Stability, Linear Convergence, and Robustness of the Wang-Elia Algorithm for Distributed Consensus Optimization

Michelangelo Bin, Ivano Notarnicola, and Thomas Parisini

Abstract—We revisit an algorithm for distributed consensus optimization proposed in 2010 by J. Wang and N. Elia. By means of a Lyapunov-based analysis, we prove input-to-state stability of the algorithm relative to a closed invariant set composed of optimal equilibria and with respect to perturbations affecting the algorithm’s dynamics. In the absence of perturbations, this result implies linear convergence of the local estimates and Lyapunov stability of the optimal steady state. Moreover, we unveil fundamental connections with the well-known Gradient Tracking and with distributed integral control. Overall, our results suggest that a control theoretic approach can have a considerable impact on (distributed) optimization, especially when robustness is considered.

I. INTRODUCTION

A. Problem Overview and Literature Review

We consider $N$ agents communicating through a connected, undirected network represented by a simple graph $(\mathcal{N}, \mathcal{E})$, with $\mathcal{N} = \{1, \ldots, N\}$ and $\mathcal{E} \subset \mathbb{N}^2$. By exchanging information with neighbors, agents cooperatively seek a consensual solution $\theta^\ast \in \mathbb{R}$ to the optimization problem

$$\min_{\theta \in \mathbb{R}} \sum_{i \in \mathcal{N}} f_i(\theta) \quad (1)$$

where, for each $i \in \mathcal{N}$, the function $f_i : \mathbb{R} \to \mathbb{R}$ is known to agent $i$ only. Problem (1) is known as a cost-coupled or consensus optimization problem, since agents minimize a global cost function $\sum_{i \in \mathcal{N}} f_i$ in a common decision variable. As each agent $i \in \mathcal{N}$ has only access to its own private function $f_i$, and not to the global cost function to be optimized, a distributed solution of Problem (1) is nontrivial.

Cost-coupled problems have been extensively investigated in the last decades starting with the pioneering works [1]–[3]. A detailed account for the large amount of research on this topic can be found in the recent survey papers [4]–[8]. In particular, an important step forward in the algorithmic solution of (1) was the introduction of a “tracking” protocol

in the distributed gradient method. See, e.g., [9]–[17] and the subsequent extensions [18], [19]. The algorithms based on this tracking protocol are known as Gradient Tracking algorithms. According to the early interpretations, the tracking protocol aims at reconstructing, in a distributed way, the gradient of the global cost function. A recent interpretation, instead, looks at the Gradient Tracking algorithms as embedding a distributed integral action [20]. As we discuss in Section III, this is one of the connection points with the Wang-Elia algorithm [3] introduced later.

A main drawback of the Gradient Tracking algorithm is that it needs a specific initialization (see Section III) to work properly. As we clarify later in Section III, such initialization requirement makes the Gradient Tracking methods fragile with respect to uncertainties in the dynamics, such as those introduced by quantization, numerical errors in the computation of the gradients, or uncertainties affecting the communication with the neighbors. In particular, as the example in Section V shows, even a small quantization error can make the Gradient Tracking diverge to infinity, with a divergence rate that worsens for smaller stepsize values.

In [3], a distributed algorithm was proposed for problem (1) that does not require any specific initialization. We refer to it as the Wang-Elia algorithm. A continuous-time version of this algorithm was also studied in [21] from a passivity-theoretic viewpoint. The discrete-time version, instead, represents the main subject of this work. In particular, in the Wang-Elia algorithm, each agent $i \in \mathcal{N}$ maintains a pair of state variables $(x_i, z_i) \in \mathbb{R}^2$ that are updated as

$$x_i^+ = x_i + \sum_{j \in \mathcal{N}_i} \beta a_{ij} (x_j - x_i + z_j - z_i) - \alpha \beta \nabla f_i(x_i) \quad (2)$$

in which $\mathcal{N}_i := \{j \in \mathcal{N} \mid (i, j) \in \mathcal{E}\}$ is the neighborhood of $i$ (we stress that $i \notin \mathcal{N}_i$) in the communication network $(\mathcal{N}, \mathcal{E})$, $a_{ij} = a_{ji} > 0$ for all $(i, j) \in \mathcal{E}$, and $\alpha, \beta > 0$ are design parameters. The variable $x_i$ is the estimate Agent $i$ has of the optimal solution $\theta^\ast$ of Problem (1), and $z_i$ is an auxiliary state variable. It was proved in [3] that, if each function $f_i$ is convex and $\alpha, \beta$ are chosen small enough, then all estimates $x_i$ converge to $\theta^\ast$.

B. Contribution

We study the Wang-Elia algorithm (2) in the presence of additive perturbations. We prove global input-to-state stability (ISS) [22] of the algorithm with respect to such perturbations and relative to a closed invariant set $\mathcal{A}^\ast \subset \mathbb{R}^{2N}$. 

---

1In this paper we focus on the single-variable case in which $\theta \in \mathbb{R}$. This simplifies the technical derivations without sacrificing generality, since all results reported in the paper directly extend to the case where $\theta \in \mathbb{R}^m$, for some $m > 1$, by properly introducing Kronecker products.
The elements \((x^*, z^*) \in A^*\) are all optimal in the sense that all estimates \(x^*_i\) equal the minimizer \(\theta^*\). In this way, we prove that, unlike the Gradient Tracking methods, the Wang-Elia algorithm is robust with respect to perturbations. Moreover, in the absence of perturbations as in \([3]\), our results establish Lyapunov stability of \(A^*\) and linear convergence of the local estimates \(x_i\) to \(\theta^*\), which are stronger properties than only convergence as shown in \([3]\). Finally, we compare the Wang-Elia and the Gradient Tracking algorithms, unveiling their similarities and differences, and making a connection with (distributed) integral control. The developed analysis is based on Lyapunov arguments and provides further insights on the structure and functioning of the algorithm.

C. Notation

We denote by \(\sigma(M)\) the spectrum of a matrix \(M\) and we call it Schur if \(\sigma(M)\) lies in the open unit disk. The vector and matrix-induced 2-norms are denoted by \(|\cdot|\). The distance of \(x \in \mathbb{R}^n\) to a closed set \(A \subset \mathbb{R}^n\) is denoted by \(|x|_A := \inf_{a \in A} |x - a|\). If \(s : \mathbb{N} \to \mathbb{R}^n\), we let \(s^+\) denote the shift operator \(s \mapsto s^+(\cdot) = s(\cdot + 1)\), and \(|s| := \sup_{k=0,\ldots,t} |s(k)|\). For compactness, we also write \(s^t\) in place of \(s(t)\). For a given \(N\), we let \(1 := (1, \ldots, 1) \in \mathbb{R}^N\) and we let \(S \in \mathbb{R}^{N \times (N-1)}\) be a matrix satisfying

\[
S^T 1 = 0, \quad S^T S = I_{N-1}. \quad (3)
\]

We define the matrix \(T \in \mathbb{R}^{N \times N}\) and its inverse as

\[
T = \begin{bmatrix} 1^T / N \\ S^T \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} 1 & S \end{bmatrix}. \quad (4)
\]

From (3)-(4), we deduce that the identity matrix \(I_N\) satisfies

\[
I_N = 11^T / N + SS^T, \quad |S| = 1. \quad (5)
\]

Given a \(\chi \in \mathbb{R}^N\), we define its average-dispersion decomposition as the pair \((\chi_m, \chi_\perp) = T \chi\), where \(\chi_m := 1^T \chi / N \in \mathbb{R}\) and \(\chi_\perp := S^T \chi \in \mathbb{R}^{N-1}\) are called, respectively, the average and the dispersion components of \(\chi\). Also, it holds \(\chi = T^{-1}(\chi_m, \chi_\perp) = 1 \chi_m + S \chi_\perp\), and \(|\chi|^2 = N \chi_m^2 + |\chi_\perp|^2\).

II. THE WANG-ELIA ALGORITHM REVISITED

A. The Perturbed Wang-Elia Algorithm

In this paper, we study the following system

\[
x^+_i = x_i + \sum_{j \in N_i} k_{ij} (x_j - x_i + z_j - z_i) - \gamma \nabla f_i(x_i) + w_i
\]

\[
z^+_i = z_i - \sum_{j \in N_i} k_{ij} (x_j - x_i) + \nu_i, \quad (6)
\]

for all \(i \in N\), with arbitrary initial conditions \((x^0_i, z^0_i)\), with \(\gamma > 0\), and with \(k_{ij} = k_{ji} > 0\) for all \((i, j) \in E\). The signals \(w := (w_1, \ldots, w_N) : \mathbb{N} \to \mathbb{R}^N\) and \(\nu := (\nu_1, \ldots, \nu_N) : \mathbb{N} \to \mathbb{R}^N\) are perturbations modeling, e.g., uncertainties in the state measurements, in the computation of \(\nabla f_i(x_i)\), and in the exchange of the neighboring states, or representing quantization errors and generic unmodeled dynamics. The aggregate version of (6) reads as

\[
x^+ = (I - K)x - Kz - \gamma \Phi(x) + w, \quad x^0 \in \mathbb{R}^N, \quad (7a)
\]

\[
z^+ = z + Kx + \nu, \quad z^0 \in \mathbb{R}^N, \quad (7b)
\]

in which \(K \in \mathbb{R}^{N \times N}\) is defined in such a way that \(K_{ij} = -k_{ij}\) for all \((i, j) \in E\), \(K_{ii} = \sum_{j \in N_i} k_{ij}\) for all \(i \in N\), and \(K_{ij} = 0\) otherwise, and where \(x := (x_1, \ldots, x_N)\), \(z := (z_1, \ldots, z_N)\), and \(\Phi(x) := (\nabla f_1(x_1), \ldots, \nabla f_N(x_N))\).

Unlike [3], we do not factor \(k_{ij}\) and \(\gamma\) in terms of \(\beta\) and \(\alpha\) (cf. (2)). We only assume that the coefficients \(k_{ij}\) are chosen in such a way that \(K\) satisfies the following conditions

\[
K = K^T, \quad \ker K = \text{span} 1, \quad \sigma(K) \subset [0, 1), \quad (8)
\]

while the gain \(\gamma\) is a small positive number to be chosen according to Theorem 1 presented later in Section II-D.

We underline that the last condition of (8) is possible since the communication network is connected. Moreover, (8) implies \(1^T K = 0\) and that \(S^T KS\) is invertible and Schur.

B. Standing Assumptions

We study System (2) under the following assumptions.

Assumption 1: For each \(i \in N\), \(f_i\) is continuously differentiable and \(\nabla f_i\) is Lipschitz continuous.

Assumption 2: The global cost function \(\sum_{i \in N} f_i\) is strongly convex.

Assumption 2 does not directly compare to the assumptions of \([3]\), \([21]\), where convexity of each \(f_i\) is asked. Indeed, while Assumption 2 asks for strong convexity, such property is only required to the global cost function (as, e.g., in \([17]\)), and not to each function \(f_i\) individually.

Assumption 2 is not necessary to prove convergence of the estimates \(x_i\) produced by (7). However, when it holds, there is a natural choice among the optimal equilibria leading to a well-defined error system characterized by a simple structure. This supports a Lyapunov-based analysis allowing to establish, in addition to convergence, stronger stability and robustness properties.

C. Existence of an Optimal Steady-State Locus

Throughout the paper, when referring to an equilibrium of (7), we always implicitly assume \((w, \nu) = 0\). We say that a state \((x, z) \in \mathbb{R}^{2N}\) is consensually optimal if \(x_i = \bar{\theta}\) for all \(i \in N\), where \(\bar{\theta} \in \mathbb{R}\) is a critical point of the global cost function \(\sum_{i \in N} f_i\) (i.e., \(\sum_{i \in N} \nabla f_i(\theta) = 0\)). The equilibria of (7) are characterized by the following lemma.

Lemma 1: Suppose that \(K\) in (7) satisfies (8). Then, every equilibrium of (7) is consensually optimal. Conversely, if \(\theta\) is a critical point of the global cost function, there exists an equilibrium \((x, z)\) of (7) satisfying \(x_i = \theta\) for all \(i \in N\).

Proof: Consider (7) with \((w, \nu) = 0\). Then, \((x^e, z^e)\) is an equilibrium of (7) if and only if

\[
K x^e + K z^e + \gamma \Phi(x^e) = 0, \quad K x^e = 0. \quad (9)
\]

In view of (8), the second equation of (9) is equivalent to \(x^e \in \text{span} 1\). Hence, \((x^e, z^e)\) is an equilibrium of (7) if and only if \(x^e\) is a consensus point for the estimates \(x_i\).

Let \(\theta^e \in \mathbb{R}\) be such that \(x^e = 10^{\theta^e}\). Then, the first claim follows by noticing that, in view of (8), the first equation of (9) implies \(1^T \Phi(x^e) = \sum_{i \in N} \nabla f_i(\theta^e) = 0\).
For the converse direction, let $\bar{\theta} \in \mathbb{R}$ be a stationary point of the global cost function, and let $x^e := 1\bar{\theta}$. Then, $1^T\Phi(x^e) = 0$ and $Kx^e = 0$. Let $z^e := -\gamma S(S^TKS)^{-1}S^T\Phi(x^e)$, where $S^TKS$ is invertible in view of (8). Then, by repeatedly using (5), and in view of (8), we get $Kz^e = SS^Tz^e = -\gamma S(S^TKS)(S^TKS)^{-1}S^T\Phi(x^e) = -\gamma SS^T\Phi(x^e) = -\gamma \Phi(x^e)$. Hence, $(x^e, z^e)$ satisfies (9).

As the proof of Lemma 1 shows, the set of all equilibria of (7) (each of which is consensualy optimal) can be expressed as follows

$$A^* = \{(x, z) \in \mathbb{R}^{2N} | \exists \theta \in \mathbb{R}, \sum_{i \in N} \nabla f_i(\theta) = 0, x = 1\theta, z \in -\gamma S(S^TKS)^{-1}S^T\Phi(\theta) + \text{span}1\}$$

which is closed but not compact. We point out that Lemma 1 does not rely on the smoothness and convexity assumptions.

The set $A^*$ is the target steady-state locus of the forthcoming stability results and analysis. We stress that we cannot target a compact subset of $A^*$ if global convergence is sought. Indeed, in view of (8), even with $(w, \nu) = 0$ the average component $z_m = 1^Tz/N$ of $z$ remains constant along every solution of (7). We stress that the same holds also for the original algorithm (2) as well as for the continuous-time counterpart, which therefore cannot have a compact attractor. We underline that this property holds also for the Gradient Tracking algorithm, see Section III.

**D. Main Result and Discussion**

A tuple $(x, z, w, \nu) : \mathbb{N} \rightarrow \mathbb{R}^{4N}$ satisfying (7) is called a solution tuple of (7). We say that $\nu$ is integral-average bounded if $t \rightarrow \sum_{i \in \mathbb{N}} \nu_m(t)$ is bounded, where $\nu_m$ denotes the average component of $\nu$ (Section I-C).

**Theorem 1:** Suppose that Assumptions 1 and 2 hold and that $K$ in (7) satisfies (8). Then, there exist $\gamma^*, \alpha > 0$ and, for each $\gamma \in (0, \gamma^*)$, there exist $\mu_0 \in (0, 1)$ and $\rho_\gamma, \tau_\gamma > 0$, such that every solution tuple $(x, z, w, \nu)$ of (7) satisfies

$$|(x^t, z^t)|_{A^*} \leq \alpha \mu_0^{|t|} (|x^0, z^0|_{A^*} + \rho_\gamma |w_m|_{t-1} + \tau_\gamma |(w_\perp, \nu_\perp)|_{t-1})$$

for all $t \in \mathbb{N}$. In particular, if $w$ and $\nu$ are bounded, then $x$ and $z_\perp$ are bounded. Moreover, if and only if $\nu$ is integral-average bounded, also $z_m$ (hence, $(x, z)$) is bounded.

Theorem 1 is proved in Section IV. Under Assumption 2, $|x - 10\theta^*| \leq |(x, z)|_{A^*}$. Hence, when $(w, \nu) = 0$, Theorem 1 implies exponential convergence of the estimates $x_i$ to the optimum $\theta^*$ with convergence rate $\mu_\gamma = \sqrt{1 - c_0^2}$, being $c_0$ related to the convexity parameter of the global cost function (see Section IV). We stress that convergence is global in the initial conditions, unlike the Gradient Tracking (see Section III below). Moreover, by means of standard ISS arguments [22], one can show that (10) implies

$$\lim_{t \rightarrow \infty} \sup_{A^*} |(x^t, z^t)|_{A^*} \leq \lim_{t \rightarrow \infty} \sup_{A^*} \left( \rho_\gamma |w_m|_{t} + \tau_\gamma |(w_\perp, \nu_\perp)|_{t} \right).$$

Thus, in particular, the estimates converge to $\theta^*$ at front of every vanishing perturbation.

Furthermore, Theorem 1 implies that the set $A^*$ is Lyapunov stable when $(w, \nu) = 0$, and **strongly stable** when $(w, \nu) \neq 0$. Namely, for every $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $\max \{|(x^t, z^0)|_{A^*}, \sup_{t \in \mathbb{N}} |w_m|_{t}, \sup_{t \in \mathbb{N}} |(w_\perp, \nu_\perp)|_{t} \} < \delta_\epsilon$ implies $|(x^t, z^t)|_{A^*} < \epsilon$ for all $t \in \mathbb{N}$. Nevertheless, we stress that the average component $z_m$ of $z$ may become unbounded when $\nu_m$ is not integral-average bounded even if $|(x^t, z^t)|_{A^*} \rightarrow 0$ and $\nu_m$ is small. Indeed, $z_m$ is Lyapunov stable when $\nu_m = 0$ but not strongly stable when $\nu_m \neq 0$. It is, however, integral- ISS [23] (see the proof of Theorem 1).

Regarding the asymptotic gain property, we underline that, as shown in Section IV, the gain $\rho_\gamma$ is $O(\gamma^{-1})$ and $\tau_\gamma$ is $O(\gamma^{-1/2})$. Hence, the effect of $w_m$ and $(w_\perp, \nu_\perp)$ is, in general, amplified by taking smaller values of $\gamma$ ($\nu_m$, instead, is unaffected by $\gamma$). Nevertheless, in the relevant case where $\nu = 0$ and $w$ represents uncertainty in the computation of the gradients $\Phi(x)$, we have $w = \gamma w'$ for some $w'$, as this gives the term $-\gamma(\Phi(x) + w')$ in (7a). In this case, the gain from $w'$ to $|(x^t, z^t)|_{A^*}$ is $O(1)$.

Finally, we remark that the proof of Theorem 1 is based on a **time-scale separation**, enforced when $\gamma \ll 1$, between the average and the dispersion dynamics. In particular, the dynamics governing the consensus error is fast, while convergence of the average to the optimum is slow. In Section IV, these two dynamics are at first studied separately, and then interconnected (see Figure 1). It is interesting to notice that, while establishing stability of the average dynamics alone does put some constraints on $\gamma$, the condition $\gamma < 1$ that actually separates the time scales only arises when the two dynamics are interconnected.²

### III. CONNECTIONS WITH THE GRADIENT TRACKING

In the “canonical coordinates” formulation of [20], the Gradient Tracking algorithm employs a pair $(x_i, s_i) \in \mathbb{R}^2$ of variables for each agent $i \in \mathcal{N}$, whose (aggregate) update law reads as follows:

$$x^+ = Rx + z - \gamma \Phi(x),$$

$$z^+ = Cz - \gamma(C-I)\Phi(x),$$

$$(10a)$$

$$(10b)$$

where $R \in \mathbb{R}^{N \times N}$ (resp. $C \in \mathbb{R}^{N \times N}$) is a row (resp. column) stochastic matrix matching the communication network $(\mathcal{N}, \mathcal{E})$, i.e., $R_{ij} = 0$ (resp. $C_{ij} = 0$) if $(i, j) \notin \mathcal{E}$. Like algorithm (7), convergence to $\theta^*$ of the estimates $x_i$ produced by (11) is obtained, at an exponential rate, under Assumptions 1, and 2.

It is interesting to compare the Gradient Tracking (11) to algorithm (7) considered here. First, we notice that also in (7a) the matrix $I - K$ multiplying $x$ is row stochastic in view of (8). Indeed, it is doubly stochastic. Likewise, the identity matrix multiplying $z$ in (7b) is column stochastic, and the exogenous term $Kx$ sums to zero as $-\gamma(C-I)\Phi(x)$.

²In particular, $\gamma \leq \gamma^*$ in Section IV-D implies $\gamma \leq (2c_0)^{-1}$, and $c_\gamma \leq c_2 = \sqrt{c_1^2 + c_0^2} = \sqrt{c_0^2 + c_0^2}$. Thus, the change of coordinate $c_{i,j} = \sqrt{c_i^2 + c_j^2}$ is equivalent, and $\gamma \leq (2c_0)^{-1}$ implies $\gamma \leq 1/(2\sqrt{N})$.

System (11) differs from the original formulation of the Gradient Tracking (see, e.g., [9]-[17]) by a change of coordinates and it is therefore equivalent. Nevertheless, (11) is causal and has the advantage of not requiring the computation of $\nabla f_i$ for the initialization.
does in (11). Indeed, this implies that, like algorithm (7),
also the Gradient Tracking has the property that $z_m = 1^T z/N$ is constant along every solution. Hence, the need
of the initialization $1^T z^0 = 0$ in (11), which is the most
significant difference between (7) and (11). As clear from
the analysis in Section IV (see, in particular, Equations (13)
and (14)), a similar initialization is not required for (7)
because the uncontrolled dynamics $z_m$ is decoupled from
the other components of (7). We notice, indeed, that (8) implies
$K z = K S z_\perp$. Hence, $z_m$ is always filtered out in (7a).

As for what concerns robustness, we underline that the
unavoidable initialization and the coupling between $z_m$ and
the remaining states make the Gradient Tracking (11) fragile
if disturbances are added as in (7). Indeed, like in (7), the
uncontrolled dynamics $z_m$ of the Gradient Tracking can be
destabilized by means of a bounded and arbitrarily small
additive perturbation $\nu$. However, unlike (7), in the case of
the Gradient Tracking $z_m$ affects all the other state variables,
as clear from (11). Hence, in general, an ISS result as
that established by Theorem 1 cannot not hold for (11). A
counterexample in this direction is given in Section V.

Finally, we notice that, when $\nu = 0$, Equation (7b) takes
the form of an integrator processing the term $K x$. From (7b),
by using (8), we can derive the following equation for the
dispersion component $z_\perp$ of $z$

$$z_\perp^+ = z_\perp + S^T K (x - 1 x_m).$$  \hspace{1cm} (12)

Since $S^T K$ is full row rank, (12) is an integral action
processing the consensus error $x - 1 x_m$. Therefore, the Wang-
Elia algorithm can be seen as a distributed proportional-integral
(PI) controller (the proportional part being $(I - K) x - \gamma \Phi(x)$
and the integral part $K z = K S z_\perp$) regulating the “plant” $x^+ = u$ to the optimal equilibrium $1 \theta^*$.

Interestingly, it can be shown that the same distributed
PI structure is shared also by the Gradient Tracking algo-
rithm (11), where the integrator processes the term $(C - I)(R - I)(x - 1 x_m)$ and only shows up in the coordinates
$(x, z) \mapsto (x, (C - I) x - z)$. However, it is worth noticing
that, differently from the Gradient Tracking, the additional
dynamics $z_m$ never contributes to the PI controller in (7),
regardless of how $z$ is initialized. Nevertheless, it still plays a
crucial role since it enables the distributed implementation of
the integral action otherwise impossible. In fact, (12) cannot
be implemented in a distributed way since $S^T K$ does not
match the sparsity constraints imposed by the communication
structure.

IV. STABILITY ANALYSIS

In this section, we prove Theorem 1. For ease of exposi-
tion, the proof is split in four parts.

A. The Reduced Error Subsystem

Under Assumption 2, there exists a unique $\theta^* \in \mathbb{R}$
such that $(x, z) \in A^*$ if and only if $x = 1 \theta^*$ and $z \in
-\gamma S(S^T K S)^{-1} S^T \Phi(1 \theta^*) + \text{span} 1$. Thus, we can define
without ambiguity the equilibrium $(x^*, z^*)$ as

$$x^* := 1 \theta^*, \hspace{1cm} z^* := -\gamma S(S^T K S)^{-1} S^T \Phi(1 \theta^*),$$

and, with $T$ defined in (4), change variables in (7) as

$$(x, z) \mapsto (\xi, \zeta) = (T(x - x^*), T(z - z^*)).$$

These new variables represent the average-dispersion com-
ponents (Section I-C) of the errors $x - x^*$ and $z - z^*$.
Indeed, $\xi = (\xi_m, \xi_\perp)$ and $\zeta = (\zeta_m, \zeta_\perp)$, with $\xi_m = 1^T (x - x^*)/N = x_m - \theta^*, \xi_\perp = S^T (x - x^*) = S^T x = x_\perp,$
$\zeta_m = 1^T (z - z^*)/N = 1^T z/N = z_m$, and $\zeta_\perp = S^T (z - z^*)$.

In addition, we have

$$x = 1 (\xi_m + \theta^*) + S \xi_\perp, \hspace{1cm} z = 1 \zeta_m + S (\zeta_\perp + S^T z^*).$$

The previous change of variables leads to the “error system”

$$\xi^+ = (I - T K T^{-1}) \xi - T K T^{-1} \zeta - \gamma T^T \Phi(x) + T w$$

$$\zeta^+ = \zeta + T K T^{-1} \xi + T \nu$$

in which $\Phi(x) := \Phi(x) - \Phi(1 \theta^*)$. From (4) and (8), we get

$$TKT^{-1} = \begin{bmatrix} 1^T K1/N & 1^T KS/N \\ S^T K1 & S^T KS \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & S^T KS \end{bmatrix}.$$  \hspace{1cm} (13)

Hence, the error system can be expanded as follows

$$\xi_m^+ = \xi_m - \gamma 1^T \Phi(x)/N + w_m$$

$$\zeta_\perp^+ = A (\zeta_\perp, \zeta_\perp) - \gamma B \Phi(x) + (w_\perp, \nu_\perp)$$

$$\zeta_m^+ = \zeta_m + \nu_m,$$  \hspace{1cm} (14a, 14b, 14c)

in which

$$A := \begin{bmatrix} I - S^T KS & -S^T KS \\ S^T KS & I \end{bmatrix}, \hspace{1cm} B := \begin{bmatrix} S^T \\ 0 \end{bmatrix}.$$  \hspace{1cm} (15)

As clear from (14), the average component $z_m = \zeta_m$ of
$z$, which is marginally stable, is decoupled from the rest
of the system. Indeed, $\zeta_m$ is not influenced by any other component
of the state, nor it influences them. Moreover, under Assumption 2, $\zeta_m$ does not contribute to the distance
of $(x, z)$ to $A^*$, as indeed we have

$$||(x, z)||_{A^*} = \inf_{(a, b) \in A^*} ||(x - a, z - b)||$$

$$= \inf_{c \in \mathbb{R}} ||(x - x^*, z - z^* + 1c)||$$

$$= ||(\sqrt{N} \xi_m, \zeta_\perp)||.$$  \hspace{1cm} (16)

Therefore, we shall now drop Equation (14c) and focus
on (14a)-(14b), to which we refer as the “reduced error
subsystem”. In the forthcoming Sections IV-B and IV-C, we
analyze the two subsystems (14a) and (14b) separately, and
characterize their stability properties. Later in Section IV-D,
we study their interconnection.

B. The “Average” Subsystem $\xi_m$

Define the function $V_1(\xi_m) := \xi_m^2$. The increment $\Delta V_1 := V_1(\xi_m^2) - V_1(\xi_m^2)$ satisfies (here and in the following, we drop the time dependency when no confusion may arise)

$$\Delta V_1 = -2 \gamma N \xi_m^2 \Phi(x) \hspace{1cm} (17a)$$

$$+ 2 \left( \xi_m^2 - \frac{\gamma}{N} \Phi(x) \right) w_m + |w_m|^2 + \frac{\gamma^2}{N^2} |1^T \Phi(x)|^2.$$  \hspace{1cm} (17b)
In view of strong convexity in Assumption 2, we can write
\[
\xi_m \mathbf{1}^T \Phi(1(\xi_m + \theta^*)) = \xi_m \sum_{i \in \mathcal{N}} (\nabla f_i(\xi_m + \theta^*) - \nabla f_i(\theta^*)) \geq 2c_0 N|\xi_m|^2
\] (18)
for some \(c_0 > 0\). Moreover, in view of Assumption 1, \(\Phi\) is Lipschitz continuous. Let \(\ell\) be its Lipschitz constant. Then, by adding and subtracting \(\gamma \xi_m \mathbf{1}^T \Phi(1(\xi_m + \theta^*))\) to (17a), we obtain
\[
(17a) = -2\frac{\gamma}{N} \xi_m \mathbf{1}^T \left( \Phi(1(\xi_m + \theta^*)) + \Phi(x) - \Phi(1(\xi_m + \theta^*)) \right)
\leq -4c_0 \gamma |\xi_m|^2 + 2c_1 \gamma |\xi_m| \cdot |\xi|, (19)
\]
with \(c_1 := \ell / \sqrt{N}, c_2 := c_1^2 / c_0\), and where we used the Young’s inequality.
\[
2ab \leq a^2 b^2 / \epsilon
\] (20)
with \(a = |\xi_m|, b = |\xi|, \) and \(\epsilon = c_0 / c_1\). Similarly, we obtain
\[
(17b) \leq 2 \left( (1 + \gamma \ell)(|\xi_m| + \sqrt{N} |\xi|) |w_m| + |w_m|^2 + 2 \ell^2 \gamma^2 (|\xi_m|^2 + |\xi|)^2 / N \right)
\leq (c_0 \gamma^2 / c_1) |\xi_m|^2 + (c_2 \gamma + c_5 \gamma^2) |\xi| |w_m|^2
c_4 := c_3 \ell / 2 + 2c_2 \ell / N, c_3 := 2c_2 / N, c_5 := 1 + 2c_0 / c_1 + (2c_0 \gamma + \sqrt{N} \gamma^{-1})
\] (21)
in which \(c_3 := c_0 \ell / 2 + 2c_2 \ell / N\). Thus, (21) implies that the subsystem \(\xi_m\) is exponentially ISS relative to the origin and with respect to the inputs \(\xi_m\) and \(w_m\).

C. The “Dispersion” Subsystem \((\xi_\perp, \zeta_\perp)\)

We now turn the attention to system (14b). First, we establish that the matrix \(A\) in (15) is Schur, and hence that also (14b) is ISS. Let \(\lambda \in \sigma(A)\) and \(e = (e_1, e_2)\) a corresponding unitary eigenvector. Then
\[
\lambda = |\lambda| e^2 = e^T (\lambda e) = e^T Ae = 1 - e_1^T S^T K Se_1.
\] (23)
Next, we claim that \(e_1 \neq 0\) for every eigenvector of \(A\). Indeed, if \(e = (0, e_2)\) and \(\lambda \in \sigma(A)\), the equation \(Ae = \lambda e\) implies \((S^T KS)e_2 = 0\), which implies \(0 \in \sigma(S^T KS)\) and thus contradicts (8). Therefore, since (8) also implies that \(S^T KS\) is positive definite, we obtain from (23) that \(\lambda < 1\) for all \(\lambda \in \sigma(A)\). Finally, \(e_1^T S^T K Se_1 \leq \max \sigma(S^T KS)|e_1|^2 < 1\), which together with (23) implies \(\lambda > 0\) for all \(\lambda \in \sigma(A)\). Thus, \(\sigma(A) \subset (0, 1)\) and \(A\) is Schur.

Let \(\eta_\perp := (\xi_\perp, \zeta_\perp)\) and \(\delta_\perp := (\nu_\perp, \nu_\perp)\). Define \(V_2(\eta_\perp) := \eta_\perp^T P \eta_\perp\), with \(P = P^T > 0\) being the unique solution to the Lyapunov equation \(A^T PA - P = -3I\). The increment \(\Delta V_2 := V_2(\eta_\perp^{t+1}) - V_2(\eta_\perp^t)\) satisfies
\[
\Delta V_2 = -2|\eta_\perp|^2 - 2(\lambda |\eta_\perp| + \delta_\perp)^T PB \Phi(x) + (2A|\eta_\perp| + \delta_\perp)^T P \delta_\perp + \gamma^2 \Phi(x)^T B^T PB \Phi(x).
\] (24a)
By using (20) twice with \(a = |\xi_\perp|, b = \delta_\perp, \epsilon = c_0 / |A|PB/\ell (\sqrt{N})\) and \(a = |\xi_\perp|, b = |\delta_\perp|, \epsilon = c_0 / |A|\), we obtain
\[
(24a) \leq (c_0 \gamma - 3)|\eta_\perp|^2 + c_0 \gamma^2 |\xi_\perp|^2 + c_0 \gamma |\delta_\perp|^2,
\]
with \(c_8 := |A|PB/\ell^2 N/c_0 + 2|A|PB/\ell |\|PB\| \ell + |P|\|\ell\|\) and \(c_9 := |A|PB/\ell^2 N/c_0 + |P|\|\ell\|\). Similarly, by using (20) with \(a = |\eta_\perp|, b = |\delta_\perp|, \epsilon = c_0 / (2A|\ell|)\), we obtain
\[
(24b) \leq (1/2 + c_0 |\gamma|^2)|\eta_\perp|^2 + c_1 |\delta_\perp|^2 + c_2 |\xi_\perp|^2,
\]
with \(c_10 := 2|B|PB/\ell^2, c_11 := 2|A|P^2/\ell + |P|\|\ell\|\) and \(c_12 := 2|B|PB/\ell^2 N\). Pick \(\gamma > 0\) such that
\[
|\gamma | < \gamma^*_1 := \min \{\gamma^*_1, (4c_10)^{-1/2}, (4c_8)^{-1}, c_0 (2c_12)^{-1}\}.
\]
Then, with \(c_13(\gamma) := c_{11} + c_9 \gamma, \) it holds
\[
\Delta V_2 \leq -2|\eta_\perp|^2 + c_0 \gamma |\xi_\perp|^2 + c_13(\gamma) |\delta_\perp|^2.
\] (25)
Similarly to (22), inequality (25) establishes ISS of the dispersion subsystem with respect to the average error \(\xi_m\) and the dispersion component of the disturbances.

D. The Interconnection Between \(\xi_m\) and \((\xi_\perp, \zeta_\perp)\)

A block diagram representing the interconnection between (14a) and (14b) is represented in Figure 1, underlining the time-scale separation in the overall dynamics. Let \(\lambda, \lambda > 0\) denote, respectively, the smallest and largest eigenvalues of \(P\). Pick \(\gamma \in (0, \gamma^*)\), where
\[
\gamma^* = \min \{\gamma^*_1, (2c_7)^{-1}, (2c_8)^{-1/2}, (c_0 \lambda)^{-1}, c_0^{-1}\}.
\]
Define the function \(V(\xi_m, \eta_\perp) := V_1(\xi_m) + V_2(\eta_\perp)\) and \(V(\xi_m, \eta_\perp) := |\xi_m|^2 + \eta_\perp^T P \eta_\perp\). Then, in view of (22) and (25), the increment \(\Delta V^t := V(\xi_m^{t+1}, \eta_\perp^{t+1}) - V(\xi_m^t, \eta_\perp^t)\) satisfies
\[
\Delta V \leq -c_0 |\xi_m|^2 + c_0 |\xi_\perp|^2 + c_0 \gamma |\delta_\perp|^2.
\] (26)
Equation (26) shows that the reduced error system is ISS with respect to the disturbances \(w\) and \(\nu_\perp\).

Then, notice that \(|\xi_m|^2 + c_0 |\xi_\perp|^2 \leq V(\xi_m, \eta_\perp) \leq |\xi_m|^2 + \lambda |\eta_\perp|^2\) and, thus, by iterating (26), one obtains
\[
V(\xi_m^t, \eta_\perp^t) \leq c_0 V(\xi_m^{t-1}, \eta_\perp^{t-1}) + c_14(\gamma)|\xi_m^{t-1}|^2 + c_15(\gamma)|\delta_\perp^{t-1}|^2.
\]
where $q_γ := 1 - \min\{c_0 γ, \tilde{λ}^{-1}\} = 1 - c_0 γ \in [0, 1)$, $c_{14}(γ) := c_6(γ)/(c_0 γ)$ and $c_{15}(γ) := c_{13}(γ)/(c_0 γ)$.

Since, in view of (16), $(\max N, \tilde{λ}^{-1})^{-1}(x, z)_{2A}^2 \leq V(\tilde{σ}_m, \nu_λ) \leq \max\{N^{-1}, \tilde{λ}\}(x, z)_{2A}^2$, we finally obtain the sought inequality (10) by setting $\alpha = \sqrt{\max\{N^{-1}, \tilde{λ}\} \max\{N, \tilde{λ}^{-1}\}}$, $\mu_γ = \sqrt{q_γ}$, $ρ_γ = \sqrt{c_{14}(γ) \max\{N, \tilde{λ}^{-1}\}}$, $τ_γ = \sqrt{c_{15}(γ) \max\{N, \tilde{λ}^{-1}\}}$.

Finally, the boundedness claims directly follow from (10) and (14c), respectively.

V. ILLUSTRATIVE EXAMPLE

We present a toy example showing the fragility of the Gradient Tracking. We consider (11) for $N = 2$ agents, with $C = R$, $R_{11} = R_{22} = 0.8$, $R_{12} = R_{21} = 0.2$, $f_1(θ) := (θ - 1)^2$ and $f_2(θ) := (θ - 4)^2$. We modify (11b) to

$$z^+ = C Q(z) - γ(C - I)Φ(x),$$

(27)

where $Q(z) := 10^{-5} \cdot |z| \cdot 10^5$ models a quantization effect (|$\cdot$| is the componentwise floor function). We can look at (27) as the original (11b) subject to the quantization error $ν = C(Q(z) - z)$, which satisfies $ν_m = 1^T ν/N \leq 0$.

Figure 2 shows four simulations obtained with stepsize $γ = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ and with the same initial condition $(x^0, z^0) = (0, 0)$. As discussed in Section III, the Gradient Tracking is not ISS. Indeed, the average quantization error $ν_m$ destabilizes the state $x$. We stress that the smaller the $γ$ the higher is the divergence rate. This is explained by the same arguments given in Section II-D.

Figure 2 also shows the solutions of algorithm (7) (for the same values of $γ$ and with the same initial condition), in which (7b) is modified to $z^+ = Q(z) + K x$. Consistently with Theorem 1, we observe that the estimation error has a stable behavior, despite a small steady-state error.

VI. CONCLUSIONS

We studied a perturbed version of the Wang-Elia algorithm, and we proved exponential ISS relative to an optimal steady state. We compared the algorithm to the Gradient Tracking, showing that the latter does not enjoy a similar ISS property due to the need of initialization. Overall, our arguments underline the impact that a control theoretic approach can have on the analysis of (distributed) optimization, especially when robustness is taken into account.

REFERENCES

[1] A. Nedić and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” IEEE Trans. Autom. Control, vol. 54, no. 1, pp. 48–61, 2009.
[2] A. Nedić, A. Ozdaglar, and P. A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” IEEE Trans. Autom. Control, vol. 55, no. 4, pp. 922–938, 2010.
[3] J. Wang and N. Elia, “Control approach to distributed optimization,” in IEEE 48th Annu. Allerton Conf. Commun. Control Comput., 2010, pp. 557–561.
[4] D. K. Molzahn, F. Dörfler, H. Sandberg, S. H. Low, S. Chakrabarti, R. Baldick, and J. Lavaei, “A survey of distributed optimization and control algorithms for electric power systems,” IEEE Trans. Smart Grid, vol. 8, no. 6, pp. 2941–2962, 2017.
[5] A. Nedić, A. Olshlevsky, and M. G. Rabbat, “Network topology and communication-computation tradeoffs in decentralized optimization,” Proc. IEEE, vol. 106, no. 5, pp. 953–976, 2018.
[6] A. Nedić and J. Liu, “Distributed optimization for control,” Ann. Rev. Control Robot. Auton. Syst., vol. 1, pp. 77–103, 2018.
[7] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, “A survey of distributed optimization,” Ann. Rev. Control, vol. 47, pp. 278–305, 2019.
[8] G. Notarstefano, I. Notarnicola, and A. Camisa, “Distributed optimization for smart cyber-physical networks,” Found. Trends Syst. Control, vol. 7, no. 3, pp. 253–383, 2019.
[9] D. Varagnolo, F. Zanella, A. Cenedese, G. Pilonetto, and L. Schenato, “Newton-Raphson consensus for distributed convex optimization,” IEEE Trans. Autom. Control, vol. 61, no. 4, pp. 994–1009, 2016.
[10] P. Di Lorenzo and G. Scutari, “Next: In-network nonconvex optimization,” IEEE Trans. Signal Inf. Process. Netw., vol. 2, no. 2, pp. 120–136, 2016.
[11] A. Nedić, A. Olshlevsky, and W. Shi, “Achieving geometric convergence for distributed optimization over time-varying graphs,” SIAM J. Optim., vol. 27, no. 4, pp. 2597–2633, 2017.
[12] G. Qu and N. Li, “Harnessing smoothness to accelerate distributed optimization,” IEEE Trans. Control Netw. Syst., vol. 5, no. 3, pp. 1245–1260, 2018.
[13] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, “Convergence of asynchronous distributed gradient methods over stochastic networks,” IEEE Trans. Autom. Control, vol. 63, no. 2, pp. 434–448, 2018.
[14] C. Xi, R. Xin, and U. A. Khan, “ADD-OPT: Accelerated distributed directed optimization,” IEEE Trans. Autom. Control, vol. 63, no. 5, pp. 1329–1339, 2018.
[15] R. Xin and U. A. Khan, “A linear algorithm for optimization over directed graphs with geometric convergence,” IEEE Contr. Syst. Lett., vol. 2, no. 3, pp. 315–320, 2018.
[16] S. Pu, W. Shi, J. Xu, and A. Nedic, “Push-pull gradient methods for distributed optimization in networks,” IEEE Trans. Autom. Control, vol. 66, no. 1, pp. 1–16, 2021.
[17] Y. Tian, Y. Sun, and G. Scutari, “Achieving linear convergence in distributed asynchronous multiagent optimization,” IEEE Trans. Autom. Control, vol. 65, no. 12, pp. 5264–5279, 2020.
[18] G. Qu and N. Li, “Accelerated distributed Nesterov gradient descent,” Proc. IEEE Autom. Control, vol. 65, no. 6, pp. 2566–2581, 2019.
[19] R. Xin and U. A. Khan, “Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking,” IEEE Trans. Autom. Control, vol. 65, no. 6, pp. 2627–2633, 2019.
[20] M. Bin, I. Notarnicola, L. Marconi, and G. Notarstefano, “A system theoretical perspective to gradient-tracking algorithms for distributed quadratic optimization,” in IEEE 58th Conf. Decis. Control, 2019, pp. 2994–2999.
[21] T. Hatanaka, N. Chopra, T. Ishizaki, and N. Li, “Passivity-Based Distributed Optimization With Communication Delays Using PI Consensus Algorithm,” IEEE Trans. Autom. Control, vol. 63, no. 12, pp. 4421–4428, 2018.
[22] F. Sontag, “Smooth stabilization implies coprime factorization,” IEEE Trans. Autom. Control, vol. 34, no. 4, pp. 435–443, 1989.
[23] D. Angeli, E. Sontag, and Y. Wang, “A characterization of integral input-to-state stability,” IEEE Trans. Autom. Control, vol. 45, no. 6, pp. 1082–1097, 2000.