \\ & \times \mathbb{P}(\{v, w\} \text{ available at time } t | \mathcal{F}_{t_0}, \{v, w\} \text{ rings at a time } t \in [t_0 + \tau'_{\max}(\{v, w\}), t_0 + \tau'_{\max}(\{v, w\}) + p_{\{v,w\}}^{-1}]) \\ & \geq (1 - e^{-1})e^{-1}, \end{aligned}$$

where we use the memoriless property of exponential random variables. Take $k \in \mathbb{N}$ such that $(1 - (1 - e^{-1})e^{-1})^k \leq \delta$, leading to $k = \log(6|E|) / \log(1 - (1 - e^{-1})e^{-1})$. Let

$$\ell_{\{v,w\}} = k(p_{\{v,w\}}^{-1} + \tau'_{\max}(\{v, w\})).$$

Then we have a.s.:

$$\mathbb{P}(\{v, w\} \text{ not activated in } [t_0, t_0 + \ell_{\{v,w\}}] | \mathcal{F}_{t_0}) \leq \delta. \quad (45)$$

□

Let $t \in \mathbb{N}$ be fixed, and B_t be the event: "in the activations $t, t+1, \dots, t+T-1$, all edges are activated". Let then $C_t(\{v, w\}, s)$ for $t \leq s < t+T$ be the event $\min(T_{\{v,w\}}(s), t+T-s, s-t) \leq a\ell_{\{v,w\}}$ and $D_t(\{u, u'\}, \{v, w\}, s)$ be the event $N(\{u, u'\}, \{v, w\}, s) \leq \lceil b\ell_{\{v,w\}}/\ell_{\{u,u'\}} \rceil$, where $N(\{u, u'\}, \{v, w\}, s)$ is the number of activations of $\{u, u'\}$ between two activations of $\{v, w\}$, around time s , where we only take into account the activations between activations t and $t+T-1$. Let then $A_t = B_t \cap (\cap_{\{u,u'\}, \{v,w\} \in E, t \leq s < t+T} C_t(\{v, w\}, s) \cap D_t(\{u, u'\}, \{v, w\}, s))$.

We want $\mathbb{P}(A_t) \geq 1/2$ for correct constants a, b, T and $\ell_{\{v,w\}}$ (that can differ from $\tau'_{\{v,w\}}$) in order to apply Theorems 6 and 7. Note that this event is \mathcal{F}_{t+T-1} -measurable, as desired. We first study the length of time $\ell_{\{v,w\}}$ edge $\{v, w\}$ must wait in order to be activated with high probability (*high* meaning more than $1 - \frac{1}{12|E|}$). This result is Lemma 4. Then, we use this length to determine the constants $T, a, b, \ell_{\{v,w\}}$ needed.

Lemma 5. For any continuous time $t_0 \geq 0$, $\{v, w\} \in \mathcal{E}$, if $p_{\{v,w\}} = \frac{1}{2\max(d_i, d_j) - 1} (\tau'_{\{v,w\}})^{-1}$ and $\tau'_{\max}(\{v, w\}) = \max_{\{u,u'\} \sim \{v,w\}} \tau'_{\{u,u'\}}$, let $\ell_{\{v,w\}} = \frac{\log(6|E|)}{\log(1 - (1 - e^{-1})e^{-1})} (p_{\{v,w\}}^{-1} + \tau'_{\max}(\{v, w\}))$. We have, almost surely:

$$\mathbb{P}(\{v, w\} \text{ not activated in } [t_0, t_0 + \ell_{\{v,w\}}] | \mathcal{F}_{t_0}) \leq \frac{1}{6|E|}. \quad (46)$$

Proof of Lemma 4. Let $\{v, w\} \in E$ and $t_0 \geq 0$ fixed. We use tools from queuing theory [Tanner, 1995] ($M/M/\infty/\infty$ queues) in order to compute the probability that edge $\{v, w\}$ is activable at a time t or not. More formally, we define a process $N_{\{v,w\}}(t)$ with values in \mathbb{N} , such that $N_{\{v,w\}}(t_0) = 1$ if $\{v, w\}$ non-available at time t_0 and 0 otherwise. Then, when an edge $\{u, u'\}$, $\{u, u'\} \sim \{v, w\}$ is activated, we make an increment of 1 on $N_{\{v,w\}}(t)$ (a *customer* arrives). This customer stays for a time $\tau'_{\{u,u'\}}$ and when he leaves we make $N_{\{v,w\}}$

decrease by 1. We have $N_{\{v,w\}} \geq 0$ a.s., and if $N_{\{v,w\}} = 0$, $\{v, w\}$ is available. For $t \geq \max_{\{u,u'\} \sim \{v,w\}} \tau'_{\{u,u'\}} + t_0$, $N_{\{v,w\}}(t)$ follows a Poisson law of parameter $\sum_{\{u,u'\} \sim \{v,w\}} p_{\{u,u'\}} \tau'_{\{u,u'\}}$. For any $t \geq \max_{\{u,u'\} \sim \{v,w\}} \tau'_{\{u,u'\}} + t_0$:

$$\mathbb{P}(\{v, w\} \text{ available at time } t | \mathcal{F}_{t_0}) \geq \mathbb{P}(N_i(t) = 0) = \exp\left(- \sum_{\{u,u'\} \sim \{v,w\}} p_{\{u,u'\}} \tau'_{\{u,u'\}}\right). \quad (47)$$

That leads to taking $p_{\{u,u'\}} = \frac{1}{2 \max(d_k, d_l) - 1} (\tau'_{\{u,u'\}})^{-1}$ for all edges, in order to have $\mathbb{P}(\{v, w\} \text{ available at time } t | \mathcal{F}_{t_0}) \geq 1/e$. Then, $\mathbb{P}(\{v, w\} \text{ rings in } [t, t + p_{\{v,w\}}^{-1}]) = 1 - e^{-1}$, giving:

$$\mathbb{P}(\{v, w\} \text{ activated in } [t_0, t_0 + \tau'_{\max}(\{v, w\}) + p_{\{v,w\}}^{-1}] | \mathcal{F}_{t_0}) = \mathbb{P}(\{v, w\} \text{ rings in } [t, t + p_{\{v,w\}}^{-1}]) \quad (48)$$

$$\times \mathbb{P}(\{v, w\} \text{ available at time } t | \mathcal{F}_{t_0}, \{v, w\} \text{ rings at a time} \quad (49)$$

$$t \in [t_0 + \tau'_{\max}(\{v, w\}), t_0 + \tau'_{\max}(\{v, w\}) + p_{\{v,w\}}^{-1}]) \quad (50)$$

$$\geq (1 - e^{-1})e^{-1}, \quad (51)$$

where we use the fact that exponential random variables have no memory. Take $k \in \mathbb{N}$ such that $(1 - (1 - e^{-1})e^{-1})^k \leq \frac{1}{6|E|}$, leading to $k \approx \log(6|E|) / \log(1 - (1 - e^{-1})e^{-1})$. Let $\ell_{\{v,w\}} = k(p_{\{v,w\}}^{-1} + \tau'_{\max}(\{v, w\}))$. Then we have a.s.:

$$\mathbb{P}(\{v, w\} \text{ not activated in } [t_0, t_0 + \ell_{\{v,w\}}] | \mathcal{F}_{t_0}) \leq \frac{1}{6|E|}. \quad (52)$$

□

Bounding T : A direct application of Lemma 4 leads, with $L = \max_{\{v,w\}} \ell_{\{v,w\}}$, to:

$$T = 2 \sum_{\{v,w\}} \frac{L}{\tau'_{\{v,w\}}}. \quad (53)$$

Indeed, for all $\{v, w\}$, not being activated in activations $t, t + 1, \dots, t + T - 1$ means not being activated for a continuous interval of time of length more than $\ell_{\{v,w\}}$. Hence:

$$\mathbb{P}(\exists (\{v, w\}) \in E : (\{v, w\}) \text{ not activated in } \{t, \dots, t + T - 1\} | \mathcal{F}_t) \quad (54)$$

$$\leq \sum_{\{v,w\} \in E} \mathbb{P}((\{v, w\}) \text{ not activated in } \{t, \dots, t + T - 1\} | \mathcal{F}_t) \quad (55)$$

$$\leq \sum_{\{v,w\} \in E} \mathbb{P}((\{v, w\}) \text{ not activated in } [t, t + \ell_{\{v,w\}}] | \mathcal{F}_t) \quad (56)$$

$$\leq |E| \times \frac{1}{6|E|} \quad (57)$$

$$= 1/6. \quad (58)$$

Bounding $T_{\{v,w\}}$: Applying Lemma 4 with $12|E|T$ instead of $6|E|$ leads to controlling all the inactivation lengths by a length $\ell'_{\{v,w\}}$, with a probability more than $1 - 1/(12|E|T)$. Let $\{v, w\} \in E$ and $s \in \mathbb{N}$, $t \leq s < t + T$. Let $\alpha > 0$ to tune later. Denote by $\delta_{\{v,w\}}(s)$ the (random) inactivation time of $\{v, w\}$, around iteration s . Note that conditionnaly on the inactivation period $\delta_{\{v,w\}}(s)$, $T_{\{v,w\}}(s)$ is dominated in law by a Poisson variable of parameter $I\delta_{\{v,w\}}(s)$, hence line (60):

$$\mathbb{P}(T_{\{v,w\}}(s) \geq \alpha \ell'_{\{v,w\}} | \mathcal{F}_t) \leq \mathbb{P}(T_{\{v,w\}}(s) \geq \alpha \ell'_{\{v,w\}} | \mathcal{F}_t, \delta_{\{v,w\}} \leq \ell'_{\{v,w\}}) \times \mathbb{P}(\delta_{\{v,w\}} \leq \ell'_{\{v,w\}}) + \mathbb{P}(\delta_{\{v,w\}} \geq \ell'_{\{v,w\}}) \quad (59)$$

$$\leq \mathbb{P}(\text{Poisson}(I\ell'_{\{v,w\}}) \geq \alpha \ell'_{\{v,w\}}) + \frac{1}{12|E|T} \quad (\text{where } I = \sum_{\{v,w\} \in E} p_{\{v,w\}}) \quad (60)$$

$$\leq \frac{1}{12|E|T} + \frac{1}{12|E|T} \quad (61)$$

$$= \frac{1}{6|E|T}, \quad (62)$$

for some $\alpha > 0$ big enough, to determine with the following large deviation inequality:

Lemma 6 (A Large Deviation Inequality on discrete Poisson variables.). *Let $Z \sim \text{Poisson}(\lambda)$, for some $\lambda > 0$. Then, for all $u \geq 0$:*

$$\mathbb{P}(Z \geq u) \leq \exp(-u + \lambda(e - 1)). \quad (63)$$

This large deviation leads to taking $\alpha = 2eI$ for (61) to be true. Finally, we get:

$$\mathbb{P}(T_{\{v,w\}}(s) \geq \alpha \ell'_{\{v,w\}} | \mathcal{F}_t) \leq \frac{1}{6|E|T}. \quad (64)$$

Bounding $N(\{u, u'\}, \{v, w\}, s)$: If $\delta_{\{v,w\}}(s) \leq \ell'_{\{v,w\}}$, this random variable is dominated by a Poisson variable of parameter $p_{\{u,u'\}} \ell'_{\{v,w\}}$. Hence, still with Lemma 6, with probability more than $1 - \frac{1}{12|E|^{2T}}$, we can bound $N(\{u, u'\}, \{v, w\})$ by $e \log(12|E|^{2T}) + p_{\{u,u'\}} \ell'_{\{v,w\}}(e - 1) \leq 2ep_{\{u,u'\}} L_{\{v,w\}}$.

Explicit writing of the union bound on A_t^C : $A_t^C = B_t^C \cup (\cup_{\{u,u'\}, \{v,w\} \in E, t \leq s < t+T} C_t(\{v, w\}, s)^C \cup D_t(\{u, u'\}, \{v, w\}, s)^C) \in \mathcal{F}_{t+T-1}$. Thanks to the previous considerations, we have that $\mathbb{P}^{\mathcal{F}^t}(B_t^C) \leq 1/6$ with (58), $\mathbb{P}^{\mathcal{F}^t}(C_t(\{v, w\}, s)^C) \leq \frac{1}{6|E|T}$ with (64) and $\mathbb{P}(D_t(\{u, u'\}, \{v, w\}, s)^C | \mathcal{F}_t) \leq \frac{1}{6|E|^{2T}}$, for the following constants and weights:

- $\tilde{\tau}'_{\{v,w\}}^{-1} = p_{\{v,w\}} = \min(\frac{1}{\tau'_{\max(\{v,w\})}}, \frac{1}{2(\max(d_i, d_j) - 1)} \frac{1}{\tau'_{\{v,w\}}})$;
- $T = 2I \max_{\{v,w\} \in E} \tau'_{\{v,w\}} \frac{\log(6|E|)}{\log(1 - (1 - e^{-1})e^{-1})}$;
- $a = 2eI \frac{\log(6|E|T)}{\log(1 - (1 - e^{-1})e^{-1})}$;
- $b = 2e \frac{\log(6|E|T)}{\log(1 - (1 - e^{-1})e^{-1})}$.

The union bound is the following:

$$\mathbb{P}^{\mathcal{F}^t}(A_t^C) \leq \mathbb{P}^{\mathcal{F}^t}(B_t^C) + \sum_{s, \{v,w\}} \mathbb{P}^{\mathcal{F}^t}(C_t(\{v, w\}, s)^C) + \sum_{s, \{v,w\}} \mathbb{P}^{\mathcal{F}^t}(\cup_{\{u,u'\}} D_t(\{u, u'\}, \{v, w\}, s)^C) \quad (65)$$

$$\leq 1/6 + |E|T/(6|E|T) \times 2 \quad (66)$$

$$\leq 1/2. \quad (67)$$

The rate of convergence γ is then defined as the smallest non null eigenvalue of the laplacian of the graph, weighted by:

$$\nu_{\{v,w\}} = \frac{p_{\{v,w\}} \min_{\{u,u'\} \sim \{v,w\}} \frac{\tau'_{\{v,w\}}}{\tau'_{\{u,u'\}}}}{8a(1 + d^2b)} = \frac{\min_{\{u,u'\} \sim \{v,w\}} p_{\{u,u'\}}}{c_1 \ln(6|E|T)(1 + d^2 \ln(6|E|T)^2) \sum_{\{u,u'\} \in \mathcal{E}} p_{\{u,u'\}}} \quad (68)$$

C.4 Concluding

What we have proved so far, is that for any $k \geq 0$, any $\mathbf{x} \in \mathbb{R}^V$, we have:

$$\mathbb{E}[\Lambda_{k+2T}(\mathbf{x}) | \mathcal{F}_k] \leq \left(\frac{1}{4}(1 - \gamma)^{T/3} + \frac{3}{4} \right) \mathbb{E}[\Lambda_k(\mathbf{x}) | \mathcal{F}_k],$$

where γ is defined in Equation (68). Then, $\Lambda_{k+2T}(\mathbf{x}) \geq \frac{1}{2} \|W^{(0,k+2T)}(\mathbf{x} - \bar{\mathbf{x}})\|^2$ and $\Lambda_k(\mathbf{x}) \leq \frac{1}{2} \|W^{(0,k)}(\mathbf{x} - \bar{\mathbf{x}})\|^2$, so that applying this for $k = 0$, almost surely conditioned on \mathcal{F}_0 ,

$$\mathbb{E} \left[\left\| W^{(0,2T)}(\mathbf{x} - \bar{\mathbf{x}}) \right\|^2 \middle| \mathcal{F}_0 \right] \leq \left(\frac{1}{4}(1 - \gamma)^{T/3} + \frac{3}{4} \right) \mathbb{E}[\|\mathbf{x} - \bar{\mathbf{x}}\|^2 | \mathcal{F}_0],$$

Now, noticing that our analysis holds almost surely for any configuration \mathcal{F}_0 , doing a time translation and starting from a configuration \mathcal{F}_k for any k , we get that:

$$\mathbb{E} \left[\left\| W^{(k, k+2T)}(\mathbf{x} - \bar{\mathbf{x}}) \right\|^2 \middle| \mathcal{F}_k \right] \leq \left(\frac{1}{4}(1-\gamma)^{T/3} + \frac{3}{4} \right) \mathbb{E}[\|\mathbf{x} - \bar{\mathbf{x}}\|^2 | \mathcal{F}_k],$$

so that Assumption 2 holds for $\rho = \frac{1}{4}(1 - (1 - \gamma)^{T/3})$, $k_\rho = 2T$, and hence $\frac{\rho}{k_\rho} = \mathcal{O}(\gamma)$, which leads to Theorem 5: γ is the eigengap of the graph, with weights of order $\tilde{\mathcal{O}}\left(\frac{\min_{\{u, u'\} \sim \{v, w\}} p_{\{u, u'\}}}{d^2 \sum_{\{u, u'\} \in \mathcal{E}} p_{\{u, u'\}}}\right)$.

D Proof of Theorem 1: Convex-Lipchitz case

D.1 Homogeneous setting, Lipschitz (bounded gradients) and convex without sampling

Proof. Studying the virtual sequence, we expand:

$$\begin{aligned} \mathbb{E} \left[\|\hat{x}^{k+1} - x^*\|^2 \right] &= \mathbb{E} \left[\|\hat{x}^k - x^*\|^2 - \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \hat{x}^k - x^* \rangle + \frac{\gamma^2}{n^2} \left\| \sum_{v \in \mathcal{I}_k} g_v^k \right\|^2 \right] \\ &\leq \mathbb{E} \left[\|\hat{x}^k - x^*\|^2 - \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - x^* \rangle + \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - \bar{x}^k \rangle \right. \\ &\quad \left. + \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \bar{x}^k - \hat{x}^k \rangle \right] \\ &\quad + \frac{\gamma^2 B^2 |\mathcal{I}_k|^2}{n^2}, \end{aligned}$$

where we used the Lipschitz assumption, $\mathbb{E} g_v^k = \nabla f_v(x_v^k)$ and boundness of gradients. Denote:

$$\begin{aligned} T_1 &= -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - x^* \rangle \\ T_2^k &= \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - \bar{x}^k \rangle \\ T_3 &= \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \bar{x}^k - \hat{x}^k \rangle. \end{aligned}$$

Using convexity of f ,

$$T_1 \leq -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)).$$

Using the Lipschitz assumption and Equation (9) that controls $\|\bar{x}^k - \hat{x}^k\|$, we bound T_3 ::

$$T_3 \leq \frac{2\gamma^2 B^2 |\mathcal{I}_k|}{n}.$$

Using the Lipschitz assumption and our consensus bound from Equation (13), we bound T_2^k :

$$\begin{aligned} \sum_{k < K} T_2^k &\leq \sum_{k < K} \frac{2\gamma B}{n} \sqrt{\sum_{v \in \mathcal{I}_k} \mathbb{E} [\|x_v^k - \bar{x}^k\|^2]} \\ &\leq \sum_{k < K} \frac{2\gamma B}{n} \sqrt{\mathbb{E} [\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2]} \\ &\leq \sum_{k < K} \frac{\gamma^2 B^2}{n\bar{\rho}} + \frac{\bar{\rho}}{B} \mathbb{E} [\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2] \\ &\leq \frac{3\gamma^2 B^2}{n\bar{\rho}} \sum_{k < K} |\mathcal{I}_k|. \end{aligned}$$

Consequently, denoting $\eta = \frac{\gamma}{n}$ and summing over $k < K$,

$$\begin{aligned} 2\eta \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} [f(x_v^k) - f(x^*)] &\leq \mathbb{E} [\|\hat{x}^0 - x^*\|^2] + \eta^2 B^2 (\mathbb{E} |\mathcal{I}_k| + 2n + 3n\bar{\rho}^{-1}) \sum_{k < K} |\mathcal{I}_k| \\ &\leq \mathbb{E} [\|\hat{x}^0 - x^*\|^2] + \eta^2 B^2 (3n + 3n\bar{\rho}^{-1}) \sum_{k < K} |\mathcal{I}_k|. \end{aligned}$$

Dividing by $2\eta \sum_{k < K} |\mathcal{I}_k|$,

$$\mathbb{E} \left[\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k < K} \sum_{v \in \mathcal{I}_k} f(x_v^k) - f(x_*) \right] \leq \frac{\mathbb{E} [\|\hat{x}^0 - x^*\|^2]}{2\eta \sum_{k < K} |\mathcal{I}_k|} + \frac{\eta B^2}{2} (3n + 3n\bar{\rho}^{-1}),$$

and

$$\begin{aligned} \mathbb{E} [\|\hat{x}^0 - x^*\|^2] &\leq \|x^0 - x^*\|^2 - 2\eta \sum_{v \in \mathcal{V}} \langle \nabla f(x^0), x^0 - x^* \rangle + \eta^2 G^2/n \\ &\leq \|x^0 - x^*\|^2 + \eta^2 B^2/K, \end{aligned}$$

provided that $K \geq n$. Optimizing over η , we obtain that for $\eta = \sqrt{\frac{D^2}{2KB^2(3n+2n\bar{\rho}^{-1})}}$,

$$\mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x_*) \right] \leq 2\sqrt{\frac{2B^2 D^2 (3n + 2n\bar{\rho}^{-1})}{\sum_{k < K} |\mathcal{I}_k|}}.$$

□

D.2 Lipschitz (bounded gradients) and convex with sampling

Proof. Taking the proof just above, we still have

$$\mathbb{E} [\|\hat{x}^{k+1} - x^*\|^2] \leq \mathbb{E} [\|\hat{x}^k - x^*\|^2 + T_1^k + T_2^k + T_3] + \frac{\gamma^2 B^2 |\mathcal{I}_k|^2}{n^2}.$$

We have, using convexity and then Lipschitzness:

$$\begin{aligned} T_1^k &= -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - x^* \rangle \\ &\leq -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} p_v f_v(x_v^k) - f(x^*) \\ &= -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} p_v f_v(\bar{x}^k) - f(x^*) + f_v(x_v^k) - f(\bar{x}^k) \\ &\leq -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} f_v(\bar{x}^k) - f(x^*) - B \|x_v^k - \bar{x}^k\|, \end{aligned}$$

so that

$$\begin{aligned} \mathbb{E} [T_1^k] &\leq -\frac{2\gamma\bar{p}}{n} (\mathbb{E} f(\bar{x}^k) - f(x^*)) + \frac{2\gamma B}{n} \sum_{v \in \mathcal{V}} p_v \|x_v^k - \bar{x}^k\| \\ &\leq -\frac{2\gamma\bar{p}}{n} (\mathbb{E} f(\bar{x}^k) - f(x^*)) + \frac{2\gamma B p_{\max}}{n} \sqrt{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|. \end{aligned}$$

We then have that:

$$\sum_{k < K} \frac{2\gamma B p_{\max}}{n} \sqrt{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\| \leq 2\sqrt{2} \frac{\gamma^2 B^2 p_{\max}}{\sqrt{n}} \sqrt{K \sum_{k < K} |\mathcal{I}_k|}.$$

Then,

$$T_3 \leq \frac{2\gamma^2 B^2}{n}.$$

We handle the consensus term differently. For some $\alpha > 0$ to be fix later, and taking the expectation conditionally on \mathbf{x}^k ,

$$\begin{aligned} \sum_{k < K} \mathbb{E} [T_2^k] &\leq \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\frac{\gamma^2}{n\alpha} \|\nabla f(x_v^k)\|^2 + \frac{\alpha}{n} \|x_v^k - \bar{x}^k\|^2 \right] \\ &\leq \sum_{k < K} \frac{\gamma^2 B^2 |\mathcal{I}_k|}{\alpha n} + \frac{\alpha}{n} \sum_{v \in \mathcal{V}} p_v \|x_v^k - \bar{x}^k\|^2 \\ &\leq \frac{\gamma^2 B^2}{\alpha n} \sum_{k < K} |\mathcal{I}_k| + \frac{\alpha p_{\max}}{n} \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\ &\leq \left(\frac{\gamma^2 B^2}{\alpha n} + \frac{\alpha p_{\max}}{n} \frac{2\gamma^2 B^2}{\bar{\rho}^2} \right) \sum_{k < K} |\mathcal{I}_k|. \end{aligned}$$

We set $\alpha = 1/\sqrt{p_{\max}\bar{\rho}^{-2}}$, so that:

$$\sum_{k < K} \mathbb{E} [T_2^k] \leq 2 \frac{\gamma^2 B^2}{n^2} \times \sqrt{p_{\max} n \bar{\rho}^{-1}} \times \sum_{k < K} |\mathcal{I}_k|.$$

The rest of the proof then follows as before, and we obtain

$$\mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] = \mathcal{O} \left(\sqrt{\frac{B^2 D^2}{\sum_{k < K} |\mathcal{I}_k|} (n + (p_{\max})^{1/2} n \bar{\rho}^{-1} + n^{3/2} p_{\max} \sqrt{\frac{K}{\sum_{k < K} |\mathcal{I}_k|}})} \right).$$

To conclude, we notice that $\frac{nK}{\sum_{k < K} |\mathcal{I}_k|}$ is of order $1/\bar{p}$ where $\bar{p} = \frac{1}{n} \sum_{v \in \mathcal{V}}$. \square

E Proof of Theorem 2: smooth-Lipschitz-convex rates

E.1 Smooth-Lipschitz-convex rates without sampling, homogeneous case

Proof. As before, we have:

$$\mathbb{E} \left[\|\hat{x}^{k+1} - x^*\|^2 \right] \leq \mathbb{E} \left[\|\hat{x}^k - x^*\|^2 + T_1 + T_2^k + T_3 \right] + \frac{\gamma^2 \sigma^2 |\mathcal{I}_k| + \gamma^2 \mathbb{E} \left[\sum_{v \in \mathcal{I}_k} \|\nabla f_v(x_v^k)\|^2 \right]}{n^2},$$

with

$$\begin{aligned} T_1 &= -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - x^* \rangle \\ T_2^k &= \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - \bar{x}^k \rangle \\ T_3 &= \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \bar{x}^k - \hat{x}^k \rangle. \end{aligned}$$

First, using convexity of $f_v \equiv f$,

$$T_1 \leq -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} (f_v(x_v^k) - f_v(x^*)) = -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)).$$

Using Assumption 6 we have, where $C > 0$ can be arbitrary:

$$\begin{aligned} \mathbb{E} [T_2^k] &\leq \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} [\|\nabla f(x_v^k)\| \|x_v^k - \bar{x}^k\|] \\ &\leq \frac{C\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} [\|\nabla f(x_v^k)\|^2] + \frac{\gamma}{Cn} \mathbb{E} \left[\sum_{v \in \mathcal{I}_k} \|x_{v_k}^k - \bar{x}^k\|^2 \right] \\ &\leq \frac{2LC\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} [(f(x_v^k) - f(x^*)))] + \frac{\gamma}{Cn} \mathbb{E} [\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2]. \end{aligned}$$

We also have:

$$\begin{aligned} T_3 &\leq \frac{\gamma}{n} \left(C \sum_{v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 + \frac{1}{C} \|\bar{x}^k - \hat{x}^k\|^2 \right) \\ &\leq \frac{\gamma}{n} \left(2LC \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)) + \frac{\gamma^2 B^2}{C} |\mathcal{I}_k| \right). \end{aligned}$$

Thus,

$$\begin{aligned} \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) &\leq -\mathbb{E} [\|\hat{x}^{k+1} - x^*\|^2] + \mathbb{E} [\|\hat{x}^k - x^*\|^2] + \frac{\gamma^2 \sigma^2 |\mathcal{I}_k|}{n^2} + \frac{2\gamma^2 L |\mathcal{I}_k|}{n^2} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) \\ &\quad + \frac{2LC\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} [(f(x_v^k) - f(x^*)))] + \frac{\gamma}{Cn} \mathbb{E} [\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2] \\ &\quad + \frac{\gamma}{n} \left(2LC \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)) + \frac{\gamma^2 B^2}{C} |\mathcal{I}_k| \right). \end{aligned}$$

Summing over $k < K$ and using Lemma 2, we obtain:

$$\begin{aligned} \frac{2\gamma}{n} \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) &\leq \mathbb{E} [\|\hat{x}^0 - x^*\|^2] + \frac{\gamma^2 \sigma^2}{n^2} \sum_{k < K} |\mathcal{I}_k| + \frac{2\gamma L}{n} (2C + \gamma) \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) \\ &\quad + \sum_{k < K} \frac{\gamma}{Cn} \mathbb{E} [\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2] + \frac{\gamma^3 B^2}{Cn} \sum_{k < K} |\mathcal{I}_k| \\ &\leq \mathbb{E} [\|\hat{x}^0 - x^*\|^2] + \frac{\gamma^2 \sigma^2}{n^2} \left(1 + \frac{2\gamma \bar{\rho}^{-1} n}{C} \right) \sum_{k < K} |\mathcal{I}_k| + \sum_{k < K} \frac{\gamma}{Cn} \mathbb{E} [\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2] \\ &\quad + \frac{\gamma^3 B^2}{Cn} \sum_{k < K} |\mathcal{I}_k| + \frac{2\gamma L}{n} \left(2C + \gamma + \frac{4\gamma^2}{\bar{\rho}^2} \right) \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)). \end{aligned}$$

Hence, provided that $2C + \gamma + \frac{4\gamma^2}{\bar{\rho}^2} \leq \frac{1}{2L}$, which is verified for $C = \frac{1}{8L}$ and $\gamma \leq \frac{1}{4L} \times \frac{1}{1+2\bar{\rho}^{-1}}$, we have:

$$\frac{\gamma}{n} \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) \leq \mathbb{E} [\|\hat{x}^0 - x^*\|^2] + \frac{\gamma^2 \sigma^2}{n^2} (1 + 16L\gamma\bar{\rho}^{-1}n) \sum_{k < K} |\mathcal{I}_k| + \frac{8L\gamma^3 B^2}{n} \sum_{k < K} |\mathcal{I}_k|,$$

leading to, for $\eta = \gamma/n$:

$$\mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] \leq \frac{\mathbb{E} [\|\hat{x}^0 - x^*\|^2]}{\eta \sum_{k < K} |\mathcal{I}_k|} + \eta \sigma^2 + \eta^2 (16L\sigma^2 n^2 \bar{\rho}^{-1} + 8LB^2 n^2).$$

Optimizing over $\eta \leq \frac{1}{4L} \times \frac{1}{n(1+2\bar{\rho}^{-1})}$, we thus obtain that:

$$\begin{aligned} &\mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] \\ &= \mathcal{O} \left(\frac{LD^2 n \bar{\rho}^{-1}}{\sum_{k < K} |\mathcal{I}_k|} + \sqrt{\frac{D\sigma^2}{\sum_{k < K} |\mathcal{I}_k|}} + \left[\frac{D^2 \sqrt{LB^2 n^2 + L\sigma^2 n^2 \bar{\rho}^{-1}}}{\sum_{k < K} |\mathcal{I}_k|} \right]^{2/3} \right). \end{aligned}$$

□

E.2 Smooth-Lipschitz-convex rates with sampling, heterogeneous case

Proof. We have:

$$\mathbb{E} \left[\|\hat{x}^{k+1} - x^*\|^2 \right] \leq \mathbb{E} \left[\|\hat{x}^k - x^*\|^2 - \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \hat{x}^k - x^* \rangle \right] + \frac{\gamma^2 \sigma^2 |\mathcal{I}_k| + \gamma^2 \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2}{n^2},$$

and we will handle the middle term differently than before. Using $-\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \hat{x}^k - x^* \rangle = -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - x^* \rangle - \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \hat{x}^k - x_v^k \rangle$ and then convexity for the first term and smoothness for the second, we obtain:

$$\begin{aligned} -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \hat{x}^k - x^* \rangle &\leq -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \left(f_v(x_v^k) - f_v(x^*) - \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} f_v(\hat{x}^k) - f_v(x_v^k) - \frac{L}{2} \|x_v^k - \hat{x}^k\|^2 \right) \\ &= -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} f_v(\hat{x}^k) - f_v(x^*) + \frac{\gamma L}{n} \sum_{v \in \mathcal{I}_k} \|x_v^k - \hat{x}^k\|^2. \end{aligned}$$

Taking the expectation wrt \mathcal{I}_k :

$$\begin{aligned} \mathbb{E} \left[-\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \hat{x}^k - x^* \rangle \right] &\leq -\frac{2\gamma}{n} \sum_{v \in \mathcal{V}} p_v (f_v(\hat{x}^k) - f_v(x^*)) + \frac{\gamma L}{n} \sum_{v \in \mathcal{V}} p_v \|x_v^k - \hat{x}^k\|^2 \\ &\leq -\frac{2\gamma n \bar{p}}{n} (f(\hat{x}^k) - f(x^*)) + \frac{2\gamma L p_{\max}}{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \frac{\gamma L}{n} \sum_{v \in \mathcal{V}} p_v \|\hat{x}^k - \bar{x}^k\|^2 \\ &\leq -\frac{2\gamma n \bar{p}}{n} (f(\hat{x}^k) - f(x^*)) + \frac{2\gamma L p_{\max}}{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2\gamma L \bar{p} \|\hat{x}^k - \bar{x}^k\|^2. \end{aligned}$$

Then, for the variance term, we need to bound $\mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2$. For any $(z_v)_{v \in \mathcal{V}}$, we have $\mathbb{E} \left[\left\| \sum_{v \in \mathcal{I}_k} z_v \right\|^2 \right] = \mathbb{E} \left[\sum_{v, v' \in \mathcal{V}} \mathbf{1}_{v \in \mathcal{I}_k} \mathbf{1}_{v' \in \mathcal{I}_k} \langle z_v, z_{v'} \rangle \right] = \sum_{v \neq v' \in \mathcal{V}} \mathbf{1}_{v \in \mathcal{V}} p_v p_{v'} \langle z_v, z_{v'} \rangle + \sum_{v \in \mathcal{V}} p_v \|z_v\|^2 \leq \sum_{v \in \mathcal{V}} p_v \|z_v\|^2 + \left\| \sum_{v \in \mathcal{V}} p_v z_v \right\|^2$. And finally, using convexity of the squared norm, $\left\| \sum_{v \in \mathcal{V}} p_v z_v \right\|^2 \leq n \bar{p} \sum_{v \in \mathcal{V}} p_v \|z_v\|^2$. Hence, we have

$$\mathbb{E}_{\mathcal{I}_k} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2 \leq \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(x_v^k)\|^2 + \left\| \sum_{v \in \mathcal{V}} p_v \nabla f_v(x_v^k) \right\|^2.$$

Thus, plugging this in the first inequality,

$$\begin{aligned} \frac{2\gamma n \bar{p}}{n} (\mathbb{E} f(\hat{x}^k) - f(x^*)) &\leq \mathbb{E} \left[\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^*\|^2 \right] + \mathbb{E} \left[\frac{\gamma^2 \sigma^2 |\mathcal{I}_k| + \gamma^2 \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2}{n^2} \right] \\ &\quad + \mathbb{E} \left[\frac{2\gamma L p_{\max}}{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2\gamma L \bar{p} \|\hat{x}^k - \bar{x}^k\|^2 \right] \\ &= \mathbb{E} \left[\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^*\|^2 \right] \\ &\quad + \frac{\gamma^2 \sigma^2 n \bar{p} + \gamma^2 \sum_{v \in \mathcal{V}} p_v \mathbb{E} \|\nabla f_v(x_v^k)\|^2 + \gamma^2 \left\| \sum_{v \in \mathcal{V}} p_v \nabla f_v(x_v^k) \right\|^2}{n^2} \\ &\quad + \mathbb{E} \left[\frac{2\gamma L p_{\max}}{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2\gamma L \bar{p} \|\hat{x}^k - \bar{x}^k\|^2 \right]. \end{aligned}$$

Then, using smoothness, we have that $f(\hat{x}^k) - f(x^*) \leq f(\hat{x}^k) - f(x^*) + \langle \nabla f(\hat{x}^k), \hat{x}^k - \bar{x}^k \rangle + \frac{L}{2} \|\bar{x}^k - \hat{x}^k\| \leq 2(f(\hat{x}^k) - f(x^*)) + 2L\|\bar{x}^k - \hat{x}^k\|^2$, leading to:

$$\begin{aligned} \frac{2\gamma n\bar{p}}{n} (\mathbb{E}f(\bar{x}^k) - f(x^*)) &\leq \frac{4\gamma n\bar{p}}{n} (\mathbb{E}f(\hat{x}^k) - f(x^*)) + \frac{4L\gamma n\bar{p}}{n} \|\bar{x}^k - \hat{x}^k\|^2 \\ &\leq 2\mathbb{E} \left[\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^*\|^2 \right] \\ &\quad + \frac{2\gamma^2\sigma^2 n\bar{p} + 2\gamma^2 \left(\sum_{v \in \mathcal{V}} p_v \mathbb{E} \|\nabla f_v(x_v^k)\|^2 + \|\sum_{v \in \mathcal{V}} p_v \nabla f_v(x_v^k)\|^2 \right)}{n^2} \\ &\quad + \mathbb{E} \left[\frac{4\gamma L p_{\max}}{n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 8\gamma L \bar{p} \|\hat{x}^k - \bar{x}^k\|^2 \right]. \end{aligned}$$

We have $\|\hat{x}^k - \bar{x}^k\|^2 \leq \gamma^2 B^2$. Now,

$$\begin{aligned} \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(x_v^k)\|^2 &\leq 2 \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(x_v^k) - \nabla f_v(\bar{x}^k)\|^2 + p_v \|\nabla f_v(\bar{x}^k)\|^2 \\ &\leq 2L^2 p_{\max} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2 \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(\bar{x}^k)\|^2 \\ &\leq 2L^2 p_{\max} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2n\bar{p} \|\nabla f(\bar{x}^k)\|^2 + 2n\bar{p}\zeta^2. \end{aligned}$$

Then,

$$\begin{aligned} \left\| \sum_{v \in \mathcal{V}} p_v \nabla f_v(x_v^k) \right\|^2 &\leq 2 \left\| \sum_{v \in \mathcal{V}} p_v \nabla f_v(\bar{x}^k) \right\|^2 + 2 \left\| \sum_{v \in \mathcal{V}} p_v (\nabla f_v(x_v^k) - \nabla f_v(\bar{x}^k)) \right\|^2 \\ &\leq 2(n\bar{p})^2 \|\nabla f(\bar{x}^k)\|^2 + 2(n\bar{p}) \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(x_v^k) - \nabla f_v(\bar{x}^k)\|^2 \\ &\leq 2(n\bar{p})^2 \|\nabla f(\bar{x}^k)\|^2 + 2(n\bar{p}) \sum_{v \in \mathcal{V}} p_v L^2 \|x_v^k - \bar{x}^k\|^2 \\ &\leq 2(n\bar{p})^2 \|\nabla f(\bar{x}^k)\|^2 + 2(n\bar{p}) p_{\max} L^2 \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \end{aligned}$$

Thus, this leads to:

$$\begin{aligned} \frac{2\gamma n\bar{p}}{n} (\mathbb{E}f(\bar{x}^k) - f(x^*)) &\leq 2\mathbb{E} \left[\|\hat{\mathbf{x}}^k - \mathbf{x}^*\|^2 - \|\hat{\mathbf{x}}^{k+1} - \mathbf{x}^*\|^2 \right] + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)n\bar{p} + 8\gamma^2 n^2 \bar{p}^2 \|\nabla f(\bar{x}^k)\|^2}{n^2} \\ &\quad + \mathbb{E} \left[\left(\frac{4\gamma L p_{\max}}{n} + \frac{2\gamma^2 L^2 p_{\max}(1+n\bar{p})}{n} \right) \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 8\gamma L \bar{p} \|\hat{x}^k - \bar{x}^k\|^2 \right]. \end{aligned}$$

We now use the following lemma.

Lemma 7. For stepsizes $\gamma \leq \frac{\bar{\rho}}{4L\sqrt{p_{\max}}}$, we have:

$$\sum_{k < K} \mathbb{E} \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \right] \leq 4\gamma^2 \sigma^2 \bar{\rho}^{-1} n\bar{p}K + 8\gamma^2 \bar{\rho}^{-2} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 + 16\gamma^2 \bar{\rho}^{-2} n\bar{p} \sum_{k < K} (\|\nabla f(\bar{x}^k)\|^2 + \zeta^2).$$

Proof of the lemma. Denoting $C_K = \sum_{k < K} \mathbb{E} \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \right]$ and using Lemma 2, we have

$$\begin{aligned} C_k &\leq 2\gamma^2 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| + \frac{4\gamma^2}{\bar{\rho}^2} \sum_{k < K} \mathbb{E} \left[\sum_{v \in \mathcal{I}_k} \|\nabla f_v(x_v^{k-\tau(k,v)})\|^2 \right] \\ &\leq 2\gamma^2 \sigma^2 \bar{\rho}^{-1} n\bar{p}K + 8\gamma^2 \bar{\rho}^{-2} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 + \frac{4\gamma^2}{\bar{\rho}^2} \sum_{k < K} \sum_{v \in \mathcal{V}} p_v \mathbb{E} \left[\|\nabla f_v(x_v^k)\|^2 \right], \end{aligned}$$

using $\sum_{k < K} \sum_{v \in \mathcal{I}_k} \left\| \nabla f_v(x_v^{k-\tau(k,v)}) \right\|^2 \leq \sum_{k < K} \sum_{v \in \mathcal{I}_k} \left\| \nabla f_v(x_v^k) \right\|^2 + \sum_{v \in \mathcal{V}} \left\| \nabla f_v(x_v^0) \right\|^2$. Then, $\sum_{v \in \mathcal{V}} p_v \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \right] \leq 2 \sum_{v \in \mathcal{V}} p_v \mathbb{E} \left[\left\| \nabla f_v(\bar{x}^k) \right\|^2 \right] + 2 \sum_{v \in \mathcal{V}} p_v \mathbb{E} \left[\left\| \nabla f_v(\bar{x}^k) - \nabla f_v(x_v^k) \right\|^2 \right] \leq 2n\bar{p}\zeta^2 + 2n\bar{p}\mathbb{E} \left\| \nabla f(\bar{x}^k) \right\|^2 + 2L^2 p_{\max} \mathbb{E} \left\| \mathbf{x}^k - \bar{\mathbf{x}}^k \right\|^2$, which leads to:

$$C_K \leq 2\gamma^2 \sigma^2 \bar{\rho}^{-1} n \bar{p} K + 4\gamma^2 \bar{\rho}^{-2} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{x}^0) \right\|^2 + 8\gamma^2 \bar{\rho}^{-2} n \bar{p} \sum_{k < K} \left(\left\| \nabla f(\bar{x}^k) \right\|^2 + \zeta^2 \right) + 8\gamma^2 L^2 p_{\max} \bar{\rho}^{-2} C_K,$$

leading to the desired result for $\gamma \leq \frac{\bar{\rho}}{4L\sqrt{p_{\max}}}$. \square

Using Lemma 1 and Lemma 7, we thus have:

$$\begin{aligned} \frac{2\gamma n \bar{p}}{n} \sum_{k < K} \left(\mathbb{E} f(\bar{x}^k) - f(x^*) \right) &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + \frac{2\gamma^2 (\sigma^2 + 2\zeta^2) \bar{p} K}{n} + 4\gamma^2 \bar{p} \sum_{k < K} \mathbb{E} \left[\left\| \nabla f(\bar{x}^k) \right\|^2 \right] \\ &+ \mathbb{E} \left[\left(\frac{4\gamma L p_{\max}}{n} + \frac{2\gamma^2 L^2 p_{\max}}{n} \right) \sum_{k < K} \left\| \mathbf{x}^k - \bar{\mathbf{x}}^k \right\|^2 \right] + 8\gamma^3 L B^2 \bar{p} K \\ &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + \frac{2\gamma^2 (\sigma^2 + 2\zeta^2) \bar{p} K}{n} + 4\gamma^2 \bar{p} \sum_{k < K} \mathbb{E} \left[\left\| \nabla f(\bar{x}^k) \right\|^2 \right] + 8\gamma^3 L B^2 \bar{p} K \\ &+ \frac{6\gamma L p_{\max}}{n} \left[4\gamma^2 \sigma^2 \bar{\rho}^{-1} n \bar{p} K + 8\gamma^2 \bar{\rho}^{-2} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{x}^0) \right\|^2 + 16\gamma^2 \bar{\rho}^{-2} n \bar{p} \sum_{k < K} \left(\left\| \nabla f(\bar{x}^k) \right\|^2 + \zeta^2 \right) \right] \\ &= 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + (8\gamma^2 L \bar{p} + 96\gamma^3 L^2 p_{\max} \bar{p} \bar{\rho}^{-2}) \sum_{k < K} \mathbb{E} [f(\bar{x}^k) - f(x^*)] + \frac{2\gamma^2 (\sigma^2 + 2\zeta^2) \bar{p} K}{n} \\ &+ \gamma^3 K (8L B^2 \bar{p} + 24L \sigma^2 p_{\max} \bar{p} \bar{\rho}^{-1} + 96L \zeta^2 p_{\max} \bar{p} \bar{\rho}^{-2}) + \frac{48\gamma^2 L p_{\max} \bar{\rho}^{-2}}{n} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{x}^0) \right\|^2. \end{aligned}$$

Hence, for stepsizes satisfying $8\gamma L \bar{p} + 96\gamma^2 L^2 p_{\max} \bar{p} \bar{\rho}^{-2} \leq \bar{p}$, which is verified for $\gamma \leq \min \left(\frac{1}{16L}, \frac{\bar{\rho}}{14L\sqrt{p_{\max}}} \right)$, we obtain:

$$\begin{aligned} \sum_{k < K} \left(\mathbb{E} f(\bar{x}^k) - f(x^*) \right) &\leq \frac{2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right]}{\gamma \bar{p}} + \frac{2\gamma (\sigma^2 + 2\zeta^2) K}{n} + \gamma^2 K (8L B^2 + 24L \sigma^2 p_{\max} \bar{\rho}^{-1} + 96L \zeta^2 p_{\max} \bar{\rho}^{-2}) \\ &+ \frac{48\gamma L p_{\max} \bar{\rho}^{-2}}{n \bar{p}} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{x}^0) \right\|^2. \end{aligned}$$

Optimizing over $\gamma \leq \min \left(\frac{1}{16L}, \frac{\bar{\rho}}{14L\sqrt{p_{\max}}}, \frac{\bar{p}}{L} \right)$, this leads to:

$$\begin{aligned} \frac{1}{K} \sum_{k < K} \left(\mathbb{E} f(\bar{x}^k) - f(x^*) \right) &= \mathcal{O} \left(\frac{LD^2 \left(\frac{1}{\bar{p}} + \sqrt{\frac{p_{\max}}{\bar{p}^2} \bar{\rho}^{-1}} \right)}{K} + \left[\frac{D^2 \sqrt{LB^2 + L\sigma^2 p_{\max} \bar{\rho}^{-1} + L\zeta p_{\max} \bar{\rho}^{-2}}}{\bar{p} K} \right]^{\frac{2}{3}} \right. \\ &\left. + \sqrt{\frac{D^2 (\sigma^2 + \zeta^2)}{n \bar{p} K}} + \frac{\bar{\rho}^{-1} p_{\max}}{K} \frac{1}{n} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{x}^0) \right\|^2 \right). \end{aligned}$$

\square

F Proof of Theorem 3: smooth-convex case

F.1 Homogeneous without sampling

Proof. As before, we have:

$$\mathbb{E} \left[\|\hat{x}^{k+1} - x^*\|^2 \right] \leq \mathbb{E} \left[\|\hat{x}^k - x^*\|^2 + T_1 + T_2^k + T_3 \right] + \frac{\gamma^2 \sigma^2 |\mathcal{I}_k| + \gamma^2 \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2}{n^2},$$

with

$$T_1 = -\frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - x^* \rangle$$

$$T_2^k = \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), x_v^k - \bar{x}^k \rangle$$

$$T_3 = \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f_v(x_v^k), \bar{x}^k - \hat{x}^k \rangle,$$

We will bound T_1, T_2 as in the proof with the Lipschitz assumption. For the term T_3 , using convexity and Lemma 1:

$$\begin{aligned} \mathbb{E} T_3 &\leq \frac{\gamma}{n} \left(C \sum_{v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 + \frac{1}{C} \mathbb{E} \|\bar{x}^k - \hat{x}^k\|^2 \right) \\ &\leq \frac{\gamma}{n} \left(2LC \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)) + \frac{2\gamma^2}{Cn} |\mathcal{I}_k| (\sigma^2 + \sum_{v \in \mathcal{V}} \|\nabla f(x_v^{k-\tau(v,k)})\|^2) \right) \\ &\leq \frac{2\gamma^2}{Cn^2} |\mathcal{I}_k| \sigma^2 + \frac{\gamma}{n} \left(2LC \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)) + \frac{2\gamma^2}{Cn} |\mathcal{I}_k| \sum_{v \in \mathcal{V}} \|\nabla f(x_v^{k-\tau(v,k)})\|^2 \right). \end{aligned}$$

for $\gamma \leq 1/(nL)$. Then,

$$\begin{aligned} \sum_{k < K} |\mathcal{I}_k| \sum_{v \in \mathcal{V}} \|\nabla f(x_v^{k-\tau(v,k)})\|^2 &\leq \sum_{v \in \mathcal{V}} \sum_{k < K: v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 \sum_{\ell=k}^{\text{next}(v,k+1)-1} |\mathcal{I}_\ell| \\ &\leq \tau_{\max} \sum_{v \in \mathcal{V}} \sum_{k < K: v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 \\ &\leq 2L\tau_{\max} \sum_{v \in \mathcal{V}} \sum_{k < K: v \in \mathcal{I}_k} f(x_v^k) - f(x^*), \end{aligned}$$

where τ_{\max} is an upper bound on the maximal compute delay defined as $\tau_{\max} \geq \sup_{k < K} \sum_{\ell=k}^{\text{next}(v,k+1)-1} |\mathcal{I}_\ell|$.

Thus,

$$\begin{aligned} \frac{2\gamma}{n} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) &\leq -\mathbb{E} \left[\|\hat{x}^{k+1} - x^*\|^2 \right] + \mathbb{E} \left[\|\hat{x}^k - x^*\|^2 \right] \\ &\quad + \frac{\gamma^2 \sigma^2 |\mathcal{I}_k|}{n^2} \\ &\quad + \frac{2\gamma^2 L |\mathcal{I}_k|}{n^2} \sum_{v \in \mathcal{I}_k} (\mathbb{E} f(x_v^k) - f(x^*)) \\ &\quad + \frac{2LC\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} [(f(x_v^k) - f(x^*))] \\ &\quad + \frac{\gamma}{Cn} \mathbb{E} \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \right] \\ &\quad + \frac{2\gamma^2 L}{Cn^2} |\mathcal{I}_k| \sigma^2 + \frac{\gamma}{n} \left(2LC \sum_{v \in \mathcal{I}_k} (f(x_v^k) - f(x^*)) + \frac{2\gamma^2}{Cn} |\mathcal{I}_k| \sum_{v \in \mathcal{V}} \|\nabla f(x_v^{k-\tau(v,k)})\|^2 \right). \end{aligned}$$

Summing over $k < K$, using Lemma 2 and our bound on T_3 , we obtain:

$$\begin{aligned}
 \frac{2\gamma}{n} \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E}f(x_v^k) - f(x^*)) &\leq \mathbb{E} \left[\|\hat{x}^0 - x^*\|^2 \right] + \frac{3\gamma^2\sigma^2}{n^2} \sum_{k < K} |\mathcal{I}_k| \\
 &+ \frac{2\gamma L}{n} \left(2C + \gamma + \frac{2\tau_{\max}\gamma^2}{Cn} \right) \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E}f(x_v^k) - f(x^*)) \\
 &+ \sum_{k < K} \frac{\gamma}{Cn} \mathbb{E} \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \right] \\
 &\leq \mathbb{E} \left[\|\hat{x}^0 - x^*\|^2 \right] + \frac{\gamma^2\sigma^2}{n^2} \left(1 + \frac{2\gamma\bar{\rho}^{-1}n}{C} \right) \sum_{k < K} |\mathcal{I}_k| + \sum_{k < K} \frac{\gamma}{Cn} \mathbb{E} \left[\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \right] \\
 &+ \frac{\gamma^3 B^2}{Cn} \sum_{k < K} |\mathcal{I}_k| \\
 &+ \frac{2\gamma L}{n} \left(2C + \gamma + \frac{2\tau_{\max}\gamma^2}{Cn} + \frac{4\gamma^2}{\bar{\rho}^2} \right) \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E}f(x_v^k) - f(x^*)),
 \end{aligned}$$

using Lemma 7 to handle the sum of the terms $\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2$.

Hence, provided that $2C + \gamma + \frac{2\tau_{\max}\gamma^2}{Cn} + \frac{4\gamma^2}{\bar{\rho}^2} \leq \frac{1}{2L}$, which is verified for $C = \frac{1}{8L}$ and $\gamma \leq \frac{1}{4L} \times \frac{1}{1+2\bar{\rho}^{-1}+4\sqrt{\tau_{\max}/n}}$, we have:

$$\frac{\gamma}{n} \sum_{k < K} \sum_{v \in \mathcal{I}_k} (\mathbb{E}f(x_v^k) - f(x^*)) \leq \mathbb{E} \left[\|\hat{x}^0 - x^*\|^2 \right] + \frac{\gamma^2\sigma^2}{n^2} (3 + 16L\gamma\bar{\rho}^{-1}n) \sum_{k < K} |\mathcal{I}_k|,$$

leading to, for $\eta = \gamma/n$:

$$\mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] \leq \frac{\mathbb{E} \left[\|\hat{x}^0 - x^*\|^2 \right]}{\eta \sum_{k < K} |\mathcal{I}_k|} + 3\eta\sigma^2 + \eta^2 16L\sigma^2 n^2 \bar{\rho}^{-1}.$$

Optimizing over $\eta \leq \frac{1}{4L} \times \frac{1}{n(1+2\bar{\rho}^{-1})+4\sqrt{n\tau_{\max}}}$, we thus obtain that:

$$\begin{aligned}
 &\mathbb{E} \left[f \left(\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k=0}^{K-1} \sum_{v \in \mathcal{I}_k} x_v^k \right) - f(x^*) \right] \\
 &= \mathcal{O} \left(\frac{LD^2(n\bar{\rho}^{-1} + \sqrt{n\tau_{\max}})}{\sum_{k < K} |\mathcal{I}_k|} + \sqrt{\frac{D\sigma^2}{\sum_{k < K} |\mathcal{I}_k|}} + \left[\frac{D^2\sqrt{L\sigma^2 n^2 \bar{\rho}^{-1}}}{\sum_{k < K} |\mathcal{I}_k|} \right]^{2/3} \right).
 \end{aligned}$$

□

F.2 Heterogeneous setting under sampling

Proof. As in the Lipschitz case, we have:

$$\begin{aligned}
 \frac{2\gamma n\bar{p}}{n} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) &\leq 2\mathbb{E} \left[\|\hat{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \right] + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)\bar{p}K}{n} + 4\gamma^2\bar{p} \sum_{k < K} \mathbb{E} \left[\|\nabla f(\bar{x}^k)\|^2 \right] \\
 &+ \mathbb{E} \left[\frac{6\gamma L p_{\max}}{n} \sum_{k < K} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \right] + 8\gamma L\bar{p} \mathbb{E} \left[\sum_{k < K} \|\bar{x}^k - \hat{x}^k\|^2 \right].
 \end{aligned}$$

Since losses are no longer assumed to be Lipschitz, we cannot bound this last term $\mathbb{E} \left[\sum_{k < K} \|\bar{x}^k - \hat{x}^k\|^2 \right]$ by $\gamma^2 B^2$. However, using Lemma 1,

$$\mathbb{E} \left[\sum_{k < K} \|\bar{x}^k - \hat{x}^k\|^2 \right] \leq \frac{2\gamma^2\sigma^2 K}{n} + \frac{2\gamma^2}{n} \mathbb{E} \left[\sum_{v \in \mathcal{V}} \sum_{k < K} \left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right].$$

Then,

$$\begin{aligned}
 \mathbb{E} \left[\sum_{v \in \mathcal{V}} \sum_{k < K} \left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right] &= \sum_{v \in \mathcal{V}} \sum_{k < K} \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \mathbb{1}_{v \in \mathcal{I}_k} (\text{next}(k, v) - k) \right] \\
 &= \sum_{v \in \mathcal{V}} \sum_{k < K} \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \times \frac{1}{p_v} \times p_v \right] \\
 &= \sum_{v \in \mathcal{V}} \sum_{k < K} \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \right] \\
 &\leq \frac{1}{p_{\min}} \sum_{v \in \mathcal{V}} \sum_{k < K} p_v \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \right].
 \end{aligned}$$

since the random variables $\left\| \nabla f_v(x_v^k) \right\|^2$, $\mathbb{1}_{v \in \mathcal{I}_k}$ and $\text{next}(k, v) - k$ are independent, $\mathbb{E}[\mathbb{1}_{v \in \mathcal{I}_k}] = p_v$ (Bernoulli random variable) and $\mathbb{E}[\text{next}(k, v) - k] = \frac{1}{p_v}$ (geometric random variable). And then, as we proved before, $\sum_{v \in \mathcal{V}} \sum_{k < K} p_v \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \right] \leq 2L^2 p_{\max} \left\| \mathbf{x}^k - \bar{\mathbf{x}}^k \right\|^2 + 2n\bar{p} \left\| \nabla f(\bar{\mathbf{x}}^k) \right\|^2 + 2n\bar{p}\zeta^2$. Consequently,

$$\begin{aligned}
 \frac{2\gamma n\bar{p}}{n} \sum_{k < K} (\mathbb{E}f(\bar{\mathbf{x}}^k) - f(x^*)) &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)\bar{p}K}{n} \\
 &\quad + (4\gamma^2\bar{p} + 32\gamma^3 L\bar{p} \frac{p_{\max}}{p_{\min}}) \sum_{k < K} \mathbb{E} \left[\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|^2 \right] \\
 &\quad + \mathbb{E} \left[\left(\frac{6\gamma L p_{\max}}{n} + \frac{32\gamma^3 L^3 \bar{p} p_{\max}}{n p_{\min}} \right) \sum_{k < K} \left\| \mathbf{x}^k - \bar{\mathbf{x}}^k \right\|^2 \right] + \frac{16\gamma^3 \sigma^2 L \bar{p} K}{n} + \frac{32\gamma^3 L \zeta^2 \bar{p}^2}{p_{\min}} \\
 &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)\bar{p}K}{n} \\
 &\quad + (4\gamma^2\bar{p} + 32\gamma^3 L\bar{p} \frac{p_{\max}}{p_{\min}}) \sum_{k < K} \mathbb{E} \left[\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|^2 \right] \\
 &\quad + \mathbb{E} \left[\frac{12\gamma L p_{\max}}{n} \sum_{k < K} \left\| \mathbf{x}^k - \bar{\mathbf{x}}^k \right\|^2 \right] + \frac{16\gamma^3 \sigma^2 L \bar{p} K}{n} + \frac{32\gamma^3 L \zeta^2 \bar{p}^2}{p_{\min}}.
 \end{aligned}$$

provided that $\gamma \leq \sqrt{\frac{6p_{\min}}{32L^2 p_{\max}}}$. Plugging Lemma 7 in here, we obtain:

$$\begin{aligned}
 \frac{2\gamma n\bar{p}}{n} \sum_{k < K} (\mathbb{E}f(\bar{\mathbf{x}}^k) - f(x^*)) &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)\bar{p}K}{n} \\
 &\quad + (4\gamma^2\bar{p} + 32\gamma^3 L\bar{p} \frac{p_{\max}}{p_{\min}}) \sum_{k < K} \mathbb{E} \left[\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|^2 \right] \\
 &\quad + \frac{16\gamma^3 \sigma^2 L \bar{p} K}{n} + \frac{32\gamma^3 L \zeta^2 \bar{p}^2}{p_{\min}} \\
 &\quad + \frac{12\gamma L p_{\max}}{n} \left[4\gamma^2 \sigma^2 \bar{\rho}^{-1} n \bar{p} K + 8\gamma^2 \bar{\rho}^{-2} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{\mathbf{x}}^0) \right\|^2 + 16\gamma^2 \bar{\rho}^{-2} n \bar{p} \sum_{k < K} (\left\| \nabla f(\bar{\mathbf{x}}^k) \right\|^2 + \zeta^2) \right] \\
 &= 2\mathbb{E} \left[\left\| \hat{\mathbf{x}}^0 - \mathbf{x}^* \right\|^2 \right] + (8\gamma^2 L \bar{p} + 192\gamma^3 L^2 p_{\max} \bar{p} \bar{\rho}^{-2} + 64\gamma^3 L^2 \bar{p} \frac{p_{\max}}{p_{\min}}) \sum_{k < K} \mathbb{E} [f(\bar{\mathbf{x}}^k) - f(x^*)] \\
 &\quad + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)\bar{p}K}{n} + \gamma^3 K (8LB^2\bar{p} + 24L\sigma^2 p_{\max} \bar{p} \bar{\rho}^{-1} + 96L\zeta^2 p_{\max} \bar{p} \bar{\rho}^{-2}) \\
 &\quad + \frac{96\gamma^3 L p_{\max} \bar{\rho}^{-2}}{n} \sum_{v \in \mathcal{V}} \left\| \nabla f_v(\bar{\mathbf{x}}^0) \right\|^2.
 \end{aligned}$$

For $8\gamma^2 L\bar{\rho} + 192\gamma^3 L^2 p_{\max} \bar{\rho} \bar{\rho}^{-2} + 64\gamma^3 L^2 \bar{\rho} \frac{p_{\max}}{p_{\min}} \leq \gamma \bar{\rho}$ which is verified for $\gamma \leq \min \left(\frac{1}{24L}, \frac{\bar{\rho}}{24L\sqrt{p_{\max}}}, \frac{1}{14L\sqrt{\frac{p_{\max}}{p_{\min}}}} \right)$,

we have:

$$\begin{aligned} \gamma \bar{\rho} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) &\leq 2\mathbb{E} \left[\|\hat{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \right] + \frac{96\gamma^3 L p_{\max} \bar{\rho}^{-2}}{n} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 \\ &\quad + \frac{2\gamma^2(\sigma^2 + 2\zeta^2)\bar{\rho}K}{n} + \gamma^3 K (8LB^2\bar{\rho} + 24L\sigma^2 p_{\max} \bar{\rho} \bar{\rho}^{-1} + 96L\zeta^2 p_{\max} \bar{\rho} \bar{\rho}^{-2}), \end{aligned}$$

and thus:

$$\begin{aligned} \frac{1}{K} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) &\leq \frac{2\mathbb{E} \left[\|\hat{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \right]}{\bar{\rho}\gamma K} + \frac{96\gamma^2 L p_{\max} \bar{\rho}^{-2}}{n\bar{\rho}K} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 \\ &\quad + \frac{2\gamma(\sigma^2 + 2\zeta^2)}{n} + \gamma^2 (8LB^2 + 24L\sigma^2 p_{\max} \bar{\rho}^{-1} + 96L\zeta^2 p_{\max} \bar{\rho}^{-2}). \end{aligned}$$

Now, we use

$$\sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 \leq \sum_{v \in \mathcal{V}} \|\nabla f(\bar{x}^0)\|^2 + \zeta^2 \leq \sum_{v \in \mathcal{V}} 2L(f(x_0) - f(x^*)) + \zeta^2,$$

so that

$$\begin{aligned} \frac{96\gamma^2 L p_{\max} \bar{\rho}^{-2}}{n\bar{\rho}K} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 &\leq \frac{192\gamma^2 L^2 p_{\max} \bar{\rho}^{-2}}{\bar{\rho}K} (f(x_0) - f(x^*)) + \frac{96\zeta^2 \gamma^2 L p_{\max} \bar{\rho}^{-2}}{\bar{\rho}K} \\ &\leq \frac{192\gamma^2 L^2 p_{\max} \bar{\rho}^{-2}}{\bar{\rho}} \frac{1}{K} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) + \frac{96\zeta^2 \gamma^2 L p_{\max} \bar{\rho}^{-2}}{\bar{\rho}K} \\ &\leq \frac{1}{2} \frac{1}{K} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) + 96\zeta^2 \gamma^2 L p_{\max} \bar{\rho}^{-2}, \end{aligned}$$

for $K \geq \frac{1}{\bar{\rho}}$ and $\gamma \leq \frac{1\bar{\rho}}{384L\sqrt{p_{\max}}}$. Thus,

$$\begin{aligned} \frac{1}{2K} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) &\leq \frac{2\mathbb{E} \left[\|\hat{\mathbf{x}}^0 - \mathbf{x}^*\|^2 \right]}{\bar{\rho}\gamma K} \\ &\quad + \frac{2\gamma(\sigma^2 + 2\zeta^2)}{n} + \gamma^2 (8LB^2 + 24L\sigma^2 p_{\max} \bar{\rho}^{-1} + 192L\zeta^2 p_{\max} \bar{\rho}^{-2}). \end{aligned}$$

Optimizing over admissible γ 's leads to:

$$\begin{aligned} &\frac{1}{K} \sum_{k < K} (\mathbb{E}f(\bar{x}^k) - f(x^*)) \\ &= \mathcal{O} \left(\frac{LD^2 \left(\frac{1}{\bar{\rho}} \sqrt{\frac{p_{\max}}{p_{\min}}} + \sqrt{\frac{p_{\max}}{\bar{\rho}^2} \bar{\rho}^{-1}} \right)}{K} + \sqrt{\frac{D^2(\sigma^2 + \zeta^2)}{n\bar{\rho}K}} + \left[\frac{D^2 \sqrt{LB^2 + L\sigma^2 p_{\max} \bar{\rho}^{-1} + L\zeta^2 p_{\max} \bar{\rho}^{-2}}}{\bar{\rho}K} \right]^{\frac{2}{3}} \right). \end{aligned}$$

□

G Proof of Theorem 4: smooth non-convex case

G.1 Homogeneous without sampling

Proof. Using L -smoothness and a virtual sequence \hat{x} defined in Section B.1, we have

$$\mathbb{E}_{k+1} f(\hat{x}^{k+1}) \leq f(\hat{x}^k) - \underbrace{\frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\hat{x}^k), \nabla f(x_v^k) \rangle}_{:=T_1} + \frac{L\gamma^2}{2n^2} \left(\sigma^2 |\mathcal{I}_k| + \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f(x_v^k) \right\|^2 \right) \quad (69)$$

We separately estimate the middle term as

$$\begin{aligned}
 T_1 &= -\frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\hat{x}^k), \nabla f(x_v^k) \rangle = -\frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\bar{x}^k), \nabla f(x_v^k) \rangle + \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\bar{x}^k) - \nabla f(\hat{x}^k), \nabla f(x_v^k) \rangle \\
 &\leq \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \left(-\frac{1}{2} \|\nabla f(\bar{x}^k)\|^2 - \frac{1}{2} \|\nabla f(x_v^k)\|^2 + \frac{L^2}{2} \|x_v^k - \bar{x}^k\|^2 \right) + \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \left(\frac{1}{4} \|\nabla f(x_v^k)\|^2 + L^2 \|\bar{x}^k - \hat{x}^k\|^2 \right) \\
 &\leq -\frac{\gamma}{4n} \sum_{v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 - \frac{|\mathcal{I}_k| \gamma}{2n} \|\nabla f(\bar{x}^k)\|^2 + \frac{L^2 \gamma}{2n} \sum_{v \in \mathcal{I}_k} \|x_v^k - \bar{x}^k\|^2 + \frac{\gamma L^2 |\mathcal{I}_k|}{n} \|\bar{x}^k - \hat{x}^k\|^2
 \end{aligned}$$

where we used that for any vectors $a, b \in \mathbb{R}^d$ it holds that $-\langle a, b \rangle = -\frac{1}{2} \|a\|^2 - \frac{1}{2} \|b\|^2 + \frac{1}{2} \|a - b\|^2$ and also it holds that $2\langle a, b \rangle \leq \gamma \|a\|^2 + \gamma^{-1} \|b\|^2$ for any $\gamma > 0$ and we chose $\gamma = 2$.

We further use Lemma 1 to estimate the last term

$$\begin{aligned}
 T_1 &\leq -\frac{\gamma}{4n} \sum_{v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 - \frac{|\mathcal{I}_k| \gamma}{2n} \|\nabla f(\bar{x}^k)\|^2 + \frac{L^2 \gamma}{2n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\
 &\quad + \frac{2L^2 \gamma^3 |\mathcal{I}_k|}{n^2} \left(\sigma^2 + \sum_{v \in \mathcal{V}} \mathbb{E} \left[\left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right] \right)
 \end{aligned}$$

Putting this estimate of T_1 back into (69) we get

$$\begin{aligned}
 \mathbb{E}_{k+1} f(\hat{x}^{k+1}) &\leq f(\hat{x}^k) + \frac{L\gamma^2 \sigma^2 |\mathcal{I}_k|}{2n^2} + \frac{L\gamma^2}{2n} \sum_{v \in \mathcal{I}_k} \mathbb{E} \|\nabla f(x_v^k)\|^2 - \frac{\gamma}{4n} \sum_{v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 - \frac{|\mathcal{I}_k| \gamma}{2n} \|\nabla f(\bar{x}^k)\|^2 \\
 &\quad + \frac{L^2 \gamma}{2n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \frac{2L^2 \gamma^3 |\mathcal{I}_k|}{n^2} \left(\sigma^2 + \sum_{v \in \mathcal{V}} \mathbb{E} \left[\left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right] \right)
 \end{aligned}$$

Using that $\gamma < \frac{1}{4L}$ we estimate

$$\begin{aligned}
 \mathbb{E}_{k+1} f(\hat{x}^{k+1}) &\leq f(\hat{x}^k) - \frac{\gamma}{8n} \sum_{v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 - \frac{|\mathcal{I}_k| \gamma}{2n} \|\nabla f(\bar{x}^k)\|^2 + \frac{L^2 \gamma}{2n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\
 &\quad + \frac{2L^2 \gamma^3 |\mathcal{I}_k|}{n^2} \left(\sigma^2 + \sum_{v \in \mathcal{V}} \mathbb{E} \left[\left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right] \right) + \frac{L\gamma^2 \sigma^2 |\mathcal{I}_k|}{2n^2}
 \end{aligned}$$

Taking the full expectation and summing over all the iterations k , we get

$$\begin{aligned}
 \sum_{k < K} \frac{|\mathcal{I}_k| \gamma}{2n} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) - \frac{\gamma}{8n} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \|\nabla f(x_v^k)\|^2 + \frac{L^2 \gamma}{2n} \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\
 &\quad + \frac{L\gamma^2 \sigma^2 \sum_{k < K} |\mathcal{I}_k|}{2n^2} (1 + 4L\gamma) + \frac{2L^2 \gamma^3}{n^2} \sum_{k < K} \sum_{v \in \mathcal{V}} |\mathcal{I}_k| \mathbb{E} \left[\left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right]
 \end{aligned}$$

For the third term we use Lemma 2, and for the last term we use that

$$\begin{aligned}
 \sum_{k < K} |\mathcal{I}_k| \sum_{v \in \mathcal{V}} \left\| \nabla f(x_v^{\text{prev}(v,k)}) \right\|^2 &\leq \sum_{v \in \mathcal{V}} \sum_{k < K: v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2 \sum_{\ell=k}^{\text{next}(v,k+1)-1} |\mathcal{I}_\ell| \\
 &\leq \tau_{\max} \sum_{v \in \mathcal{V}} \sum_{k < K: v \in \mathcal{I}_k} \|\nabla f(x_v^k)\|^2
 \end{aligned}$$

where τ_{\max} is an upper bound on the maximal compute delay defined as $\tau_{\max} \geq \sup_{k < K} \sum_{\ell=k}^{\text{next}(v,k+1)-1} |\mathcal{I}_\ell|$. For estimating the third term with Lemma 2, we also use that

$$\sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\left\| \nabla f_v(x_v^{k-\tau(k,v)}) \right\|^2 \right] \leq \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \right]$$

We therefore get

$$\begin{aligned} \sum_{k < K} \frac{|\mathcal{I}_k| \gamma}{2n} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) - \frac{\gamma}{8n} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \|\nabla f(x_v^k)\|^2 + \frac{L^2 \gamma}{2n} 2\gamma^2 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| \\ &\quad + \frac{2L^2 \gamma^3}{n \bar{\rho}^2} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \left[\|\nabla f(x_v^k)\|^2 \right] \\ &\quad + \frac{L\gamma^2 \sigma^2 \sum_{k < K} |\mathcal{I}_k|}{2n^2} (1 + 4L\gamma) + \frac{2L^2 \gamma^3}{n^2} \tau_{\max} \sum_{k < K} \sum_{v \in \mathcal{I}_k} \mathbb{E} \|\nabla f(x_v^k)\|^2 \end{aligned}$$

We further use that the stepsize $\gamma < \frac{1}{8L} (\sqrt{\frac{n}{\tau_{\max}}} + \bar{\rho})$

$$\sum_{k < K} \frac{|\mathcal{I}_k| \gamma}{2n} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 \leq (f(x^0) - f^*) + \frac{L^2}{n} \gamma^3 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| + \frac{L\gamma^2 \sigma^2 \sum_{k < K} |\mathcal{I}_k|}{n^2}$$

Therefore,

$$\sum_{k < K} |\mathcal{I}_k| \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 \leq \frac{2n}{\gamma} (f(x^0) - f^*) + 2L^2 \gamma^2 \sigma^2 \bar{\rho}^{-1} \sum_{k < K} |\mathcal{I}_k| + \frac{2L\gamma \sigma^2 \sum_{k < K} |\mathcal{I}_k|}{n}$$

Denoting $T = \sum_{k < K} |\mathcal{I}_k|$ and tuning over the stepsize γ , we get

$$\frac{1}{\sum_{k < K} |\mathcal{I}_k|} \sum_{k < K} |\mathcal{I}_k| \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 \leq \frac{16LF_0(\sqrt{n\tau_{\max}} + n\bar{\rho}^{-1})}{T} + 4 \left(\frac{L\sigma^2 F_0}{T} \right)^{\frac{1}{2}} + 4 \left(\frac{L\sigma n F_0}{T\sqrt{\bar{\rho}}} \right)^{\frac{2}{3}}$$

where $F_0 = (f(x^0) - f^*)$. □

G.2 Heterogeneous with sampling

Proof. Using L -smoothness of f ,

$$\mathbb{E}_{k+1} f(\hat{x}^{k+1}) \leq f(\hat{x}^k) - \underbrace{\frac{\gamma}{n} \mathbb{E} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\hat{x}^k), \nabla f_v(x_v^k) \rangle}_{:= T_1} + \frac{L\gamma^2}{2n^2} \left(\sigma^2 \mathbb{E} |\mathcal{I}_k| + \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2 \right) \quad (70)$$

We separately estimate the T_1 term

$$\begin{aligned} T_1 &= -\frac{\gamma}{n} \mathbb{E} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\hat{x}^k), \nabla f_v(x_v^k) \rangle = -\frac{\gamma}{n} \mathbb{E} \sum_{v \in \mathcal{I}_k} \langle \nabla f(\hat{x}^k), \nabla f_v(\bar{x}^k) \rangle + \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} \langle \nabla f(\hat{x}^k), \nabla f_v(\bar{x}^k) - \nabla f_v(x_v^k) \rangle \\ &= -\gamma \bar{p} \langle \nabla f(\hat{x}^k), \nabla f(\bar{x}^k) \rangle + \frac{\gamma}{n} \sum_{v \in \mathcal{I}_k} \mathbb{E} \langle \nabla f(\hat{x}^k), \nabla f_v(\bar{x}^k) - \nabla f_v(x_v^k) \rangle \\ &\leq \gamma \bar{p} \left(-\frac{1}{2} \|\nabla f(\hat{x}^k)\|^2 - \frac{1}{2} \|\nabla f(\bar{x}^k)\|^2 + \frac{L^2}{2} \|\hat{x}^k - \bar{x}^k\|^2 \right) + \frac{\gamma}{n} \left(\frac{1}{2} n \bar{p} \|\nabla f(\hat{x}^k)\|^2 + \frac{L^2}{2} \mathbb{E} \sum_{v \in \mathcal{I}_k} \|x_v^k - \bar{x}^k\|^2 \right) \end{aligned}$$

Since $\mathbb{E} \sum_{v \in \mathcal{I}_k} \nabla f_v(\bar{x}^k) = n \bar{p} \nabla f(\bar{x}^k)$, and $\mathbb{E} |\mathcal{I}_k| = n \bar{p}$. We further use that $\mathbb{E} \sum_{v \in \mathcal{I}_k} \|x_v^k - \bar{x}^k\|^2 = \sum_{v \in \mathcal{V}} p_v \|x_v^k - \bar{x}^k\|^2 \leq p_{\max} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2$. Therefore,

$$T_1 \leq -\frac{\gamma \bar{p}}{2} \|\nabla f(\bar{x}^k)\| + \frac{\gamma \bar{p} L^2}{2} \|\hat{x}^k - \bar{x}^k\|^2 + \frac{\gamma L^2 p_{\max}}{2n} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2$$

Putting this back to (70) and summing it up over K , we get

$$\begin{aligned} \frac{\gamma\bar{p}}{2} \sum_{k < K} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) + \frac{\gamma\bar{p}L^2}{2} \sum_{k < K} \mathbb{E} \|\hat{x}^k - \bar{x}^k\|^2 \\ &\quad + \frac{\gamma L^2 p_{\max}}{2n} \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \frac{L\gamma^2 \sigma^2 \bar{p}K}{2n} \\ &\quad + \frac{L\gamma^2}{2n^2} \sum_{k < K} \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2 \end{aligned}$$

We use calculations from Section E.2 to further estimate the last term

$$\begin{aligned} \mathbb{E} \left\| \sum_{v \in \mathcal{I}_k} \nabla f_v(x_v^k) \right\|^2 &\leq \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(x_v^k)\|^2 + \left\| \sum_{v \in \mathcal{V}} p_v \nabla f_v(x_v^k) \right\|^2 \\ \sum_{v \in \mathcal{V}} p_v \|\nabla f_v(x_v^k)\|^2 &\leq 2L^2 p_{\max} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2n\bar{p} \|\nabla f(\bar{x}^k)\|^2 + 2n\bar{p}\zeta^2. \end{aligned} \quad (71)$$

$$\left\| \sum_{v \in \mathcal{V}} p_v \nabla f_v(x_v^k) \right\|^2 \leq 2(n\bar{p})^2 \|\nabla f(\bar{x}^k)\|^2 + 2(n\bar{p})p_{\max}L^2 \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2$$

We therefore get

$$\begin{aligned} \frac{\gamma\bar{p}}{2} \sum_{k < K} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) + \frac{\gamma\bar{p}L^2}{2} \sum_{k < K} \mathbb{E} \|\hat{x}^k - \bar{x}^k\|^2 \\ &\quad + \frac{\gamma L^2 p_{\max}}{2n} \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + \frac{L\gamma^2 \sigma^2 \bar{p}K}{2n} \\ &\quad + \frac{L\gamma^2}{n^2} \left((L^2 p_{\max} + (n\bar{p})p_{\max}L^2) \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + (n\bar{p} + (n\bar{p})^2) \|\nabla f(\bar{x}^k)\|^2 + n\bar{p}\zeta^2 \right) \\ &\leq (f(x^0) - f^*) + \frac{\gamma\bar{p}L^2}{2} \sum_{k < K} \mathbb{E} \|\hat{x}^k - \bar{x}^k\|^2 \\ &\quad + \frac{\gamma L^2 p_{\max}}{2n} \left[1 + 2L\gamma \left(\frac{1}{n} + \bar{p} \right) \right] \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\ &\quad + \frac{L\gamma^2 \bar{p}K\sigma^2}{2n} + \frac{L\gamma^2 n\bar{p}(1 + n\bar{p})}{n^2} \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 + \frac{L\gamma^2 \bar{p}\zeta^2 K}{n} \end{aligned}$$

We further use Lemma 1 to estimate the term with $\mathbb{E} \|\hat{x}^k - \bar{x}^k\|^2$:

$$\mathbb{E} \left[\|\hat{x}^k - \bar{x}^k\|^2 \right] \leq \frac{2\gamma^2}{n} \left(\sigma^2 + \sum_{v \in \mathcal{V}} \mathbb{E} \left[\left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right] \right)$$

And we use calculations from Section F.2 estimating

$$\mathbb{E} \left[\sum_{v \in \mathcal{V}} \sum_{k < K} \left\| \nabla f_v(x_v^{\text{prev}(v,k)}) \right\|^2 \right] \leq \frac{1}{p_{\min}} \sum_{v \in \mathcal{V}} \sum_{k < K} p_v \mathbb{E} \left[\left\| \nabla f_v(x_v^k) \right\|^2 \right].$$

And (71) to estimate the last term. Therefore we get

$$\sum_{k < K} \mathbb{E} \left[\|\hat{x}^k - \bar{x}^k\|^2 \right] \leq \frac{2\gamma^2}{n} \left(\sigma^2 K + \frac{1}{p_{\min}} \sum_{k < K} 2L^2 p_{\max} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 + 2n\bar{p} \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 + 2n\bar{p}\zeta^2 K \right)$$

And

$$\begin{aligned} \frac{\gamma\bar{p}}{2} \sum_{k < K} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) + \frac{\gamma L^2 p_{\max}}{2n} \left[1 + 2L\gamma \left(\frac{1}{n} + \bar{p} \right) + \frac{4\gamma^2 L^2 \bar{p}}{p_{\min}} \right] \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\ &\quad + \frac{L\gamma^2 \bar{p} K \sigma^2}{2n} (1 + \gamma L) + \frac{L\gamma^2 n \bar{p} (1 + n\bar{p} + 2\gamma L n \bar{p} / p_{\min})}{n^2} \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 \\ &\quad + \frac{L\gamma^2 \bar{p} \zeta^2 K}{n} (1 + 2n\bar{p}\gamma L) \end{aligned}$$

We further use that $\gamma < \min \left\{ \frac{1}{4L}, \frac{\sqrt{\bar{p}}}{4L\sqrt{p_{\min}}} \right\}$

$$\begin{aligned} \frac{\gamma\bar{p}}{2} \sum_{k < K} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) + \frac{\gamma L^2 p_{\max}}{n} \sum_{k < K} \mathbb{E} \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2 \\ &\quad + \frac{L\gamma^2 \bar{p} K \sigma^2}{2n} (1 + \gamma L) + 3L\gamma^2 \bar{p} \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 + \frac{L\gamma^2 \bar{p} \zeta^2 K}{n} (1 + 2n\bar{p}\gamma L) \end{aligned}$$

We further use Lemma 7:

$$\begin{aligned} \frac{\gamma\bar{p}}{2} \sum_{k < K} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq \frac{\gamma L^2 p_{\max}}{n} \left[4\gamma^2 \sigma^2 \bar{\rho}^{-1} n \bar{p} K + 8\gamma^2 \bar{\rho}^{-2} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 \right. \\ &\quad \left. + 16\gamma^2 \bar{\rho}^{-2} n \bar{p} \sum_{k < K} (\|\nabla f(\bar{x}^k)\|^2 + \zeta^2) \right] \\ &\quad + \frac{L\gamma^2 \bar{p} K \sigma^2}{2n} (1 + \gamma L) + 3L\gamma^2 \bar{p} \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 + \frac{L\gamma^2 \bar{p} \zeta^2 K}{n} (1 + 2n\bar{p}\gamma L) + (f(x^0) - f^*) \\ &\leq (f(x^0) - f^*) + 4\gamma^3 L^2 p_{\max} \bar{p} \bar{\rho}^{-1} K \sigma^2 + \frac{8\gamma^3 L^2 p_{\max} \bar{\rho}^{-2}}{n} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 \\ &\quad + \frac{L\gamma^2 \bar{p} K \sigma^2}{2n} (1 + \gamma L) \\ &\quad + L\gamma^2 \bar{p} (3 + 16\gamma L \bar{\rho}^{-2} p_{\max}) \sum_{k < K} \|\nabla f(\bar{x}^k)\|^2 + \frac{2L\gamma^2 \bar{p} \zeta^2 K}{n} (1 + 8\bar{\rho}^{-2} n \gamma L p_{\max}) \end{aligned}$$

Taking the stepsize $\gamma < \min \left\{ \frac{1}{24L}, \frac{\bar{\rho}}{16L\sqrt{p_{\max}}} \right\}$ we get:

$$\begin{aligned} \frac{\gamma\bar{p}}{4} \sum_{k < K} \mathbb{E} \|\nabla f(\bar{x}^k)\|^2 &\leq (f(x^0) - f^*) + 4\gamma^3 L^2 p_{\max} \bar{p} \bar{\rho}^{-1} K \sigma^2 \\ &\quad + \frac{8\gamma^3 L^2 p_{\max} \bar{\rho}^{-2}}{n} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2 + \frac{2L\gamma^2 \bar{p} K \sigma^2}{2n} \\ &\quad + \frac{2L\gamma^2 \bar{p} \zeta^2 K}{n} (1 + 8\bar{\rho}^{-2} n \gamma L p_{\max}) \end{aligned}$$

We conclude as in the smooth convex case by tuning the stepsize and getting rid of the $\frac{8\gamma^3 L^2 p_{\max} \bar{\rho}^{-2}}{n} \sum_{v \in \mathcal{V}} \|\nabla f_v(\bar{x}^0)\|^2$. \square