SECOND-ORDER GUARANTEES OF DISTRIBUTED GRADIENT ALGORITHMS*

AMIR DANESHMAND†, GESUALDO SCUTARI‡, AND VYACHESLAV KUNGURTSEV §

Abstract. We consider distributed smooth nonconvex unconstrained optimization over networks, modeled as a connected graph. We examine the behavior of distributed gradient-based algorithms near strict saddle points. Specifically, we establish that (i) the renowned Distributed Gradient Descent (DGD) algorithm likely converges to a *neighborhood* of a Second-order Stationary (SoS) solution; and (ii) the more recent class of distributed algorithms based on gradient tracking–implementable also over digraphs–likely converges to *exact* SoS solutions, thus avoiding (strict) saddle-points. Furthermore, a convergence rate is provided for the latter class of algorithms.

Key words. Distributed Gradient Methods, Gradient Tracking, Nonconvex Optimization.

AMS subject classifications. 68Q25, 68R10, 68U05

1. Introduction. We consider smooth unconstrained nonconvex optimization over networks, in the following form:

$$\text{(P)} \qquad \min_{\boldsymbol{\theta} \in \mathbb{R}^m} F(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{n} f_i(\boldsymbol{\theta}),$$

where n is the number of agents in the network; and $f_i : \mathbb{R}^m \to \mathbb{R}$ is the cost function of agent i, assumed to be smooth and known only to agent i. Agents are connected through a communication network, modeled as a (possibly directed, strongly) connected graph. No specific topology is assumed for the graph (such as star or hierarchical structure). In this setting, agents seek to cooperatively solve Problem (P) by exchanging information with their immediate neighbors in the network.

Distributed *nonconvex* optimization in the form (P) has found a wide range of applications in several areas, including network information processing, machine learning, communications, and multi-agent control; see, e.g., [27]. For instance, this is the typical scenario of in-network data-intensive (e.g., sensor-network) applications wherein data are scattered across the agents (e.g., sensors, clouds, robots), and the sheer volume and spatial/temporal disparity of data render centralized processing and storage infeasible or inefficient. Communication networks modeled as *directed* graphs capture simplex communications between adjacent nodes. This is the case, e.g., in several wireless (sensor) networks wherein nodes transmit at different power and/or communication channels are not symmetric.

Main objective: We call $\boldsymbol{\theta}$ a critical point of F if $\nabla F(\boldsymbol{\theta}) = \mathbf{0}$; a critical point $\boldsymbol{\theta}$ is a *strict saddle* of F if $\nabla^2 F(\boldsymbol{\theta})$ has at least one negative eigenvalue; and it is a *Second-order Stationary* (SoS) solution if $\nabla^2 F(\boldsymbol{\theta})$ is positive semidefinite. Critical points that are not minimizers are of little interest in the nonconvex setting. It is thus desirable to consider methods for (P) that are not attracted to such points. When

†School of Industrial Engineering, Purdue University, West-Lafayette, IN, USA (adaneshm@purdue.edu, http://web.ics.purdue.edu/~adaneshm/).

‡School of Industrial Engineering, Purdue University, West-Lafayette, IN, USA (gscutari@purdue.edu, https://engineering.purdue.edu/~gscutari/).

§Czech Technical University in Prague, Department of Computer Science, Faculty of Electrical Engineering, Prague, Czech Republic (vyacheslav.kungurtsev@fel.cvut.cz).

F has a favorable structure, stronger guarantees can be claimed. For instance, a wide range of salient objective functions arising from applications in machine learning and signal processing have been shown to enjoy the so-called *strict saddle* property: all the critical points of F are either strict saddles or local minimizers. Examples include principal component analysis and fourth order tensor factorization [8], low-rank matrix completion [9], and some instances of neural networks [13], just to name a few. In all these cases, converging to SoS solutions–and thus circumventing strict saddles–guarantees finding a local minimizer. The goal of this paper is to study second-order guarantees of existing decentralizations of the gradient algorithm for Problem (P) over undirected and directed graphs.

Literature review. Second-order guarantees of *centralized* gradient-based methods have been extensively studied in the literature, and briefly summarized next. Early works (e.g., [24]) showed that the gradient descent algorithm escapes strict saddle points, provided that the direction is perturbed by unbiased noise. More recently, [12] derived the convergence rate of the noisy gradient algorithm converging to SoS solutions. In [18], it was proved that running the gradient descent algorithm with a random initialization is sufficient to escape strict saddles. The elegant analysis of [18], based on tools from topology of dynamical systems, has been later extended in [17] to establish second-order guarantees of a variety of first-order methods. Finally, [23] studied the behavior of some momentum-based gradient methods near strict saddle points. However, all these schemes are *centralized*.

A natural question is whether *distributed* instantiations of the gradient descent algorithm over (di-)graphs enjoy similar second-order guarantees. This paper provides a first answer to this open question. In fact, there is a rich convergence theory of distributed algorithms for convex optimization, but little is known in the nonconvex case, let alone second-order guarantees. The Distributed Gradient Descent (DGD) is among the first attempts to decentralize the gradient algorithm [21, 22]. Roughly speaking, in the DGD (and subsequent variants), the update of each agent i is a linear combination of two components: i) the gradient ∇f_i evaluated at the latest agent's iterate (recall that agents do not have access to the entire gradient ∇F); and ii) a convex combination of the current iterates of the neighbors of agent i (including agent i itself). The latter term (a.k.a. consensus step) is instrumental to enforce asymptotically an agreement among the agents' local variables. When (P) is convex, convergence of the DGD algorithm is fully understood. With a diminishing step-size, agents' iterates converge to a consensual *exact* solution; if a constant step-size is used, convergence is generally faster but only to a neighborhood of the solution, and exact consensus is not achieved. When (P) is nonconvex, little is known about the convergence of the DGD algorithm. Specifically, [34] showed that, if a constant step-size is used, agents' iterates converge to a critical point of an auxiliary function [the Lyapunov function used to prove convergence–see (4.1) in Sec. 4] while reaching approximate consensus (see Theorem 4.1 in Sec. 4.1 for a formal statement of these results). Exact consensus can be achieved using a diminishing step-size. However, nothing is known about the connection of the critical points of the aforementioned auxiliary function and the critical points of F, let alone second-order guarantees.

The extension of the DGD algorithm to *digraphs* was proposed in [19] for convex unconstrained optimization, and later extended in [31] to nonconvex objectives. The algorithm, which combines a local gradient step with the push-sum scheme [5], converges to an exact stationary solution of (P), when a diminishing step-size is employed; and its noisy perturbed version almost surely converges to local minimizers, provided that F does not have any saddle point [31].

87   To cope with the speed-accuracy dilemma of DGD, [6, 7] proposed a new class
88 of distributed gradient-based methods that converge to an *exact* consensual solution
89 of nonconvex (constrained) problems while using a *fixed step-size*. The algorithmic
90 framework, termed NEXT, introduces the idea of *gradient tracking* to correct the
91 DGD direction and cancel the steady state error in it while using a fixed step-size:
92 each agent updates its own local variables along a surrogate direction that tracks
93 the gradient ∇F of the entire objective (the same idea was proposed independently
94 in [33] for convex unconstrained smooth problems). The generalization of NEXT
95 to digraphs–the SONATA algorithm–was proposed in [30, 27, 26, 29], with [26, 29]
96 proving convergence of the agents' iterates to consensual stationary solutions of the
97 nonconvex problem at a sublinear rate. Extensions of the SONATA family based on
98 different choices of the weight matrices were later introduced in [32, 25] for convex
99 smooth unconstrained problems. Since some schemes contain others as special cases
100 (see, e.g., [26, Sec. 5]), hereafter we will refer collectively to this family of algorithms
101 as Distributed Optimization with Gradient Tracking (DOGT) algorithms.
102   **Summary of the contributions:** We study second-order guarantees of the
103 available decentralizations of the plain gradient algorithm, namely: the DGD algo-
104 rithm (over undirected graphs) and the DOGT schemes (over undirected and directed
105 graphs). For DGD employing a constant step-size, we establish the following:
106 **(i)** Convergence of the iterates to a neighborhood of critical points of F, for all
107   initializations, is proved; this complements the convergence results in [34];
108 **(ii)** For sufficiently small step-sizes, the critical points in (i) are almost surely SoS
109   solutions of (P), where the probability is taken over the initialization.
110 For the DOGT algorithms using a constant step-size, our results are the following:
111 **(i)** Convergence to an *exact* stationary solution of (P) over digraphs at a sublin-
112   ear rate is proved. The analysis is based on a Lyapunov-like function that
113   properly combines average dynamics, consensus and tracking disagreements;
114 **(ii)** If F is a Kurdyka-Łojasiewicz (KŁ) function [16, 15], *global* convergence is estab-
115   lished (i.e., for an arbitrary starting point, the algorithm generates a sequence
116   that converges to a critical point of F);
117 **(iii)** Convergence to *exact* SoS solutions of (P) over undirected graphs and over di-
118   graphs when $m = 1$ is proved, for almost all initializations, drawn from a
119   suitably chosen subspace.
120   To our knowledge these are the first guarantees proved for distributed *gradient*
121 algorithms over (undirected and directed) graphs. We notice that, recently, [10] stud-
122 ied second-order guarantees of a primal-dual method applied to linearly constrained
123 nonconvex problems. The scheme can be customized to Problem (P) if the graph
124 is *undirected*, and thus convergence to SoS solutions follows (under a suitably chosen
125 initialization, not discussed in [10]). Results in [10] neither extend to the decentralized
126 gradient algorithms studied in this paper nor to digraphs.
127   **1.1. Notation.** All vectors are denoted by bold letters and assumed to be col-
128 umn vectors; the tuple $\mathbf{x} = (\mathbf{x}_i)_{i=1}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denotes a column vector whose i-th
129 (column) block component is $\mathbf{x}_i$; $\mathcal{V}_{\mathbf{x}}$ and $\mathcal{B}(\mathbf{x}, r)$ denote a neighborhood of $\mathbf{x}$ and the
130 closed ball of radius $r > 0$ centered at $\mathbf{x}$, respectively; $\mathbf{x}$ is called *stochastic* if all its
131 components are non-negative and sum to one; and $\mathbf{1}$ is the vector of all ones (we write
132 $\mathbf{1}_m$ for the m–dimensional vector, if the dimension is not clear from the context).
133 Given $\mathcal{X} \subseteq \mathbb{R}^m$, $\overline{\mathcal{X}}$ denotes the complement of $\mathcal{X}$; we use the shorthand $[I]$ for the set
134 $\{1, 2, \ldots, I\}$. The set of nonnegative integers is denoted by $\mathbb{N}_+$. The Euclidean projec-
135 tion of $\mathbf{x} \in \mathbb{R}^m$ onto the convex closed set $\mathcal{X} \subseteq \mathbb{R}^m$ is $\mathrm{proj}_{\mathcal{X}}(\mathbf{x}) \triangleq \arg\min_{\mathbf{y} \in \mathcal{X}} ||\mathbf{x} - \mathbf{y}||$.
136   Matrices are denoted by capital bold letters; the (i, j)-th element of $\mathbf{A}$ is denoted

by A_{ij}; $\mathcal{M}_m(\mathbb{R})$ is the set of all $m \times m$ real matrices; $\mathbf{I}$ is the identity matrix (if the dimension is not clear from the context, we write $\mathbf{I}_m$ for the $m \times m$ identity matrix); $\mathbf{A} \geq 0$ denotes a nonnegative matrix, that is, a matrix with all the entries being nonnegative numbers; and $\mathbf{A} \geq \mathbf{B}$ stands for $\mathbf{A} - \mathbf{B} \geq 0$. The spectrum of a square real matrix $\mathbf{M}$ is denoted by $\mathrm{spec}(\mathbf{M})$; its spectral radius is $\mathrm{spradii}(\mathbf{M}) \triangleq \max\{|\lambda| : \lambda \in \mathrm{spec}(\mathbf{M})\}$; and its minimum singular value and minimum eigen value are denoted by $\sigma_{min}(\mathbf{M})$ and $\lambda_{min}(\mathbf{M})$, respectively. With a slight abuse of notation, we will use the same symbol $||\cdot||$ to denote vector norms in $\mathbb{R}^n$ and their induced matrix norms.

1.2. Paper organization. The rest of the paper is organized as follows. The main assumptions on the optimization problem and network setting are introduced in Sec. 2 along with the description of the DGD algorithm (cf. Sec. 2.2) and DOGT algorithms (cf. Sec. 2.3). Convergence of DOGT algorithms is studied in Sec. 3 along the following steps: i) Sub-sequence convergence is proved in Sec. 3.1; Sec. 3.2 establishes global convergence under the KL property of F; and Sec. 3.3 derives second-order guarantees over undirected and directed graphs. Sec. 4 studies guarantees of the DGD algorithm over undirected graphs, along the following steps: i) existing convergence results are discussed in Sec. 4.1; ii) Sec. 4.2 studies convergence to a neighborhood of a critical point of F; and iii) Sec. 4.3 establishes second-order guarantees. Finally, Sec. 5 presents some preliminary numerical experiments.

2. Problem/network setting and distributed algorithms.

2.1. Problem/network setting. We study Problem (P) under the following blanket assumptions.

Assumption 2.1 (On Problem P). Given Problem (P), suppose that
(i) Each $f_i : \mathbb{R}^m \to \mathbb{R}$ is $r + 1$ times continuously differentiable, for some integer $r \geq 1$, and ∇f_i is L_i-Lipschitz continuous;
(ii) F is bounded from below.

We also tacitly assume that agent i knows only its own function f_i. The set of critical (a.k.a. stationary) points of F is denoted by crit $F \triangleq \{\boldsymbol{\theta} : \nabla F(\boldsymbol{\theta}) = \mathbf{0}\}$.

Network model: The network of the n agents is modeled as a (possibly) directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices $\mathcal{V} = [n]$ coincides with the set of agents, and the set of edges $\mathcal{E}$ represents the agents' communication links: $(i, j) \in \mathcal{E}$ if and only if there is link directed from agent i to agent j. The in-neighborhood of agent i is defined as $\mathcal{N}_i^{\mathrm{in}} = \{j | (j, i) \in \mathcal{E}\} \cup \{i\}$ and represents the set of agents that can send information to agent i (including agent i itself, for notational simplicity). The out-neighborhood of agent i is similarly defined $\mathcal{N}_i^{\mathrm{out}} = \{j | (i, j) \in \mathcal{E}\} \cup \{i\}$. When the graph is undirected, these two sets coincide and we use $\mathcal{N}_i$ to denote the neighborhood of agent i (with a slight abuse of notation, we use the same symbol $\mathcal{G}$ to denote either directed or undirected graphs). Given a nonnegative matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, the directed graph induced by $\mathbf{A}$ is defined as $\mathcal{G}_A = (\mathcal{V}_A, \mathcal{E}_A)$, where $\mathcal{V}_A \triangleq [n]$ and $(j, i) \in \mathcal{E}_A$ if and only if $A_{ij} > 0$. The set of roots of all the directed spanning trees in $\mathcal{G}_A$ is denoted by $\mathcal{R}_A$. We make the following blanket standard assumptions on $\mathcal{G}$.

Assumption 2.2 (On the network). The graph (resp. digraph) $\mathcal{G}$ is connected (resp. strongly connected).

2.2. The DGD algorithm. Consider Problem (P) and assume that the network is modeled as an undirected graph $\mathcal{G}$. The DGD algorithm is based on the following decentralization of the gradient algorithm. Each agent i holds a copy $\mathbf{x}_i$ of the optimization variable $\boldsymbol{\theta}$, which is updated iteratively; its value at iteration ν is denoted by

$\mathbf{x}_i^\nu$. All the local copies are stacked column-wise in the vector $\mathbf{x} \triangleq (\mathbf{x}_i)_{i=1}^n$; similarly, we write $\mathbf{x}^\nu \triangleq (\mathbf{x}_i^\nu)_{i=1}^n$. We define the *aggregate* function $F_c(\mathbf{x}) \triangleq \sum_{i=1}^n f_i(\mathbf{x}_i)$. The update of the DGD algorithm reads: given $\mathbf{x}^\nu$,

$$(2.1) \qquad \mathbf{x}^{\nu+1} = \mathbf{W}_D\,\mathbf{x}^\nu - \alpha\nabla F_c(\mathbf{x}^\nu),$$

where $\alpha > 0$ is a step-size; $\mathbf{W}_D \triangleq \mathbf{D} \otimes \mathbf{I}_m$, and $\mathbf{D} \in \mathcal{M}_n(\mathbb{R})$ is a symmetric doubly-stochastic matrix compliant to the graph $\mathcal{G}$, that is, $D_{ij} > 0$ iff $(j,i) \in \mathcal{E}$; and $D_{ij} = 0$, otherwise. Breaking (2.1) into per-agent steps gives some insight on agents' updates: $\mathbf{x}_i^{\nu+1} = \sum_{j\in\mathcal{N}_i} D_{ij}\,\mathbf{x}_j^\nu - \alpha\nabla f_i(\mathbf{x}_i^\nu)$. Thus each agent is updating using only the gradient of its own local objective (linearly) combined with the mixing $\sum_{j\in\mathcal{N}_i} D_{ij}\,\mathbf{x}_j^\nu$, which is necessary for reaching consensus across the $\mathbf{x}_i$'s.

2.3. DOGT algorithms. In this class of algorithms, each agent i, in addition to its local optimization variable $\mathbf{x}_i$, also owns an auxiliary variable $\mathbf{y}_i \in \mathbb{R}^m$ that works as a local proxy of the sum-gradient $\sum_i \nabla f_i(\mathbf{x}_i^\nu)$, aiming thus at correcting the direction $-\nabla f_i(\mathbf{x}_i^\nu)$ (as used in the DGD algorithm). Let $\mathbf{y}_i^\nu$ be the value of $\mathbf{y}_i$ at iteration ν. The DOGT update reads:

$$\begin{cases} \mathbf{x}_i^{\nu+1} = \displaystyle\sum_{j\in\mathcal{N}_i^{\mathrm{in}}} R_{ij}\mathbf{x}_j^\nu \;-\alpha\,\mathbf{y}_i^\nu, \\[2mm] \mathbf{y}_i^{\nu+1} = \displaystyle\sum_{j\in\mathcal{N}_i^{\mathrm{in}}} C_{ij}\mathbf{y}_j^\nu + \nabla f_i\big(\mathbf{x}_i^{\nu+1}\big) - \nabla f_i\big(\mathbf{x}_i^\nu\big), \quad \text{(Gradient Tracking)} \end{cases}$$

for all $i = 1,\ldots,n$, where each $\mathbf{x}_i^0 \in \mathbb{R}^m$ is arbitrarily chosen and $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0)$; $\mathbf{R} \triangleq (R_{ij})_{i,j=1}^n$ and $\mathbf{C} \triangleq (C_{ij})_{i,j=1}^n$ are suitably chosen non-negative weight matrices (cf. Assumption 2.3 below); and $\alpha > 0$ is the step-size. Note that the update of the y-variables along with the consensus mix $\sum_{j\in\mathcal{N}_i^{\mathrm{in}}} R_{ij}\mathbf{x}_j^\nu$ ensures the aforementioned distributed tracking of the sum-gradient ∇F.

Denoting $\mathbf{y}^\nu \triangleq (\mathbf{y}_i^\nu)_{i=1}^n$, $\mathbf{W}_R \triangleq \mathbf{R}\otimes\mathbf{I}_m$ and $\mathbf{W}_C \triangleq \mathbf{C}\otimes\mathbf{I}_m$, and using $\mathbf{x}^\nu \triangleq (\mathbf{x}_i^\nu)_{i=1}^n$ and function $F_c(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x}_i)$, the algorithm can be written in compact form as

$$(2.2) \qquad \begin{cases} \mathbf{x}^{\nu+1} = \mathbf{W}_R\,\mathbf{x}^\nu - \alpha\,\mathbf{y}^\nu, \\[1mm] \mathbf{y}^{\nu+1} = \mathbf{W}_C\,\mathbf{y}^\nu + \nabla F_c\big(\mathbf{x}^{\nu+1}\big) - \nabla F_c\big(\mathbf{x}^\nu\big). \end{cases}$$

Different choices for $\mathbf{R}$ and $\mathbf{C}$ are possible, resulting in different existing algorithms. For instance, if $\mathbf{R} = \mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ are doubly-stochastic matrices compliant to the graph $\mathcal{G}$, (2.2) reduces to the NEXT algorithm [6, 7] (or the one in [33], when (P) is convex). If $\mathbf{R}$ and $\mathbf{C}$ are allowed to be time-varying (suitably chosen) (2.2) reduces to the SONATA algorithm applicable to (possibly time-varying) digraphs [30, 27, 26, 29] [or the one later proposed in [20] for strongly convex instances of (P)]. Finally, if $\mathbf{R}$ and $\mathbf{C}$ are chosen according to Assumption 2.3 below, the scheme (2.2) becomes the algorithm proposed independently in [25] and [32], for strongly convex objectives in (P), and implementable over fixed digraphs.

Assumption 2.3. (On the matrices $\mathbf{R}$ and $\mathbf{C}$) The weight matrices $\mathbf{R}, \mathbf{C} \in \mathbb{R}^{n\times n}$ satisfy the following:
   (i) $\mathbf{R}$ is nonnegative row-stochastic and $R_{ii} > 0$, for all $i \in [n]$;
   (ii) $\mathbf{C}$ is nonnegative column-stochastic and $C_{ii} > 0$, for all $i \in [n]$;
   (iii) The graphs $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$ each contain at least one spanning tree; and $\mathcal{R}_R \cap \mathcal{R}_{C^\top} \neq \emptyset$.

It is not difficult to check that matrices $\mathbf{R}$ and $\mathbf{C}$ above exist if and only if the digraph $\mathcal{G}$ is strongly connected; however, $\mathcal{G}_R$ and $\mathcal{G}_{C^\top}$ need not be so. Several choices for such matrices are discussed in [25, 32]. Here, we only point out the following property of $\mathbf{R}$ and $\mathbf{C}$, as a consequence of Assumption 2.3, which will be used in our analysis.

LEMMA 2.4 ([25]). *Given $\mathbf{R}$ and $\mathbf{C}$ satisfying Assumption 2.3, there exist matrix norms $\|\cdot\|_R$ and $\|\cdot\|_C$ such that $\rho_R \triangleq \|\mathbf{R}-\mathbf{1}\mathbf{r}^\top\|_R < 1$ and $\rho_C \triangleq \|\mathbf{C}-\mathbf{c}\mathbf{1}^\top\|_C = \rho_C < 1$, where $\mathbf{r}$ (resp. $\mathbf{c}$) is the stochastic left-eigenvector (resp. right-eigenvector) of $\mathbf{R}$ (resp. $\mathbf{C}$) associated with the eigenvalue one. Furthermore, $\mathbf{r}^\top\mathbf{c} > 0$.*

Moreover, the matrix norms $||\cdot||_R$ and $||\cdot||_C$ have the following properties.

LEMMA 2.5. *Given $\mathbf{R}$ and $\mathbf{C}$ satisfying Assumption 2.3, the vector norms induced by the matrix norms $||\cdot||_R$ and $||\cdot||_C$ in Lemma 2.4 can be written as $||\mathbf{x}||_R = ||\mathbf{H}_R\mathbf{x}||_2$ and $||\mathbf{x}||_C = ||\mathbf{H}_C\mathbf{x}||_2$, respectively, where $\mathbf{H}_R \in \mathcal{M}_m(\mathbb{R})$ and $\mathbf{H}_C \in \mathcal{M}_m(\mathbb{R})$ are invertible matrices (dependent on $\mathbf{R}$ and $\mathbf{C}$, respectively). Furthermore, the vector-norm functions $||\cdot||_R^2$ and $||\cdot||_C^2$ are real-analytic.*

Proof. See Appendix A.1. □

3. DOGT Algorithms: convergence and second-order guarantees. Convergence of DOGT algorithms in the form (2.2) (with $\mathbf{R}$ and $\mathbf{C}$ satisfying Assumption 2.3) has not been studied in the literature when F is nonconvex. In this section we fill this gap and provide a full characterization of the convergence behavior of DOGT including its second order guarantees.

3.1. Subsequence convergence & rate analysis. We begin studying asymptotic convergence; we assume $m = 1$ (scalar optimization variables); while this simplifies the notation, all the conclusions hold for the general case $m > 1$. As in [25], define the weighted sums

$$(3.1) \qquad \bar{x}^\nu \triangleq \mathbf{r}^\top\mathbf{x}^\nu, \quad \bar{y}^\nu \triangleq \mathbf{1}^\top\mathbf{y}^\nu, \quad \text{and} \quad \bar{g}^\nu \triangleq \mathbf{1}^\top\nabla F_c(\mathbf{x}^\nu),$$

where we recall that $\mathbf{r}$ is the Perron vector associated with $\mathbf{R}$ (cf. Lemma 2.4). Define also $L_{\max} \triangleq \max_i L_i$, where L_i is the Lipschitz constant of ∇f_i.

Using (2.2), it is not difficult to check that the following holds

$$(3.2) \qquad \bar{x}^{\nu+1} = \bar{x}^\nu - \zeta\alpha\bar{y}^\nu - \alpha\mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right) \quad \text{and} \quad \bar{y}^\nu = \bar{g}^\nu,$$

where $\mathbf{c}$ is the Perron vector associated with $\mathbf{C}$, and $\zeta \triangleq \mathbf{r}^\top\mathbf{c} > 0$ (cf. Lemma 2.4).

3.1.1. Descent on F. Using the descent lemma along with (3.2) yields

$$
\begin{aligned}
F(\bar{x}^{\nu+1}) &= F\left(\bar{x}^\nu - \zeta\alpha\bar{y}^\nu - \alpha\mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right) \\
&\leq F(\bar{x}^\nu) - \zeta\alpha\left\langle\nabla F(\bar{x}^\nu), \bar{y}^\nu\right\rangle - \alpha\left\langle\nabla F(\bar{x}^\nu), \mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right\rangle \\
&\quad + \frac{L}{2}\left\|\zeta\alpha\bar{y}^\nu + \alpha\mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right\|^2.
\end{aligned}
$$

Adding/subtracting suitably chosen terms we obtain

$$
\begin{aligned}
F(\bar{x}^{\nu+1}) \leq &F(\bar{x}^\nu) - \zeta\alpha\left\langle\nabla F(\bar{x}^\nu) - \bar{y}^\nu, \bar{y}^\nu\right\rangle - \zeta\alpha|\bar{y}^\nu|^2 \\
&- \alpha\left\langle\nabla F(\bar{x}^\nu) - \bar{y}^\nu, \mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right\rangle - \alpha\left\langle\bar{y}^\nu, \mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right\rangle \\
&+ L\zeta^2\alpha^2|\bar{y}^\nu|^2 + L\alpha^2\left\|\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right\|^2
\end{aligned}
$$

257

$$
\leq F(\bar{x}^\nu) + \frac{\zeta\alpha}{2\epsilon_1}|\nabla F(\bar{x}^\nu) - \bar{y}^\nu|^2 + \frac{\zeta\alpha\epsilon_1}{2}|\bar{y}^\nu|^2 - \zeta\alpha|\bar{y}^\nu|^2
$$

$$
+ \frac{\alpha}{2}|\nabla F(\bar{x}^\nu) - \bar{y}^\nu|^2 + \frac{\alpha}{2}\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|^2 + \frac{\alpha\epsilon_2}{2}|\bar{y}^\nu|^2 + \frac{\alpha}{2\epsilon_2}\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|^2
$$

258 (3.3)
$$
+ L\zeta^2\alpha^2|\bar{y}^\nu|^2 + L\alpha^2\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|^2
$$

$$
= F(\bar{x}^\nu) + \left(\frac{\zeta\alpha\epsilon_1}{2} - \zeta\alpha + \frac{\alpha\epsilon_2}{2} + L\zeta^2\alpha^2 \right)|\bar{y}^\nu|^2
$$

$$
+ \left(\frac{\zeta\alpha}{2\epsilon_1} + \frac{\alpha}{2} \right)|\nabla F(\bar{x}^\nu) - \bar{y}^\nu|^2 + \left(\frac{\alpha}{2} + \frac{\alpha}{2\epsilon_2} + L\alpha^2 \right)\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|^2 ,
$$

259 where ϵ_1 and ϵ_2 are some arbitrary positive quantities. Noting that ∇F_c is L_c-Lipschitz
260 continuous, with $L_c \triangleq L_{max}$, and using $\bar{y}^\nu = \bar{g}^\nu$ [cf. (3.2)], we can write

261 (3.4)
$$
|\nabla F(\bar{x}^\nu) - \bar{y}^\nu| = \left| \sum_{i=1}^n \nabla f_i(x_i^\nu) - \sum_{i=1}^n \nabla f_i(\bar{x}^\nu) \right| \leq L_c \left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\| .
$$

262 Combining (3.3) and (3.4) yields

(3.5)
$$
F(\bar{x}^{\nu+1}) \leq F(\bar{x}^\nu) + \left(\frac{\zeta\alpha\epsilon_1}{2} - \zeta\alpha + \frac{\alpha\epsilon_2}{2} + L\zeta^2\alpha^2 \right)|\bar{y}^\nu|^2
$$

263
$$
+ L_c^2 K_{||}^2 \left(\frac{\zeta\alpha}{2\epsilon_1} + \frac{\alpha}{2} \right)\left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R^2 + K_{||}^2 \left(\frac{\alpha}{2} + \frac{\alpha}{2\epsilon_2} + L\alpha^2 \right)\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|_C^2 ,
$$

264 where $K_{||}$ is a positive constant such that $\left\| \mathbf{z} \right\|_a \leq K_{||} \left\| \mathbf{z} \right\|_b$, for all $\mathbf{z} \in \mathbb{R}^n$ and $a, b \in$
265 $\{2, R, C\}$. Note that such a constant exists, due to the equivalence of the norms (we
266 omit their specific expression).

267 **3.1.2. Bounding the consensus and gradient tracking errors.** Let us
268 bound first the consensus error $\left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R$. Using $\left\| \mathbf{z} + \mathbf{w} \right\|_R^2 \leq (1 + \epsilon)\left\| \mathbf{x} \right\|_R^2 +$
269 $(1 + 1/\epsilon)\left\| \mathbf{y} \right\|_R^2$, for arbitrary $\mathbf{z}, \mathbf{w} \in \mathbb{R}^m$ and $\epsilon > 0$, along with Lemma 2.4, yields

(3.6)
$$
\left\| \mathbf{x}^{\nu+1} - \mathbf{1}\bar{x}^{\nu+1} \right\|_R^2 = \left\| \left(\mathbf{R} - \mathbf{1}\mathbf{r}^\top \right)\left(\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right) - \alpha\left(\mathbf{I} - \mathbf{1}\mathbf{r}^\top \right)\left(\mathbf{y}^\nu - \mathbf{1}\bar{y}^\nu \right) \right\|_R^2
$$

$$
\leq (1 + \epsilon_x)\left\| \left(\mathbf{R} - \mathbf{1}\mathbf{r}^\top \right)\left(\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right) \right\|_R^2 + \alpha^2\left(1 + \frac{1}{\epsilon_x} \right)\left\| \left(\mathbf{I} - \mathbf{1}\mathbf{r}^\top \right)\left(\mathbf{y}^\nu - \mathbf{1}\bar{y}^\nu \right) \right\|_R^2
$$

270
$$
\leq \rho_R^2(1 + \epsilon_x)\left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R^2 + \alpha^2\left(1 + \frac{1}{\epsilon_x} \right)\left\| \mathbf{I} - \mathbf{1}\mathbf{r}^\top \right\|_R^2\left\| \mathbf{y}^\nu - \mathbf{1}\bar{y}^\nu \right\|_R^2
$$

$$
\leq \rho_R^2(1 + \epsilon_x)\left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R^2 + 2\alpha^2 K_1\left(1 + \frac{1}{\epsilon_x} \right)\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|_R^2
$$

$$
+ 2\alpha^2 K_1(1 + \frac{1}{\epsilon_x})\left\| (\mathbf{1} - \mathbf{c})\bar{y}^\nu \right\|_R^2
$$

$$
\leq \rho_R^2(1 + \epsilon_x)\left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R^2 + \alpha^2 K_2\left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|_C^2 + \alpha^2 K_3|\bar{y}^\nu|_2^2,
$$

271 for some positive constants K_1, K_2, K_3 (whose expression is omitted) and some $\epsilon_x > 0$.
272 Similarly, the tracking error can be bounded as

273
$$
\left\| \mathbf{y}^{\nu+1} - \mathbf{c}\bar{y}^{\nu+1} \right\|_C^2 = \left\| \left(\mathbf{C} - \mathbf{c}\mathbf{1}^\top \right)\mathbf{y}^\nu + \left(\mathbf{I} - \mathbf{c}\mathbf{1}^\top \right)\left(\nabla F_c(\mathbf{x}^{\nu+1}) - \nabla F_c(\mathbf{x}^\nu) \right) \right\|_C^2
$$

(3.7)

$$\leq (1 + \epsilon_y) \left\| \left(\mathbf{C} - \mathbf{c}\mathbf{1}^\top \right) \left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right) \right\|_C^2$$

$$+ (1 + \frac{1}{\epsilon_y}) \left\| \left(\mathbf{I} - \mathbf{c}\mathbf{1}^\top \right) \left(\nabla F_c(\mathbf{x}^{\nu+1}) - \nabla F_c(\mathbf{x}^\nu) \right) \right\|_C^2$$

$$\leq \rho_C^2 (1 + \epsilon_y) \left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|_C^2 + \left(1 + \frac{1}{\epsilon_y} \right) \left\| \mathbf{I} - \mathbf{c}\mathbf{1}^\top \right\|_C^2 \left\| \left(\nabla F_c(\mathbf{x}^{\nu+1}) - \nabla F_c(\mathbf{x}^\nu) \right) \right\|_C^2$$

$$\leq \left(\rho_C^2 + \frac{\alpha^2 K_4}{\epsilon_y} \right) (1 + \epsilon_y) \left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|_C^2 + \alpha^2 K_5 |\bar{y}^\nu|_2^2 + K_6 \left(1 + \frac{1}{\epsilon_y} \right) \left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R^2,$$

for some positive constants K_4, K_5, K_6 (whose expression is omitted) and some $\epsilon_y > 0$.

3.1.3. Lyapunov function. Let us introduce now the candidate Lyapunov function: denoting $\mathbf{J}_R \triangleq \mathbf{1}\mathbf{r}^\top$ and $\mathbf{J}_C \triangleq \mathbf{c}\mathbf{1}^\top$, let

(3.8) $$L(\mathbf{x}, \mathbf{y}) \triangleq F_c(\mathbf{J}_R \mathbf{x}) + \left\| (\mathbf{I} - \mathbf{J}_R)\mathbf{x} \right\|_R^2 + \varkappa \left\| (\mathbf{I} - \mathbf{J}_C)\mathbf{y} \right\|_C^2,$$

where $\varkappa > 0$ is a positive constant (to be properly chosen). Combining (3.5), (3.6), and (3.7) leads to the following descent property for L:

(3.9) $$L(\mathbf{x}^{\nu+1}, \mathbf{y}^{\nu+1}) \leq L(\mathbf{x}^\nu, \mathbf{y}^\nu) - (d^\nu)^2,$$

where

(3.10) $$d^\nu \triangleq \sqrt{(1 - \tilde{\rho}_R) \left\| \mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu \right\|_R^2 + \varkappa (1 - \tilde{\rho}_C) \left\| \mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu \right\|_C^2 + \Gamma |\bar{y}^\nu|_2^2},$$

and

$$\tilde{\rho}_R \triangleq \rho_R^2 (1 + \epsilon_x) + \frac{\alpha L_c^2 K_\|^2}{2} \left(1 + \frac{\zeta}{\epsilon_1} \right) + \varkappa K_6 \left(1 + \frac{1}{\epsilon_y} \right),$$

$$\tilde{\rho}_C \triangleq \rho_C^2 (1 + \epsilon_y) + \frac{\alpha K_\|^2}{2\varkappa} \left(1 + \frac{1}{\epsilon_2} \right) + \alpha^2 \left(\frac{L K_\|^2 + K_2}{\varkappa} + K_4 \left(1 + \frac{1}{\epsilon_y} \right) \right),$$

$$\Gamma \triangleq \left(\zeta - \frac{\epsilon_1 \zeta}{2} - \frac{\epsilon_2}{2} \right) \alpha - \left(L\zeta^2 + K_3 + K_5 \varkappa \right) \alpha^2.$$

To ensure $\tilde{\rho}_R < 1$, $\tilde{\rho}_C < 1$, and $\Gamma > 0$ in d^ν, we choose the free parameters ϵ_x, ϵ_y, ϵ_1, ϵ_2, and $\varkappa$ as follows:

(3.11)
$$0 < \epsilon_x < \frac{1 - \rho_R^2}{2\rho_R^2}, \qquad 0 < \epsilon_y < \frac{1 - \rho_C^2}{\rho_C^2},$$

$$\epsilon_1 = \epsilon_2 = \epsilon, \qquad 0 < \epsilon < \frac{2\zeta}{1 + \zeta}, \qquad 0 < \varkappa \leq \frac{\rho_R^2 \epsilon_x}{K_6(1 + 1/\epsilon_y)}.$$

Finally, $\alpha > 0$ satisfies

(3.12)
$$\alpha < \frac{2}{L_c^2 K_\|^2 \left(1 + \frac{\zeta}{\epsilon} \right)} \left(1 - \rho_R^2(1 + 2\epsilon_x) \right),$$

$$\alpha < \frac{1 - \rho_C^2(1 + \epsilon_y)}{\frac{1}{2\varkappa} K_\|^2 \left(1 + \frac{1}{\epsilon} + 2L \right) + \frac{K_2}{\varkappa} + K_4 \left(1 + \frac{1}{\epsilon_y} \right)},$$

$$\alpha < \frac{\zeta - \frac{\epsilon}{2} (\zeta + 1)}{L\zeta^2 + K_3 + K_5 \varkappa}.$$

The descent property (3.9) readily implies the following convergence result for $\{L(\mathbf{x}^\nu, \mathbf{y}^\nu)\}_\nu$ and $\{d^\nu\}_\nu$.

LEMMA 3.1. *In the setting above, there hold:*

(i) The sequence $\{L(\mathbf{x}^\nu, \mathbf{y}^\nu)\}_\nu$ converges;

(ii) $\sum_{\nu=0}^\infty (d^\nu)^2 < \infty$, and thus $\lim_{\nu \to \infty} d^\nu = 0$.

We end this subsection stating the following property of the Lyapunov function, which will be used later in our derivations.

LEMMA 3.2. *Let $\nabla L(\mathbf{x}^\nu, \mathbf{y}^\nu) \triangleq (\nabla_\mathbf{x} L(\mathbf{x}^\nu, \mathbf{y}^\nu), \nabla_\mathbf{y} L(\mathbf{x}^\nu, \mathbf{y}^\nu))$, where $\nabla_\mathbf{x} L$ (resp. $\nabla_\mathbf{y} L$) are the gradient of L with respect to the first (resp. second) argument. In the setting above, there holds*

(3.13) $$\|\nabla L(\mathbf{x}^\nu, \mathbf{y}^\nu)\| \leq M d^\nu, \quad \nu \geq 0,$$

for some finite $M > 0$.

Proof. Using Lemma 2.5, we can write

$$\begin{aligned} \nabla_\mathbf{x} L(\mathbf{x}^\nu, \mathbf{y}^\nu) &= \mathbf{J}_R^\top \nabla F_c(\mathbf{J}_R \mathbf{x}^\nu) + 2(\mathbf{I} - \mathbf{J}_R)^\top \mathbf{H}_R^\top \mathbf{H}_R (\mathbf{I} - \mathbf{J}_R) \mathbf{x}^\nu \\ &= \mathbf{r}\, \bar{y}^\nu + \mathbf{J}_R^\top (\nabla F_c(\mathbf{J}_R \mathbf{x}^\nu) - \nabla F_c(\mathbf{x}^\nu)) \\ &\quad + 2(\mathbf{I} - \mathbf{J}_R)^\top \mathbf{H}_R^\top \mathbf{H}_R (\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu), \\ \nabla_\mathbf{y} L(\mathbf{x}^\nu, \mathbf{y}^\nu) &= 2\varkappa (\mathbf{I} - \mathbf{J}_C)^\top \mathbf{H}_C^\top \mathbf{H}_C (\mathbf{I} - \mathbf{J}_C) \mathbf{y}^\nu \\ &= 2\varkappa (\mathbf{I} - \mathbf{J}_C)^\top \mathbf{H}_C^\top \mathbf{H}_C (\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu), \end{aligned}$$

(3.14)

where $\mathbf{H}_C$ and $\mathbf{H}_R$ are invertible matrices. Eq. (3.13) follows readily from (3.14) and the Lipschitz continuity of ∇F_c. $\square$

3.1.4. Main result. We can now state the main convergence result.

THEOREM 3.3. *Consider Problem (P), and suppose that Assumptions 2.1 and 2.2 are satisfied. Let $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}_\nu$ be the sequence generated by the DOGT Algorithm (2.2), with $\mathbf{R}$ and $\mathbf{C}$ satisfying Assumption 2.3, and α chosen according to (3.12); let $\{\bar{x}^\nu\}_\nu$ and $\{\bar{y}^\nu\}_\nu$ be defined in (3.1); and let $\{d^\nu\}_\nu$ be defined in (3.10). Given $\epsilon > 0$, let $T_\epsilon = \min\{\nu \in \mathbb{N}_+ : d^\nu \leq \epsilon\}$. Then, there hold*

(i) [consensus]: $\lim_{\nu \to \infty} \|\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu\| = 0$ and $\lim_{\nu \to \infty} \bar{y}^\nu = 0$;

(ii) [stationarity]: let $\mathbf{x}^\infty$ be a limit point of $\{\mathbf{x}^\nu\}_\nu$; then, $\mathbf{x}^\infty = \theta^\infty \mathbf{1}$, for some $\theta^\infty \in \operatorname{crit} F$;

(iii) [sublinear rate]: $T_\epsilon = \mathcal{O}(1/\epsilon^2)$.

Proof. (i) follows readily from Lemma 3.1(ii).

We prove (ii). Let $(\mathbf{x}^\infty, \mathbf{y}^\infty)$ be a limit point of $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}_\nu$. By (i), it must be $(\mathbf{I} - \mathbf{J}_R)\mathbf{x}^\infty = \mathbf{0}$, implying $\mathbf{x}^\infty = \mathbf{1}\theta^\infty$, for some $\theta^\infty \in \mathbb{R}$. Also, $\lim_{\nu \to \infty} \mathbf{1}^\top \nabla F_c(\mathbf{x}^\nu) = \lim_{\nu \to \infty} \bar{g}^\nu = \lim_{\nu \to \infty} \bar{y}^\nu = 0$, which together with the continuity of ∇F_c, yields $0 = \mathbf{1}^\top \nabla F_c(\mathbf{1}\theta^\infty) = \nabla F(\theta^\infty)$. Therefore, $\theta^\infty \in \operatorname{crit} F$.

Finally, (iii) follows readily from the inequality below, due to (3.9) and the definition of T_ϵ:

$$T_\epsilon\, \epsilon^2 \leq \sum_{t=0}^{T_\epsilon} (d^t)^2 \leq l^0 - l^{T_\epsilon + 1} < \infty,$$

where we used the shorthand $l^\nu \triangleq L(\mathbf{x}^\nu, \mathbf{y}^\nu)$. $\square$

Note that, as a direct consequence of Lemma 3.2, one can infer the following further property of the limit points $(\mathbf{x}^\infty, \mathbf{y}^\infty)$ of the sequence $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}_\nu$: any such a $(\mathbf{x}^\infty, \mathbf{y}^\infty)$ is a critical point of L [defined in (3.8)].

3.2. Global convergence under the KŁ property. We now strengthen the subsequence convergence result in Theorem 3.3, proving that the sequence $\{\mathbf{x}^\nu\}_\nu$ is globally convergent to a critical point of F, under the additional assumption that F is a KŁ function [16, 15]. Our analysis blends the approach to *centralized* nonconvex optimization provided in [4] with the subsequence convergence analysis developed in the previous section. We begin introducing the definition of the KŁ property along with some basic facts (cf. Sec. 3.2.1); in Sec. 3.2.2, we then proceed to apply the KŁ property to obtain the global convergence result. Throughout the section we still assume $m = 1$, without loss of generality.

3.2.1. Preliminaries: KŁ properties and basic facts. Let $U : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function; we set $[a < U < b] \triangleq \{\mathbf{z} \in \mathbb{R}^N : a < U(\mathbf{z}) < b\}$. The KŁ property is reviewed below [16, 15].

DEFINITION 3.4 (Kurdyka-Łojasiewicz property).
(a) The function $U : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is said to have the KŁ property at $\acute{\mathbf{z}} \in$ dom ∂U if there exists $\eta \in (0, +\infty]$, a neighborhood $\mathcal{V}_{\acute{\mathbf{z}}}$, and a continuous concave function $\phi : [0, \eta) \to \mathbb{R}_+$ such that:
  (i) $\phi(0) = 0$,
  (ii) ϕ is $\mathcal{C}^1$ on $(0, \eta)$,
  (iii) for all $s \in (0, \eta)$, $\phi'(s) > 0$,
  (iv) for all $\mathbf{z} \in \mathcal{V}_{\acute{\mathbf{z}}} \cap [U(\acute{\mathbf{z}}) < U < U(\acute{\mathbf{z}}) + \eta]$, the KŁ inequality holds:

$$(3.15) \qquad \phi'\left(U(\mathbf{z}) - U(\acute{\mathbf{z}})\right) \operatorname{dist}(0, \partial U(\mathbf{z})) \geq 1.$$

(b) Proper lower semicontinuous functions which satisfy the KŁ inequality at each point of dom ∂U are called KŁ functions.

A lot of functions are known to satisfy the KŁ property; we refer the reader to the recent work [4] (and references therein) for many specific examples of such functions.

To proceed, we make the following extra assumption on F in (P).

Assumption 3.5. The objective function F is a KŁ function.

Since the convergence analysis in Sec. 3.1 leverages the Lyapunov function L defined in (3.8), to build on the KŁ property, we need L to be a KŁ function. Lemma 2.5 together with Assumption 3.5 ensure L to be so.

3.2.2. Convergence analysis. We begin proving the following abstract intermediate result similar to [4], which is at the core of the subsequent analysis.

PROPOSITION 3.6. *In the setting of Theorem 3.3, let L defined in (3.8) have the KŁ property at some $\acute{\mathbf{z}} \triangleq (\acute{\mathbf{x}}, \acute{\mathbf{y}})$. Denote by $\mathcal{V}_{\acute{\mathbf{z}}}$, η, and $\phi : [0, \eta) \to \mathbb{R}_+$ the objects appearing in Definition 3.4. Let $\rho > 0$ be such that $\mathcal{B}(\acute{\mathbf{z}}, \rho) \subseteq \mathcal{V}_{\acute{\mathbf{z}}}$. Consider the sequence $\{\mathbf{z}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{y}^\nu)\}_\nu$ generated by the DOGT Algorithm (2.2), with initialization $\mathbf{z}^0 \triangleq (\mathbf{x}^0, \mathbf{y}^0)$; and define $\acute{l} \triangleq L(\acute{\mathbf{z}})$ and $l^\nu \triangleq L(\mathbf{z}^\nu)$. Suppose that*

$$(3.16) \qquad \acute{l} < l^\nu < \acute{l} + \eta, \quad \forall \nu \geq 0,$$

and

$$(3.17) \qquad K M \phi(l^0 - \acute{l}) + \left\| \mathbf{z}^0 - \acute{\mathbf{z}} \right\| < \rho,$$

where

$$(3.18) \qquad K = \sqrt{3}(1 + L_c) \max\left(\frac{4nK_{\|}^2}{1 - \tilde{\rho}_R}, \frac{K_{\|}^2}{\varkappa(1 - \tilde{\rho}_C)} \left(\alpha + \frac{2\sqrt{n}}{1 + L_c} \right)^2, \alpha^2/\Gamma \right)^{1/2},$$

and $M > 0$ is defined in (3.13) *(cf. Lemma 3.2).*

Then, $\{\mathbf{z}^\nu\}_\nu$ satisfies:

(i) $\mathbf{z}^\nu \in \mathcal{B}(\acute{\mathbf{z}}, \rho)$, for all $\nu \geq 0$;

(ii) $\sum_{t=k}^\nu \left\| \mathbf{z}^{t+1} - \mathbf{z}^t \right\| \leq KM \left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l}) \right)$ for all $\nu, k \geq 0$ and $\nu \geq k$;

(iii) $l^\nu \to \acute{l}$, as $\nu \to \infty$.

Proof. Let $d^\nu > 0$ for all integer $\nu \geq 0$; otherwise, $\{\mathbf{x}^\nu\}_\nu$ converges in a finite number of steps, and its limit point is $\mathbf{x}^\infty = \mathbf{1}\theta^\infty$, for some $\theta^\infty \in \text{crit } F$.

We first bound the "length" $\sum_{t=k}^\nu \left\| \mathbf{z}^{t+1} - \mathbf{z}^t \right\|$. By (2.2), there holds

$$\mathbf{x}^{\nu+1} - \mathbf{x}^\nu = (\mathbf{R} - \mathbf{I})(\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu) - \alpha(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu) - \alpha\mathbf{c}\bar{y}^\nu,$$

$$\mathbf{y}^{\nu+1} - \mathbf{y}^\nu = (\mathbf{C} - \mathbf{I})(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu) + \nabla F_c(\mathbf{x}^{\nu+1}) - \nabla F_c(\mathbf{x}^\nu).$$

Using $||\mathbf{A}||_2 \leq \sqrt{n}||\mathbf{A}||_\infty$ and $||\mathbf{A}||_2 \leq \sqrt{n}||\mathbf{A}||_1$, with $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$; and $||\mathbf{R}-\mathbf{I}||_\infty \leq 2$ and $||\mathbf{C}-\mathbf{I}||_1 \leq 2$, we get

$$\sum_{t=k}^\nu \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\| \leq \sum_{t=k}^\nu 2\sqrt{n} \left\| \mathbf{x}^t - \mathbf{1}\bar{x}^t \right\| + \alpha \left\| \mathbf{y}^t - \mathbf{c}\bar{y}^t \right\| + \alpha|\bar{y}^t|,$$

$$\sum_{t=k}^\nu \left\| \mathbf{y}^{t+1} - \mathbf{y}^t \right\| \leq \sum_{t=k}^\nu 2\sqrt{n} \left\| \mathbf{y}^t - \mathbf{c}\bar{y}^t \right\| + L_c \sum_{t=k}^\nu \left\| \mathbf{x}^{t+1} - \mathbf{x}^t \right\|,$$

where L_c is the Lipschitz constant of ∇F_c. The above inequalities imply

$$\sum_{t=k}^\nu \left\| \mathbf{z}^{t+1} - \mathbf{z}^t \right\|$$

(3.19)
$$\leq \sum_{t=k}^\nu 2(1 + L_c)\sqrt{n}K_{||} \left\| \mathbf{x}^t - \mathbf{1}\bar{x}^t \right\|_R + K_{||} \left(\alpha(1 + L_c) + 2\sqrt{n} \right) \left\| \mathbf{y}^t - \mathbf{c}\bar{y}^\nu \right\|_C$$

$$+ \alpha(1 + L_c)|\bar{y}^t| \leq K \sum_{t=k}^\nu d^t,$$

where K is defined in (3.18).

We prove now the proposition, starting from statement (ii). Multiplying both sides of (3.9) by $\phi'(l^\nu - \acute{l})$ and using $\phi'(l^\nu - \acute{l}) > 0$ [due to property (iii) in Definition 3.4 and (3.16)] and the concavity of ϕ, yield

(3.20) $(d^\nu)^2 \, \phi'(l^\nu - \acute{l}) \leq \phi'(l^\nu - \acute{l}) \, (l^\nu - l^{\nu+1}) \leq \phi(l^\nu - \acute{l}) - \phi(l^{\nu+1} - \acute{l}).$

For all $\mathbf{z} \in \mathcal{V}_{\acute{\mathbf{z}}} \cap [\acute{l} < L < \acute{l} + \eta]$, the KŁ inequality (3.15) holds; hence, assuming $\mathbf{z}^t \in \mathcal{B}(\acute{\mathbf{z}}, \rho)$ for all $t = 0, \ldots, \nu$, yields

(3.21) $\phi'(l^t - \acute{l})||\nabla L(\mathbf{z}^t)|| \geq 1, \quad t = 0, \ldots, \nu,$

which together with (3.20) and (3.13) (cf. Lemma 3.2), gives

$$M \left(\phi(l^t - \acute{l}) - \phi(l^{t+1} - \acute{l}) \right) \geq d^t, \quad t = 0, \ldots, \nu,$$

and thus

(3.22) $M \left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l}) \right) \geq \sum_{t=k}^\nu d^t.$

398   Combining (3.22) with (3.19), we obtain

399 (3.23) $$\sum_{t=k}^{\nu} \left\| \mathbf{z}^{t+1} - \mathbf{z}^{t} \right\| \leq KM \left(\phi(l^{k} - \acute{l}) - \phi(l^{\nu+1} - \acute{l}) \right).$$

400 Ineq. (3.23) proves (ii) if $\mathbf{z}^{\nu} \in \mathcal{B}(\acute{\mathbf{z}}, \rho)$ for all $\nu \geq 0$, which is shown next.
401   Now let us prove statement (i). Letting $k = 0$ in (3.23), by (3.17), we obtain

402          $$\left\| \mathbf{z}^{\nu+1} - \acute{\mathbf{z}} \right\| \leq KM \left(\phi(l^{0} - \acute{l}) - \phi(l^{\nu+1} - \acute{l}) \right) + \left\| \mathbf{z}^{0} - \acute{\mathbf{z}} \right\| < \rho.$$

403 Therefore, $\mathbf{z}^{\nu} \in \mathcal{B}(\acute{\mathbf{z}}, \rho)$, for all $\nu \geq 0$.
404   We finally prove statement (iii). Inequalities (3.13) (cf. Lemma 3.2) and (3.21)
405 imply

406 (3.24) $$\phi'(l^{\nu} - \acute{l}) \, d^{\nu} \geq 1/M, \quad \nu \geq 0.$$

407 On the other hand, by Lemma 3.1-(i), as $\nu \to \infty$, we have $l^{\nu} \to p$, for some $p \geq \acute{l}$.
408 In fact, $p = \acute{l}$, otherwise $p - \acute{l} > 0$, which would contradict (3.24) (because $d^{\nu} \to 0$ as
409 $\nu \to \infty$ and $\phi'(p - \acute{l}) < \infty$).   □

410   Roughly speaking, Proposition 3.6 states that, if the algorithm is initialized in
411 a suitably chosen neighborhood of a point at which L satisfies the KŁ property,
412 then it will converge to that point. Combining this property with the subsequence
413 convergence proved in Theorem 3.7 we can obtain global convergence of the sequence
414 to critical points of F, as stated next.

415   THEOREM 3.7. *Consider the setting of Theorem 3.3, and further assume that*
416 *Assumption 3.5 holds. Each bounded sequence $\{(\mathbf{x}^{\nu}, \mathbf{y}^{\nu})\}_{\nu}$ generated by the DOGT*
417 *Algorithm (2.2) converges to some $(\mathbf{x}^{\infty}, \mathbf{y}^{\infty}) \in \mathrm{crit}\, L$. Furthermore, $\mathbf{x}^{\infty} = \mathbf{1}\theta^{\infty}$, for*
418 *some $\theta^{\infty} \in \mathrm{crit}\, F$.*

419   *Proof.* Let $\mathbf{z}^{\infty} \triangleq (\mathbf{x}^{\infty}, \mathbf{y}^{\infty})$ be a limit point of $\{\mathbf{z}^{\nu} \triangleq (\mathbf{x}^{\nu}, \mathbf{y}^{\nu})\}_{\nu}$. Since $\{l^{\nu} \triangleq$
420 $L(\mathbf{z}^{\nu})\}_{\nu}$ is convergent (cf. Lemma 3.1) and L is continuous, we deduce $l^{\nu} \to l^{\infty} \triangleq$
421 $L(\mathbf{z}^{\infty})$. The function L has the KŁ property at $\mathbf{z}^{\infty}$; set $\acute{\mathbf{z}} = \mathbf{z}^{\infty}$ and $\acute{l} = l^{\infty}$;
422 denote by $\mathcal{V}_{\acute{\mathbf{z}}}$, η, and $\phi : [0, \eta) \to \mathbb{R}_{+}$ the objects appearing in Definition 3.4; and
423 let $\rho > 0$ be such that $\mathcal{B}(\acute{\mathbf{z}}, \rho) \subseteq \mathcal{V}_{\acute{\mathbf{z}}}$. By the continuity of ϕ and the properties
424 above, we deduce that there exists an integer ν_{0} such that i) $l^{\nu} \in (\acute{l}, \acute{l} + \eta)$, for all
425 $\nu \geq \nu_{0}$; and ii) $K\,M\,\phi(l^{\nu_{0}} - \acute{l}) + \|\mathbf{z}^{\nu_{0}} - \acute{\mathbf{z}}\| < \rho$, with K and M defined in (3.18) and
426 (3.13), respectively. Global convergence of the sequence $\{\mathbf{z}^{\nu}\}_{\nu}$ follows by applying
427 Proposition 3.6 to the sequence $\{\mathbf{z}^{\nu+\nu_{0}}\}_{\nu}$.
428   Finally, by Lemma 3.1(ii), $d^{\nu} \to 0$ as $\nu \to \infty$. Invoking the continuity of ∇L and
429 Lemma 3.2, we have $\nabla L(\mathbf{x}^{\infty}, \mathbf{y}^{\infty}) = \mathbf{0}$, thus $(\mathbf{x}^{\infty}, \mathbf{y}^{\infty}) \in \mathrm{crit}\, L$. By Theorem 3.3(ii),
430 $\mathbf{x}^{\infty} = \mathbf{1} \otimes \theta^{\infty}$, with $\theta^{\infty} \in \mathrm{crit}\, F$.   □

431   In the following theorem, we provide some convergence rate estimates.

432   THEOREM 3.8. *In the setting of Theorem 3.7, let L be a KL function with $\phi(s) =$*
433 $cs^{1-\theta}$*, for some constant $c > 0$ and $\theta \in [0, 1)$. Let $\{\mathbf{z}^{\nu} \triangleq (\mathbf{x}^{\nu}, \mathbf{y}^{\nu})\}_{\nu}$ be a bounded*
434 *sequence generated by the DOGT Algorithm (2.2). Then, there hold:*
435   *(i) If $\theta = 0$, $\{\mathbf{z}^{\nu}\}_{\nu}$ converges to $\mathbf{z}^{\infty}$ in a finite number of iterations;*
436   *(ii) If $\theta \in (0, 1/2]$, then $\|\mathbf{z}^{\nu} - \mathbf{z}^{\infty}\| \leq C\tau^{\nu}$, for some $\tau \in [0, 1)$, $C > 0$, and all*
437   *$\nu \geq 0$;*
438   *(iii) If $\theta \in (1/2, 1)$, then $\|\mathbf{z}^{\nu} - \mathbf{z}^{\infty}\| \leq C\nu^{-\frac{1-\theta}{2\theta-1}}$, for some $C > 0$ and all $\nu \geq 0$.*

439   *Proof.* Define $D^\nu \triangleq \sum_{t=\nu}^\infty d^t$. By (3.19), we have

440   (3.25)
$$\left\| \mathbf{z}^{\nu+1} - \mathbf{z}^\infty \right\| \leq \sum_{t=\nu}^\infty \left\| \mathbf{z}^{t+1} - \mathbf{z}^t \right\| \leq K D^\nu.$$

441   It is then sufficient to establish the convergence rates for the sequence $\{D^\nu\}_\nu$.

442   By KŁ inequality (3.15) and (3.13), we have

443   (3.26)
$$M d^\nu \phi'(l^\nu - l^\infty) \geq 1 \implies \tilde{M}(d^\nu)^{(1-\theta)/\theta} \geq (l^\nu - l^\infty)^{1-\theta},$$

444   where $\tilde{M} = (Mc(1-\theta))^{(1-\theta)/\theta}$, $l^\nu \triangleq L(\mathbf{z}^\nu)$, and $l^\infty \triangleq L(\mathbf{z}^\infty)$. In addition, by (3.22)
445   (setting $\acute{l} = l^\infty$), we have $D^\nu \leq M\phi(l^\nu - l^\infty) = Mc(l^\nu - l^\infty)^{1-\theta}$, which together with
446   (3.26), yields

447   (3.27)
$$D^\nu \leq \tilde{M} M c (d^\nu)^{(1-\theta)/\theta} = \tilde{M} M c (D^\nu - D^{\nu+1})^{(1-\theta)/\theta}.$$

448   The convergence rate estimates as stated in the theorem can be derived from (3.27),
449   using the same line of analysis introduced in [3]. The remaining part of the proof is
450   provided in Appendix A.2 for completeness.   □

451   **3.3. Second-order guarantees.** We prove that the DOGT algorithm almost
452   surely converges to SoS solutions of (P), under a suitably chosen initialization and
453   some additional conditions on the weight matrices $\mathbf{R}$ and $\mathbf{C}$. Following a path first
454   established in [18] and further developed in [17], the key to our argument for the non-
455   convergence to strict saddle points of F lies in formulating the DOGT algorithm as
456   a dynamical system while leveraging an instantiation of the stable manifold theorem,
457   as given in [17, Theorem 2]. Our analysis is organized in the following three steps:
458   1) Sec. 3.3.1 introduces the preparatory background; 2) Sec. 3.3.2 tailors the results
459   of Step 1 to the DOGT algorithm; and 3) finally, Sec. 3.3.3 states our main results
460   about convergence of the DOGT algorithm to SoS solutions of (P).

461   **3.3.1. The stable manifold theorem and unstable fixed-points.** Let $g :$
462    $\mathcal{S} \to \mathcal{S}$ be a mapping from $\mathcal{S}$ to itself, where $\mathcal{S}$ is a manifold without boundary.
463   Consider the dynamical system $\mathbf{u}^{\nu+1} = g(\mathbf{u}^\nu)$, with $\mathbf{u}^0 \in \mathcal{S}$; we denote by g^ν the ν-
464   fold composition of g. Our focus is on the analysis of the trajectories of the dynamical
465   system around the fixed points of g; in particular we are interested in the set of
466   unstable fixed points of g. We begin introducing the following definition.

467   DEFINITION 3.9 (Chapter 3 of [1]). *The differential of the mapping* $g : \mathcal{S} \to \mathcal{S}$,
468   *denoted as* $\mathrm{D}g(\mathbf{u})$, *is a linear operator from* $\mathcal{T}(\mathbf{u}) \to \mathcal{T}(g(\mathbf{u}))$, *where* $\mathcal{T}(\mathbf{u})$ *is the tan-*
469   *gent space of* $\mathcal{S}$ *at* $\mathbf{u} \in \mathcal{S}$. *Given a curve* γ *in* $\mathcal{S}$ *with* $\gamma(0) = \mathbf{u}$ *and* $\frac{d\gamma}{dt}(0) = \mathbf{v} \in \mathcal{T}(\mathbf{u})$,
470   *the linear operator is defined as* $\mathrm{D}g(\mathbf{u})\mathbf{v} = \frac{d(g \circ \gamma)}{dt}(0) \in \mathcal{T}(g(\mathbf{u}))$. *The determinant of*
471   *the linear operator* $\det(\mathrm{D}g(\mathbf{u}))$ *is the determinant of the matrix representing* $\mathrm{D}g(\mathbf{u})$
472   *with respect to a standard basis.*[1]

473   We can now introduce the definition of the set of unstable fixed points of g.

474   DEFINITION 3.10 (Unstable fixed points). *The set of unstable fixed points of* g
475   *is defined as*

476   (3.28)
$$\mathcal{A}_g = \left\{ \mathbf{u} : g(\mathbf{u}) = \mathbf{u}, \ \mathrm{spradii}(\mathrm{D}g(\mathbf{u})) > 1 \right\}.$$

---

[1] This determinant may not be uniquely defined, in the sense of being completely invariant to the
basis used for the geometry. In this work, we are interested in properties of the determinant that are
independent of scaling, and thus the potentially arbitrary choice of a standard basis does not affect
our conclusions.

The theorem below, which is based on the stable manifold theorem [28, Theorem III.7], provides tools to let us connect $\mathcal{A}_g$ with the set of limit points which $\{\mathbf{u}^\nu\}_\nu$ can escape from.

THEOREM 3.11 (Theorem 2 of [17]). *Let $g : \mathcal{S} \to \mathcal{S}$ be a $\mathcal{C}^1$ mapping and $\det(\mathbf{D}g(\mathbf{u})) \neq 0$, for all $\mathbf{u} \in \mathcal{S}$. Then, the set of initial points that converge to an unstable fixed point (termed stable set of $\mathcal{A}_g$) is zero measure in $\mathcal{S}$. Therefore,*

$$\mathbb{P}_{\mathbf{u}^0}\left(\lim_{\nu \to \infty} g^\nu(\mathbf{u}^0) \in \mathcal{A}_g\right) = 0,$$

where the probability is taken over the starting point $\mathbf{u}^0 \in \mathcal{S}$.

3.3.2. DOGT as a dynamical system. Theorem 3.11 sets the path to the analysis of the convergence of the DOGT algorithm to SoS solutions of F: it is sufficient to describe the DOGT algorithm by a proper mapping $g : \mathcal{S} \to \mathcal{S}$ satisfying the assumptions in the theorem and such that the non-convergence of $g^\nu(\mathbf{u}^0)$, $\mathbf{u}^0 \in \mathcal{S}$, to $\mathcal{A}_g$ implies the non-convergence of the DOGT algorithm to strict saddles of F.

We begin rewriting the DOGT in an equivalent and more convenient form. Define $\mathbf{h}^\nu \triangleq \mathbf{y}^\nu - \nabla F_c(\mathbf{x}^\nu)$; (2.2) can be rewritten as

$$(3.29) \qquad \begin{cases} \mathbf{x}^{\nu+1} = \mathbf{W}_R\mathbf{x}^\nu - \alpha\left(\mathbf{h}^\nu + \nabla F_c(\mathbf{x}^\nu)\right); \\ \mathbf{h}^{\nu+1} = \mathbf{W}_C\mathbf{h}^\nu + (\mathbf{W}_C - \mathbf{I})\nabla F_c(\mathbf{x}^\nu), \end{cases}$$

with arbitrary $\mathbf{x}^0 \in \mathbb{R}^{mn}$ and $\mathbf{h}^0 = \mathbf{0}$. By Theorem 3.3, every limit point $(\mathbf{x}^\infty, \mathbf{h}^\infty)$ of $\{(\mathbf{x}^\nu, \mathbf{h}^\nu)\}_\nu$ has the form $\mathbf{x}^\infty = \mathbf{1}_n \otimes \boldsymbol{\theta}^\infty$ and $\mathbf{h}^\infty = -\nabla F_c(\mathbf{1}_n \otimes \boldsymbol{\theta}^\infty)$, for some $\boldsymbol{\theta}^\infty \in \text{crit } F$. We are interested in the non-convergence of (3.29) to such points whenever $\boldsymbol{\theta}^\infty \in \text{crit } F$ is a strict saddle of F. This motivates the following definition.

DEFINITION 3.12 (Consensual strict saddle points). *Let $\Theta_{ss}^* = \{\boldsymbol{\theta}^* \in \text{crit } F : \lambda_{min}(\nabla^2 F(\boldsymbol{\theta}^*)) < 0\}$ denote the set of strict saddles of F. The set of consensual strict saddle points is defined as*

$$(3.30) \qquad \mathcal{U}^* \triangleq \left\{\begin{bmatrix} \mathbf{1}_n \otimes \boldsymbol{\theta}^* \\ -\nabla F_c(\mathbf{1}_n \otimes \boldsymbol{\theta}^*) \end{bmatrix} : \boldsymbol{\theta}^* \in \Theta_{ss}^*\right\}.$$

Roughly speaking, $\mathcal{U}^*$ represents the candidate set of "adversarial" limit points which any sequence generated by (3.29) should escape from. The next step is then to write (3.29) as a proper dynamical system whose mapping satisfies conditions in Theorem 3.11 and its set of unstable fixed points $\mathcal{A}_g$ is such that $\mathcal{U}^* \subseteq \mathcal{A}_g$.

Identification of g and $\mathcal{S}$. Define $\mathbf{u} \triangleq (\mathbf{x}, \mathbf{h})$, where $\mathbf{x} \triangleq (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_n)$, and each $\mathbf{x}_i, \mathbf{h}_i \in \mathbb{R}^m$; its value at iteration ν is denoted by $\mathbf{u}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{h}^\nu)$. Consider the dynamical system

$$(3.31) \qquad \mathbf{u}^{\nu+1} = g(\mathbf{u}^\nu), \quad \text{with} \quad g(\mathbf{u}) \triangleq \begin{bmatrix} \mathbf{W}_R\mathbf{x} - \alpha\nabla F_c(\mathbf{x}) - \alpha\mathbf{h} \\ \mathbf{W}_C\mathbf{h} + (\mathbf{W}_C - \mathbf{I})\nabla F_c(\mathbf{x}) \end{bmatrix},$$

and $\mathbf{u}^0 = (\mathbf{x}^0, \mathbf{0})$. Cleary (3.31) describes the trajectory generated by the DOGT algorithm (3.29). However, the initialization imposed by the DOGT scheme leads to a g that maps $\mathbb{R}^{nm} \times \{\mathbf{0}\}$ into $\mathbb{R}^{nm} \times \mathbb{R}^{nm}$. We show next how to change the domain and codomain of g to a subspace $\mathcal{S} \subseteq \mathbb{R}^{nm} \times \mathbb{R}^{nm}$, without affecting the convergence of (3.31) to critical points of F, and consequently that of the DOGT algorithm (2.2).

Applying (3.29) telescopically to the update of the h-variables yields: $\mathbf{h}^\nu = \mathbf{W}_C^\nu \mathbf{h}^0 + (\mathbf{W}_C - \mathbf{I})\mathbf{g}_{\text{acc}}^\nu$, for all $\nu \geq 1$, where $\mathbf{g}_{\text{acc}}^\nu \triangleq \sum_{t=0}^{\nu-1} \mathbf{W}_C^t \nabla F_c(\mathbf{x}^{\nu-t-1})$. Denoting $\bar{\mathbf{h}}^\nu \triangleq (\mathbf{1}_n^\top \otimes \mathbf{I}_m)\mathbf{h}^\nu$, we have

$$(3.32) \qquad \bar{\mathbf{h}}^\nu = \cdots = \bar{\mathbf{h}}^0, \quad \text{and} \quad \mathbf{h}^\nu \in \mathbf{W}_C^\nu \mathbf{h}^0 + \text{span}(\mathbf{W}_C - \mathbf{I}) \quad \forall \nu \geq 1.$$

518   The initialization $\mathbf{h}^0 = \mathbf{0}$ in (3.29) was meant to preserve the tracking property of the
519   h-variables, namely $\bar{\mathbf{h}}^\nu = \mathbf{0}$, for all $\nu \geq 1$. This property still holds if we let instead
520   $\mathbf{h}^0 \in \mathrm{span}\,(\mathbf{W}_C - \mathbf{I})$ (due to the column stochasticity of $\mathbf{C}$). This naturally suggests
521   the following $(2n-1)m$-dimensional subspace as candidate set $\mathcal{S}$:

522   (3.33) $$\mathcal{S} \triangleq \mathbb{R}^{nm} \times \mathrm{span}\,(\mathbf{W}_C - \mathbf{I})\,.$$

523   Such an $\mathcal{S}$ also ensures that $g : \mathcal{S} \to \mathcal{S}$. In fact, by (3.32), $\mathbf{h}^\nu \in \mathrm{span}(\mathbf{W}_C - \mathbf{I})$, for all
524   $\nu \geq 1$, provided that $\mathbf{h}^0 \in \mathrm{span}\,(\mathbf{W}_C - \mathbf{I})$. Therefore, $\{g^\nu(\mathbf{u}^0)\}_\nu \subseteq \mathcal{S}$, for all $\mathbf{u}^0 \in \mathcal{S}$.

525   *Remark* 3.13. The choice of the set $\mathcal{S}$ results in the new initialization of the
526   DOGT iterate (3.29), that is $\mathbf{u}^0 \in \mathcal{S}$. This however does not affect its convergence
527   properties, and the conclusions in Theorem 3.7 (cf. Section 3) still hold. In fact, one
528   can check that the proof of the theorem does not change, since the gradient tracking
529   property, $\bar{\mathbf{h}}^0 = \mathbf{0}$ for all $\nu \geq 0$ [which in (3.2) reads $\bar{g}^\nu = \bar{y}^\nu$ for all $\nu \geq 0$], holds also
530   under the new initialization. Note that such a new initialization can be enforced in
531   a distributed way, with minimal coordination: first, agents choose independently a
532   vector $\mathbf{h}_i^{-1} \in \mathbb{R}^m$; then they run one step of consensus on $\mathbf{h}_i^{-1}$'s with weights matrix
533   $\mathbf{C}$, and set $\mathbf{h}_i^0 = \sum_{j \in \mathcal{N}_i^{in}} C_{ij}\mathbf{h}_j^{-1} - \mathbf{h}_i^{-1}$, resulting in $\mathbf{h}^0 \in \mathrm{span}(\mathbf{W}_C - \mathbf{I})$.

534   Equipped with the mapping g in (3.31) and $\mathcal{S}$ defined in (3.33), we check next
535   that the condition in Theorem 3.11 is satisfied; we then prove that $\mathcal{U}^* \subseteq \mathcal{A}_g$.
536   **1) g is a diffeomorphism:** To establish this property, we add the following
537   extra assumption on the weight matrices $\mathbf{R}$ and $\mathbf{C}$.

538   *Assumption* 3.14. Matrices $\mathbf{R} \in \mathcal{M}_n(\mathbb{R})$ and $\mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ are nonsingular.

539   The above condition is not particularly restrictive and it is compatible with As-
540   sumption 2.3. A rule of thumb is to choose $\mathbf{R} = (\tilde{\mathbf{R}}+\mathbf{I})/2$ and $\mathbf{C} = (\tilde{\mathbf{C}}+\mathbf{I})/2$, with $\mathbf{R}$
541   and $\mathbf{C}$ satisfying Assumption 2.3. The new matrices still satisfy Assumption 2.3 due
542   to the following fact: given two nonnegative matrices $\mathbf{A}, \mathbf{B} \in \mathcal{M}_n(\mathbb{R})$, if the directed
543   graph associated with matrix $\mathbf{A}$ has a spanning tree and $\mathbf{B} \geq \rho\mathbf{A}$, for some $\rho > 0$,
544   then the directed graph associated with matrix $\mathbf{B}$ has a spanning tree as well.
545   We build now the differential of g. Let $\tilde{g}$ be a smooth extension of (3.31) to
546   $\mathbb{R}^{mn} \times \mathbb{R}^{mn}$, that is $g = \tilde{g}|_{\mathcal{S}}$. The differential $\mathrm{D}\tilde{g}(\mathbf{u})$ of $\tilde{g}$ at $\mathbf{u} \in \mathcal{S}$ reads

547   (3.34) $$\mathrm{D}\tilde{g}(\mathbf{u}) = \begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c(\mathbf{x}) & -\alpha\mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\nabla^2 F_c(\mathbf{x}) & \mathbf{W}_C \end{bmatrix};$$

548   $\mathrm{D}\tilde{g}(\mathbf{u})$ is related to the differential of g by $\mathrm{D}g(\mathbf{u}) = \mathrm{D}\tilde{g}(\mathbf{u})\mathbf{P}_{\mathcal{T}(\mathbf{u})}$ [2], where $\mathbf{P}_{\mathcal{T}(\mathbf{u})}$
549   is the orthogonal projector onto $\mathcal{T}(\mathbf{u})$. Using $\mathcal{T}(\mathbf{u}) = \mathcal{S}$, for all $\mathbf{u} \in \mathcal{S}$ (recall that
550   $\mathcal{S}$ is a linear subspace) and denoting by $\mathbf{U}_h \in \mathbb{R}^{mn \times m(n-1)}$ an orthonormal basis of
551   $\mathrm{span}(\mathbf{W}_C - \mathbf{I})$, $\mathrm{D}g(\mathbf{u})$ reads

552   (3.35) $$\mathrm{D}g(\mathbf{u}) = \begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c(\mathbf{x}) & -\alpha\mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\nabla^2 F_c(\mathbf{x}) & \mathbf{W}_C \end{bmatrix}\mathbf{U}\mathbf{U}^\top, \quad \text{with} \quad \mathbf{U} \triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_h \end{bmatrix}.$$

553   Note that $\mathbf{P}_{\mathcal{S}} = \mathbf{U}\mathbf{U}^\top$. We establish next the conditions for g to be a $\mathcal{C}^1$ diffeomor-
554   phism, as stated in Theorem 3.11.

555   PROPOSITION 3.15. *Consider the mapping $g : \mathcal{S} \to \mathcal{S}$ defined in (3.31), under*
556   *Assumptions 2.1(i), 2.3, and 3.14, with $\mathcal{S}$ defined in (3.33). If the step-size is chosen*
557   *according to*

558   (3.36) $$0 < \alpha < \frac{\sigma_{min}(\mathbf{CR})}{L_c},$$

559   *where $L_c = L_{max}$, then $\det(\mathrm{D}g(\mathbf{u})) \neq 0$, for all $\mathbf{u} \in \mathcal{S}$.*

Proof. Since $\mathrm{D}g(\mathbf{u}) : \mathcal{S} \to \mathcal{S}$, it is sufficient to verify that $\mathrm{D}g(\mathbf{u})$ is an invertible linear transformation for every $\mathbf{u} \in \mathcal{S}$. Using the definition of $\mathbf{U}$, this is equivalent to show that $\mathbf{U}^T \mathrm{D}g(\mathbf{u})\mathbf{U}$ is invertible, for all $\mathbf{u} \in \mathcal{S}$. Invoking (3.35), $\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U}$ reads

$$(3.37) \quad \mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U} = \mathbf{U}^\top \mathrm{D}\tilde{g}(\mathbf{u})\mathbf{U} = \begin{bmatrix} \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) & -\alpha \mathbf{U}_h \\ \mathbf{U}_h^\top \left(\mathbf{W}_C - \mathbf{I} \right) \nabla^2 F_c(\mathbf{x}) & \mathbf{U}_h^T \mathbf{W}_C \mathbf{U}_h \end{bmatrix}.$$

Since $\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h$ is non-singular, we can use the Schur complement of $\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U}$ with respect to $\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h$ and write

$$(3.38)$$
$$\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U} = \mathbf{S}_1 \begin{bmatrix} \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) + \alpha \boldsymbol{\Phi} \left(\mathbf{W}_C - \mathbf{I} \right) \nabla^2 F_c(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h \end{bmatrix} \mathbf{S}_2,$$

where $\boldsymbol{\Phi} \triangleq \mathbf{U}_h \left(\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h \right)^{-1} \mathbf{U}_h^\top$, and $\mathbf{S}_1$ and $\mathbf{S}_2$ are some nonsingular matrices. By (3.38), it is sufficient to show that

$$(3.39) \quad \begin{aligned} \mathbf{S} &\triangleq \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) + \alpha \boldsymbol{\Phi} \left(\mathbf{W}_C - \mathbf{I} \right) \nabla^2 F_c(\mathbf{x}) \\ &= \mathbf{W}_R - \alpha \mathbf{W}_C^{-1} \nabla^2 F_c(\mathbf{x}) + \alpha \left(\boldsymbol{\Phi} - \mathbf{W}_C^{-1} \right) \left(\mathbf{W}_C - \mathbf{I} \right) \nabla^2 F_c(\mathbf{x}). \end{aligned}$$

is non-singular. Using $\mathbf{W}_C - \mathbf{I} = \mathbf{U}_h \boldsymbol{\Delta}$, for some $\boldsymbol{\Delta} \in \mathbb{R}^{m(n-1) \times mn}$ (recall that $\mathbf{U}_h$ is an orthonormal basis of span$(\mathbf{W}_C - \mathbf{I})$), we can write

$$(3.40) \quad \begin{aligned} \boldsymbol{\Phi} &= \mathbf{U}_h \left(\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h \right)^{-1} \mathbf{U}_h^\top = \mathbf{U}_h \left(\mathbf{I} + \boldsymbol{\Delta} \mathbf{U}_h \right)^{-1} \mathbf{U}_h^\top \\ &\stackrel{(a)}{=} \mathbf{U}_h \mathbf{U}_h^\top - \mathbf{U}_h \boldsymbol{\Delta} \left(\mathbf{I} + \mathbf{U}_h \boldsymbol{\Delta} \right)^{-1} \mathbf{U}_h \mathbf{U}_h^\top \\ &= \mathbf{U}_h \mathbf{U}_h^\top - \left(\mathbf{W}_C - \mathbf{I} \right) \mathbf{W}_C^{-1} \mathbf{U}_h \mathbf{U}_h^\top \\ &= \mathbf{W}_C^{-1} \mathbf{U}_h \mathbf{U}_h^\top, \end{aligned}$$

where (a) we used the Woodbury identity of inverse matrices. Using (3.40) in (3.39), we obtain

$$\mathbf{S} = \mathbf{W}_R - \alpha \mathbf{W}_C^{-1} \nabla^2 F_c(\mathbf{x}) - \alpha \mathbf{W}_C^{-1} \underbrace{\left(\mathbf{I} - \mathbf{U}_h \mathbf{U}_h^\top \right) \left(\mathbf{W}_C - \mathbf{I} \right)}_{=\mathbf{0}} \nabla^2 F_c(\mathbf{x})$$

Therefore, if $\alpha < \frac{\sigma_{\min}(\mathbf{CR})}{L_c}$, $\mathbf{S}$ is invertible, and consequently, so is $\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U}$. $\square$

2) The consensual strict saddle points are unstable fixed points of g $(\mathcal{U}^* \subseteq \mathcal{A}_g)$: First of all, note that every limit point of the sequence generated by (3.29) is a fixed point of g on $\mathcal{S}$; the converse might not be true. The next result establishes the desired connection between the set $\mathcal{A}_g$ of unstable fixed points of g (cf. Definition 3.10) and the set $\mathcal{U}^*$ of consensual strict saddle points (cf. Definition 3.12). This will let us infer the instability of $\mathcal{U}^*$ from that of $\mathcal{A}_g$.

PROPOSITION 3.16. *Suppose that Assumption 2.3 holds along with one of the following two conditions*
   (i) The weight matrices $\mathbf{R}$ and $\mathbf{C}$ are symmetric;
   (ii) $m = 1$.
Then, any consensual strict saddle point is an unstable fixed point of g, i.e.,

$$(3.41) \qquad\qquad\qquad \mathcal{U}^* \subseteq \mathcal{A}_g,$$

with $\mathcal{A}_g$ and $\mathcal{U}^*$ defined in (3.28) and (3.30), respectively.

590   *Proof.* Let $\mathbf{u}^* \in \mathcal{U}^*$; $\mathbf{u}^*$ is a fixed point of g defined in (3.31). It is thus sufficient
591   to show that $\mathrm{D}g(\mathbf{u}^*)$ has an eigenvalue with magnitude greater than one.
592   To do so, we begin showing that the differential $\mathrm{D}\tilde{g}(\mathbf{u}^*)$ of $\tilde{g}$ at $\mathbf{u}^*$ has an eigenvalue
593   greater than one. Using (3.34), $\mathrm{D}\tilde{g}(\mathbf{u}^*)$ reads

594   (3.42)
$$\mathrm{D}\tilde{g}(\mathbf{u}^*) = \begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c^* & -\alpha\mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\nabla^2 F_c^\star & \mathbf{W}_C \end{bmatrix},$$

595   where we defined the shorthand $\nabla^2 F_c^* \triangleq \nabla^2 F_c(\mathbf{1}\otimes\boldsymbol{\theta}^*)$, and $\boldsymbol{\theta}^* \in \Theta_{ss}^*$. We need to
596   prove

597   (3.43)
$$\det(\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I}) = 0, \quad \text{for some} \quad |\lambda_u| > 1.$$

598   If $|\lambda_u| > 1$, $\mathbf{W}_C - \lambda_u\mathbf{I}$ is nonsingular (since spradii$(\mathbf{C}) = 1$). Using the Schur com-
599   plement of $\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I}$ with respect to $\mathbf{W}_C - \lambda_u\mathbf{I}$, we have

600   (3.44)
$$\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I} = \tilde{\mathbf{S}}_1 \begin{bmatrix} (\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I})/(\mathbf{W}_C - \lambda_u\mathbf{I}) & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_C - \lambda_u\mathbf{I} \end{bmatrix} \tilde{\mathbf{S}}_2,$$

601   for some $\tilde{\mathbf{S}}_1, \tilde{\mathbf{S}}_2 \in \mathcal{M}_{2mn}(\mathbb{R})$, with $\det(\tilde{\mathbf{S}}_1) = \det(\tilde{\mathbf{S}}_2) = 1$. Given (3.44), (3.43) holds
602   if and only if

603
$$\det \begin{bmatrix} \mathbf{W}_R - \lambda_u\mathbf{I} - \alpha\nabla^2 F_c^\star + \alpha(\mathbf{W}_C - \lambda_u\mathbf{I})^{-1}(\mathbf{W}_C - \mathbf{I})\nabla^2 F_c^* & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_C - \lambda_u\mathbf{I} \end{bmatrix} = 0,$$

604   or equivalently

605   (3.45)
$$\det\left(\mathbf{W}_R - \lambda_u\mathbf{I} - \alpha\nabla^2 F_c^* + \alpha(\mathbf{W}_C - \lambda_u\mathbf{I})^{-1}(\mathbf{W}_C - \mathbf{I})\nabla^2 F_c^*\right) = 0.$$

606   Multiplying both sides of (3.45) by $\det(\mathbf{W}_C - \lambda_u\mathbf{I})$ yields

607   (3.46)
$$Q(\lambda_u) \triangleq \det\left(\underbrace{(\mathbf{W}_C - \lambda_u\mathbf{I})(\mathbf{W}_R - \lambda_u\mathbf{I}) + \alpha(\lambda_u - 1)\nabla^2 F_c^*}_{\triangleq\mathbf{T}(\lambda_u)}\right) = 0.$$

608   Trivially $Q(\lambda_u) > 0$, if $\lambda_u \gg 1$. Therefore, to show that (3.43) holds, it is sufficient to
609   prove that there exists some $\lambda_u > 1$ such that $Q(\lambda_u) \leq 0$. Next, we prove this result
610   under either condition (i) or (ii).
611   Suppose (i) holds; $\mathbf{R}$ and $\mathbf{C}$ are symmetric. Define $\tilde{\boldsymbol{v}} \triangleq \mathbf{1}\otimes\boldsymbol{v}$, where $\boldsymbol{v}$ is
612   the unitary eigenvector associated with a negative eigenvalue of $\nabla^2 F(\boldsymbol{\theta}^*)$, and let
613    $\lambda_{min}(\nabla^2 F(\boldsymbol{\theta}^*)) = -\delta$; we can write

614   (3.47)
$$\tilde{\boldsymbol{v}}^\top \mathbf{T}(\lambda_u)\tilde{\boldsymbol{v}} = n(\lambda_u - 1)(\lambda_u - 1 - \alpha\delta/n) < 0,$$

615   for all $1 < \lambda_u < 1 + \alpha\delta/n$. By Rayleigh-Ritz theorem, $\mathbf{T}(\lambda_u)$ has a negative eigenvalue,
616   implying that there exists some real value $\bar{\lambda}_u > 1$ such that $Q(\bar{\lambda}_u) = 0$.
617   Suppose now that conditions (ii) holds; $\mathbf{W}_R$ and $\mathbf{W}_C$ reduce to $\mathbf{R}$ and $\mathbf{C}$, respec-
618   tively. Note that $\mathbf{R}$ and $\mathbf{C}$ are now not symmetric. Let $\lambda_u = 1 + \epsilon$, and consider the
619   Taylor expansion of

620   (3.48)
$$Q(1 + \epsilon) = \det\left((\mathbf{C} - \mathbf{I})(\mathbf{R} - \mathbf{I}) + \epsilon\left(\alpha\nabla^2 F_c^* + 2\mathbf{I} - \mathbf{C} - \mathbf{R}\right) + \epsilon^2\mathbf{I}\right),$$

621   around $\epsilon = 0$. Define $\mathbf{M} \triangleq (\mathbf{C} - \mathbf{I})(\mathbf{R} - \mathbf{I})$ and $\mathbf{N} \triangleq \alpha\nabla^2 F_c^* + 2\mathbf{I} - \mathbf{C} - \mathbf{R}$. It is clear
622   that $Q(1) = 0$; then, by the Jacobi's formula, we have

(3.49) $$Q(1 + \epsilon) = \mathrm{tr}\Big(\mathrm{adj}\,(\mathbf{M})\,\mathbf{N}\Big)\epsilon + \mathcal{O}(\epsilon^2).$$

Expanding (3.49) yields

(3.50)
$$Q(1 + \epsilon) = \mathrm{tr}\Big(\mathrm{adj}\,(\mathbf{R} - \mathbf{I})\,\mathrm{adj}\,(\mathbf{C} - \mathbf{I})\,\mathbf{N}\Big)\epsilon + \mathcal{O}(\epsilon^2)$$
$$= \mathrm{tr}\Big(\mathbf{1}\tilde{\mathbf{r}}^\top \tilde{\mathbf{c}}\mathbf{1}^\top \mathbf{N}\Big)\epsilon + \mathcal{O}(\epsilon^2) = (\tilde{\mathbf{r}}^\top \tilde{\mathbf{c}})\mathbf{1}^\top \mathbf{N}\mathbf{1}\epsilon + \mathcal{O}(\epsilon^2),$$

where $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{c}}$ are the Perron vectors of $\mathbf{R}$ and $\mathbf{C}$, respectively. The second equality in (3.50) is due to the following fact: a rank-$(n - 1)$ matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ has rank-1 adjugate matrix $\mathrm{adj}\,(\mathbf{A}) = \mathbf{a}\mathbf{b}^\top$, where $\mathbf{a}$ and $\mathbf{b}$ are non-zero vectors belonging to the 1-dimensional null space of $\mathbf{A}$ and $\mathbf{A}^\top$, respectively [11, Sec. 0.8.2]. We also have $\tilde{\zeta} \triangleq \tilde{\mathbf{r}}^\top \tilde{\mathbf{c}} > 0$, due to Lemma 2.4. Furthermore, since $\boldsymbol{\theta}^* \in \Theta_{ss}^*$, $\mathbf{1}^\top \nabla^2 F_c^* \mathbf{1} \leq -\delta$, for some $\delta > 0$, and

(3.51) $$Q(1 + \epsilon) \leq -\delta\tilde{\zeta}\alpha\epsilon + \mathcal{O}(\epsilon^2),$$

which implies the existence of a sufficiently small $\epsilon > 0$ such that $Q(1 + \epsilon) < 0$. Consequently, there must exist some $\bar{\lambda}_u > 1$ such that (3.43) holds. Moreover, such $\bar{\lambda}_u$ is a real eigenvalue of $\mathrm{D}\tilde{g}(\mathbf{u}^*)$.

To summarize, we proved that there exists an eigenpair $(\bar{\lambda}_u, \mathbf{v}_u)$ of $\mathrm{D}\tilde{g}(\mathbf{u}^*)$, with $\bar{\lambda}_u > 1$. Next we show that $(\bar{\lambda}_u, \mathbf{v}_u)$ is also an eigenpair of $\mathrm{D}g(\mathbf{u}^*)$. Let us partition $\mathbf{v}_u \triangleq (\mathbf{v}_u^x, \mathbf{v}_u^h)$ such that

(3.52) $$\begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c\,(\mathbf{x}^*) & -\alpha\mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\,\nabla^2 F_c\,(\mathbf{x}^*) & \mathbf{W}_C \end{bmatrix} \begin{bmatrix} \mathbf{v}_u^x \\ \mathbf{v}_u^h \end{bmatrix} = \bar{\lambda}_u \begin{bmatrix} \mathbf{v}_u^x \\ \mathbf{v}_u^h \end{bmatrix}.$$

In particular, we have $(\mathbf{W}_C - \mathbf{I})\,(\nabla^2 F_c\,(\mathbf{x}^*)\,\mathbf{v}_u^x + \mathbf{v}_u^h) = (\bar{\lambda}_u - 1)\mathbf{v}_u^h$, which implies $\mathbf{v}_u^h \in \mathrm{span}(\mathbf{W}_C - \mathbf{I})$, since $\bar{\lambda}_u - 1 \neq 0$. Therefore, $\mathbf{v}_u \in \mathcal{S}$.

Now, let $\mathbf{P}_\mathcal{S}$ be the orthogonal projection matrix onto $\mathcal{S}$. Since $\mathbf{v}_u \in \mathcal{S}$, we have

(3.53) $$\mathrm{D}\tilde{g}(\mathbf{u}^*)\mathbf{v}_u = \bar{\lambda}_u\mathbf{v}_u \implies \mathrm{D}\tilde{g}(\mathbf{u}^*)\mathbf{P}_\mathcal{S}^\top \mathbf{v}_u = \bar{\lambda}_u\mathbf{v}_u \overset{(a)}{\implies} \mathrm{D}g(\mathbf{u}^*)\mathbf{v}_u = \bar{\lambda}_u\mathbf{v}_u,$$

where (a) is due to $\mathrm{D}g(\mathbf{u}^*) = \mathrm{D}\tilde{g}(\mathbf{u}^*)\mathbf{P}_\mathcal{S}^\top$ [cf. (3.35)]. Hence $(\bar{\lambda}_u, \mathbf{v}_u)$ is also an eigenpair of $\mathrm{D}g(\mathbf{u}^*)$, which completes the proof. $\square$

Remark 3.17. Note that condition (i) in Proposition 3.16 implies that that $\mathcal{G}_C$ and $\mathcal{G}_R$ are undirected graphs. Condition (ii) relaxes this assumption to directed network topologies when $m = 1$.

3.3.3. DOGT likely converges to SoS solutions of (P). Combining Theorem 3.11, Proposition 3.15, and Proposition 3.16, we can readily obtain the following second-order guarantees of the DOGT algorithms.

THEOREM 3.18. *Consider Problem* (P), *under Assumptions 2.1 and 2.2; and let* $\{\mathbf{u}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{h}^\nu)\}_\nu$ *be the sequence generated by the DOGT Algorithm (3.29) under the following tuning: i) the step-size* α *satisfies* (3.12) *and* (3.36); *the weight matrices* $\mathbf{C}$ *and* $\mathbf{R}$ *are chosen according to Assumptions 2.3 and 3.14; and the initialization is set to* $\mathbf{u}^0 \in \mathcal{S}$, *with* $\mathcal{S}$ *defined in* (3.33). *Furthermore, suppose that either (i) or (ii) in Proposition 3.16 holds. Then, we have*

(3.54) $$\mathbb{P}_{\mathbf{u}^0}\Big(\lim_{\nu \to \infty}\mathbf{u}^\nu \in \mathcal{U}^*\Big) = 0,$$

where the probability is taken over $\mathbf{u}^0 \in \mathcal{S}$.

In addition, if F is a KL function, then $\{\mathbf{x}^\nu\}_\nu$ *converges almost surely to* $\mathbf{1} \otimes \boldsymbol{\theta}^\infty$ *at a rate determined in Theorem 3.8, where* $\boldsymbol{\theta}^\infty$ *is a SoS solution of* (P).

Note that (3.54) implies the desired second-order guarantees only when the sequence $\{\mathbf{u}^\nu\}_\nu$ convergences [i.e., the limit in (3.54) exists]; otherwise (3.54) is trivially satisfied, and some limit point of $\{\mathbf{u}^\nu\}_\nu$ can belong to $\mathcal{U}^*$ with non-zero probability. A sufficient condition for the required global convergence of $\{\mathbf{u}^\nu\}_\nu$ is that F is a KL function, which is stated in the second part of the above theorem.

4. The DGD Algorithm: Second-order guarantees. Consider the DGD algorithm in the setting of Sec. 2.2. The iterate (2.1) can be interpreted as an instance of the gradient descent (GD) algorithm applied to $L_\alpha : \mathbb{R}^{nm} \to \mathbb{R}$, defined as

(4.1)
$$L_\alpha(\mathbf{x}) \triangleq F_c(\mathbf{x}) + \frac{1}{2\alpha} \mathbf{x}^\top (\mathbf{I} - \mathbf{W}_D)\,\mathbf{x}.$$

Therefore, (2.1) can also be written as

(4.2)
$$\mathbf{x}^{\nu+1} = \mathbf{x}^\nu - \alpha \nabla L_\alpha(\mathbf{x}^\nu).$$

This connection has been extensively used in the literature to establish convergence of the DGD algorithm. The next section summarizes the existing results, which will be the starting point of our analysis on second-order guarantees.

4.1. Existing convergence results. Convergence of the DGD algorithm applied to the nonconvex problem (P) has been established [34], under the following assumption, which is slightly more restrictive than Assumption 2.1.

Assumption 2.1' (On Problem P). Given Problem (P), suppose that: (i) Assumption 2.1(i) is satisfied; and (ii) each f_i is bounded from below.

The convergence properties of the DGD algorithm are summarized below.

THEOREM 4.1 ([34]). *Consider Problem* (P), *under Assumptions 2.1', 2.2. Let* $\{\mathbf{x}^\nu = (\mathbf{x}_i^\nu)_{i=1}^n\}_\nu$ *be a bounded sequence generated by the DGD algorithm* (2.1) *with* $0 < \alpha < \sigma_{min}(\mathbf{I} + \mathbf{D})/L_c$, *and let* $\bar{\mathbf{x}}^\nu \triangleq (1/n) \sum_{i=1}^n \mathbf{x}_i^\nu$. *Then, the following hold*

(i) [*almost consensus*]: *for all* $i = 1, \ldots, n$ *and* $\nu \in \mathbb{N}_+$,

$$\|\mathbf{x}_i^\nu - \bar{\mathbf{x}}^\nu\| \leq \frac{\alpha R}{1 - \sigma_2},$$

where σ_2 *is the second largest singular value of* $\mathbf{D}$, *and* R *is a universal bound of* $\|\nabla F_c(\mathbf{x}^\nu)\|$;

(ii) [*stationarity*]: *every limit point* $\mathbf{x}^\infty$ *of* $\{\mathbf{x}^\nu\}_\nu$ *is such that* $\mathbf{x}^\infty \in \mathrm{crit}\, L_\alpha$.

In addition, if L_α *is a KL function, then* $\{\mathbf{x}^\nu\}_\nu$ *is globally convergent to some* $\mathbf{x}^\infty \in \mathrm{crit}\, L_\alpha$.

Since (2.1) represents the gradient algorithm applied to L_α, non-convergence of the DGD algorithm to strict saddle points of L_α can be established by a direct application of [17, Corollary 2] to (4.2). The following extra assumption on the weight matrix $\mathbf{D}$ is needed, which is similar to Assumption 3.14 for the DOGT schemes.

Assumption 4.2. The matrix $\mathbf{D} \in \mathcal{M}_n(\mathbb{R})$ is nonsingular.

THEOREM 4.3. *Consider Problem* (P), *under Assumptions 2.1', 2.2, and further assume that each* f_i *is a KL function. Let* $\{\mathbf{x}^\nu\}_\nu$ *be the sequence generated by the DGD algorithm with step-size* $0 < \alpha < \frac{\sigma_{min}(\mathbf{D})}{L_c}$ *and weight matrix* $\mathbf{D}$ *satisfying Assumption 4.2. Then, the stable set of strict saddle points of* L_α *is zero measure in* $\mathbb{R}^{nm}$. *Therefore,* $\{\mathbf{x}^\nu\}_\nu$ *convergences almost surely to a SoS solution of* L_α, *where the probability is taken over the random initialization* $\mathbf{x}^0 \in \mathbb{R}^{nm}$.

Convergence results as stated in Theorems 4.1 and 4.3 are not fully satisfactory, as they do not provide any information on the behavior of DGD near critical points

of F (unless all the functions f_i have the same unique minimizer), including the strict saddles of F. In the following, we fill this gap. Specifically, we first show that the DGD algorithm convergences to a(n arbitrarily small) neighborhood of the critical points of F, for sufficiently small $\alpha > 0$ (cf. Section 4.2). Then, we prove that, under some further (mild) assumptions, such critical points are almost surely SoS solutions of (P), where the randomization is taken on the initial point (cf. Section 4.3).

4.2. DGD converges to a neighborhood of critical points of F. We begin introducing the definition of ϵ-critical points of F.

DEFINITION 4.4. *A point $\boldsymbol{\theta} \in \mathbb{R}^m$ such that $\|\nabla F(\boldsymbol{\theta})\| \leq \epsilon$, with $\epsilon > 0$, is called ϵ-critical point of F. Given $\epsilon > 0$, the set of ϵ-critical points of F is denoted by $\operatorname{crit}_\epsilon F$.*

We establish next a relation between the critical points of L_α and the ϵ-critical points of F.

LEMMA 4.5. *Consider Problem (P), under Assumptions 2.1' and 2.2. Every limit point $\mathbf{x}^\infty = (\mathbf{x}_i^\infty)_{i=1}^n$ of $\{\mathbf{x}^\nu\}_\nu$, generated by the DGD algorithm for $0 < \alpha < \sigma_{min}(\mathbf{I} + \mathbf{D})/L_c$, is such that $\bar{\mathbf{x}}^\infty \in \operatorname{crit}_{\mathcal{O}(\alpha)} F$, with $\bar{\mathbf{x}}^\infty \triangleq (1/n) \sum_{i=1}^n \mathbf{x}_i^\infty$.*

Proof. By Theorem 4.1(ii), $(\mathbf{1} \otimes \mathbf{I})^\top \nabla L_\alpha(\mathbf{x}^\infty) = \mathbf{0}$, which using (4.1) and the column stochasticity of $\mathbf{D}$ yields $(\mathbf{1} \otimes \mathbf{I})^\top \nabla F_c(\mathbf{x}^\infty) = \mathbf{0}$. We can then write

$$\|\nabla F(\bar{\mathbf{x}}^\infty)\| = \left\|(\mathbf{1} \otimes \mathbf{I})^\top \left(\nabla F_c(\mathbf{1} \otimes \bar{\mathbf{x}}^\infty) - \nabla F_c(\mathbf{x}^\infty)\right)\right\|$$

$$\leq L_c \sqrt{n} \|\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty\| \overset{(a)}{\leq} \alpha \cdot \frac{n\sqrt{n}L_c R}{1 - \sigma_2},$$

where in (a) we used Theorem 4.1(i). Therefore, $\bar{\mathbf{x}}^\infty \in \operatorname{crit}_{K'\alpha} F$, with $K' = n\sqrt{n}L_c R/(1 - \sigma_2)$. $\square$

Lemma 4.5 shows that $\bar{\mathbf{x}}^\infty \in \operatorname{crit}_{K'\alpha} F$, for some $K' > 0$. A natural question is whether $\operatorname{dist}(\bar{\mathbf{x}}^\infty, \operatorname{crit} F)$ can be made arbitrarily small by reducing $\alpha > 0$; Lemma 4.7 below provides a positive answer to the question (under a mild assumption on F–see Assumption 4.6). This result, together with Theorem 4.1(i), are enough to show (subsequence) convergence of $\{\bar{\mathbf{x}}^\nu \triangleq (1/n) \sum_{i=1}^n \mathbf{x}_i^\nu\}_\nu$ to a(n arbitrary small) neighborhood of critical points of F with all $\mathbf{x}_i^\nu$'s being almost consensual; this is stated in Theorem 4.8 below.

Assumption 4.6. There exist $R, \varrho > 0$ such that $\|\nabla F(\mathbf{x})\| \geq \varrho$, for all $\mathbf{x} \notin \operatorname{crit} F$ and $\|\mathbf{x}\| > R$.

LEMMA 4.7. *Let $F : \mathbb{R}^m \to \mathbb{R}$ be defined in (P), and satisfies Assumptions 2.1'(i) and 4.6. Then, there holds*

$$\lim_{\epsilon \to 0} \max_{\mathbf{q} \in \operatorname{crit}_\epsilon F} \operatorname{dist}(\mathbf{q}, \operatorname{crit} F) = 0.$$

Proof. By Assumption 4.6, there exists a $\tilde{\epsilon} > 0$ such that for all $\epsilon \leq \tilde{\epsilon}$, $\operatorname{crit}_\epsilon F \cap \overline{\mathcal{B}(\mathbf{0}, R)} = \operatorname{crit} F \cap \overline{\mathcal{B}(\mathbf{0}, R)}$. Thus, it is enough to show

(4.3) $$\lim_{\epsilon \to 0} \max_{\mathbf{q} \in \operatorname{crit}_\epsilon F \cap \mathcal{B}(\mathbf{0}, R)} \operatorname{dist}(\mathbf{q}, \operatorname{crit} F \cap \mathcal{B}(\mathbf{0}, R)) = 0.$$

We prove (4.3) by contradiction. Suppose

(4.4) $$\limsup_{\epsilon \to 0} \max_{\mathbf{q} \in \operatorname{crit}_\epsilon F \cap \mathcal{B}(\mathbf{0}, R)} \operatorname{dist}(\mathbf{q}, \operatorname{crit} F \cap \mathcal{B}(\mathbf{0}, R)) = \gamma > 0.$$

Then, one can find a sequence $\{\mathbf{q}^\nu\}_\nu$, with $\mathbf{q}^\nu \in \operatorname{crit}_{1/\nu} F \cap \mathcal{B}(\mathbf{0}, R)$, such that $\operatorname{dist}(\mathbf{q}^\nu, \operatorname{crit} F \cap \mathcal{B}(\mathbf{0}, R)) \geq \gamma$, for all $\nu \in \mathbb{N}$. Since ∇F is continuous, $\operatorname{crit}_1 F$ is closed, thus $\operatorname{crit}_1 F \cap \mathcal{B}(\mathbf{0}, R)$ is compact. Since $\{\mathbf{q}^\nu\}_\nu \subseteq \operatorname{crit}_{1/\nu} F \cap \mathcal{B}(\mathbf{0}, R) \subseteq \operatorname{crit}_1 F \cap \mathcal{B}(\mathbf{0}, R)$,

$\{\mathbf{q}^\nu\}_n$ is bounded. Let $\{\mathbf{q}^{t_\nu}\}_\nu$ be a convergent subsequence of $\{\mathbf{q}^\nu\}_\nu$; its limit point $\mathbf{q}^\infty$ satisfies $\mathrm{dist}(\mathbf{q}^\infty, \mathrm{crit}\, F) \geq \gamma$. For every given $\acute{\nu} \in \mathbb{N}$, $\{\mathbf{q}^{t_\nu}\}_\nu$ eventually will belong to $\mathrm{crit}_{1/\acute{\nu}}\, F \cap \mathcal{B}(\mathbf{0}, R)$, and thus $\mathbf{q}^\infty \in \mathrm{crit}_{1/\acute{\nu}}\, F \cap \mathcal{B}(\mathbf{0}, R)$. This means that $\|\nabla F(\mathbf{q}^\infty)\| \leq 1/\acute{\nu}$, for all $\acute{\nu} \in \mathbb{N}$, implying $\|\nabla F(\mathbf{q}^\infty)\| = 0$. We thus have $\mathrm{dist}(\mathbf{q}^\infty, \mathrm{crit}\, F) = 0$, which contradicts (4.4) as $\mathbf{q}^\infty \in \mathcal{B}(\mathbf{0}, R)$. □

Combining Lemma 4.5, Lemma 4.7 and Theorem 4.1(i), we readily obtain the desired local convergence result of the DGD algorithm, as summarized next.

THEOREM 4.8. *Consider the setting of Lemma 4.5 and let Assumption 4.6 hold. Then, for any $\epsilon > 0$, there exists a sufficiently small $\alpha > 0$, such that every limit point $\mathbf{x}^\infty = (\mathbf{x}_i^\infty)_{i=1}^n$ of a bounded sequence $\{\mathbf{x}^\nu\}_\nu$ generated by the DGD algorithm with that α (or smaller), satisfies*

$$(4.5) \qquad \mathrm{dist}(\bar{\mathbf{x}}^\infty, \mathrm{crit}\, \mathrm{F}) < \epsilon \quad and \quad \|\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty\| < \epsilon.$$

4.3. DGD likely converges to a neighborhood of SoS solutions of F. Theorem 4.8 proves convergence of the DGD algorithm to a neighborhood of critical points of F where all agents' variables are almost consensual; however nothing is said about the second-order nature of these points. In this section, we show that in fact it is unlikely that DGD gets close to strict saddles of F.

We prove our result under the following extra assumptions.

Assumption 4.9. Each $f_i : \mathbb{R}^m \to \mathbb{R}$ is twice differentiable and $\nabla^2 f_i$ is $L_{\nabla_i^2}$-Lipschitz continuous. The Lipschitz constant of $\nabla^2 F$ and $\nabla^2 F_c$ are $L_{\nabla^2} = \sum_{i=1}^n L_{\nabla_i^2}$ and $L_{\nabla_c^2} = \max_i\, L_{\nabla_i^2}$, respectively.

Assumption 4.10. There exists $\delta > 0$ such that $\lambda_{\min}(\nabla^2 F(\boldsymbol{\theta}^*)) \leq -\delta$, for all $\boldsymbol{\theta}^* \in \Theta_{ss}^*$ (Θ_{ss}^* is the set of strict saddle of F–see Definition 3.12).

Intuition: Our path to prove almost sure convergence of the DGD algorithm to a neighborhood of SoS solutions of (P) will be derived from the non-convergence of DGD to strict saddles of L_α (cf. Theorem 4.5). Roughly speaking, our idea consists in showing that whenever $\bar{\mathbf{x}}^\infty = (1/n) \sum_{i=1}^n \mathbf{x}_i^\infty$ belongs to a sufficiently small neighborhood of a strict saddle of F inside the region (4.5), $\mathbf{x}^\infty = (\mathbf{x}_i^\infty)_{i=1}^n$ must be a *strict saddle of L_α*. The escaping properties of DGD from strict saddles of L_α will then ensure that $\{\bar{\mathbf{x}}^\nu = (1/n) \sum_{i=1}^n \mathbf{x}_i^\nu\}_\nu$ unlikely gets trapped in a neighborhood of a strict saddle of F, ending thus in a neighborhood of a SoS solution of (P).

Proposition 4.11 makes this argument formal; in particular, conditions (i)-(iii) identify the neighborhood of a strict saddle of F with the mentioned escaping properties.

PROPOSITION 4.11. *Consider the setting of Theorem 4.1 and further assume that Assumptions 4.9 and 4.10 hold. Let $\{\mathbf{x}^\nu\}_\nu$ be a bounded sequence generated by the DGD algorithm (2.1) such that its limit point $\mathbf{x}^\infty = (\mathbf{x}_i^\infty)_{i=1}^n$, along with $\bar{\mathbf{x}}^\infty \triangleq (1/n) \sum_{i=1}^n \mathbf{x}_i^\infty$, satisfy*

(i) $\mathrm{dist}(\bar{\mathbf{x}}^\infty, \mathrm{crit}\, \mathrm{F}) < \dfrac{\delta}{2L_{\nabla^2}}$;

(ii) $\|\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty\| < \dfrac{\delta}{2nL_{\nabla_c^2}}$;

(iii) There exists $\boldsymbol{\theta}^* \in \mathrm{proj}_{\mathrm{crit}\, \mathrm{F}}(\bar{\mathbf{x}}^\infty) \cap \Theta_{ss}^*$.

Then, $\mathbf{x}^\infty$ is a strict saddle point of L_α.

Proof. Given $\boldsymbol{\theta} \in \mathbb{R}^m$, let $\boldsymbol{v}(\boldsymbol{\theta})$ denote the unitary eigenvector of $\nabla^2 F(\boldsymbol{\theta})$ associated with the smallest eigenvalue, and define $\tilde{\boldsymbol{v}}(\boldsymbol{\theta}) \triangleq \mathbf{1} \otimes \boldsymbol{v}(\boldsymbol{\theta})$. Then, we have

(4.6)
$$\tilde{\boldsymbol{v}}(\boldsymbol{\theta})^\top \nabla^2 L_\alpha(\mathbf{x}^\infty) \tilde{\boldsymbol{v}}(\boldsymbol{\theta}) \stackrel{(a)}{=} \tilde{\boldsymbol{v}}(\boldsymbol{\theta})^\top \nabla^2 F_c(\mathbf{x}^\infty) \tilde{\boldsymbol{v}}(\boldsymbol{\theta})$$
$$\leq \boldsymbol{v}(\boldsymbol{\theta})^\top \nabla^2 F(\boldsymbol{\theta}) \boldsymbol{v}(\boldsymbol{\theta})$$
$$+ ||\nabla^2 F(\bar{\mathbf{x}}^\infty) - \nabla^2 F(\boldsymbol{\theta})|| \, ||\boldsymbol{v}(\boldsymbol{\theta})||^2 + ||\nabla^2 F_c(\mathbf{x}^\infty) - \nabla^2 F_c(\mathbf{1} \otimes \bar{\mathbf{x}}^\infty)|| \, ||\tilde{\boldsymbol{v}}(\boldsymbol{\theta})||^2$$
$$\stackrel{(b)}{\leq} \boldsymbol{v}(\boldsymbol{\theta})^\top \nabla^2 F(\boldsymbol{\theta}) \boldsymbol{v}(\boldsymbol{\theta}) + L_{\nabla^2} ||\bar{\mathbf{x}}^\infty - \boldsymbol{\theta}|| + n \, L_{\nabla_c^2} ||\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty||$$

where (a) follows from $\tilde{\boldsymbol{v}}(\boldsymbol{\theta}) \in \text{null}(\mathbf{W}_D - \mathbf{I})$; and (b) is due to Assumption 4.9. Let us now evaluate (4.6) at some $\boldsymbol{\theta}^*$ as defined in condition (iii) of the proposition; using $\boldsymbol{v}(\boldsymbol{\theta}^*)^\top \nabla^2 F(\boldsymbol{\theta}^*) \boldsymbol{v}(\boldsymbol{\theta}^*) \leq -\delta$ and conditions (i) and (ii), yields $\tilde{\boldsymbol{v}}(\boldsymbol{\theta}^*)^\top \nabla^2 L_\alpha(\mathbf{x}^\infty) \tilde{\boldsymbol{v}}(\boldsymbol{\theta}^*) < 0$. By the Rayleigh-Ritz theorem, it must be $\lambda_{min}(\nabla^2 L_\alpha(\mathbf{x}^\infty)) < 0$. This, together with Theorem 4.1(ii), proves the thesis. $\square$

Invoking now Theorem 4.8, we infer that there exists a sufficiently small $\alpha > 0$ such that conditions (i) and (ii) of Proposition 4.11 are always satisfied, implying that $\mathbf{x}^\infty$ is a strict saddle of L_α if there exists a strict saddle of F "close" to $\bar{\mathbf{x}}^\infty$ [in the sense of (iii)]. This is formally summarized next.

COROLLARY 4.12. *Consider the setting of Proposition 4.11 and let Assumption 4.6 hold. There exists a sufficiently small $\alpha > 0$ such that, if $\text{proj}_{\text{crit } F}(\bar{\mathbf{x}}^\infty) \cap \Theta_{ss}^* \neq \emptyset$, then $\mathbf{x}^\infty$ is a strict saddle of L_α.*

To state our final result, let us introduce the following merit function: given $\mathbf{x} = (\mathbf{x}_i)_{i=1}^n$, let

$$M(\mathbf{x}) \triangleq \max \Big(\text{dist}(\bar{\mathbf{x}}, \mathcal{X}_{SoS}), ||\mathbf{x} - \mathbf{1} \otimes \bar{\mathbf{x}}|| \Big),$$

where $\mathcal{X}_{SoS}$ denotes the set of SoS solutions of (P), and $\bar{\mathbf{x}} \triangleq (1/n) \sum_{i=1}^n \mathbf{x}_i$. $M(\mathbf{x})$ capture the distance of the average $\bar{\mathbf{x}}$ from the set of SoS solutions of (P) and well as the consensus disagreement of the agents' local variables $\bar{\mathbf{x}}_i$. Using Theorem 4.3 in conjunction with Corollary 4.12, we obtain our final result.

THEOREM 4.13. *Consider Problem (P) under Assumptions 2.1', 2.2, 4.6, 4.9, and 4.10; further assume that each f_i is a KL function. For every $\epsilon > 0$, there exits a sufficiently small $0 < \bar{\alpha} < \frac{\sigma_{min}(\mathbf{D})}{L_c}$ such that*

$$\mathbb{P}_{\mathbf{x}^0} \big(M(\mathbf{x}^\infty) \leq \epsilon \big) = 1,$$

where $\mathbf{x}^\infty = (\mathbf{x}_i^\infty)_{i=1}^n$, with $\bar{\mathbf{x}}^\infty = (1/n) \sum_{i=1}^n \mathbf{x}_i^\infty$, is the limit point of the sequence $\{\mathbf{x}^\nu\}_\nu$ generated by the DGD algorithm (2.1) with $\alpha \in (0, \bar{\alpha}]$, the weight matrix $\mathbf{D}$ satisfying Assumption 4.2, and starting point $\mathbf{x}^0 \in \mathbb{R}^{mn}$; and the probability is taken over the initialization $\mathbf{x}^0 \in \mathbb{R}^{nm}$. Furthermore, any $\boldsymbol{\theta}^* \in \text{proj}_{\text{crit } F}(\bar{\mathbf{x}}^\infty)$ is a SoS solution of F almost surely.

Proof. By Corollary 4.12, for sufficiently small $\alpha < \bar{\alpha}_1$, if $\text{proj}_{\text{crit } F}(\bar{\mathbf{x}}^\infty)$ contains a strict saddle point of F, then $\mathbf{x}^\infty$ is also a strict saddle point of L_α. Let $\bar{\alpha}_2$ be such that for DGD with $\alpha < \bar{\alpha}_2$, by Theorem 4.8, every limit point $\mathbf{x}^\infty$ satisfies $\text{dist}(\bar{\mathbf{x}}^\infty, \text{crit } F) \leq \epsilon$ and $||\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty|| \leq \epsilon$. Consider now the DGD algorithm with $\alpha < \min\{\bar{\alpha}_1, \bar{\alpha}_2\}$. Let $\mathbf{x}^0 \in \mathbb{R}^{mn}$ be drawn randomly from the set of probability one measure defined by Theorem 4.3 for which the algorithm converges to a SoS solution of L_α. By the above properties of α, it holds that $M(\mathbf{x}^\infty) \leq \epsilon$ and $\text{proj}_{\text{crit } F}(\bar{\mathbf{x}}^\infty)$
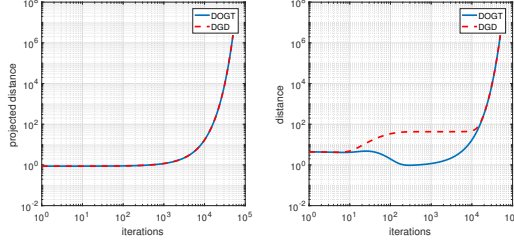
FIG. 1. *Escaping properties of the DGD and DOGT algorithms, applied to Problem (5.1). Left plot: distance of the average iterates from $\boldsymbol{\theta}^*$ projected onto the unstable manifold E_u versus the number of iterations. Right plot: distance of the average iterates from $\boldsymbol{\theta}^*$ versus the number of iterations.*

must contain only SoS solutions of F. Therefore, there exists a $\boldsymbol{\theta}^* \in \text{crit } F$ such that $\boldsymbol{\theta}^* \in \mathcal{X}_{SoS}$ and $\|\bar{\mathbf{x}}^\infty - \boldsymbol{\theta}^*\| \leq \epsilon$. □

5. Numerical Results. In this section we present some preliminary tests showing the behavior of DGD and DOGT algorithms around strict saddle points of a quadratic nonconvex minimization problem.

Consider the following minimization

(5.1)
$$\min_{\boldsymbol{\theta} \in \mathbb{R}^m} F(\boldsymbol{\theta}) \triangleq \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{\theta} - \mathbf{b}_i)^\top \mathbf{Q}_i (\boldsymbol{\theta} - \mathbf{b}_i),$$

where $m = 20$; $n = 10$; $\mathbf{b}_i$'s are i.i.d Gaussian zero mean random vectors with standard deviation 10^3; and $\mathbf{Q}_i$'s are $m \times m$ randomly generated symmetric matrices where $\sum_{i=1}^{n} \mathbf{Q}_i$ has $m - 1$ eigenvalues $\{\lambda_i\}_{i=1}^{m-1}$ uniformly distributed over $(0, n]$ and one negative eigenvalue $\lambda_m = -n\delta$, with $\delta = 0.01$. Clearly (5.1) is an instance of Problem (P), with F having a unique strict saddle point $\boldsymbol{\theta}^* = (\sum_i \mathbf{Q}_i)^{-1} \sum_i \mathbf{Q}_i \mathbf{b_i}$. The network of n agents is modeled as a ring; the weight matrix $\mathbf{W} \triangleq \{w_{ij}\}_{i,j=1}^{n}$, compliant to the graph topology, is generated doubly stochastic.

To test the escaping properties of DGD and DOGT from the strict saddle of F, we initialize the algorithms in a randomly generated neighborhood of $\boldsymbol{\theta}^*$. More specifically, every agent's initial point $\mathbf{x}_i^0 = \boldsymbol{\theta}^* + \boldsymbol{\epsilon}_{x,i}$, $i = 1, \ldots, n$; in addition, for the DOGT algorithm we set $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0) + (w_{ii} - 1)\boldsymbol{\epsilon}_{y,i} + \sum_{j \neq i} w_{ij} \boldsymbol{\epsilon}_{y,j}$, where $\boldsymbol{\epsilon}_{x,i}$'s and $\boldsymbol{\epsilon}_{y,i}$'s are realizations of i.i.d. Gaussian random vectors with standard deviation equal to 1. Both algorithms use the same step-size $\alpha = 0.99 \, \sigma_{min}(\mathbf{I} + \mathbf{W})/L_c$, with $L_c = \max_i\{|\lambda_i|\}$; this is the largest theoretical step-size guaranteeing convergence of the DGD algorithm (cf. Theorem 4.1).

In the left panel of Fig. 1, we plot the distance of the average iterates $\bar{\mathbf{x}}^\nu = (1/n) \sum_{i=1}^{n} \mathbf{x}_i^\nu$ from the critical point $\boldsymbol{\theta}^*$ projected on the unstable manifold $E_u = \text{span}(\mathbf{u}^u)$, where $\mathbf{u}^u$ is the eigenvector associated with the negative eigenvalue $\lambda_m = -n\delta$. In the right panel, we plot $\|\bar{\mathbf{x}}^\nu - \boldsymbol{\theta}^*\|$ versus the number of iterations. All the curves are averaged over 50 independent initializations. Figure in the left panel shows that, as predicted by our theory, both algorithms almost surely escapes from the unstable subspace E_u, at an indistinguishable practical rate. The right panel shows that DOGT gets closer to the strict saddle; this can be justified by the fact that, differently from DGD, DOGT exhibits *exact* convergence to critical points of F.

Appendix A. Appendix.

A.1. Proof of Lemma 2.5. The lemma is a direct consequence of the following result, whose proof is quite standard and reported for completeness.

LEMMA A.1. *Given any* $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ *and* $\epsilon > 0$, *there exists a matrix norm* $||\cdot||_\epsilon$ *such that* $\mathrm{spradii}(\mathbf{A}) \leq ||\mathbf{A}||_\epsilon \leq \mathrm{spradii}(\mathbf{A}) + \epsilon$. *The induced vector norm* $||\cdot||_\epsilon$ *can be written as* $||\mathbf{x}||_\epsilon = ||\mathbf{Hx}||_2$, *where* $\mathbf{H}$ *is an invertible matrix dependent on* $\mathbf{A}$ *and* ϵ. *Furthermore, the vector-norm function* $||\cdot||_\epsilon^2$ *is real-analytic.*

Proof. The proof of the first part is similar to that of [11, Lemma 5.6.10]. The Schur form of the matrix $\mathbf{A}$ is $\mathbf{A} = \mathbf{U}^H \mathbf{T} \mathbf{U}$, where $\mathbf{U}$ is a unitary matrix and $\mathbf{T}$ is an upper triangular matrix with $(\mathbf{T})_{ii} = \lambda_i$ and $(\mathbf{T})_{ij} = d_{ij}$, $i < j$. Define $\mathbf{D}_t \triangleq \mathrm{diag}(t, t^2, t^3, \ldots, t^n)$ and let

(A.1)
$$\boldsymbol{\Delta}_t \triangleq \mathbf{D}_t \mathbf{T} \mathbf{D}_t^{-1} = \begin{bmatrix} \lambda_1 & \frac{d_{12}}{t} & \frac{d_{13}}{t^2} & \cdots & \frac{d_{1n}}{t^{n-1}} \\ 0 & \lambda_2 & \frac{d_{23}}{t} & \cdots & \frac{d_{2n}}{t^{n-2}} \\ 0 & 0 & \lambda_3 & \cdots & \frac{d_{3n}}{t^{n-3}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Now let us define

(A.2)
$$||\mathbf{A}|| \triangleq \left\| (\mathbf{D}_t \mathbf{U}) \, \mathbf{A} \, (\mathbf{D}_t \mathbf{U})^{-1} \right\|_2 = \lambda_{max} \left(\boldsymbol{\Delta}_t \boldsymbol{\Delta}_t^H \right)^{\frac{1}{2}},$$

which is a matrix norm. It is not difficult to check that
(A.3)
$$\boldsymbol{\Delta}_t \boldsymbol{\Delta}_t^H = \begin{bmatrix} |\lambda_1|^2 + \mathcal{O}(\frac{1}{t^2}) & \mathcal{O}(\frac{1}{t}) & \mathcal{O}(\frac{1}{t^2}) & \cdots & \mathcal{O}(\frac{1}{t^{n-1}}) \\ \mathcal{O}(\frac{1}{t}) & |\lambda_2|^2 + \mathcal{O}(\frac{1}{t^2}) & \mathcal{O}(\frac{1}{t}) & \cdots & \mathcal{O}(\frac{1}{t^{n-2}}) \\ \mathcal{O}(\frac{1}{t^2}) & \mathcal{O}(\frac{1}{t}) & |\lambda_3|^2 + \mathcal{O}(\frac{1}{t^2}) & \cdots & \mathcal{O}(\frac{1}{t^{n-3}}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}(\frac{1}{t^{n-1}}) & \mathcal{O}(\frac{1}{t^{n-2}}) & \mathcal{O}(\frac{1}{t^{n-3}}) & \cdots & |\lambda_n|^2 + \mathcal{O}(\frac{1}{t^2}) \end{bmatrix},$$

where big-O notation $\mathcal{O}(\cdot)$ is defined and trivially extended for complex functions. Using the Geršgorin lemma, we conclude that there exists a sufficiently large $t > 0$ such that $||\mathbf{A}|| \leq \mathrm{spradii}(\mathbf{A}) + \epsilon$.

To prove the second part of the lemma, let $\mathbf{H}$ be a nonsingular matrix, and define the vector-norm $||\mathbf{x}||_{H,2} \triangleq ||\mathbf{Hx}||_2$ and the matrix norm $||\mathbf{A}||_H \triangleq ||\mathbf{HAH}^{-1}||_2$. By definition

(A.4) $$||\mathbf{A}||_H = \max_{||\mathbf{x}||_2=1} \left\| \mathbf{HAH}^{-1}\mathbf{x} \right\|_2 = \max_{||\mathbf{Hy}||_2=1} ||\mathbf{HAy}||_2 = \max_{||\mathbf{y}||_{H,2}=1} ||\mathbf{Ay}||_{H,2},$$

implying that the matrix norm $||\mathbf{A}||_H$ induces the vector-norm $||\mathbf{x}||_{H,2}$. Since $||\cdot||_2^2$ is analytic and composition of real-analytic funtions are real-analytic (see Proposition 2.2.8 in [14]), therefore $||\mathbf{x}||_{H,2}^2$ is real-analytic too. The specific matrix norm used in (A.2) is an instance of (A.4) and $\mathbf{D}_t\mathbf{U}$ is full-rank, thus $||\mathbf{x}||_{D_t U,2}^2$ is real-analytic. $\square$

A.2. Supplement for the proof of Theorem 3.8. We first show that, if there exists some ν_0 such that $d^{\nu_0} = 0$, $\mathbf{z}^\nu = \mathbf{z}^{\nu_0}$, for all $\nu \geq \nu_0$ [see updates in (2.2)]; this means that $\{\mathbf{z}^\nu\}_{\nu \in \mathbb{N}_+}$ converges in finitely many iterations. Define $\mathcal{D} \triangleq \{\nu : d^\nu \neq 0\}$ and take ν in $\mathcal{D}$. Let $\theta = 0$, then the KŁ inequality yields $||\nabla L(\mathbf{x}^\nu, \mathbf{y}^\nu)|| \geq 1/c$, for

all $\nu \in \mathcal{D}$. This together with (3.9) and Lemma 3.2, lead to $l^{\nu+1} \leq l^{\nu} - 1/(Mc)^2$, which by Assumption 2.1-(ii), implies that $\mathcal{D}$ must be finite and $\{\mathbf{z}^{\nu}\}_{\nu \in \mathbb{N}_+}$ converges in a finite number of iterations.

Consider (3.27). Let $\theta \in (0, 1/2]$, then $(1-\theta)/\theta \geq 1$. Since $D^{\nu} \to 0$ as $\nu \to \infty$ [by Lemma 3.1-(ii)], there exists a sufficiently large ν_0 such that $(D^{\nu} - D^{\nu+1})^{(1-\theta)/\theta} \leq D^{\nu} - D^{\nu+1}$. By (3.27), we have

$$D^{\nu+1} \leq \frac{\tilde{M}Mc - 1}{\tilde{M}Mc}D^{\nu},$$

which proves case (ii).

Finally, let us assume $\theta \in (1/2, 1)$, then $\theta/(1-\theta) > 1$. Eq. (3.27) implies

$$1 \leq \frac{\bar{M}(D^{\nu} - D^{\nu+1})}{(D^{\nu})^{\theta/(1-\theta)}}$$

where $\bar{M} = (M\tilde{M}c)^{\theta/(1-\theta)}$. Define $h : (0, +\infty) \to \mathbb{R}$ by $h(s) \triangleq s^{-\frac{\theta}{1-\theta}}$. Since h is monotonically decreasing over $[D^{\nu+1}, D^{\nu}]$, we get

(A.5)
$$1 \leq \bar{M}(D^{\nu} - D^{\nu+1})h(D^{\nu}) \leq \bar{M}\int_{D^{\nu+1}}^{D^{\nu}} h(s)ds = \bar{M}\frac{1-\theta}{1-2\theta}\left((D^{\nu})^p - (D^{\nu+1})^p\right),$$

with $p = \frac{1-2\theta}{1-\theta} < 0$. By (A.5) one infers that there exists a constant $\mu > 0$ such that $(D^{\nu+1})^p - (D^{\nu})^p \geq \mu$. The following chain of implications then holds: $(D^{\nu+1})^p \geq \mu\nu + (D^1)^p \implies D^{\nu+1} \leq \left(\mu\nu + (D^1)^p\right)^{1/p} \implies D^{\nu+1} \leq C_0\nu^{1/p}$, for some constant $C_0 > 0$. This proves case (iii).

REFERENCES

[1] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, USA, 2007.

[2] P. A. ABSIL, R. MAHONY, AND J. TRUMPF, *An extrinsic look at the riemannian hessian*, in Geom. Sci. Inf., Springer Berlin Heidelberg, 2013, pp. 361–368.

[3] H. ATTOUCH AND J. BOLTE, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program., 116 (2009), pp. 5–16.

[4] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods*, Math. Program., 137 (2013), pp. 91–129.

[5] F. BÉNÉZIT, V. BLONDEL, P. THIRAN, J. TSITSIKLIS, AND M. VETTERLI, *Weighted gossip: Distributed averaging using non-doubly stochastic matrices*, in IEEE Intern. Symp. on Inf. Theory, 2010, pp. 1753–1757.

[6] P. DI LORENZO AND G. SCUTARI, *Distributed nonconvex optimization over networks*, in IEEE Intern. Conf. on Comput. Advances in Multi-Sensor Adapt. Process., 2015, pp. 229–232.

[7] P. DI LORENZO AND G. SCUTARI, *NEXT: In-network nonconvex optimization*, IEEE Trans. Signal Inf. Process. Netw., 2 (2016), pp. 120–136.

[8] R. GE, F. HUANG, C. JIN, AND Y. YUAN, *Escaping from saddle points — online stochastic gradient for tensor decomposition*, in Proc. of the 28th Conf. on Learn. Theory, 2015, pp. 797–842.

[9] R. GE, J. D. LEE, AND T. MA, *Matrix completion has no spurious local minimum*, in Proc. of the 30th Intern. Conf. on Neural Inf. Process. Syst., 2016, pp. 2981–2989.

[10] M. HONG, J. D. LEE, AND M. RAZAVIYAYN, *Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization*, arXiv:1802.08941, (2018).

[11] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, NY, USA, 2nd ed., 2012.

[12] C. JIN, R. GE, P. NETRAPALLI, S. M. KAKADE, AND M. I. JORDAN, *How to escape saddle points efficiently*, in Proc. of the 34th Intern. Conf. on Mach. Learn., vol. 70, 2017, pp. 1724–1732.

[13] K. KAWAGUCHI, *Deep learning without poor local minima*, in Proc. of the Advances in Neural Inf. Process. Syst. 29, 2016, pp. 586–594.

[14] S. KRANTZ AND H. PARKS, *A Primer of Real Analytic Functions*, Birkhäuser Boston, 2002.

[15] K. KURDYKA, *On gradients of functions definable in o-minimal structures*, Annales de l'institut Fourier, 48 (1998), pp. 769–783.

[16] S. Ł OJASIEWICZ, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles (Paris, 1962), 1963, pp. 87–89.

[17] J. D. LEE, I. PANAGEAS, G. PILIOURAS, M. SIMCHOWITZ, M. I. JORDAN, AND B. RECHT, *First-order methods almost always avoid saddle points*, arXiv:1710.07406, (2017).

[18] J. D. LEE, M. SIMCHOWITZ, M. I. JORDAN, AND B. RECHT, *Gradient descent only converges to minimizers*, in Proc. of 29th Conf. on Learn. Theory, 2016, pp. 1246–1257.

[19] A. NEDIĆ AND A. OLSHEVSKY, *Distributed optimization over time-varying directed graphs*, IEEE Trans. Autom. Control, 60 (2015), pp. 601–615.

[20] A. NEDIĆ, A. OLSHEVSKY, AND W. SHI, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM J. on Optim., 27 (2017), pp. 2597–2633.

[21] A. NEDIĆ AND A. OZDAGLAR, *Distributed subgradient methods for multi-agent optimization*, IEEE Trans. Autom. Control, 54 (2009), pp. 48–61.

[22] A. NEDIĆ, A. OZDAGLAR, AND P. A. PARRILO, *Constrained consensus and optimization in multi-agent networks*, IEEE Trans. Autom. Control, 55 (2010), pp. 922–938.

[23] M. O'NEILL AND S. J. WRIGHT, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, arXiv:1706.07993, (2017).

[24] R. PEMANTLE, *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18 (1990), pp. 698–712.

[25] S. PU, W. SHI, J. XU, AND A. NEDICH, *A push-pull gradient method for distributed optimization in networks*, arXiv:1803.07588v1, (2018).

[26] G. SCUTARI AND Y. SUN, *Distributed nonconvex constrained optimization over time-varying digraphs*, arXiv:1809.01106, (2017).

[27] G. SCUTARI AND Y. SUN, *Parallel and distributed successive convex approximation methods for big-data optimization*, in Multi-Agent Optimization, F. Facchinei and J.-S. Pang, eds., Springer, C.I.M.E. Foundation Subseries (Lecture Notes in Mathematics), 2018, pp. 1–158, arXiv:1805.06963.

[28] M. SHUB, *Global stability of dynamical systems*, Springer-Verlag, 1987.

[29] Y. SUN, A. DANESHMAND, AND G. SCUTARI, *Convergence rate of distributed optimization algorithms with gradient tracking*, arxiv preprint, (2018).

[30] Y. SUN, G. SCUTARI, AND D. PALOMAR, *Distributed nonconvex multiagent optimization over time-varying networks*, in Proc. of the 50th Asilomar Conf. on Sign., Syst., and Comp., 2016, pp. 788–794.

[31] T. TATARENKO AND B. TOURI, *Non-convex distributed optimization*, IEEE Trans. on Autom. Control, 62 (2017), pp. 3744–3757.

[32] R. XIN AND U. A. KHAN, *A linear algorithm for optimization over directed graphs with geometric convergence*, IEEE Control Syst. Lett., 2 (2018), pp. 325–330.

[33] J. XU, S. ZHU, Y. C. SOH, AND L. XIE, *Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes*, in IEEE Conf. on Decis. Control (CDC), 2015, pp. 2055–2060.

[34] J. ZENG AND W. YIN, *On nonconvex decentralized gradient descent*, IEEE Trans. on Signal Process., 66 (2018), pp. 2834–2848.