Asynchrony and Acceleration in Gossip Algorithms

Mathieu Even, Hadrien Hendrikx, Laurent Massoulié

Asynchrony and Acceleration in Gossip Algorithms

Mathieu Even, Hadrien Hendrikx, and Laurent Massoulié

Inria - DI ENS - PSL Research University - DMA ENS

November 5, 2020

Abstract

This paper considers the minimization of a sum of smooth and strongly convex functions dispatched over the nodes of a communication network. Previous works on the subject either focus on synchronous algorithms, which can be heavily slowed down by a few slow nodes (the *straggler problem*), or consider a historical asynchronous setting (Boyd et al., 2006), which relies on a communication model that cannot be readily implemented in practice, as it does not capture important aspects of asynchronous communications such as non-instantaneous computations and communications. We have two main contributions. 1) We introduce a new communication scheme, based on *Loss-Networks*, that is programmable in a fully asynchronous and decentralized fashion. We establish empirically and theoretically that it improves over existing synchronous algorithms by depending on local communication delays in the analysis instead of global worst-ones. 2) We provide an acceleration of the standard gossip algorithm in the historical asynchronous model without requiring any additional synchronization.

1 INTRODUCTION

A broad cast of problems require to find the minimizer of a certain function in order to compute an estimator. Often, this function takes the form $f(x) = \sum_{i=1}^{n} f_i(x)$, where $x \in \mathbb{R}^d$ is the variable to optimize over and each f_i depends on a subset of the data. In this paper we study the case where data are distributed among the nodes of a known communication network. Tang et al. (2018); Zhang et al. (2013); Boyd et al. (2011); Scaman et al. (2017); Nedich et al. (2016); Sun et al. (2018) present different approaches for this problem, involving gossip communications and first order local gradient steps. Our work focuses on the relaxation of synchrony in these distributed algorithms: we aim at improving rates of convergence when the communication graph has high fluctuations in terms of delays, while preserving the same speed when all delays are of the same magnitude.

1.1 Gossip Algorithms and Asynchrony

In gossip averaging algorithms (Boyd et al., 2006; Dimakis et al., 2010), nodes of the network communicate with their neighbors without any central coordinator in order to compute the global average of local vectors. These algorithms are of interest to us, as they can be generalized to our distributed optimization problem, where nodes of the network possess a local function f_i. Two types of gossip algorithms appear in the literature: synchronous ones, where all nodes communicate with each other simultaneously (Scaman et al., 2017; Dimakis et al., 2010; Berthier et al., 2018), and asynchronous ones also called randomized gossip (Boyd et al., 2006; Nedic and Ozdaglar, 2009; Hendrikx et al., 2018), where at a defined time $t \geq 0$, only a pair of adjacent nodes can communicate. In the synchronous framework, the communication speed is limited by the slowest node. This paper aims at developing asynchronous algorithms that alleviate this issue. Our focus on asynchrony is motivated by empirical execution speed: we build a framework for the

analysis of asynchronous algorithms in order to show their efficiency over synchronous ones. Such a construction enables us to extract the quantities of interest, giving us a better understanding of the communication network and the quantities at stake, while being programmable in a fully asynchronous and distributed way.

In the historical asynchronous model (Boyd et al., 2006), each edge (ij) of the network has a local clock that ticks at a Poisson rate of intensity $p_{ij} > 0$ (Klenke, 2014). When clock (ij) ticks, nodes i and j communicate. This model is referred to as the *Poisson point process (P.p.p.) model*. Although qualified as so, this model cannot be programmed in a fully distributed and asynchronous structure: it assumes that communications and computations are made instantly. This modelling issue can be dealt with using two different approaches: (i) when a node i receives information from a neighbor j at a time $t \geq 0$, assume that this information is delayed, or (ii) forbid communications with a *busy* (*i.e.* communicating or computing) edge to avoid delayed information. These two modellings of asynchrony are respectively inspired by existing works done in an asynchronous but centralized framework with *perturbed iterates* for (i) (Leblond et al., 2016; Niu et al., 2011), and (ii) by *Loss-Networks*, initially considered for telecommunication networks (Kelly, 1991), yet also adequate to reflect primitives in distributed computing such as *locks* and *atomic transactions*. In the *perturbed iterate* modelling, a central unit delegates computations to workers. Asynchrony lies in the fact that these workers do not wait for the central unit to perform updates on the model: they send computed gradients whenever they can. In order to update the parameter on the central unit, the steps available are thus perturbed (*delayed*) gradients (Mania et al., 2015). Our work focuses on the second modelling: nodes behave as in the *P.p.p. model*, but are made *busy* and hence non-available for other nodes for a time $\tau_{ij} > 0$ after their activation. The system is asynchronous since it does not rely on global coordination and nodes do not wait for specific neighbours, but received gradients are never out of date since communicating and computing nodes are made *busy*.

1.2 Acceleration in an Asynchronous Setting

Our second main contribution is introducing a new accelerated gossip algorithm, the first of its kind in the historical *P.p.p. model*. Acceleration means gaining order of magnitudes in terms of convergence speed, compared to classical algorithms. Accelerating gossip algorithms has been studied in previous works in the synchronous framework: *SSDA* (Scaman et al., 2017), Chebyshev acceleration (Montijano et al., 2011) Jacobi-Polynomial acceleration in the first iterations (Berthier et al., 2018)), or in the asynchronous *P.p.p. model*: Geographic Gossip (Dimakis et al., 2008) , shift registers (Liu et al., 2013). However, no algorithm in the *P.p.p. model* gets a provably accelerated rate for general graphs. Inspired by *ACDM* (Nesterov and Stich, 2017a), Hendrikx et al. (2018) introduced *ESDACD* where at each iteration, only a pair of adjacent nodes communicate, but all nodes need to make local contractions and thus know that an update is taking place somewhere else in the graph. This last fact, also present in *Stochastic Heavy Balls* methods (Loizou and Richtárik, 2018), makes their method inapplicable in the *P.p.p. model*. Section 3 presents a continuous alternative to *ACDM* in the *P.p.p. model*, where the contractions previously cited are made continuously. Our algorithm (*CACDM*, for Continuously Accelerated Coordinate Descent Method) gets the same accelerated rate as Dimakis et al. (2008); Loizou and Richtárik (2018); Hendrikx et al. (2018) for any graph, without assuming access to any global counter. Although our analysis of *CACDM* does not extend to more general communication models such as those presented in Section 2, *CACDM* improves empirically over non-accelerated gossip in the Loss-Network model.

1.3 Problem Formulation and Notations

The communication network is represented by an undirected graph $G = (V, E)$ on the set of nodes $V = [n]$, and is assumed to be connected. Two nodes are said to be neighbors in the graph,

and we write $i \sim j$, if $(ij) \in E$. Each node $i \in V$ has access to a local function f_i defined on $\mathbb{R}^d$, L_i-smooth and σ_i-strongly convex (Bubeck, 2014), *i.e.* $\forall x, y \in \mathbb{R}^d$:

$$f_i(x) \leq f_i(y) + \langle \nabla f_i(y), x - y \rangle + \frac{L_i}{2}||x - y||^2,$$
$$f_i(x) \geq f_i(y) + \langle \nabla f_i(y), x - y \rangle + \frac{\sigma_i}{2}||x - y||^2.$$

Let us denote $f(z) = \sum_{i \in [n]} f_i(z)$ for $z \in \mathbb{R}^d$, $F(x) = \sum_{i \in [n]} f_i(x_i)$ for $x \in \mathbb{R}^{n \times d}$ the *augmented problem* where $x_i \in \mathbb{R}^d$ is stacked at node i, $L_{\max} = \max_i L_i$ and $\sigma_{\min} = \min_i \sigma_i$ the global complexity numbers. Computing gradients and communicating them between two neighboring nodes $i \sim j$ is assumed to take time $\tau_{ij} > 0$. This constant takes into account both the communication and computation times.
The problem can be formulated as follows:

$$\min_{x \in \mathbb{R}^{n \times d} : x_1 = ... = x_n} F(x), \tag{1.1}$$

where $x_1 = ... = x_n$ enforces consensus on all the nodes. We add the following structural constraints:

(i) *Local computations:* node i can compute first-order characteristics, such as ∇f_i or ∇f_i^*;

(ii) *Local communications:* node i can send information only to neighboring nodes $j \sim i$.

These operations may be performed asynchronously and in parallel, and each node possesses a local version $x_i \in \mathbb{R}^d$ of the global parameter x. The rate of convergence of our algorithms will be controlled by the smallest positive eigenvalue γ of the Laplacian of graph G (Mohar et al., 1991), weighted by some constants ν_{ij} that will depend on the local communication and computation delays. γ is non-decreasing in every parameter ν_{ij}, a result proved in the appendix.

Definition 1 (Graph Laplacian)**.** *Let $(\nu_{ij})_{(ij) \in E}$ a set of non-negative real numbers. The Laplacian of the graph G weighted by the ν_{ij}'s is the matrix with entries $-\nu_{ij}$ for $(ij) \in E$, $\sum_{j \sim i} \nu_{ij}$ for (ii) and 0 otherwise. In this paper, the notation ν_{ij} will stand for the weights of the Laplacian. This matrix is symmetric and non-negative. We denote $\gamma(\nu_{ij})$ its smallest non-null eigenvalue.*

For $g : \mathbb{R}^p \to \mathbb{R}$ some function, we define its *Fenchel conjugate* on $\mathbb{R}^p$, noted g^*, by:

$$\forall y \in \mathbb{R}^p, g^*(y) = \sup_{x \in \mathbb{R}^p} \langle x, y \rangle - g(x) \in \mathbb{R} \cup \{+\infty\}.$$

1.4 Quantitative Motivations for Asynchrony

Synchronous Communication Cost: In Synchronous Gossip Algorithm iterations (Dimakis et al., 2010), all nodes update their value synchronously by taking a weighted average of their neighbors. Their linear rate of convergence is given by the smallest eigenvalue of the Laplacian of the graph weighted by weights $\nu_{ij} \leq 1$. Every iteration taking a time τ_{max}, synchronous Gossip algorithms have an exponential rate of convergence $\gamma_{synch} = \gamma(\nu_{ij})$ with weights $\nu_{ij} \leq \tau_{max}^{-1}$ for all $(ij) \in E$ (Definition 1). More precisely, if $x(t) \in \mathbb{R}^{n \times d}$ is the vector stacked on the nodes $(x_i(t)$ at node $i)$ and $\bar{c}$ the consensus ($\frac{1}{n} \sum_i x_i(0)$ at each node), for $t \in \mathbb{R}^+$, we have:

$$||x(t) - \bar{c}||^2 \leq \exp(-(t - \tau_{\max})\gamma_{synch})||x(0) - \bar{c}||^2. \tag{1.2}$$

Asynchronous Cost in the *P.p.p. model*: the continuous rate of convergence γ_{asynch} is $\gamma(\nu_{ij})$ with weights $\nu_{ij} = p_{ij}$. For all $t \in \mathbb{R}^+$:

$$\mathbb{E}[||x(t) - \bar{c}||^2] \leq \exp(-t\gamma_{asynch})||x(0) - \bar{c}||^2. \tag{1.3}$$

Proofs of (1.2) and (1.3) can be found in Appendix A. (1.3) uses ideas from Boyd et al. (2006), combined with a study of infinitesimal intervals of times $[t, t+dt]$. Working with infinitesimal increments dt and continuous time leads to a more elegant formulation of the continuous time bound than previous works. Moreover, this *continuous increments trick* is a key idea for accelerating gossip in the *P.p.p. model* (Section 3).

Since the p_{ij}'s are expected to be of order τ_{ij}^{-1}, the asynchronous speed-up is quantitatively translated in the Laplacian of the graph, by taking local weights τ_{ij}^{-1} instead of the global worst-case one τ_{max}^{-1}. Intuitively, $\nu_{ij}dt$ for edge ij symbolizes the flow of information that can be sent in an infinitesimal interval of time dt through this edge. This explains the importance to have local constraints and weights in the Laplacian, instead of worst and global ones.

In Section 2, we present our new asynchronous communication scheme, its analysis and empirical results. Then in Section 3, we introduce *CACDM*, an accelerated Gossip Algorithm in the *P.p.p. model*. Due to space limitations, proofs are deferred to the appendix.

2 GOSSIP ALGORITHMS IN LOSS-NETWORK MODELS

We first introduce our new communication schemes (Section 2.1) and the related optimization algorithm. Then, the convergence bound and empirical results are presented (Section 2.2). Material for the analysis and intuition behind the algorithm is then provided (Section 2.3): a dual formulation of the problem, and general Theorems. We believe these results to be of independent interest, as they can be used for any communication scheme that involves pairwise operations.

2.1 Loss-Network Communication Scheme

The *P.p.p. model*, qualified as asynchronous, helps us understand quantitatively why asynchronous algorithms can outperform synchronous ones, but it assumes that communications and computations are done instantly. To alleviate this issue, we forbid communications between *busy* nodes. Our model is inspired from classical Loss Network models (Kelly, 1991). In this model, edges are activated following the same procedure as in the *P.p.p. model*, with processes of intensity p_{ij} (tuned in Section 2.2). Each node has an exponential clock of intensity $\frac{1}{2}\sum_{j\sim i} p_{ij}$. At each clock-ticking, if i is not busy, it selects a neighbor j with probability $p_{ij}/\sum_{k\sim i} p_{ik}$. i first checks if j is currently *busy*, an operation that takes a time $\varepsilon\tau_{ij}$ for some *small* $\varepsilon > 0$ ($\varepsilon \ll 1$ if sending a simple request if much faster than sending a whole vector). If j is not *busy*, i and j can compute and exchange information, becoming busy for a duration τ_{ij}. We can think of this procedure as classical gossip on an underlying random graph (Figure 1), that follows a Markov-Chain process if we extend the space of states with the inactivation time. We call our model the ***Refined Loss Network Model of parameter*** ε ***(RLNM(*** ε***))***. It is *refined* as the operation that consists in checking on its neighbors is not present in classical Loss Networks.

2.2 Algorithm and Main Theorem

Asynchronous gossip on the Refined Loss-Network communication model runs as follows: given local delays τ_{ij} defined in Section 1.3, each node has a local clock and a *Poisson Point Process* of intensity $\frac{1}{2}\sum_{j\sim i} p_{ij}$, where, with d_i the degree of node i and $\tau_{\max}(ij) = \max_{kl\sim ij}\tau_{kl}$:

$$p_{ij} = \min\left(\frac{1}{\tau_{\max}(ij)}, \frac{1}{2(\max(d_i,d_j)-1)}\frac{1}{\tau_{ij}}\right). \tag{2.1}$$

Let $I = \sum_{ij\in E} p_{ij}$ the global activation intensity.

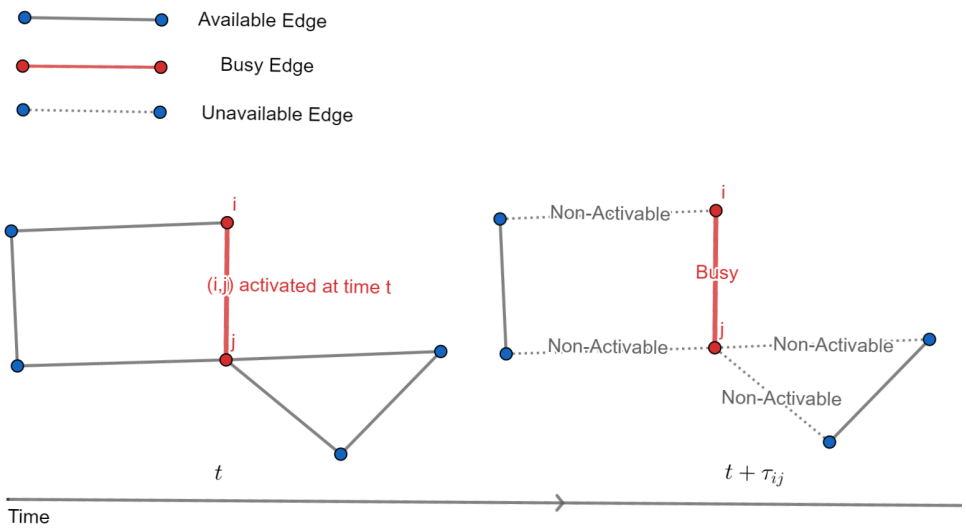


Figure 1: Underlying Markov Process for the Graph: edge (ij) activated at time t implies that while ij busy *i.e.* between times t and $t + \tau_{ij}$, all edges kl adjacent to ij are unavailable.

1) "Busy-Checking" Operation: when clock i rings at time t, select $j \sim i$ with probability $\frac{p_{ij}}{\sum_{k\sim i} p_{ki}}$ and check if j is currently busy. Operation makes i busy for a timelapse of length $\varepsilon\tau_{ij}$.

2) Gradient Exchange: if chosen neighbor j is available, make both nodes busy for a time τ_{ij}, and i sends $\nabla f_i^*(x_i)$ to j (and reciprocally).

3) Gradient Step: when i receives gradient $\nabla f_j^*(x_j)$ from j, it updates its local value x_i by a gradient step:

$$x_i \overset{t}{\leftarrow} x_i - \frac{\nabla f_i^*(x_j) - \nabla f_j^*(x_i)}{\sigma_i^{-1} + \sigma_j^{-1}}. \tag{2.2}$$

The desired output at node i at time t is then $\nabla f_i^*(x_i)$. Note that in the gossip averaging problem, these operations are equivalent to local averagings. Operations 2) and 3) both happen in the timelapse of length τ_{ij}, leading to no asynchrony issues and no delayed gradients. Define the following constants:

$$\begin{aligned} \tilde{\tau}_{ij} &= (1+\varepsilon)p_{ij}^{-1} \\ \tilde{\tau}_{\max} &= \max_{(ij)\in E} \tilde{\tau}_{ij} \\ T &= \frac{2\log(6|E|)}{\log(1-(1-e^{-1})e^{-1})} I\tilde{\tau}_{\max}. \end{aligned}$$

Define for $k \in \mathbb{N}$, $\mathcal{E}_k = \|x_{t_k} - \bar{x}^*\|^2$ the error to the consensus, where $\bar{x}^*$ is the minimizer of the augmented problem (1.1) and $t_k \in \mathbb{R}^+$ is the time of the k^{th} activation. Let, for $k \in \mathbb{N}$:

$$\mathcal{L}_k = \sum_{l=k}^{k+T-1} \mathcal{E}_l.$$

Theorem 1 (Discrete-time rate of convergence in the Loss-Network model)**.** *Let* $\Gamma_{RLNM} = \gamma(\nu_{ij})$ *(see Definition 1) with:*

$$\nu_{ij} = \alpha \times \frac{\tilde{\tau}_{ij}^{-1} \min_{(kl)\sim(ij)} \frac{\tilde{\tau}_{ij}}{\tilde{\tau}_{kl}}}{I d_{\max}^2 \left(\log(|E|) + \log(I\tilde{\tau}_{\max})\right)^2},$$

where $\alpha = \frac{32e^2}{\log(1-(1-e^{-1})e^{-1})^2}$ *is a universal constant. Then, for all* $k \in \mathbb{N}$*:*

$$\mathbb{E}[\mathcal{L}_k] \leq \left(\frac{1}{4}(1 - \frac{\sigma_{\min}}{L_{\max}}\Gamma_{RLNM})^{T/3} + \frac{3}{4}\right)^{\lceil \frac{k}{2T} \rceil} \mathbb{E}[\mathcal{L}_0].$$

Assumption 1 (Delay Constraints). *Let $\gamma_1 = \gamma(\nu_{ij})$ for $\nu_{ij} = 1, (ij) \in E$ (Definition 1). Assume that:*

$$\frac{\tau_{\max}}{\tau_{\min}} \leq \frac{L_{\max}}{\sigma_{\min}} \times \frac{\alpha d_{\max}^2 \log(|E|)}{\gamma_1}. \tag{2.3}$$

Notice that the right-hand side of (2.3) reflects the complexity of the optimization problem through the first factor (generally referred to as the condition number of the optimization problem), and the topology of the graph through γ_1 without the delays. The more difficult the problem is, the bigger the right-hand side is. Assumption 1 will then be verified more easily for graphs with slow mixing times (γ_1^{-1} bigger) and for complex local functions. The order of magnitude of γ_1^{-1} is n^2 for the grid, and n for the line or the cyclic graph. More generally, the right-hand side of (2.3) is always of order bigger than n.

Corollary 1 (Asymptotic Rate). *Under Assumption 1, Theorem 1 gives:*

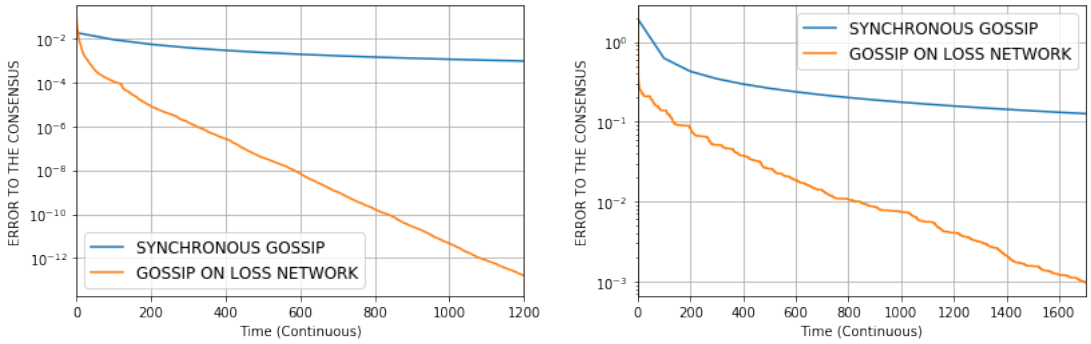
$$\limsup_{k\to\infty} \frac{1}{k} \log\left(\mathbb{E}[\mathcal{E}_k]\right) \leq -\frac{\sigma_{\min}}{L_{\max}} \times \frac{\Gamma_{RLNM}}{24e}.$$

Comments on the obtained rate of convergence: Theorem 1 and Corollary 1 are formulated in discrete time. The continuous exponential rate of convergence is obtained by multiplying the global *P.p.p.* intensity I, up to a constant factor of order 1. The factor $\frac{1}{I}$ is hence simply a normalization factor, due to a study in discrete time. As desired, the communication cost factor in the rate of convergence (Γ_{RLNM}) is captured by the Laplacian of the graph, weighted by *local* delays, instead of $\tau_{\max}^{-1}$. We however observe slowdowns due to other factors. 1) Having $\tilde{\tau}_{ij}$ instead of τ_{ij} (as in the *P.p.p. model* (1.3)) means that the effective waiting time of edge ij between two activations is of order $\tilde{\tau}_{ij}$ and no longer τ_{ij}, which was expected since p_{ij} is tuned accordingly. 2) Adding the factor $\min_{(kl)\sim(ij)} \frac{\tilde{\tau_{ij}}}{\tilde{\tau_{kl}}}$ to the local weight in the Laplacian is a local slowdown: a node with a slow neighbor becomes less effective. These first two remarks 1) and 2) lead to an interesting phenomenon: deleting some edges could improve the rate of convergence. A similar phenomenon occurs in road-trafficking (Bean et al., 1997; Steinberg and Zangwill, 1983), where deleting some roads can lead to more fluidity (*Braess's paradox*). 3) The global factor $\frac{1}{d_{\max}}$ is not intuitive at first: the more connected the graph is, the higher the rate should be. We hence have a trade-off between $\frac{1}{d_{\max}}$ that decreases when adding edges, and the smallest eigenvalue of the Laplacian of the graph Γ that increases with connectivity. We believe that $\frac{1}{d_{\max}}$ is an artifact of the proof, but acknowledge our difficulty in alleviating this factor. 4) If some nodes are *stragglers* (*i.e.* with high delays compared to the others), the rate of convergence stated for *RLNM* improves over synchronous algorithms, as it takes into account *local* delays. If the delays are all of the same order of magnitude, a case favourable to synchrony, the rate obtained is the same as in synchronous algorithms, up to a factor of order $\frac{1}{d^2 \log(n)}$. The log factor comes from exponential tails of our random variables.

Empirical results: In Figure 1, we modelled our Loss-Network scheme on two graphs: the circle with 50 nodes and the 2D-Grid with 225 nodes. In both cases, 10% of the nodes have a delay $\tau = 100$ time units, while the others have a delay equal to 1 time unit. These 10% are chosen uniformly at random. The local functions for the gossip problems are chosen as $f_i(x) = \|x - c_i\|^2$, with $c_0 = 1$ and $c_i = 0$ otherwise (worst case scenario in terms of mixing). We compare our algorithm on the Loss-Network to synchronous gossip. Time is indexed in a continuous way. Synchronous iterations are done every 100 units of time. The speed-up is significant when the fluctuation in term of delays in the graph is high, which illustrates the discussion at the end of Section 1.4.

2.3 Elements of Analysis

Section 2.3.1 introduces a classical dual formulation of the problem, while Section 2.3.2 provides general theorems that we derived in order to prove Theorem 1.



(a) 2D-Grid (b) Cyclic Graph

Figure 2: Asynchronous Speed-Up

2.3.1 Dual Formulation Of The Problem

A standard way to deal with the constraint $x_i = ... = x_n$, is to use a dual formulation (Scaman et al., 2017; Hendrikx et al., 2018; Uribe et al., 2020), by introducing a dual variable λ indexed by the edges. We first introduce a matrix $A \in \mathbb{R}^{n\times E}$ such that $\mathrm{Ker}(\mathrm{A}^\top) = \mathrm{Vect}(\mathbb{I})$ where $\mathbb{I}$ is the constant vector $(1, ..., 1)^\top$ of dimension n. A is chosen such that:

$$\forall (ij) \in E, Ae_{ij} = \mu_{ij}(e_i - e_j). \tag{2.4}$$

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

$$\begin{aligned}&\min_{x\in\mathbb{R}^{n\times d}, A^\top x=0} \sum_{i=1}^n f_i(x)\\ &= \min_{x\in\mathbb{R}^{n\times d}} \max_{\lambda\in\mathbb{R}^E} \sum_{i=1}^n f_i(x) - \langle A^\top x, \lambda\rangle.\end{aligned}$$

Let $F_A^*(\lambda) = F^*(A\lambda)$ for $\lambda \in \mathbb{R}^{E\times d}$ (notations in Section 1.3 for Fenchel conjugates). The dual problem reads:

$$\min_{x\in\mathbb{R}^{n\times d}, x_1=...=x_n} F(x) = \max_{\lambda\in\mathbb{R}^{E\times d}} -F_A^*(\lambda).$$

$F_A^*(\lambda) = \sum_{i=1}^n f_i^*((A\lambda)_i)$ is thus to be minimized over the dual variable $\lambda \in \mathbb{R}^{E\times d}$. The goal being to minimize F under consensus and local update constraints, a parallel needs to be made between minimization methods on the dual problem and on the primal one. As $F_A^*(\lambda) = \max_{x\in\mathbb{R}^{n\times d}} -F(x) + \langle A\lambda, x\rangle$, for any $\lambda \in \mathbb{R}^{E\times d}$ a primal variable $x \in \mathbb{R}^{n\times d}$ is uniquely associated through the formula $\nabla F(x) = A\lambda$. A local gradient on the dual variable λ alongside coordinate (ij) is:

$$\begin{aligned}\nabla_{ij} F_A^*(\lambda) &= (Ae_{ij})^\top \nabla F^*(A\lambda)\\ &= \mu_{ij}(\nabla f_i^*((A\lambda)_i) - \nabla f_j^*((A\lambda)_j)),\end{aligned}$$

The quantities $\nabla f_i^*((A\lambda)_i)$ are locally computable at each node, hence the choice of A made in (2.4). A gradient step on the dual variable λ alongside coordinate ij, when ij activated at iteration k (corresponding to time a t_k), is given by, where $U_{ij} = e_{ij}e_{ij}^\top$:

$$\lambda_{t_{k+1}} = \lambda_{t_k} - \frac{1}{(\sigma_i^{-1} + \sigma_j^{-1})\mu_{ij}^2} U_{ij}\nabla_{ij} F_A^*(\lambda_{t_k}).$$

Denoting $y_k = A\lambda_{t_k} \in \mathbb{R}^{n\times d}$, we get the following (local) formula, when ij activated:

$$y_{k+1,i} = y_{k,i} - \frac{\nabla f_i^*(y_{k,i}) - \nabla f_j^*(y_{k,j})}{\sigma_i^{-1} + \sigma_j^{-1}}, \tag{2.5}$$

$$y_{k+1,j} = y_{k,j} + \frac{\nabla f_i^*(y_{k,i}) - \nabla f_j^*(y_{k,j})}{\sigma_i^{-1} + \sigma_j^{-1}}. \tag{2.6}$$

While $\lambda \in \mathbb{R}^{E\times d}$ is a dual variable on the edge, $y \in \mathbb{R}^{n\times d}$ is still a dual variable, but on the nodes. The primal surrogate of y is $x = \nabla F^*(y)$ *i.e.* $x_i = \nabla f_i^*(y_i)$ at node i, that can hence be computed with local updates on y ((2.5) and (2.6)). A dual formulation hence enabled us to derive *local* updates on the primal problem out of simple coordinate gradient descent updates on the dual problem.

2.3.2 General Theorems

We analyze communication schemes that are defined through *edge activation processes*: each edge (ij) has a *Point Process* $\mathcal{P}_{ij} \subset \mathbb{R}^+$ that defines activation times of (ij). This is a generalization of both *P.p.p. model* and *RLNM*. When an edge is activated, the same update is performed as in (2.2) at nodes i and j. The delay of an edge is defined as its (random) waiting time between two activations. Two ergodic 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. In this section, we will work in discrete time. More precisely, discrete time $t \in \mathbb{N}$ stands for the t^{th} edge activation.

Definition 2 (Quantities of interest)**.** *In what follows, $t = 0, 1, 2, ...$ denotes the consecutive edge activations. Let $s \in \mathbb{N}$, ij and $kl \in E$. Let $s_{ij} < t_{ij}$ such that $s_{ij} \leq s < t_{ij}$ consecutive activation times (in discrete time) of ij. Denote $T_{ij}(s) = t_{ij} - s_{ij} - 1$ the total number of edge activations between the two consecutive activations of ij. Denote $N(kl, ij, s)$ the number of activations of edge kl in the activations $\{s_{ij}, s_{ij} + 1, ..., t_{ij} - 1\}$.*

Assumption 2 (Delay Assumptions)**.** *There exist $T \in \mathbb{N}^*$, $a, b > 0$, and $L_{ij} > 0, ij \in E$ such that:*

(i) *For all $t \in \mathbb{N}$, all edges are activated between times t and $t + T - 1$.*
(ii) *$\forall s \geq 0, \forall (ij) \in E, T_{ij}(s) \leq aL_{ij}$: ij is activated at least every aL_{ij} activations.*
(iii) *$\forall s \geq 0, \forall (ij), (kl) \in E$ such that $(kl) \sim (ij)$, $N(kl, ij, s) \leq \lceil \frac{bL_{ij}}{L_{kl}} \rceil$.*

Assumptions (i) and (ii) are a control over the inactivation period of all edges (the first one being a global one, the second one being a local version), while (iii) controls the local variance in terms of activation delays. The key technical difficulty lies in the fact that at a defined activation time, not all edges are available, meaning that when performing our coordinate gradient step on the dual variable, some coordinates are missing, as in *Markov-Chain Coordinate Gradient Descent* (Sun et al., 2018). The Lyapunov function Λ_t we study aims at alleviating this issue, by taking the value at T consecutive activation times. It is defined as follows on the dual variable:

$$\forall t \in \mathbb{N}, \Lambda_t = \frac{1}{T} \sum_{s=t}^{t+T-1} F_A^*(\lambda_s) - F_A^*(\lambda^*).$$

Note that a continuous analog is introduced by Fridman (2001) to study dynamic systems with time delay.

Theorem 2. *Assume that Assumption 2 holds for our edge-activation process. Let γ be the smallest positive eigenvalue of the Laplacian of the graph with:*

$$\nu_{ij} = CL_{ij}^{-1} \min_{kl\sim ij} \frac{L_{kl}}{L_{ij}},$$

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

$$\Lambda_t \leq \left(1 - \frac{\sigma_{\min}}{L_{\max}} \times \gamma\right)^t \Lambda_0.$$

In order to deal with stochasticity in the activation delays, we present another version of the theorem, that relies on the same properties. Define $(\mathcal{F}_s)_{s\geq 0}$ the filtration induced by the activation processes on the edges. If t_k is the k^{th} activation time, we define $(\mathcal{F}_{t_k})_{k\in\mathbb{N}}$, and when there is no doubt wether we work in continuous or discrete time, we write $(\mathcal{F}_k)$ or even $(\mathcal{F}_t)$ in what follows. Theorem 1 is obtained by applying the following result, with adequate constants.

Theorem 3 (Adding Stochasticity)**.** *Assume that, for all* $t \in \mathbb{N}$*, there exists a* $\mathcal{F}_{t+T-1}$*-measurable event* A_t*, such that* $\boldsymbol{p}(A_t|\mathcal{F}_t) \geq \frac{1}{2}$ *almost surely, and that under* A_t*, Assumption 2 holds for* $t \leq s \leq t+T-1$*. Then, we have the following bound on* L_t*, :*

$$\mathbb{E}[\Lambda_t] \leq \left(\frac{1}{4}(1 - \frac{\sigma_{\min}}{L_{\max}}\gamma)^{T/3} + \frac{3}{4}\right)^{\lceil \frac{t}{2T} \rceil} \mathbb{E}[\Lambda_0].$$

Theorems 2 and 3 are proved in Appendix B.Then, Appendix C applies these theorems to our Loss-Network model. Next section presents an accelerated gossip algorithm, in order to improve obtained communication rates.

3 ACCELERATED GOSSIP ALGORITHM

Inspired by previous works (Nesterov and Stich, 2017a; Hendrikx et al., 2018), we propose an accelerated gossip algorithm. We prove a rigorous accelerated rate of convergence (Theorem 4) for this algorithm in the historical *P.p.p. model*. Applying our algorithm to *RLNM* communication schemes lead to an empirical accelerated rate of convergence. We call our algorithm *CACDM* (Continuously Accelerated Coordinate Descent Method). Edge activations are ruled by local independent *P.p.p.* of intensity p_{ij} for edge (ij). We denote $I = \sum_{ij} p_{ij}$ the global activation intensity.

3.1 The *CACDM* algorithm

Time $t \in \mathbb{R}^+$ is indexed continuously. Our algorithm involves two different types of operations: continuous contractions and local updates when a *P.p.p.* ticks. Node i needs to stack two vectors $x_{i,t}, y_{i,t} \in \mathbb{R}^d$, $y_{i,t}$ being the *momentum* variable. These variables are dual variables on the nodes, introduced in Section 2.3.1.

1) Continuous Contractions: For all times $t \in \mathbb{R}^+$ and for some fixed $\theta > 0$ to determine, make the infinitesimal contraction

$$\begin{pmatrix} x_{i,t+dt} \\ y_{i,t+dt} \end{pmatrix} = \begin{pmatrix} 1 - dtI\theta & dtI\theta \\ dtI\theta & 1 - dtI\theta \end{pmatrix} \begin{pmatrix} x_{i,t} \\ y_{i,t} \end{pmatrix},$$

between times t and $t + dt$. Between times $s < t$, if there is no activation of i, it consists in performing the contraction:

$$\begin{pmatrix} x_{i,t} \\ y_{i,t} \end{pmatrix} = \exp\left((t-s)I \begin{pmatrix} -\theta & \theta \\ \theta & -\theta \end{pmatrix}\right) \begin{pmatrix} x_{i,s} \\ y_{i,s} \end{pmatrix}, \tag{3.1}$$

where we have:

$$\exp\left(tI \begin{pmatrix} -\theta & \theta \\ \theta & -\theta \end{pmatrix}\right) = \begin{pmatrix} \frac{1+e^{-2I\theta t}}{2} & \frac{1-e^{-2I\theta t}}{2} \\ \frac{1-e^{-2I\theta t}}{2} & \frac{1+e^{-2I\theta t}}{2} \end{pmatrix}.$$

2) Local Updates: When edge (ij) is activated at time $t \geq 0$, perform the local update between nodes i and j, where σ_A is the strong convexity parameter of F_A^* on the orthogonal of Ker(A):

$$x_{i,t} \overset{t}{\leftarrow} x_{i,t} - \frac{\nabla f_i^*(x_t(i)) - \nabla f_j^*(x_t(j))}{\sigma_i^{-1} + \sigma_j^{-1}}, \tag{3.2}$$

$$y_{i,t} \overset{t}{\leftarrow} y_{i,t} - \frac{\theta}{\sigma_A}\left(\nabla f_i^*(x_t(i)) - \nabla f_j^*(x_t(j)\right). \tag{3.3}$$

The desired output at node i and at time t is then $\nabla f_i^*(x_{i,t})$ (Section 2.3.1).
More formally, the stochastic process defined above is the following, where $X_t = (x_t, y_t)^T$, $\eta_{ij,t}$ is the gradient step on coordinates i, j as in (3.2) and (3.3), and N_{ij} are independent *P.p.p.* of intensities p_{ij}:

$$dX_t = I \begin{pmatrix} -\theta & \theta \\ \theta & -\theta \end{pmatrix} X_t dt + \sum_{(ij)\in E} dN_{ij}(t)\eta_{ij,t}.$$

This procedure is an asynchronous one: the length $t - s$ between two activations of an edge that appears in the exponential contraction (3.1) is a local variable, and only needs a local clock to be computed.

3.2 Convergence Theorem for *CACDM*

The convergence theorem involves dual variables $\lambda_t, \omega_t \in \mathbb{R}^{E\times d}$ (Section 2.3.1), respectively edge conjugates of x_t and y_t (*i.e.* $A\lambda_t = x_t$ and $A\omega_t = y_t$). Denote:

$$L_t = \|\omega_t - \lambda^*\|^{*2} + \frac{2\theta^2 S^2}{\sigma_A^2}\left(F_A^*(\lambda_t) - F_A^*(\lambda^*)\right)$$

the Lyapunov function we study, with λ^* an optimizer of F_A^*, $\theta, S > 0$ to be defined, and $\|.\|^*$ the euclidian norm on the orthogonal of Ker(A). Matrix A (2.4) is tuned with $\mu_{ij}^2 = \frac{p_{ij}}{I}$ where $I = \sum_{kl} p_{kl}$.

Theorem 4. *For the* CACDM *algorithm, if* $\theta = \sqrt{\frac{\sigma_A}{IS^2}}$ *for* S *verifying the inequality* $S^2 \geq \sup_{(ij)\in E} \frac{e_{ij}^T A^* A e_{ij}(\sigma_i^{-1}+\sigma_j^{-1})}{2p_{ij}/I}$, *we have for all* $t \in \mathbb{R}^+$*:*

$$\mathbb{E}[L_t] \leq L_0 e^{-I\theta t}.$$

Proof of this theorem (Appendix D) uses ideas introduced for (1.3) (study of intervals $[t, t + dt]$ with infinitesimal increments dt), combined with inequalities of the same type as Nesterov and Stich (2017a). Note that although formulated in terms of dual variables, the exponential rate of convergence still applies for primal variables.

Remarks on the bound: σ_A is the strong convexity parameter of F_A^*. It can be lower-bounded by $\frac{\gamma_{\text{asynch}}/I}{L_{\max}}$, where γ_{asynch} is the smallest eigenvalue of the laplacian of the graph weighted by $\nu_{ij} = p_{ij}$ (non accelerated *P.p.p.* rate of convergence). It is divided by I so that the entries p_{ij} sum to 1. If there exists a constant $c > 0$ such that:

$$\forall (ij) \in E, \frac{p_{ij}}{I} \geq \frac{c}{|E|},$$

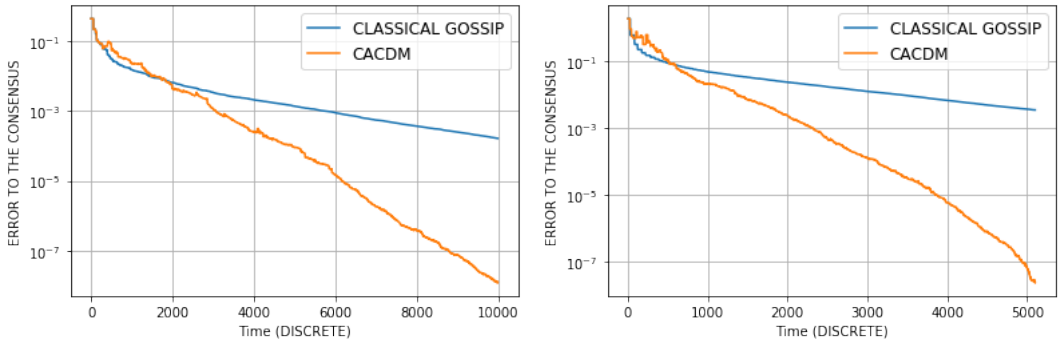
then $S^2 \geq \sigma_{\min}^{-1}|E|/c$, leading to the following rate of convergence:

$$I \times \sqrt{c\frac{\sigma_{\min}}{L_{\max}} \times \frac{\gamma_{\text{asynch}}}{I|E|}}.$$

Taking $I = 1$ (re-normalizing time) and the simple averaging problem, leads to an improved rate n^{-2} on the line graph instead of n^{-3}. For the 2D-Grid, we have $n^{-3/2}$ instead of n^{-2}. However, there is no improvement on the complete graph ($1/n$ instead of $1/n$). These rates are the same as Dimakis et al. (2008); Hendrikx et al. (2018); Loizou and Richtárik (2018). However, our algorithm does not require to know the number of activations performed on the whole network, and only requires local clocks. Moreover, similarly to Hendrikx et al. (2018), it works for any graph and for the more general problem of distributed optimization of smooth and strongly convex functions provided dual gradients of local functions are computable.

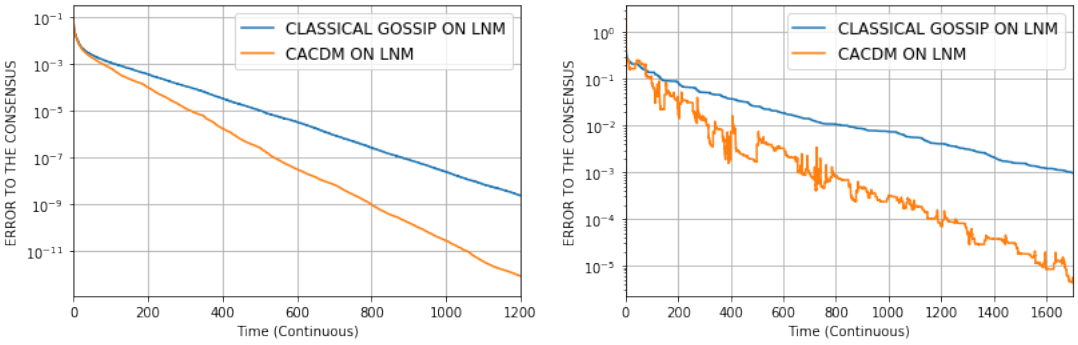
Empirical Results: in Figure 3, the setting is the same as for Figure 2, in order to compare the *CACDM* acceleration on the *P.p.p. model*, with the classical gossip in the *P.p.p. model*. Time is indexed in a discrete way.

Acceleration in *RLNM*(ε): The analysis of *CACDM* does not extend to more general models than the *P.p.p. model*. However, applying it to *RLNM* leads to an accelerated rate of convergence in Figure 4, showing us that our algorithm is quite robust to changes in edge activation statistics. In order to tune the algorithm, we take values p_{ij} as in (2.1). Time is indexed in a continuous way. 1000 units of time hence correspond to approximately $I \times 1000 \approx 10^5 - 10^6$ edge activations.



(a) 2D-Grid (b) Cyclic Graph

Figure 3: *CACDM* Vs Gossip in the *P.p.p. Model*



(a) 2D-Grid (b) Cyclic Graph

Figure 4: *CACDM* vs Gossip in **LNM**

4 CONCLUSION

This paper introduces a new way to deal with asynchrony in distributed optimization, leading to improved rates of convergence in terms of communication compared to synchronous algorithms, determined by local communication delays instead of worst-case ones. The generality of our framework makes it possible to consider any asynchronous communication scheme on the graph for the analysis. We highlighted quantities such as local graph degrees or local fluctuation in terms of communication delays, that seem to be involved in the execution speed of our asynchronous gossip. An interesting problem would be to study the optimality of our communication scheme. If it held, we would have in hand necessary tools in order to construct an optimal communication network knowing the delays between nodes. Finally, we proposed an accelerated gossip algorithm in the historical asynchronous gossip framework introduced by Boyd et al. (2006) in which nodes do not need need to know the global number of updates in the graph. We believe that both contributions pave the way for fast asynchronous gossip algorithms with theoretical guarantees. Yet, we leave the theory of acceleration in the loss network model as a hard but interesting open problem.

References

Assran, M., Aytekin, A., Feyzmahdavian, H., Johansson, M., and Rabbat, M. (2020). Advances in asynchronous parallel and distributed optimization.

Aybat, N. and Gürbüzbalaban, M. (2017). Decentralized computation of effective resistances and acceleration of consensus algorithms.

Bean, N. G., Kelly, F. P., and Taylor, P. G. (1997). Braess's paradox in a loss network. *Journal of Applied Probability*, 34(1):155–159.

Berthier, R., Bach, F., and Gaillard, P. (2018). Accelerated gossip in networks of given dimension using jacobi polynomial iterations.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122.

Bubeck, S. (2014). Convex optimization: Algorithms and complexity.

Colin, I., Bellet, A., Salmon, J., and Clémençon, S. (2016). Gossip Dual Averaging for Decentralized Optimization of Pairwise Functions. In *International Conference on Machine Learning (ICML 2016)*, New York, United States.

Dimakis, A. D. G., Sarwate, A. D., and Wainwright, M. J. (2008). Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3):1205–1216.

Dimakis, A. G., Kar, S., Moura, J. M. F., Rabbat, M. G., and Scaglione, A. (2010). Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864.

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

Fridman, E. (2001). New lyapunov–krasovskii functionals for stability of linear retarded and neutral type systems. *Systems & Control Letters*, 43(4):309 – 319.

Hannah, R., Feng, F., and Yin, W. (2018). A2bcd: An asynchronous accelerated block coordinate descent algorithm with optimal complexity.

Hendrikx, H., Bach, F., and Massoulié, L. (2018). Accelerated decentralized optimization with local updates for smooth and strongly convex objectives.

Kelly, F. P. (1991). Loss networks. *The Annals of Applied Probability*, 1(3):319–378.

Klenke, A. (2014). *The Poisson Point Process*, pages 543–561. Springer London, London.

Le Gall, J.-F. (2016). *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274.

Leblond, R., Pedregosa, F., and Lacoste-Julien, S. (2016). Asaga: Asynchronous parallel saga.

Lee, S. and Nedich, A. (2013). Asynchronous gossip-based random projection algorithms over networks.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017a). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent.

Lian, X., Zhang, W., Zhang, C., and Liu, J. (2017b). Asynchronous decentralized parallel stochastic gradient descent.

Liu, J., Anderson, B. D., Cao, M., and Morse, A. S. (2013). Analysis of accelerated gossip algorithms. *Automatica*, 49(4):873 – 883.

Liu, J. and Wright, S. J. (2014). Asynchronous stochastic coordinate descent: Parallelism and convergence properties.

Loizou, N. and Richtárik, P. (2018). Accelerated gossip via stochastic heavy ball method.

Mania, H., Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., and Jordan, M. I. (2015). Perturbed iterate analysis for asynchronous stochastic optimization.

Mohar, Alavi, Y., Chartrand, G., and Oellermann, O. (1991). The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*.

Mokhtari, A. and Ribeiro, A. (2016). Dsa: Decentralized double stochastic averaging gradient algorithm. *Journal of Machine Learning Research*, 17(61):1–35.

Montijano, E., Montijano, J., and Sagues, C. (2011). Chebyshev polynomials in distributed consensus applications. *IEEE Transactions on Signal Processing*, 61.

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

Nedich, A., Olshevsky, A., and Shi, W. (2016). Achieving geometric convergence for distributed optimization over time-varying graphs.

Nesterov, Y. and Stich, S. U. (2017a). Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123.

Nesterov, Y. and Stich, S. U. (2017b). Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123.

Niu, F., Recht, B., Re, C., and Wright, S. J. (2011). Hogwild!: A lock-free approach to parallelizing stochastic gradient descent.

Oreshkin, B. N., Coates, M. J., and Rabbat, M. G. (2010). Optimization and analysis of distributed averaging with short node memory. *IEEE Transactions on Signal Processing*, 58(5):2850–2865.

Pu, S., Shi, W., Xu, J., and Nedic, A. (2020). Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, pages 1–1.

Ram, S., Nedić, A., and Veeravalli, V. (2009). Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48th IEEE Conference on Decision and Control held jointly with 2009 28th Chinese Control Conference, CDC/CCC 2009*, Proceedings of the IEEE Conference on Decision and Control, pages 3581–3586. 48th IEEE Conference on Decision and Control held jointly with 2009 28th Chinese Control Conference, CDC/CCC 2009 ; Conference date: 15-12-2009 Through 18-12-2009.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks.

Steinberg, R. and Zangwill, W. I. (1983). The prevalence of braess' paradox. *Transportation Science*, 17(3):301–318.

Sun, T., Sun, Y., and Yin, W. (2018). On markov chain gradient descent.

Sundhar Ram, S., Nedic, A., and Veeravalli, V. V. (2009). Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3581–3586.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. (2018). D^2: Decentralized training over decentralized data.

Tanner, M. (1995). *Practical queueing analysis*. IBM McGraw-Hill. McGraw-Hill, London.

Uribe, C. A., Lee, S., Gasnikov, A., and Nedić, A. (2020). A dual approach for optimal algorithms in distributed optimization over networks.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(68):3321–3363.

Zheng, S., Meng, Q., Wang, T., Chen, W., Yu, N., Ma, Z.-M., and Liu, T.-Y. (2016). Asynchronous stochastic gradient descent with delay compensation.

- Appendix A presents known results on synchronous and asynchronous gossip algorithms. The asynchronous result is formulated in a different way than is usually done, with a study in continuous time. A result on the monotonicity of the second smallest eigenvalue of the Laplacian of the graph is also presented.
- Appendix B proves the two theorems of Section 2.3.2: first with the deterministic assumptions, then the stochastic version. Preliminary inequalities on the function F_A^* are also presented.
- Appendix C studies our Refined Loss Network model in order to derive Theorem 1.
- Appendix D proves the accelerated rate of convergence for *CACDM*.

A GOSSIP ALGORITHMS: SYNCHRONOUS AND ASYNCHRONOUS BOUNDS PROOFS

A.1 Synchronous Gossip

In the synchronous setting, all nodes are allowed to share a common clock, which enables them to perform operations synchronously. Formally, a *gossip matrix* is defined as follows:

Definition 3 (Gossip Matrix)**.** *A gossip matrix is a matrix $W \in \mathbb{R}^{n\times n}$ such that:*

- $\forall (i,j) \in [n]^2$, $W_{i,j} > 0 \implies i \sim j$ *or* $i = j$ *(supported by G),*
- $\forall i \in [n], \sum_{j\sim i} W_{i,j} = 1$ *(stochastic),*
- $\forall (i,j) \in [n]^2, W_{i,j} = W_{j,i}$ *(symmetric).*

Iteratively, at times $t = 0, 1, 2, ...$, if $x(t) = (x_i(t))_i \in \mathbb{R}^{n\times d}$ describes the information stacked locally at each node ($x_i(t)$ being the vector at node i), we perform the operation $x(t+1) = Wx(t)$. It is to be noted that, thanks to the sparsity of the gossip matrix, this operation is local: for all node i,

$$x_i(t+1) = \sum_{j\sim i} W_{ij} x_j(t), \tag{A.1}$$

where $i \sim j$ if they are neighbors or if $i = j$. The convergence bound will be stated below. Intuitively, at each iteration, each node i sends a proportion of its mass to each one of its neighbour, the condition $\sum_{j\sim i} W_{ij} = 1$ being the mass conservation.

Proposition 1 (Synchronous Gossip)**.** *Let γ_W be the eigengap of the laplacian of G weighted by W_{ij} at each edge. Then, for all $k = 0, 1, 2...$:*

$$||x(k) - \bar{c}|| \leq (1 - \gamma_W)^k ||c - \bar{c}||, \tag{A.2}$$

where $x(0) = c$, and $\bar{c}$ is when consensus is reached

Proof. For $k \geq 0$,

$$\begin{aligned} &x(k+1) - \bar{c} = W(x(k) - \bar{c}) \\ &\implies ||x(k+1) - \bar{c}|| \leq \lambda_2(W)||x(k) - \bar{c}||, \end{aligned}$$

where λ_2 is the second largest eigenvalue of W, 1 being the largest (W is stochastic symmetric), and $\bar{c}$ being in the corresponding eigenspace. We conclude by saying that $\lambda_2(W) = 1 - \gamma_W$ where γ_W is the smallest non null eigenvalue of $Id - W$. Notice that $Id - W$ is the laplacian of the graph weighted by $\nu_{ij} = 1 - W_{ij}$. □

Then, since every iteration takes a time τ_{max}, denoting time in a continuous way by $t \in \mathbb{R}^+$, we have:

$$||x(t) - \bar{c}|| \leq (1 - \gamma_W)^{t/\tau_{max} - 1} \leq \exp\left(-\frac{\gamma_W}{\tau_{max}}(t - \tau_{max})\right). \tag{A.3}$$

A.2 Asynchronous Gossip

Time is indexed in a continuous way, by $\mathbb{R}^+$. For every edge $e = (ij) \in E$, let $\mathcal{P}_{ij}$ be a Poisson point process (P.p.p.) of constant intensity $p_{ij} > 0$ that we will call "clocks", all independent from each other. Updates will be ruled by these processes: at every clock ticking of $\mathcal{P}_{ij}$, nodes i and j update the value they stack by the mean $\frac{x_i + x_j}{2}$. If we write $\mathcal{P} = \cup_{ij\in E} \mathcal{P}_{ij}$, $\mathcal{P}$ is a P.p.p. of intensity $I := \sum_{ij\in E} p_{ij}$.

Proposition 2 (Asynchronous Continuous Time Bound)**.** *Let $(x_t(i))_i$ be the vector stacked on the graph. Let σ_{asynch} be the smallest non null eigenvalue of the laplacian of the graph, weighted by the p_{ij}'s. For $t \geq 0$, we have:*

$$\mathbb{E}[||x(t) - \bar{c}||^2] \leq \exp(-t\sigma_{asynch})||c - \bar{c}||^2.$$

Proof. First, it is to be noted that, if $\mathcal{P}$ is a *P.p.p.* of intensity $\lambda > 0$, for all $t \in \mathbb{R}$ and $dt \to 0$:

$$\mathbb{P}([t, t+dt] \cap \mathcal{P} \neq \emptyset) = \lambda dt + o(dt). \tag{A.4}$$

When ij activated at time t, multiply $x(t)$ by $W_{ij} = I_n - \frac{^t(e_i - e_j)(e_i - e_j)}{2}$. By observing that $W_{ij}^2 = W_{ij}$ and that $\sum_{ij} p_{ij} W_{ij} = II_n - L$, where L is the laplacian of the graph weighted by the p_{ij}, we get that, with R_t^2 the squared error to the consensus at time t, up to a $o(dt)$:

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_t}[R_{t+dt}^2] =& (1 - Idt)\mathbb{E}^{\mathcal{F}_t}\left[R_{t+dt}^2 | \text{no activations in } [t, t+dt]\right] \\
&+ dt \sum_{ij} p_{ij} \mathbb{E}^{\mathcal{F}_t}\left[R_{t+dt}^2 | ij \text{ activated in } [t, t+dt]\right] + o(dt) \\
&= R_t^2 - dt(x(t) - \bar{c})^\top \sum_{ij} W_{ij}(x(t) - \bar{c}) \\
&\leq R_t^2 - dt\sigma_p R_t^2.
\end{aligned}$$

Then, taking the mean, dividing by $dt \to 0$ and integrating concudes the proof. □

A.3 Laplacian Monotonicity

We finish by proving the following intuitive result:

Proposition 3 (Monotonicity of the Laplacian)**.** *Let $\Lambda(\lambda_{ij}, (ij) \in E)$ be the laplacian of the graph weighted by λ_{ij}. Then, its second smallest eigenvalue σ is a non decreasing function of each weight λ_{ij}.*

Proof. First compute $\langle \Lambda u, u \rangle$, the weights λ_{ij} being fixed:

$$\begin{aligned}
\langle \Lambda u, u \rangle &= \sum_i \sum_{j \sim i} u_i(u_i - u_j)\lambda_{ij} \\
&= \frac{1}{2} \sum_i \sum_{j \sim i} (u_i - u_j)^2 \lambda_{ij}.
\end{aligned}$$

It appears that for any $u \in \mathbb{R}^n$, these are non decreasing quantities in each λ_{ij}. If we take Λ and Λ' two laplacians with weights $\lambda_{ij} \leq \lambda'_{ij}$, we get, for all $u \in \mathbb{R}^n$, $\langle \Lambda u, u \rangle \leq \langle \Lambda' u, u \rangle$. Then, using that $\sigma = \min_{||u||=1, \langle u, \mathbb{I} \rangle = 0} \langle \Lambda u, u \rangle$ (as $\mathbb{I}$ is a eigenvector associated to the eigenvalue 0), we have $\sigma' \leq \sigma$ the desired result. □

B GENERAL ASYNCHRONOUS COMMUNICATION SCHEMES: PROOF OF BOTH THEOREMS

B.1 Preliminary Inequalities

We first present preliminary inequalities using properties on our function F_A^*. These properties were also proven in **?** (except for Lemma 5) but we present them here for the paper to be self-contained.

Lemma 1. *For $\lambda \in \mathbb{R}^{E \times d}$ and $ij \in E$, we have:*

$$F_A^*\left(\lambda - \frac{1}{\mu_{ij}^2(\sigma_i^{-1} + \sigma_j^{-1})} U_{ij} \nabla_{ij} F_A^*(\lambda)\right) - F_A^*(\lambda) \leq -\frac{1}{2\mu_{ij}^2(\sigma_i^{-1} + \sigma_j^{-1})} ||\nabla_{ij} F_A^*(\lambda)||^2. \tag{B.1}$$

Proof. Let us define $h_{ij} = -\frac{1}{\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})} U_{ij}\nabla_{ij}F_A^*(\lambda)$.

$$\begin{aligned} F_A^*(\lambda + h_{ij}) - F_A^*(\lambda) &= \sum_k f_k^*((A\lambda)_k + (Ah_{ij})_k)) - f_k^*((A\lambda)_k) \\ &= f_i^*((A\lambda)_i + (Ah_{ij})_i) - f_i^*((A\lambda)_i) + f_j^*((A\lambda)_j + (Ah_{ij})_j) - f_j^*((A\lambda)_j), \end{aligned}$$

as (Ah_{ij}) is supported only by coordinates i and j. Moreover, as f_i^* is σ_i-smooth, we have:

$$f_i^*((A\lambda)_i + (Ah_{ij})_i) - f_i^*((A\lambda)_i) \leq \langle \nabla f_i^*((A\lambda)_i), (Ah_{ij})_i\rangle + \frac{\sigma_i^{-1}}{2}||(Ah_{ij})_i||^2,$$

and by summing for i and j and noticing that $(Ah_{ij})_i = \mu_{ij}\nabla_{ij}F_A^*(\lambda)$:

$$\begin{aligned} F_A^*(\lambda + h_{ij}) - F_A^*(\lambda) &\leq \langle \nabla_{ij}F_A(\lambda), h_{ij}\rangle + \frac{(\sigma_i^{-1}+\sigma_j^{-1})\mu_{ij}^2}{2}\left(\frac{1}{\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})}\right)^2 ||\nabla_{ij}F_A^*(\lambda)||^2 \\ &= -\frac{1}{2\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})}||\nabla_{ij}F_A^*(\lambda)||^2. \end{aligned}$$

□

Lemma 2. *σ_A the strong convexity parameter of F_A^* on the orthogonal of $Ker(A)$ is lower bounded by $\lambda_{min}^+(A^TA)/L_{max}$, where $\lambda_{min}^+(A^TA)$ is the smallest non null eigenvalue of A^TA.*

Proof. Let $\lambda, \lambda' \in \mathbb{R}^{E\times d}$. By L_i^{-1} and thus L_{max}^{-1}-strong convexity of f_i^*:

$$f_i^*((A\lambda)_i) - f_i^*((A\lambda')_j) \geq \langle \nabla f_i^*((A\lambda')_i), (A(\lambda - \lambda'))_i\rangle - \frac{1}{2L_{max}}||(A(\lambda - \lambda')||^2$$

Summing over all $i \in [n]$ and using $\nabla F_A^*(\lambda') =^t A(\nabla_i f_i^*((A\lambda')_i))_i$ leads to:

$$\begin{aligned} F_A^*(\lambda) - F_A^*(\lambda') &\geq \langle \nabla F_A^*(\lambda'), \lambda - \lambda'\rangle - \frac{1}{2L_{\max}}||A(\lambda' - \lambda||^2 \\ &\geq \langle \nabla F_A^*(\lambda'), \lambda - \lambda'\rangle - \frac{\lambda_{\min}^+(A^TA)}{2L_{max}}||\lambda - \lambda'||^{*2}. \end{aligned}$$

where $||.||^*$ is the euclidian norm on the orthogonal of $Ker(A)$. □

Lemma 3. *AA^T is the laplacian of the graph G weighted by μ_{ij}^2 on the edges.*

Proof.

$$A^Te_i = \sum_{j\sim i}\mu_{ij}e_{ij}$$

For the diagonal, we have:

$$\begin{aligned} e_iAA^Te_i &= \sum_{k\sim i}\sum_{l\sim i}\mu_{ik}\mu_{il}\langle e_{ik}, e_{il}\rangle \\ &= \sum_{j\sim i}\mu_{ij}^2. \end{aligned}$$

Then, for $i \sim j, i \neq j$:

$$\begin{aligned} e_iAA^Te_j &= \sum_{k\sim i}\sum_{l\sim i}\mu_{ik}\mu_{jl}\langle e_{ik}, e_{jl}\rangle \\ &= \mu_{ij}\mu_{ji} \\ &= -\mu_{ij}^2. \end{aligned}$$

□

Lemma 4. *For $x, x' \in R^{E\times d}$, and $ij \in E$, we have:*

$$||\nabla_{ij}F_A^*(x) - \nabla_{ij}F_A^*(x')||^2 \leq 2(\sigma_i^{-1} + \sigma_j^{-1})^2 d_{ij}\mu_{ij}^2 \sum_{(kl)\sim(ij)} \mu_{kl}^2 ||x_{kl} - x'_{kl}||^2. \tag{B.2}$$

Proof. First, notice that $\nabla_{ij}F_A^*(x) = \mu_{ij}(\nabla f_i^*((Ax)_i) - \nabla f_j^*((Ax)_j))$. Then:

$$\begin{aligned}
||\nabla f_i^*((Ax)_i) - \nabla f_i^*((Ax')_j)|| &\leq \sigma_i^{-1}||(A(x-x'))_i|| \text{ (smoothness)} \\
&\leq \sigma_i^{-1}|| \sum_{kl\sim ij} \mu_{kl}(x-x')_{kl}|| \\
&\leq \sigma_i^{-1} \sum_{kl\sim ij} \mu_{kl}||(x-x')_{kl}||
\end{aligned}$$

Conclude by taking the square and summing for i and j. □

Lemma 5 (Distance to Optimum)**.** *For any $\lambda \in \mathbb{R}^{E\times d}$ and for λ^* minimizing F_A^*, we have:*

$$F_A^*(\lambda) - F_A^*(\lambda^*) \leq \frac{1}{2\sigma_A}||\nabla F_A^*(\lambda)||^2 \tag{B.3}$$

Proof. We introduce Bregman divergences, which make the proof straightforward. For ϕ any real-valued function, differentiable, defined on an euclidian space $\mathcal{V}$, we define its Bregman divergence D_ϕ on $\mathcal{V}^2$ by:

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle. \tag{B.4}$$

ϕ is thus L-smooth if and only if $D_\phi \leq LD_{||.||^2/2}$. An important equality is the following, under convexity assumption for ϕ:

$$D_\phi(x, y) = D_{\phi^*}(\nabla\phi(y), \nabla\phi(x)). \tag{B.5}$$

Applying this to $\phi = F_A^*$, $x = \lambda, y = \lambda^*$, together with the fact that $(F_A^*)^*$ is σ_A^{-1}-smooth with respect to $||.||^{*2}$ (**?**), the squared norm on the orthogonal of $Ker(A)$ leads to:

$$D_{F_A^*}(\lambda, \lambda^*) = D_{F_A^{**}}(\nabla F_A^*(\lambda^*), \nabla F_A^*(\lambda)) \leq \frac{1}{\sigma_A} D_{||.||^{*2}/2}(\nabla F_A^*(\lambda^*), \nabla F_A^*(\lambda)),$$

and the result follows since $\nabla F_A^*(\lambda^*) = 0$ and $||\nabla F_A^*(\lambda)||^{*2} = ||\nabla F_A^*(\lambda)||^2$. □

B.2 Proof Of Theorem 2

To prove the theorem, we need to study every gradient step involved. At iteration s, not every coordinates is available, hence the need to study the impact of T gradient steps together. A gradient step alongside edge ij only involves edges in its neighborhood (thanks to the sparsity of the matrix A), a key element that will need to be explicited. The proof involves three main steps.

Step 1: Applying Lemma 1 (local smoothness) gives, where ij is the t^{th} activated edge:

$$F_A^*(\lambda(t+1)) - F_A^*(\lambda(t)) \leq -\frac{1}{2(\sigma_i^{-1} + \sigma_j^{-1})\mu_{ij}^2}||\nabla_{ij}F_A^*(\lambda(t))||^2. \tag{B.6}$$

Hence, we get an inequality between L_t and L_{t+1}:

$$\Lambda_{t+1} = \frac{1}{T}\sum_{t\leq s<t+T}(F_A^*(\lambda(s+1)) - F_A^*(\lambda^*)) \leq \Lambda_t - \frac{1}{T}\sum_{t\leq s<t+T} \frac{1}{2(\sigma_i^{-1} + \sigma_j^{-1})\mu_{(ij)_s}^2}||\nabla_{(ij)_s}F_A^*(\lambda(s))||^2 \tag{B.7}$$

where $(ij)_s$ is the edge activated during activation s. Let's introduce the following quantity:

$$\frac{1}{T}\sum_{t\leq s<t+T}\sum_{ij\in E}||\nabla_{ij}F_A^*(\lambda(s))||^2=\frac{1}{T}\sum_{t\leq s<t+T}||\nabla F_A^*(\lambda(s))||^2\geq\sigma_A\Lambda_t \tag{B.8}$$

where where we used Lemma 5 (gradient domination), and σ_A is the strong convexity parameter of F_A^* (lower bounded by $\lambda_{min}^+(A^TA)/L_{max}$). Hence, if an inequality of the type

$$\frac{C}{T}\sum_{t\leq s<t+T}\sum_{ij\in E}||\nabla_{ij}F_A^*(\lambda(s))||^2\leq\frac{1}{T}\sum_{t\leq s<t+T}\frac{1}{2(\sigma_i^{-1}+\sigma_j^{-1})\mu_{(ij)_s}^2}||\nabla_{(ij)_s}F_A^*(\lambda(s))||^2 \tag{B.9}$$

holds, we have (using (B.3)):

$$\Lambda_{t+1}\leq L_t-C\frac{1}{T}\sum_{t\leq s<t+T}||\nabla F_A^*(\lambda(s))||^2\leq(1-C\sigma_A)\Lambda_t. \tag{B.10}$$

We thus need to tune correctly the μ_{ij}^2 and C in order to have (B.9) verified.

Step 2: We are looking for necessary conditions for (B.9) to hold. In the left term, every coordinate is present at each time s. 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 $||\nabla_{ij}F_A^*(\lambda(s))||^2$, that only make activated coordinates intervene. Let $s\in\{t,...,t+T-1\}$, and suppose that there exists $t\leq r\leq s<r+t_{ij}\leq t+T-1$ such that ij is activated at times r and $r+t_{ij}$. Thanks to the asumption 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 $r+t_{ij}$. Thanks to our asumptions, we know that $t_{ij}\leq aL_{ij}$. We have the following basic inequalities:

$$||\nabla_{ij}F_A^*(\lambda(s))||^2\leq(||\nabla_{ij}F_A^*(\lambda(r))||+||\nabla_{ij}F_A^*(\lambda(s))-\nabla_{ij}F_A^*(\lambda(r))||)^2 \tag{B.11}$$

$$\leq 2(||\nabla_{ij}F_A^*(\lambda(r))||^2+||\nabla_{ij}F_A^*(\lambda(s))-\nabla_{ij}F_A^*(\lambda(r))||^2). \tag{B.12}$$

The quantity $||\nabla_{ij}F_A^*(\lambda(s))-\nabla_{ij}F_A^*(\lambda(r))||^2$ then needs to be controlled. We know that thanks to (B.2), for $x,x'\in\mathbb{R}^{E\times d}$, we have

$$||\nabla_{ij}F_A^*(x)-\nabla_{ij}F_A^*(x')||^2\leq 2(\sigma_i^{-1}+\sigma_j^{-1})^2d_{ij}\mu_{ij}^2\sum_{(kl)\sim(ij)}\mu_{kl}^2||x_{kl}-x'_{kl}||^2. \tag{B.13}$$

Using this with

$$||x_{kl}-x'_{kl}||^2=||\sum_{r<u<s:(ij)_u=(kl)}\frac{1}{(\sigma_k^{-1}+\sigma_l^{-1})\mu_{kl}^2}\nabla_{kl}F_A^*(\lambda(u))||^2 \tag{B.14}$$

$$\leq\sum_{r<u<r+t_{ij}:(ij)_u=(kl)}\left(\frac{1}{(\sigma_k^{-1}+\sigma_l^{-1})\mu_{kl}^2}\right)^2N(kl,ij,u)||\nabla_{kl}F_A^*(\lambda(u))||^2, \tag{B.15}$$

where we used (and will widely use again below) that $||x_1+...+x_n||^2\leq n(||x_1||^2+...+||x_n||^2)$ (convexity of the squared norm), leads to:

$$||\nabla_{ij}F_A^*(\lambda(s))||^2\leq 2||\nabla_{ij}F_A^*(\lambda(r))||^2 \tag{B.16}$$

$$+2d_{ij}\sum_{r<u<r+t_{ij}}N((ij)_u,ij,u)\frac{\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})^2}{\mu_{(ij)_u}^2(\sigma_{i_u}^{-1}+\sigma_{j_u}^{-1})^2}||\nabla_{(ij)_u}F_A^*(\lambda(u))||^2 \tag{B.17}$$

$$\leq 2||\nabla_{ij}F_A^*(\lambda(r))||^2 \tag{B.18}$$

$$+2d_{ij}\sum_{r<u<r+t_{ij}}\left\lceil b\frac{L_{ij}}{L_{(ij)_u}}\right\rceil\frac{\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})^2}{\mu_{(ij)_u}^2(\sigma_{i_u}^{-1}+\sigma_{j_u}^{-1})^2}||\nabla_{(ij)_u}F_A^*(\lambda(u))||^2 \tag{B.19}$$

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 $t \leq s < t+T$, $ij \in E$. When summing, each $||\nabla_{(ij)_r} F_A^*(\lambda(r))||^2$ appears on the right hand-side of the inequality, with a factor upper-bounded by ($(ij)_r$ noted (ij)):

$$2aL_{ij} + 2d_{ij} \sum_{kl \sim ij} aL_{kl} \left\lceil \frac{bL_{kl}}{L_{ij}} \right\rceil \frac{\mu_{kl}^2 (\sigma_k^{-1} + \sigma_l^{-1})^2}{\mu_{ij}^2 (\sigma_i^{-1} + \sigma_j^{-1})^2}. \tag{B.20}$$

We want the expression above multiplied by C defined in Step 1 to be upper-bounded by $\frac{1}{2(\sigma_i^{-1}+\sigma_j^{-1})\mu_{ij}^2}$, in order for (B.9) to be verified. This is possible if and only if:

$$C \left(2aL_{ij}\mu_{ij}^2(\sigma_i^{-1} + \sigma_j^{-1}) + 2d_{ij} \sum_{kl \sim ij} a \left\lceil \frac{bL_{kl}}{L_{ij}} \right\rceil L_{kl}\mu_{kl}^2 \frac{(\sigma_k^{-1} + \sigma_l^{-1})^2}{\sigma_i^{-1} + \sigma_j^{-1}} \right) \leq \frac{1}{2}, \tag{B.21}$$

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

$$C \left(2aL_{ij}\mu_{ij}^2(\sigma_i^{-1} + \sigma_j^{-1}) + 2d_{ij} \sum_{kl \sim ij} a \frac{bL_{kl}^2}{L_{ij}} \mu_{kl}^2 \frac{(\sigma_k^{-1} + \sigma_l^{-1})^2}{\sigma_i^{-1} + \sigma_j^{-1}} \right) \leq \frac{1}{4} \text{ if } \forall kl \sim ij, L_{ij} \leq bL_{kl}, \tag{B.22}$$

where we bounded $\left\lceil b\frac{L_{ij}}{L_{kl}} \right\rceil$ by $2\frac{bL_{ij}}{L_{kl}}$ here. We here see that in this case, if

$$\mu_{ij}^2 = \frac{1}{L_{ij}(\sigma_i^{-1} + \sigma_j^{-1})} \times \min_{kl \sim ij} \frac{L_{kl}(\sigma_k^{-1} + \sigma_l^{-1})}{L_{ij}(\sigma_i^{-1} + \sigma_j^{-1})} \tag{B.23}$$

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 kl a neighbor of ij such that $L_{ij} > bL_{kl}$. As $L_{ij} > bL_{kl}$, we have $\lceil \frac{bL_{kl}}{L_{ij}} \rceil = 1$, leading to $a\lceil \frac{bL_{kl}}{L_{ij}} \rceil L_{kl}\mu_{kl}^2 = aL_{kl}\mu_{kl}^2 \leq a \leq ab$. 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: $\mu_{ij}^2 = \frac{1}{L_{ij}(\sigma_i^{-1}+\sigma_j^{-1})} \times \min_{kl \sim ij} \frac{L_{kl}(\sigma_k^{-1}+\sigma_l^{-1})}{L_{ij}(\sigma_i^{-1}+\sigma_j^{-1})}$. The final rate thus depends on the smallest eigenvalue of the laplacian weighted by:

$$\frac{1}{2a + 8d_{max}^2 ab} \frac{1}{L_{max}} \frac{1}{L_{ij}(\sigma_i^{-1} + \sigma_j^{-1})} \times \min_{kl \sim ij} \frac{L_{kl}(\sigma_k^{-1} + \sigma_l^{-1})}{L_{ij}(\sigma_i^{-1} + \sigma_j^{-1})}. \tag{B.24}$$

However, having local complexity constraints is not really of much interest to us, as the parameters σ_i entered in the algorithm are generally taken to be the same on all nodes. We thus formulate Theorem 2 with σ_{min} for simplicity (which is slightly weaker in general) which gives as final rate of convergence the smallest eigenvalue of the laplacian weighted by:

$$\nu_{ij} = \frac{1}{2a + 8d_{\max}^2 ab} \frac{\sigma_{\min}}{2L_{\max}} \frac{1}{L_{ij}} \times \min_{kl \sim ij} \frac{L_{kl}}{L_{ij}}. \tag{B.25}$$

B.3 Proof Of Theorem 3: Adding Stochasticity

We now prove the other theorem, where we assume the existence of events A_t for $t \in \mathbb{N}$, under which the asumptions are true. Using the same arguments as in the proof of Theorem 2, we obtain:

$$\mathbb{E}[\Lambda_{t+1} - \Lambda_t | \mathcal{F}_t, A_t] \leq -\sigma \Lambda_t. \tag{B.26}$$

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

$$\mathbb{E}[\Lambda_{t+1}|\mathcal{F}_t] \leq (1 - \sigma \mathbb{I}_{A_t})\Lambda_t. \tag{B.27}$$

And then, by induction:

$$\mathbb{E}[\Lambda_t] \leq \mathbb{E}[P_t \Lambda_0], \text{ where } P_t = \prod_{s=0}^{t-1}(1 - \sigma \mathbb{I}_{A_s}). \tag{B.28}$$

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:

$$\log(P_t) = \log(1 - \sigma) \sum_{s=0}^{t-1} \mathbb{I}_{A_s}, \tag{B.29}$$

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:

$$\mathbb{P}(X_t \geq 1/3|\mathcal{F}_t) + 1/3\mathbb{P}(X_t \leq 1/3|\mathcal{F}_t) \geq \mathbb{E}[X_t|\mathcal{F}_t] \geq 1/2 \implies \mathbb{P}(X_t \geq 1/3|\mathcal{F}_t) \geq 1/4. \tag{B.30}$$

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:

$$\begin{aligned}
\mathbb{E}\left[\prod_{s=t}^{t+3T-1}(1 - \mathbb{I}_{A_s}\sigma)\right] &= \mathbb{E}\left[\prod_{s=t}^{t+T-1}(1 - \mathbb{I}_{A_s}\sigma)\prod_{s=t+T}^{t+2T-1}(1 - \mathbb{I}_{A_s}\sigma)\prod_{s=t+2T}^{t+3T-1}(1 - \mathbb{I}_{A_s}\sigma)\right] && \text{(B.31)}\\
&\leq \mathbb{E}\left[\prod_{s=t}^{t+T-1}(1 - \mathbb{I}_{A_s}\sigma)\prod_{s=t+2T}^{t+3T-1}(1 - \mathbb{I}_{A_s}\sigma)\right] && \text{(B.32)}\\
&\leq \mathbb{E}\left[\prod_{s=t}^{t+T-1}(1 - \mathbb{I}_{A_s}\sigma)\mathbb{E}^{\mathcal{F}_{t+2T}}\left\{\prod_{s=t+2T}^{t+3T-1}(1 - \mathbb{I}_{A_s}\sigma)\right\}\right] && \text{(B.33)}
\end{aligned}$$

as in the last right hand side, the first big product is $\mathcal{F}_{t+2T}$-measurable (our asumption 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:

$$\mathbb{E}[P_t] \leq \left(\frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4}\right)^{\lfloor t/(2T)\rfloor}, \tag{B.34}$$

which is the desired bound. For the asymptotic one, $(1 - \sigma)^{T/3} \leq e^{-\sigma T/3}$. For σT small enough (less than $\log(2)$), we have $e^{-\sigma T/3} \leq 1 - \sigma T/3$, leading to $(\frac{1}{4}(1 - \sigma)^{T/3} + \frac{3}{4})^{\lfloor t/(2T)} \leq (1 - T\sigma/12)^{\lfloor t/(2T)} \leq e^{-(t+o(t))\sigma/24}$. The asymptotic rate of convergence thus holds if the assumption made in Corollary 1 holds.

C STUDY IN THE LOSS NETWORK MODEL: PROOF OF THE STATED RATE OF CONVERGENCE

We first assume to be in the case $\varepsilon = 0$. We generalize to $\varepsilon > 0$ at the end. Let $t \in \mathbb{N}$ be fixed, and B_t be the event: "in the activations $t, t+1, ..., t+T-1$, all edges are ativated". Let then $C_t(ij, s)$ for $t \leq s < t+T$ be the event $\min(T_{ij}(s), t+T-s, s-t) \leq aL_{ij}$ and $D_t(kl, ij, s)$ be the event $N(kl, ij, s) \leq \lceil bL_{ij}/L_{kl} \rceil$, where $N(kl, ij, s)$ is the number of activations of kl between two activations of ij, around time s, where we only take into account the activations between times t and $t+T-1$. Let then $A_t = B_t \cap (\cap_{kl,ij \in E, t \leq s < t+T} C_t(ij, s) \cap D_t(kl, ij, s))$. We want $\mathbb{P}(A_t) \geq 1/2$ for correct constants a, b, T and L_{ij} (that can differ from τ_{ij}). Note that this event is $\mathcal{F}_{t+T-1}$-measurable, as desired. We first study the length of time L_{ij} edge ij must wait in order to be activated with high probability (*high* meaning more that $1 - \frac{1}{12|E|}$). This result is Lemma 6. Then, we use this length to determine the constants T, a, b, L_{ij} needed.

Lemma 6. *For any $t_0 \geq 0$, $ij \in E$, if $p_{ij} = \frac{1}{2\max(d_i, d_j)-1}\tau_{ij}^{-1}$ and $\tau_{max}(ij) = \max_{kl \sim ij} \tau_{kl}$, let $L_{ij} = \frac{\log(6|E|)}{\log(1-(1-e^{-1})e^{-1})}(p_{ij}^{-1} + \tau_{max}(ij))$. We have:*

$$\mathbb{P}(ij \text{ not activated in } [t_0, t_0 + L_{ij}] | \mathcal{F}_{t_0}) \leq \frac{1}{6|E|}. \tag{C.1}$$

Proof of Lemma 6. Let $ij \in E$ 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 ij is activable at a time t or not. More formally, we define a process $N_{ij}(t)$ with values in $\mathbb{N}$, such that $N_{ij}(t_0) = 1$ if ij non-available at time t_0 and 0 otherwise. Then, when an edge $kl, kl \sim ij$ is activated, we make an increment of 1 on $N_{ij}(t)$ (a *customer* arrives). This customer stays for a time τ_{kl} and when he leaves we make N_{ij} decrease by 1. We have $N_{ij} \geq 0$ a.s., and if $N_{ij} = 0$, ij is available. For $t \geq \max_{kl \sim ij} \tau_{kl} + t_0$, $N_{ij}(t)$ follows a Poisson law of parameter $\sum_{kl \sim ij} p_{kl}\tau_{kl}$. For any $t \geq \max_{kl \sim ij} \tau_{kl} + t_0$:

$$\mathbb{P}(ij \text{ available at time } t | \mathcal{F}_{t_0}) \geq \mathbb{P}(N_i(t) = 0) = \exp(-\sum_{kl \sim ij} p_{kl}\tau_{kl}). \tag{C.2}$$

That leads to taking $p_{kl} = \frac{1}{2}\frac{1}{\max(d_k, d_l)-1}\tau_{kl}^{-1}$ for all edges, in order to have $\mathbb{P}(ij \text{ available at time } t | \mathcal{F}_{t_0}) \geq 1/e$. Then, $\mathbb{P}(ij \text{ rings in } [t, t + p_{ij}^{-1}]) = 1 - e^{-1}$, giving:

$$\begin{aligned}
&\mathbb{P}(ij \text{ activated in } [t_0, t_0 + \tau_{\max}(ij) + p_{ij}^{-1}] | \mathcal{F}_{t_0}) = \mathbb{P}(ij \text{ rings in } [t, t + p_{ij}^{-1}]) && \text{(C.3)}\\
&\quad \times \mathbb{P}(ij \text{ available at time } t | \mathcal{F}_{t_0}, ij \text{ rings at a time} && \text{(C.4)}\\
&\quad t \in [t_0 + \tau_{\max}(ij), t_0 + \tau_{\max}(ij) + p_{ij}^{-1}]) && \text{(C.5)}\\
&\quad \geq (1 - e^{-1})e^{-1}, && \text{(C.6)}
\end{aligned}$$

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 $L_{ij} = k(p_{ij}^{-1} + \tau_{max}(ij))$. Then we have a.s.:

$$\mathbb{P}(ij \text{ not activated in } [t_0, t_0 + L_{ij}] | \mathcal{F}_{t_0}) \leq \frac{1}{6|E|}. \tag{C.7}$$

□

Bounding T: A direct application of Lemma 6 leads, with $L = \max_{ij} L_{ij}$, to:

$$T = 2\sum_{ij} \frac{L}{\tau_{ij}}. \tag{C.8}$$

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

$$\begin{aligned}
&\mathbb{P}(\exists (ij) \in E : (ij) \text{ not activated in } \{t, ..., t+T-1\}|\mathcal{F}_t) && \text{(C.9)}\\
&\leq \sum_{ij \in E} \mathbb{P}((ij) \text{ not activated in } \{t, ..., t+T-1\}|\mathcal{F}_t) && \text{(C.10)}\\
&\leq \sum_{ij \in E} \mathbb{P}((ij) \text{ not activated in } [t, t+L_{ij}]|\mathcal{F}_t) && \text{(C.11)}\\
&\leq |E| \times \frac{1}{6|E|} && \text{(C.12)}\\
&= 1/6. && \text{(C.13)}
\end{aligned}$$

Bounding T_{ij}: Applying Lemma 6 with $12|E|T$ instead of $6|E|$ leads to controlling all the inactivation lengths by a length L'_{ij}, with a probability more than $1 - 1/(12|E|T)$. Let $ij \in E$ and $s \in \mathbb{N}$, $t \leq s < t+T$. Let $\alpha > 0$ to tune later. Denote by $\delta_{ij}(s)$ the (random) inactivation time of ij, around iteration s. Note that conditionnaly on the inactivation period $\delta_{ij}(s)$, $T_{ij}(s)$ is dominated in law by a Poisson variable of parameter $I\delta_{ij}(s)$, hence line (C.15):

$$\begin{aligned}
\mathbb{P}(T_{ij}(s) \geq \alpha L'_{ij}|\mathcal{F}_t) &\leq \mathbb{P}(T_{ij}(s) \geq \alpha L'_{ij}|\mathcal{F}_t, \delta_{ij} \leq L'_{ij}) \times \mathbb{P}(\delta_{ij} \leq L'_{ij}) + \mathbb{P}(\delta_{ij} \geq L'_{ij}) && \text{(C.14)}\\
&\leq \mathbb{P}(Poisson(IL'_{ij}) \geq \alpha L'_{ij}) + \frac{1}{12|E|T} && \text{(C.15)}\\
&\leq \frac{1}{12|E|T} + \frac{1}{12|E|T} && \text{(C.16)}\\
&= \frac{1}{6|E|T}, && \text{(C.17)}
\end{aligned}$$

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

Lemma 7 (A Large Deviation Inequality on discrete Poisson variables.)**.** *Let $Z \sim Poisson(\lambda)$, for some $\lambda > 0$. Then, for all $u \geq 0$:*

$$\mathbb{P}(Z \geq u) \leq \exp(-u + \lambda(e-1)). \tag{C.18}$$

This large deviation leads to taking $\alpha = 2eI$ for (C.16) to be true. Finally, we get:

$$\mathbb{P}(T_{ij}(s) \geq \alpha L'_{ij}|\mathcal{F}_t) \leq \frac{1}{6|E|T}. \tag{C.19}$$

Bounding $N(kl, ij, s)$: If $\delta_{ij}(s) \leq L'_{ij}$, this random variable is dominated by a Poisson variable of parameter $p_{kl}L'_{ij}$. Hence, still with Lemma 7, with probability more than $1 - \frac{1}{12|E|^2T}$, we can bound $N(kl, ij)$ by $e\log(12|E|^2T) + p_{kl}L_{ij}(e-1) \leq 2ep_{kl}L_{ij}$.

Explicit writing of the union bound on A_t^C: $A_t^C = B_t^C \cup (\cup_{kl,ij \in E, t \leq s < t+T} C_t(ij,s)^C \cup D_t(kl,ij,s)^C) \in \mathcal{F}_{t+T-1}$. Thanks to the previous considerations, we have that $\mathbb{P}^{\mathcal{F}_t}(B_t^C) \leq 1/6$ with (C.13), $\mathbb{P}^{\mathcal{F}_t}(C_t(ij,s)^C) \leq \frac{1}{6|E|T}$ with (C.19) and $\mathbb{P}(D_t(kl,ij,s)^C|\mathcal{F}_t) \leq \frac{1}{6|E|^2T}$, for the following constants and weights:

- $\tilde{\tau}_{ij}^{-1} = p_{ij} = \min(\frac{1}{\tau_{\max(ij)}}, \frac{1}{2(\max(d_i,d_j)-1)}\frac{1}{\tau_{ij}})$;
- $T = 2I \max_{ij \in E} \tilde{\tau_{ij}} \frac{\log(6|E|)}{\log(1-(1-e^{-1})e^{-1})}$;
- $a = 2eI \frac{\log(6|E|T)}{\log(1-(1-e^{-1})e^{-1})}$;
- $b = 2e \frac{\log(6|E|T)}{\log(1-(1-e^{-1})e^{-1})}$.

The union bound is the following:

$$\mathbb{P}^{\mathcal{F}_t}(A_t^C) \leq \mathbb{P}^{\mathcal{F}_t}(B_t^C) + \sum_{s,ij} \mathbb{P}^{\mathcal{F}_t}(C_t(ij,s)^C) + \sum_{s,ij} \mathbb{P}^{\mathcal{F}_t}(\cup_{kl} D_t(kl,ij,s)^C) \tag{C.20}$$

$$\leq 1/6 + |E|T/(6|E|T) \times 2 \tag{C.21}$$

$$\leq 1/2. \tag{C.22}$$

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

$$\nu_{ij} = \frac{\sigma_{min}}{L_{max}} \times \frac{\tilde{\tau_{ij}} \min_{kl \sim ij} \frac{\tau_{ij}}{\tau_{kl}}}{8a(1+d^2b)}. \tag{C.23}$$

Note that this analysis works for $\varepsilon = 0$, but also for **RLNM($\varepsilon > 0$)** by replacing τ_{ij} by $(1+\varepsilon)\tau_{ij}$. Indeed, Lemma 6 still holds with $(1+\varepsilon)\tau_{ij}$: the queuing construction still works.

D PROOF OF THE ACCELERATED CACDM RATE

D.1 *CACDM* Formulated on the Dual Variables

Section 3 of the paper presents *CACDM* formulated on node dual variables x, y. The analysis is done with edge dual variables λ, ω verifying $A\lambda = x, A\omega = y$. Local updates on node variables are equivalent to coordinate gradient steps on edge variables. Here are the operations done no λ, ω.

1) Continuous Contractions: For all times $t \in \mathbb{R}^+$, make the infinitesimal contraction

$$\begin{pmatrix} \lambda_{t+dt} \\ \omega_{t+dt} \end{pmatrix} = \begin{pmatrix} 1 - dtI\theta & dtI\theta \\ dtI\theta & 1 - dtI\theta \end{pmatrix} \begin{pmatrix} \lambda_t \\ \omega_t \end{pmatrix}, \tag{D.1}$$

between times t and $t + dt$, on the dual variables.

2) Local Updates: When edge (ij) is activated at time $t \geq 0$, define the coordinate gradient step:

$$\eta_{ij,t} = -\begin{pmatrix} \frac{1}{2\mu_{ij}^2} U_{ij} \nabla_{ij} F_A^*(\lambda_t) \\ \frac{\theta}{\sigma_A p_{ij}} U_{ij} \nabla_{ij} F_A^*(\lambda_t) \end{pmatrix} \tag{D.2}$$

where σ_A is the strong convexity parameter of F_A^*, $U_{ij} = e_{ij} e_{ij}^T$, and perform the gradient step:

$$\begin{pmatrix} \lambda_t \\ \omega_t \end{pmatrix} \overset{t}{\leftarrow} \begin{pmatrix} \lambda_t \\ \omega_t \end{pmatrix} + \eta_{ij,t} \tag{D.3}$$

on the dual variables λ_t and ω_t.

D.2 Proof of the Accelerated *CACDM* Rate of Convergence

Proof of the continuous bound on Continuous ACDM. The proof closely follows the lines of Nesterov and Stich (2017b); **?**, adapted to fit our continuous time algorithm. Without loss of generality, we can assume that $I = 1$ i.e. that the p_{ij} sum to 1 (by rescaling time with $t' = tI$). Note $r_t = ||\omega_t - \lambda^*||^*$, and $f_t = F_A^*(\lambda_t) - F_A^*(\lambda^*)$, such that $L_t = r_t^2 + \frac{2\theta^2 S^2}{\sigma_A^2} f_t$. Let $t \geq 0$ and $dt > 0$. The following equalities and inequalities are true at a $o(dt)$ approximation. Let's start with the term r_t^2:

$$\mathbb{E}^{\mathcal{F}_t}[r_{t+dt}^2] = (1 - dt)\mathbb{E}^{\mathcal{F}_t}[r_{t+dt}^2 | \text{no activations between t and t+dt}] \tag{D.4}$$

$$+ dt\mathbb{E}^{\mathcal{F}_t}[r_{t+dt}^2 | \text{1 activation between t and t+dt}] \tag{D.5}$$

For the first term, we get:

$$\mathbb{E}^{\mathcal{F}_t}[r^2_{t+dt}|\text{no activation in } [t,t+dt]] = ||(1-\theta dt)\omega_t + \theta dt\lambda_t - \lambda^*||^{*2} \tag{D.6}$$
$$\leq (1-\theta dt)r_t^2 + \theta dt||\lambda_t - \lambda^*||^{*2} \tag{D.7}$$

where the inequality uses convexity of the squared function. For the other term, we decompose the event "1 activation between t and t+dt" in the disjoint events "ij activated between t and t+dt", of probability $p_{ij}dt$, to get the following, true at a $o(1)$ approximation (enough because we multiply by dt afterwise):

$$\mathbb{E}^{\mathcal{F}_t}[r^2_{t+dt}|\text{1 activation between t and t+dt}] = \sum_{(ij)\in E} p_{ij}||\omega_t - \frac{\theta}{p_{ij}\sigma_A}U_{ij}\nabla_{ij}F_A^*(\lambda_t) - \lambda^*||^{*^2} \tag{D.8}$$
$$= ||\omega_t - \lambda^*||^{*2} \tag{D.9}$$
$$+ \sum_{ij} p_{ij}\frac{\theta^2}{\sigma_A^2 p_{ij}^2}||U_{ij}\nabla_{ij}F_A^*(\lambda_t)||^{*2} \tag{D.10}$$
$$- 2\sum_{ij} p_{ij}\frac{\theta}{p_{ij}\sigma_A}\langle U_{ij}\nabla_{ij}F_A^*(\lambda_t), \omega_t - \lambda^*\rangle \tag{D.11}$$

For the term $\sum_{ij} p_{ij}\frac{\theta^2}{\sigma_A^2 p_{ij}^2}||U_{ij}\nabla_{ij}F_A^*(\lambda_t)||^{*2}$, we get by definition of S^2, and by a local smoothness inequality (namely, $\forall y, F_A^*(y) - F_A^*(y - \frac{1}{\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})}U_{ij}\nabla_{ij}F_A^*(y)) \geq \frac{1}{2\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})}||\nabla_{ij}F_A^*(y)||^2$):

$$\sum_{ij} p_{ij}\frac{\theta^2}{\sigma_A^2 p_{ij}^2}||U_{ij}\nabla_{ij}F_A^*(\lambda_t)||^{*2} \leq \sum_{ij} p_{ij}\frac{2\theta^2 S^2}{\sigma_A^2\mu_{ij}^2(\sigma_i^{-1}+\sigma_j^{-1})}||U_{ij}\nabla_{ij}F_A^*(\lambda_t)||^2 \tag{D.12}$$
$$\leq \sum_{ij} p_{ij}\frac{2\theta^2 S^2}{\sigma_A^2}(F_A^*(\lambda_t) - F_A^*(\lambda_t - \frac{\theta}{\sigma_A p_{ij}}U_{ij}\nabla_{ij}F_A^*(\lambda_t))) \tag{D.13}$$
$$= \frac{2\theta^2 S^2}{\sigma_A^2}(F_A^*(\lambda_t) - \mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt})|\text{1 activation in [t,t+dt]}]). \tag{D.14}$$

For the term $-2\sum_{ij} p_{ij}\frac{\theta}{p_{ij}\sigma_A}\langle U_{ij}\nabla_{ij}F_A^*(\lambda_t), \omega_t - \lambda^*\rangle$, we get, by adding and subtracting a λ_t in the bracket, and by convexity of F_A^* (σ_A is the strong convexity parameter of F_A^*):

$$-2dt\frac{\theta}{\sigma_A}\langle\nabla F_A^*(\lambda_t), \omega_t - \lambda^*\rangle = -2dt\frac{\theta}{\sigma_A}\langle\nabla F_A^*(\lambda_t), \omega_t - \lambda_t\rangle - 2dt\frac{\theta}{\sigma_A}\langle\nabla F_A^*(\lambda_t), \lambda_t - \lambda^*\rangle \tag{D.15}$$
$$\leq -2\frac{1}{\sigma_A}\langle\nabla F_A^*(\lambda_t), \theta dt(\omega_t - \lambda_t)\rangle \tag{D.16}$$
$$- 2dt\frac{\theta}{\sigma_A}(F_A^*(\lambda_t) - F_A^*(\lambda^*) + \sigma_A/2||\lambda_t - \lambda^*||^{*2}) \tag{D.17}$$
$$\tag{D.18}$$

Then, let's define $\lambda'_{t+dt} = (1-\theta dt)\lambda_t + \theta dt\omega_t = \mathbb{E}^{\mathcal{F}_t}[\lambda_{t+dt}|\text{no activations in } [t,t+dt]]$. By noticing that $\theta dt(\omega_t - \lambda_t) = \lambda'_{t+dt} - \lambda_t$, we get:

$$-2\frac{1}{\sigma_A}\langle\nabla F_A^*(\lambda_t), \theta dt(\omega_t - \lambda_t)\rangle = -2\frac{1}{\sigma_A}\langle\nabla F_A^*(\lambda_t), \lambda'_{t+dt} - \lambda_t\rangle \tag{D.19}$$
$$= -2\frac{1}{\sigma_A}\langle\nabla F_A^*(\lambda'_{t+dt}), \lambda'_{t+dt} - \lambda_t\rangle \tag{D.20}$$
$$\leq -2\frac{1}{\sigma_A}(F_A^*(\lambda'_{t+dt}) - F_A^*(\lambda_t)), \tag{D.21}$$

where from (D.19) to (D.20), the equality holds at $o(dt)$, as the left part of the bracket is true at $o(1)$ precision, and the right part of the bracket is a $O(dt)$. Then equation (D.20) to (D.21) is a convexity inequality. By combining these inequalities, and deleting the terms that compensate themselves, we get:

$$\mathbb{E}^{\mathcal{F}_t}[r_{t+dt}^2] - r_t^2 \leq -dt\theta r_t^2 + dt\theta||\lambda_t - \lambda^*||^{*2} \tag{D.22}$$

$$+ dt\frac{2\theta^2 S^2}{\sigma_A^2}(F_A^*(\lambda_t) - \mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt})|\text{1 activation in [t,t+dt]}]) \tag{D.23}$$

$$- 2dt\frac{\theta}{\sigma_A}(F_A^*(\lambda_t) - F_A^*(\lambda^*) + \sigma_A/2||\lambda_t - \lambda^*||^{*2}) \tag{D.24}$$

$$- 2\frac{1}{\sigma_A}(F_A^*(\lambda'_{t+dt}) - F_A^*(\lambda_t)) \tag{D.25}$$

Studying $\mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt})]$, we get:

$$\mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt})] = (1-dt)F_A^*(\lambda'_{t+dt}) + dt\mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt}) - F_A^*(\lambda^*)|\text{1 activation in [t,t+dt]}] \tag{D.26}$$

Using $\theta^2 = \sigma_A/S^2$ (i.e $\theta^2 S^2/\sigma_A^2 = 1/\sigma_A$) and the above equality, equations (D.8) to (D.11) become:

$$\mathbb{E}^{\mathcal{F}_t}[r_{t+dt}^2] - r_t^2 \leq -dt\theta r_t^2 \tag{D.27}$$

$$+ dt\frac{2}{\sigma_A}(F_A^*(\lambda_t) - \mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt})|\text{1 activation in [t,t+dt]}]) \tag{D.28}$$

$$- 2dt\frac{\theta}{\sigma_A}(F_A^*(\lambda_t) - F_A^*(\lambda^*)) \tag{D.29}$$

$$- 2\frac{1}{\sigma_A}(F_A^*(\lambda'_{t+dt}) - F_A^*(\lambda_t)) \tag{D.30}$$

$$= -dt\theta r_t^2 \tag{D.31}$$

$$- \frac{2}{\sigma_A}(\mathbb{E}^{\mathcal{F}_t}[F_A^*(\lambda_{t+dt}) - F_A^*(\lambda^*)] - F_A^*(\lambda_t) - F_A^*(\lambda^*)) \tag{D.32}$$

$$- 2dt\frac{\theta}{\sigma_A}(F_A^*(\lambda_t) - F_A^*(\lambda^*)) \tag{D.33}$$

$$+ 2dt\frac{\theta}{\sigma_A}(F_A^*(\lambda_t) - F_A^*(\lambda'_{t+dt})) \tag{D.34}$$

As line (D.34) is a $o(dt)$, we get the desired equation, namely:

$$\mathbb{E}^{\mathcal{F}_t}[L_{t+dt}] - L_t \leq -\theta dt L_t \tag{D.35}$$

Taking the mean, dividing by dt that we make tend to zero, we get $\frac{d}{dt}\mathbb{E}L_t \leq -\theta\mathbb{E}L_t$, and by integrating:

$$\forall t \geq 0, \mathbb{E}L_t \leq \exp(-\theta t)L_0$$

□