SIAM J. Optim.
Vol. 30, No. 4, pp. 3029–3068

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 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, new convergence rate results for first-order critical points is established for the latter class of algorithms.

Key words. distributed gradient methods, gradient tracking, nonconvex optimization

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

DOI. 10.1137/18M121784X

1. Introduction.

We consider smooth unconstrained nonconvex optimization over networks in the following form:

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

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 multiagent control; see, e.g., [58]. 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

[*] Received by the editors October 24, 2018; accepted for publication (in revised form) May 24, 2020; published electronically October 22, 2020. Part of this work has been presented at the 56th Annual Allerton Conference on Communication, Control, and Computing, IEEE, Piscataway, NJ, 2018, pp. 359–365 [19].
https://doi.org/10.1137/18M121784X
Funding: The work of the first and second authors has been supported by the USA National Science Foundation under grants CIF 1564044, CIF 1719205, and CMMI 1832688; and by the Army Research Office under grant W911NF1810238. The work of the third author was supported by the OP VVV project CZ.02.1.01/0.0/0.0/16 019/0000765, "Research Center for Informatics."

[†] School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA (adaneshm@purdue.edu, http://web.ics.purdue.edu/~adaneshm/, gscutari@purdue.edu, https://engineering.purdue.edu/~gscutari/).
[‡] Department of Computer Science, Czech Technical University in Prague, Department of Computer Science, Faculty of Electrical Engineering, Prague, Czech Republic (vyacheslav.kungurtsev@fel.cvut.cz).

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 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 [26], low-rank matrix completion [27], and some instances of neural networks [37], just to name a few. In all these cases, converging to SoS solutions—and thus circumventing strict saddles–guarantees finding a local minimizer.

This paper studies for the first time second-order guarantees of two renowned distributed gradient-based algorithms for problem (P), namely, the distributed gradient descent (DGD) [49, 50] and the family of distributed algorithms based on gradient tracking [21, 22, 68]. The former is implementable on undirected graphs while the latter is suitable also for directed graphs. Convergence of these schemes applied to *convex* instances of (P) is well understood; however, less is known in the nonconvex case, let alone second-order guarantees; the relevant works are discussed next.

1.1. Literature review.

Recent years have witnessed many studies proving asymptotic solution—and convergence rate—grantees for a variety of algorithms for specific classes of nonconvex optimization problems (e.g., satisfying suitable regularity conditions); a good overview can be found in [16]. Since these analyses are heavily tailored to specific applications and it is unclear how to generalize them to a wider class of nonconvex functions, we omit further details and discuss next only results of centralized and distributed algorithms for *general* nonconvex instances of (P).

1.1.1. Second-order guarantees of centralized optimization algorithms.

Second-order guarantees of centralized solution methods for general nonconvex optimization (P) have been extensively studied in the literature.

Hessian-based methods: Algorithms based on *second-order* information have long been known to converge to SoS solutions of (P); they rely on computing the Hessian to distinguish between first- and SoS points. The classical cubic regularization [29, 52, 14, 15, 3] and trust region (e.g., [46, 55, 17, 20]) methods can provably find *approximate* SoS solutions in polynomial time (by approximate SoS we mean $\boldsymbol{\theta}$ such that $||\nabla F(\boldsymbol{\theta})|| \leq \epsilon_g$ and $\lambda_{\min}(\nabla^2 F(\boldsymbol{\theta})) \geq -\epsilon_h$ for small $\epsilon_g, \epsilon_h > 0$); they however require access to the full Hessian matrix. A recent line of works [13, 4, 12] show that the requirement of full Hessian access can be relaxed to Hessian-vector products in each iteration, hence solving simpler subproblems per iteration, but at the cost of requiring more iterations to reach approximate SoS solutions.

First-order methods: For general nonconvex problems, gradient descent (GD) is known to find a stationary point in polynomial time [51]. In [42], it was proved that randomly initialized GD with a fixed step size converges to SoS solutions almost surely. The elegant analysis of [42], leveraging tools from the theory of dynamical systems (e.g., the stable manifold theorem), has been later extended in a number of follow-up works establishing the same type of second-order guarantees for a variety of first-order methods, including the proximal point algorithm, block coordinate descent, mirror descent [41]; the heavy-ball method and Nesterov's accelerated method [53]; block coordinate descent and alternating minimization [43]; and a primal-dual optimization procedure for solving linear equality constrained nonconvex optimization problems [33]. These results are all asymptotic in nature and it is unclear whether polynomial convergence rates can be obtained for these methods. In [23] it was actually proven that, even with fairly natural random initialization schemes and for nonpathological functions, GD can be significantly slowed down by saddle points, taking exponential time to escape. Recent work has analyzed variations of GD that include stochastic perturbations. It has been shown that when perturbations are incorporated into GD at each step the resulting algorithm can escape strict saddle points in polynomial time [26]; the same conclusion was earlier established in [54] for stochastic gradient methods, although without escape time guarantees. It has also been shown that episodic perturbations suffice; in particular, [35] introduced an algorithm that occasionally adds a perturbation to GD, and proved that the number of iterations to escape saddle points depends only polylogarithmically on dimension (i.e., it is nearly dimension independent). Fruitful follow-up results show that other first-order perturbed algorithms escape from strict saddle points efficiently [36, 45].

1.1.2. Distributed algorithms for (P) and guarantees. Distributed algorithms for convex instances of (P) have a long history; fewer results are available for nonconvex objectives. Since the focus on this paper is on nonconvex problems, next, we mainly comment on distributed algorithms for minimizing nonconvex objectives.

• **DGD and its variants:** DGD (and its variants) is unquestionably among the first and most studied decentralizations of the GD algorithm for (P) [49, 50]. The instance of DGD considered in this paper reads given $\mathbf{x}_i^0 \in \mathbb{R}^m$, $i \in [n]$,

$$\mathbf{x}_i^{\nu+1} = \sum_{j=1}^{n} D_{ij}\, \mathbf{x}_j^{\nu} - \alpha \nabla f_i(\mathbf{x}_i^{\nu}), \quad i \in [n], \tag{1.1}$$

where $\mathbf{x}_i^{\nu}$ is the agent i's estimate at iteration ν of the vector variable $\boldsymbol{\theta}$; the $\{D_{ij}\}_{i,j}$ are a suitably chosen set of nonnegative weights (cf. Assumption 3.1), matching the graph topology (i.e., $D_{ij} > 0$ if there is a link between node i and j, and $D_{ij} = 0$ otherwise); and $\alpha > 0$ is the step size. Roughly speaking, the update of each agent i in (1.1) is the linear combination of two components: (i) the gradient ∇f_i evaluated at the agent's latest 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 (also known as the consensus step) is instrumental to asymptotically enforcing agreement among the agents' local variables.

When each f_i in (P) is (strongly) convex, convergence of DGD is well 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, the available convergence guarantees are weaker. In [70] it was shown that if a constant step size is employed, every limit point $(\mathbf{x}_1^{\infty}, \ldots, \mathbf{x}_n^{\infty})$ of the sequence generated by (1.1) satisfies $\sum_{i=1}^{n} \nabla_{x_i} f_i(\mathbf{x}_i^{\infty}) = \mathbf{0}$; the limit points of agents' iterates are not consensual; asymptotic consensus is achieved only by using a diminishing step size. Since in general f_i are all different, such limit points are *not* critical points of F. Nothing is known about the connection of the critical points of $\sum_{i=1}^{n} f_i(\mathbf{x}_i)$ and those of F, *let alone its second-order guarantees.* A first contribution of this paper is to establish second-order guarantees of DGD (1.1) applied to (P) over undirected graphs.

Several extensions/variants of the vanilla DGD followed the seminal works [49, 50]. The projected (stochastic) DGD for nonconvex constrained instances of (P) was proposed in [11]; with a diminishing step size, the algorithm converges to a stationary solution of the problem (almost surely, if noisy instances of the local gradients are used). The extension of DGD to *digraphs* was studied in [47] for convex unconstrained optimization, and later extended in [63] to nonconvex objectives. The algorithm, termed push-sum DGD, combines a local gradient step with the push-sum algorithm [9]. When a diminishing step size is employed, push-sum DGD converges to an exact stationary solution of (P); and its noisy perturbed version almost surely converges to local minimizers, provided that F does not have any saddle point [63]. To our knowledge, no other guarantees are known for DGD-like algorithms in the nonconvex setting. In particular, it is unclear whether DGD (1.1) escapes strict saddles of F.

• *Gradient tracking-based methods.* To cope with the speed-accuracy dilemma of DGD, [21, 22] proposed a new class of distributed gradient-based methods that converge to an exact consensual solution of nonconvex (constrained) problems while using a *fixed step size.* The algorithmic framework, termed NEXT, introduces the idea of *gradient tracking* to correct the DGD direction and cancel the steady state error in it while using a fixed step size: each agent updates its own local variables along a surrogate direction that tracks the gradient ∇F of the entire objective (the same idea was proposed independently in [68] for convex unconstrained smooth problems). The generalization of NEXT to digraphs—the SONATA algorithm—was proposed in [62, 58, 59, 61], with [59, 61] proving convergence of the agents' iterates to consensual stationary solutions of nonconvex problems at a sublinear rate. *No second-order guarantees have been established for these methods.* Extensions of the SONATA family based on different choices of the weight matrices were later introduced in [66, 56] for convex smooth unconstrained problems. In this paper we consider the following family of distributed algorithms based on gradient tracking, which encompasses the majority of the above schemes (see, e.g., [59, section 5]), and refer to it as distributed optimization with gradient gracking (DOGT):

$$\mathbf{x}_i^{\nu+1} = \sum_{j=1}^{n} R_{ij}\mathbf{x}_j^{\nu} \; - \alpha\,\mathbf{y}_i^{\nu}, \tag{1.2}$$

$$\mathbf{y}_i^{\nu+1} = \sum_{j=1}^{n} C_{ij}\mathbf{y}_j^{\nu} + \nabla f_i\big(\mathbf{x}_i^{\nu+1}\big) - \nabla f_i\big(\mathbf{x}_i^{\nu}\big) \quad \text{(gradient tracking)}, \tag{1.3}$$

where $(R_{ij})_{i,j}$ and $(C_{ij})_{i,j}$ are suitably chosen nonnegative weights compliant with the graph structure (cf. Assumption 4.1); and $\mathbf{y}_i \in \mathbb{R}^m$ is an auxiliary variable, controlled by agent i via the update (1.3), which aims at tracking locally the gradient sum $\sum_i \nabla f_i(\mathbf{x}_i^{\nu})$. Overall, the update (1.3) in conjunction with the consensus step in (1.2) is meant to "correct" the local gradient direction $-\nabla f_i(\mathbf{x}_i^{\nu})$ (as instead used in the DGD algorithm) and thus nulls asymptotically the steady error $\nabla f_i(\mathbf{x}_i^{\nu}) - \nabla F(\mathbf{x}_i^{\nu})$. This permits the use of a constant step size α while still achieving exact consensus without penalizing the convergence rate. Another important difference between DOGT and DGD in (1.1) is that the former serves as a unified platform for distributed algorithms applicable over both undirected and directed graphs. Convergence of DOGT in the form (1.2)–(1.3) when F is nonconvex remains an open problem, let alone second-order guarantees. A second contribution of this paper is to fill this gap and provide a first- and second-order convergence analysis of DOGT.

• *Primal-dual distributed algorithms.* We conclude this literature review by commenting on distributed algorithms for nonconvex (P) using a primal-dual form [72, 32, 30]. Because of their primal-dual nature, all these schemes are implementable only over *undirected* graphs. In [72] a distributed approximate dual (sub)gradient algorithm, coupled with a consensus step, is introduced. Assuming zero-duality gap, the algorithm is proved to asymptotically find a pair of primal-dual solutions of an auxiliary problem, which however might not be critical points of F; also, consensus is not guaranteed. No rate analysis is provided. In [32], a proximal primal-dual algorithm is proposed; the algorithm, termed Prox-PDA, employs either a constant or an increasing penalty parameter (which plays the role of the step size); a sublinear convergence rate of a suitably defined primal-dual gap is proved. A perturbed version of Prox-PDA, P-Prox-PDA, was introduced in [30], which can also deal with nonsmooth convex, additive functions in the objective of (P). P-Prox-PDA converges to an ϵ-critical point (and thus also to *inexact* consensus), under a proper choice of the penalty parameters that depends on ϵ. A sublinear convergence rate is also proved. No second-order guarantees have been established for the above schemes. The only primal-dual algorithms we are aware of with provable convergence to SoS solutions is the one in [33], proposed for a linearly constrained nonconvex optimization problem. When linear constraints are used to enforce consensus, the primal-dual method [33] becomes distributed and applicable to problem (P), but only for undirected graphs (DOGT is instead implementable also over digraphs). Second-order guarantees of such a scheme are established under slightly stronger assumptions than those required for DOGT (cf. Remark 4.18, section 4.3.3). Finally, notice that, since [33] substantially differs from DGD and DOGT—the former is a primal-dual scheme while the latter are primal methods—the convergence analysis put forth in [33] is not applicable to DGD and DOGT. Since DGD and DOGT in their general form encompass two classic algorithms for distributed optimization, the open problem of their second-order properties leaves a significant gap in the literature.

1.2. Major results.

We establish for the first time second-order guarantees of DGD (1.1) and DOGT (1.2)–(1.3). The main results are summarized next.

1.2.1. DGD (1.1).

We prove the following:

(i) For a sufficiently small step size α, agents' iterates $\{\mathbf{x}^\nu\}$ generated by (1.1) converge to an $O(\alpha)$-critical point of F for all initializations; see Lemma 3.6; neighborhood convergence to critical points is also established (cf. Theorem 3.9). This complements the convergence results in [70].

(ii) The average sequence $\{\overline{\mathbf{x}}^\nu \triangleq (1/n)\sum_{i=1}^n \mathbf{x}_i^\nu\}$ converges almost surely to a neighborhood of an SoS solution of (P), where the probability is taken over the initializations; see Theorem 3.12.

To prove (ii), we employ a novel analysis, which represents a major technical contribution of this work. In fact, existing techniques developed to established second-order guarantees of the centralized GD are not readily applicable to DGD—roughly speaking, this is due to the fact that DGD (1.1) converges only to a neighborhood of critical points of F (fixed points of (1.1) are not critical points of F). We elaborate next on this challenge and outline our analysis.

The elegant roadmap developed in [42, 41] to establish second-order guarantees of the centralized GD builds on the stable manifold theorem: roughly speaking, fixed points of the gradient map corresponding to strict saddles of the objective function are "unstable" (more formally, the stable set[1] of strict saddles has zero measure), implying almost sure convergence of GD iterates to SoS points [41, Corollary 2]. It is known that the DGD iterates (1.1) can be interpreted as instances of the GD applied to the following auxiliary function [69, 70]: denoting $\mathbf{x} \triangleq [\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top]^\top$,

[1] Given $\mathcal{X} \subseteq \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}^m$, and the fixed-point iterate $\mathbf{x}^{\nu+1} = g(\mathbf{x}^\nu)$, the stable set of $\mathcal{X}$ is $\{\mathbf{x} : \lim_\nu g^\nu(\mathbf{x}) \in \mathcal{X}\}$, i.e., the set of initial points such that $\{\mathbf{x}^\nu\}$ converges to a member of $\mathcal{X}$.

$$L_\alpha(\mathbf{x}) \triangleq \underbrace{\sum_{i=1}^{n} f_i(\mathbf{x}_i)}_{\triangleq F_c(\mathbf{x})} + \frac{1}{2\alpha}\sum_{i=1}^{n}\sum_{i=j}^{n}(e_{ij} - D_{ij})\mathbf{x}_i^\top \mathbf{x}_j, \tag{1.4}$$

where $e_{ij} = 1$ if there is an edge in the graph between agent i and agent j; and $e_{ij} = 0$ otherwise. Using (1.4), (1.1) can be rewritten as: denoting $\mathbf{x}^\nu \triangleq [\mathbf{x}_1^{\nu\,\top}, \ldots, \mathbf{x}_n^{\nu\,\top}]^\top$,

$$\mathbf{x}^{\nu+1} = \mathbf{x}^\nu - \alpha \nabla L_\alpha(\mathbf{x}^\nu). \tag{1.5}$$

One can then apply the above argument (cf. [41, Corollary 2]) to (1.5) and readily establish the following result (see Theorem 3.6 for the formal statement).

Fact 1 (informal): For sufficiently small $\alpha > 0$, randomly initialized DGD (1.5) [and thus (1.1)] converges almost surely to second-order critical point of L_α.

Unfortunately, this result alone is not satisfactory, as no connection is known between the critical points of L_α and those of F (note that $L_\alpha : \mathbb{R}^{n\cdot m} \to \mathbb{R}$ whereas $F : \mathbb{R}^m \to \mathbb{R}$). To cope with this issue we prove the following two facts.

Fact 2 (informal): Every limit point $\overline{\mathbf{x}}^\infty$ of the average sequence $\overline{\mathbf{x}}^\nu = 1/n \sum_{i=1}^n \mathbf{x}_i^\nu$ can be made arbitrarily close to a critical point of F by using a sufficiently small $\alpha > 0$ (Theorem 3.9).

Fact 3 (informal): Whenever the limit point $\bar{\mathbf{x}}^\infty = 1/n \sum_{i=1}^n \mathbf{x}_i^\infty$ belongs to a sufficiently small neighborhood of a strict saddle of F, $\mathbf{x}^\infty = [\mathbf{x}_1^{\infty\top}, \ldots, \mathbf{x}_n^{\infty\top}]^\top$ must be a strict saddle of L_α (Proposition 3.10 and Corollary 3.11).

The above three facts will then ensure that, for sufficiently small $\alpha > 0$, with almost complete certainty, $\{\bar{\mathbf{x}}^\nu\}$ will not get trapped in a neighborhood of a strict saddle of F—as $\mathbf{x}^\infty$ would be a strict saddle of L_α—thus landing in an neighborhood of an SoS solution of (P).

Facts 2 and 3 above are proved under a regularity condition on F which recalls (albeit slightly weaker than) [28]. Roughly speaking, the gradient flow over some *annulus* must be uniformly positively correlated with any outward (from the origin) direction (cf. Assumption 2.4). This condition is quite mild and is satisfied by functions arising, e.g., from several machine learning applications, including distributed principal component analysis (PCA), matrix sensing, and binary classification problems; see section 2 for more details. Furthermore, this condition is also sufficient to prove convergence of DGD without assuming the objective function to be globally L-smooth (but just locally L-smooth, LC^1 for short), a requirement that instead is common to existing (first-order) convergence conditions of DGD. Notice that the loss functions arising from many of the aforementioned machine learning problems are not globally L-smooth.

1.2.2. DOGT (1.2)–(1.3). For DOGT, we establish the following three results.

(i) When F is nonconvex and the graph is either undirected or directed, it is proved that every limit point of the sequence generated by DOGT is a critical point of F. Furthermore, a merit function, measuring distance of the iterates from stationarity, and consensus disagreement is introduced, and proved to vanish at a sublinear rate; see Theorem 4.5. This extends convergence results [56, 66], established only for convex functions. To deal with nonconvexity, our analysis builds on a novel Lyapunov-like function (cf. (4.20)), which properly combines optimization error dynamics, consensus, and tracking disagreements. While these three terms alone do not "sufficiently" decrease along the iterates—as local optimization and consensus/tracking steps might act as competing forces—a suitable combination of them, as captured by the Lyapunov function, does monotonically decrease.

(ii) When F satisfies the Kurdyka–Łojasiewicz (KŁ) property [40, 39] at any of its critical points, convergence of the entire sequence to a critical point of F is proved (cf. Theorem 4.7), and a convergence rate is provided (cf. Theorem 4.8). Although inspired by [7], establishing similar convergence results (but no rate analysis) for centralized first-order methods, our proof follows a different path building on the descent of the Lyapunov function introduced in (i), which does not satisfy [7, conditions H1–H2]); see section 4.2 for details.

(iii) The sequence of iterates generated by DOGT is shown to converge to SoS solutions of (P) almost surely, when initial points are randomly drawn from a suitably chosen linear subspace; see Theorem 4.17. This result is proved for undirected and directed networks. The proofs build on the stable manifold theorem, based upon the interpretation of DOGT dynamics as fixed-point iterates of a suitably defined map. The challenge in finding such a map is ensuring that the stable set of its undesirable fixed-points—those associated with the strict saddles of F—has measure zero in the subspace where the initialization of DOGT takes place. Note that this subspace is not full dimensional.

While our paper was under review after its initial arXiv posting [18] and its companion conference version [19], we became aware of a follow-up line of related works [64, 65]. These schemes study second-order guarantees of variations of the DGD algorithm (1.1). Specifically, [64, 65] studied the behavior of (the adapt-then-combine version of) DGD wherein exact gradients are replaced by stochastic approximations; the algorithm is proved to return approximate SoS points in a polynomial number of iterations. Finally, [44] proposed a variant of DGD to solve the distributed low-rank matrix factorization problem and they prove almost sure convergence to global minima of the problem.

1.3. Paper organization. The rest of the paper is organized as follows. The main assumptions on the optimization problem and network are introduced in section 2. Section 3 studies guarantees of DGD over undirected graphs, along the following steps: (i) existing convergence results are discussed in section 3.1; (ii) section 3.2 studies convergence to a neighborhood of a critical point of F; and (iii) section 3.3 establishes second-order guarantees. DOGT algorithms are studied in section 4 along the following steps: (i) subsequence convergence is proved in section 4.1; (ii) section 4.2 establishes global convergence under the KŁ property of F; and (iii) section 4.3 derives second-order guarantees over undirected and directed graphs. Finally, section 5 presents some numerical results.

1.4. Notation. The set of nonnegative integers is denoted by $\mathbb{N}_+$ and we use $[n]$ as a shorthand for $\{1, 2, \ldots, n\}$. All vectors are denoted by bold letters and assumed to be column vectors; given a vector $\mathbf{x}$, $||\mathbf{x}||$ denotes the ℓ_2 norm of $\mathbf{x}$; any other specific vector norm is subscripted accordingly. $\mathbf{x}$ is called *stochastic* if all its components are nonnegative and sum to one; and $\mathbf{1}$ is the vector of all ones (we write $\mathbf{1}_m$ for the m–dimensional vector, if the dimension is not clear from the context). Given sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^m$, we denote $\mathcal{X} \setminus \mathcal{Y} \triangleq \{x \in \mathcal{X} : x \notin \mathcal{Y}\}$, $\overline{\mathcal{X}} \triangleq \mathbb{R}^m \setminus \mathcal{X}$ (complement of $\mathcal{X}$), and $\mathbf{x} + \mathcal{X} = \{\mathbf{x} + \mathbf{z} : \mathbf{z} \in \mathcal{X}\}$. $\mathcal{V}_{\mathbf{x}}$ and $\mathcal{B}(\mathbf{x}, r)^d$ denote a neighborhood of $\mathbf{x}$ and the d-dimensional closed ball of radius $r > 0$ centered at $\mathbf{x}$, respectively; when the ball is centered at $\mathbf{0}$, we will write $\mathcal{B}_r^d$. We further define an annulus by $\mathcal{S}_{r,\epsilon} \triangleq \mathcal{B}_r^d \setminus \mathcal{B}_{r-\epsilon}^d$ with some $r > \epsilon > 0$. The Euclidean projection of $\mathbf{x} \in \mathbb{R}^m$ onto the convex closed set $\mathcal{X} \subseteq \mathbb{R}^m$ is $\text{proj}_{\mathcal{X}}(\mathbf{x}) \triangleq \arg\min_{\mathbf{y} \in \mathcal{X}} ||\mathbf{x} - \mathbf{y}||$. The sublevel set of a function U at u is denoted by $\mathcal{L}_U(u) \triangleq \{\mathbf{x} : U(\mathbf{x}) \leq u\}$.

Matrices are denoted by capital bold letters; A_{ij} is the the (i,j)th element of $\mathbf{A}$; $\mathcal{M}_m(\mathbb{R})$ 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; 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 $\operatorname{spec}(\mathbf{M})$ and its spectral radius is $\operatorname{spradii}(\mathbf{M})\triangleq\max\{|\lambda|:\lambda\in\operatorname{spec}(\mathbf{M})\}$; the spectral norm is $||\mathbf{M}||\triangleq\max_{||\mathbf{x}||\neq 0}||\mathbf{M}\mathbf{x}||/||\mathbf{x}||$, and any other matrix norm is subscripted accordingly. Finally, the minimum (resp., maximum) singular value are denoted by $\sigma_{\min}(\mathbf{M})$ (resp., $\sigma_{\max}(\mathbf{M})$) and minimum (resp., maximum) eigenvalue by $\lambda_{\min}(\mathbf{M})$ (resp., $\lambda_{\max}(\mathbf{M})$).

The sequence generated by DGD (and DOGT) depends on the step-size α and the initialization $\mathbf{x}^0$. When necessary, we write $\{\mathbf{x}^\nu(\alpha,\mathbf{x}^0)\}$ for $\{\mathbf{x}^\nu\}$.

Throughout the paper, we assume that all the probability measures are absolutely continuous with respect to the Lebesgue measure.

2. Problem and network setting.

In this section, we introduce the various assumptions on the functions f_i and the graph, under which our results are derived.

Assumption 2.1 (on problem P). Given problem (P),

(i) f_i $(\forall i)$ is $r+1$ times continuously differentiable for some $r\geq 1$, and ∇f_i is L_i-Lipschitz continuous. Denote $L_{\max}\triangleq\max_i\ L_i$;

(ii) F is coercive.

For some convergence results of DGD we need the following slightly stronger condition.

Assumption 2.1′. Assumption 2.1(i) is satisfied and (ii) each f_i is coercive.

We also make the blanket assumption that each agent i knows only its own f_i but not the rest of the objective function.

Note that Assumption 2.1, particularly the global Lipschitz gradient continuity of f_i, is quite standard in the literature. Motivated by some applications of interest (see examples below), we will also prove convergence of DGD under LC^1 only and the mild condition (2.2) below (cf. Assumption 2.4). Although strictly not necessary, coercivity in Assumptions 2.1 and 2.1′ simplifies some of our derivations; our results can be extended under the weaker assumption that (P) has a solution.

Some of the convergence results of DGD and DOGT are established under the assumption that F satisfies the KŁ inequality [39, 40].

DEFINITION 2.2 (KŁ property). *Given a function $U:\mathbb{R}^N\to\mathbb{R}\cup\{+\infty\}$, we set* $[a<U<b]\triangleq\{\mathbf{z}\in\mathbb{R}^N\ :\ a<U(\mathbf{z})<b\}$, *and*

(a) *the function U has the KŁ property at $\acute{\mathbf{z}}\in\operatorname{dom}\partial U$ 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:*

$$\phi'\left(U(\mathbf{z})-U(\acute{\mathbf{z}})\right)\operatorname{dist}(0,\partial U(\mathbf{z}))\geq 1; \tag{2.1}$$

(b) *a proper lower-semicontinuous function U is called KŁ if it satisfies the KŁ inequality at every point in* dom ∂U*.*

Many problems involve functions satisfying the KŁ inequality; real semialgebraic functions provide a very rich class of functions satisfying the KŁ; see [6] for a thorough discussion.

Second-order guarantees of DGD are obtained under the following two extra assumptions; Assumption 2.3 is quite standard and widely used in the literature to establish second-order guarantees of centralized algorithms (e.g., [26, 35, 52, 14, 15, 3, 17]) as well as of distributed algorithms [33, 64, 65]. Assumption 2.4 is introduced for this paper and commented on below.

Assumption 2.3. 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, where F_c is defined in (1.4).

Assumption 2.4. (i) Each f_i is LC^1; and (ii) there exist $0 < \epsilon < R$ and $\delta > 0$ such that

$$\inf_{\boldsymbol{\theta} \in \mathcal{S}_{R,\epsilon}} \langle \nabla f_i(\boldsymbol{\theta}), \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\| \rangle \geq \delta \quad \forall i \in [n]. \tag{2.2}$$

Roughly speaking, the condition above postulates that the gradient $\nabla f_i(\boldsymbol{\theta})$ is positively correlated with any radial direction $\boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|$ for all $\boldsymbol{\theta}$ in the annulus $\mathcal{S}_{R,\epsilon}$. A slightly more restrictive form of the above assumption has appeared in [28, Assumption A3]. Many functions of practical interest satisfy this assumption; some examples arising from machine learning applications are listed below.

Distributed PCA [25]: Given matrices $\mathbf{M}_i \in \mathbb{R}^{m\times m}$, $i \in [n]$, the distributed PCA problem is to find the leading eigenvector of $\sum_{i=1}^n \mathbf{M}_i$ by solving

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \quad \frac{1}{4}\left\|\boldsymbol{\theta}\boldsymbol{\theta}^\top - \sum_{i=1}^n \mathbf{M}_i\right\|_F^2, \tag{2.3}$$

which can be rewritten in the form (P).

Phase retrieval [16]: Let $\{(\mathbf{a}_i, y_i)\}_{i=1}^n$, with $\mathbf{a}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ such that $y_i = \mathbf{a}_i^\top \mathbf{M}^* \mathbf{a}_i = (\mathbf{a}_i^\top \boldsymbol{\theta}^*)^2$ and $\mathbf{M}^* = \boldsymbol{\theta}^*\boldsymbol{\theta}^{*\top} \in \mathbb{R}^{m\times m}$. The phase retrieval problem reads

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \quad \frac{1}{4}\sum_{i=1}^n \left(||\mathbf{a}_i^\top\boldsymbol{\theta}||^2 - y_i\right)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||^2, \tag{2.4}$$

where $\lambda > 0$ is a given parameter.

Matrix sensing [16]: Let $\{(\mathbf{A}_i, y_i)\}_{i=1}^n$, with $\mathbf{A}_i \in \mathbb{R}^{m\times m}$ and $y_i \in \mathbb{R}$ such that $y_i = \langle \mathbf{A}_i, \mathbf{M}^*\rangle$ and $\mathbf{M}^* = \boldsymbol{\Theta}^*\boldsymbol{\Theta}^{*\top} \in \mathbb{R}^{m\times m}$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{m\times r}$. The matrix sensing problem reads

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{m\times r}} \quad \frac{1}{4}\sum_{i=1}^n \left(\left\langle \mathbf{A}_i, \boldsymbol{\Theta}\boldsymbol{\Theta}^\top \right\rangle - y_i\right)^2 + \frac{\lambda}{2}\left\|\boldsymbol{\Theta}\right\|_F^2, \tag{2.5}$$

where $\lambda > 0$ is a given parameter.

Gaussian mixture model [44]: Let $\{\mathbf{z}_i\}_{i=1}^n$ be n points drawn from a mixture of q Gaussian distributions, i.e., $\mathbf{z}_i \sim \sum_{d=1}^q \mathcal{N}(\boldsymbol{\mu}_d^*, \boldsymbol{\Sigma})$, where $\mathcal{N}(\boldsymbol{\mu}_d^*, \boldsymbol{\Sigma})$ is the Gaussian distribution with mean $\boldsymbol{\mu}_d^* \in \mathbb{R}^m$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$. The goal is to estimate the mean values $\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_q^*$ by solving the maximum likelihood problem

$$\min_{\{\boldsymbol{\theta}_d\in\mathbb{R}^m\}_{d=1}^q} -\sum_{i=1}^n \log\left(\sum_{d=1}^q \phi_m(\mathbf{z}_i - \boldsymbol{\theta}_d)\right) + \frac{\lambda}{2}\left\|\boldsymbol{\theta}_d\right\|^2, \tag{2.6}$$

where $\phi_m(\boldsymbol{\theta})$ is the multivariate normal distribution with $\mathbf{0}$ mean and covariance $\boldsymbol{\Sigma}$.

Bilinear logistic regression [24]: The description of the problem along with some numerical results can be found in section 5.2.

Artificial neuron [8, 71]: Let $\{(\mathbf{s}_i, \xi_i)\}_{i=1}^n$ be n samples with $\mathbf{s}_i \in \mathbb{R}^m$, $\xi_i \in \mathbb{R}$, and measurement model $\xi_i = \sigma(\mathbf{s}_i^\top \boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is the optimal weight and $\sigma(\cdot)$ is a *transfer* function; e.g., the logistic regression function $\sigma(\theta) = 1/(1+\exp(-\theta))$. The goal is to estimate $\boldsymbol{\theta}^*$ by solving

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \quad \sum_{i=1}^n \frac{1}{2n}\left[\left(\xi_i - \sigma(\mathbf{s}_i^\top\boldsymbol{\theta})\right)^2 + \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|^2\right], \tag{2.7}$$

where $\lambda > 0$ is a given parameter. Further binary classification models satisfying Assumption 2.4 include f_i functions such as [71]

$$\begin{aligned} f_i(\boldsymbol{\theta}) &= 1 - \tanh \xi_i \mathbf{s}_i^\top \boldsymbol{\theta} + \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|^2, \\ f_i(\boldsymbol{\theta}) &= \left(1 - \sigma(\xi_i \mathbf{s}_i^\top\boldsymbol{\theta})\right)^2 + \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|^2, \\ f_i(\boldsymbol{\theta}) &= -\ln\sigma(\xi_i\mathbf{s}_i^\top\boldsymbol{\theta}) + \ln\sigma(\xi_i\mathbf{s}_i^\top\boldsymbol{\theta} + \mu) + \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|^2, \end{aligned} \tag{2.8}$$

where $\lambda > 0$ and $\mu > 0$ are given parameters.

In all these examples, Assumption 2.4 is satisfied for any sufficiently large R and $R-\epsilon$; the proof can be found in Appendix A.1. Note that many of the functions listed above are not L-smooth on their entire domain, violating thus (part of) Assumption 2.1(i). Motivated by these examples, we will extend existing convergence results of DGD, replacing Assumption 2.1(i) with Assumption 2.4.

Network model: The network 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 a link directed from agent i to agent j. The in-neighborhood of agent i is defined as $\mathcal{N}_i^{\text{in}} = \{j|(j,i) \in \mathcal{E}\} \cup \{i\}$ 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 $\mathcal{N}_i^{\text{out}} = \{j|(i,j) \in \mathcal{E}\} \cup \{i\}$. 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 $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$, 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 2.5 (on the network). The graph (resp., digraph) $\mathcal{G}$ is connected (resp., strongly connected).

3. The DGD algorithm.

Consider Problem (P) and assume that the network is modeled as an undirected graph $\mathcal{G}$. As described in section 1, the DGD algorithm is based on a decentralization of GD as described in (1.1). It is convenient to rewrite the update (1.1) in the matrix/vector form: Using the definition of *aggregate* function $F_c(\mathbf{x})$ (cf. (1.4)) and $\mathbf{x}^\nu \triangleq [\mathbf{x}_1^{\nu\top}, \ldots, \mathbf{x}_n^{\nu\top}]^\top$, we have

$$\mathbf{x}^{\nu+1} = \mathbf{W}_D\, \mathbf{x}^\nu - \alpha\nabla F_c(\mathbf{x}^\nu), \tag{3.1}$$

given $\mathbf{x}^0 \in \mathbb{R}^{mn}$, where $\mathbf{W}_D \triangleq \mathbf{D} \otimes \mathbf{I}_m$ and $\mathbf{D} \in \mathcal{M}_n(\mathbb{R})$ satisfying the following assumption.

Assumption 3.1. $\mathbf{D} \in \mathcal{M}_n(\mathbb{R})$ is nonnegative, doubly stochastic, and compliant to $\mathcal{G}$, i.e., $D_{ij} > 0$ if and only if $(j,i) \in \mathcal{E}$, and $D_{ij} = 0$ otherwise.

3.1. Existing convergence results. Convergence of DGD applied to the nonconvex problem (P) has been established [69, 70], and is summarized below.

THEOREM 3.2 (see [69, 70]). *Let Assumptions* 2.1′, 2.5 *hold. Given arbitrary* $\mathbf{x}^0 \in \mathbb{R}^{mn}$ *and* $0 < \alpha < \alpha_{\max} \triangleq \sigma_{\min}(\mathbf{I} + \mathbf{D})/L_c$, *let* $\{\mathbf{x}^\nu\}$ *be the sequence generated by the DGD algorithm* (3.1) *under Assumption* 3.1. *Then* $\{\mathbf{x}^\nu\}$ *is bounded and*

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

$$\|\mathbf{x}_i^\nu - \bar{\mathbf{x}}^\nu\| \leq (\sigma_2)^\nu \|\mathbf{x}_i^0\| + \frac{\alpha H}{1 - \sigma_2},$$

where $\sigma_2 < 1$ *is the second largest singular value of* $\mathbf{D}$, *and* H *is a universal upper bound of* $\{||\nabla F_c(\mathbf{x}^\nu)||\}$;

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

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

Although L-smoothness of f_i's is a common assumption in the literature, above convergence results can also be established without this condition but under Assumption 2.4; see Remark 3.13 and Appendix A.2 for details.

Since (3.1) is the gradient update applied to L_α (cf. (1.5)), nonconvergence of the DGD algorithm to strict saddle points of L_α can be established by applying [41, Corollary 2] to (1.5); the statement is given in Theorem 3.4 below. The following extra assumption on the weight matrix $\mathbf{D}$ is needed.

Assumption 3.3. The matrix $\mathbf{D} \in \mathcal{M}_n(\mathbb{R})$ is nonsingular.

THEOREM 3.4. *Consider problem* (P), *under Assumptions* 2.1′, 2.5, *and further assume that each* f_i *is a KL function. Let* $\{\mathbf{x}^\nu\}$ *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 Assumptions* 3.1 *and* 3.3. *Then, the stable set of strict saddles has measure zero. Therefore,* $\{\mathbf{x}^\nu\}$ *convergences almost surely to an SoS solution of* L_α, *where the probability is taken over the random initialization* $\mathbf{x}^0 \in \mathbb{R}^{mn}$.

As anticipated in section 1.2.1, the above second-order guarantees are not satisfactory as they do not provide any information on the behavior of DGD near critical points of F, including the strict saddles of F. In the following, we fill this gap. We first show that the DGD algorithm convergences to a neighborhood of the critical points of F, whose size is controlled by the step size $\alpha > 0$ (cf. section 3.2). Then, we prove that, for sufficiently small $\alpha > 0$, such critical points are almost surely SoS solutions of (P), where the randomization is taken on the initial point (cf. section 3.3).

3.2. DGD converges to a neighborhood of critical points of *F*. Let us begin with introducing the definition of ϵ-critical points of F.

DEFINITION 3.5. *A point* $\boldsymbol{\theta} \in \mathbb{R}^m$ *such that* $||\nabla F(\boldsymbol{\theta})|| \leq \varepsilon$ *with* $\varepsilon > 0$, *is called an* ε*-critical point of* F. *The set of* ε*-critical points of* F *is denoted by* $\text{crit}_\varepsilon F$.

In this section, we prove that when the step size is sufficiently small and DGD is initialized in a compact set, the iterates $\{\mathbf{x}_i^\nu\}$, $i \in [n]$, converge to an arbitrarily small neighborhood of critical points of F—the result is formally stated in Theorem 3.9. Roughly speaking, this is proved chaining the following intermediate results:

(i) Lemma 3.6: Every limit point of DGD is an $\mathcal{O}(\alpha)$-critical point of F.

(ii) Lemma 3.7: Every sequence generated by DGD for given $\alpha > 0$ and initialization in a compact set, is enclosed in some compact set, for all $\alpha \downarrow 0$.

(iii) Lemma 3.8: Any ϵ-critical point of F achievable by DGD is arbitrarily close to a critical point of F, when ϵ is sufficiently small.

Lemma 3.6 implies that, for any given $\epsilon > 0$, one can find an arbitrarily small $\alpha > 0$ so that every limit point of each $\{\mathbf{x}_i^\nu\}$ (whose existence is guaranteed by Lemma 3.7) is an ϵ-critical point of F. Finally, Lemma 3.8 guarantees that every such ϵ-critical point can be made arbitrarily close to a critical point of F as $\epsilon \downarrow 0$. The proof of the above three lemmas follows.

LEMMA 3.6. *Let Assumptions* 2.1$'$ *and* 2.5 *hold. Given arbitrary* $\mathbf{x}^0 \in \mathbb{R}^{mn}$ *and* $0 < \alpha < \sigma_{\min}(\mathbf{I}+\mathbf{D})/L_c$, *every limit point* $\mathbf{x}^\infty = [\mathbf{x}_1^{\infty\top}, \ldots, \mathbf{x}_n^{\infty\top}]^\top$ *of* $\{\mathbf{x}^\nu\}$ *generated by the DGD algorithm satisfies* $\bar{\mathbf{x}}^\infty \in \mathrm{crit}_{K'\alpha} F$ *with* $\bar{\mathbf{x}}^\infty \triangleq (1/n)\sum_{i=1}^n \mathbf{x}_i^\infty$ *and* $K' = n\sqrt{n}L_c H/\,(1-\sigma_2)$, *where* H *and* σ_2 *are defined in Theorem* 3.2.

Proof. By Theorem 3.2(ii), $(\mathbf{1}\otimes\mathbf{I})^\top \nabla L_\alpha(\mathbf{x}^\infty) = \mathbf{0}$, which using (1.4) and the column stochasticity of $\mathbf{D}$ yields $(\mathbf{1}\otimes\mathbf{I})^\top\, \nabla F_c(\mathbf{x}^\infty) = \mathbf{0}$. Hence,

$$\begin{aligned}\|\nabla F(\bar{\mathbf{x}}^\infty)\| &= \left\|(\mathbf{1}\otimes\mathbf{I})^\top \left(\nabla F_c(\mathbf{1}\otimes\bar{\mathbf{x}}^\infty) - \nabla F_c(\mathbf{x}^\infty)\right)\right\| \\ &\le L_c\sqrt{n}\,\|\mathbf{x}^\infty - \mathbf{1}\otimes\bar{\mathbf{x}}^\infty\| \overset{(a)}{\le} \alpha\cdot\frac{n\sqrt{n}L_c H}{1-\sigma_2},\end{aligned} \tag{3.2}$$

where in (a) we used Theorem 3.2(i). □

To proceed, we limit DGD initialization to $\mathbf{x}_i^0 \in \mathcal{X}_i$, $i \in [n]$, where $\mathcal{X}_i^0 \subseteq \mathbb{R}^m$ is some compact set with positive Lebesgue measure.

LEMMA 3.7. *Consider problem* (P)*, under Assumptions* 2.1$'$, 2.4, *and* 2.5. *Let* $\{\mathbf{x}^\nu(\alpha,\mathbf{x}^0)\}$ *be any sequence generated by DGD under Assumption* 3.1, *with step size* α *and initialization* $\mathbf{x}^0$. *Then, there exists a bounded set* $\mathcal{Y}$ *such that* $\{\mathbf{x}^\nu(\alpha,\mathbf{x}^0)\} \subseteq \mathcal{Y}$ *for all* $0 < \alpha \le \alpha_{\max} = \sigma_{\min}(\mathbf{I}+\mathbf{D})/L_c$ *and* $\mathbf{x}_i^0 \in \mathcal{X}_i \subseteq \mathcal{B}_R^m, i \in [n]$, *where* R *is defined in Assumption* 2.4.

Proof. We proceed by induction. For the sake of notation, throughout the proof, we will use for $\mathbf{x}^\nu(\alpha,\mathbf{x}^0)$ the shorthand $\mathbf{x}^\nu$. Define $h \triangleq \max_{i\in[n],\boldsymbol{\theta}\in\mathcal{B}_R^m} ||\nabla f_i(\boldsymbol{\theta})||$. By assumption, there holds $\mathbf{x}_i^0 \in \mathcal{B}_R^m$ for all i. Suppose $\mathbf{x}_i^\nu \in \mathcal{B}_R^m$ for all i. If $\mathbf{x}_i^\nu \in \mathcal{B}_{R-\epsilon}^m$ and $\alpha \le \epsilon D_{ii}/h$, then $\mathbf{x}_i^\nu - \frac{\alpha}{D_{ii}}\nabla f_i(\mathbf{x}_i^\nu) \in \mathcal{B}_R^m$, since

$$\left\|\mathbf{x}_i^\nu - \frac{\alpha}{D_{ii}}\nabla f_i(\mathbf{x}_i^\nu)\right\| \le \|\mathbf{x}_i^\nu\| + \frac{\alpha}{D_{ii}}\|\nabla f_i(\mathbf{x}_i^\nu)\| \le R - \epsilon + \frac{\alpha h}{D_{ii}}. \tag{3.3}$$

If $\mathbf{x}_i^\nu \in \mathcal{S}_{R,\epsilon}$ and $\alpha \le 2D_{ii}\delta(R-\epsilon)/h^2$, then $\mathbf{x}_i^\nu - \frac{\alpha}{D_{ii}}\nabla f_i(\mathbf{x}_i^\nu) \in \mathcal{B}_R^m$, since

$$\begin{aligned}\left\|\mathbf{x}_i^\nu - \frac{\alpha}{D_{ii}}\nabla f_i(\mathbf{x}_i^\nu)\right\|^2 &= \|\mathbf{x}_i^\nu\|^2 - \frac{2\alpha||\mathbf{x}_i^\nu||}{D_{ii}}\left\langle\frac{\mathbf{x}_i^\nu}{||\mathbf{x}_i^\nu||}, \nabla f_i(\mathbf{x}_i^\nu)\right\rangle + \frac{\alpha^2}{D_{ii}^2}\|\nabla f_i(\mathbf{x}_i^\nu)\|^2 \\ &\le R^2 - \frac{2\alpha\delta(R-\epsilon)}{D_{ii}} + \frac{\alpha^2 h^2}{D_{ii}^2}.\end{aligned} \tag{3.4}$$

By agents' updates $\mathbf{x}_i^{\nu+1} = \sum_{j\neq i} D_{ij}\mathbf{x}_j^\nu + D_{ii}(\mathbf{x}_i^\nu - \frac{\alpha}{D_{ii}}\nabla f_i(\mathbf{x}_i^\nu))$ and convexity of the norm, we conclude that if $\mathbf{x}_i^\nu \in \mathcal{B}_R^m$ for all i, and $0 < \alpha \le \alpha_b \triangleq \min_i \min\{\epsilon D_{ii}/h, 2D_{ii}\delta(R-\epsilon)/h^2\}$, then $\mathbf{x}_i^{\nu+1} \in \mathcal{B}_R^m$. This proves that, for $\alpha \in (0,\alpha_b]$, any sequence $\{\mathbf{x}_i^\nu\}$ initialized in $\mathcal{B}_R^m$ lies in $\mathcal{B}_R^m$ for all i.

We prove now the same result for $\alpha \in [\alpha_b, \sigma_{\min}(\mathbf{I}+\mathbf{D})/L_c]$. Note that since each f_i is coercive (cf. Assumption 2.1$'$(ii)), any sublevel set of L_α is compact. Also, since $\{L_\alpha(\mathbf{x}^\nu)\}$ is nonincreasing for all $\alpha \in (0, \sigma_{\min}(\mathbf{I}+\mathbf{D})/L_c]$ (cf. [69, Lemma 2]), then

$\{\mathbf{x}^\nu\} \subseteq \mathcal{L}_{L_\alpha}(F_c(\mathbf{x}^0) + \frac{1}{2\alpha}||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}})$ and, furthermore,

(3.5)
$$\begin{aligned}&\mathcal{L}_{L_\alpha}\left(F_c(\mathbf{x}^0) + \frac{1}{2\alpha}||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}}\right) \subseteq \mathcal{L}_{L_\alpha}\left(F_c(\mathbf{x}^0) + \frac{1}{2\alpha_b}||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}}\right)\\ &\subseteq \mathcal{L}_{F_c}\left(F_c(\mathbf{x}^0) + \frac{1}{2\alpha_b}||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}}\right) \subseteq \mathcal{L}_{F_c}\left(\max_{\mathbf{x}_i^0\in\mathcal{B}_R^m, i\in[n]}\left\{F_c(\mathbf{x}^0) + \frac{1}{2\alpha_b}||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}}\right\}\right).\end{aligned}$$

Since $||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}} \le 2||\mathbf{x}^0||^2$, it follows that

$$\mathcal{L}_{L_\alpha}\left(F_c(\mathbf{x}^0) + \frac{1}{2\alpha}||\mathbf{x}^0||^2_{\mathbf{I}-\mathbf{W}}\right) \subseteq \underbrace{\mathcal{L}_{F_c}\left(\max_{\mathbf{x}_i^0\in\mathcal{B}_R^m, i\in[n]}\left\{\sum_{i=1}^n f_i(\mathbf{x}_i^0)\right\} + \frac{R^2}{\alpha_b}\right)}_{\triangleq\bar{\mathcal{L}}}. \tag{3.6}$$

The statement of the lemma holds with $\mathcal{Y} = \bar{\mathcal{L}} \cup \prod_{i=1}^n \mathcal{B}_R^m$. □

The following lemma shows that any ϵ-critical point of F achievable by DGD (i.e., any point in $\mathrm{crit}_\varepsilon F \cap \bar{\mathcal{Y}}$) can be made arbitrarily close to a critical point of F, when $\epsilon > 0$ (and thus $\alpha > 0$) is sufficiently small.

LEMMA 3.8. *Suppose $F : \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable. For any given compact set $\bar{\mathcal{Y}} \subseteq \mathbb{R}^m$, there holds*

$$\lim_{\varepsilon\to 0} \max_{\mathbf{q}\in\mathrm{crit}_\varepsilon F\cap\bar{\mathcal{Y}}} \mathrm{dist}(\mathbf{q}, \mathrm{crit}\ F) = 0. \tag{3.7}$$

Proof. We prove the lemma by contradiction. Suppose

$$\limsup_{\varepsilon\to 0} \max_{\mathbf{q}\in\mathrm{crit}_\varepsilon F\cap\bar{\mathcal{Y}}} \mathrm{dist}(\mathbf{q}, \mathrm{crit}\ F) = \gamma > 0. \tag{3.8}$$

Then, one can construct $\{\mathbf{q}^\nu\}$ with $\mathbf{q}^\nu \in \mathrm{crit}_{1/\nu}F \cap \bar{\mathcal{Y}}$ such that $\mathrm{dist}(\mathbf{q}^\nu, \mathrm{crit}F) \ge \gamma$ for all $\nu \in \mathbb{N}$. Since ∇F is continuous, $\mathrm{crit}_1 F$ is closed and $\mathrm{crit}_1 F \cap \bar{\mathcal{Y}}$ is compact. Note that $\{\mathbf{q}^\nu\} \subseteq \mathrm{crit}_1 F \cap \bar{\mathcal{Y}}$, which ensures $\{\mathbf{q}^\nu\}$ is bounded. Let $\{\mathbf{q}^{t_\nu}\}$ be a convergent subsequence of $\{\mathbf{q}^\nu\}$; its limit point $\mathbf{q}^\infty$ satisfies $\mathrm{dist}(\mathbf{q}^\infty, \mathrm{crit}\ F) \ge \gamma$. By construction, for any $\acute{\nu} \in \mathbb{N}$, $\{\mathbf{q}^{t_\nu}\}$ eventually settles in $\mathrm{crit}_{1/\acute{\nu}}\ F \cap \bar{\mathcal{Y}}$, thus $\mathbf{q}^\infty \in \mathrm{crit}_{1/\acute{\nu}}\ F \cap \bar{\mathcal{Y}}$. This means that $||\nabla F(\mathbf{q}^{\acute{\nu}})|| \le 1/\acute{\nu}$, for all $\acute{\nu} \in \mathbb{N}$, implying $||\nabla F(\mathbf{q}^\infty)|| = 0$. Hence $\mathrm{dist}(\mathbf{q}^\infty, \mathrm{crit}\ F) = 0$, which contradicts (3.8). □

We can now combine Lemmas 3.6–3.8 with Theorem 3.2(i) and state the main result of this section.

THEOREM 3.9. *Let Assumptions* 2.1′, 2.4, *and* 2.5 *hold. Let $\epsilon > 0$. There exists $\bar{\alpha} > 0$ (which depends on ϵ) such that with any initialization $\mathbf{x}_i^0 \in \mathcal{X}_i^0 \subseteq \mathcal{B}_R^m$ ($R > 0$ is defined in Assumption* 2.4*), $i \in [n]$, and any step size $0 < \alpha \le \bar{\alpha}$, all the limit points $\mathbf{x}^\infty(\alpha, \mathbf{x}^0) = [\mathbf{x}_1^\infty(\alpha, \mathbf{x}^0)^\top, \ldots, \mathbf{x}_n^\infty(\alpha, \mathbf{x}^0)^\top]^\top$ of the sequence $\{\mathbf{x}^\nu(\alpha, \mathbf{x}^0)\}$, generated by DGD satisfy*

$$\mathrm{dist}\big(\bar{\mathbf{x}}^\infty(\alpha, \mathbf{x}^0), \mathrm{crit\,F}\big) < \epsilon \quad \textit{and} \quad \big\|\mathbf{x}^\infty(\alpha, \mathbf{x}^0) - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty(\alpha, \mathbf{x}^0)\big\| < \epsilon, \tag{3.9}$$

where $\bar{\mathbf{x}}^\infty(\alpha, \mathbf{x}^0) \triangleq (1/n) \sum_{i=1}^n \mathbf{x}_i^\infty(\alpha, \mathbf{x}^0)$.

Proof. Combining Lemmas 3.6–3.8 proves that there exists some $\alpha_1 > 0$ such that $\mathrm{dist}(\bar{\mathbf{x}}^\infty(\alpha, \mathbf{x}^0), \mathrm{crit\,F}) < \epsilon$ for all $\alpha \le \alpha_1$. In addition, Theorem 3.2(i) with $H = \sup_{\mathbf{x}\in\mathcal{Y}} F_c(\mathbf{x})$, implies that there exists some $\alpha_2 > 0$ such that $||\mathbf{x}^\infty(\alpha, \mathbf{x}^0) - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty(\alpha, \mathbf{x}^0)|| < \epsilon$ for all $\alpha \le \alpha_2$. Hence, choosing $\bar{\alpha} = \min\{\alpha_1, \alpha_2\}$ proves (3.9). □

3.3. DGD likely converges to a neighborhood of SoS solutions of $\boldsymbol{F}$.

We study now second-order guarantees of DGD. Our path to prove almost sure convergence to a neighborhood of SoS solutions of (P) will pass through the nonconvergence of DGD to strict saddles of L_α (cf. Theorem 3.6). Roughly speaking, our idea is to show 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 (3.9), $\mathbf{x}^\infty = [\mathbf{x}_1^{\infty\top}, \ldots, \mathbf{x}_n^{\infty\top}]^\top$ must be a *strict saddle of* L_α. The escaping properties of DGD from strict saddles of L_α will then ensure that it is unlikely that $\{\bar{\mathbf{x}}^\nu = 1/n \sum_{i=1}^n \mathbf{x}_i^\nu\}$ gets trapped in a neighborhood of a strict saddle of F, thus ending in a neighborhood of an SoS solution of (P). Proposition 3.10 makes this argument formal; in particular, conditions (i)–(iii) identify the neighborhood of a strict saddle of F with the mentioned escaping properties.

PROPOSITION 3.10. *Consider the setting of Lemma* 3.7 *and further assume that Assumption* 2.3 *holds. Let* $\bar{\mathcal{Y}}$ *be the image of the compact set* $\mathcal{Y}$ *(defined in Lemma* 3.7*) through the linear operator* $(\mathbf{1}_n \otimes \mathbf{I}_m)^\top$. *Suppose that the limit point* $\mathbf{x}^\infty = [\mathbf{x}_1^{\infty\top}, \ldots, \mathbf{x}_n^{\infty\top}]^\top$ *of* $\{\mathbf{x}^\nu\}$*, along with* $\bar{\mathbf{x}}^\infty = 1/n \sum_{i=1}^n \mathbf{x}_i^\infty$*, satisfy*

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

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

(iii) *there exists* $\boldsymbol{\theta}^* \in \operatorname{proj}_{\operatorname{crit} \mathrm{F}}(\bar{\mathbf{x}}^\infty) \cap \Theta^*_{ss}$

for some δ *such that* $\delta \leq -\lambda_{\min}\left(\nabla^2 F(\boldsymbol{\theta}^*)\right)$ *for all* $\boldsymbol{\theta}^* \in \Theta^*_{ss} \cap \bar{\mathcal{Y}}$. *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

(3.10)

$$\begin{aligned}
\tilde{\boldsymbol{v}}(\boldsymbol{\theta})^\top \nabla^2 L_\alpha(\mathbf{x}^\infty) \tilde{\boldsymbol{v}}(\boldsymbol{\theta}) &\overset{(a)}{=} \tilde{\boldsymbol{v}}(\boldsymbol{\theta})^\top \nabla^2 F_c(\mathbf{x}^\infty) \tilde{\boldsymbol{v}}(\boldsymbol{\theta}) \\
&\leq \boldsymbol{v}(\boldsymbol{\theta})^\top \nabla^2 F(\boldsymbol{\theta}) \boldsymbol{v}(\boldsymbol{\theta}) \\
&\quad + ||\nabla^2 F(\bar{\mathbf{x}}^\infty) - \nabla^2 F(\boldsymbol{\theta})|| \, \|\boldsymbol{v}(\boldsymbol{\theta})\|^2 + ||\nabla^2 F_c(\mathbf{x}^\infty) - \nabla^2 F_c(\mathbf{1} \otimes \bar{\mathbf{x}}^\infty)|| \, \|\tilde{\boldsymbol{v}}(\boldsymbol{\theta})\|^2 \\
&\overset{(b)}{\leq} \boldsymbol{v}(\boldsymbol{\theta})^\top \nabla^2 F(\boldsymbol{\theta}) \boldsymbol{v}(\boldsymbol{\theta}) + L_{\nabla^2} \|\bar{\mathbf{x}}^\infty - \boldsymbol{\theta}\| + n \, L_{\nabla^2_c} \|\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty\|,
\end{aligned}$$

where (a) follows from $\tilde{\boldsymbol{v}}(\boldsymbol{\theta}) \in \operatorname{null}(\mathbf{W}_D - \mathbf{I})$ and (b) is due to Assumption 2.3. Let us now evaluate (3.10) 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 $\mathbf{x}^\infty \in \operatorname{crit} L_\alpha$ (cf. Theorem 3.2(ii)), proves the proposition. □

Invoking now Theorem 3.9, we infer that there exists a sufficiently small $\alpha > 0$ such that conditions (i) and (ii) of Proposition 3.10 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 stated next.

COROLLARY 3.11. *Consider the setting of Theorem* 3.9 *and Proposition* 3.10. *There exists a sufficiently small* $\alpha > 0$ *such that, if* $\operatorname{proj}_{\operatorname{crit} \mathrm{F}}(\bar{\mathbf{x}}^\infty) \cap \Theta^*_{ss} \neq \emptyset$*, then* $\mathbf{x}^\infty$ *is a strict saddle of* L_α.

To state our final result, let us introduce the following merit function: given $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top]^\top$ let

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

where $\mathcal{X}_{SOS}$ denotes the set of SoS solutions of (P), and $\bar{\mathbf{x}} = 1/n \sum_{i=1}^n \mathbf{x}_i$. $M(\mathbf{x})$

captures the distance of the average $\bar{\mathbf{x}}$ from the set of SoS solutions of (P) as well as the consensus disagreement of the agents' local variables $\bar{\mathbf{x}}_i$.

THEOREM 3.12. *Consider problem* (P) *under Assumptions* 2.1′, 2.3, 2.4, *and* 2.5*; further assume that each f_i is a KŁ function. For every $\epsilon > 0$, there exists sufficiently small $0 < \bar{\alpha} < \frac{\sigma_{\min}(\mathbf{D})}{L_c}$ such that*

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

where $\mathbf{x}^\infty = [\mathbf{x}_1^{\infty\top}, \ldots, \mathbf{x}_n^{\infty\top}]^\top$ is the limit point of the sequence $\{\mathbf{x}^\nu\}$ generated by the DGD algorithm (3.1) *with $\alpha \in (0, \bar{\alpha}]$, the weight matrix $\mathbf{D}$ satisfying Assumptions* 3.1 *and* 3.3*, and initialization $\mathbf{x}^0 \in \prod_{i=1}^n \mathcal{X}_i^0 \subseteq \prod_{i=1}^n \mathcal{B}_R^m$; R is defined in Assumption* 2.4 *and each $\mathcal{X}_i^0$ has positive Lebesgue measure; and the probability is taken over the initialization $\mathbf{x}^0 \in \prod_{i=1}^n \mathcal{X}_i^0$. Furthermore, any $\boldsymbol{\theta}^* \in \operatorname{proj}_{\text{crit F}}(\bar{\mathbf{x}}^\infty)$ is almost surely an SoS solution of F, where $\bar{\mathbf{x}}^\infty = (1/n)\sum_{i=1}^n \mathbf{x}_i^\infty$.*

Proof. For sufficiently small $\alpha < \bar{\alpha}_1$, if $\operatorname{proj}_{\text{crit F}}(\bar{\mathbf{x}}^\infty)$ contains a strict saddle point of F, then $\mathbf{x}^\infty$ is also a strict saddle point of L_α (by Corollary 3.11). Let also $\bar{\alpha}_2$ be a sufficiently small step size such that every limit point $\mathbf{x}^\infty$ satisfies $\operatorname{dist}(\bar{\mathbf{x}}^\infty, \operatorname{crit} F) \le \epsilon$ and $\|\mathbf{x}^\infty - \mathbf{1} \otimes \bar{\mathbf{x}}^\infty\| \le \epsilon$ (by Theorem 3.9). Now consider DGD update (3.1) with $\alpha < \min\{\bar{\alpha}_1, \bar{\alpha}_2\}$ and $\mathbf{x}^0$ being drawn randomly from the set of probability one measures $\prod_{i=1}^n \mathcal{X}_i^0$ for which the algorithm converges to an SoS solution of L_α (by[2] Theorem 3.4). Finally, by the above properties of α, it holds that $M(\mathbf{x}^\infty) \le \epsilon$ and $\operatorname{proj}_{\text{crit F}}(\bar{\mathbf{x}}^\infty)$ must contain only SoS solutions of F. Therefore, there exists a $\boldsymbol{\theta}^* \in \operatorname{crit} F$ such that $\boldsymbol{\theta}^* \in \mathcal{X}_{SoS}$ and $\|\bar{\mathbf{x}}^\infty - \boldsymbol{\theta}^*\| \le \epsilon$. □

Remark 3.13. All (first- and second-order) convergence results of DGD established in this section remain valid when ∇f_i's are not globally Lipschitz continuous (Assumption 2.1(i)) but Assumption 2.4 holds. Specifically, Theorems 3.2, 3.4, 3.9, and Lemmas 3.6–3.7 hold if one replaces Assumption 2.1(i) with Assumption 2.4 and the global Lipschitz constant L_c with the Lipschitz constant of ∇F_c *restricted* to the compact set $\widetilde{\mathcal{Y}}$, defined in Appendix A.2, which we refer to for the technical details.

4. DOGT algorithms.

The family of DOGT algorithms is introduced in section 1.1.2. We begin here rewriting (1.2)–(1.3) in matrix/vector form. Denoting $\mathbf{x}^\nu \triangleq [\mathbf{x}_1^{\nu\top}, \ldots, \mathbf{x}_n^{\nu\top}]^\top$ and $\mathbf{y}^\nu \triangleq [\mathbf{y}_1^{\nu\top}, \ldots, \mathbf{y}_n^{\nu\top}]^\top$, we have

$$\begin{cases} \mathbf{x}^{\nu+1} = \mathbf{W}_R\, \mathbf{x}^\nu - \alpha\, \mathbf{y}^\nu, \\ \mathbf{y}^{\nu+1} = \mathbf{W}_C\, \mathbf{y}^\nu + \nabla F_c\big(\mathbf{x}^{\nu+1}\big) - \nabla F_c\big(\mathbf{x}^\nu\big), \end{cases} \tag{4.1}$$

where $\mathbf{W}_R \triangleq \mathbf{R} \otimes \mathbf{I}_m$ and $\mathbf{W}_C \triangleq \mathbf{C} \otimes \mathbf{I}_m$ with $\mathbf{R} \triangleq (R_{ij})_{i,j=1}^n$ and $\mathbf{C} \triangleq (C_{ij})_{i,j=1}^n$ being some *column-stochastic* and *row-stochastic* matrices (respectively) compliant to the graph $\mathcal{G}$ (cf. Assumption 4.1 below). The initialization of (4.1) is set to $\mathbf{x}^0 \in \mathbb{R}^{mn}$ and $\mathbf{y}^0 \in \nabla F_c(\mathbf{x}^0) + \operatorname{span}(\mathbf{W}_C - \mathbf{I})$. Note that the latter condition is instrumental to preserve the *total-sum* of the y-variables, namely, $\sum_i \mathbf{y}_i^\nu = \sum_i f_i(\mathbf{x}_i^\nu)$ (which holds due to the column-stochasticity of matrix $\mathbf{C}$; cf. Assumption 4.1). This property is imperative for the y-variables to track the sum-gradient. Notice that the condition used in the literature [22, 57, 59, 48, 66]—$\mathbf{y}^0 = \nabla F_c(\mathbf{x}^0)$—is a special case of the proposed initialization. On the practical side, this initialization can be enforced in a distributed way, with minimal coordination. For instance, agents first choose independently

[2] Note that the conclusion of Theorem 3.4 is valid also when the set of initial points is restricted to $\prod_{i=1}^n \mathcal{X}_i^0$, as $\prod_{i=1}^n \mathcal{X}_i^0$ has positive measure (the Cartesian product of sets with positive measure; has positive measure; cf. [31, section 35]).

a vector $\mathbf{y}_i^{-1} \in \mathbb{R}^m$; then they run one step of consensus on the y-variables using the values y_i^{-1} and weights matrix $\mathbf{C}$, and set $\mathbf{y}_i^0 = \nabla f_i(\mathbf{x}_i^0) + \sum_{j \in \mathcal{N}_i^{in}} C_{ij}\mathbf{y}_j^{-1} - \mathbf{y}_i^{-1}$, resulting in $\mathbf{y}^0 \in \nabla F_c(\mathbf{x}^0) + \operatorname{span}(\mathbf{W}_C - \mathbf{I})$.

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 $\mathcal{G}$, (4.1) reduces to the NEXT algorithm [21, 22] (or the one in [68], when (P) is convex). If $\mathbf{R}$ and $\mathbf{C}$ are allowed to be time varying (suitably chosen) (4.1) reduces to the SONATA algorithm applicable to (possibly time-varying) digraphs [62, 58, 59, 61] (or the one later proposed in [48] for strongly convex instances of (P)). Finally, if $\mathbf{R}$ and $\mathbf{C}$ are chosen according to Assumption 4.1 below, the scheme (4.1) becomes the algorithm proposed independently in [56] and [66], for strongly convex objectives in (P), and implementable over fixed digraphs.

Assumption 4.1 (on the matrices $\mathbf{R}$ and $\mathbf{C}$). The weight matrices $\mathbf{R}, \mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ 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 [56, 66]. Here, we only point out the following property of $\mathbf{R}$ and $\mathbf{C}$, as a consequence of Assumption 4.1, which will be used in our analysis. The result is a consequence of [67, Lemma 1].

LEMMA 4.2. *Given* $\mathbf{R}$ *and* $\mathbf{C}$ *satisfying Assumption* 4.1 *with stochastic left eigenvector* $\mathbf{r}$ *(resp., right eigenvector* $\mathbf{c}$*) of* $\mathbf{R}$ *(resp.,* $\mathbf{C}$*) associated with the eigenvalue one, then there exist matrix norms*

$$||\mathbf{X}||_R \triangleq ||\operatorname{diag}(\sqrt{\mathbf{r}})\mathbf{X}\operatorname{diag}(\sqrt{\mathbf{r}})^{-1}||_2, \tag{4.2}$$

$$||\mathbf{X}||_C \triangleq ||\operatorname{diag}(\sqrt{\mathbf{c}})^{-1}\mathbf{X}\operatorname{diag}(\sqrt{\mathbf{c}})||_2, \tag{4.3}$$

such that $\rho_R \triangleq \|\mathbf{R} - \mathbf{1}\mathbf{r}^\top\|_R < 1$ *and* $\rho_C \triangleq \|\mathbf{C} - \mathbf{c}\mathbf{1}^\top\|_C < 1$. *Furthermore,* $\mathbf{r}^\top\mathbf{c} > 0$.

Using Lemma 4.2, it is not difficult to check that the following properties hold:

$$\rho_R = \sigma_2\left(\operatorname{diag}(\sqrt{\mathbf{r}})\mathbf{R}\operatorname{diag}(\sqrt{\mathbf{r}})^{-1}\right), \tag{4.4}$$

$$\rho_C = \sigma_2\left(\operatorname{diag}(\sqrt{\mathbf{c}})^{-1}\mathbf{C}\operatorname{diag}(\sqrt{\mathbf{c}})\right), \tag{4.5}$$

$$||\mathbf{R}||_R = ||\mathbf{1}\mathbf{r}^T||_R = ||\mathbf{I} - \mathbf{1}\mathbf{r}^T||_R = 1, \tag{4.6}$$

$$||\mathbf{C}||_C = ||\mathbf{c}\mathbf{1}^T||_C = ||\mathbf{I} - \mathbf{c}\mathbf{1}^T||_R = 1. \tag{4.7}$$

The vector norms associated with the above matrix norms are

$$||\mathbf{x}||_R = ||\operatorname{diag}(\sqrt{\mathbf{r}})\mathbf{x}||_2, \tag{4.8}$$

$$||\mathbf{x}||_C = ||\operatorname{diag}(\sqrt{\mathbf{c}})^{-1}\mathbf{x}||_2; \tag{4.9}$$

and $||\cdot||_a \leq K_{a,b}||\cdot||_b$ holds for $a, b \in \{R, C, 2\}$ with

$$\begin{aligned} &K_{R,2} = \sqrt{r_{\max}}, \quad K_{2,R} = 1/\sqrt{r_{\min}}, \\ &K_{C,2} = 1/\sqrt{c_{\min}}, \quad K_{2,C} = \sqrt{c_{\max}}, \\ &K_{R,C} = \sqrt{r_{\max}c_{\max}}, \quad K_{C,R} = 1/\sqrt{c_{\min}r_{\min}}, \end{aligned} \tag{4.10}$$

where $r_{\min}$ (resp., $c_{\min}$) and $r_{\max}$ (resp., $c_{\max}$) are minimum and maximum elements of $\mathbf{r}$ (resp., $\mathbf{c}$).

Convergence of DOGT algorithms in the form (4.1) (with $\mathbf{R}$ and $\mathbf{C}$ satisfying Assumption 4.1) has not been studied in the literature when F is nonconvex. In next subsection we fill this gap and provide a full characterization of the convergence behavior of DOGT including its second-order guarantees.

4.1. First-order convergence and rate analysis.

In this section, we study asymptotic convergence to first-order stationary solutions; we assume $m = 1$ (scalar optimization variables); while this simplifies the notation, all the conclusions hold for the general case $m > 1$. As in [56], define the weighted sums

$$\bar{x}^\nu \triangleq \mathbf{r}^\top \mathbf{x}^\nu, \quad \bar{y}^\nu \triangleq \mathbf{1}^\top \mathbf{y}^\nu, \quad \text{and} \quad \bar{g}^\nu \triangleq \mathbf{1}^\top \nabla F_c(\mathbf{x}^\nu), \tag{4.11}$$

where we recall that $\mathbf{r}$ is the Perron vector associated with $\mathbf{R}$ (cf. Lemma 4.2). Note that ∇F_c is L_c-Lipschitz continuous with $L_c \triangleq L_{\max}$.

Using (4.1), it is not difficult to check that the following holds:

$$\bar{x}^{\nu+1} = \bar{x}^\nu - \zeta\alpha\bar{y}^\nu - \alpha\mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right) \quad \text{and} \quad \bar{y}^\nu = \bar{g}^\nu, \tag{4.12}$$

where $\mathbf{c}$ is the Perron vector associated with $\mathbf{C}$, and $\zeta \triangleq \mathbf{r}^\top\mathbf{c} > 0$ (cf. Lemma 4.2).

4.1.1. Descent on F.

Using the descent lemma along with (4.12) yields

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

Adding/subtracting suitably chosen terms we obtain

$$\begin{aligned} F(\bar{x}^{\nu+1}) &\le F(\bar{x}^\nu) - \zeta\alpha\left\langle\nabla F(\bar{x}^\nu) - \bar{y}^\nu, \bar{y}^\nu\right\rangle - \zeta\alpha|\bar{y}^\nu|^2 \\ &\quad - \alpha\left\langle\nabla F(\bar{x}^\nu) - \bar{y}^\nu, \mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right\rangle - \alpha\left\langle\bar{y}^\nu, \mathbf{r}^\top\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right)\right\rangle \\ &\quad + L\zeta^2\alpha^2|\bar{y}^\nu|^2 + L\alpha^2\left\|\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right\|^2 \\ &\le F(\bar{x}^\nu) + \frac{\zeta\alpha}{2\epsilon_1}|\nabla F(\bar{x}^\nu) - \bar{y}^\nu|^2 + \frac{\zeta\alpha\epsilon_1}{2}|\bar{y}^\nu|^2 - \zeta\alpha|\bar{y}^\nu|^2 \\ &\quad + \frac{\alpha}{2}|\nabla F(\bar{x}^\nu) - \bar{y}^\nu|^2 + \frac{\alpha}{2}\left\|\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right\|^2 + \frac{\alpha\epsilon_2}{2}|\bar{y}^\nu|^2 + \frac{\alpha}{2\epsilon_2}\left\|\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right\|^2 \\ &\quad + L\zeta^2\alpha^2|\bar{y}^\nu|^2 + L\alpha^2\left\|\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right\|^2 \\ &= F(\bar{x}^\nu) + \left(\frac{\zeta\alpha\epsilon_1}{2} - \zeta\alpha + \frac{\alpha\epsilon_2}{2} + L\zeta^2\alpha^2\right)|\bar{y}^\nu|^2 \\ &\quad + \left(\frac{\zeta\alpha}{2\epsilon_1} + \frac{\alpha}{2}\right)|\nabla F(\bar{x}^\nu) - \bar{y}^\nu|^2 + \left(\frac{\alpha}{2} + \frac{\alpha}{2\epsilon_2} + L\alpha^2\right)\left\|\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right\|^2, \end{aligned} \tag{4.13}$$

where ϵ_1 and ϵ_2 are some arbitrary positive quantities (to be chosen). By $\bar{y}^\nu = \bar{g}^\nu$ (cf. (4.12)), it holds that

$$|\nabla F(\bar{x}^\nu) - \bar{y}^\nu| = \left|\sum_{i=1}^n \nabla f_i(\bar{x}^\nu) - \sum_{i=1}^n \nabla f_i(x_i^\nu)\right| \le L_c\sqrt{n}\left\|\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu\right\|. \tag{4.14}$$

Combining (4.13) and (4.14) yields

$$
\begin{aligned}
&F(\bar{x}^{\nu+1})\\
&\leq F(\bar{x}^{\nu}) + \left(\frac{\zeta\alpha\epsilon_1}{2} - \zeta\alpha + \frac{\alpha\epsilon_2}{2} + L\zeta^2\alpha^2\right)|\bar{y}^{\nu}|^2\\
&\quad + nL_c^2K_{2,R}^2\left(\frac{\zeta\alpha}{2\epsilon_1} + \frac{\alpha}{2}\right)\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|_R^2 + K_{2,C}^2\left(\frac{\alpha}{2} + \frac{\alpha}{2\epsilon_2} + L\alpha^2\right)\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2\,,
\end{aligned}
\tag{4.15}
$$

where $K_{2,R} = 1/\sqrt{r_{\min}}$ and $K_{2,C} = \sqrt{c_{\max}}$ (cf. (4.36)).

4.1.2. Bounding the consensus and gradient tracking errors. Let us bound the consensus error $\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|_R$. Using $\|\mathbf{z}+\mathbf{w}\|_R^2 \leq (1+\epsilon)\,\|\mathbf{x}\|_R^2 + (1+1/\epsilon)\,\|\mathbf{y}\|_R^2$ for arbitrary $\mathbf{z}, \mathbf{w} \in \mathbb{R}^m$ and $\epsilon > 0$, along with Lemma 4.2, yields

$$
\begin{aligned}
&\left\|\mathbf{x}^{\nu+1} - \mathbf{1}\bar{x}^{\nu+1}\right\|_R^2 = \left\|\left(\mathbf{R} - \mathbf{1}\mathbf{r}^{\top}\right)(\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}) - \alpha\left(\mathbf{I} - \mathbf{1}\mathbf{r}^{\top}\right)(\mathbf{y}^{\nu} - \mathbf{1}\bar{y}^{\nu})\right\|_R^2\\
&\leq (1+\epsilon_x)\left\|\left(\mathbf{R} - \mathbf{1}\mathbf{r}^{\top}\right)(\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu})\right\|_R^2 + \alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\left\|\left(\mathbf{I} - \mathbf{1}\mathbf{r}^{\top}\right)(\mathbf{y}^{\nu} - \mathbf{1}\bar{y}^{\nu})\right\|_R^2\\
&\leq \rho_R^2(1+\epsilon_x)\,\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|_R^2 + \alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\|\mathbf{I} - \mathbf{1}\mathbf{r}^{\top}\|_R^2\,\|\mathbf{y}^{\nu} - \mathbf{1}\bar{y}^{\nu}\|_R^2\\
&\overset{(4.6)}{\leq} \rho_R^2(1+\epsilon_x)\,\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|_R^2 + 2\alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_R^2\\
&\qquad + 2\alpha^2\left(1 + \frac{1}{\epsilon_x}\right)\|(\mathbf{1} - \mathbf{c})\bar{y}^{\nu}\|_R^2\\
&\leq \rho_R^2(1+\epsilon_x)\,\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|_R^2 + \alpha^2K_2\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2 + \alpha^2K_3|\bar{y}^{\nu}|_2^2,
\end{aligned}
\tag{4.16}
$$

where $\epsilon_x > 0$ is arbitrary and we defined

$$
K_2 \triangleq 2K_{R,C}^2\left(1 + \frac{1}{\epsilon_x}\right), \quad K_3 \triangleq 2n\left(1 + \frac{1}{\epsilon_x}\right).
\tag{4.17}
$$

Similarly, the tracking error can be bounded as

$$
\begin{aligned}
&\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\\
&\leq (1+\epsilon_y)\left\|\left(\mathbf{C} - \mathbf{c}\mathbf{1}^{\top}\right)(\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu})\right\|_C^2\\
&\quad + \left(1 + \frac{1}{\epsilon_y}\right)\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\\
&\overset{(4.7)}{\leq} \rho_C^2(1+\epsilon_y)\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2 + K_{C,2}^2L_c^2\left(1 + \frac{1}{\epsilon_y}\right)\left\|\mathbf{x}^{\nu+1} - \mathbf{x}^{\nu}\right\|^2\\
&\overset{(a)}{=} \rho_C^2(1+\epsilon_y)\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2\\
&+ 3K_{C,2}^2L_c^2\left(1 + \frac{1}{\epsilon_y}\right)\left[\|(\mathbf{R} - \mathbf{I})(\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu})\|^2 + \alpha^2\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|^2 + \alpha^2|\bar{y}^{\nu}|^2\,\|\mathbf{c}\|^2\right]\\
&\overset{(4.6)}{\leq} \rho_C^2(1+\epsilon_y)\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2\\
&+ 3K_{C,2}^2L_c^2\left(1 + \frac{1}{\epsilon_y}\right)\left[K_{2,R}^2\,\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|^2 + K_{2,C}^2\alpha^2\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2 + \alpha^2|\bar{y}^{\nu}|^2\right]\\
&= \rho_C^2(1+\epsilon_y)\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2\\
&+ 3K_{C,2}^2L_c^2\left(1 + \frac{1}{\epsilon_y}\right)\left[K_{2,R}^2\,\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|^2 + K_{2,C}^2\alpha^2\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2 + \alpha^2|\bar{y}^{\nu}|^2\right]\\
&\leq \left(\rho_C^2 + \frac{\alpha^2K_4}{\epsilon_y}\right)(1+\epsilon_y)\,\|\mathbf{y}^{\nu} - \mathbf{c}\bar{y}^{\nu}\|_C^2 + \alpha^2K_5|\bar{y}^{\nu}|_2^2 + K_6\left(1 + \frac{1}{\epsilon_y}\right)\|\mathbf{x}^{\nu} - \mathbf{1}\bar{x}^{\nu}\|_R^2\,,
\end{aligned}
\tag{4.18}
$$

where in (a) we used $\mathbf{x}^{\nu+1}-\mathbf{x}^{\nu}=(\mathbf{R}-\mathbf{I})(\mathbf{x}^{\nu}-\mathbf{1}\bar{x}^{\nu})-\alpha(\mathbf{y}^{\nu}-\mathbf{c}\bar{y}^{\nu})-\alpha\mathbf{c}\bar{y}^{\nu}$ and Jensen's inequality; and in the last inequality we defined

$$K_4=3K_{C,2}^2K_{2,C}^2L_c^2, \quad K_5=3K_{C,2}^2L_c^2, \quad K_6=3K_{C,2}^2K_{2,R}^2L_c^2. \tag{4.19}$$

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

$$L(\mathbf{x},\mathbf{y}) \triangleq F_c(\mathbf{J}_R\mathbf{x})+\|(\mathbf{I}-\mathbf{J}_R)\mathbf{x}\|_R^2+\varkappa\,\|(\mathbf{I}-\mathbf{J}_C)\mathbf{y}\|_C^2\,, \tag{4.20}$$

where $\varkappa>0$ is a positive constant (to be properly chosen). Combining (4.15)–(4.18) and using $\bar{y}^{\nu}=\bar{g}^{\nu}=\sum_{i=1}^{n}\nabla f_i(x_i^{\nu})$ (cf. (4.12)) leads to the following descent property for L:

$$L(\mathbf{x}^{\nu+1},\mathbf{y}^{\nu+1}) \leq L(\mathbf{x}^{\nu},\mathbf{y}^{\nu})-d(\mathbf{x}^{\nu},\mathbf{y}^{\nu})^2, \tag{4.21}$$

where

$$d(\mathbf{x},\mathbf{y}) \triangleq \sqrt{(1-\tilde{\rho}_R)\,\|(\mathbf{I}-\mathbf{J}_R)\mathbf{x}\|_R^2+\varkappa(1-\tilde{\rho}_C)\,\|(\mathbf{I}-\mathbf{J}_C)\mathbf{y}\|_C^2+\Gamma\,\Big|\sum_{i=1}^{n}\nabla f_i(x_i)\Big|^2} \tag{4.22}$$

and

$$\begin{aligned}\tilde{\rho}_R &\triangleq \rho_R^2(1+\epsilon_x)+\frac{\alpha n L_c^2K_{2,R}^2}{2}\left(1+\frac{\zeta}{\epsilon_1}\right)+\varkappa K_6\left(1+\frac{1}{\epsilon_y}\right),\\ \tilde{\rho}_C &\triangleq \rho_C^2(1+\epsilon_y)+\frac{\alpha K_{2,C}^2}{2\varkappa}\left(1+\frac{1}{\epsilon_2}\right)+\alpha^2\left(\frac{LK_{2,C}^2+K_2}{\varkappa}+K_4\left(1+\frac{1}{\epsilon_y}\right)\right),\\ \Gamma &\triangleq \left(\zeta-\frac{\epsilon_1\zeta}{2}-\frac{\epsilon_2}{2}\right)\alpha-\left(L\zeta^2+K_3+K_5\varkappa\right)\alpha^2.\end{aligned} \tag{4.23}$$

Note that the function $d(\bullet,\bullet)$ is a valid measure of optimality/consensus for DOGT: (i) it is continuous and (ii) $d(\mathbf{x},\mathbf{y})=0$ implies $x_i=x_j=x^*$ for all $i,j\in[n]$ and some x^* such that $\sum_{i=1}^{n}\nabla f_i(x^*)=0$, meaning that all x_i are consensual and equal to a critical point of F.

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

$$\begin{aligned}&0<\epsilon_x<\frac{1-\rho_R^2}{2\rho_R^2}, \qquad 0<\epsilon_y<\frac{1-\rho_C^2}{\rho_C^2},\\ &\epsilon_1=\epsilon_2=\epsilon, \qquad 0<\epsilon<\frac{2\zeta}{1+\zeta}, \qquad 0<\varkappa\leq\frac{\rho_R^2\epsilon_x}{K_6(1+1/\epsilon_y)},\end{aligned} \tag{4.24}$$

and, finally, $\alpha>0$ must satisfy

$$\begin{aligned}\alpha &< \frac{2}{nL_c^2K_{2,R}^2\left(1+\frac{\zeta}{\epsilon}\right)}\left(1-\rho_R^2(1+2\epsilon_x)\right),\\ \alpha &< \frac{1-\rho_C^2(1+\epsilon_y)}{\frac{1}{2\varkappa}K_{2,C}^2\left(1+\frac{1}{\epsilon}+2L\right)+\frac{K_2}{\varkappa}+K_4\left(1+\frac{1}{\epsilon_y}\right)},\\ \alpha &< \frac{\zeta-\frac{\epsilon}{2}(\zeta+1)}{L\zeta^2+K_3+K_5\varkappa}.\end{aligned} \tag{4.25}$$

Substituting (4.10), (4.17), and (4.19) in (4.25) and setting for simplicity

$$\epsilon_x = \frac{1-\rho_R^2}{4\rho_R^2}, \quad \epsilon_y = \frac{1-\rho_C^2}{2\rho_C^2}, \quad \epsilon = \frac{\xi}{1+\xi}, \quad \varkappa = \frac{c_{\min} r_{\min}}{24 L_c^2}(1-\rho_R^2)(1-\rho_C^2), \tag{4.26}$$

we obtain the following sufficient conditions for (4.25):

$$\begin{aligned} \alpha &\leq \tilde{\alpha}_1 \triangleq \frac{r_{\min}(1-\rho_R^2)}{3nL_c^2}, \\ \alpha &\leq \tilde{\alpha}_2 \triangleq \frac{(1-\rho_R^2)^2(1-\rho_C^2)^2 r_{\min}^2 c_{\min}^2}{1152 L_c^2(2+L)}, \\ \alpha &\leq \tilde{\alpha}_3 \triangleq \frac{r_{\min} c_{\min}(1-\rho_R^2)}{2(L+16n)}. \end{aligned} \tag{4.27}$$

A further simplification leads to the following final more restrictive condition on α:

$$0 < \alpha \leq \frac{(1-\rho_R^2)^2(1-\rho_C^2)^2 r_{\min}^2 c_{\min}^2}{1152 L_c^2(L+16n)}. \tag{4.28}$$

The descent property (4.21) readily implies the following convergence result for $\{L(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ and $\{d(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$.

Lemma 4.3. *Under Assumptions* 2.1, 2.5, *and* 4.1, *and the above choice of parameter, there hold*

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

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

We conclude this subsection by lower bounding $d(\mathbf{x}^\nu, \mathbf{y}^\nu)$ by the magnitude of the gradient of the Lyapunov function L. This will allow us to transfer the convergence properties of $\{d(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ to $\{||\nabla L(\mathbf{x}^\nu, \mathbf{y}^\nu)||\}$. The lemma below will also be useful to establish global convergence of DOGT under the KŁ property (cf. section 4.2.1).

Lemma 4.4. *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*

$$\|\nabla L(\mathbf{x}^\nu, \mathbf{y}^\nu)\| \leq M d(\mathbf{x}^\nu, \mathbf{y}^\nu), \quad \nu \geq 0, \tag{4.29}$$

with

$$M = \sqrt{2} \max\left(\frac{(2r_{\max} + L_c\sqrt{n})^2}{r_{\min}(1-\tilde{\rho}_R)}, \frac{2\varkappa c_{\max}}{c_{\min}^2(1-\tilde{\rho}_C)}, \frac{1}{\Gamma}\right)^{\frac{1}{2}}. \tag{4.30}$$

Proof. Recall that $\mathbf{J}_R = \mathbf{1}\mathbf{r}^\top$ and $\mathbf{J}_C = \mathbf{c}\mathbf{1}^\top$. By definition (4.20) and Lemma 4.2, we can write

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}^\nu, \mathbf{y}^\nu) &= \mathbf{J}_R^\top \nabla F_c(\mathbf{J}_R \mathbf{x}^\nu) + 2(\mathbf{I} - \mathbf{J}_R)^\top \operatorname{diag}(\mathbf{r})(\mathbf{I} - \mathbf{J}_R)\mathbf{x}^\nu \\ &\overset{(a)}{=} \mathbf{r}\,\bar{y}^\nu + \mathbf{J}_R^\top\left(\nabla F_c(\mathbf{J}_R \mathbf{x}^\nu) - \nabla F_c(\mathbf{x}^\nu)\right) \\ &\quad + 2(\mathbf{I} - \mathbf{J}_R)^\top \operatorname{diag}(\mathbf{r})(\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu), \\ \nabla_{\mathbf{y}} L(\mathbf{x}^\nu, \mathbf{y}^\nu) &= 2\varkappa(\mathbf{I} - \mathbf{J}_C)^\top \operatorname{diag}(\mathbf{c})^{-1}(\mathbf{I} - \mathbf{J}_C)\mathbf{y}^\nu \\ &= 2\varkappa(\mathbf{I} - \mathbf{J}_C)^\top \operatorname{diag}(\mathbf{c})^{-1}(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu), \end{aligned} \tag{4.31}$$

where (a) is due to $\bar{y}^\nu = \bar{g}^\nu$ (cf. (4.12)). Thus there holds

$$\begin{aligned} ||\nabla_{\mathbf{x}} L(\mathbf{x}^\nu, \mathbf{y}^\nu)|| &\leq ||\mathbf{r}||\,|\bar{y}^\nu| + ||\mathbf{J}_R^\top\left(\nabla F_c(\mathbf{J}_R \mathbf{x}^\nu) - \nabla F_c(\mathbf{x}^\nu)\right)|| \\ &\quad + 2||(\mathbf{I} - \mathbf{J}_R)^\top \operatorname{diag}(\mathbf{r})(\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu)|| \\ &\overset{(b)}{\leq} |\bar{y}^\nu| + K_{2,R}\left(2r_{\max} + L_c\sqrt{n}\right)||\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu||_R, \\ ||\nabla_{\mathbf{y}} L(\mathbf{x}^\nu, \mathbf{y}^\nu)|| &\overset{(c)}{\leq} 2\varkappa K_{2,C} c_{\min}^{-1}||\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu||_C, \end{aligned} \tag{4.32}$$

where (b) holds due to $||\operatorname{diag}(\mathbf{r})||_R = ||\operatorname{diag}(\mathbf{r})||_2 = r_{\max}$, $||\mathbf{r}|| \leq 1$, $||\mathbf{J}_R||_2 \leq \sqrt{n}$, and (4.6); (c) is due to $||\operatorname{diag}(\mathbf{c})^{-1}||_C = ||\operatorname{diag}(\mathbf{c})^{-1}||_2 = c_{\min}^{-1}$ and (4.7). Equation (4.29) follows readily from (4.32). ◻

4.1.4. Main result.

We can now state the main convergence result of DOGT to critical points of F.

THEOREM 4.5. *Consider problem* (P)*, and suppose that Assumptions* 2.1 *and* 2.5 *are satisfied. Let* $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ *be the sequence generated by the DOGT Algorithm* (4.1) *with* $\mathbf{R}$ *and* $\mathbf{C}$ *satisfying Assumption* 4.1*, and* α *chosen according to* (4.28) *(or* (4.26)*); let* $\{\bar{x}^\nu\}$ *and* $\{\bar{y}^\nu\}$ *be defined in* (4.11)*; and let* $\{d(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ *be defined in* (4.22)*. Given* $\epsilon > 0$*, let* $T_\epsilon = \min\{\nu \in \mathbb{N}_+ : d(\mathbf{x}^\nu, \mathbf{y}^\nu) \leq \epsilon\}$*. 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* $\{\mathbf{x}^\nu\}$*; then,* $\mathbf{x}^\infty = \theta^\infty \mathbf{1}$ *for some* $\theta^\infty \in \operatorname{crit} F$*;*

(iii) (sublinear rate): $T_\epsilon = o(1/\epsilon^2)$.

Proof. (i) follows readily from Lemma 4.3(ii).

We prove (ii). Let $(\mathbf{x}^\infty, \mathbf{y}^\infty)$ be a limit point of $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$. 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, $\theta^\infty \in \operatorname{crit} F$.

We now prove (iii). Using (4.21) and the definition of T_ϵ, we can write

$$\frac{T_\epsilon}{2}\epsilon^2 \leq \sum_{t=\lfloor \frac{T_\epsilon}{2} \rfloor + 1}^{T_\epsilon} d(\mathbf{x}^t, \mathbf{y}^t)^2 \leq l^{\lfloor \frac{T_\epsilon}{2} \rfloor + 1} - l^{T_\epsilon + 1}, \tag{4.33}$$

where we used the shorthand $l^\nu \triangleq L(\mathbf{x}^\nu, \mathbf{y}^\nu)$. Consider the following two cases: (1) $T_\epsilon \to \infty$ as $\epsilon \to 0$, then $l^{\lfloor \frac{T_\epsilon}{2} \rfloor + 1} - l^{T_\epsilon + 1} \to 0$ (recall that $\{l^\nu\}$ converges; cf. Lemma 4.3(i)); and (2) $T_\epsilon < \infty$ as $\epsilon \to 0$, then $\{l^\nu\}$ converges in a finite number of iterations. Therefore, by (4.33), we have $T_\epsilon = o(1/\epsilon^2)$. ◻

Note that, as a direct consequence of Lemma 4.4, one can infer the following further property of the limit points $(\mathbf{x}^\infty, \mathbf{y}^\infty)$ of the sequence $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$: any such $(\mathbf{x}^\infty, \mathbf{y}^\infty)$ is a critical point of L [defined in (4.20)].

4.2. Convergence under the KŁ property.

We now strengthen the subsequence convergence result in Theorem 4.5, under the additional assumption that F is a KŁ function [40, 39]: we prove that the entire sequence $\{\mathbf{x}^\nu\}$ converges to a critical point of F (cf. Theorem 4.7), and establish asymptotic convergence rates (cf. Theorem 4.8). We extend the analysis developed in [5, 7] for centralized first-order methods to our distributed setting and complement it with a rate analysis. The major difference with [7] is that the *sufficient descent* condition postulated in [7] is neither satisfied by the objective value sequence $\{F(\mathbf{x}^\nu)\}$ (as requested in [7]), due to consensus and gradient tracking errors, nor by the Lyapunov function sequence $\{L(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$, which instead satisfies (4.21). A key step to cope with this issue is to establish necessary connections between $\nabla L(\mathbf{x}, \mathbf{y})$ and $d(\mathbf{x}, \mathbf{y})$ (defined in (4.20) and (4.22), respectively); see Lemma 4.29 and Proposition 4.6.

4.2.1. Convergence analysis.

We begin proving the following abstract intermediate results similar to [7] but extended to our distributed setting, which is at the core of the subsequent analysis; we still assume $m = 1$ without loss of generality.

PROPOSITION 4.6. *In the setting of Theorem* 4.5*, let L defined in* (4.20) *is KŁ 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* 2.2*. Let $\rho > 0$ be such that $\mathcal{B}(\acute{\mathbf{z}}, \rho)^{2mn} \subseteq \mathcal{V}_{\acute{\mathbf{z}}}$. Consider the sequence $\{\mathbf{z}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ generated by the DOGT algorithm* (4.1)*, 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*

$$\acute{l} < l^\nu < \acute{l} + \eta \quad \forall \nu \geq 0, \tag{4.34}$$

and

$$K\, M\, \phi(l^0 - \acute{l}) + \left\|\mathbf{z}^0 - \acute{\mathbf{z}}\right\| < \rho, \tag{4.35}$$

where

$$K = \sqrt{3}(1 + L_c) \max\left(\frac{4nK_{||}^2}{1 - \tilde{\rho}_R}, \frac{K_{||}^2}{\varkappa(1 - \tilde{\rho}_C)}\left(\alpha + \frac{2\sqrt{n}}{1 + L_c}\right)^2, \alpha^2/\Gamma\right)^{1/2}, \tag{4.36}$$

and $M > 0$ is defined in (4.29) *(cf. Lemma* 4.4*).*

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

(i) $\mathbf{z}^\nu \in \mathcal{B}(\acute{\mathbf{z}}, \rho)^{2mn}$ *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. Throughout the proof, we will use the following shorthand $d^\nu \triangleq d(\mathbf{x}^\nu, \mathbf{y}^\nu)$. Let $d^\nu > 0$, for all integers $\nu \geq 0$; otherwise, $\{\mathbf{x}^\nu\}$ converges in a finite number of steps, and its limit point is $\mathbf{x}^\infty = \mathbf{1}\theta^\infty$ for some $\theta^\infty \in \operatorname{crit} F$.

We first bound the "length" $\sum_{t=k}^{\nu} \left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\|$. By (4.1), there holds

$$\begin{aligned}
\mathbf{x}^{\nu+1} - \mathbf{x}^\nu &= (\mathbf{R} - \mathbf{I})\left(\mathbf{x}^\nu - \mathbf{1}\bar{x}^\nu\right) - \alpha\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right) - \alpha\mathbf{c}\bar{y}^\nu, \\
\mathbf{y}^{\nu+1} - \mathbf{y}^\nu &= (\mathbf{C} - \mathbf{I})\left(\mathbf{y}^\nu - \mathbf{c}\bar{y}^\nu\right) + \nabla F_c(\mathbf{x}^{\nu+1}) - \nabla F_c(\mathbf{x}^\nu).
\end{aligned}$$

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

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

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

$$\begin{aligned}
&\sum_{t=k}^{\nu} \left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\| \\
&\leq \sum_{t=k}^{\nu} 2(1 + L_c)\sqrt{n}K_{||}\left\|\mathbf{x}^t - \mathbf{1}\bar{x}^t\right\|_R + K_{||}\left(\alpha(1 + L_c) + 2\sqrt{n}\right)\left\|\mathbf{y}^t - \mathbf{c}\bar{y}^\nu\right\|_C \\
&\quad + \alpha(1 + L_c)|\bar{y}^t| \leq K\sum_{t=k}^{\nu} d^t,
\end{aligned} \tag{4.37}$$

where K is defined in (4.36).

We prove now the proposition, starting from statement (ii). Multiplying both sides of (4.21) by $\phi'(l^\nu - \acute{l})$ and using $\phi'(l^\nu - \acute{l}) > 0$ [due to property (iii) in Definition 2.2 and (4.34)] and the concavity of ϕ, yield

$$(d^\nu)^2\ \phi'(l^\nu - \acute{l}) \leq \phi'(l^\nu - \acute{l})\ \left(l^\nu - l^{\nu+1}\right) \leq \phi(l^\nu - \acute{l}) - \phi(l^{\nu+1} - \acute{l}). \tag{4.38}$$

For all $\mathbf{z} \in \mathcal{V}_{\acute{\mathbf{z}}} \cap (\acute{l} < L < \acute{l} + \eta)$, the KŁ inequality (2.1) holds; hence, assuming $\mathbf{z}^t \in \mathcal{B}(\acute{\mathbf{z}}, \rho)^{2mn}$ for all $t = 0, \ldots, \nu$, yields

$$\phi'(l^t - \acute{l})||\nabla L(\mathbf{z}^t)|| \geq 1, \quad t = 0, \ldots, \nu, \tag{4.39}$$

which together with (4.38) and (4.29) (cf. Lemma 4.4), 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

$$M\left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l})\right) \geq \sum_{t=k}^{\nu} d^t. \tag{4.40}$$

Combining (4.40) with (4.37), we obtain

$$\sum_{t=k}^{\nu} \left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\| \leq KM\left(\phi(l^k - \acute{l}) - \phi(l^{\nu+1} - \acute{l})\right). \tag{4.41}$$

Inequality (4.41) proves (ii) if $\mathbf{z}^\nu \in \mathcal{B}(\acute{\mathbf{z}}, \rho)^{2mn}$ for all $\nu \geq 0$, which is shown next.

Now let us prove statement (i). Letting $k = 0$ in (4.41), by (4.35), we obtain

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

Therefore, $\mathbf{z}^\nu \in \mathcal{B}(\acute{\mathbf{z}}, \rho)^{2mn}$ for all $\nu \geq 0$.

We finally prove statement (iii). Inequalities (4.29) (cf. Lemma 4.4) and (4.39) imply

$$\phi'(l^\nu - \acute{l})\ d^\nu \geq 1/M, \quad \nu \geq 0. \tag{4.42}$$

On the other hand, by Lemma 4.3(i), as $\nu \to \infty$, we have $l^\nu \to p$ for some $p \geq \acute{l}$. In fact, $p = \acute{l}$, otherwise $p - \acute{l} > 0$, which would contradict (4.42) (because $d^\nu \to 0$ as $\nu \to \infty$ and $\phi'(p - \acute{l}) < \infty$). □

Roughly speaking, Proposition 4.6 states that, if the algorithm is initialized in a suitably chosen neighborhood of a point at which L satisfies the KŁ property, then it will converge to that point. Combining this property with the subsequence convergence proved in Theorem 4.7 we can obtain global convergence of the sequence to critical points of F, as stated next.

THEOREM 4.7. *Consider the setting of Theorem* 4.5 *and, furthermore, assume that* F *is real-analytic. Any sequence* $\{(\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ *generated by the DOGT algorithm* (4.1) *converges to some* $(\mathbf{x}^\infty, \mathbf{y}^\infty) \in \text{crit } L$. *Furthermore,* $\mathbf{x}^\infty = \mathbf{1} \otimes \theta^\infty$ *for some* $\theta^\infty \in \text{crit } F$.

Proof. Let $\mathbf{z}^\infty \triangleq (\mathbf{x}^\infty, \mathbf{y}^\infty)$ be a limit point of $\{\mathbf{z}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{y}^\nu)\}$. Since $\{l^\nu \triangleq L(\mathbf{z}^\nu)\}$ is convergent (cf. Lemma 4.3) and L is continuous, we deduce $l^\nu \to l^\infty \triangleq L(\mathbf{z}^\infty)$. Since F is real-analytic, L is real-analytic (due to Lemma 4.2 and the fact that summation/composition of functions preserve the real-analytic property [38, Proposition 2.2.8]) and thus KŁ at at $\mathbf{z}^\infty$ [40]. Set $\acute{\mathbf{z}} = \mathbf{z}^\infty$ and $\acute{l} = l^\infty$; denote by $\mathcal{V}_{\acute{\mathbf{z}}}$, η, and $\phi : [0, \eta) \to \mathbb{R}_+$ the objects appearing in Definition 2.2; and let $\rho > 0$ be such that $\mathcal{B}(\acute{\mathbf{z}}, \rho)^{2mn} \subseteq \mathcal{V}_{\acute{\mathbf{z}}}$. By the continuity of ϕ and the properties above, we deduce that there exists an integer ν_0 such that (i) $l^\nu \in (\acute{l}, \acute{l} + \eta)$ for all $\nu \geq \nu_0$ and (ii) $K\ M\ \phi(l^{\nu_0} - \acute{l}) + \|\mathbf{z}^{\nu_0} - \acute{\mathbf{z}}\| < \rho$, with K and M defined in (4.36) and (4.29), respectively. Global convergence of the sequence $\{\mathbf{z}^\nu\}$ follows by applying Proposition 4.6 to the sequence $\{\mathbf{z}^{\nu+\nu_0}\}$.

Finally, by Lemma 4.3(ii), $d(\mathbf{x}^\nu, \mathbf{y}^\nu) \to 0$ as $\nu \to \infty$. Invoking the continuity of ∇L and Lemma 4.4, we have $\nabla L(\mathbf{x}^\infty, \mathbf{y}^\infty) = \mathbf{0}$, thus $(\mathbf{x}^\infty, \mathbf{y}^\infty) \in \text{crit } L$. By Theorem 4.5(ii), $\mathbf{x}^\infty = \mathbf{1} \otimes \theta^\infty$ with $\theta^\infty \in \text{crit } F$. ∎

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

Theorem 4.8. *In the setting of Theorem* 4.7*, let L be a KŁ function with $\phi(s) = cs^{1-\theta}$ for some constant $c > 0$ and $\theta \in [0, 1)$. Let $\{\mathbf{z}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{y}^\nu)\}$ be a sequence generated by DOGT algorithm* (4.1)*. Then, there hold*

(i) *if $\theta = 0$, $\{\mathbf{z}^\nu\}$ converges to $\mathbf{z}^\infty$ in a finite number of iterations;*

(ii) *if $\theta \in (0, 1/2]$, then $||\mathbf{z}^\nu - \mathbf{z}^\infty|| \le C\tau^\nu$ for all $\nu \ge \bar{\nu}$ for some $\tau \in [0, 1)$, $\bar{\nu} \in \mathbb{N}_+$, $C > 0$;*

(iii) *if $\theta \in (1/2, 1)$, then $||\mathbf{z}^\nu - \mathbf{z}^\infty|| \le C\nu^{-\frac{1-\theta}{2\theta-1}}$ for all $\nu \ge \bar{\nu}$ for some $\bar{\nu} \in \mathbb{N}_+$, $C > 0$.*

Proof. For the sake of simplicity of notation, denote $d^\nu \triangleq d(\mathbf{x}^\nu, \mathbf{y}^\nu)$ and define $D^\nu \triangleq \sum_{t=\nu}^\infty d^t$. By (4.37), we have

$$\left\|\mathbf{z}^{\nu+1} - \mathbf{z}^\infty\right\| \le \sum_{t=\nu}^{\infty} \left\|\mathbf{z}^{t+1} - \mathbf{z}^t\right\| \le KD^\nu. \tag{4.43}$$

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

By KŁ inequality (2.1) and (4.29), we have

$$Md^\nu \phi'(l^\nu - l^\infty) \ge 1 \implies \tilde{M}(d^\nu)^{(1-\theta)/\theta} \ge (l^\nu - l^\infty)^{1-\theta} \quad \forall \nu \ge \bar{\nu} \tag{4.44}$$

for sufficiently large $\bar{\nu}$, 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 (4.40) (setting $\acute{l} = l^\infty$), we have $D^\nu \le M\phi(l^\nu - l^\infty) = Mc(l^\nu - l^\infty)^{1-\theta}$, which together with (4.44), yields

$$D^\nu \le \tilde{M}Mc(d^\nu)^{(1-\theta)/\theta} = \tilde{M}Mc(D^\nu - D^{\nu+1})^{(1-\theta)/\theta} \quad \forall \nu \ge \bar{\nu}. \tag{4.45}$$

The convergence rate estimates as stated in the theorem can be derived from (4.45), using the same line of analysis introduced in [5]. The remaining part of the proof is provided in Appendix A.3 for completeness. ∎

4.3. Second-order guarantees.

We prove that the DOGT algorithm almost surely converges to SoS solutions of (P) under a suitably chosen initialization and some additional conditions on the weight matrices $\mathbf{R}$ and $\mathbf{C}$. Following a path first established in [42] and further developed in [41], the key to our argument for the nonconvergence to strict saddle points of F lies in formulating the DOGT algorithm as a dynamical system while leveraging an instantiation of the stable manifold theorem, as given in [41, Theorem 2]. The nontrivial task is finding a self-map representing DOGT so that the stable set of the strict saddles of F is zero measure with respect to the domain of the mapping; note that the domain of the map—which is the set of initialization points—is not full dimensional and is the same as the support of the probability measure.

Our analysis is organized in the following three steps: (1) section 4.3.1 introduces the preparatory background; (2) section 4.3.2 tailors the results of step 1 to the DOGT algorithm; and (3) finally, section 4.3.3 states our main results about convergence of the DOGT algorithm to SoS solutions of (P).

4.3.1. The stable manifold theorem and unstable fixed points.

Let $g : \mathcal{S} \to \mathcal{S}$ be a mapping from $\mathcal{S}$ to itself, where $\mathcal{S}$ is a manifold without boundary.

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-fold composition of g. Our focus is on the analysis of the trajectories of the dynamical system around the fixed points of g; in particular we are interested in the set of unstable fixed points of g. We begin by introducing the following definition.

DEFINITION 4.9 (Chapter 3 of [1]). *The differential of the mapping $g : \mathcal{S} \to \mathcal{S}$, 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 tangent space of $\mathcal{S}$ at $\mathbf{u} \in \mathcal{S}$. Given a curve γ in $\mathcal{S}$ with $\gamma(0) = \mathbf{u}$ and $\frac{d\gamma}{dt}(0) = \mathbf{v} \in \mathcal{T}(\mathbf{u})$, 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 the linear operator $\det(\mathrm{D}g(\mathbf{u}))$ is the determinant of the matrix representing $\mathrm{D}g(\mathbf{u})$ with respect to a standard basis.*[3]

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

DEFINITION 4.10 (unstable fixed points). *The set of unstable fixed points of g is defined as*

$$\mathcal{A}_g = \Big\{ \mathbf{u} : g(\mathbf{u}) = \mathbf{u},\ \mathrm{spradii}\big(\mathrm{D}g(\mathbf{u})\big) > 1 \Big\}. \tag{4.46}$$

The theorem below, which is based on the stable manifold theorem [60, Theorem III.7], provides tools to let us connect $\mathcal{A}_g$ with the set of limit points which $\{\mathbf{u}^{\nu}\}$ can escape from.

THEOREM 4.11 (see [41, Theorem 2]). *Let $g : \mathcal{S} \to \mathcal{S}$ be a $\mathcal{C}^1$ mapping and*

$$\det\left(\mathrm{D}g(\mathbf{u})\right) \neq 0 \quad \forall \mathbf{u} \in \mathcal{S}.$$

Consider any nonatomic probability measure $\mathbb{P}_{\mathbf{u}^0}$ on $\mathcal{S}$ defining the choice of an initial point. Then, the set of initial points that converge to an unstable fixed point (termed stable set of $\mathcal{A}_g$) is of measure zero, i.e.,

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

4.3.2. DOGT as a dynamical system. Theorem 4.11 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 nonconvergence of $g^{\nu}(\mathbf{u}^0)$, $\mathbf{u}^0 \in \mathcal{S}$, to $\mathcal{A}_g$ implies the nonconvergence of the DOGT algorithm to strict saddles of F.

We begin rewriting the DOGT in an equivalent and more convenient form. Define $\mathbf{h}^{\nu} \triangleq \mathbf{y}^{\nu} - \nabla F_c(\mathbf{x}^{\nu})$; (4.1) can be rewritten as

$$\begin{cases} \mathbf{x}^{\nu+1} = \mathbf{W}_R\mathbf{x}^{\nu} - \alpha\left(\mathbf{h}^{\nu} + \nabla F_c(\mathbf{x}^{\nu})\right), \\ \mathbf{h}^{\nu+1} = \mathbf{W}_C\mathbf{h}^{\nu} + \left(\mathbf{W}_C - \mathbf{I}\right)\nabla F_c(\mathbf{x}^{\nu}), \end{cases} \tag{4.47}$$

with arbitrary $\mathbf{x}^0 \in \mathbb{R}^{nm}$ and $\mathbf{h}^0 \in \mathrm{span}(\mathbf{W}_C - \mathbf{I})$. By Theorem 4.5, every limit point $(\mathbf{x}^{\infty}, \mathbf{h}^{\infty})$ of $\{(\mathbf{x}^{\nu}, \mathbf{h}^{\nu})\}$ 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 $\boldsymbol{\theta}^{\infty} \in \mathrm{crit}\ F$. We are interested in the nonconvergence of (4.47) to such points whenever $\boldsymbol{\theta}^{\infty} \in \mathrm{crit}\ F$ is a strict saddle of F. This motivates the following definition.

DEFINITION 4.12 (consensual strict saddle points). *Let*

$$\Theta^*_{ss} = \{\boldsymbol{\theta}^* \in \mathrm{crit}\ F \,:\, \lambda_{min}(\nabla^2 F(\boldsymbol{\theta}^*)) < 0\}$$

[3] 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.

denote the set of strict saddles of F. The set of consensual strict saddle points *is defined as*

$$\mathcal{U}^* \triangleq \left\{ \begin{bmatrix} \mathbf{1}_n \otimes \boldsymbol{\theta}^* \\ -\nabla F_c(\mathbf{1}_n \otimes \boldsymbol{\theta}^*) \end{bmatrix} : \boldsymbol{\theta}^\star \in \Theta^*_{ss} \right\}. \tag{4.48}$$

Roughly speaking, $\mathcal{U}^*$ represents the candidate set of "adversarial" limit points which any sequence generated by (4.47) should escape from. The next step is then to write (4.47) as a proper dynamical system whose mapping satisfies conditions in Theorem 4.11 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 $\mathbf{x} \triangleq [\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top]^\top$, $\mathbf{h} \triangleq [\mathbf{h}_1^\top, \dots, \mathbf{h}_n^\top]^\top$, 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

$$\mathbf{u}^{\nu+1} = g(\mathbf{u}^\nu) \quad \text{with} \quad g(\mathbf{u}) \triangleq \begin{bmatrix} \mathbf{W}_R\mathbf{x} - \alpha\nabla F_c(\mathbf{x}) - \alpha\mathbf{h} \\ \mathbf{W}_C\mathbf{h} + (\mathbf{W}_C - \mathbf{I})\nabla F_c(\mathbf{x}) \end{bmatrix}, \tag{4.49}$$

and $\mathbf{u}^0 \in \mathbb{R}^{nm} \times \operatorname{span}(\mathbf{W}_C - \mathbf{I})$. The fixed-point iterate (4.49) describes the trajectory generated by the DOGT algorithm (4.47). However, the initialization imposed by DOGT leads to a g that maps $\mathbb{R}^{nm} \times \operatorname{span}(\mathbf{W}_C - \mathbf{I})$ into $\mathbb{R}^{nm} \times \mathbb{R}^{nm}$. We show next how to unify the domain and codomain of g to a subspace $\mathcal{S} \subseteq \mathbb{R}^{nm} \times \mathbb{R}^{nm}$ as in the form of the mapping in Theorem 4.11.

Applying (4.47) telescopically to the update of the h-variables yields $\mathbf{h}^\nu = \mathbf{W}_C^\nu\mathbf{h}^0 + (\mathbf{W}_C - \mathbf{I})\,\mathbf{g}^\nu_{\text{acc}}$ for all $\nu \geq 1$, where $\mathbf{g}^\nu_{\text{acc}} \triangleq \sum_{t=0}^{\nu-1} \mathbf{W}_C^t \nabla F_c\left(\mathbf{x}^{\nu-t-1}\right)$. Denoting $\bar{\mathbf{h}}^\nu \triangleq (\mathbf{1}_n^\top \otimes \mathbf{I}_m)\mathbf{h}^\nu$, we have

$$\bar{\mathbf{h}}^\nu = \cdots = \bar{\mathbf{h}}^0 \quad \text{and} \quad \mathbf{h}^\nu \in \mathbf{W}_C^\nu\mathbf{h}^0 + \operatorname{span}(\mathbf{W}_C - \mathbf{I}) \quad \forall \nu \geq 1. \tag{4.50}$$

The initialization $\mathbf{h}^0 \in \operatorname{span}(\mathbf{W}_C - \mathbf{I})$ in (4.47) naturally suggests the following $(2n-1)m$-dimensional linear subspace as the candidate set $\mathcal{S}$:

$$\mathcal{S} \triangleq \mathbb{R}^{nm} \times \operatorname{span}(\mathbf{W}_C - \mathbf{I}). \tag{4.51}$$

Such an $\mathcal{S}$ also ensures that $g : \mathcal{S} \to \mathcal{S}$. In fact, by (4.50), $\mathbf{h}^\nu \in \operatorname{span}(\mathbf{W}_C - \mathbf{I})$ for all $\nu \geq 1$, provided that $\mathbf{h}^0 \in \operatorname{span}(\mathbf{W}_C - \mathbf{I})$. Therefore, $\{g^\nu(\mathbf{u}^0)\} \subseteq \mathcal{S}$, for all $\mathbf{u}^0 \in \mathcal{S}$.

Equipped with the mapping g in (4.49) and $\mathcal{S}$ defined in (4.51), we check next that the condition in Theorem 4.11 is satisfied; we then prove that $\mathcal{U}^* \subseteq \mathcal{A}_g$.

(1) *g is a diffeomorphism.* To establish this property, we add the following extra assumption on the weight matrices $\mathbf{R}$ and $\mathbf{C}$, which is similar to Assumption 3.3 for the DGD scheme.

Assumption 4.13. Matrices $\mathbf{R} \in \mathcal{M}_n(\mathbb{R})$ and $\mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ are nonsingular.

The above condition is not particularly restrictive and it is compatible with Assumption 4.1. 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 $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{C}}$ satisfying Assumption 4.1. The new matrices still satisfy Assumption 4.1 due to the following fact: given two nonnegative matrices $\mathbf{A}, \mathbf{B} \in \mathcal{M}_n(\mathbb{R})$, if the directed graph associated with matrix $\mathbf{A}$ has a spanning tree and $\mathbf{B} \geq \rho\mathbf{A}$ for some $\rho > 0$, then the directed graph associated with matrix $\mathbf{B}$ has a spanning tree as well.

We build now the differential of g. Let $\tilde{g}$ be a smooth extension of (4.49) to $\mathbb{R}^{mn} \times \mathbb{R}^{mn}$, 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

$$\mathrm{D}\tilde{g}(\mathbf{u}) = \begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c(\mathbf{x}) & -\alpha\mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\nabla^2 F_c(\mathbf{x}) & \mathbf{W}_C \end{bmatrix}; \tag{4.52}$$

$\mathrm{D}\tilde{g}(\mathbf{u})$ 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 $\mathbf{P}_{\mathcal{T}(\mathbf{u})}$ 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 $\mathcal{S}$ is a linear subspace) and denoting by $\mathbf{U}_h \in \mathbb{R}^{mn \times m(n-1)}$ an orthonormal basis of $\operatorname{span}(\mathbf{W}_C - \mathbf{I})$, $\mathrm{D}g(\mathbf{u})$ reads

$$\mathrm{D}g(\mathbf{u}) = \begin{bmatrix} \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) & -\alpha \mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\, \nabla^2 F_c(\mathbf{x}) & \mathbf{W}_C \end{bmatrix} \mathbf{U}\mathbf{U}^\top \quad \text{with} \quad \mathbf{U} \triangleq \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_h \end{bmatrix}. \tag{4.53}$$

Note that $\mathbf{P}_{\mathcal{S}} = \mathbf{U}\mathbf{U}^\top$. We establish next the conditions for g to be a $\mathcal{C}^1$ diffeomorphism, as stated in Theorem 4.11.

Proposition 4.14. *Consider the mapping $g : \mathcal{S} \to \mathcal{S}$ defined in* (4.49)*, under Assumptions* 2.1(i), 4.1*, and* 4.13*, with $\mathcal{S}$ defined in* (4.51)*. If the step size is chosen according to*

$$0 < \alpha < \frac{\sigma_{\min}(\mathbf{CR})}{L_c}, \tag{4.54}$$

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 (4.53), $\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U}$ reads

$$\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U} = \mathbf{U}^\top \mathrm{D}\tilde{g}(\mathbf{u})\mathbf{U} = \begin{bmatrix} \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) & -\alpha \mathbf{U}_h \\ \mathbf{U}_h^\top (\mathbf{W}_C - \mathbf{I})\, \nabla^2 F_c(\mathbf{x}) & \mathbf{U}_h^T \mathbf{W}_C \mathbf{U}_h \end{bmatrix}. \tag{4.55}$$

Since $\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h$ is nonsingular, 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

$$\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U} = \mathbf{S}_1 \begin{bmatrix} \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) + \alpha \mathbf{\Phi} \left(\mathbf{W}_C - \mathbf{I}\right) \nabla^2 F_c(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h \end{bmatrix} \mathbf{S}_2, \tag{4.56}$$

where $\mathbf{\Phi} \triangleq \mathbf{U}_h \left(\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h\right)^{-1} \mathbf{U}_h^\top$, and $\mathbf{S}_1$ and $\mathbf{S}_2$ are some nonsingular matrices. By (4.56), it is sufficient to show that

$$\begin{aligned} \mathbf{S} &\triangleq \mathbf{W}_R - \alpha \nabla^2 F_c(\mathbf{x}) + \alpha \mathbf{\Phi} \left(\mathbf{W}_C - \mathbf{I}\right) \nabla^2 F_c(\mathbf{x}) \\ &= \mathbf{W}_R - \alpha \mathbf{W}_C^{-1} \nabla^2 F_c(\mathbf{x}) + \alpha \left(\mathbf{\Phi} - \mathbf{W}_C^{-1}\right) \left(\mathbf{W}_C - \mathbf{I}\right) \nabla^2 F_c(\mathbf{x}) \end{aligned} \tag{4.57}$$

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

$$\begin{aligned} \mathbf{\Phi} &= \mathbf{U}_h \left(\mathbf{U}_h^\top \mathbf{W}_C \mathbf{U}_h\right)^{-1} \mathbf{U}_h^\top = \mathbf{U}_h \left(\mathbf{I} + \mathbf{\Delta} \mathbf{U}_h\right)^{-1} \mathbf{U}_h^\top \\ &\overset{(a)}{=} \mathbf{U}_h \mathbf{U}_h^\top - \mathbf{U}_h \mathbf{\Delta} \left(\mathbf{I} + \mathbf{U}_h \mathbf{\Delta}\right)^{-1} \mathbf{U}_h \mathbf{U}_h^\top \\ &= \mathbf{U}_h \mathbf{U}_h^\top - \left(\mathbf{W}_C - \mathbf{I}\right) \mathbf{W}_C^{-1} \mathbf{U}_h \mathbf{U}_h^\top \\ &= \mathbf{W}_C^{-1} \mathbf{U}_h \mathbf{U}_h^\top, \end{aligned} \tag{4.58}$$

where for (a) we used the Woodbury identity of inverse matrices. Using (4.58) in (4.57), 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 $\mathbf{U}^\top \mathrm{D}g(\mathbf{u})\mathbf{U}$. □

(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 (4.47) 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 4.10) and the set $\mathcal{U}^*$ of consensual strict saddle points (cf. Definition 4.12). This will let us infer the instability of $\mathcal{U}^*$ from that of $\mathcal{A}_g$.

PROPOSITION 4.15. *Suppose that Assumptions* 2.1(i) *and* 4.1 *hold 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.,

$$\mathcal{U}^* \subseteq \mathcal{A}_g \tag{4.59}$$

with $\mathcal{A}_g$ *and* $\mathcal{U}^*$ *defined in* (4.46) *and* (4.48), *respectively.*

Proof. Let $\mathbf{u}^* \in \mathcal{U}^*$; $\mathbf{u}^*$ is a fixed point of g defined in (4.49). It is thus sufficient to show that $\mathrm{D}g(\mathbf{u}^*)$ has an eigenvalue with magnitude greater than one.

To do so, we begin showing that the differential $\mathrm{D}\tilde{g}(\mathbf{u}^*)$ of $\tilde{g}$ at $\mathbf{u}^*$ has an eigenvalue greater than one. Using (4.52), $\mathrm{D}\tilde{g}(\mathbf{u}^*)$ reads

$$\mathrm{D}\tilde{g}(\mathbf{u}^*) = \begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c^* & -\alpha\mathbf{I} \\ (\mathbf{W}_C - \mathbf{I})\,\nabla^2 F_c^\star & \mathbf{W}_C \end{bmatrix}, \tag{4.60}$$

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 prove

$$\det\left(\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I}\right) = 0 \quad \text{for some} \quad |\lambda_u| > 1. \tag{4.61}$$

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

$$\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I} = \tilde{\mathbf{S}}_1 \begin{bmatrix} \left(\mathrm{D}\tilde{g}(\mathbf{u}^*) - \lambda_u\mathbf{I}\right) / \left(\mathbf{W}_C - \lambda_u\mathbf{I}\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_C - \lambda_u\mathbf{I} \end{bmatrix} \tilde{\mathbf{S}}_2 \tag{4.62}$$

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 (4.62), (4.61) holds if and only if

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

or, equivalently,

$$\det\left(\mathbf{W}_R - \lambda_u\mathbf{I} - \alpha\nabla^2 F_c^* + \alpha\left(\mathbf{W}_C - \lambda_u\mathbf{I}\right)^{-1}\left(\mathbf{W}_C - \mathbf{I}\right)\nabla^2 F_c^*\right) = 0. \tag{4.63}$$

Multiplying both sides of (4.63) by $\det(\mathbf{W}_C - \lambda_u\mathbf{I})$ yields

$$Q(\lambda_u) \triangleq \det\Bigg(\underbrace{\left(\mathbf{W}_C - \lambda_u\mathbf{I}\right)\left(\mathbf{W}_R - \lambda_u\mathbf{I}\right) + \alpha(\lambda_u - 1)\nabla^2 F_c^*}_{\triangleq\mathbf{T}(\lambda_u)}\Bigg) = 0. \tag{4.64}$$

Trivially $Q(\lambda_u) > 0$ if $\lambda_u \gg 1$. Therefore, to show that (4.61) holds, it is sufficient to prove that there exists some $\lambda_u > 1$ such that $Q(\lambda_u) \leq 0$. Next, we prove this result under either condition (i) or (ii).

Suppose (i) holds; $\mathbf{R}$ and $\mathbf{C}$ are symmetric. Define $\tilde{\boldsymbol{v}} \triangleq \mathbf{1} \otimes \boldsymbol{v}$, where $\boldsymbol{v}$ is the unitary eigenvector associated with a negative eigenvalue of $\nabla^2 F(\boldsymbol{\theta}^*)$, and let $\lambda_{\min}(\nabla^2 F(\boldsymbol{\theta}^*)) = -\delta$; we can write

$$\tilde{\boldsymbol{v}}^\top \mathbf{T}(\lambda_u)\tilde{\boldsymbol{v}} = n(\lambda_u - 1)\left(\lambda_u - 1 - \alpha\delta/n\right) < 0 \tag{4.65}$$

for all $1 < \lambda_u < 1 + \alpha\delta/n$. By the Rayleigh–Ritz theorem, $\mathbf{T}(\lambda_u)$ has a negative eigenvalue, implying that there exists some real value $\bar{\lambda}_u > 1$ such that $Q(\bar{\lambda}_u) = 0$.

Suppose now that conditions (ii) holds; $\mathbf{W}_R$ and $\mathbf{W}_C$ reduce to $\mathbf{R}$ and $\mathbf{C}$, respectively. Note that $\mathbf{R}$ and $\mathbf{C}$ are now not symmetric. Let $\lambda_u = 1 + \epsilon$, and consider the Taylor expansion of

$$Q(1+\epsilon) = \det\Big(\left(\mathbf{C} - \mathbf{I}\right)\left(\mathbf{R} - \mathbf{I}\right) + \epsilon\left(\alpha\nabla^2 F_c^* + 2\mathbf{I} - \mathbf{C} - \mathbf{R}\right) + \epsilon^2\mathbf{I}\Big) \tag{4.66}$$

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

$$Q(1+\epsilon) = \operatorname{tr}\Big(\operatorname{adj}\left(\mathbf{M}\right)\mathbf{N}\Big)\epsilon + O(\epsilon^2). \tag{4.67}$$

Expanding (4.67) yields

$$\begin{aligned} Q(1+\epsilon) &= \operatorname{tr}\Big(\operatorname{adj}\left(\mathbf{R} - \mathbf{I}\right)\operatorname{adj}\left(\mathbf{C} - \mathbf{I}\right)\mathbf{N}\Big)\epsilon + O(\epsilon^2) \\ &= \operatorname{tr}\Big(\mathbf{1}\tilde{\mathbf{r}}^\top\tilde{\mathbf{c}}\mathbf{1}^\top\mathbf{N}\Big)\epsilon + O(\epsilon^2) = (\tilde{\mathbf{r}}^\top\tilde{\mathbf{c}})\mathbf{1}^\top\mathbf{N}\mathbf{1}\epsilon + O(\epsilon^2), \end{aligned} \tag{4.68}$$

where $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{c}}$ are the Perron vectors of $\mathbf{R}$ and $\mathbf{C}$, respectively. The second equality in (4.68) is due to the following fact: a rank-$(n-1)$ matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ has rank-1 adjugate matrix $\operatorname{adj}\left(\mathbf{A}\right) = \mathbf{a}\mathbf{b}^\top$, where $\mathbf{a}$ and $\mathbf{b}$ are nonzero vectors belonging to the 1-dimensional null space of $\mathbf{A}$ and $\mathbf{A}^\top$, respectively [34, section 0.8.2]. We also have $\tilde{\zeta} \triangleq \tilde{\mathbf{r}}^\top\tilde{\mathbf{c}} > 0$, due to Lemma 4.2. Furthermore, since $\boldsymbol{\theta}^* \in \Theta_{ss}^*$, $\mathbf{1}^\top\nabla^2 F_c^*\mathbf{1} \leq -\delta$ for some $\delta > 0$, and

$$Q(1+\epsilon) \leq -\delta\tilde{\zeta}\alpha\epsilon + O(\epsilon^2), \tag{4.69}$$

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 (4.61) 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

$$\begin{bmatrix} \mathbf{W}_R - \alpha\nabla^2 F_c\left(\mathbf{x}^*\right) & -\alpha\mathbf{I} \\ \left(\mathbf{W}_C - \mathbf{I}\right)\nabla^2 F_c\left(\mathbf{x}^*\right) & \mathbf{W}_C \end{bmatrix} \begin{bmatrix} \mathbf{v}_u^x \\ \mathbf{v}_u^h \end{bmatrix} = \bar{\lambda}_u \begin{bmatrix} \mathbf{v}_u^x \\ \mathbf{v}_u^h \end{bmatrix}. \tag{4.70}$$

In particular, we have $\left(\mathbf{W}_C - \mathbf{I}\right)\left(\nabla^2 F_c\left(\mathbf{x}^*\right)\mathbf{v}_u^x + \mathbf{v}_u^h\right) = (\bar{\lambda}_u - 1)\mathbf{v}_u^h$, which implies $\mathbf{v}_u^h \in \operatorname{span}(\mathbf{W}_C - \mathbf{I})$, 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

$$\mathrm{D}\tilde{g}(\mathbf{u}^*)\mathbf{v}_u = \bar{\lambda}_u\mathbf{v}_u \implies \mathrm{D}\tilde{g}(\mathbf{u}^*)\mathbf{P}_{\mathcal{S}}^\top\mathbf{v}_u = \bar{\lambda}_u\mathbf{v}_u \overset{(a)}{\Longrightarrow} \mathrm{D}g(\mathbf{u}^*)\mathbf{v}_u = \bar{\lambda}_u\mathbf{v}_u, \tag{4.71}$$

where (a) is due to $\mathrm{D}g(\mathbf{u}^*) = \mathrm{D}\tilde{g}(\mathbf{u}^*)\mathbf{P}_{\mathcal{S}}^\top$ (cf. (4.53)). Hence $(\bar{\lambda}_u, \mathbf{v}_u)$ is also an eigenpair of $\mathrm{D}g(\mathbf{u}^*)$, which completes the proof. ◻

Remark 4.16. Note that condition (i) in Proposition 4.15 implies that $\mathcal{G}_C$ and $\mathcal{G}_R$ are undirected graphs. Condition (ii) extends the network model to directed topologies under assumption $m = 1$. For sake of completeness, we relax condition (ii) in Appendix A.4 to arbitrary $m \in \mathbb{N}$, under extra (albeit mild) assumptions on the set of strict saddle points and the weight matrices $\mathbf{R}$ and $\mathbf{C}$.

4.3.3. DOGT likely converges to SoS solutions of (P).

Combining Theorem 4.11 and Propositions 4.14, and 4.15, we can readily obtain the following second-order guarantees of the DOGT algorithms.

THEOREM 4.17. *Consider problem* (P) *under Assumptions* 2.1 *and* 2.5 *and let* $\{\mathbf{u}^\nu \triangleq (\mathbf{x}^\nu, \mathbf{h}^\nu)\}$ *be the sequence generated by the DOGT algorithm* (4.47) *under the following tuning: the step size* α *satisfies* (4.25) *(or* (4.28)*) and* (4.54)*; the weight matrices* $\mathbf{C}$ *and* $\mathbf{R}$ *are chosen according to Assumptions* 4.1 *and* 4.13*; and the initialization is set to* $\mathbf{u}^0 \in \mathcal{S}$*, with* $\mathcal{S}$ *defined in* (4.51)*. Furthermore, suppose that either* (i) *or* (ii) *in Proposition* 4.15 *holds. Then, we have*

$$\mathbb{P}_{\mathbf{u}^0}\left(\lim_{\nu\to\infty} \mathbf{u}^\nu \in \mathcal{U}^*\right) = 0. \tag{4.72}$$

In addition, if F *is a KŁ function, then* $\{\mathbf{x}^\nu\}$ *converges almost surely to* $\mathbf{1} \otimes \boldsymbol{\theta}^\infty$ *at a rate determined in Theorem* 4.8*, where* $\boldsymbol{\theta}^\infty$ *is an SoS solution of* (P)*.*

Note that (4.72) implies the desired second-order guarantees only when the sequence $\{\mathbf{u}^\nu\}$ converges (i.e., the limit in (4.72) exists); otherwise (4.72) is trivially satisfied, and some limit point of $\{\mathbf{u}^\nu\}$ can belong to $\mathcal{U}^*$ with nonzero probability. A sufficient condition for the required global convergence of $\{\mathbf{u}^\nu\}$ is that F is a KŁ function, which is stated in the second part of the above theorem.

Remark 4.18 (comparison with [33]). As already discussed in section 1.1.2, the primal-dual method in [33] is applicable to (P); it is proved to almost surely converge to SoS solutions. Convergence of [33] is proved under stricter conditions on the problem than DOGT, namely, (i) the network must be undirected; and (ii) the Hessian of each local f_i must be Lipschitz continuous. It does not seem possible to extend the analysis of [33] beyond this assumption.

5. Numerical results.

In this section we test the behavior of DGD and DOGT around strict saddles on three classes of nonconvex problems, namely, (i) a quadratic function (cf. section 5.1), (ii) a classification problem based on the cross-entropy risk function using sigmoid activation functions (cf. section 5.2), and (iii) a two Gaussian mixture model (cf. section 5.3).

5.1. Nonconvex quadratic optimization.

Consider

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

where $m = 20$, $n = 10$, $\mathbf{b}_i$'s are independent and identically distributed (i.i.d.) Gaussian zero mean random vectors with standard deviation 10^3, and the $\mathbf{Q}_i$'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 $\boldsymbol{\theta}^* = \left(\sum_i \mathbf{Q}_i\right)^{-1} \sum_i \mathbf{Q}_i \mathbf{b_i}$. 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 to be 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 is $\mathbf{x}_i^0 = \boldsymbol{\theta}^* + \boldsymbol{\epsilon}_{x,i}$, $i \in [n]$. 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

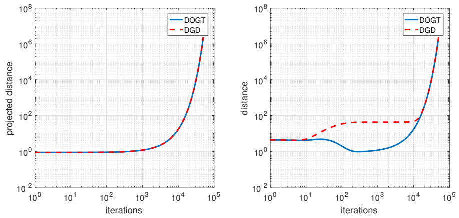

FIG. 1. *Escaping properties of DGD and DOGT, applied to Problem* (5.1). *Left plot: distance of the average iterates from $\boldsymbol{\theta}^*$ projected onto the unstable manifold E_u versus the number of iterations. Right plot: distance of the average iterates from $\boldsymbol{\theta}^*$ versus the number of iterations.*

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 3.2).

In the left panel of Figure 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 $E_u = \operatorname{span}(\mathbf{u}^u)$, 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. The figure in the left panel shows that, as predicted by our theory, both algorithms almost surely escape 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.

5.2. Bilinear logistic regression [24]. Consider a classification problem with distributed training data set $\{\mathbf{s}_i,\xi_i\}_{i=1}^n$, where $\mathbf{s}_i\in\mathbb{R}^d$ is the feature vector associated with the binary class label $\xi_i\in\{0,1\}$. The bilinear logistic regression problem aims at finding the bilinear classifier $\zeta_i(\mathbf{Q},\mathbf{w};\mathbf{s}_i) = \mathbf{s}_i^\top\mathbf{Q}\mathbf{w}$, with $\mathbf{Q}\in\mathbb{R}^{d\times p}$ and $\mathbf{w}\in\mathbb{R}^p$ that best separates data with distinct labels. Let $(\mathbf{s}_i,\xi_i)$ be private information for agent i. Using the sigmoid activation function $\sigma(x)\triangleq 1/(1+e^{-x})$ together with the *cross-entropy risk* function, the optimization problem reads

(5.2)
$$\min_{\mathbf{Q},\mathbf{w}} -\frac{1}{n}\sum_{i=1}^n\Big[\xi_i\ln\big(\sigma(\mathbf{s}_i^\top\mathbf{Q}\mathbf{w})\big)+(1-\xi_i)\ln\big(1-\sigma(\mathbf{s}_i^\top\mathbf{Q}\mathbf{w})\big)\Big]+\frac{\tau}{2}\Big(\|\mathbf{Q}\|_F^2+\|\mathbf{w}\|^2\Big).$$

It is not difficult to show that (5.2) is equivalent to the following instance of (P):

$$\min_{\mathbf{Q},\mathbf{w}}\quad F(\mathbf{Q},\mathbf{w}) = \sum_{i=1}^n \underbrace{\frac{1}{n}\Big[-\ln\big(\sigma(\tilde{\xi}_i\mathbf{s}_i^\top\mathbf{Q}\mathbf{w})\big)+\frac{\tau}{2}\Big(\|\mathbf{Q}\|_F^2+\|\mathbf{w}\|^2\Big)\Big]}_{=f_i(\mathbf{Q},\mathbf{w})} \tag{5.3}$$

with

$$\tilde{\xi}_i \triangleq \begin{cases} -1 & \text{if} \quad \xi_i = 0,\\ 1 & \text{if} \quad \xi_i = 1.\end{cases}$$

To visualize the landscape of $F(\mathbf{Q},\mathbf{w})$ (2-dimensional plot), we consider the following setting for the free parameters. We set $d=p=1$; $\tau=0.2$; $n=5$; and we generate uniformly random $\tilde{\xi}_i\in\{0,1\}$, and we draw s_i from a normal distribution with mean ξ_i and variance 1. The gradient of the local loss f_i reads

$$\begin{bmatrix}\nabla_Q f_i(Q,w)\\ \nabla_w f_i(Q,w)\end{bmatrix} = \frac{1}{n}\begin{bmatrix}\tau Q - \tilde{\xi}_i s_i w\sigma(-\tilde{\xi}_i s_i Q w)\\ \tau w - \tilde{\xi}_i s_i Q\sigma(-\tilde{\xi}_i s_i Q w)\end{bmatrix}.$$

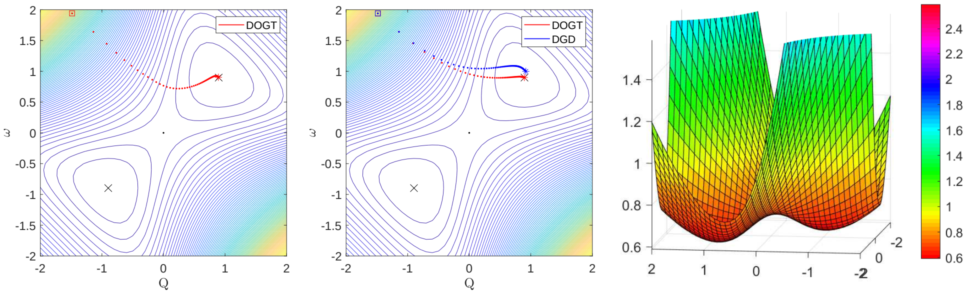

FIG. 2. *Escaping properties of the DGD and DOGT, applied to the bilinear logistic regression problem* (5.2)*. Left (resp., middle) plot: directed (resp., undirected) network; trajectory of the average iterates on the contour of F ($(0,0)$ is the strict saddle point and the $\times$ are the local minima); DGD and DOGT are initialized at $\square$ and terminated after* 100 *iterations at* $*$*. Right plot: plot of F.*

A surface plot of $F(Q,w)$ in the above setting is plotted in the right panel of Figure 2. Note that such an F has three critical points, two of which are local minima (see the location of minima in the left or middle panel of Figure 2 marked by $\times$) and one strict saddle point at $(0,0)$—the Hessian at $(0,0)$,

$$\nabla^2 F(0,0) = \begin{bmatrix} \tau & -\frac{1}{2n}\sum_i \tilde{\xi}_i s_i \\ -\frac{1}{2n}\sum_i \tilde{\xi}_i s_i & \tau \end{bmatrix},$$

has an eigenvalue at $\tau - \frac{1}{2n}\sum_i \tilde{\xi}_i s_i = -0.26$.

We test DGD and DOGT over a network of $n=5$ agents; for DGD we considered undirected graphs whereas we run DOGT on both undirected and directed graphs. Both algorithms are initialized at the same random point and terminated after 100 iterations; the step size is set to $\alpha = 0.9$. We denote by Q_i^ν and w_i^ν the agent in i's νth iterate of the local copies of Q and w, respectively. The trajectories of the average iterates $(\bar{Q}^\nu, \bar{w}^\nu) \triangleq \frac{1}{n}(\sum_i Q_i^\nu, \sum_i w_i^\nu)$ are plotted in Figure 2; the left panel refers to the directed graph while the middle panel reports the same results for the undirected network. As expected, the DOGT algorithm converges to an exact critical point (local minimum) avoiding the strict saddle $(0,0)$ while DGD converges to a neighborhood of the local minimum. The consensus error is $1/n\sqrt{\sum_{i=1}^n ||(Q_i^\nu, w_i^\nu) - (\bar{Q}^\nu, \bar{w}^\nu)||^2}$; at the termination, it reads 2.33×10^{-4} for DOGT over the directed network, and 2.18×10^{-4} and 9.74×10^{-2} for DOGT and DGD, respectively, over undirected networks.

5.3. Gaussian mixture model. Consider the Gaussian mixture model defined in section 2. The data $\{\mathbf{z}_i\}_{i=1}^n$, where $\mathbf{z}_i \in \mathbb{R}^m$ are realizations of the mixture model $\mathbf{z}_i \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Let each agent i own $\mathbf{z}_i$. Both parameters $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ and $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ are unknown. The goal is to approximate $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ while $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ is set to an estimate $(\tilde{\boldsymbol{\Sigma}}, \tilde{\boldsymbol{\Sigma}})$. The problem reads

$$\min_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^m} -\sum_{i=1}^n \log\left(\phi_m(\mathbf{z}_i - \boldsymbol{\theta}_1) + \phi_m(\mathbf{z}_i - \boldsymbol{\theta}_2)\right), \tag{5.4}$$

where $\phi_m(\boldsymbol{\theta})$ is the m-dimensional normal distribution with mean $\mathbf{0}$ and covariance $\tilde{\boldsymbol{\Sigma}}$. Consider the case of a mixture of two scalar Gaussians, i.e., $m=1$. We draw $\{\mathbf{z}_i\}_{i=1}^5$ from this model, with means $\mu_1 = 0$, $\mu_2 = -5$, and variance $\sigma_1 = \sigma_2 = 25$. The estimate of variance in problem (5.4) is pessimistically set to $\tilde{\sigma} = 1$. A surface plot of a random instance of the above problem is depicted in the right panel of Figure 3. Note

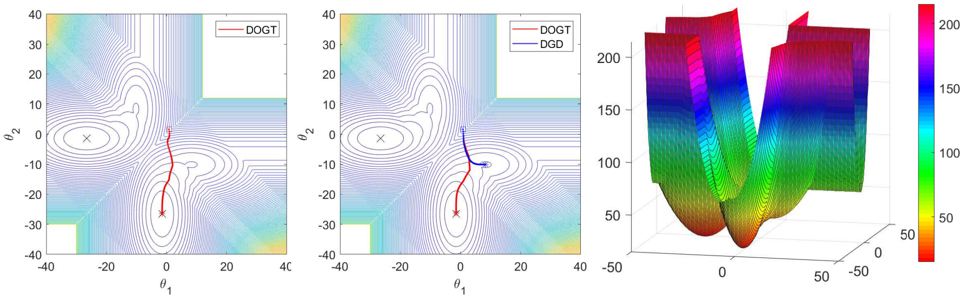

FIG. 3. *Escaping properties of the DGD and DOGT applied to the Gaussian mixture problem* (5.4). *Left (resp., middle) plot: directed (resp., undirected) network; trajectory of the average iterates on the contour of* F *(the global minima are marked by* $\times$*); DGD and DOGT are initialized at* $\square$ *and terminated after* 250 *iterations at* $*$. *Right plot: plot of* F.

that this instance of problem has 2 global minima (marked by $\times$) and multiple local minima. We test DGD and DOGT on the above problem over the same networks as described in section 5.2. Both algorithms are initialized at the same random point and terminated after 250 iterations; the step size is set to $\alpha = 0.1$. In Figure 3, we plot the trajectories of the average iterates $(\bar{\theta}_1^\nu, \bar{\theta}_2^\nu) \triangleq \frac{1}{n}(\sum_i \theta_{1,i}^\nu, \sum_i \theta_{2,i}^\nu)$, where $\theta_{1,i}^\nu$ and $\theta_{2,i}^\nu$ are the agent i's νth iterate of the local copies of θ_1 and θ_2, respectively; the left (resp., middle) panel refers to the undirected (resp., directed) network. DOGT converges to the global minimum while DGD happens to converge to a neighborhood of a local minima. The consensus error is measured by $(1/n)\sqrt{\sum_{i=1}^n ||(\theta_{1,i}^\nu, \theta_{2,i}^\nu) - (\bar{\theta}_1^\nu, \bar{\theta}_2^\nu)||^2}$ and at the termination it is equal to 1.9×10^{-3} for DOGT on the directed graph; and 2.8×10^{-3} and 1.135 for DOGT and DGD, respectively over the undirected graph.

Appendix A. Appendix.

A.1. On the problems satisfying Assumption 2.4.

We prove that all the functions arising from the examples in section 2 satisfy Assumption 2.4, for sufficiently large R and $R - \epsilon$. To do so, for each function, we establish lower bounds implying $\langle \nabla f_i(\boldsymbol{\theta}), \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|\rangle \to \infty$ as $||\boldsymbol{\theta}|| \to \infty$.

(a) *Distributed PCA.* Let us expand the objective function in (2.3) as

$$
\begin{aligned}
F(\boldsymbol{\theta}) &= \frac{1}{4}\mathrm{tr}\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\boldsymbol{\theta}\boldsymbol{\theta}^\top\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\theta}^\top\sum_{i=1}^n \mathbf{M}_i\boldsymbol{\theta}\right) + \frac{1}{4}\mathrm{tr}\left(\sum_{i=1}^n \mathbf{M}_i^\top \sum_{i=1}^n \mathbf{M}_i\right)\\
&= \sum_{i=1}^n \underbrace{\frac{1}{4}\left\{\frac{1}{n}\|\boldsymbol{\theta}\boldsymbol{\theta}^\top\|_F^2 - 2\mathrm{tr}\left(\boldsymbol{\theta}^\top\mathbf{M}_i\boldsymbol{\theta}\right)\right\}}_{\triangleq f_i(\boldsymbol{\theta})} + \frac{1}{4}\mathrm{tr}\left(\sum_{i=1}^n \mathbf{M}_i^\top \sum_{i=1}^n \mathbf{M}_i\right).
\end{aligned}
$$

We have

$$
\begin{aligned}
\langle \nabla f_i(\boldsymbol{\theta}), \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|\rangle &= \left\langle \frac{1}{n}\boldsymbol{\theta}\boldsymbol{\theta}^\top\boldsymbol{\theta} - \mathbf{M}_i\boldsymbol{\theta}, \boldsymbol{\theta}\right\rangle / \left\|\boldsymbol{\theta}\right\|\\
&= \frac{1}{n}\left\|\boldsymbol{\theta}\boldsymbol{\theta}^\top\right\|_F^2 / \left\|\boldsymbol{\theta}\right\| - \boldsymbol{\theta}^\top\mathbf{M}_i\boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|\\
&\geq \frac{1}{nK_{2,4}^4}\left\|\boldsymbol{\theta}\right\|^3 - \sigma_{\max}(\mathbf{M}_i)\left\|\boldsymbol{\theta}\right\|
\end{aligned}
$$

for some $K_{2,4} > 0$, where in the last inequality we used the equivalence of ℓ_4 and ℓ_2 norms, i.e., $||\boldsymbol{\theta}||_2 \leq K_{2,4}||\boldsymbol{\theta}||_4$ for all $\boldsymbol{\theta} \in \mathbb{R}^m$.

(b) *Phase retrieval.* It is not difficult to show that for the objective function in (2.4), it holds that

$$\begin{aligned}\langle \nabla f_i(\boldsymbol{\theta}), \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|\rangle &= \left(||\mathbf{a}_i^\top \boldsymbol{\theta}||^2 - y_i\right) ||\mathbf{a}_i^\top \boldsymbol{\theta}||^2 / \left\|\boldsymbol{\theta}\right\| + \lambda \left\|\boldsymbol{\theta}\right\| \\ &= \left(||\mathbf{a}_i^\top \boldsymbol{\theta}||^2 - y_i/2\right)^2 / \left\|\boldsymbol{\theta}\right\| - \frac{y_i^2}{4\left\|\boldsymbol{\theta}\right\|} + \lambda \left\|\boldsymbol{\theta}\right\|.\end{aligned}$$

(c) *Matrix sensing.* Consider the objective function in (2.4); it is not difficult to show that

$$\begin{aligned}\langle \nabla f_i(\boldsymbol{\Theta}), \boldsymbol{\Theta}/\left\|\boldsymbol{\Theta}\right\|_F\rangle &= \left(\operatorname{tr}\left(\boldsymbol{\Theta}^\top \mathbf{A}_i \boldsymbol{\Theta}\right) - y_i\right) \operatorname{tr}\left(\boldsymbol{\Theta}^\top \mathbf{A}_i \boldsymbol{\Theta}\right) / \left\|\boldsymbol{\Theta}\right\|_F + \lambda \left\|\boldsymbol{\Theta}\right\|_F \\ &= \operatorname{tr}\left(\boldsymbol{\Theta}^\top \mathbf{A}_i \boldsymbol{\Theta}\right)^2 / \left\|\boldsymbol{\Theta}\right\|_F - y_i \operatorname{tr}\left(\boldsymbol{\Theta}^\top \mathbf{A}_i \boldsymbol{\Theta}\right) / \left\|\boldsymbol{\Theta}\right\|_F + \lambda \left\|\boldsymbol{\Theta}\right\|_F \\ &= \left(\operatorname{tr}\left(\boldsymbol{\Theta}^\top \mathbf{A}_i \boldsymbol{\Theta}\right) - y_i/2\right)^2 / \left\|\boldsymbol{\Theta}\right\|_F - \frac{y_i^2}{4\left\|\boldsymbol{\Theta}\right\|_F} + \lambda \left\|\boldsymbol{\Theta}\right\|_F.\end{aligned}$$

(d)–(f). We prove the property only for the Gaussian mixture model; a similar proof applies also to the other classes of problems. Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_d)_{d=1}^q$. Since ϕ_m is a bounded function, we have

$$\langle \nabla_{\theta_d} f_i(\boldsymbol{\theta}_d), \boldsymbol{\theta}_d\rangle \geq -C_d + \lambda \left\|\boldsymbol{\theta}_d\right\|^2$$

for some $C_d > 0$. Hence, $\langle \nabla_\theta f_i(\boldsymbol{\theta}), \boldsymbol{\theta}/\left\|\boldsymbol{\theta}\right\|\rangle \geq -C/\left\|\boldsymbol{\theta}\right\| + \lambda \left\|\boldsymbol{\theta}\right\|$ with $C = \sum_d C_d$.

A.2. Convergence of DGD without L–smoothness of f_i's. We sketch here how to extend the convergence results of DGD stated in section 3 to the case when the gradient of the agents' loss functions is not globally Lipschitz continuous (i.e., removing Assumption 2.1(i)). Due to the space limitation, we prove only the counterpart of Theorem 3.2; the other results in section 3 can be extended following similar arguments.

We begin introducing some definitions. Under Assumptions 2.4 and 3.1, define the set $\tilde{\mathcal{Y}} \triangleq \mathcal{Y} + \mathcal{B}_b^{mn}$ with $\mathcal{Y} = \bar{\mathcal{L}} \cup \prod_{i=1}^n \mathcal{B}_R^m$ and

$$\bar{\mathcal{L}} = \mathcal{L}_{F_c}\left(\max_{\mathbf{x}_i^0 \in \mathcal{B}_R^m, i \in [n]} \left\{\sum_{i=1}^n f_i(\mathbf{x}_i^0)\right\} + \frac{R^2}{\alpha_b}\right), \tag{A.1}$$

where

$$\begin{gathered}\alpha_b = \min_{i \in [n]} \min\{\epsilon D_{ii}/h, 2D_{ii}\delta(R-\epsilon)/h^2\} > 0, \\ h = \max_{i \in [n], \mathbf{z} \in \mathcal{B}_R^m} ||\nabla f_i(\mathbf{z})||, \quad \text{and} \quad b = \max_{\alpha \in [\alpha_b, 1], \boldsymbol{\theta} \in \mathcal{Y}} ||\nabla L_\alpha(\boldsymbol{\theta})||.\end{gathered} \tag{A.2}$$

Note that, under Assumption 2.1′ (ii), $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ are compact. Hence, ∇F_c is globally Lipschitz on $\tilde{\mathcal{Y}}$, and so is ∇L_α; we denote such Lipschitz constants as $\tilde{L}_{\nabla F_c}$ and $\tilde{L}_{\nabla L_\alpha}$, respectively; it is not difficult to check that

$$\tilde{L}_{\nabla L_\alpha} = \tilde{L}_{\nabla F_c} + \frac{1 - \sigma_{\min}(\mathbf{D})}{\alpha_b}. \tag{A.3}$$

The following result replaces Theorem 3.2 in the above setting.

THEOREM A.1. *Consider problem* (P) *under Assumptions* 2.1′ (ii), 2.4*, and* 2.5. *Let* $\{\mathbf{x}^\nu\}$ *be the sequence generated by DGD in* (3.1) *under Assumption* 3.1*, with* $\mathbf{x}_i^0 \in \mathcal{B}_R^m, i \in [n]$*, and* $0 < \alpha < \bar{\alpha}_{\max} \triangleq \sigma_{\min}(\mathbf{I}+\mathbf{D})/\tilde{L}_{\nabla F_c}$*. Then the same conclusion of Theorem* 3.2 *holds.*

Proof. It is sufficient to show that $\{\mathbf{x}^\nu\} \subseteq \mathcal{Y}$; the rest of the proof follows similar steps as those in [69, Lemma 2] replacing L_c with $\tilde{L}_{\nabla F_c}$.

When $\alpha < \alpha_b$, $\{\mathbf{x}^\nu\} \subseteq \mathcal{Y}$ can be proved leveraging the same arguments used in the proof of Lemma 3.7. Therefore, in the following, we consider only the case $\alpha_b \le \alpha < \sigma_{\min}(\mathbf{I}+\mathbf{D})/\tilde{L}_{\nabla F_c}$ with $\alpha_b < \sigma_{\min}(\mathbf{I}+\mathbf{D})/\tilde{L}_{\nabla F_c}$. We prove the theorem by induction. Clearly $\mathbf{x}^0 \in \mathcal{L}_{L_\alpha}(L_\alpha(\mathbf{x}^0))$ and, by (3.6) (cf. Lemma 3.7),

$$\mathcal{L}_{L_\alpha}(L_\alpha(\mathbf{x}^0)) \subseteq \bar{\mathcal{L}} \subseteq \mathcal{Y} \quad \forall \alpha \in [\alpha_b, 1].$$

Assume $\mathcal{L}_{L_\alpha}(L_\alpha(\mathbf{x}^\nu)) \subseteq \mathcal{Y}$. Since $\mathbf{x}^\nu \in \mathcal{Y}$, there hold $\mathbf{x}^{\nu+1} = \mathbf{x}^\nu - \alpha \nabla L_\alpha(\mathbf{x}^\nu) \in \tilde{\mathcal{Y}}$ and $\theta \mathbf{x}^\nu + (1-\theta)\mathbf{x}^{\nu+1} \in \tilde{\mathcal{Y}}$ for all $\theta \in [0,1]$. Invoking the descent lemma on L_α at $\mathbf{x}^{\nu+1}$ [recall that L_α is $\tilde{L}_{\nabla L_\alpha}$-smooth on $\tilde{\mathcal{Y}}$], we have

$$L_\alpha(\mathbf{x}^{\nu+1}) \le L_\alpha(\mathbf{x}^\nu) - \alpha \left(\frac{\sigma_{\min}(\mathbf{I}+\mathbf{D}) - \alpha \tilde{L}_{\nabla F_c}}{2} \right) \|\nabla L_\alpha(\mathbf{x}^\nu)\|^2 \le L_\alpha(\mathbf{x}^\nu). \tag{A.4}$$

Therefore, $\mathcal{L}_{L_\alpha}(L_\alpha(\mathbf{x}^{\nu+1})) \subseteq \mathcal{L}_{L_\alpha}(L_\alpha(\mathbf{x}^\nu)) \subseteq \mathcal{Y}$, which completes the induction. ☐

A.3. Proof of Theorem 4.8: Supplement. 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 $\nu \ge \nu_0$ [see updates in (4.1)]; this means that $\{\mathbf{z}^\nu\}$ 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 $||\nabla L(\mathbf{x}^\nu, \mathbf{y}^\nu)|| \ge 1/c$ for all $\nu \in \mathcal{D}$. This, together with (4.21) and Lemma 4.4, lead to $l^{\nu+1} \le l^\nu - 1/(Mc)^2$, which by Assumption 2.1(ii), implies that $\mathcal{D}$ must be finite and $\{\mathbf{z}^\nu\}$ converges in a finite number of iterations.

Consider (4.45). Let $\theta \in (0, 1/2]$, then $(1-\theta)/\theta \ge 1$. Since $D^\nu \to 0$ as $\nu \to \infty$ (by Lemma 4.3(ii)), there exists a sufficiently large ν_0 such that $(D^\nu - D^{\nu+1})^{(1-\theta)/\theta} \le D^\nu - D^{\nu+1}$. By (4.45), we have

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

which proves case (ii).

Finally, let us assume $\theta \in (1/2, 1)$, then $\theta/(1-\theta) > 1$. Equation (4.45) implies

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

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

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

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

A.4. Extension of Proposition 4.15. We relax conditions (i)–(ii) of Proposition 4.15 under the following additional mild assumptions on the set of strict saddle points and the weight matrices $\mathbf{R}$ and $\mathbf{C}$.

Assumption A.2. 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 points of F; cf. Definition 4.12).

Assumption A.3. The matrices $\mathbf{R}$ and $\mathbf{C}$ are chosen according to

$$\mathbf{R} = \frac{\widetilde{\mathbf{R}} + (t-1)\mathbf{I}}{t}, \quad \mathbf{C} = \frac{\widetilde{\mathbf{C}} + (t-1)\mathbf{I}}{t}$$

for some $t \geq 1$ and some matrices $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{C}}$ satisfying Assumption 4.1.

Note that $\mathbf{R}$ and $\mathbf{C}$ satisfy Assumption 4.1 as well. The main result is given in Proposition A.5. Before proceeding, we recall the following result on spectral variation of nonnormal matrices.

THEOREM A.4 (see [10, Theorem VIII.1.1]). *For arbitrary $d \times d$ matrices $\mathbf{A}$ and $\mathbf{B}$, it holds that*

$$s\left(\sigma(\mathbf{A}), \sigma(\mathbf{B})\right) \leq \left(\|\mathbf{A}\| + \|\mathbf{B}\|\right)^{1-1/d} \|\mathbf{A} - \mathbf{B}\|^{1/d}$$

with

$$s\left(\sigma(\mathbf{A}), \sigma(\mathbf{B})\right) \triangleq \max_j \min_i |\alpha_i - \beta_j|,$$

where $\alpha_1, \ldots, \alpha_d$ and $\beta_1, \ldots, \beta_d$ are the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, respectively.

Following the same reasoning as in the proof of the proposition, it is sufficient to show that for any $\mathbf{u}^* \in \mathcal{U}^*$, the Jacobian matrix (recall from (4.60)),

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

has an eigenvalue with absolute value strictly greater than 1; proving that such an eigenpair is also a member of $\sigma(\mathrm{D}g(\mathbf{u}^*))$ follows equivalent steps as in the proof of the proposition and thus is omitted. Decompose $\mathrm{D}\tilde{g}(\mathbf{u}^*)$ as

$$\mathrm{D}\tilde{g}(\mathbf{u}^*) = \underbrace{\begin{bmatrix} \mathbf{I} - \alpha \nabla^2 F^*_c & -\alpha \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}}_{\triangleq \mathbf{Q}} + \underbrace{\frac{1}{t} \begin{bmatrix} \widetilde{\mathbf{W}}_R - \mathbf{I} & \mathbf{0} \\ (\widetilde{\mathbf{W}}_C - \mathbf{I}) \nabla F^\star_c & \widetilde{\mathbf{W}}_C - \mathbf{I} \end{bmatrix}}_{\triangleq \mathbf{P}_t}, \tag{A.6}$$

where $\widetilde{\mathbf{W}}_R \triangleq \widetilde{\mathbf{R}} \otimes \mathbf{I}_m$ and $\widetilde{\mathbf{W}}_C \triangleq \widetilde{\mathbf{C}} \otimes \mathbf{I}_m$. Equation (A.6) reads the Jacobian matrix $\mathrm{D}\tilde{g}(\mathbf{u}^*)$ as a variation of $\mathbf{Q}$ by perturbation $\mathbf{P}_t$. For any $\mathbf{u}^* \in \mathcal{U}^*$, the spectrum of $\mathbf{Q}$ consists of $n \cdot m$ counts of 1 along with the eigenvalues of $\mathbf{I} - \alpha \nabla^2 F^*_c$, which contains a real eigenvalue $\lambda_1 \geq 1 + \alpha\delta/(mn)$, since $\boldsymbol{\theta}^* \in \Theta^*_{ss}$. Theorem A.4 guarantees that the spectrum variation of any perturbed arbitrary nonnormal matrix is bounded by the norm of the perturbation matrix. Thus it is sufficient to show that the perturbed λ_1, as a member of $\sigma(\mathrm{D}\tilde{g}(\mathbf{u}^*))$, is strictly greater than 1.

Applying Theorem A.4 gives the following sufficient conditions: denote $\tilde{d} \triangleq 2mn$,

$$\left(\|\mathbf{Q} + \mathbf{P}_t\| + \|\mathbf{Q}\|\right)^{1-1/\tilde{d}} \|\mathbf{P}_t\|^{1/\tilde{d}} < 2\alpha\delta/\tilde{d}. \tag{A.7}$$

By subadditivity of the matrix norm, it is sufficient for (A.7) that

$$\left(\|\mathbf{P}_t\| + 2\|\mathbf{Q}\|\right)^{1-1/\tilde{d}} \|\mathbf{P}_t\|^{1/\tilde{d}} \leq \frac{\alpha\delta}{\tilde{d}}. \tag{A.8}$$

Since each ∇f_i is Lipschitz continuous (cf. Assumption 2.1), there exist constants $C_Q > 0$ and $C_P > 0$ such that $\max_{\mathbf{u}^* \in \mathcal{U}^*} \|\mathbf{Q}\| \leq C_Q$ and $\max_{\mathbf{u}^* \in \mathcal{U}^*} \|\mathbf{P}_t\| \leq C_P/t$. It is not difficult to show that a sufficient condition for (A.8) is

$$t \geq \frac{(C_P + 2C_Q)^{\tilde{d}-1} C_P}{(\alpha\delta/\tilde{d})^{\tilde{d}}}, \qquad \tilde{d} = 2mn. \tag{A.9}$$

Proposition A.5. *Let Assumptions* 4.1 *and* A.2 *hold, and matrices* $\mathbf{R}$ *and* $\mathbf{C}$ *be chosen according to Assumption* A.3 *with* t *satisfying* (A.9)*. Then, any consensual strict saddle point is an unstable fixed point of* g*, i.e.,* $\mathcal{U}^* \subseteq \mathcal{A}_g$ *with* $\mathcal{A}_g$ *and* $\mathcal{U}^*$ *defined in* (4.46) *and* (4.48)*, respectively.*

Note that the above proposition ensures $\mathcal{U}^* \subseteq \mathcal{A}_g$ under (A.9) and given step size α. Convergence of the sequence is proved under (4.28) and (4.54) for step size α. However, (4.28) may not hold for some large t (there can be instances where the set of step-size satisfying conditions (A.9) and (4.28) is empty). Hence, when $m > 1$, the statement in Theorem 4.17 is conditioned to the convergence of the algorithm.

REFERENCES

[1] P. A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2007.

[2] P. A. Absil, R. Mahony, and J. Trumpf, *An extrinsic look at the Riemannian Hessian*, in Geometric Science of Information, Springer, Berlin, 2013, pp. 361–368.

[3] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, *Finding approximate local minima faster than gradient descent*, in Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, New York, ACM, New York, 2017, pp. 1195–1199.

[4] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, *Finding approximate local minima faster than gradient descent*, in the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC), ACM, New York, 2017, pp. 1195–1199.

[5] 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.

[6] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Lojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[7] 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.

[8] P. Auer, M. Herbster, and M. K. Warmuth, *Exponentially many local minima for single neurons*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 1996, pp. 316–322.

[9] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, *Weighted gossip: Distributed averaging using non-doubly stochastic matrices*, in IEEE International Symposium on Information Theory, IEEE, Piscataway, NJ, 2010, pp. 1753–1757.

[10] R. Bhatia, *Matrix Analysis*, Vol. 169, Springer, New York, 1997.

[11] P. Bianchi and J. Jakubowicz, *Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization*, IEEE Trans. Automat. Control, 58 (2013), pp. 391–405.

[12] Y. Carmon and J. Duchi, *Gradient descent finds the cubic-regularized non-convex newton step*, SIAM J. Optim., 29 (2019), pp. 2146–2178.

[13] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, *Accelerated methods for nonconvex optimization*, SIAM J. Optim., 28 (2018), pp. 1751–1772.

[14] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* I: *Motivation, convergence and numerical results*, Math. Program., 127 (2011), pp. 245–295.

[15] C. Cartis, N. I. M. Gould, and P. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* II: *Worst-case function- and derivative-evaluation complexity*, Math. Program., 130 (2011), pp. 295–319.

[16] Y. Chi, Y. M. Lu, and Y. Chen, *Nonconvex optimization meets low-rank matrix factorization: An overview*, IEEE Trans. Signal Process., 67 (2019), pp. 5239–5269.

[17] F. E. Curtis, D. P. Robinson, and M. Samadi, *A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization*, Math. Program., 162 (2017), pp. 1–32.

[18] A. Daneshmand, G. Scutari, and V. Kungurtsev, *Second-Order Guarantees of Distributed Gradient Algorithms*, preprint, https://arxiv.org/abs/1809.08694v1 (2018).

[19] A. Daneshmand, G. Scutari, and V. Kungurtsev, *Second-order guarantees of gradient algorithms over networks*, in 2018 56th Annual Allerton Conference on Communcation, Control, and Computing, IEEE, Piscataway, NJ, 2018, pp. 359–365.

[20] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, in Proceedings of the 27th International Conference on Neural Information Processing System, Vol. 2, Cambridge, MA, 2014, MIT Press, Cambridge, MA, 2014, pp. 2933–2941.

[21] P. Di Lorenzo and G. Scutari, *Distributed nonconvex optimization over networks*, in IEEE International Conference on Computational Advances in Multi-Sensor Adaptive Processing, IEEE, Piscataway, NJ, 2015, pp. 229–232.

[22] P. Di Lorenzo and G. Scutari, *NEXT: In-network nonconvex optimization*, IEEE Trans. Signal Inform. Process. Netw., 2 (2016), pp. 120–136.

[23] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, *Gradient descent can take exponential time to escape saddle points*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2017, pp. 1067–1077.

[24] M. Dyrholm, C. Christoforou, and L. C. Parra, *Bilinear discriminant component analysis*, J. Mach. Learn. Res., 8 (2007), pp. 1097–1111.

[25] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[26] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points — online stochastic gradient for tensor decomposition*, Proc. Mach. Learn. Res. (PMLR), 40, (2015), pp. 797–842.

[27] R. Ge, J. D. Lee, and T. Ma, *Matrix completion has no spurious local minimum*, in Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2016, pp. 2981–2989.

[28] S. B. Gelfand and S. K. Mitter, *Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$*, SIAM J. Control Optim., 29 (1991), pp. 999–1018.

[29] A. Griewank, *The Modification of Newton's Method for Unconstrained Optimization by Bounding Cubic Terms*, Technical report NA/12, University of Cambridge, England, 1981.

[30] D. Hajinezhad and M. Hong, *Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization*, Math. Program. Ser. B, 176 (2019), pp. 207–245.

[31] P. Halmos, *Measure Theory*, Grad. Texts Math., Springer New York, 1976.

[32] M. Hong, D. Hajinezhad, and M. Zhao, *Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks*, in Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Vol. 70, International Machine Learning Society, Stroudsburg, PA, 2017, pp. 1529–1538.

[33] M. Hong, J. D. Lee, and M. Razaviyayn, *Gradient Primal-Dual Algorithm Converges to Second-Order Stationary Solutions for Nonconvex Distributed Optimization*, preprint, https://arxiv.org/abs/1802.08941 (2018).

[34] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, New York, 2012, pp. 207–245.

[35] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, *How to escape saddle points efficiently*, Proc. Mach. Learn. Res. (PMLR), 70 (2017), pp. 1724–1732.

[36] C. Jin, P. Netrapalli, and M. I. Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, Proc. Mach. Learn. Res. (PMLR), 75 (2018), pp. 1042–1085.

[37] K. Kawaguchi, *Deep learning without poor local minima*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2016, pp. 586–594.

[38] S. Krantz and H. Parks, *A Primer of Real Analytic Functions*, Birkhäuser, Boston, 2002.

[39] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. lnst. Fourier (Grenoble), 48 (1998), pp. 769–783.

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

[41] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid strict saddle points*, Math. Program., 176 (2019), pp. 311–337.

[42] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, *Gradient descent only converges to minimizers*, Proc. Mach. Learn. Res. (PMLR), 49 (2016), pp. 1246–1257.

[43] Q. Li, Z. Zhu, and G. Tang, *Alternating minimizations converge to second-order optimal solutions*, Proc. Mach. Learn. Res. (PMLR), 97 (2019), pp. 3935–3943.

[44] S. Li, G. Tang, and M. B. Wakin, *The landscape of non-convex empirical risk with degenerate population risk*, in Advances in Neural Information Processing Systems, Curran Associates, Red Hook, NY, 2019, pp. 3502–3512.

[45] S. Lu, M. Hong, and Z. Wang, *Accelerated gradient descent escapes saddle points faster than gradient descent*, in Proceedings of the 36th International Conference on Machine Learning, Vol. 97, Curran Associates, Red Hook, NY, 2018, pp. 4134–4143.

[46] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 553–572.

[47] A. Nedić and A. Olshevsky, *Distributed optimization over time-varying directed graphs*, IEEE Trans. Automat. Control, 60 (2015), pp. 601–615.

[48] A. Nedić, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM J. Optim., 27 (2017), pp. 2597–2633.

[49] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Trans. Automat. Control, 54 (2009), pp. 48–61.

[50] A. Nedić, A. Ozdaglar, and P. A. Parrilo, *Constrained consensus and optimization in multi-agent networks*, IEEE Trans. Automat. Control, 55 (2010), pp. 922–938.

[51] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Opt., Kluwer, Boston, 2004.

[52] Y. Nesterov and B. Polyak, *Cubic regularization of Newton method and its global performance*, Math. Program., 108 (2006), pp. 177–205.

[53] M. ONeill and S. J. Wright, *Behavior of accelerated gradient methods near critical points of nonconvex functions*, Math. Program., 176 (2019), pp. 403–427.

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

[55] M. J. D. Powell, *On the global convergence of trust region algorithms for unconstrained minimization*, Math. Program., 29 (1984), pp. 297–303.

[56] S. Pu, W. Shi, J. Xu, and A. Nedić, *A push-pull gradient method for distributed optimization in networks*, in 2018 IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 2018, pp. 3385–3390.

[57] G. Qu and N. Li, *Harnessing smoothness to accelerate distributed optimization*, IEEE Trans. Control Netw. Syst., 5 (2017), pp. 1245–1260.

[58] 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, Lecture Notes in Mathematics 2224, Springer, Cham, Switzerland, 2018, pp. 1–158.

[59] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Math. Program., 176 (2019), pp. 497–544.

[60] M. Shub, *Global Stability of Dynamical Systems*, Springer, New York, 1987.

[61] Y. Sun, A. Daneshmand, and G. Scutari, *Convergence Rate of Distributed Optimization Algorithms Based on Gradient Tracking*, preprint, https://arxiv.org/abs/1905.02637, 2019.

[62] Y. Sun, G. Scutari, and D. Palomar, *Distributed nonconvex multiagent optimization over time-varying networks*, in Proceedings of the 50th Asilomar Conference on Signals, Systems, and Computers, IEEE, Piscataway, NJ, 2016, pp. 788–794.

[63] T. Tatarenko and B. Touri, *Non-convex distributed optimization*, IEEE Trans. Automat. Control, 62 (2017), pp. 3744–3757.

[64] S. Vlaski and A. H. Sayed, *Distributed Learning in Non-Convex Environments–Part* I: *Agreement at a Linear Rate*, preprint, https://arxiv.org/abs/1907.01848 (2019).

[65] S. Vlaski and A. H. Sayed, *Distributed Learning in Non-Convex Environments–Part* II: *Polynomial Escape from Saddle-Points*, preprint, https://arxiv.org/abs/1907.01849 (2019).

[66] 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.

[67] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, *Distributed Stochastic Optimization with Gradient Tracking over Strongly-Connected Networks*, in 2019 IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 2019, pp. 8353–8358.

[68] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, *Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes*, in IEEE Conference on Decision Control (CDC), IEEE, Piscataway, NJ, 2015, pp. 2055–2060.

[69] K. Yuan, Q. Ling, and W. Yin, *On the convergence of decentralized gradient descent*, SIAM J. Optim., 26 (2016), pp. 1835–1854.

[70] J. Zeng and W. Yin, *On nonconvex decentralized gradient descent*, IEEE Trans. Signal Process., 66 (2018), pp. 2834–2848.

[71] L. Zhao, M. Mammadov, and J. Yearwood, *From convex to nonconvex: a loss function analysis for binary classification*, in 2010 IEEE International Conference on Data Mining Workshops, IEEE, Piscataway, NJ, 2010, pp. 1281–1288.

[72] M. Zhu and S. Martinez, *An approximate dual subgradient algorithm for multi-agent nonconvex optimization*, IEEE Trans. Automat. Control, 58 (2013), pp. 1534–1539.