

Submitted to *Bernoulli*

Distributed gradient-based optimization in the presence of dependent aperiodic communication

ADRIAN REDDER*¹, ARUNSELVAN RAMASWAMY² and HOLGER KARL³

¹*Department of Computer Science, Paderborn University. E-mail: aredder@mail.upb.de*

²*Department of Computer Science, Karlstad University. E-mail: arunselvan.ramaswamy@kau.se*

³*Hasso-Plattner-Institute, University of Potsdam. E-mail: holger.karl@hpi.de*

Iterative distributed optimization algorithms involve multiple agents that communicate with each other, over time, in order to minimize/maximize a global objective. In the presence of unreliable communication networks, the Age-of-Information (AoI), which measures the freshness of data received, may be large and hence hinder algorithmic convergence. In this paper, we study the convergence of general distributed gradient-based optimization algorithms in the presence of communication that neither happens periodically nor at stochastically independent points in time. We show that convergence is guaranteed provided the random variables associated with the AoI processes are stochastically dominated by a random variable with finite first moment. This improves on previous requirements of boundedness of more than the first moment. We then introduce stochastically strongly connected (SSC) networks, a new stochastic form of strong connectedness for time-varying networks. We show: If for any $p \geq 0$ the processes that describe the success of communication between agents in a SSC network are α -mixing with $n^{p-1}\alpha(n)$ summable, then the associated AoI processes are stochastically dominated by a random variable with finite p -th moment. In combination with our first contribution, this implies that distributed stochastic gradient descent converges in the presence of AoI, if $\alpha(n)$ is summable.

Keywords: Distributed Optimization; Stochastic Gradient Descent; Time-Varying Networks; Age of Information; α -Mixing; Stochastic dominance

1. Introduction

Distributed optimization of stochastically approximated loss functions lies at the heart of many system-level problems that arise in multi-agent learning [26], resource allocation for data centers [12], or decentralized control of power systems [18]. In these scenarios, distributed implementations have many advantages such as balanced workload or the avoidance of a single point of failure. However, this usually comes with high communication costs for coordination [29], entailing that information can only be exchanged rarely, causing local versions of global information to be significantly outdated. Hence, it is of high interest to characterize conditions such that a distributed optimization algorithm can converge when only significantly outdated information with sporadic updates is available.

We therefore consider distributed stochastic optimization problems (SOPs) where the choice of local optimization variables has to be coordinated over an uncertain time-varying communication network. A typical distributed SOP can take the following form:

$$x^* = (x_1^*, \dots, x_D^*) = \operatorname{argmin}_{x \in \mathbb{R}^d} \mathbb{E}_\xi [f(x, \xi)] \quad (1.1)$$

The objective is to minimize a real-valued function $f : \mathbb{R}^d \times \mathcal{S} \rightarrow \mathbb{R}$, which is function of an optimization variable $x \in \mathbb{R}^d$ and a random variable ξ representing noise or uncertainty taken from a set \mathcal{S} . The optimization variable x is composed of local components $x_i \in \mathbb{R}^{d_i}$ that are associated with local

agents of a distributed system. Hence, no global control of x is possible. Moreover, the distribution of ξ is typically unknown in practical scenarios and the agents can only observe samples ξ^n of the uncertainty ξ at discrete time steps $n \in \mathbb{N}_0$. Thus, the problem is that each agent i has to coordinate the local choice for its variable x_i with all other agents by exchanging information over a network and iteratively refine the choice based on the observed samples of uncertainty ξ^n .

To solve this problem, we propose the following solution. Suppose every agent i runs a local distributed stochastic gradient descent (SGD) algorithm that generates a sequence $\{x_i^n\}_{n=0}^\infty$ to solve problem (1.1). Ideally, every agent i would like to have direct access to every new element of the sequences $\{x_j^n\}_{n=0}^\infty$ from every other agent $j \neq i$ during run-time of its own local algorithm. However, due to the distributed nature of the considered setting, the agents have to communicate their updates for their local optimization variables to other agents via a communication network. Because of the uncertainty of communication networks, each agent i can therefore only use delayed versions $x_j^{n-\tau_{ij}(n)}$ for all $j \neq i$ to update its own local variable x_i^n . Here, $x_j^{n-\tau_{ij}(n)}$ denotes the newest update of x_j available at agent i at time n and $\tau_{ij}(t)$ is its corresponding age. We refer to the $\tau_{ij}(t)$'s as the *Age of Information (AoI) variables*. The resulting distributed algorithm is therefore in essence a “straightforward” implementation of SGD, where the true values of local variables are replaced by their aged counterparts. Due to the size of generated information in large distributed systems, and the uncertainty and high cost for communication over networks, the AoI variables can not be expected to be bounded and should therefore be modelled as an unbounded sequence of random variables. *The problem is therefore to formulate mild network and communication assumptions that are representative and easily verifiable such that this SGD algorithm that uses highly aged information will still converge.*

A major challenge for this problem is the multitude of potential factors that affect the AoI random variables. Information exchange between some pairs of agents might experience unbounded delays; mobility of agents or network scheduling algorithms can induce a varying set of network topologies. This can create dependencies among successive network transmissions, preventing agents from exchanging data for extended periods of time. In general, transmissions that happen close in some domain (e.g. time, frequency, or space in wireless communication) are expected to be highly correlated. It is therefore important to formulate a communication network model and associated assumptions that can represent these cases while being mathematically tractable for analysis. Notably, the assumption of guaranteed periodic or stochastically independent communication is practically unrealistic.

1.1. Network models in the literature

One of the most common models in the distributed optimization literature is a time-varying network model that is represented by a time-varying graph (Definition 5). For this graph, the most common assumption is that there is a constant M such that the union graph associated with all time intervals $[n, n + M]$ is strongly connected [32, 1, 34, 15]. A network with this property is typically called uniformly strongly connected [21], M -strongly connected [22, 28] or jointly strongly connected [33]. This model implies guaranteed periodic communication. Another common model is to assume a time-varying network graph whose expected union graph is strongly connected, where the events that describe the success of communication across network edges are independent across time [3, 16, 14, 25].

In ref. [16, 22, 32, 28, 34, 14, 3, 15] the objective is that agents come to a consensus on one global optimization variable to minimize the sum of real-valued functions, each of which associated with one of the local agents. Although such consensus-type problems might appear quite different to (1.1), it turns out that an algorithm for (1.1) can also find a solution for consensus problems after a minor reformulation at the cost of additional communication, which we discuss in [27]. In contrast to the

consensus type problems, ref. [33, 1, 25] and this work consider distributed optimization problems where each agent has to select a local optimization variable, such that the combination of all local variables solves a global optimization problem.

Observe that literature exclusively considers network models that either guarantee periodic communication or require communication based on independent events. We believe that these are restrictive assumptions that do not represent real-world communication networks well. To close this gap, we present a less restrictive network model and verifiable network conditions that guarantee that an SGD algorithm finds a solution to problems of the form (1.1). We also show the aforementioned typical network assumptions from the literature are stronger versions of our new set of network assumptions (Assumptions 5 and 6). Our assumptions only require a stochastic form of strong connectivity and a dependency decay (mixing) property. *To the best of our knowledge, ours is the first work that guarantees asymptotic convergence of a distributed optimization scheme under such mildly restrictive conditions, connecting an abstract optimization theory with a wide range verifiable network conditions.* However, it must be noted that other papers (such as those discussed above) can provide rate of convergence results while we merely give an almost sure convergence analysis.

1.2. Summary of contribution

Our work contributes to the literature of network conditions that guarantee asymptotic convergence to the set of stationary points of a distributed stochastic optimization problems with potentially non-convex objective function. Our work, builds on our previous work on SGD for time-varying networks [25]. However, whereas in [25] the focus was on the optimization iteration, with a strong and restrictive i.i.d. network assumption, this work focuses on guaranteeing convergence under significantly weaker network conditions. Most importantly, our network conditions cover time-varying network topologies, unbounded communication delays, non-independent aperiodic communication, asynchronous local updates and event-driven communication.

As the first step, we describe a distributed stochastic gradient descend algorithm (Algorithm 1) that instead of true local variables uses aged variables as a consequence of network communication. The AoI variables $\tau_{ij}(n)$ therefore induce gradient errors when comparing Algorithm 1 with and without AoI. As our first major contribution, we show in Lemma 2 that the aforementioned gradient errors vanish asymptotically under an asymptotic growth conditions for the AoI variables. Specifically, we require that all $\tau_{ij}(n)$ for all $n \in \mathbb{N}_0$ are stochastically dominated by a non-negative integer-valued random variable with *at least finite first moment*. This provides a significant weakening of traditional assumptions from the stochastic approximation literature in the present setting, since traditionally a dominating random variable with at least a bounded moment greater than one was required. With Lemma 2 we then show the convergence of Algorithm 1 in Theorem 1.

Our second contribution is a universally applicable time-varying network model and associated assumptions to generally verify the existence of dominating random variables with an arbitrary required moment conditions. Our time-varying network model is formulated using events A_{ij}^n , each of which represents successful information exchange from some agent i to another agent j during some time slot n . We then introduce the notion of (ε, κ) -stochastically strongly connected (SSC) network with $\varepsilon \in (0, 1)$ and $\kappa \in \mathbb{N}_0$. This notion requires that there is a set of network edges that form a strongly connected graph for which $\mathbb{P}\left(\bigcup_{n=n}^{n+\kappa} A_{ij}^n\right) \geq \varepsilon$ for all $n \in \mathbb{N}_0$. In other words, for those edges communication is successful at least ones over every interval of length κ with at least probability ε . We then present a general recipe to validate stochastic dominance properties with required moment conditions. Afterwards, Theorem 2 presents our main result: Fix any $p \geq 0$ and consider a (ε, κ) -SSC network. If there exists some $\eta \in \mathbb{N}_0$, such that the processes $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^k}$ are α -mixing with $\sum_{n=0}^{\infty} n^{p-1} \alpha(n) < \infty$,

then all AoI variables $\tau_{ij}(n)$ are stochastically dominated by non-negative integer-valued random variable $\bar{\tau}$ with $\mathbb{E}[\bar{\tau}^p] < \infty$. This result, together with Theorem 1, will therefore imply our final convergence result for Algorithm 1 under the minimal requirement of a SSC network with summable α -mixing coefficients.

The rest of the paper is structured as follows: In Section 2 we state notation and preliminaries from probability and graph theory. In Section 3 we discuss the problem formulation and our distributed SGD algorithm. Afterwards we prove the almost sure convergence of Algorithm 1 in Section 4 under asymptotic growth conditions for the AoI variables. We then introduce our time-varying network model and associated assumptions in Section 5. Section 6 then presents our construction to validate stochastic dominance properties and our main results. Finally, we discuss the verifiability of our network assumptions and future work in Section 7.

2. Notation, definitions and preliminaries

This section presents notation, and preliminaries from probability and graph theory. Throughout our work, discrete points in time are indicated by superscript letters n . We refer to a time slot n as the time interval from time step $n - 1$ to n . We use $n \in \mathbb{N}_0$ to denote $n \in \mathbb{N} \cup \{0\}$.

We make frequent use of the big \mathcal{O} notation: Consider two real-valued sequences x^n, y^n . Then $x^n \in \mathcal{O}(y^n)$, if $\limsup_{n \rightarrow \infty} \frac{x^n}{y^n} < \infty$.

From probability theory we need the concepts of stochastic dominance, expectation of non-negative integer-valued random variables, measure of dependency and α -mixing:

Definition 1. A non-negative integer-valued random variable τ is said to be **stochastically dominated** by a random variable $\bar{\tau}$ if $\mathbb{P}(\tau > m) \leq \mathbb{P}(\bar{\tau} > m)$ for all $m \geq 0$.

Proposition 1 ([9]). *Suppose τ is non-negative integer-valued random variable, then*

$$\mathbb{E}[\tau^p] = \sum_{m=0}^{\infty} ((m+1)^p - m^p) \mathbb{P}(\tau > m), \quad p > 0. \quad (2.1)$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let \mathcal{A} and \mathcal{B} be two sub- σ -algebras of \mathcal{F} . The measure of dependency α between \mathcal{A} and \mathcal{B} is defined as

$$\alpha(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|. \quad (2.2)$$

Consider a stochastic process $X = \{X_n\}_{n \in \mathbb{N}_0}$. For $0 \leq l \leq m \leq \infty$, define the sub- σ -algebra generated from X_l up to X_m by

$$\mathcal{F}_l^m := \sigma(X_n \mid l \leq n \leq m), \quad (2.3)$$

Informally, the σ -algebra generated by a stochastic process from a time interval describes the information that can be extracted from the associated process realizations, see [11] for details.

Definition 2. The α -mixing coefficients of the process X are

$$\alpha(n) := \sup_{l \geq 0} \alpha(\mathcal{F}_0^l, \mathcal{F}_{l+n}^\infty). \quad (2.4)$$

for every $n \in \mathbb{N}_0$. The process X is called **strongly mixing** (or α -mixing), if $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty$.

Mixing is a notion of asymptotic independence. We refer to [8] for a survey about different mixing notions. We now introduce a subclass of strongly mixing processes with different rates of convergence:

Definition 3. The process X is called **p -strongly mixing** for some $p \geq 0$, if

$$\sum_{n=0}^{\infty} n^{p-1} \alpha(n) < \infty. \quad (2.5)$$

We will use this new mixing property to describe dependency decay of different orders. For details on graph theory we refer the reader to [30]. We require the following concepts:

Definition 4. A directed graph is called **strongly connected**, if every pair of nodes is connected by a directed path.

Definition 5. A **time-varying network** is defined as sequence

$$\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}, \quad (2.6)$$

where each element $(\mathcal{V}, \mathcal{E}^n)$ is a directed graph.

We will use the following new connectivity notion for time-varying networks:

Definition 6. A time-varying network $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$ is called **(ε, κ) -stochastically strongly connected (SSC)** with $\varepsilon \in (0, 1)$ and $\kappa \in \mathbb{N}_0$, if there exists a strongly connected graph $(\mathcal{V}, \mathcal{E})$, such that for all $n \in \mathbb{N}_0$ and for all $(i, j) \in \mathcal{E}$

$$\mathbb{P} \left((i, j) \in \bigcup_{k=n}^{n+\kappa} \mathcal{E}^k \right) \geq \varepsilon. \quad (2.7)$$

3. Problem description

We consider a D -agent distributed optimization problem, where each agent $i \in \mathcal{V} := \{1, \dots, D\}$ has to choose values for a local variable $x_i \in \mathbb{R}^{d_i}$ to minimize a global objective function F . The global optimization variable $x = (x_1, \dots, x_D) \in \mathbb{R}^d$ is the concatenation of the local optimization variables x_i associated with the local agents. The objective function is assumed to be stochastic and given by

$$F(x) := \mathbb{E}_{\xi} [f(x, \xi)], \quad (3.1)$$

with $f : \mathbb{R}^d \times \mathcal{S} \rightarrow \mathbb{R}$ a random real-valued function, where the randomness is modeled by an \mathcal{S} -valued random variable ξ that represents noise or uncertainty.

As discussed in the introduction, if a central agent had direct control of the optimization vector x , it would be straightforward to find a local minimum of (3.1) using stochastic gradient descend (SGD) under suitable assumptions [7, Ch. 10]. However, as the components x_i of x are associated with distributed agents, we consider that the agents need to coordinate their choice for the local optimization variables by exchanging information via a communication network.

We assume a synchronized communication setting according to a global clock $n \in \mathbb{N}_0$. Each agent updates its local variable at every time step n based on a local gradient descend iteration. The iterations

will be defined in Section 3.1. For each agent $i \in \mathcal{V}$, the local iterations generate a sequence $\{x_i^n\}_{n=0}^\infty$ starting from an initial candidate value x_i^0 for the optimal value x_i^* . To execute the local gradient iteration, agent i requires a locally available estimate of the current optimization variable x_j^n of agent j for all $j \neq i$. We consider that these information have to be communicated using a communication network. Specifically, every agent will use the newest available local optimization variable from every other agent to update its own local variable. Due to the potential uncertainty of the network only aged/delayed versions of the local variables of the other agents are available at agent i at any time step. Therefore, agent i has only access to the delayed version

$$\hat{X}_i^n := \left(x_1^{n-\tau_{i1}(n)}, \dots, x_i^n, \dots, x_D^{n-\tau_{iD}(n)} \right) \quad (3.2)$$

of x^n at every time step n . Here, $x_j^{n-\tau_{ij}(n)}$ denotes the newest update of x_j available at agent i at time n and $\tau_{ij}(t)$'s are the AoI random variables. Further, we refer to \hat{X}_i^n as the *local believe vector* of agent i at time n . As the next step, we describe the gradient based iteration that uses \hat{X}_i^n instead of x^n to solve problem (3.1).

3.1. Algorithm

We consider that the agents iteratively refine their local variables using the partial derivatives $\nabla_{x_j} f(\cdot, \xi)$. We assume that the agents do not know the distribution of ξ , but during any time slot n an agent can observe an i.i.d. realisation ξ^n of ξ . For simplicity, we assume that all agents are affected by the same realisation of the random variable ξ . In other words, when agent i and agent j calculate their partial derivatives during some time slot n , they use the same realisation ξ^n of ξ , i.e. $\nabla_{x_i} f(\cdot, \xi^n)$ and $\nabla_{x_j} f(\cdot, \xi^n)$. The extension to agent-specific realisations of ξ is merely a technical reformulation that was already described in [24].

To evaluate the partial derivatives $\nabla_{x_i} f(\cdot, \xi^n)$, agent i uses the most recent available version of the optimization variable x_j^n of agent j for all $j \neq i$, i.e. it calculates $\nabla_{x_i} f(\hat{X}_i^n, \xi^n)$ instead of $\nabla_{x_i} f(x^n, \xi^n)$. The following SGD iteration is used by each agent to update its local variable:

$$x_i^{n+1} = x_i^n - a(n) \left(\nabla_{x_i} f(\hat{X}_i^n, \xi^n) + \lambda_i^n \right), \quad (3.3)$$

where $\{a(n)\}_{n \in \mathbb{N}_0}$ is a given step-size sequence and λ_i^n is a local stochastic additive error that may arise during the calculation of $\nabla_{x_i} f$. Algorithm 1 summarizes the protocol that runs on every agent locally. For now, we assume that the agents use some communication protocol to exchange their local believe vectors \hat{X}_i^n over a network. The protocol and the network properties therefore induce the distribution of the AoI variables $\tau_{ij}(n)$. In the next section, we will prove the convergence of Algorithm 1 under an abstract growth conditions for the AoI variables. Section 5 then formulates a communication network model and associated assumptions to satisfy these growth conditions.

Remark 1. In our previous work [25], we also included asynchronous gradient updates in (3.3). The agents are therefore allowed to update their local variable not at every time step $n \geq 0$. This may be included here using the associated assumptions from [25]. Our previous work, considers (3.3) for a restrictive network model with independent communication (see Section 5.4 for further details). This work resolves this issue, but we use synchronous gradient updates to avoid notational overload.

Algorithm 1: Local algorithm at agent $i \in \mathcal{V}$

```

1 Initialize local optimization variable estimate  $x_i^0$ ;
2 Initialize local belief vector  $\hat{X}_i^0$ ;
3 for all time steps  $n$  do
4   Obtain network sample  $\xi^n$ ;
5    $x_i^{n+1} \leftarrow x_i^n - a(n) \nabla_{x_i} f(\hat{X}_i^n, \xi^n)$ ;
6   Update  $i$ -th component of  $\hat{X}_i^n$  to new  $x_i^{n+1}$  with appended timestamp  $n + 1$ ;
7   Run local communication protocol to exchange  $\hat{X}_i^n$ ;
8 end

```

4. Asymptotic convergence of Algorithm 1

In this section, we will show the asymptotic convergence of Algorithm 1. Specifically, we show that the iterations in (3.3) converge to a neighbourhood of a local stationary point of (3.1). The main part of the proof is to show that the gradient errors

$$e_i^n := \nabla_x F(x_1^{n-\tau_{i1}(n)}, \dots, x_D^{n-\tau_{iD}(n)}, \xi^n) - \nabla_x F(x_1^n, \dots, x_D^n, \xi^n) \quad (4.1)$$

due to AoI vanish asymptotically. This error captures the difference of the gradient descent step some agent i would take given its local believe vector compared to the true global state.

To show that the gradient errors vanish, we require that the AoI variables $\tau_{ij}(n)$ satisfy an asymptotic growth conditions. Observe that the gradient errors depend on the AoI variables and the step size sequence $a(n)$, since $a(n)$ determines how much successive steps of iteration (3.3) differ. If the step size sequence gets smaller quick enough relative to some maximal potential growth of the AoI variables, we expect e_i^n to decay to zero. This is because even significantly outdated information stays relevant, if the steps taken during that time were comparably small. The convergence of Algorithm 1 will then follow from the convergence of (3.3) when one considers no AoI, i.e. the case $\tau_{ij}(n) = 0$.

The following assumption formalize the required trade of between the choice of the step size sequence and the required network quality.

Assumption 1. 1. There exists $p \in [1, 2)$ and a non-negative integer-valued random variable $\bar{\tau}$, such that $\bar{\tau}$ stochastically dominates (Definition 1) all $\tau_{ij}(n)$ for all $i, j \in \mathcal{V}$ and all $n \in \mathbb{N}_0$ with

$$\mathbb{E}[\bar{\tau}^p] < \infty.$$

2. The step-size sequence $\{a(n)\}_{n \in \mathbb{N}_0}$ satisfies:

$$(i) \quad \sum_{n=0}^{\infty} a(n) = \infty, \quad \sum_{n=0}^{\infty} a(n)^2 < \infty.$$

$$(ii) \quad a(n) \in \mathcal{O}(n^{-\frac{1}{p}}) \text{ with } p \text{ as in 1.}$$

Assumption 1.1 requires that the network quality is good enough, such that the tail distribution of the AoI variables $\tau_{ij}(n)$ decays rapidly enough, such that at least a dominating random variable with finite mean exists. This assumption contributes a significant weakening of the traditional assumptions required for convergence in the present setting. The traditional assumptions formulated in [6], required at least a dominating random variable with finite p -th moment for $p > 1$. In this work, we show for the first time that actually $p = 1$ is sufficient to achieve asymptotic convergence. We show that under

Assumption 1.1 the growth of each $\tau_{ij}(n)$ can not exceed any fraction of n after some potentially large time-step. We formulate this in Lemma 1.

Assumption 1.1(i) is standard in the stochastic approximation literature. Assumption 1.1(ii) requires that we choose the step size depending on the network quality. For example, if only the worst network quality can be verified, i.e. that there is only a dominating variable with finite mean, then we have to choose $a(n) \in \mathcal{O}(\frac{1}{n})$. In addition to the aforementioned weakening of assumptions, we also do not require that the step-size sequence is eventually monotonically decreasing and we only require $a(n) \in \mathcal{O}(n^{-\frac{1}{p}})$ instead of $a(n) \in o(n^{-\frac{1}{p}})$. Both conditions were traditionally assumed.

We will now present additional assumptions associated with the objective function f in (3.1) and the iterations in (3.3). After that we show the convergence of Algorithm 1. In Section 5 we will then present verifiable network conditions to ensure that Assumption 1.1 holds. We will also see that it is easy to formulate very restrictive network conditions, such that the growth of the AoI variables behave very well. For example, one can show that all moments of the AoI variables are bounded under the standard assumptions in the distributed optimization literature (see Section 5.4). That is, Assumption 1.1 would be satisfied for all $p \geq 1$.

In addition to Assumption 1, we require the following assumptions.

- Assumption 2.**
1. $\nabla_x f$ is continuous and locally Lipschitz continuous in the x -coordinate, where the associated constant may depend on ξ .
 2. $\mathbb{E}[\nabla_x f] = \nabla_x \mathbb{E}[f]$.
 3. ξ is an \mathcal{S} -valued random variable, where \mathcal{S} is a one-point compactifiable space.

Assumption 3. For all $i \in \mathcal{V}$, we have $\sup_{n \in \mathbb{N}_0} \|x_i^n\| < \infty$ a.s.

Assumption 4. Almost surely, $\limsup_{n \rightarrow \infty} \|\lambda^n\| \leq \lambda$ for some fixed $\lambda > 0$.

We refer to [25] for detailed discussion on the verifiability of Assumptions 2 to 4.

Recall the gradient errors due to the AoI variables in (4.1). Next, we will show that these gradient errors vanish asymptotically. We start with an asymptotic grow property for the AoI variables under Assumption 1.1.

Lemma 1. Under Assumption 1.1, we have for every $\varepsilon \in (0, 1)$ and for all $i, j \in \mathcal{V}$ that

$$\sum_{n=0}^{\infty} \mathbb{P}\left(\tau_{ij}(n) > \varepsilon n^{\frac{1}{p}}\right) < \infty. \quad (4.2)$$

Proof. Fix $\varepsilon \in (0, 1)$. By Assumption 1 there is a non-negative integer-valued random variable $\bar{\tau}$, such that

$$\mathbb{P}\left(\tau_{ij}(n) > \varepsilon n^{\frac{1}{p}}\right) \leq \mathbb{P}\left(\bar{\tau} > \varepsilon n^{\frac{1}{p}}\right) \quad (4.3)$$

for all $n \in \mathbb{N}_0$ and $\mathbb{E}[\bar{\tau}^p] < \infty$. Hence, we have

$$\sum_{n=0}^{\infty} \mathbb{P}\left(\tau_{ij}(n) > \varepsilon n^{\frac{1}{p}}\right) \leq \sum_{n=0}^{\infty} \mathbb{P}\left(\bar{\tau} > \varepsilon n^{\frac{1}{p}}\right) \quad (4.4)$$

$$= \sum_{m=0}^{\infty} \sum_{n \in \mathcal{N}(m)} \mathbb{P} \left(\bar{\tau} > \varepsilon n^{\frac{1}{p}} \right) \quad (4.5)$$

$$\leq \sum_{m=0}^{\infty} \sum_{n \in \mathcal{N}(m)} \mathbb{P}(\bar{\tau} > m) = \sum_{m=0}^{\infty} |\mathcal{N}(m)| \mathbb{P}(\bar{\tau} > m), \quad (4.6)$$

where the sets $\mathcal{N}(m)$ are defined as

$$\mathcal{N}(m) := \{n \in \mathbb{N}_0 : m \leq \varepsilon n^{\frac{1}{p}} < m+1\} = \{n \in \mathbb{N}_0 : m^p/\varepsilon^p \leq n < (m+1)^p/\varepsilon^p\} \quad (4.7)$$

for every $m \in \mathbb{N}_0$. We use these sets to consider all $\varepsilon n^{\frac{1}{p}}$ in every interval $[m, m+1)$. The second inequality then follows from the monotonicity of the cumulative distribution function (CDF) by definition of the sets $\mathcal{N}(m)$. Since $|\mathcal{N}(m)| \leq \frac{1}{\varepsilon^p} ((m+1)^p - m^p)$, we have therefore shown that

$$\sum_{n=0}^{\infty} \mathbb{P} \left(\tau_{ij}(n) > \varepsilon n^{\frac{1}{p}} \right) \leq \frac{1}{\varepsilon^p} \sum_{n=0}^{\infty} ((n+1)^p - n^p) \mathbb{P}(\bar{\tau} > n) = \frac{1}{\varepsilon^p} \mathbb{E}[\bar{\tau}^p] < \infty. \quad (4.8)$$

The last equality follows from Proposition 1, since $\bar{\tau}$ is a non-negative integer-valued random variable. \square

We are now ready to prove that the gradient errors due to AoI vanish asymptotically.

Lemma 2. *Under Assumptions 1 to 3, we have that $\lim_{n \rightarrow \infty} \|e_i^n\| = 0$.*

Proof. By Assumption 3, we have that $x^n \in \bar{B}_R(0)$ for some sample path dependent radius $0 < R < \infty$. Then, [25, Lemma 1] shows that $\nabla_x F$ is locally Lipschitz continuous with a constant independent of ξ . Hence, $\nabla_x F$ is globally Lipschitz continuous with a constant L when restricted to $\bar{B}_R(0)$. Using the triangular inequality, the established Lipschitz continuity of $\nabla_x F$ and Assumption 3, we have that

$$\|e_i^n\| \leq L \sum_{j \in \mathcal{V}} \sum_{m=n-\tau_{ij}(n)}^{n-1} \|x_j^{m+1} - x_j^m\| \leq C \sum_{j \in \mathcal{V}} \sum_{m=n-\tau_{ij}(n)}^{n-1} a(m), \quad (4.9)$$

for a sample path dependent constant $C > 0$. We will now show that

$$\lim_{n \rightarrow \infty} \left(\sum_{m=n-\tau_{ij}(n)}^{n-1} a(m) \right) = 0, \quad (4.10)$$

which will imply that $\lim_{n \rightarrow \infty} \|e_i^n\| = 0$.

By Assumption 1, $a(n) \in \mathcal{O}(n^{-\frac{1}{p}})$. Hence, there are constants $c > 0$ and $N \in \mathbb{N}$, such that

$$a(n) \leq cn^{-\frac{1}{p}} \text{ for all } n \geq N. \quad (4.11)$$

Also by Assumption 1, there is some $\bar{\tau}$ that stochastically dominates all $\tau_{ij}(n)$, with $\mathbb{E}[\bar{\tau}^p] < \infty$. Now fix $\varepsilon \in (0, 1)$. By Lemma 1 we have that

$$\sum_{n=0}^{\infty} \mathbb{P}\left(\tau_{ij}(n) > \varepsilon n^{\frac{1}{p}}\right) < \infty. \quad (4.12)$$

It now follows from the Borel-Cantelli Lemma that $\mathbb{P}\left(\tau_{ij}(n) > \varepsilon n^{\frac{1}{p}} \text{ i.o.}\right) = 0$. Hence, there is sample path dependent $N(\varepsilon) \in \mathbb{N}$, such that

$$\tau_{ij}(n) \leq \varepsilon n^{\frac{1}{p}} \quad \forall n \geq N(\varepsilon). \quad (4.13)$$

Equations (4.11) and (4.13) therefore yield that

$$\sum_{m=n-\tau_{ij}(n)}^{n-1} a(m) \leq c \sum_{m=n-\varepsilon n^{\frac{1}{p}}}^{n-1} m^{-\frac{1}{p}} \quad (4.14)$$

for all n with $n \geq N(\varepsilon)$ and $n - \varepsilon n^{\frac{1}{p}} \geq N$. Finally, using the monotonicity of $n^{-\frac{1}{p}}$, we have

$$\sum_{m=n-\varepsilon n^{\frac{1}{p}}}^{n-1} m^{-\frac{1}{p}} \leq \varepsilon n^{\frac{1}{p}} (n - \varepsilon n^{\frac{1}{p}})^{-\frac{1}{p}} = \varepsilon (1 - \varepsilon n^{\frac{1}{p}-1})^{-\frac{1}{p}} \xrightarrow{n \rightarrow \infty} \begin{cases} \varepsilon & p \in (1, 2), \\ \frac{\varepsilon}{1-\varepsilon} & p = 1. \end{cases} \quad (4.15)$$

Hence,

$$\limsup_{n \rightarrow \infty} \left(\sum_{m=n-\tau_{ij}(n)}^{n-1} a(m) \right) \leq \frac{c\varepsilon}{1-\varepsilon} \quad (4.16)$$

and (4.10) follows, since the choice of ε is arbitrary. \square

In [25, Theorem 1] we proved the convergence of Algorithm 1 for $\tau_{ij}(n) = 0$ for all $n \in \mathbb{N}_0$. The following theorem is now an immediate consequence of this result and Lemma 2.

Theorem 1. *Under Assumptions 1 to 4, we have that Algorithm 1 converges almost surely to a λ -neighbourhood of the set of stationary points of F , where λ is the almost sure bound of the additive errors according to Assumption 4.*

5. A new set of network conditions for distributed optimization

In the previous section, we presented a convergence proof for Algorithm 1 under the network assumption Assumption 1.1. This assumption directly requires that some p -th moment of all AoI variables is bounded. However, the distribution of all AoI variables will typically be the consequence of direct agent to agent communication. We are therefore interested in more concrete conditions on the network and the agent communication that imply the required AoI moment conditions. To achieve this, this section introduces a network model and associated assumptions to verify Assumption 1.1.

5.1. Network model

Recall that Algorithm 1 requires that the agents exchange their local variables x_i^n over a network. The network and an associated communication protocol should allow that local variables x_i^n can frequently spread across the network and reach every agent. We will now introduce a network model where the agents try to exchange their local belief vectors \hat{X}_i^n . The agents therefore try to share their latest available version of all other agents local variable with other agents. This might potentially flood the network with data, however, there well known protocols to reduce the number of possibly redundant transmissions [17].

We assume a time-varying network (Definition 5)

$$\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}, \quad (5.1)$$

which is a sequence of directed graphs. Each agent is in one-to-one correspondence with one node in the graph. For every time step $n \in \mathbb{N}_0$, an edge $(i, j) \in \mathcal{E}^n$ represents the event that agent i successfully exchanges its local believe vector \hat{X}_i^n during time slot n with agent j . We denote this event by A_{ij}^n . Therefore, the sequence of directed graphs and the sequences of events $\{A_{ij}^n\}_{n \in \mathbb{N}_0}$ are in one to one correspondence: An edge $(i, j) \in \mathcal{E}^n$ if and only if the event A_{ij}^n occurs. An edge therefore does not represent the possibility for communication, but the actual event of communication.

One may add additional complexity to the model, e.g. using a graph that represents the possibility for communication. Additionally, the model may be extended to scenarios where multiple successive events A_{ij}^n need to occur to guarantee the exchange of a single realization of a believe vector \hat{X}_i^n . This might be necessary if the dimension of \hat{X}_i^n is very large and/or the network bandwidth is small.

Note that although we defined the events A_{ij}^n for all $(i, j) \in \mathcal{V} \times \mathcal{V}$, some of those events might never occur over the whole time horizon. We will especially do not require that all agents communicate directly! However, at least some of the events A_{ij}^n should occur “frequently” enough such that the time-varying network satisfies certain connectivity properties. This will be formulated in Section 5.2 with Assumption 5.

The formulation of the time-varying communication network using the edge events A_{ij}^n has several advantages. The model allows for an underlying time-varying graph that may be the consequence of an network scheduling algorithm or the physical dynamics of the agents themselves. Each event A_{ij}^n can be represented as a multistage process. For example, (i) the availability of a channel, (ii) the use of an access protocol given the availability of a channel, (iii) the success of the transmission given the successful channel access. In general, the event-based formulation appears to be very convenient for analysis.

In the next two subsection, we will formulate our assumptions for the time-varying network $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$ using the events A_{ij}^n .

5.2. Stochastic strong connectedness

The following assumption formalizes our required network connectivity property.

Assumption 5 (Network connectivity assumption). We assume that the time-varying network is (ε, κ) -stochastically strongly connected (SSC) (Definition 6) for some $\varepsilon \in (0, 1)$ and some $\kappa \in \mathbb{N}_0$.

Using the events A_{ij}^n , a (ε, κ) -SSC network requires there exists a strongly connected graph $(\mathcal{V}, \mathcal{E})$, such that for all $n \in \mathbb{N}_0$ and for all $(i, j) \in \mathcal{E}$, we have

$$\mathbb{P} \left(\bigcup_{k=n}^{n+\kappa} A_{ij}^k \right) \geq \varepsilon. \quad (5.2)$$

A (ε, κ) -SSC network therefore requires that there are some pairs of agents $(i, j) \in \mathcal{E}$ that can communicate directly at least ones in every time-interval of the form $[n, n + \kappa]$ with positive probability ε . Notice that SSC does not require direct communication between every pair of agents. The only agents that do communicate are those given in the set \mathcal{E} . A SSC network reflects our intuition of a non-degenerate communication network. Some agents can “frequently” exchange information with positive probability and information can spread across the network since \mathcal{E} is strongly connected.

Note that a network that is SSC does not imply guaranteed transmissions periodically. We will see shortly that SSC is significantly weaker than plain guaranteed periodic communication. With stochastic strong connectivity we can not draw any conclusions about the dependency of events in the network. On the other hand, assuming guaranteed periodic communication does imply a strong form of dependency decay as shown in Section 5.4. Recall that our objective is to verify Assumption 1.1. However, using SSC alone is not sufficient to even guarantee the existence of a dominating random variable as required in Assumption 1.1. The next subsection therefore formulates dependency decay conditions using strong mixing (Definition 2).

5.3. Network dependency decay

Recall that our time-varying network is given by a sequence of directed graphs $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$. The sequence is in one-to-one correspondences with events A_{ij}^n that represent the presence of an edge at time n . We will now formulate a dependency decay assumption based on the notion of strongly mixing processes. We can then show that the AoI variables $\tau_{ij}(n)$ associated with a (ε, κ) -SSC network satisfies specific moment conditions depending on the assumed rate at which dependency decays in the network.

Assumption 6 (Dependency decay assumption). We assume that the time-varying network is such that there is some $\eta \geq 0$ such that each process $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^k}$ is p -strongly mixing (Definition 3) for some $p \in [1, 2)$.

With this assumption we do not require that the dependency of subsequent events A_{ij}^n does decay at any specific rate. However, there should be an interval size $\eta > 0$, such that the dependency of subsequent union events $\bigcup_{k=n}^{n+\eta} A_{ij}^k$ decays sufficiently fast. Notice that Assumption 6 is a dependency decay assumption for the network processes $\mathbb{1}_{\bigcup_{k=n}^{n+N} A_{ij}^k}$ associated with all network edges $(i, j) \in \mathcal{V}$.

However, we actually only require the assumption for those edges $(i, j) \in \mathcal{E}$ in an edge set \mathcal{E} according to Assumption 5. *Additionally, notice that we do not require any form of independence or dependency decay between transmissions over different edges.* The reason for this is Lemma 5. The lemma will show that the existence of a dominating random variable for the AoI variables is in a natural way a transitive property of the network.

In this work we don’t give a recipe to verify Assumption 6. However, we will see in the next subsection that the standard assumptions in the distributed optimization literature all imply Assumption 6. Another set of examples where Assumption 6 is also directly satisfied are scenarios where the network

events A_{ij}^n are driven by a geometrically ergodic Markov process [10, 8]. Of course, it can be comparatively difficult to verify this in practice. However, traditionally and also more recently it has been quite common to model network fading channels by finite Markov chains [31, 4, 23, 19, 5]. We further discuss the verifiability of Assumption 6 in Section 7.

5.4. Comparison of Assumptions 5 and 6 to assumptions in the literature

In this subsection we show that the typical network assumptions in the literature imply Assumptions 5 and 6.

First, consider the network models in [25, 3, 16, 14]. It is easy to check that network models imply the following properties:

1. There is a strongly connected graph $(\mathcal{V}, \mathcal{E})$ and some $\varepsilon > 0$, such that $\mathbb{P}\left(A_{ij}^n\right) > \varepsilon$ for all $n \in \mathbb{N}_0$ and for all $(i, j) \in \mathcal{E}$.
2. The events A_{ij}^n are independent for different time-steps or different edges.

Independence is particularly unrealistic for wireless communication systems, since transmission that occur close in time, space, frequency or code can be highly correlated. Notably, this assumptions do not show any trade off between the choice of the step size sequence $a(n)$ and some network related property. Hence, there is no trade of between the growth of the AoI variables and the choice of the step size sequence. In fact, it is easy to show that under this assumptions *all* moments of *all* $\tau_{ij}(n)$ are bounded, see Section 6.1 Example 1.

We can now show that the above properties imply Assumptions 5 and 6. Assumption 5 is directly satisfied for $\kappa = 1$. Define the σ -algebras

$$\mathcal{F}_l^m := \sigma\left(A_{ij}^n \mid l \leq n \leq m; i, j \in \mathcal{V}\right). \quad (5.3)$$

Then Assumption 6 holds trivially, since the independence of the events A_{ij}^n implies that

$$|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| = 0. \quad (5.4)$$

for $A \in \mathcal{F}_0^l$ and $B \in \mathcal{F}_{l+n}^\infty$ for all $l, n \in \mathbb{N}$. Hence, the mixing coefficients $\alpha_{ij}(n)$ for each process A_{ij}^n satisfies $\alpha_{ij}(n) = 0$ for every $n \geq 0$.

Second, consider the time-varying network in [33, 22, 32, 28, 1, 34, 15]. The authors assume that their network is M -strongly connected. Hence, they assume guaranteed periodic communication. Assumption 5 is therefore directly satisfied by choosing $\kappa = M$. Then $\mathbb{P}\left(\bigcup_{k=n}^{n+\kappa} A_{ij}^k\right) = 1$ for all $n \in \mathbb{N}_0$. Assumption 6 is also directly satisfied by choosing $\eta = M$. To see this, fix any $n, m \geq 0$ with $m \neq n$. Then

$$\mathbb{P}\left(\left(\bigcup_{k=n}^{n+\eta} A_{ij}^k\right) \cap \left(\bigcup_{k=m}^{m+\eta} A_{ij}^k\right)\right) = 1, \quad (5.5)$$

since the intersection of almost sure events is an almost sure event. Therefore,

$$\mathbb{P}\left(\left(\bigcup_{k=n}^{n+\eta} A_{ij}^k\right) \cap \left(\bigcup_{k=m}^{m+\eta} A_{ij}^k\right)\right) - \mathbb{P}\left(\bigcup_{k=n}^{n+\eta} A_{ij}^k\right)\mathbb{P}\left(\bigcup_{k=m}^{m+\eta} A_{ij}^k\right) = 0 \quad (5.6)$$

and Assumption 6 follows.

We have therefore shown that the network models in the literature satisfy Assumptions 5 and 6. Moreover, Assumptions 5 and 6 are significantly weaker, since they do not require independent communication or guaranteed periodic communication, but merely asymptotic independence.

6. Stochastic dominance properties of AoI for Time-Varying Networks

In this section, we show that Assumptions 5 and 6 imply Assumption 1.1. Recall that the AoI variables $\tau_{ij}(n)$, as defined in Section 3, are now a consequence of the network model formulated in Section 5.1. Each agent tries to send its local believe vector \hat{X}_i^n (Equation (3.2)) to some other agents. A successful transmission to some other agent j is represented by an edge $(i, j) \in \mathcal{E}^n$ of the time-varying network $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$ or equivalently by the event A_{ij}^n .

Recall that Assumption 1.1 requires finite moment properties of a random variable that stochastically dominates (Definition 1) all $\tau_{ij}(n)$. The following definition will be useful to formulate our main result and the subsequent proof.

Definition 7. We say an AoI variable $\tau_{ij}(n)$ is **stochastically dominated with finite p -th moment** for some $p \geq 0$, if there exists a non-negative integer-valued random variable $\bar{\tau}$ that stochastically dominates all $\tau_{ij}(n)$ for and all $n \in \mathbb{N}_0$ with $\mathbb{E}[\bar{\tau}^p] < \infty$.

The following theorem formulates the main result of this section.

Theorem 2. Let $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$ be a time-varying network that is (ε, κ) -SSC (Definition 6) with associated strongly connected graph $(\mathcal{V}, \mathcal{E})$. If for each $(i, j) \in \mathcal{E}$, there is some $\eta \in \mathbb{N}_0$, such that the process $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^k}$ is p -strongly mixing (Definition 3) for some $p \geq 0$, then all AoI variables $\tau_{ij}(n)$ are stochastically dominated by a single random variable with finite p -th moment.

Stochastic dominance with finite 0-th moment corresponds to the mere existence of a dominating random variable without any necessary moment condition. Theorem 2 shows a more general result as it would be required for the convergence of Algorithm 1. It is shown for all $p \in [0, \infty)$. The following corollary is now immediate and requires Theorem 2 for $p \in [1, 2)$.

Corollary 1. Under Assumptions 2 to 6, we have that Algorithm 1 converges almost surely to a λ -neighbourhood of the set of stationary points of F , where λ is the almost sure bound of the additive errors according to Assumption 4.

Proof. Under Assumption 5 and 6, it follows from Theorem 2 that Assumption 1.1 holds for some $p \in [1, 2)$. We can then choose a step size sequence $a(n)$ that is not summable, but square summable with $a(n) \in \mathcal{O}(n^{-\frac{1}{p}})$ and therefore also satisfy Assumption 1.2. The requirements of Theorem 1 are therefore satisfied and the statement of the corollary follows. \square

The rest of this section is devoted to the proof of Theorem 2. We begin by describing a general construction/recipe to establish the stochastic dominance properties for AoI variables of time-varying networks. In addition, we illustrate the recipe for the scenario where the edge events A_{ij}^n are independent. Afterwards, we give the proof of Theorem 2. Before proceeding, we show a preliminary property of the AoI variables for an (ε, κ) -SSC network.

Lemma 3. *Let $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$ be a time-varying network that is (ε, κ) -SSC with associated strongly connected graph $(\mathcal{V}, \mathcal{E})$, then for all $(i, j) \in \mathcal{E}$ we have*

$$\mathbb{P}(\tau_{ij}(n) > m) < \varepsilon, \quad \forall m, n \geq \kappa, \quad (6.1)$$

Proof. First, we have $\mathbb{P}(\tau_{ij}(n) > m) = 0$ for $m \geq n$, since $\tau_{ij}(n) \leq n$. We therefore concentrate on $m < n$. Fix $(i, j) \in \mathcal{E}$, i.e. i and j are agents that can communicate directly. Observe that successful direct communication from i to j during any time interval of the form $[n - m + 1, n]$ implies that the AoI at time n is less than m . In other words, we have the following inclusion

$$\left\{ \bigcup_{l=n-m+1}^n A_{ij}^l \right\} \subset \{\tau_{ij}(n) \leq m\}. \quad (6.2)$$

Since the network is (ε, κ) -SSC, we have that

$$\mathbb{P}(\tau_{ij}(n) \leq m) \geq \mathbb{P}\left(\bigcup_{l=n-m+1}^n A_{ij}^l\right) \geq \varepsilon, \quad \forall n > m \geq M \quad (6.3)$$

The complementary event of the previous expression therefore concludes the proof of the lemma. \square

6.1. A construction to establish stochastic dominance properties

We now describe a general construction to establish the stochastic dominance properties with some finite p -th moment (Definition 7) for an AoI variable $\tau_{ij}(n)$. The idea is to find a uniform upper bound $u : \mathbb{N}_0 \rightarrow \mathbb{R}_{\geq 0}$, such that

$$\mathbb{P}(\tau_{ij}(n) > m) \leq u(m)$$

for all $m \geq N$ independent of $n \in \mathbb{N}_0$ for some $N \in \mathbb{N}_0$ and $\lim_{m \rightarrow \infty} u(m) = 0$. We can now use this bound to define the CDF of a new random variable. Since $\lim_{m \rightarrow \infty} u(m) = 0$ there is some $M \in \mathbb{N}_0$, such that $u(m) \leq 1$ for all $m \geq M \geq N$. Now define a non-negative integer-valued random variable $\bar{\tau}_{ij}$ by describing its CDF (more precisely its complementary CDF) as follows:

$$\mathbb{P}(\bar{\tau}_{ij} > m) = 1, \quad 0 \leq m < M, \quad (6.4)$$

$$\mathbb{P}(\bar{\tau}_{ij} > m) = u(m), \quad m \geq M. \quad (6.5)$$

By definition $\bar{\tau}_{ij}$ stochastically dominates all $\tau_{ij}(n)$ for all $n \in \mathbb{N}_0$. Moreover, if

$$\sum_{m=0}^{\infty} ((m+1)^p - m^p) u(m) < \infty$$

for some $p > 0$, then it will follow from Proposition 1 that $\tau_{ij}(n)$ is stochastically dominated with finite p -th moment.

As the next step, we describe how we can find a function $u(m)$ for the above construction. Consider a (ε, κ) -SSC network. Let $(\mathcal{V}, \mathcal{E})$ be the strongly connected graph associated with the (ε, κ) -SSC network

and fix an edge $(i, j) \in \mathcal{E}$. Let $\Delta(m)$ be an increasing sequence in \mathbb{N} , with $\lim_{m \rightarrow \infty} \Delta(m) = \infty$. Now for each $n, m \in \mathbb{N}_0$ use this sequence to define time indices

$$n_1 := n - m + \Delta(m), \quad n_k := n_{k-1} + 2\Delta(m) \quad (6.6)$$

as long as $n_k \leq n$. Let $L(m)$ be the number of constructed time indices and observe that

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \mathbb{P}\left(\bigcap_{k=1}^{L(m)} \{\tau_{ij}(n_k) > \Delta(m)\}\right). \quad (6.7)$$

This follows since $\tau_{ij}(n) > m$ implies $\tau_{ij}(n_k) > \Delta(m)$ for all $k \in \{1, \dots, L(m)\}$ by the very construction of the time indices n_k . In general, we can now derive $u(m)$ as an upper bound to the right-hand side in (6.7), which we illustrate immediately for case of independent network communication, i.e. were the events A_{ij}^n are independent. For the case of dependent network communication, this will be formulated in Lemma 4 in the next section.

Example 1 (Independent network communication). Let $(\mathcal{V}, \mathcal{E})$ be the strongly connected graph associated with a (ε, κ) -SSC network and consider an edge $(i, j) \in \mathcal{E}$. Using the exemplary network independence and Lemma 3, we have from (6.7) that

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \prod_{k=1}^{L(m)} \mathbb{P}(\tau_{ij}(n_k) > \Delta(m)) < \varepsilon^{L(m)}, \quad (6.8)$$

for all m large enough such that $\Delta(m) \geq \kappa$. Now define

$$u(m) := \varepsilon^{L(m)}, \quad \Delta(m) \approx \sqrt{m}$$

and hence $L(m) \approx \sqrt{m}/2$. The construction described above then yields a dominating random variable $\overline{\tau}$ for all $\tau_{ij}(n)$ for all $n \in \mathbb{N}_0$. It is now easy to verify that

$$\mathbb{E}[\overline{\tau}^p] \leq \sum_{m=0}^{\infty} ((m+1)^p - m^p) u(m) \approx \sum_{m=0}^{\infty} ((m+1)^p - m^p) \varepsilon^{\sqrt{m}/2} < \infty$$

for all $p \geq 0$, since the series is a version of a weighted geometric series. We have therefore established that with independent communication, each AoI variable $\tau_{ij}(n)$ with $(i, j) \in \mathcal{E}$ is stochastically dominated with finite p -th moment for every $p \geq 0$. This underlines how strong the assumption of independent communication is.

6.2. Proof of Theorem 2

In the previous example, we used the independence of the edge events A_{ij}^n to establish a uniform upper bound for $\mathbb{P}(\tau_{ij}(n) > m)$ with geometric decay. Recall that $\Delta(m)$ was used in (6.6) to define the time indices n_k , such that $n_k - \Delta(m) - n_{k-1} = \Delta(m)$. Now consider the case where the edge events are not independent but merely mixing. We will see that we can then find a new upper bound to (6.7), such that

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \mathbb{P}\left(\bigcap_{k=1}^{L(m)} \{\tau_{ij}(n_k) > \Delta(m)\}\right) \leq \varepsilon^{L(m)} + \text{error}(\Delta(m)). \quad (6.9)$$

with an error term $error(\Delta(m))$ due to the non independence.

Now, if the mixing coefficients associated with processes $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^n}$ decay rapidly enough, we expect that $error(\Delta(m))$ decays sufficiently, such that the new upper bound still satisfies some summability properties and hence allows that we establish stochastic dominance properties. The following lemma makes this intuition precise. We establishes the stochastic dominance property of order $p \geq 0$ for those network edges (i, j) that ensure that the network is (ε, κ) -SSC.

Lemma 4. *Let $\{(\mathcal{V}, \mathcal{E}^n)\}_{n \in \mathbb{N}_0}$ be a time-varying network that is (ε, κ) -SSC (Definition 6) with associated strongly connected graph $(\mathcal{V}, \mathcal{E})$. If for any $(i, j) \in \mathcal{E}$ the process $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^n}$ is p -strongly mixing (Definition 3) for some $p \geq 0$ and some $\eta \in \mathbb{N}_0$, then $\tau_{ij}(n)$ is stochastically dominated with finite p -th moment (Definition 7).*

Proof. Fix an edge $(i, j) \in \mathcal{E}$. The theme of the proof is to establish a uniform upper bound to the complementary CDF of $\tau_{ij}(n)$ independent of n , such that the construction from Section 6.1 yields the required dominating random variable.

Step 1 (Reduction to $\eta = 0$): The p -strongly mixing property of the network guarantees mixing of the process $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^n}$ for some $\eta \in \mathbb{N}_0$. W.l.o.g. we can assume that $\eta = 0$. This is justified as follows. Lets denote by $\tau_{ij}^\eta(k)$ a new random variable that captures the time, since the last interval of the form $[m\eta, (m+1)\eta]$ with at least one successful transmission from i to j . The case $\eta = 0$ then yields the conclusion of the Lemma for $\tau_{ij}^\eta(k)$, i.e. there will be random variable $\bar{\tau}_{ij}^\eta$ that stochastically dominates all $\tau_{ij}^\eta(k)$ with $\mathbb{E} \left[(\bar{\tau}_{ij}^\eta)^p \right] < \infty$. For any $k \geq 0$ and $n \in \{k\eta, (k+1)\eta\}$, we have $\tau_{ij}(n) \leq \eta(\tau_N(k) + 1)$. Therefore,

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \mathbb{P}\left(N(\tau_{ij}^\eta(k) + 1) > m\right) \leq \mathbb{P}\left(\eta(\bar{\tau}_{ij}^\eta + 1) > m\right) \quad (6.10)$$

and $\mathbb{E} \left[\eta^p (\bar{\tau}_{ij}^\eta + 1)^p \right] < \infty$ by Minkowski's inequality. Hence, $\eta(\tau_N(k) + 1)$ would be the required dominating random variable for $\tau_{ij}(n)$ and we may therefore assume $\eta = 0$.

Step 2 (Initial CDF bound): Fix $m \in \mathbb{N}_0$ and recall the definition of $\Delta(m)$ and the associated sequence n_k for each $n \in \mathbb{N}_0$ from Section 6.1. We have

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \mathbb{P}\left(\bigcap_{k=1}^{L(m)} \{\tau_{ij}(n_k) > \Delta(m)\}\right). \quad (6.7 \text{ recalled})$$

With a slide abuse of notation we will now refer with $\tau_{ij}(n)$ to the age of information associated with the direct information exchange from i to j . The age of information associated with direct information exchange by definition stochastically dominates the actual AoI. Without this step we would technically require a stronger mixing requirement, specifically, one for the events generated by all A_{ij}^n and not only for the events generated by A_{ij}^n for the pair (i, j) . Note that Lemma 3 also directly holds for this case, since we anyway used the direct information exchange to prove it.

We will now establish an upper bound to (6.7 recalled) using that $\mathbb{1}_{A_{ij}^n}$ is p -strongly mixing. For this, define the following sub- σ -algebras generated by the events A_{ij}^n :

$$\mathcal{F}_l^s := \sigma\left(A_{ij}^n \mid l \leq n \leq s\right), \quad l \in \mathbb{N}_0, s \in \mathbb{N}_0 \cup \{\infty\}. \quad (6.11)$$

The important generated events are, whether the AoI variables at some time step $s \in \mathbb{N}_0$ exceed a threshold $l \in \mathbb{N}_0$, i.e. whether $\{\tau_{ij}(s) > l\}$. Since the event $\{\tau_{ij}(s) > l\}$ is generated by the events A_{ij}^k with $k \in \{s-l+1, \dots, s-1, s\}$, we have that

$$\{\tau_{ij}(s) > l\} \in \mathcal{F}_{s-l+1}^s. \quad (6.12)$$

For this, we required the reduction to age of information associated with direct information exchange. It then follows by definition of the time indices n_k that

$$\{\tau_{ij}(n_{L(m)}) > \Delta(m)\} \in \mathcal{F}_{n_{L(m)}-\Delta(m)+1}^{n_{L(m)}} \subset \mathcal{F}_{n_{L(m)}-\Delta(m)}^\infty \quad (6.13)$$

and

$$\{\tau_{ij}(n_k) > \Delta(m)\} \in \mathcal{F}_{n_k-\Delta(m)+1}^{n_k} \subset \mathcal{F}_0^{n_{L(m)}-1} \quad (6.14)$$

for every $k \in \{1, \dots, L(m)-1\}$. Hence,

$$\bigcap_{k=1}^{L(m)-1} \{\tau_{ij}(n_k) > \Delta(m)\} \in \mathcal{F}_0^{n_{L(m)}-1}. \quad (6.15)$$

By construction of the indices n_k , we have $n_{L(m)} - \Delta(m) - n_{L(m)-1} = \Delta(m)$. The strong mixing property of the process $\mathbb{1}_{A_{ij}^n}$ therefore implies that

$$\begin{aligned} \mathbb{P} \left(\bigcap_{k=1}^{L(m)} \{\tau_{ij}(n_k) > \Delta(m)\} \right) &\leq \mathbb{P} \left(\{\tau_{ij}(n_{L(m)}) > \Delta(m)\} \right) \mathbb{P} \left(\bigcap_{k=1}^{L(m)-1} \{\tau_{ij}(n_k) > \Delta(m)\} \right) \\ &\quad + \alpha(\Delta(m)), \end{aligned} \quad (6.16)$$

where $\alpha(n)$ are the mixing coefficients associated with the process $\mathbb{1}_{A_{ij}^n}$. It now follows from Lemma 3 that $\mathbb{P}(\tau_{ij}(n_k) > \Delta(m)) < \varepsilon$ for $\Delta(m) \geq \kappa$, since the network is (ε, κ) -SSC. Hence,

$$\mathbb{P} \left(\bigcap_{k=1}^{L(m)} \{\tau_{ij}(n_k) > \Delta(m)\} \right) \leq \varepsilon \mathbb{P} \left(\bigcap_{k=1}^{L(m)-1} \{\tau_{ij}(n_k) > \Delta(m)\} \right) + \alpha(\Delta(m)). \quad (6.17)$$

Applying (6.16) and (6.17) successively yields:

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \prod_{k=1}^{L(m)} \mathbb{P}(\{\tau_{ij}(n_k) > \Delta(m)\}) + \sum_{k=1}^{L(m)-1} \varepsilon^{k-1} \alpha(\Delta(m)) \quad (6.18)$$

$$\leq \varepsilon^{L(m)} + \frac{1}{1-\varepsilon} \alpha(\Delta(m)). \quad (6.19)$$

for $\Delta(m) \geq \kappa$.

For $p = 0$, we can now apply the construction presented in Section 6.1 with the bound (6.19) to obtain a dominating random variable. Here we may choose $\Delta(m)$ as in Example 1. For $p > 0$ it is now crucial to choose $\Delta(m)$, such that both terms in (6.19) decay rapidly enough to obtain the required

stochastic dominance property with finite p -th moment. However, it turns out that the bound (6.19) is only sufficient to achieve this for all $q < p$, due to the merely geometric decay of the first term. The next step therefore uses (6.19) to obtain a better upper bound for (6.18).

Step 3: To improve the CDF bound for $p > 0$, we use that $\sum_{m=0}^{\infty} m^{p-1} \alpha(m)$ is summable. It then follows that for $p > 1$ we have

$$\alpha(m) \in \mathcal{O}(m^{-(p-1)}) \quad (6.20)$$

and for $0 < p \leq 1$ we have

$$\alpha(m) \in \mathcal{O}(m^{-p}), \quad (6.21)$$

since for this case $m^{p-1} \alpha(m)$ is guaranteed to be decreasing as $p - 1 < 0$. Both cases show that there is a constant c and some $\tilde{\mu} > 0$, such that

$$\alpha(\Delta(m)) \leq C(\Delta(m))^{-\tilde{\mu}} \quad (6.22)$$

for sufficiently large m . With (6.19) it then follows that

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \varepsilon^{L(m)} + c(\Delta(m))^{-\tilde{\mu}}. \quad (6.23)$$

for sufficiently large m . Since the first term above is exponential and the second is rational, we can find a new $\mu > 0$, such that

$$\mathbb{P}(\tau_{ij}(n) > m) < m^{-\mu} \quad (6.24)$$

for m sufficiently large. For this, one may again choose $\Delta(m) \approx \sqrt{m}$.

Step 4 - (Verifying the stochastic dominance property with finite p -th moment):

We now insert the CDF bound from step 4 in (6.18) and obtain

$$\mathbb{P}(\tau_{ij}(n) > m) \leq \Delta(m)^{-\mu L(m)} + \frac{1}{1-\varepsilon} \alpha(\Delta(m)) \quad (6.25)$$

for m sufficiently large. Now choose $\delta \in (0, 1)$, such that

$$\mu \left(\frac{1}{4\delta} - 1 \right) \geq p + 1 \quad (6.26)$$

and then choose $\Delta(m) = \lceil \delta m \rceil$. We choose this to guarantee the required summability property of the first term in (6.25), since

$$L(m) = \lfloor \frac{m}{2\Delta(m)} \rfloor \geq \frac{1}{4\delta} - 1 \quad (6.27)$$

for $m \geq \frac{1}{2\delta}$. Hence, we have

$$\mathbb{P}(\tau_{ij}(n) > m) \leq (\delta m)^{-(p+1)} + \frac{1}{1-\varepsilon} \alpha(\lceil \delta m \rceil) \quad (6.28)$$

for $m \geq \frac{1}{2\delta}$.

Now define

$$u(m) := (\delta m)^{-(p+1)} + \frac{1}{1-\varepsilon} \alpha(\lceil \delta m \rceil)$$

and apply the construction presented in Section 6.1. This yields a non-negative integer-valued random variable $\bar{\tau}_{ij}$ that stochastically dominates $\tau_{ij}(n)$ for all $n \in \mathbb{N}$. Moreover, we have $\mathbb{E} \left[\bar{\tau}_{ij}^p \right] < \infty$, if

$$\sum_{m=0}^{\infty} ((m+1)^p - m^p) u(m) < \infty. \quad (6.29)$$

The first part of the series is finite, since

$$\sum_{m=1}^{\infty} ((m+1)^p - m^p) (\delta m)^{-(p+1)} \leq \frac{2^p}{\delta^{p+1}} \sum_{m=1}^{\infty} m^{-2} < \infty, \quad (6.30)$$

where we used that $((m+1)^p - m^p) \leq 2^p m^{p-1}$ for $m \in \mathbb{N}$.

For the second part of the series, note that $\alpha(n)$ is by construction a monotonically decreasing function from \mathbb{N}_0 to $[0, \frac{1}{4}]$ [8]. Now extend $\alpha(n)$ by linear interpolation to a monotonically decreasing function from $\mathbb{R}_{\geq 0}$ to $[0, \frac{1}{4}]$. Then for all $m \in \mathbb{N}_0$, we have $\alpha(\lceil \delta m \rceil) \leq \alpha(\delta m)$ by monotonicity. Hence the second part is finite, since

$$\begin{aligned} \sum_{m=1}^{\infty} ((m+1)^p - m^p) \alpha(\delta m) &\leq 2^p \sum_{m=1}^{\infty} m^{p-1} \alpha(\delta m) \\ &\leq \frac{2^{2p-1}}{\delta^p} \sum_{m=1}^{\infty} m^{p-1} \alpha(m) < \infty \end{aligned} \quad (6.31)$$

The second inequality can be shown using a similar construction as in Lemma 1. Finally, the finiteness of the last summation follows from the assumed p -strongly mixing property. \square

We have thus established the stochastic dominance property of order $p \geq 0$ for those network edges that ensure that the network is SSC under the p -strongly mixing condition. As the next step, we show an elementary lemma associated with the AoI variables of a time-varying network. The lemma shows that the existence of stochastically dominating random variables associated with the AoI variables of a time-varying network is a transitive property.

Lemma 5. *For nodes $i, j, k \in \mathcal{V}$ of a time-varying network suppose $\tau_{ij}(n)$ and $\tau_{jk}(n)$ are stochastically dominated by $\bar{\tau}_{ij}$ and $\bar{\tau}_{jk}$, respectively. Then*

1. *There is random variable $\bar{\tau}_{ik}$ that stochastically dominates $\tau_{ik}(n)$.*
2. *If moreover $\mathbb{E} \left[\bar{\tau}_{ij}^p \right] + \mathbb{E} \left[\bar{\tau}_{jk}^p \right] < \infty$ for some $p > 0$, then also $\mathbb{E} \left[\bar{\tau}_{ik}^p \right] < \infty$.*

Proof. Fix $i, j, k \in \mathcal{V}$ and some $m \geq 2$. Now observe the following inclusion associated with events of the three AoI variables $\tau_{ij}(n)$, $\tau_{jk}(n)$ and $\tau_{ik}(n)$:

$$\left\{ \tau_{ij}(n - \frac{m}{2}) \leq \frac{m}{2} \right\} \cap \left\{ \tau_{jk}(n) \leq \frac{m}{2} \right\} \subset \left\{ \tau_{ik}(n) \leq m \right\}, \quad (6.32)$$

The inclusion states that the two events

1. The AoI is less than $\frac{m}{2}$ for information received at node j from node i at time $n - \frac{m}{2}$

2. The AoI is less than $\frac{m}{2}$ for information received at node k from node j at time n

imply the event that the AoI is less than m for information received at node k from node i at time n . By taking the complement of the inclusion in (6.32), we have that

$$\mathbb{P}(\tau_{ik}(n) > m) \leq \mathbb{P}\left(\{\tau_{ij}(n - \frac{m}{2}) > \frac{m}{2}\} \cup \{\tau_{jk}(n) > \frac{m}{2}\}\right) \quad (6.33)$$

$$\leq \mathbb{P}\left(\tau_{ij}(n - \frac{m}{2}) > \frac{m}{2}\right) + \mathbb{P}\left(\tau_{jk}(n) > \frac{m}{2}\right) \quad (6.34)$$

$$< \mathbb{P}\left(\bar{\tau}_{ij} > \frac{m}{2}\right) + \mathbb{P}\left(\bar{\tau}_{jk} > \frac{m}{2}\right). \quad (6.35)$$

In the last step, we used the assumption that there are random variables $\bar{\tau}_{ij}$ and $\bar{\tau}_{jk}$ that stochastically dominate $\tau_{ij}(n)$ and $\tau_{jk}(n)$, respectively, for all n .

Now $\bar{\tau}_{ij}$ and $\bar{\tau}_{jk}$ are integer-valued, so there is some $M \in \mathbb{N}$ such that

$$\mathbb{P}\left(\bar{\tau}_{ij} > \frac{m}{2}\right) + \mathbb{P}\left(\bar{\tau}_{jk} > \frac{m}{2}\right) < 1 \quad (6.36)$$

for all $m \geq M$. Define a non-negative integer-valued random variable $\bar{\tau}_{ik}$ by defining its CDF:

$$\mathbb{P}(\bar{\tau}_{ik} > m) := 1, \quad \text{for all } 0 \leq m < M, \quad (6.37)$$

$$\mathbb{P}(\bar{\tau}_{ik} > m) := \mathbb{P}\left(\bar{\tau}_{ij} > \frac{m}{2}\right) + \mathbb{P}\left(\bar{\tau}_{jk} > \frac{m}{2}\right), \quad \text{otherwise.} \quad (6.38)$$

This proves part (a) of the lemma.

Now suppose $\mathbb{E}\left[\bar{\tau}_{ij}^p\right] + \mathbb{E}\left[\bar{\tau}_{jk}^p\right] < \infty$ for some $p > 0$. We can now write the p -th moment of $\bar{\tau}_{ik}$ using its CDF from above:

$$\mathbb{E}\left[\bar{\tau}_{ik}^p\right] = \sum_{m=0}^{\infty} ((m+1)^p - m^p) \mathbb{P}(\bar{\tau}_{ik} > m) \quad (6.39)$$

$$\leq \sum_{m=0}^{\infty} ((m+1)^p - m^p) \mathbb{P}\left(\bar{\tau}_{ij} > \frac{m}{2}\right) + \sum_{m=0}^{\infty} ((m+1)^p - m^p) \mathbb{P}\left(\bar{\tau}_{jk} > \frac{m}{2}\right) \quad (6.40)$$

$$= 2^p \left(\mathbb{E}\left[\bar{\tau}_{ij}^p\right] + \mathbb{E}\left[\bar{\tau}_{jk}^p\right] \right) < \infty. \quad (6.41)$$

Where the equality follows from Proposition 1, since $2\bar{\tau}_{ij}$ and $2\bar{\tau}_{jk}$ are non-negative integer-valued random variables. This proves part (b) of the lemma. \square

Lemma 5 allows that we extend the stochastic dominance properties from Lemma 4 for node pairs $(i, j) \in \mathcal{E}$ to arbitrary node pairs $(i, j) \in \mathcal{V}^2$. We are now ready to prove Theorem 2.

Proof of Theorem 2. First, fix an arbitrary pairs of nodes $(i, j) \in \mathcal{V}^2$. Since the network is SSC, it follows from Lemma 4 there is a sequence of edges $\{(i_k, i_{k+1})\}_{k=1}^{K-1} \in \mathcal{E}$ for some $K \geq 1$, with $i_1 = i$ and $i_K = j$, such that for each $\tau_{i_k i_{k+1}}$, there is non-negative integer-valued random variable $\bar{\tau}_{i_k i_{k+1}}$ that stochastically dominates all $\tau_{i_k i_{k+1}}(n)$ for all $n \in \mathbb{N}_0$, with $\mathbb{E}\left[\bar{\tau}_{i_k i_{k+1}}^p\right] < \infty$. It now follows by

induction using the transitive property of the AoI variables from Lemma 5(b), that there is a non-negative integer-valued random variable $\bar{\tau}_{ij}$ that stochastically dominates all $\tau_{ij}(n)$ for all $n \in \mathbb{N}_0$, with $\mathbb{E} \left[\bar{\tau}_{ij}^p \right] < \infty$.

It is now left to verify that there is a single dominating random variables for all pairs $(i, j) \in \mathcal{V}^2$. This essentially follows since we consider finitely many agents. For every $m \geq 0$, define

$$h(m) := \sum_{(i,j) \in \mathcal{V}^2} \mathbb{P}(\bar{\tau}_{ij} > m). \quad (6.42)$$

Since $|\mathcal{V}^2| < \infty$, there is some $M \geq 0$, such that $h(m) \leq 1$ for all $m \geq M$. Define a non-negative integer-valued random variable $\bar{\tau}$ by describing its CDF as follows:

$$\mathbb{P}(\bar{\tau}_{ij} > m) = 1, \quad 0 \leq m < M, \quad (6.43)$$

$$\mathbb{P}(\bar{\tau}_{ij} > m) = h(m), \quad m \geq M. \quad (6.44)$$

By construction $\bar{\tau}$ stochastically dominates all $\tau_{ij}(n)$ for all $(i, j) \in \mathcal{V}^2$ and for all $n \in \mathbb{N}_0$. Finally, we have

$$\mathbb{E}[\bar{\tau}^p] \leq \sum_{m=0}^{\infty} ((m+1)^p - m^p) h(m) = \sum_{(i,j) \in \mathcal{V}^2} \mathbb{E} \left[\bar{\tau}_{ij}^p \right] < \infty, \quad (6.45)$$

where the equality simply follows from continuity of addition and since all $\mathbb{E} \left[\bar{\tau}_{ij}^p \right]$ are convergent. \square

7. Conclusions and future work

In this work, we presented an asymptotic convergence analysis of distributed stochastic gradient descent that uses aged information. The required network assumptions have been weakened to the mere existence of non-negative integer-valued random variable with finite first moment that stochastically dominates all age of information random variables variables. This assumption can be satisfied with the new network Assumptions 5 and 6. These assumptions are significantly weaker then the common network assumptions in the literature. We hope that our assumptions penalize future work in distributed optimization under less restrictive network assumptions. Notably, instead of periodic or independent communication, we merely require asymptotically independent communication formulated using α -mixing with the minimal requirement that $\sum_{n=0}^{\infty} \alpha(n) < \infty$.

It would be interesting to see, whether summability properties of α -mixing coefficients indeed hold for representative physical wireless communication system. This might be possible when the underlying physical system has a mixing property in an ergodic sense. For example, hyperbolic systems are common models to describe electro magnetic wave propagation and it was shown in [2] that hyperbolic systems admit a strong mixing property in an ergodic sense.

To apply Assumption 6 in practice, it would be most desirable if the α -mixing coefficients (or an upper bound) for the network processes $\mathbb{1}_{\bigcup_{k=n}^{n+\eta} A_{ij}^k}$ could be estimated from data. Unfortunately, there are only a handful of methods that estimate or approximate the mixing coefficients from data. One method that uses an approximation method based on histograms was presented in [20]. However, this method suffers from high complexity. Very recently, a new method was presented in [13]. Most notably, the work presents a hypothesis test to decide, whether the sum of the alpha mixing coefficients is

below an upper bound. With this, it is therefore now possible to verify with high confidence, whether Assumption 6 holds for $p = 1$ using data.

Funding

Adrian Redder was supported by the German Research Foundation (DFG) - 315248657 and SFB 901.

References

- [1] AYBAT, N. S. and HAMEDANI, E. Y. (2019). A distributed ADMM-like method for resource sharing over time-varying networks. *SIAM Journal on Optimization* **29** 3036–3068.
- [2] BABILLOT, M. (2002). On the mixing property for hyperbolic systems. *Israel journal of mathematics* **129** 61–76.
- [3] BASTIANELLO, N., CARLI, R., SCHENATO, L. and TODESCATO, M. (2020). Asynchronous distributed optimization over lossy networks via relaxed ADMM: Stability and linear convergence. *IEEE Transactions on Automatic Control* **66** 2620–2635.
- [4] BIANCHI, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on selected areas in communications* **18** 535–547.
- [5] BOBAN, M., GONG, X. and XU, W. (2016). Modeling the evolution of line-of-sight blockage for V2V channels. In *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)* 1–7. IEEE.
- [6] BORKAR, V. S. (1998). Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization* **36** 840–851.
- [7] BORKAR, V. S. (2009). *Stochastic approximation: A dynamical systems viewpoint*. Springer.
- [8] BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys* **2** 107–144.
- [9] CHAKRABORTI, S., JARDIM, F. and EPPRECHT, E. (2018). Higher-order moments using the survival function: The alternative expectation formula. *The American Statistician*.
- [10] DAVYDOV, Y. A. (1974). Mixing conditions for Markov chains. *Theory of Probability & Its Applications* **18** 312–328.
- [11] DURRETT, R. (2019). *Probability: Theory and Examples*. Cambridge University Press.
- [12] HAGHSHENAS, K., PAHLEVAN, A., ZAPATER, M., MOHAMMADI, S. and ATIENZA, D. (2019). Magnetic: Multi-agent machine learning-based approach for energy efficient dynamic consolidation in data centers. *IEEE Transactions on Services Computing*.
- [13] KHALEGHI, A. and LUGOSI, G. (2021). Inferring the mixing properties of an ergodic process. *arXiv preprint arXiv:2106.07054*.
- [14] KOLOSKOVA, A., LOIZOU, N., BOREIRI, S., JAGGI, M. and STICH, S. (2020). A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning* 5381–5393. PMLR.
- [15] KOVALEV, D., SHULGIN, E., RICHTÁRIK, P., ROGOZIN, A. and GASNIKOV, A. (2021). ADOM: Accelerated decentralized optimization method for time-varying networks. In *International Conference on Machine Learning*. PMLR.
- [16] LEI, J., CHEN, H.-F. and FANG, H.-T. (2018). Asymptotic Properties of Primal-Dual Algorithm for Distributed Stochastic Optimization over Random Networks with Imperfect Communications. *SIAM Journal on Control and Optimization* **56** 2159–2188.
- [17] LIM, H. and KIM, C. (2001). Flooding in wireless ad hoc networks. *Computer Communications* **24** 353–363.

- [18] LIN, W. and BITAR, E. (2017). Decentralized stochastic control of distributed energy resources. *IEEE Transactions on Power Systems* **33** 888–900.
- [19] LIN, S., KONG, L., HE, L., GUAN, K., AI, B., ZHONG, Z. and BRISO-RODRÍGUEZ, C. (2015). Finite-state Markov modeling for high-speed railway fading channels. *IEEE Antennas and Wireless Propagation Letters* **14** 954–957.
- [20] MCDONALD, D. J., SHALIZI, C. R. and SCHERVISH, M. (2015). Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics* **9** 2855–2883.
- [21] NEDIĆ, A. and OLSHEVSKY, A. (2014). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control* **60** 601–615.
- [22] NEDIC, A., OLSHEVSKY, A. and SHI, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization* **27** 2597–2633.
- [23] PIMENTEL, C., FALK, T. H. and LISBÔA, L. (2004). Finite-state Markov modeling of correlated Rician-fading channels. *IEEE transactions on vehicular technology* **53** 1491–1501.
- [24] RAMASWAMY, A., REDDER, A. and QUEVEDO, D. E. (2019). Optimization over time-varying networks with unbounded delays. *arXiv preprint arXiv:1912.07055*.
- [25] RAMASWAMY, A., REDDER, A. and QUEVEDO, D. E. (2021). Distributed optimization over time-varying networks with stochastic information delays. *IEEE Transactions on Automatic Control*.
- [26] REDDER, A., RAMASWAMY, A. and KARL, H. (2022a). Asymptotic Convergence of Deep Multi-Agent Actor-Critic Algorithms. *arXiv preprint arXiv:2201.00570*.
- [27] REDDER, A., RAMASWAMY, A. and KARL, H. (2022b). Multi-agent gradient-based resource allocation for networked systems (To appear).
- [28] SCUTARI, G. and SUN, Y. (2019). Distributed nonconvex constrained optimization over time-varying digraphs. *Math. Program.* **176** 497–544.
- [29] TANG, H., GAN, S., ZHANG, C., ZHANG, T. and LIU, J. (2018). Communication Compression for Decentralized Training. *Advances in Neural Information Processing Systems* **31** 7652–7662.
- [30] TRUDEAU, R. J. (1993). *Introduction to Graph Theory*. Courier Corporation.
- [31] WANG, H. S. and MOAYERI, N. (1995). Finite-state Markov channel—a useful model for radio communication channels. *IEEE transactions on vehicular technology* **44** 163–171.
- [32] WANG, Y., ZHAO, W., HONG, Y. and ZAMANI, M. (2019). Distributed Subgradient-Free Stochastic Optimization Algorithm for Nonsmooth Convex Functions over Time-Varying Networks. *SIAM Journal on Control and Optimization* **57** 2821–2842.
- [33] XU, Y., HAN, T., CAI, K., LIN, Z., YAN, G. and FU, M. (2017). A distributed algorithm for resource allocation over dynamic digraphs. *IEEE Transactions on Signal Processing* **65** 2600–2612.
- [34] YU, Z., HO, D. W. and YUAN, D. (2020). Distributed Stochastic Optimization over Time-Varying Noisy Network. *arXiv preprint arXiv:2005.03982*.