Quantized average consensus via dynamic coding/decoding schemes

Ruggero Carli[1,*], Francesco Bullo[1], and Sandro Zampieri[2]

[1] *Center for Control, Dynamical Systems and Computation, University of California at Santa Barbara,*
Santa Barbara, CA 93106, USA,
[2] *Department of Information Engineering, University of Padova, Via Gradenigo 6/a, 35131 Padova, Italy*

SUMMARY

In the average consensus a set of linear systems has to be driven to the same final state which corresponds to the average of their initial states. This mathematical problem can be seen as the simplest example of coordination task. In fact it can be used to model both the control of multiple autonomous vehicles which all have to be driven to the centroid of the initial positions, and to model the decentralized estimation of a quantity from multiple measure coming from distributed sensors. This contribution presents a consensus strategy in which the systems can exchange information among themselves according to a fixed strongly connected digital communication network. Beside the decentralized computational aspects induced by the choice of the communication network, we here have also to face the quantization effects due to the digital links. We here present and discuss two different encoding/decoding strategies with theoretical and simulation results on their performance.

KEY WORDS: Average consensus, quantization, coding/decoding schemes

1. Introduction

A basic aspect in the analysis and in the design of cooperative agent systems is related to the effect of the agents' information exchange on the coordination performance. A coordination task which is widely treated in the literature is the so called average consensus. This is the problem of driving states of a set of dynamic systems to a final common state which corresponds to the average of initial states of each system. This mathematical problem can be shown to be relevant in the control of multiple autonomous vehicles which all have to be driven to the centroid of the initial positions, and in the decentralized estimation of a quantity from multiple measures coming from distributed sensors. The way in which the information flow on the network influences the consensus performance has been already considered in the literature [1, 2], where the communication cost is modeled simply by the number of active links in the network which admit the transmission of real numbers. However, this model can

[*] Correspondence to: carlirug@engineering.ucsb.edu

be too rough when the network links represent actual digital communication channels. Indeed the transmission over a finite alphabet requires the design of efficient ways to translate real numbers into digital information, namely smart quantization techniques.

The investigation of consensus under quantized communication started with [3]. In this paper the authors study systems having (and transmitting) integer-valued states and propose a class of gossip algorithms which preserve the average of states and are guaranteed to converge up to one quantization bin. Besides the fact there is not precise consensus, since the algorithm requires the use of a single link per iteration, the convergence is very slow. The authors in [4] analyzed the impact of the quantization noise through modification of the consensus algorithm proposed in [5], where the case of noisy communication links is addressed. Precisely, the authors in [5] consider the case in which the information transmitted by each system is corrupted by additive zero-mean noise. In [4] it is noted that the noise component can be considered as the quantization noise and by simulations, it is shown for small N that, if the increasing correlation among the states of the systems is taken into account, the variance of the quantization noise diminishes and systems converge to a consensus. In [6] the authors propose a distributed algorithm that uses quantized values and preserves the average at each iteration. They showed favorable convergence properties using simulations on some static topologies, and provided performance bounds for the limit points of the generated iterates. The authors in [7] adopt the probabilistic quantization scheme to quantize the information before transmitting to the neighboring sensors. By proposing a iterative scheme to update the state at each sensor node utilizing only quantized information communication, they show that, almost surely, the node states reach consensus to a quantized level; only in expectation they converge to the desired average. Moreover if the quantization step size is large this approach will lead to large residual errors. Of note is that all the papers mentioned above considered quantized strategy that either maintain the average of the state but do not converge to the consensus, or converge to a consensus that, since average is not preserve, does not coincide with the average of the initial conditions.

The main contribution of this paper is to propose a novel quantized strategy that permits both to maintain the initial average and to reach it asymptotically. A similar approach has been introduced in the context of the multi-agent coordination laws, rendezvous and deployment [8]. Precisely, this novel strategy adapts coding/decoding strategies, that were proposed for centralized control and communication problems, to the distributed consensus problem. In particular, two coding/decoding strategies, one based on the exchange of logarithmically quantized information, the other on a zoom in - zoom out strategy (this latter involves the use of uniform quantizers) are considered. In this paper we provide analytical and simulations results illustrating the convergence properties of these strategies. In particular we show that the convergence factors depend smoothly on the accuracy parameter of the quantizers used and, remarkably, that the critical quantizer accuracy sufficient to guarantee convergence is independent from the network dimension.

The paper is organized as follows. Section 2 briefly reviews the standard average consensus algorithm. In Section 3 we present two strategies of coding/decoding of the data throughout reliable digital channels: one based on logarithmic quantizers, the other on uniform quantizers. We analyze the former from a theoretical point in Section 4 and Section 5. We provide simulations results for the latter in Section 6. Finally, we gather our conclusions in Section 7.

Mathematical Preliminaries

Before proceeding, we collect some useful definitions and notations. In this paper, $\mathcal{G} = (V, E)$ denotes an *undirected graph* where $V = \{1, \dots, N\}$ is the set of vertices and E is the set of (directed) edges, i.e., a subset of $V \times V$. Clearly, if $(i, j) \in E$ also $(j, i) \in E$. A *path* in $\mathcal{G}$ consists of a sequence of vertices $(i_1, i_2, \dots, i_r)$ such that $(i_j, i_{j+1}) \in E$ for every $j \in \{1, \dots, r-1\}$. A graph $\mathcal{G}$ is *connected* if for any pair of vertices (i, j) there exists a path connecting i to j. A matrix M is *nonnegative* if $M_{ij} \geq 0$ for all i and j. A square matrix M is *stochastic* if it is nonnegative and the sum along each row of M is equal to 1. Moreover, a square matrix M is *doubly stochastic* if it is stochastic and the sum along each column of M is equal to 1. Given a nonnegative matrix $M \in \mathbb{R}^{N \times N}$, we define the induced graph $\mathcal{G}_M$ by taking N nodes and putting an edge (j, i) in E if $M_{ij} > 0$. Given a graph $\mathcal{G}$ on V, the matrix M is *adapted* to, or *compatible* with, $\mathcal{G}$ if $\mathcal{G}_M \subset \mathcal{G}$.

Now we give some notational conventions. Given a vector $v \in \mathbb{R}^N$ and a matrix $M \in \mathbb{R}^{N \times N}$, we let v^T and M^T respectively denote the transpose of v and of M. We let $\sigma(M)$ denote the set of eigenvalues of M. In particular, if M is symmetric and stochastic we will assume that

$$\sigma(M) = \{1, \lambda_1(M), \dots, \lambda_{N-1}(M)\},$$

where $1, \lambda_1(M), \dots, \lambda_{N-1}(M)$ denote the eigenvalues of M and are such that $\lambda_1(M) \geq \lambda_2(M) \geq \dots \geq \lambda_{N-1}(M)$. We define

$$\lambda_{\max}(M) = \lambda_1(M), \qquad \lambda_{\min}(M) = \lambda_{N-1}(M).$$

If there is no risk of confusion regarding the matrix we are considering, we will use the easy notation λ_i, for $i \in \{1, \dots, N-1\}$, and $\lambda_{\min}$ and $\lambda_{\max}$. With the symbols $\mathbf{1}$ and $\mathbf{0}$ we denote the N-dimensional vectors having respectively all the components equal to 1 and equal to 0. Given $v = [v_1, \dots, v_N]^T \in \mathbb{R}^N$, $\operatorname{diag}\{v\}$ or $\operatorname{diag}\{v_1, \dots, v_N\}$ mean a diagonal matrix having the components of v as diagonal elements. Moreover, $\|v\|$ and $< v >$ denote the Euclidean norm of v and the subspace generated by v, respectively. Finally, for $f, g : \mathbb{N} \to \mathbb{R}$, we say that $f \in o(g)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

2. Problem Formulation

We start this section by briefly describing the standard discrete-time consensus algorithm. Assume that we have a set of agents V and a graph $\mathcal{G}$ on V describing the feasible communications among the agents. For each agent $i \in V$ we denote by $x_i(t)$ the estimate of the average of agent i at time t. Standard consensus algorithm are constructed by choosing a doubly stochastic matrix $P \in \mathbb{R}^{N \times N}$ compatible with $\mathcal{G}$ and assuming that at every time t agent i updates its estimate according to

$$x_i(t+1) = \sum_{j=1}^{N} P_{ij} x_j(t). \tag{1}$$

More compactly we can write

$$x(t+1) = Px(t), \tag{2}$$

where $x(t)$ is the column vector whose entries $x_i(t)$ represent the agents states. In our treatment we will restrict to the case in which P is symmetric, i.e., $P^T = P$. Note that a stochastic symmetric matrix P is automatically doubly stochastic.

It is well known in the literature [1] that, if P is a symmetric stochastic matrix with positive diagonal entries and such that $\mathcal{G}_P$ is connected, then the algorithm (2) solves the *average consensus problem*, namely

$$\lim_{t \to +\infty} x(t) = \left(\frac{1}{N} \sum_{i=1}^{N} x_i(0) \right) \mathbf{1}.$$

From now on we will assume the following property.

Assumption 2.1. *P is a symmetric stochastic matrix such that $P_{ii} > 0$, for $i \in \{1, \dots, N\}$, and $\mathcal{G}_P$ is connected.*

Note that the algorithm (2) relies upon a crucial assumption: each agent transmits to its neighboring agents the precise value of its state. This implies the exchange of perfect information through the communication network. In what follows, we consider a more realistic case, i.e., we assume that the communication network is constituted only of rate-constrained digital links. Accordingly, the main objectives of this paper are to understand

(1) how the standard consensus algorithm may be modified to overcome the forced quantization effects due to the digital channel, and
(2) how much does its performance degrade.

We note that the presence of a rate constraint prevents the agents from having a precise knowledge about the state of the other agents. In fact, through a digital channel, the i-th agent can only send to the j-th agent symbolic data in a finite or countable alphabet; using only this data, the j-th agent can build at most an estimate of the i-th agent's state. To tackle this problem we take a two step approach. First, we introduce a coding/decoding scheme; each agent uses this scheme to estimate the state of its neighbors. Second, we consider the standard consensus algorithm where, in place of the exact knowledge of the states of the agents, we substitute estimates calculated according to the proposed coding/decoding scheme.

3. Coder/decoder pairs for digital channels

In this section we discuss a general and two specific coder/decoder models for reliable digital channels; we follow the treatment in the survey [9]. We will later adopt this coder/decoder structure to define communication protocols between the agents.

Suppose a source wants to communicate to a receiver some time-varying data $x : \mathbb{N} \to \mathbb{R}$ via repeated transmissions at time instants in $\mathbb{N}$. Each transmission takes place through a digital channel, i.e., messages can only be symbols in a finite or countable set (to be designed). The channel is assumed to be reliable, that is, each transmitted symbol is received without error. A coder/decoder pair for a digital channel is defined by the sets:

(i) a set Ξ, serving as *state space* for the coder/decoder; a fixed $\xi_0 \in \Xi$ is the *initial coder/decoder state*;

(ii) a finite or countable set $\mathcal{A}$, serving as *transmission alphabet*; elements $\alpha \in \mathcal{A}$ are called messages;

and by the maps:

(i) a map $F : \Xi \times \mathcal{A} \to \Xi$, called the *coder/decoder dynamics*;
(ii) a map $Q : \Xi \times \mathbb{R} \to \mathcal{A}$, being the *quantizer function*;
(iii) a map $H : \Xi \times \mathcal{A} \to \mathbb{R}$, called the *decoder function.*

The coder computes the symbols to be transmitted according to, for $t \in \mathbb{N}$,

$$\xi(t+1) = F(\xi(t), \alpha(t)), \quad \alpha(t) = Q(\xi(t), x(t)).$$

Correspondingly, the decoder implements, for $t \in \mathbb{N}$,

$$\xi(t+1) = F(\xi(t), \alpha(t)), \quad \hat{x}(t) = H(\xi(t), \alpha(t)).$$

Coder and decoder are jointly initialized at $\xi(0) = \xi_0$. Note that an equivalent representation for the coder is $\xi(t+1) = F(\xi(t), Q(\xi(t), x(t)))$, and $\alpha(t) = Q(\xi(t), x(t))$. In summary, the coder/decoder dynamics is given by

$$\begin{aligned} \xi(t+1) &= F(\xi(t), \alpha(t)), \\ \alpha(t) &= Q(\xi(t), x(t)), \\ \hat{x}(t) &= H(\xi(t), \alpha(t)). \end{aligned} \tag{3}$$

In what follows we present two interesting coder/decoder pairs: the logarithmic quantizer strategy and the “zoom in - zoom out” uniform quantizer strategy.

3.1. Zoom in - zoom out uniform coder

In this strategy the information transmitted from source to receiver is quantized by a scalar uniform quantizer which can be described as follows. For $L \in \mathbb{N}$, define the *uniform set of quantization levels*

$$S_L = \Big\{ -1 + \frac{2\ell - 1}{L} \mid \ell \in \{1, \dots, L\} \Big\} \cup \{-1\} \cup \{1\}$$

and the corresponding *uniform quantizer* (see Figure 1) $\mathrm{unq}_L : \mathbb{R} \to S_L$ by

$$\mathrm{unq}_L(x) = -1 + \frac{2\ell - 1}{L}$$

if $\ell \in \{1, \dots, L\}$ satisfies $-1 + \frac{2(\ell-1)}{L} \leq x \leq -1 + \frac{2\ell}{L}$, otherwise $\mathrm{unq}_L(x) = 1$ if $x > 1$ or $\mathrm{unq}_L(x) = -1$ if $x < -1$. Note that larger values of the parameter L correspond to more accurate uniform quantizers unq_L. Moreover note that, if we define m to be the number of quantization levels we have that $m = L + 2$.

For $L \in \mathbb{N}$, $k_{in} \in]0, 1[$, and $k_{out} \in]1, +\infty[$, the *zoom in - zoom out uniform coder/decoder* has the state space $\Xi = \mathbb{R} \times \mathbb{R}_{>0}$, the initial state $\xi_0 = (0, 1)$, and the alphabet $\mathcal{A} = S_L$. The coder/decoder state is written as $\xi = (\hat{x}_{-1}, f)$ and the coder/decoder dynamics are

$$\begin{aligned} \hat{x}_{-1}(t+1) &= \hat{x}_{-1}(t) + f(t)\alpha(t), \\ f(t+1) &= \begin{cases} k_{\text{in}}\, f(t), & \text{if } |\alpha(t)| < 1, \\ k_{\text{out}}\, f(t), & \text{if } |\alpha(t)| = 1. \end{cases} \end{aligned}$$

The quantizer and decoder functions are, respectively,

$$\alpha(t) = \mathrm{unq}_L\Big(\frac{x(t) - \hat{x}_{-1}(t)}{f(t)}\Big),$$
$$\hat{x}(t) = \hat{x}_{-1}(t) + f(t)\alpha(t).$$

The coder/decoder pair is analyzed as follows. One can observe that $\hat{x}_{-1}(t+1) = \hat{x}(t)$ for $t \in \mathbb{Z}_{\geq 0}$, that is, the first component of the coder/decoder state contains the estimate of the data x. The transmitted messages contain a quantized version of the estimate error $x - \hat{x}_{-1}$ scaled by factor f. Accordingly, the second component of the coder/decoder state f is referred to as the *scaling factor*: it grows when $|x - \hat{x}_{-1}| \geq f$ ("zoom out step") and it decreases when $|x - \hat{x}_{-1}| < f$ ("zoom in step").

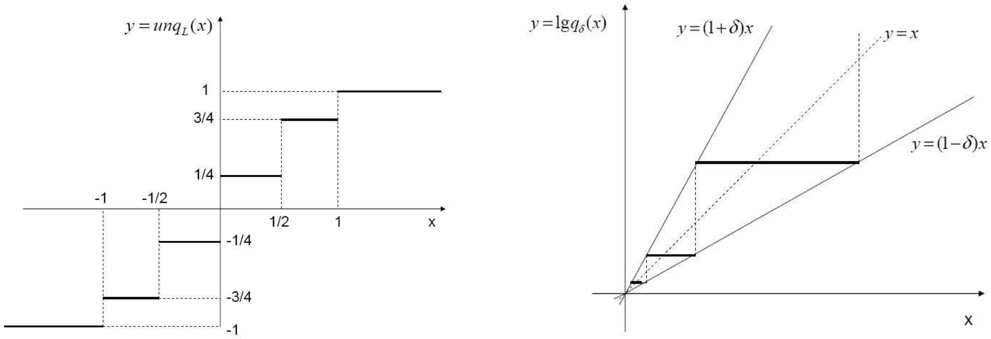

Figure 1. The uniform quantizer ($m = 6$) ***(left)***; the logarithmic quantizer ***(right)***.

3.2. Logarithmic coder

This strategy is presented for example in [10]. Given an *accuracy parameter* $\delta \in\,]0, 1[$, define the *logarithmic set of quantization levels*

$$S_\delta = \Big\{\Big(\frac{1+\delta}{1-\delta}\Big)^\ell\Big\}_{\ell\in\mathbb{Z}} \cup \{0\} \cup \Big\{-\Big(\frac{1+\delta}{1-\delta}\Big)^\ell\Big\}_{\ell\in\mathbb{Z}}, \tag{4}$$

and the corresponding *logarithmic quantizer* (see Figure 1) $\mathrm{lgq}_\delta : \mathbb{R} \to S_\delta$ by

$$\mathrm{lgq}_\delta(x) = \Big(\frac{1+\delta}{1-\delta}\Big)^\ell,$$

if $\ell \in \mathbb{Z}$ satisfies $\frac{(1+\delta)^{\ell-1}}{(1-\delta)^\ell} \leq x \leq \frac{(1+\delta)^\ell}{(1-\delta)^{\ell+1}}$, otherwise $\mathrm{lgq}_\delta(x) = 0$ if $x = 0$ or $\mathrm{lgq}_\delta(x) = -\mathrm{lgq}_\delta(-x)$ if $x < 0$. Note that smaller values of the parameter δ correspond to more accurate logarithmic quantizers lgq_δ. For $\delta \in\,]0, 1[$, the *logarithmic coder/decoder* is defined by the state space $\Xi = R$, initial state $\xi_0 = 0$, the alphabet $\mathcal{A} = S_\delta$, and by the maps

$$\begin{aligned}\xi(t+1) &= \xi(t) + \alpha(t),\\ \alpha(t) &= \mathrm{lgq}_\delta(x(t) - \xi(t)),\\ \hat{x}(t) &= \xi(t) + \alpha(t).\end{aligned} \tag{5}$$

The coder/decoder pair is analyzed as follows. One can observe that $\xi(t+1) = \hat{x}(t)$ for $t \in \mathbb{N}$, that is, the coder/decoder state contains the estimate of the data x. The transmitted messages contain a quantized version of the estimate error $x - \xi$. The estimate $\hat{x} : N \to \mathbb{R}$ satisfies the recursive relation

$$\hat{x}(t+1) = \hat{x}(t) + \mathrm{lgq}_\delta \left(x(t+1) - \hat{x}(t) \right),$$

with initial condition $\hat{x}(0) = \mathrm{lgq}_\delta \left(x(0) \right)$ determined by $\xi(0) = 0$. Finally, define the function $r : \mathbb{R} \to \mathbb{R}$ by $r(y) = \frac{\mathrm{lgq}_\delta(y) - y}{y}$ for $y \neq 0$ and $r(0) = 0$. Some elementary calculations show that $|r(y)| \leq \delta$ for all $y \in \mathbb{R}$. Accordingly, if we define the trajectory $\omega : \mathbb{N} \to [-\delta, +\delta]$ by $\omega(t) = r(x(t+1) - \hat{x}(t))$, then

$$\hat{x}(t+1) = \hat{x}(t) + (1 + \omega(t))\big(x(t+1) - \hat{x}(t)\big). \tag{6}$$

This is called the *multiplicative noise* model for the logarithmic quantizer.

Remark 3.1. When communicating through digital channels, the use of the logarithmic quantizer described in the above Section, presents an evident drawback with respect to the zoom in- zoom out strategy, due to the fact that the logarithmic set of quantization levels S_δ is countable and not finite as the uniform set of quantization levels. This implementation issue could be overcome by truncating the map lgq_δ as follows. Let $a, b \in \mathbb{R}$ satisfy $0 < a < b$; if $a \leq |x| \leq b$, then

$$\mathrm{lgq}_\delta(x) = \mathrm{sgn}(x) \left(\frac{1+\delta}{1-\delta} \right)^\ell,$$

where $\ell \in \mathbb{Z}$ is such that $\frac{(1+\delta)^{\ell-1}}{(1-\delta)^\ell} \leq |x| \leq \frac{(1+\delta)^\ell}{(1-\delta)^{\ell+1}}$, otherwise

$$\mathrm{lgq}_\delta(x) = \begin{cases} 0, & \text{if } |x| < a, \\ \mathrm{sgn}(x)\ \mathrm{lgq}_\delta(b), & \text{if } |x| > b. \end{cases}$$

Again, if m denotes the number of quantization levels, it is possible to see (see [11]) that, for the truncated logarithmic quantizer,

$$m = \frac{2 \log C}{\log \frac{1+\delta}{1-\delta}}.$$

We will come back on this remark later on.

4. Consensus algorithm with exchange of quantized information

We consider now the same algorithm previously illustrated with the assumption that the agents can communicate only through digital channels. Here, we adopt the logarithmic coder/decoder scheme (3) described in Subsection 3.2; we analyze the zoom in - zoom out strategy via simulations in Section 6.

4.1. Algorithm description

Here is an informal description of our proposed scheme. We envision that along each communication edge we implement a logarithmic coder/decoder; in other words, each agent transmits through a dynamic encoding scheme to all its neighbors the quantized information regarding its state. Once state estimates of all agent's neighbors are available, each agent will then implement the average consensus algorithm.

Next, we provide a formal description of the proposed algorithm. Let $P \in \mathbb{R}^{N\times N}$ be a stochastic symmetric matrix with positive diagonal elements and with connected induced graph $\mathcal{G}_{\mathcal{P}}$. Assume there are digital communication channels along all edges of $\mathcal{G}_{\mathcal{P}}$ capable of carrying a countable number of symbols. Pick an accuracy parameter $\delta \in]0,1[$. The *consensus algorithm with dynamic coder/decoder* is defined as follows:

Processor states: For each $i \in \{1,\dots,N\}$, agent i has a state variable $x_i \in \mathbb{R}$ and state estimates $\hat{x}_j \in \mathbb{R}$ of the states of all neighbors j of i in $\mathcal{G}_{\mathcal{P}}$. Furthermore, agent i maintains a copy of $\hat{x}_i$.

Initialization: The state $x(0) = (x_1(0),\dots,x_N(0))^T \in \mathbb{R}^N$ is given as part of the problem. All estimates $\hat{x}_j(0)$, for $j \in \{1,\dots,N\}$, are initialized to 0.

State iteration: At time $t \in \mathbb{N}$, for each i, agent i performs three actions in the following order:

(1) Agent i updates its own state by

$$x_i(t) = x_i(t-1) + \sum_{j=1}^{N} P_{ij}\left(\hat{x}_j(t-1) - \hat{x}_i(t-1)\right). \tag{7}$$

(2) Agent i transmits to all its neighbors the symbol

$$\alpha_i(t) = \mathrm{lgq}_\delta(x_i(t) - \hat{x}_i(t-1)).$$

(3) Node i updates its estimates

$$\hat{x}_j(t) = \hat{x}_j(t-1) + \alpha_j(t), \tag{8}$$

for j being equal to all neighbors of i and to i itself.

Before the algorithm analysis, we clarify a few points.

Remark 4.1 (Clarifications and variations)

(i) Agent i and all its neighbors j maintain in memory an estimate $\hat{x}_i$ of the state x_i. We denote all these estimates by the same symbol because they are all identical: they are initialized in the same manner and they are updated through the same equation with the same information. On the other hand, it would be possible to adopt distinct quantizer accuracies δ_{ij} for each communication channel (i,j). In such a case then we would have to introduce variables $\hat{x}_{ij}$ that node i and j would maintain for the estimate of x_i.

(ii) We could define a different state update equation where each agent i uses the exact knowledge of its own state x_i instead of the estimate $\hat{x}_i$, that is, we could adopt

$$x_i(t) = x_i(t-1) + \sum_{j=1}^{N} P_{ij}\left(\hat{x}_j(t-1) - x_i(t-1)\right) = P_{ii}x_i(t-1) + \sum_{j\neq i} P_{ij}\hat{x}_j(t-1), \tag{9}$$

instead of equation (7). We will discuss the drawback of this choice below.

4.2. Main convergence result

We now analyze the algorithm. First, we write the closed-loop system in matrix form. Equation (7) is written as

$$x(t+1) = x(t) + (P - I)\hat{x}(t). \tag{10}$$

The N-dimensional vector of state estimates $\hat{x} = (\hat{x}_1, \dots, \hat{x}_N)^T$ is updated according to the multiplicative-noise model in equation (6). In other words, there exist $\omega_j \colon \mathbb{N} \to [-\delta, +\delta]$, for $j \in \{1, \dots, N\}$, such that

$$\hat{x}_j(t+1) = \hat{x}_j(t) + (1 + \omega_j(t))\big(x_j(t+1) - \hat{x}_j(t)\big),$$

and, for $\Omega(t) := \operatorname{diag}\{\omega_1(t), \dots, \omega_N(t)\}$,

$$\hat{x}(t+1) = \hat{x}(t) + (I + \Omega(t))\big(x(t+1) - \hat{x}(t)\big). \tag{11}$$

Equations (10) and (11) with multiplicative noise Ω determine the closed-loop system.

Next, we define the estimate error $e = \hat{x} - x \in \mathbb{R}^N$ and rewrite the close-loop system in terms of the quantities x and e. Straightforward calculations show that, for $t \in \mathbb{Z}_{\geq 0}$,

$$\begin{bmatrix} x(t+1) \\ e(t+1) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \Omega(t) \end{bmatrix} \begin{bmatrix} P & P - I \\ P - I & P - 2I \end{bmatrix} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix}. \tag{12}$$

Initial conditions are $x(0)$ and $e(0) = -x(0)$.

Finally, we are ready to state the main properties of our quantized consensus algorithm.

Theorem 4.2. *Assume $P \in \mathbb{R}^{N \times N}$ satisfies Assumption 2.1 and define $\bar{\delta} \in \mathbb{R}$ by*

$$\bar{\delta} := \frac{1 + \lambda_{\min}(P)}{3 - \lambda_{\min}(P)}. \tag{13}$$

The solution $t \mapsto (x(t), e(t))$ of the consensus algorithm with dynamic coder/decoder has the following two properties:

(i) the state average is maintained constant by the algorithm, that is, $\frac{1}{N}\sum_{i=1}^{N} x_i(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(0)$ for all $t \in \mathbb{N}$;

(ii) if $0 < \delta < \bar{\delta}$, then the state variables converge to their average value and the estimate error vanishes, that is,

$$\lim_{t \to \infty} x(t) = \Big(\frac{1}{N}\sum_{i=1}^{N} x_i(0)\Big)\mathbf{1}$$

and

$$\lim_{t \to \infty} e(t) = 0.$$

Proof: Observe that $\mathbf{1}^T x(t+1) = \mathbf{1}^T P x(t) + \mathbf{1}^T (P - I) e(t) = \mathbf{1}^T x(t)$, where the second equality holds since $\mathbf{1}^T (P - I) = 0$. This proves the first statement of the theorem. The second statement is a consequence of Theorem 4.8 stated in Section 4.3. □

We here consider some remarks and examples.

Remark 4.3. Note that $\bar{\delta}$ is an increasing function on $\lambda_{\min}(P)$ and that $\bar{\delta} = 0$, if $\lambda_{\min}(P) = -1$, and $\bar{\delta} = 1$, if $\lambda_{\min}(P) = 1$ (see Figure 2).

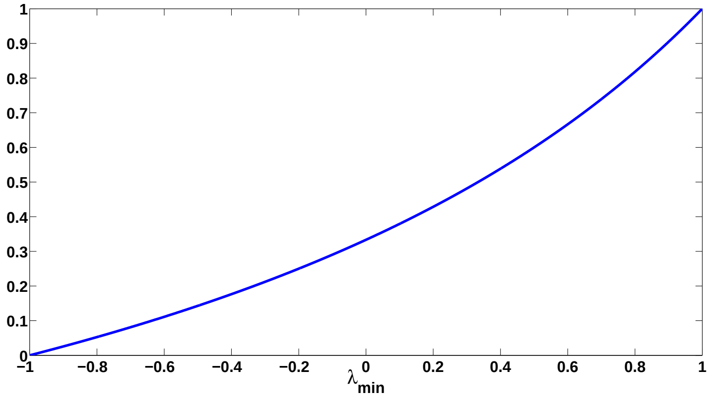

Figure 2. Behavior of $\bar{\delta}$.

Remark 4.4. The state update in equation (9) does not maintain the average. This fact motivates the choice of state update equation (7).

Example 4.5. Consider the sequence of circulant matrices $\{P_N\} \in \mathbb{R}^{N\times N}$ defined by

$$P_N = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & 0 & 0 & \cdots & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & \vdots \\ \frac{1}{3} & 0 & 0 & 0 & \cdots & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix}. \qquad (14)$$

For this sequence of symmetric stochastic matrices we have that $\lambda_{\min}(P_N) = \frac{1}{3} - \frac{2}{3}\cos\left(\frac{2\pi}{N}\left\lfloor\frac{N}{2}\right\rfloor\right)$. Hence $\lambda_{\min}(P_N) \geq -\frac{1}{3}$, implying therefore that $\bar{\delta} \geq \frac{1}{5}$ for all N. This shows that $\bar{\delta}$ is uniformly bounded away from 0. This is a remarkable property of scalability on the dimension of the network.

Remark 4.6. The fact that the critical accuracy sufficient to guarantee convergence is independent on the network dimension is more general than what seen in the previous example. Indeed, assume that $\{P_N\} \in \mathbb{R}^{N\times N}$ is a sequence of matrices of increasing size, where each P_N satisfies Assumption 2.1 and where each P_N has all the diagonal elements greater than a positive real number $\bar{p}$. Then, by Gershgorin's Theorem we have that $\lambda_{\min}(P_N) \geq -1 + 2\bar{p}$ and hence $\bar{\delta} \geq \frac{\bar{p}}{2-\bar{p}}$ for all N. It follows that the critical accuracy sufficient to guarantee convergence is bounded away from zero uniformly on the dimension of the network.

4.3. Convergence analysis

In this section we provide the analysis of the asymptotic properties of system (12). For the sake of the notational convenience, let us define

$$\mathcal{F}(t) = \begin{bmatrix} I & 0 \\ 0 & \Omega(t) \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix} \in \mathbb{R}^{2N\times 2N}. \qquad (15)$$

Consider now the system

$$z(t+1) = \mathcal{F}(t)z(t), \tag{16}$$

where $z(t) \in \mathbb{R}^{2N}$ for all $t \geq 0$ and where $z(0)$ is any vector in $\mathbb{R}^{2N}$. We start our analysis by rewriting (15) in a more suitable way. Let

$$\mathcal{E} = \left\{ \text{diag}\{e_1, \ldots, e_N\} \in \mathbb{R}^{N \times N} \,:\, e_i \in \{-1, +1\}, i \in \{1, \ldots, N\} \right\}.$$

Notice that $\mathcal{E}$ contains 2^N elements. Hence, we can write $\mathcal{E} = \{E_1, \ldots, E_{2^N}\}$, where we are assuming that some suitable way to enumerate the matrices inside $\mathcal{E}$ has been used. We assume that $E_1 = I$. We define now $\mathcal{E}_\delta = \{\delta E_1, \ldots, \delta E_{2^N}\}$. Observe that $\Omega(t) \in Co\{\mathcal{E}_\delta\}$ for all $t \geq 0$, where $Co\{\mathcal{E}_\delta\}$ denotes that convex hull of the set $\mathcal{E}_\delta$. By means of the above definitions we can introduce another set of matrices

$$\mathcal{R} = \left\{ R_i = \begin{bmatrix} I & 0 \\ 0 & \delta E_i \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix} \,:\, E_i \in \mathcal{E} \right\}. \tag{17}$$

Accordingly to the definition of E_1 we have that

$$R_1 = \begin{bmatrix} I & 0 \\ 0 & \delta I \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix}. \tag{18}$$

The set $\mathcal{R}$ is useful because it is easy to see that the matrix $\mathcal{F}(t)$, belongs to $Co\{\mathcal{R}\}$ for all $t \geq 0$, where $Co\{\mathcal{R}\}$ denote the convex hull of the set $\mathcal{R}$. In other words, for all $t \geq 0$, there exist $\nu_1(t), \ldots, \nu_{2^N}(t)$ nonnegative real numbers such that $\sum_{i=1}^{2^N} \nu_i(t) = 1$ and

$$\mathcal{F}(t) = \sum_{i=1}^{2^N} \nu_i(t) R_i.$$

We state the following result that will permit us to analyze the system (16) by means of Theorem II.1 (see Appendix).

Lemma 4.7. *For $v = \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix}^T$, we have*

$$R_i v = v, \quad \textit{and} \quad v^T R_i = v^T, \quad \textit{for all } i \in \{1, \ldots, 2^N\}.$$

Moreover, for $\bar{\delta}$ as in equation (13), the following facts are equivalent:

(i) *1 is the only eigenvalue of unit magnitude of the matrix R_1, and all its other eigenvalues are strictly inside the unit disc;*
(ii) $0 \leq \delta < \bar{\delta}$.

Proof: The first part of the lemma is easily proved by observing that

$$\begin{bmatrix} I & 0 \\ 0 & \delta E_i \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \delta E_i \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix},$$

and

$$\begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \delta E_i \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix}.$$

Consider now R_1; to compute its eigenvalues we calculate

$$\det\left(sI - R_1\right) \;=\; \det\begin{bmatrix} sI-P & -(P-I) \\ -\delta(P-I) & sI-\delta(P-2I) \end{bmatrix}.$$

Since each block of the above matrix commute with each other block, we have from [12] that

$$\begin{aligned}
\det\left(sI - R_1\right) &= \det\left[(sI-P)(sI-\delta(P-2I)) - \delta(P-I)^2\right] \\
&= \det\left[s^2 I - s\left(\delta(P-2I)+P\right) + \delta\left(P^2-2P-P^2-I+2P\right)\right] \\
&= \prod_{i=0}^{N-1}\left[s^2 - \left(\delta(\lambda_i-2)+\lambda_i\right)s - \delta\right] \\
&= \left(s^2-(1-\delta)s-\delta\right)\prod_{i=1}^{N-1}\left(s^2-(\delta(\lambda_i-2)+\lambda_i)s-\delta\right).
\end{aligned}$$

Hence the eigenvalues of R_1 are given by the solution of the following N second order equations

$$s^2-(1-\delta)s-\delta=0, \tag{19}$$

and

$$s^2-(\delta(\lambda_i-2)+\lambda_i)s-\delta=0, \qquad i\in\{1,\dots,N-1\}. \tag{20}$$

The solutions of (19) are 1 and $-\delta$. Consider now (20). Given i, let $s_1^{(i)}$ and $s_2^{(i)}$ denote the two solutions of (20). We have that

$$s_1^{(i)} = \frac{\delta(\lambda_i-2)+\lambda_i-\sqrt{(\delta(\lambda_i-2)+\lambda_i)^2+4\delta}}{2}$$

and

$$s_2^{(i)} = \frac{\delta(\lambda_i-2)+\lambda_i+\sqrt{(\delta(\lambda_i-2)+\lambda_i)^2+4\delta}}{2}.$$

Now we have to analyze the conditions $|s_1^{(i)}|<1$ and $|s_2^{(i)}|<1$, for all $i\in\{1,\dots,N-1\}$. To this purpose, we consider the bilinear transformation of the equation (20), i.e., we substitute to s the term $\frac{1+\tilde{s}}{1-\tilde{s}}$. We obtain the new equation

$$(1+\delta)(1-\lambda_i)\tilde{s}^2+2(1+\delta)\tilde{s}+1+\lambda_i+\delta(\lambda_i-3)=0. \tag{21}$$

Let $\tilde{s}_1^{(i)}$ and $\tilde{s}_2^{(i)}$ denote the two solutions of (21). From the property of the bilinear transformation, we have that $|s_1^{(i)}|<1$ and $|s_2^{(i)}|<1$ if and only if $\tilde{s}_1^{(i)}<0$ and $\tilde{s}_2^{(i)}<0$. Since $1+\delta>0$ and $(1+\delta)(1-\lambda_i)>0$ for $i\in\{1,\dots,N-1\}$, we obtain, from the Cartesian rule, that $\tilde{s}_1^{(i)}<0$ and $\tilde{s}_2^{(i)}<0$ for all $i\in\{1,\dots,N-1\}$, if and only if $1+\lambda_i+\delta(\lambda_i-3)>0$ for all $i\in\{1,\dots,N-1\}$. This last condition is verified if and only if $\delta<\bar{\delta}$. □

We are able now to state the following theorem characterizing the asymptotic stability of the system (16).

Theorem 4.8. *Consider the system (16). The following facts are equivalent:*

(a) $\delta<\bar{\delta}$;

(b) for each initial condition $z(0) \in \mathbb{R}^{2N}$ and for any sequence $\{\Omega(t)\}_{t=0}^{+\infty}$ with $\Omega(t) \in Co\{\mathcal{E}_\delta\}$ for all $t \geq 0$, we have

$$\lim_{t\to+\infty} z(t) = \left[\begin{array}{c} \alpha\mathbf{1} \\ 0 \end{array} \right], \qquad (22)$$

for $\alpha = \frac{1}{N}\left[\mathbf{1}^T \;\; \mathbf{0}^T\right] z(0)$.

Proof: We start by proving that (b) implies (a). To this aim, we consider the sequence $\mathcal{F}(0) = \mathcal{F}(1) = \mathcal{F}(2) = \ldots = R_1$. In this case $z(t)$ is the evolution of an autonomous linear time invariant discrete-time systems with updating matrix R_1. Therefore, by Lemma 4.7, (22) holds true if and only if and only if $\delta < \bar{\delta}$.

We prove now that (a) implies (b). We will show that, for $\delta < \bar{\delta}$, there exists a suitable symmetric matrix $L \in \mathbb{R}^{2N\times 2N}$ satisfying the following three properties

$$L\left[\mathbf{1}^T \;\; \mathbf{0}^T\right]^T = 0, \qquad (23)$$

$$z^T L z > 0, \qquad (24)$$

$$z^T\left(\frac{1}{2}\left(R_i^T L R_j + R_j^T L R_i\right) - L\right) z < 0, \quad \text{for all } R_i,\, R_j \in \mathcal{R}, \qquad (25)$$

$\forall\;\; z \in <\left[\mathbf{1}^T \;\; \mathbf{0}^T\right]^T >^\perp$. This fact, together with Lemma 4.7 and Theorem II.1 (see Appendix), ensures that fact (a) implies (b). As candidate matrix L we select

$$L = \left[\begin{array}{cc} I-P & 0 \\ 0 & \gamma I \end{array} \right], \qquad (26)$$

where γ is a suitable positive scalar to be determined. Observe that the eigenvalues of $I-P$ are 0 and $1-\lambda_i$ for $i \in \{1,\ldots,N-1\}$, where it is immediate to see that $1-\lambda_i > 0$ for $i \in \{1,\ldots,N-1\}$. Since $\sigma(L) = \sigma(I-P)\cup\sigma(\gamma I)$ it follows that also L has an eigenvalue equal to 0 and all other eigenvalues positive. Moreover, since $L\left[\mathbf{1}^T \;\; \mathbf{0}^T\right]^T = \left[((I-P)\mathbf{1})^T \;\; \mathbf{0}^T\right]^T = 0$, we have that the eigenspace associated to the eigenvalue 0 is generated by the vector $[\mathbf{1}^T \;\; 0^T]^T$. Hence L satisfies (23) and (24). Moreover, by the structure of L, it is easy to check that $R_i^T L R_j = R_j^T L R_i$ for all $R_i, R_j \in \mathcal{R}$. Thus, verifying (25) is equivalent to verify

$$z^T\left(R_i^T L R_j - L\right) z < 0, \quad \text{for all } R_i,\, R_j \in \mathcal{R}, \qquad (27)$$

for any nonzero $z \in <\left[\mathbf{1}^T \;\; \mathbf{0}^T\right]^T >^\perp$. By straightforward calculations, we have that

$$R_i^T L R_j - L = R_1 L R_1 - L - Q,$$

where

$$R_1^T L R_1 - L \;\; = \;\; \left[\begin{array}{cc} (I-P)^2(\gamma\delta^2 I - I - P) & (I-P)(P(P-I) - \gamma\delta^2(P-2I)) \\ (I-P)(P(P-I) - \gamma\delta^2(P-2I)) & (I-P)^3 + \gamma\delta^2(P-2I)^2 - \gamma I \end{array} \right],$$

and

$$Q = \gamma\delta^2 \left[\begin{array}{c} (P-I)K \\ (P-2I)K \end{array} \right] \left[K(P-I) \;\; K(P-2I)\right],$$

with K such that $K^2 = I - E_i E_j$[†]. Clearly, $Q = Q^T \geq 0$ and $Q \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix}^T = 0$. If (27) is satisfied for $i = j = 1$, then (27) holds also for any pair R_i, R_j belonging to $\mathcal{R}$. Finally, observe that, by Lemma III.2 (see Appendix), we immediately have that

$$z^T (R_1^T L R_1 - L) z < 0, \qquad \forall \ z \in < \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix}^T >^\perp$$

if we choose

$$\gamma = \frac{1 + \lambda_{\min} + \delta^2 (\lambda_{\min} - 3)}{2\delta^2}.$$

□

5. Exponential convergence

The objective of this section is to understand how much the quantization affects the performance of the consensus algorithm. To this aim, by means of a Lyapunov analysis, we will provide a characterization of the asymptotic speed of the convergence toward the consensus of both the ideal algorithm (2) and the algorithm (12). We start by introducing some definitions. A function $f : \mathbb{N} \to \mathbb{R}$ converges to 0 *exponentially fast* if there exist a constant $C > 0$ and another constant $\xi \in [0, 1)$ such that $|f(t)| \leq C\xi^t$, for all t; the infimum among all numbers $\xi \in [0, 1)$ satisfying the exponential convergence property is called the *exponential convergence factor* of f. In other words, the exponential convergence factor of f is given by

$$\limsup_{t \to \infty} |f(t)|^{\frac{1}{t}}.$$

Consider first the system (2). To quantify the speed of convergence of (2) toward consensus, we introduce the following variable

$$\bar{x}(t) := x(t) - x_a(0)\mathbf{1},$$

where $x_a(0) = \frac{1}{N}\mathbf{1}^* x(0)$. Note that the i-th component of $\bar{x}(t)$ represents the distance of the state of the i-th system from the initial average. Clearly, $\lim_{t\to\infty} x(t) = x_a(0)\mathbf{1}$ if and only if $\lim_{t\to\infty} \bar{x}(t) = 0$. It is easy to see that the variable $\bar{x}$ satisfies the same recursive equation of the variable x, that is,

$$\bar{x}(t + 1) = P\bar{x}(t). \tag{28}$$

Moreover note that $\mathbf{1}^T \bar{x}(t) = 0$, for all $t \geq 0$. We define the exponential convergence factor of $\bar{x}(t)$, for a given initial condition $\bar{x}_0 \in < \mathbf{1} >^\perp$, to be

$$\rho(P, \bar{x}_0) := \limsup_{t\to\infty} ||\bar{x}(t)||^{\frac{1}{t}}$$

We can get rid of the initial condition and define the *exponential convergence factor* of the system (2) as follows

$$\rho(P) := \sup_{\bar{x}_0 \in <\mathbf{1}>^\perp} \rho(P, \bar{x}_0) \tag{29}$$

[†] Of note is that $I - E_i E_j$ is a positive semidefinite matrix and hence the matrix K is well-defined.

Consider now the positive semidefinite matrix $I - P$. Notice that

$$\rho(P, \bar{x}_0) = \limsup_{t\to\infty} (\bar{x}(t)^T (I - P)\bar{x}(t))^{\frac{1}{2t}}$$

and so we can characterize the speed of convergence to 0 of the variable $\bar{x}$ by studying the exponential convergence factor of the Lyapunov function $\bar{x}(t)^T (I - P)\bar{x}(t)$.

Theorem 5.1. *Consider* (28) *with $P \in \mathbb{R}^{N\times N}$ satisfing Assumption 2.1. Then the function $t \mapsto (\bar{x}(t)^T (I - P)\bar{x}(t))^{1/2}$, defined along any trajectory $t \mapsto \bar{x}(t)$, converges exponentially fast to* 0*. Moreover, the factor $\rho(P)$, defined in equation* (29)*, satisfies*

$$\rho(P) = \max\{\lambda_{\max}(P), -\lambda_{\min}(P)\}.$$

Proof: Let $\alpha := \max\left\{\lambda^2_{\max}(P), \lambda^2_{\min}(P)\right\}$ so that $z^T P^2 z \le \alpha z^T z$ for all $z \in <\mathbf{1}>^\perp$ and, in turn,

$$z^T (P(I - P)P)z \le \alpha z^T (I - P)z, \tag{30}$$

for all $z \in <\mathbf{1}>^\perp$. This shows that the map $t \mapsto \bar{x}(t)^T (I - P)\bar{x}(t)$ converges exponentially fast to 0 along any trajectory $t \mapsto \bar{x}(t)$ and that $\rho(P) \le \sqrt{\alpha}$. Moreover, observe that, if z is equal to the eigenvector corresponding to the eigenvalue defining β, then (30) holds true as equality. Then, if $\bar{x}_0$ is equal to this eigenvector, we obtain a trajectory $t \mapsto \bar{x}(t)$ along which the function $t \mapsto \bar{x}(t)^T (I - P)\bar{x}(t)$ has exponential convergence factor equal to $\sqrt{\alpha}$. □

This concludes the analysis of the algorithm (2). In the sequel of this section, we provide a similar analysis of the system (12). To this aim we consider again the system (16), that is

$$z(t + 1) = \mathcal{F}(t)z(t), \tag{31}$$

where $z(0) = z_0$ is any vector in $\mathbb{R}^{2N}$. To perform a Lyapunov analysis of (31), it is convenient to introduce the variable

$$\bar{z}(t) = \begin{bmatrix} I - \frac{1}{N}\mathbf{1}\mathbf{1}^T & 0 \\ 0 & I \end{bmatrix} z(t).$$

Clearly, condition (b) of Theorem 4.8 holds true if and only if $\lim_{t\to\infty} \bar{z}(t) = 0$. It is straightforward to see that $\bar{z}$ satisfies the same recursive equation of $z(t)$, i.e.,

$$\bar{z}(t + 1) = \mathcal{F}(t)\bar{z}(t) \tag{32}$$

and that $\begin{bmatrix}\mathbf{1}^T & \mathbf{0}^T\end{bmatrix}^T \bar{z}(t) = 0$ for all $t \ge 0$. Consider now the matrix $L \in \mathbb{R}^{2N\times 2N}$, introduced along the proof of Theorem 4.8 and defined as

$$L = \begin{bmatrix} I - P & 0 \\ 0 & \gamma I \end{bmatrix}.$$

For each $\gamma > 0$ define

$$\tilde{\rho}\left(P, \delta, \gamma; \bar{z}_0, \{\mathcal{F}(t)\}_{t=0}^\infty\right) := \limsup_{t\to\infty} (\bar{z}(t)^T L\bar{z}(t))^{\frac{1}{2t}} \tag{33}$$

We can get rid of the initial conditions $\bar{z}_0$ and the sequences $\{\mathcal{F}(t)\}_{t=0}^\infty$ by considering

$$\tilde{\rho}(P, \delta, \gamma) := \sup_{\bar{z}_0, \{\mathcal{F}(t)\}_{t=0}^\infty} \tilde{\rho}\left(P, \delta, \gamma; \bar{z}_0, \{\mathcal{F}(t)\}_{t=0}^\infty\right) \tag{34}$$

where the initial conditions $\bar{z}_0$ belong to the set of vectors orthogonal to $\left[\mathbf{1}^T \;\; \mathbf{0}^T\right]^T$ and the sequences $\{\mathcal{F}(t)\}_{t=0}^{\infty}$ are such that $\mathcal{F}(t) \in Co\{\mathcal{R}\}$ for all $t \geq 0$. It can be shown that $\tilde{\rho}(P, \delta, \gamma)$ is independent of γ and for this reason we denote it as $\tilde{\rho}(P, \delta)$.

We characterize now $\tilde{\rho}(P, \delta)$. To this aim, consider the following semidefinite programming problem

$$\bar{\beta}(P, \delta, \gamma) := \begin{array}{cc} & \max \;\; \beta \\ \text{such that} & R_1^T L R_1 - L \leq -\beta L \end{array} \tag{35}$$

We have the following result.

Theorem 5.2. *Consider* (32) *with the matrix P satisfing Assumption 2.1. Let $\bar{\delta}$ be defined as in* (13) *and let $\delta \in \mathbb{R}$ be such that $0 \leq \delta < \bar{\delta}$. Moreover let $\gamma \in \mathbb{R}$ be such that $\gamma > 0$, and let $\bar{\beta}(P, \delta, \gamma)$ be defined as in* (35)*. Then, the function $t \to (\bar{z}(t)^T L \bar{z}(t))^{1/2}$, defined along any trajectory $t \to \bar{z}(t)$ converges exponentially fast to* 0 *and the factor $\tilde{\rho}(P, \delta)$, defined in equation* (34)*, satisfies*

$$\tilde{\rho}(P, \delta) \leq \sqrt{1 - \bar{\beta}(P, \delta, \gamma)}.$$

Proof: We start by recalling, that since $\mathcal{F}(t)$ belongs to $Co\,(\mathcal{R})$, we can write that $\mathcal{F}(t) = \sum_{i=1}^{2^N} \nu_i(t) R_i$, where $\nu_1(t), \ldots, \nu_{2^N}(t)$ are nonnegative real numbers such that $\sum_{i=1}^{2^N} \nu_i = 1$. Along the proof of Theorem 4.8, we have seen that

$$z^T (R_i^T L R_j - L) z \leq z^T (R_1^T L R_1 - L) z < 0,$$

for all $z \in \mathbb{R}^{2N}$ such that $z \in <[\mathbf{1}^T \;\; \mathbf{0}^T]^T>^{\perp}$ and for any pair of matrices R_i, R_j belonging to $\mathcal{R}$. Hence we have that

$$\begin{aligned} z^T (\mathcal{F}^T(t) L \mathcal{F}(t) - L) z &= z^T \left(\left(\sum_{i=1}^{2^N} \nu_i(t) R_i \right)^T L \left(\sum_{j=1}^{2^N} \nu_j(t) R_j \right) - L \right) z \\ &= z^T \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \Big(\nu_i(t)\nu_j(t) R_i^T L R_j - \nu_i(t)\nu_j(t) L \Big) \right) z \\ &\leq z^T \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \nu_i(t)\nu_j(t) (R_1^T L R_1 - L) \right) z = z^T (R_1^T L R_1 - L) z, \end{aligned}$$

for all $z \in \mathbb{R}^{2N}$ such that $z \in <[\mathbf{1}^T \;\; \mathbf{0}^T]^T>^{\perp}$. Observe finally that $z^T (R_1^T L R_1 - L) z \leq \bar{\beta} z^T L z < 0$, from which we can argue that $\bar{z}(t+1)^T L \bar{z}(t+1) \leq (1 - \bar{\beta}) \bar{z}(t)^T L \bar{z}(t)$ and so the theses follow. □

It is worth noting that the above Theorem relates $\tilde{\rho}(P, \delta)$ to the resolution of a LMI [13]. It is well known that the computational effort required by the resolution of a LMI strictly depends on its dimensionality. However, we can observe that Lemma III.1 (see Appendix) provides an efficient way of solving (35), that drastically reduces its computational complexity. Indeed, we have that $\bar{\beta}(P, \delta, \gamma) = \min\{\beta_{min}^{-}(\delta, \gamma), \beta_{max}^{-}(\delta, \gamma)\}$, where $\beta_{min}^{-}(\delta, \gamma), \beta_{max}^{-}(\delta, \gamma)$ are defined in Lemma III.1. This means that one has to calculate only the value of the two variables $\beta_{min}^{-}(\delta, \gamma), \beta_{max}^{-}(\delta, \gamma)$ and evaluate the minimum between them. Differently from the method based on the LMI, the complexity of this method is independent of N.

Example 5.3. In this example we consider a connected random geometric graph generated by choosing $N = 30$ points at random in the unit square, and then placing an edge between each pair of points at distance less than 0.4. The matrix P is built using the Metropolis weights [14]. In this case we have that $\lambda_{\min} = -0.013$ and $\bar{\delta} = 0.327$. In figure 3, we plot the behavior of β^{-}_{min} and β^{-}_{max} as functions of γ. The value of δ is assumed constant and precisely equal to 0.25.

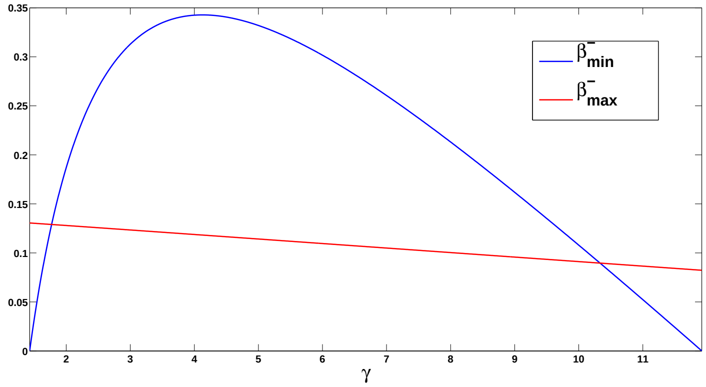

Figure 3. Behavior of $\tilde{\rho}$ as function of γ for P and δ fixed.

In general, assigned the matrix P and the value of the accuracy parameter δ, one could be interested in determining the maximum value of $\bar{\beta}$, as function of γ. Clearly, the best bound on $\tilde{\rho}(P, \delta)$ corresponds to the maximum value of $\bar{\beta}$, that is,

$$\tilde{\rho}(P, \delta) \leq \sqrt{1 - \bar{\beta}_{opt}(P, \delta)}$$

where

$$\bar{\beta}_{opt}(P, \delta) := \max_{\gamma > 0} \bar{\beta}(P, \delta, \gamma).$$

We illustrate this discussion in the following example.

Example 5.4. We consider the same matrix P generated in the previous example. In Figure 4, we depict the behavior of $\sqrt{1 - \bar{\beta}_{opt}(P, \delta)}$ as a function of δ. The dotted line represents the value of $\rho(P)$, that is, the convergence factor of the ideal algorithm (28). Notice that the convergence factor $\sqrt{1 - \bar{\beta}_{opt}(P, \delta)}$ depends smoothly on the accuracy parameter δ and that

$$\lim_{\delta \to 0} \sqrt{1 - \bar{\beta}_{opt}(P, \delta)} = \rho(P).$$

An interesting characterization of $\tilde{\rho}$ can be provided when considering a family of matrices $\{P_N\}$ of increasing size whose maximum eigenvalue converges to 1. It is worth noting that this situation is encountered in many practical situations [15, 16, 2]. We formalize this situation as follows.

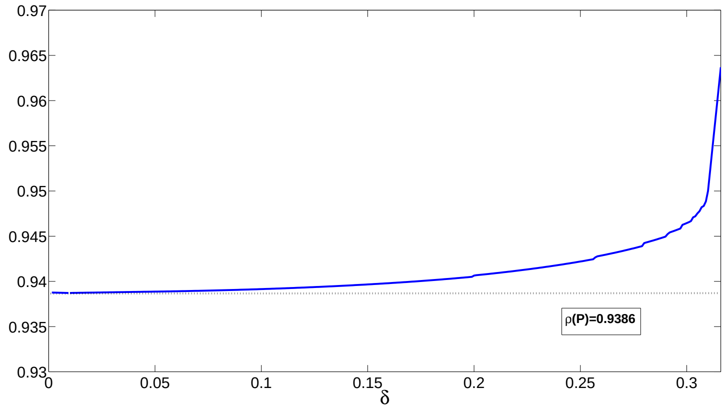

Figure 4. Behavior of $\sqrt{1-\bar{\beta}_{opt}(P,\delta)}$.

Assumption 5.5 (Vanishing spectral gap) *Assume we have a sequence of symmetric stochastic matrices* $\{P_N\} \subset \mathbb{R}^{N\times N}$ *satisfying Assumption 2.1 and the following conditions*

(i) $\lambda_{\min}(P_N) > c$ *for some* $c \in \,]-1,1[$ *and for all* $N \in \mathbb{N}$*;*
(ii) $\lambda_{\max}(P_N) = 1 - \epsilon(N) + o(\epsilon(N))$ *as* $N \to \infty$*, where* $\epsilon : \mathbb{N} \to \mathbb{R}$ *is a positive function such that* $\lim_{N\to\infty} \epsilon(N) = 0$*.*

According to Theorem 5.1, as $N \to \infty$, we have that $\rho\,(P_N) = 1 - \epsilon(N) + o\,(\epsilon(N))$. In considering the quantized version of the consensus algorithm, together with the sequence $\{P_N\}$, we have also to fix the sequence $\{\delta_N\}$. For simplicity, in the following we will assume that, $\{\delta_N\}$ is a constant sequence, i.e., $\delta_N = \delta$ with suitable δ such that $\delta < \frac{1+c}{3-c}$ which ensures the stability for all N.

Theorem 5.6. *Let* $\{P_N\} \subset \mathbb{R}^{N\times N}$ *be a family of matrices of increasing size satisfying Assumptions 2.1 and 5.5. Let* $\delta \in \mathbb{R}$ *be such that* $\delta < \frac{1+c}{3-c}$*. Then, as* $N \to \infty$*, we have that*

$$\tilde{\rho}(P_N,\delta) \leq 1 - \left(1 - \frac{1+c+\delta^2(c-3)}{4(1-\delta^2)}\right)\epsilon(N) + o\,(\epsilon(N))\,.$$

Proof: We choose

$$\gamma = \frac{1+c+\delta^2\,(c-3)}{2\delta^2}.$$

Consider the polynomial f defined in (42) and let $\beta^-_{min}(\delta,\gamma,N)$ and $\beta^-_{max}(\delta,\gamma,N)$ be as defined in Lemma III.1 (see Appendix) relatively to the matrix P_N. Notice that $f(1,\delta,\gamma,\beta) = \gamma\beta^2 + \left(\gamma\delta^2 - \gamma\right)\beta$. Then the equation $f(1,\delta,\gamma,\beta) = 0$ has solutions $\beta = 0$ and $\beta = 1-\delta^2$. This implies that, since $\lambda_{\max}(P_N) \to 1$, then $\beta^-_{max}(\delta,\gamma,N) \to 0$ as $N \to \infty$. This implies that for N big enough we have that

$$\min\{\beta^-_{min}(\delta,\gamma,N), \beta^-_{max}(\delta,\gamma,N)\} = \beta^-_{max}(\delta,\gamma,N)$$

and hence, from Theorem 5.2 and Lemma III.1 (see Appendix), it follows that for N big enough we have that

$$\tilde{\rho}(P_N, \delta) \leq \sqrt{1 - \beta^{-}_{max}(\delta, \gamma, N)}$$

Let $\lambda_N := \lambda_{\max}(P_N)$ and $\beta_N := \beta^{-}_{max}(\delta, \gamma, N)$ so then we have that $\lambda_N \to 1$ and $\beta_N \to 0$. We know that $f(\lambda_N, \delta, \gamma, \beta_N) = 0$. As $N \to \infty$, from the implicit function theorem, we have that

$$\beta_N = \left[\frac{\frac{\partial}{\partial \lambda} f}{\frac{\partial}{\partial \beta} f}\right]_{|\lambda=1, \beta=0} \epsilon(N) + o(\epsilon(N)).$$

Now notice that

$$\begin{aligned}\frac{\partial f}{\partial \lambda} = &\left(-3(1-\lambda)^2 + 2\gamma\delta^2(\lambda-2)^2 - \gamma^2\delta^2 + 2\gamma\delta)\right)\beta - \left(-\gamma^2\delta^2 + \gamma(1+\lambda+\delta^2(\lambda-3)) - (1-\lambda)^2\right) + \\ &+ (1-\lambda)(\gamma - 2\gamma\delta^2 + 2(1-\lambda))\end{aligned}$$

and that

$$\frac{\partial f}{\partial \beta} = 2\gamma\beta + (1-\lambda)^3 + \gamma\delta^2(\lambda-2)^2 - \gamma + \gamma(1-\lambda)(\gamma\delta^2 - 1 - \lambda).$$

which lead to

$$\frac{\partial f}{\partial \lambda}_{|\lambda=1,\beta=0} = -(2\gamma - 2\gamma\delta^2 - \gamma^2\delta^2)$$

and

$$\frac{\partial f}{\partial \beta}_{|\lambda=1,\beta=0} = \gamma\delta^2 - \gamma.$$

Then

$$\beta_N = \left(2 - \frac{\gamma\delta^2}{1-\delta^2}\right)\epsilon(N) + o(\epsilon(N)).$$

The thesis follows by expanding in Taylor's series the function $\sqrt{1-\beta_N}$. □

Notice that the coefficient in front of $\epsilon(N)$ is negative. Indeed, it can be seen that coefficient is negative if and only if

$$\delta^2 < \frac{3-c}{1+c}$$

and this is true since we have chosen $\delta < \frac{1+c}{3-c}$ and since $\delta < 1$.

6. Numerical simulations

In this section we consider two examples providing some numerical results illustrating the performance respectively of the Zoom in -Zoom out strategy and of the truncated version of the logarithmic quantizer discussed in Remark 3.1.

Example 6.1. In this example we consider a connected random geometric graph generated by choosing N points at random in the unit square, and then placing an edge between each pair of points at distance less than 0.25. We assume that $N = 30$ and that the initial conditions has been generated randomly inside the interval $[-100, 100]$. Again, the matrix P is built using the Metropolis weights. For all the experiments, we set the parameters k_{in} and k_{out} to the values 1/2 and 2 respectively, and initialized the scaling factor f of each agent to the value 50. Moreover we run simulations for two different values of m, $m = 5$ and $m = 10$. The results obtained are reported in Figure 5. The variable plotted is the normalized Euclidean norm of the vector $\bar{x}(t) := x(t) - x_a(0)\mathbf{1}$, that is,

$$s(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\bar{x}_i^2(t)}.$$

Note that, as depicted in Figure 5, also the zoom in- zoom out uniform coder- decoder strategy seems to be very efficient in achieving the consensus. In particular it is remarkable that this strategy works well even if the uniform quantizer has a low number of quantization levels ($m = 5$). Finally it is worth observing, that as theoretically proved in the logarithmic coder-decoder strategy, also in this case the performance degrades smoothly as the quantization becomes coarser.

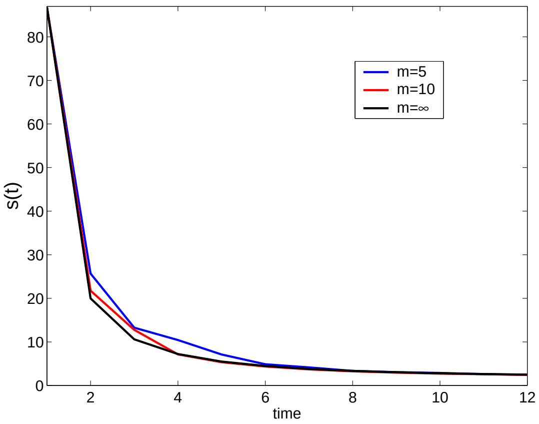

Figure 5. Zoom in- zoom out strategy

Example 6.2. In this example we consider the same matrix P of the previous example. Moreover we assume again that the initial conditions have been generated randomly inside the interval $[-100, 100]$. The information exchanged between the systems is quantized by the *truncated* logarithmic quantizer discussed in Remark 3.1. More precisely, we assume that the real numbers a, b introduced in Remark 3.1 are equal respectively to 0.5 and 100. The result obtained is reported in Figure 6. The variable plotted is

$$d(t) := \|\bar{x}(t)\|_\infty.$$

One can see that $d(t)$ does not converge asymptotically to 0. However, at the steady state, $d(t)$ oscillates inside an interval whose amplitude is comparable to 0.5, that is, the lower value at which we have truncated the logarithmic quantizer.

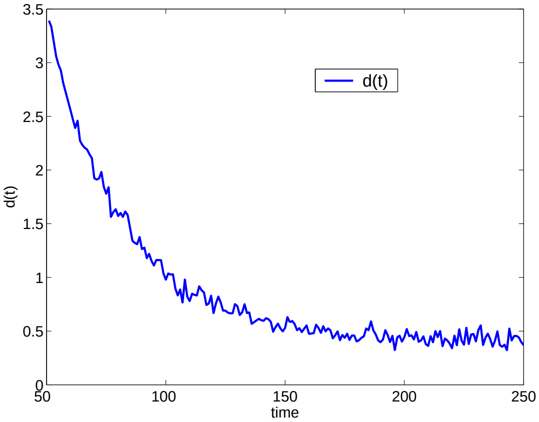

Figure 6. Zoom in- zoom out strategy

This numerical observations leads to the following consideration. Assume that our goal is to have convergence of the initial states $x_i(0) \in [-M, M]$ to a target configuration $x_i(\infty) \in [\alpha - \epsilon, \alpha + \epsilon]$, where α is a constant depending only on the initial condition $x(0)$ and ϵ describes the desired agreement precision. This is a "practical stability" requirement. In this case the contraction rate is $C := M/\epsilon$. Assume that, as in [3], the exact data transmission are substituted by transmissions of precision ϵ uniformly quantized data. In this framework it is well known [17] that each uniform quantizer needs C different levels and so the transmission of its data needs an alphabet of C different symbols. Assume now that the information is encoded by truncated logarithmic quantizers where $a = \epsilon$ and $b = M$. We have seen in Remark 3.1 that in such case each logarithmic quantizer needs

$$\frac{2 \log C}{\log \frac{1+\delta}{1-\delta}}$$

different symbols. Note that for C sufficiently large, with the logarithmic communications we obtain a significantly improvement in terms of the communication effort required. It will be the subject of future research to analyze the tradeoff between the steady state of $d(t)$ and the values of the parameters a, b at which we truncate the logarithmic quantizers.

7. Conclusions

In this paper we presented a new approach solving the average consensus problem in presence of only quantized exchanges of information. In particular we considered two strategies, one based on logarithmic quantizers, and the other one based on a zooming in-zooming out strategy. We studied them with theoretical and experimental results proving that using these schemes the average consensus problem can be efficiently solved even if the agents can share only quantized information. Additionally, we show that the convergence factors depend smoothly on the accuracy parameter of the quantized and, remarkably, that the critical quantizer accuracy sufficient to guarantee convergence is independent from the network dimension. A field of future research will be to look for encoding and decoding methods which are able to solve the average problem also with noisy digital channels.

Acknowledgements This work was supported in part by AFOSR MURI award FA9550-07-1-0528 and ONR award N00014-07-1-0721. An early version of this work appeared as [18].

REFERENCES

1. R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
2. R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, "Communication constraints in the average consensus problem," *Automatica*, vol. 44, no. 3, pp. 671–684, 2008.
3. A. Kashyap, T. Başar, and R. Srikant, "Quantized consensus," *Automatica*, vol. 43, no. 7, pp. 1192–1203, 2007.
4. M. E. Yildiz and A. Scaglione, "Differential nested lattice encoding for consensus problems," in *Symposium on Information Processing of Sensor Networks (IPSN)*, (Cambridge, MA), pp. 89–98, Apr. 2007.
5. L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.
6. R. Carli, F. Fagnani, P. Frasca, T. Taylor, and S. Zampieri, "Average consensus on networks with transmission noise or quantization," in *European Control Conference*, (Kos, Greece), pp. 1852–1857, June 2007.
7. T. C. Aysal, M. Coates, and M. Rabbat, "Distributed average consensus using probabilistic quantization," in *IEEE Workshop on Statistical Signal Processing*, (Maddison, Wisconsin), pp. 640–644, Aug. 2007.
8. R. Carli and F. Bullo, "Quantized coordination algorithms for rendezvous and deployment," *SIAM Journal on Control and Optimization*, Dec. 2007. Submitted.
9. G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans, "Feedback control under data rate constraints: An overview," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 108–137, 2007.
10. N. Elia and S. K. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1384–1400, 2001.
11. F. Fagnani and S. Zampieri, "Performance evaluations of quantized stabilizers," in *IEEE Conf. on Decision and Control*, (Maui, HI), pp. 1897–1901, Dec. 2003.
12. J. R. Silvester, "Determinants of block matrices," *The Mathematical Gazette*, vol. 84, no. 501, pp. 460–467, 2000.
13. S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, vol. 15 of *Studies in Applied Mathematics*. Philadelphia, Pennsylvania: SIAM, 1994.
14. L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Symposium on Information Processing of Sensor Networks (IPSN)*, (Los Angeles, CA), pp. 63–70, Apr. 2005.
15. S. Martínez, F. Bullo, J. Cortés, and E. Frazzoli, "On synchronous robotic networks – Part I: Models, tasks and complexity," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2199–2213, 2007.
16. S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis, and application," in *IEEE Conference on Computer Communications (INFOCOM)*, pp. 1653–1664, Mar. 2005.
17. F. Fagnani and S. Zampieri, "Quantized stabilization of linear systems: Complexity versus performances," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1534–1548, 2004.
18. R. Carli, F. Bullo, and S. Zampieri, "Quantized average consensus via dynamic coding/decoding schemes," in *IEEE Conf. on Decision and Control*, (Cancun, Mexico), Dec. 2008. To appear.

APPENDIX

II. Stability of discrete time linear parameter varying (LPV) systems

Given $A_1, \dots, A_k \in \mathbb{R}^{n\times n}$, we let $\{A(t)\}_{t=0}^{+\infty} \subset \mathrm{Co}\{A_1, \dots, A_k\}$ denote a sequence of matrices taking values in the convex hull of $\{A_1, \dots, A_k\}$. We consider the dynamical system

$$x(t+1) = A(t)x(t). \tag{36}$$

The following result is an extension to the discrete-time system (36) of the classical results stated in [13], in the context of continuous-time LPV systems. The proof can be found in [8].

Theorem II.1 (Common Lyapunov function for convergence to eigenspace) *Assume that 1 is a simple eigenvalue with left and right eigenvector $v \in \mathbb{R}^n$ for each matrix $A_1, \dots, A_k \in \mathbb{R}^{n\times n}$. If there exists a symmetric matrix $P \in \mathbb{R}^{n\times n}$ satisfying, for all nonzero $z \notin \mathrm{span}\{v\}$,*

$$Pv = 0, \tag{37}$$

$$z^T P z > 0, \tag{38}$$

and

$$z^T \left(\frac{A_i^T P A_j + A_j^T P A_i}{2} - P \right) z < 0, \quad \textit{for all } i, j \in \{1, \dots, k\}, \tag{39}$$

then, for all initial conditions $x(0) \in \mathbb{R}^n$ and sequences $\{A(t)\}_{t=0}^{+\infty} \subset \mathrm{Co}\{A_1, \dots, A_k\}$, the solution to (36) satisfies

$$\lim_{t\to+\infty} x(t) = \alpha v, \quad \alpha = \frac{1}{\|v\|^2} v^T x(0).$$

III. Solvability of a Lyapunov equation

The proof of some results contained in the paper are based on the solvability of the following Lyapunov equation

$$z^T(R_1^T L R_1 - (1-\beta)L)z < 0, \qquad \forall\ z \in < \begin{bmatrix}\mathbf{1}^T & \mathbf{0}^T\end{bmatrix}^T >^\perp \tag{40}$$

where

$$L = \begin{bmatrix} I-P & 0 \\ 0 & \gamma I \end{bmatrix}, \qquad R_1 = \begin{bmatrix} I & 0 \\ 0 & \delta I \end{bmatrix} \begin{bmatrix} P & P-I \\ P-I & P-2I \end{bmatrix}, \tag{41}$$

$P \in \mathbb{R}^{N\times N}$ satisfies Assumption 2.1, $0 < \delta < 1$ and $\gamma > 0$.

The following lemma helps to determine for what parameters γ, β, δ the Lyapunov inequality (40) holds.

Lemma III.1. *Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimum and the maximum eigenvalue in $\sigma(P)\backslash\{1\}$, respectively. Define the polynomial*

$$f(\lambda,\delta,\gamma,\beta) := \gamma\beta^2 + \left((1-\lambda)^3 + \gamma\delta^2(\lambda-2)^2 - \gamma + \gamma(1-\lambda)(\gamma\delta^2 - 1 - \lambda)\right)\beta + \\ + (1-\lambda)\left(-\gamma^2\delta^2 + \gamma(1+\lambda+\delta^2(\lambda-3)) - (1-\lambda)^2\right). \tag{42}$$

Then (40) holds true if and only if

$$\beta < \min\left\{\beta^-_{min}(\delta,\gamma), \beta^-_{max}(\delta,\gamma)\right\}, \tag{43}$$

where $\beta^-_{min}(\delta,\gamma)$ and $\beta^-_{max}(\delta,\gamma)$ are the minimum real roots of $f(\lambda_{\min},\delta,\gamma,\beta)$ and of $f(\lambda_{\max},\delta,\gamma,\beta)$ as polynomials in β.

Proof: We start by observing that

$$\begin{aligned} &R_1^T L R_1 - (1-\beta)L = \\ &= \left[\begin{array}{cc} (I-P)^2(\gamma\delta^2 I - I - P) + \beta(I-P) & (I-P)(P(P-I) - \gamma\delta^2(P-2I)) \\ (I-P)(P(P-I) - \gamma\delta^2(P-2I)) & (I-P)^3 + \gamma\delta^2(P-2I)^2 - \gamma I + \gamma\beta I \end{array}\right]. \end{aligned}$$

Note that

$$(R_1^T L R_1 - (1-\beta)L)\left[\begin{array}{c} \mathbf{1} \\ \mathbf{0} \end{array}\right] = 0$$

and hence showing (40) is equivalent to show that the symmetric matrix $-R_1^T L R_1 + (1-\beta)L$ has all positive eigenvalues except one, which is zero and has multiplicity one. If we define the polynomials $q_{11}(\lambda), q_{22}(\lambda), q_{12}(\lambda)$ as follows

$$\begin{aligned} q_{11}(\lambda) &:= (1-\lambda)^2(\gamma\delta^2 - 1 - \lambda) + \beta(1-\lambda) \\ q_{22}(\lambda) &:= (1-\lambda)^3 + \gamma\delta^2(\lambda-2)^2 - \gamma + \gamma\beta \\ q_{12}(\lambda) &:= (1-\lambda)(\lambda(\lambda-1) - \gamma\delta^2(\lambda-2)) \end{aligned}$$

we can write

$$R_1^T L R_1 - (1-\beta)L = \left[\begin{array}{cc} q_{11}(P) & q_{12}(P) \\ q_{12}(P) & q_{22}(P) \end{array}\right].$$

To compute the eigenvalues of $-R_1^T L R_1 + (1-\beta)L$ we consider its characteristic polynomial. Using the same arguments used in the proof of Lemma 4.7 we can argue that

$$\det(sI + R_1^T L R_1 - (1-\beta)L) \;=\; \prod_{i=0}^{N-1}\left[s^2 + (q_{11}(\lambda_i) + q_{22}(\lambda_i))s + (q_{11}(\lambda_i)q_{22}(\lambda_i) - q_{12}(\lambda_i)^2)\right]$$

where $\lambda_0 = 1, \lambda_1, \ldots, \lambda_{N-1}$ denote the eigenvalues of P. Observe now that, for $i = 0$ the polynomial in the previous product is

$$s(s - \gamma\delta^2 + \gamma - \gamma\beta)$$

which gives one zero eigenvalue and another eigenvalue equal to $\gamma(1-\delta^2-\beta)$. We can argue that, since this must positive, then we have this first constraint

$$\beta < 1 - \delta^2 \tag{44}$$

Moreover all the roots of the other polynomials for $i = 1, \ldots, N-1$ must be all positive. Observe that $s^2 + (q_{11}(\lambda_i) + q_{22}(\lambda_i))s + (q_{11}(\lambda_i)q_{22}(\lambda_i) - q_{12}(\lambda_i)^2)$ can be seen as the characteristic polynomial of the 2×2 matrix

$$\left[\begin{array}{cc} -q_{11}(\lambda_i) & -q_{12}(\lambda_i) \\ -q_{12}(\lambda_i) & -q_{22}(\lambda_i) \end{array}\right]$$

and, to impose that the roots of this polynomial are positive is equivalent to impose that such a matrix is positive definite and so that

$$-q_{11}(\lambda_i) > 0 \qquad\qquad q_{11}(\lambda_i)q_{22}(\lambda_i) - q_{12}^2(\lambda_i) > 0.$$

Therefore, together with condition (44), we have other $2N-2$ conditions. Some of these conditions are superfluous. We start from condition $-q_{11}(\lambda_i) > 0$. Observe that, since $1-\lambda_i > 0$ for all $i \in \{1,\dots,N-1\}$, then $q_{11}(\lambda_i) < 0$ for all $i \in \{1,\dots,N-1\}$ if and only if $\beta + \gamma\delta^2(1-\lambda_i) - 1 + \lambda_i^2 < 0$ for all $i \in \{1,\dots,N-1\}$ and this happens if and only if

$$\beta < 1 - \lambda_{\min}^2 - \gamma\delta^2(1-\lambda_{\min}) \tag{45}$$

$$\beta < 1 - \lambda_{\max}^2 - \gamma\delta^2(1-\lambda_{\max}) \tag{46}$$

Notice now that

$$q_{11}(\lambda)q_{22}(\lambda) - q_{12}^2(\lambda) = (1-\lambda)f(\lambda)$$

where $f(\lambda) = f(\lambda,\delta,\gamma,\beta)$ is defined in (42). Notice that

$$\frac{\partial^2 f}{\partial \lambda^2} = (1-\beta)(6\lambda - 6 - 2\gamma(1+\delta^2))$$

which is negative for $\lambda < 1$. This implies that $f(\lambda)$, is a concave function in λ, for $\lambda \in [-1,1]$ and so $q_{11}(\lambda_i)q_{22}(\lambda_i) - q_{12}^2(\lambda_i) > 0$ for all $i = 1,\dots,N-1$ if and only if

$$f(\lambda_{\min},\delta,\gamma,\beta) > 0 \tag{47}$$

$$f(\lambda_{\max},\delta,\gamma,\beta) > 0. \tag{48}$$

At this point we have that (40) holds true if and only if conditions (44), (45), (46), (47) and (48) hold true. Consider the condition $f(\lambda_{\min},\delta,\gamma,\beta) > 0$. Observe that, if $\beta = 1 - \lambda_{\min}^2 - \gamma\delta^2(1-\lambda_{\min})$, then $q_{11} = 0$ and so

$$f(\lambda_{\min},\delta,\gamma,\beta)_{|\beta = 1-\lambda_{\min}^2-\gamma\delta^2(1-\lambda_{\min})} = \frac{-q_{12}^2}{1-\lambda} < 0.$$

We can argue that $f(\lambda_{\min},\delta,\gamma,\beta)$ is a convex parabola in β which has always two real roots $\beta_{min}^-(\delta,\gamma)$ and $\beta_{min}^+(\delta,\gamma)$ which satisfy

$$\beta_{min}^-(\delta,\gamma) < 1 - \lambda_{\min}^2 - \gamma\delta^2(1-\lambda_{\min}) < \beta_{min}^+(\delta,\gamma) \tag{49}$$

and moreover $f(\lambda_{\min},\delta,\gamma,\beta) > 0$ if and only if

$$\beta < \beta_{min}^-(\delta,\gamma) \qquad \text{or} \qquad \beta > \beta_{min}^+(\delta,\gamma). \tag{50}$$

This implies that conditions (45) and (47) hold if and only if $\beta < \beta_{min}^-(\delta,\gamma)$. Reasoning similarly for the condition $f(\lambda_{\max},\delta,\gamma,\beta) > 0$ we obtain that conditions (46) and (48) hold if and only if $\beta < \beta_{max}^-(\delta,\gamma)$.

We prove finally that condition (44) is superfluous which would give the thesis. To prove this observe that

$$\begin{aligned} f(\lambda_{\min},\delta,\gamma,\beta)_{|\beta=0} &= (1-\lambda_{\min})[-\lambda_{\min}^2 + (2+\gamma\delta^2+\gamma)\lambda_{\min} - (1+3\gamma\delta^2+\gamma^2\delta^2-\gamma)] \\ f(\lambda_{\min},\delta,\gamma,\beta)_{|\beta=1-\delta^2} &= \delta^2(1-\lambda_{\min})[-\lambda_{\min}^2 + (2+\gamma\delta^2+\gamma)\lambda_{\min} - (1+3\gamma\delta^2+\gamma^2\delta^2-\gamma)] \end{aligned}$$

This implies that we can have three cases

1. We have $f(\lambda_{\min}, \delta, \gamma, \beta)_{|\beta=0} = 0$. In this case we have that $\beta^{-}_{min}(\delta, \gamma) = 0$ and $\beta^{+}_{min}(\delta, \gamma) = 1 - \delta^2$.
2. We have $f(\lambda_{\min}, \delta, \gamma, \beta)_{|\beta=0} < 0$. In this case we have that $\beta^{-}_{min}(\delta, \gamma) < 0 < 1 - \delta^2 < \beta^{+}_{min}(\delta, \gamma)$;
3. We have $f(\lambda_{\min}, \delta, \gamma, \beta)_{|\beta=0} > 0$. In this case we may have three situations:

   1a. $0 < \beta^{-}_{min}(\delta, \gamma) \leq \beta^{+}_{min}(\delta, \gamma) < 1 - \delta^2$;
   2b. $0 < 1 - \delta^2 < \beta^{-}_{min}(\gamma) \leq \beta^{+}_{min}(\delta, \gamma)$;
   3c. $\beta^{-}_{min}(\delta, \gamma) \leq \beta^{+}_{min}(\delta, \gamma) < 0 < 1 - \delta^2$.

However the cases 2b and 2c cannot occur since the $\beta^{-}_{min}(\delta, \gamma)$ is a continuous function of γ, while in these two cases the value of this function would pass from 0 to $1 - \delta^2$ in a neighbor of the γ's such that $f(\lambda_{\min}, \delta, \gamma, \beta)_{|\beta=0} = 0$ Notice finally that in all the cases which can occur we have that $\beta^{-}_{min}(\delta, \gamma) \leq 1 - \delta^2$. □

We provide now a consequence of the previous result.

Lemma III.2. *Assume the same assumptions of the previous lemma hold. Let $\bar{\delta}$ be defined as in* (13) *and $\delta \in \mathbb{R}$ be such that $0 \leq \delta < \bar{\delta}$. Moreover let*

$$\bar{\gamma} := \frac{1 + \lambda_{\min} + \delta^2(\lambda_{\min} - 3)}{2\delta^2}.$$

Then $\bar{\gamma} > 0$ and the following inequality

$$z^T(R_1^T L R_1 - L)z < 0, \qquad \forall \;\; z \in < \begin{bmatrix} \mathbf{1}^T & \mathbf{0}^T \end{bmatrix}^T >^{\perp}, \tag{51}$$

holds true.

Proof: Notice first that $0 \leq \delta < \bar{\delta}$ implies that $\bar{\gamma} > 0$. By the previous lemma, (51) holds true if and only if $\beta = 0$ is an admissible solution of (43) and this happens if and only if both $\beta^{-}_{min}(\delta, \bar{\gamma}) > 0$ and $\beta^{-}_{max}(\delta, \bar{\gamma}) > 0$. Notice now that $f(\lambda, \delta, \gamma, \beta)$ can be written as follows

$$f(\lambda, \delta, \gamma, \beta) = \gamma\beta^2 + [(1 - \lambda)p(\lambda, \delta, \gamma) - \gamma(1 - \delta^2)]\beta - (1 - \lambda)p(\lambda, \delta, \gamma)$$

where

$$p(\lambda, \delta, \gamma) = \delta^2\gamma^2 - [1 + \lambda + \delta^2(\lambda - 3)]\gamma + (1 - \lambda)^2.$$

Notice moreover that, $\beta^{-}_{min}(\delta, \bar{\gamma}) > 0$ if and only if $(1 - \lambda_{\min})p(\lambda_{\min}, \delta, \bar{\gamma}) - \bar{\gamma}(1 - \delta^2) < 0$ and $(1 - \lambda_{\min})p(\lambda_{\min}, \delta, \bar{\gamma}) < 0$ and these two conditions occurs if and only if $p(\lambda_{\min}, \delta, \bar{\gamma}) < 0$. Similarly we can see that $\beta^{-}_{max}(\delta, \bar{\gamma}) > 0$ if and only if $p(\lambda_{\max}, \delta, \bar{\gamma}) < 0$. Notice now that, since

$$\frac{\partial p}{\partial \lambda} = -\gamma - \gamma\delta^2 - 2(1 - \lambda),$$

is negative for $\lambda < 1$, then $p(\lambda_{\max}, \delta, \bar{\gamma}) < 0$ is implied by $p(\lambda_{\min}, \delta, \bar{\gamma}) < 0$ which is the only condition we need to prove. Notice now that $\bar{\gamma}$ is the minimizer of $p(\lambda_{\min}, \delta, \gamma)$ as a function of γ. Therefore $p(\lambda_{\max}, \delta, \bar{\gamma}) < 0$ if and only if the discriminant is positive, namely if and only if $(1 + \lambda_{\min} + \delta^2(\lambda_{\min} - 3))^2 - 4\delta^2(1 - \lambda_{\min})^2 > 0$. Observe that this last inequality holds true if and only if

$$(3 - \lambda_{\min})^2\delta^4 - 2(5 - 2\lambda_{\min} + \lambda_{\min}^2)\delta^2 + (1 + \lambda_{\min})^2 > 0. \tag{52}$$

Consider the equation $(3-\lambda_{\min})^2x^2-2(5-\lambda_{\min}+\lambda_{\min}^2)x+(1+\lambda_{\min})^2=0$. The solutions of this equation are $x_1=1$ and $x_2=\left(\frac{1+\lambda_{\min}}{3-\lambda_{\min}}\right)^2$. Since $\lambda_{\min}<1$ we have that $x_2<1$. Hence, since $(3-\lambda_{\min})^2x^2-2(5-\lambda_{\min}+\lambda_{\min}^2)x+(1+\lambda_{\min})^2>0$ is a convex parabola, we have that $(3-\lambda_{\min})^2x^2-2(5-\lambda_{\min}+\lambda_{\min}^2)x+(1+\lambda_{\min})^2>0$ if and only if $x<x_2$ and $x>x_1$. It follows that, if $\delta^2<\left(\frac{1+\lambda_{\min}}{3-\lambda_{\min}}\right)^2$, i.e., if $\delta<\bar{\delta}$, then (52) is satisfied. □