Distributed Stochastic Gradient Descent and Convergence to Local Minima

Brian Swenson∗,†, Ryan Murray∗,§, Soummya Kar‡, and H. Vincent Poor†

Abstract

In centralized settings, it is well known that stochastic gradient descent (SGD) avoids saddle points. However, similar guarantees are lacking for distributed first-order algorithms in nonconvex optimization. The paper studies distributed stochastic gradient descent (D-SGD)—a simple network-based implementation of SGD. Conditions under which D-SGD converges to local minima are studied. In particular, it is shown that, for each fixed initialization, with probability 1 we have that: (i) D-SGD converges to critical points of the objective and (ii) D-SGD avoids nondegenerate saddle points. To prove these results, we use ODE-based stochastic approximation techniques. The algorithm is approximated using a continuous-time ODE which is easier to study than the (discrete-time) algorithm. Results are first derived for the continuous-time process and then extended to the discrete-time algorithm. Consequently, the paper studies continuous-time distributed gradient descent (DGD) alongside D-SGD. Because the continuous-time process is easier to study, this approach allows for simplified proof techniques and builds important intuition that is obfuscated when studying the discrete-time process alone.

I. INTRODUCTION

Nonconvex optimization has come to the forefront of machine learning, data science, and signal processing research in recent years [1]–[4]. Applications in these areas often involve large-scale optimization problems that can be efficiently handled using first-order optimization techniques, e.g., gradient descent and its variants. Beyond their efficiency, first-order algorithms are particularly popular because they are conceptually simple, easy to implement, easy to debug, computationally tractable, and achieve excellent results in practice [5], [6].

In this paper we study distributed (i.e., network-based) variants of gradient descent for large-scale nonconvex optimization. In a distributed optimization algorithm, the problem data are assumed to be distributed among a group of networked nodes. Data is collaboratively processed in-network by the nodes without any centralized coordination or centralized aggregation of data. There is a growing need for such algorithms driven by several factors. In the modern world, data is often collected and stored in a decentralized fashion. Prominent examples include the internet of things, vehicle-to-vehicle networks, and networked cyber-physical systems. It can be impractical to collect the enormous amount of information generated by these devices to a centralized location for processing [6]. No less important, user privacy concerns can make it undesirable or impossible to collect data generated by individual users to a central location [3], [9]. In these settings, it is critical to develop effective distributed algorithms for information processing [10].

The work of B. Swenson and H. V. Poor was partially supported by the Air Force Office of Scientific Research under MURI Grant FA9550-18-1-0502.

∗These authors contributed equally.
†Department of Electrical Engineering, Princeton University, Princeton, NJ (bswenson@princeton.edu, poor@princeton.edu).
§Department of Mathematical Sciences, North Carolina State University, Raleigh, NC (rwmurray@ncsu.edu).
‡Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA (soummyak@andrew.cmu.edu).
1For example, self-driving vehicles can generate upwards of 1GB of data per second [7].
Distributed algorithms are also important from the perspective of parallel computing. This is distinct from the previously mentioned applications in that data may be centrally available, but the problem is deliberately subdivided to be handled simultaneously by many processors. Traditional approaches to parallelization involve a central node that coordinates the computation. In large-scale problems, this central node can act as a bottleneck. Network-based distributed algorithms have been studied as a powerful alternative means of parallelization that can outperform classical parallelization schemes [11].

In this paper we will consider the following distributed setup. There are \( N \) nodes, or agents. Each agent \( n = 1, \ldots, N \) possesses a local function \( f_n : \mathbb{R}^d \to \mathbb{R} \), that is smooth but possibly nonconvex, and is known only to agent \( n \). Agents are assumed to be equipped with an overlaid communication network through which they may exchange information with neighboring agents. We are interested in optimizing the sum function \( f : \mathbb{R}^d \to \mathbb{R} \) given by

\[
f(x) := \sum_{n=1}^{N} f_n(x).
\]

To illustrate how such problems can arise in practice, suppose, for example, that \( D_n = \{(x_i, y_i)\}_i \) represents a local dataset collected by agent \( n \). Let \( \ell(\cdot, \cdot) \) denote some predefined loss function (e.g., \( \ell(y, \tilde{y}) = \|y - \tilde{y}\|^2 \)) and let \( h(\cdot, \theta) \) denote a parameteric hypothesis class, with parameter \( \theta \). In empirical risk minimization, the objective is to minimize the empirical risk over the data collected by all agents, i.e., solve the optimization problem

\[
\min_{\theta} \sum_{(x,y) \in \bigcup_n D_n} \ell(h(x, \theta), y) = \min_{\theta} \sum_{n=1}^{N} \sum_{(x,y) \in D_n} \ell(h(x, \theta), y),
\]

where the objective above fits the form of (1) with \( f_n(\theta) = \sum_{(x,y) \in D_n} \ell(h(x_i, \theta), y_i) \).

In centralized settings, there exist well-developed theoretical guarantees underpinning gradient based methods for nonconvex optimization [1], [12]–[16]. In this paper we are interested in establishing theoretical guarantees for distributed gradient algorithms. Our main focus will be on showing that distributed stochastic gradient descent (D-SGD) (see Section II for a formal definition) converges to local minima (or, equivalently, does not converge to saddle points) under conditions analogous to those used to study centralized algorithms. Our secondary but complementary focus will be on characterizing distributed gradient descent (DGD) in continuous time.

Our reasons for studying continuous-time DGD alongside D-SGD are twofold. First, the continuous-time DGD dynamics are easier to study using the panoply of available analysis techniques. In this setting it is (in our view) far easier both to build intuition and prove rigorous results. Second, the continuous-time DGD dynamics are the mean-field ordinary differential equation (ODE) describing the behavior D-SGD. Using powerful stochastic approximation techniques, one may extend results from the continuous-time setting to discrete-time stochastic setting. This is the route we will take.

A common assumption used for studying saddle points in the literature is that a saddle point be nondegenerate, meaning that at the saddle point, the Hessian of \( f \) is invertible (see Definition 6 below). Because we will deal with nonconvergence to these points, to avoid frequent use of double negatives we will often refer to a nondegenerate saddle point as a regular saddle point.

The main results of the paper are informally summarized as follows. (A formal presentation of the main results will be given in Section III) Under mild assumptions we show the following, where \( f \) refers to the sum function (1):

1. Continuous-time DGD and D-SGD converge to the set of critical points of \( f \).
2. Continuous-time DGD does not typically converge to regular saddle points of \( f \). In particular, we will establish a stable-manifold theorem for continuous-time DGD.
3. With probability 1, D-SGD does not converge to regular saddle points of $f$.

A few remarks are now in order concerning these results.

**Remark 1 (Convergence to Local Minima).** An immediate implication of the above results is that if all saddle points are regular, then DGD and D-SGD converge to local minima of $f$. In some literature, this is referred to as convergence to second-order stationary points [12], [18].

**Remark 2 (Stable Manifold Theorem).** In contribution 2 we establish a stable manifold theorem for DGD. In centralized settings, the stable manifold theorem plays a critical role in the analysis of gradient-based algorithms. It is the main tool used to characterize first-order optimization dynamics near saddle points [12], [16]. However, the classical stable manifold theorem does not apply in distributed settings. This is due to the inherent non-autonomous nature of the system (see Section II-A). Consequently, there is currently a poor qualitative understanding of distributed gradient dynamics near saddle points and there exist few rigorous results characterizing these fundamental properties. The stable manifold for DGD established in this paper fills this void. The stable manifold theorem for DGD elucidates the key structural properties of DGD near saddle points, and will be the cornerstone upon which our saddle-point nonconvergence results are constructed. However, we note that the stable-manifold theorem established here only holds for a restricted class of distributed gradient dynamics (see (2)–(3) and (9)–(10)). Further research is needed to characterize stable manifolds for broader classes of distributed first-order optimization algorithms.

**Remark 3 (Decaying Step Size).** It is important to remark that, because we are interested in convergence to local minima and not merely evasion of saddle points, we will consider D-SGD with a decaying step size in this paper. By choosing an appropriate decaying step size, noise is damped and the process behaves asymptotically like the mean-field ODE, converging to local minima.

**Remark 4 (Proof Techniques).** ODE-based methods for studying optimization dynamics have grown in popularity recently [19], [20]. These powerful techniques often allow for much simpler analysis and provide deep insight by characterizing the qualitative properties of the underlying ODE. However, we note that these techniques are typically used to study convergence properties, i.e., the ODE converges to some set, ergo the discrete-time algorithm converges to the same set. In contrast, in this paper, we use ODE-based stochastic approximation techniques to study nonconvergence; i.e., the ODE does not converge to some set, therefore the discrete-time algorithm does not converge to the set. This is a nontrivial issue and requires a subtler analysis. This approach was used to study centralized dynamics in [16]. To the best of our knowledge, this paper is the first time this approach has been used to study nonconvergent or unstable behavior in distributed algorithms. The proof techniques are also significant in that they provide a method for handling multiple time scale processes (see (2)–(3)). We consider these proof techniques to be a significant contribution of independent interest. More details outlining proof techniques can be found in Section II-C and in the beginning of individual sections.

**Organization.** The remainder of the paper is organized as follows. In the remainder of this section, we discuss related literature. In Section II we present the DGD and D-SGD algorithms and present the main results of the paper. Section III presents the general optimization framework in which we will prove our results. Sections IV–VI are devoted to proofs of the main results. In Section IV we study convergence to consensus and critical points for continuous- and discrete-time dynamics. In Section V we establish the existence of a stable manifold for continuous-time dynamics. In Section VI we show that discrete-time dynamics do not typically converge to saddle points. Further details about the organization of Sections IV–VI and our strategy of proof will be given in Section II after we present the main results.
We note that a discussion of notation used throughout the paper (particularly, in the proofs) can be found in Section II-D after the presentation of the main results. By way of notation, we also remark that, following standard convention, theorems, lemmas, propositions, and definitions will all be numbered using a single common counter (e.g., Theorem 1, Definition 2, Lemma 3, etc.). However, in order to more easily track the assumptions made throughout the paper, assumptions will be numbered separately using an independent counter (i.e., Assumption 1, Assumption 2, Assumption 3, etc.).

**Literature Review.** Motivated by applications in machine learning, there has been a glut of recent research on gradient-based algorithms for nonconvex optimization in classical (centralized) settings. Research on saddle point nonconvergence and saddle point escape time in centralized gradient methods includes [1], [12]–[15]. Reference [16] considers nonconvergence to unstable points (such as saddle points) in autonomous stochastic recursive dynamical systems (such as centralized gradient descent). Some of the techniques used in this paper to study D-SGD are inspired by and build off of the techniques developed in [16]. We remark, however, that the nonautonomous nature of the distributed dynamics makes the problem here more challenging.

In this paper we will establish a stable-manifold theorem for continuous-time DGD. A review of the stable-manifold theorem for classical continuous-time dynamical systems can be found in [21], and a discussion of the stable manifold theorem for discrete-time dynamical systems can be found in [22].

Distributed gradient algorithms for convex optimization have been the subject of intensive research over the past decade, see e.g., [23]–[26] and references therein. Important considerations include time-varying vs. static communication graphs [26], directed vs. undirected communication graphs [26], communication efficiency [27], and rates of convergence [25]. Distributed algorithms for nonconvex optimization have been a subject of more recent focus. The majority of work on distributed gradient methods for nonconvex optimization have focused on addressing various issues related to convergence to critical points [28]–[34]. More recent work has focused on refined convergence guarantees. References [17], [35] consider discrete-time deterministic DGD and nonconvergence to saddle points. It is shown that for sufficiently small step size, DGD avoids regular saddle points and converges to the neighborhood of a second-order stationary point from almost all initializations. The result relies on the classical stable-manifold theorem applied to an appropriate Lyapunov function that captures both the consensus dynamics and the gradient dynamics descending (1) given a fixed step size. In a similar vein, the work [36], [37] considers a constant-step size gradient-based algorithm for distributed stochastic optimization. It is shown that the algorithm avoids saddle points, and a polynomial escape time bound is established. The work [38] considers relaxed conditions on gradient noise variance to escape saddle points. A primal-dual method for distributed nonconvex optimization with local minima convergence guarantees was considered in [18].

A preliminary version of this work studying the stable manifold for continuous-time DGD can be found in [39]. We note that the present work considers the substantially more challenging problem of studying D-SGD. We also note that the present paper fills a minor gap in the proof of the stable manifold theorem in [39] which requires Assumption 6.

We also remark that distributed gradient-based algorithms for computing global minima have recently been considered in [40]–[43]. In these algorithms, noise is deliberately added in order to escape local minima and seek out global minima.

2Since the main results use standard notational conventions, and much of the notation that needs to be introduced pertains only to the proofs, we have elected to put off the presentation of notation to this point in order to minimize the build up to the main results.
II. MAIN RESULTS

We are interested in algorithms (or dynamical systems) for optimizing (1), where \( f_n : \mathbb{R}^d \to \mathbb{R} \) is the local objective function of agent (or node) \( n \), \( d \geq 1 \) denotes the ambient dimension, and \( N \) denotes the number of agents. We assume that agents are equipped with an overlaid communication network, represented by a graph \( G = (V, E) \), where the set of vertices \( V \) is the set of agents, and an edge \((i, j) \in E \) between nodes represents the ability of agents to communicate.

We will consider two algorithms: distributed stochastic gradient descent (D-SGD) and continuous-time distributed gradient descent (DGD). We will begin by defining these algorithms and briefly discussing underlying intuition. We will then present the main results of the paper for each algorithm respectively in Sections II-A and II-B below.

**Distributed SGD.** For integers \( k \geq 1 \), let \( x_n(k) \) denote agent \( n \)'s estimate of a minimizer of (1) at iteration \( k \). D-SGD is defined agentwise by the recursion

\[
x_n(k + 1) = x_n(k) + \beta_k \sum_{\ell \in \Omega_n} (x_{\ell}(k) - x_n(k)) - \alpha_k (\nabla f_n(x(k)) + \xi_n(k + 1)) ,
\]

for \( n = 1, \ldots, N \), where \( \{\alpha_k\}_{k \geq 1}, \{\beta_k\}_{k \geq 1} \subset (0, 1] \) are weight parameters, \( \{\xi_n(k)\}_{k \geq 1} \) is zero-mean noise, and \( \Omega_n \) represents the set of neighbors of agent \( n \) in the graph \( G \).

**Continuous-Time DGD.** Let \( \bar{x}_n(t) \) represent agent \( n \)'s estimate of a minimizer of (1) at time \( t \in [0, \infty) \). Continuous-time DGD is defined agentwise by the differential equation

\[
\dot{x}_n(t) = \beta_t \sum_{\ell \in \Omega_n} (x_{\ell}(t) - x_n(t)) - \alpha_t \nabla f_n(x_n(t))
\]

for \( n = 1, \ldots, N \), where \( t \mapsto \alpha_t \in (0, 1] \) and \( t \mapsto \beta_t \in (0, 1] \) are weight parameters.

**Intuition.** We first consider D-SGD. In order to see how D-SGD relates to classical (centralized) SGD, observe that the algorithm consists of two components: a consensus term \( \beta_k \sum_{\ell \in \Omega_n} (x_{\ell}(k) - x_n(k)) \) and a local (stochastic) gradient descent term \( -\alpha_k (\nabla f_n(x(k)) + \xi_n(k + 1)) \). The consensus term is related to well-studied consensus algorithms [44] (in particular, if one sets \( f_n \equiv 0 \), then (2) reduces to a standard consensus algorithm). Intuitively, the consensus term asymptotically forces each \( x_n(k) \) towards the network mean \( \bar{x}(k) := \frac{1}{N} \sum_{n=1}^{N} x_n(k) \). In turn, the network mean behaves (nearly) like a classical stochastic gradient descent process. To see this, one takes the average over agents on both sides of (2) to obtain

\[
\bar{x}(k + 1) = -\alpha_k \left( \nabla f(\bar{x}(k)) + \xi(k + 1) + r(k) \right) ,
\]

where \( r(k) \to 0 \) is a decaying perturbation term and \( \xi(k) := \frac{1}{N} \sum_{n=1}^{N} \xi_n(k) \) is the network-averaged noise. (See Section II-C for a more detailed derivation). In summary, the consensus dynamics push all agents states towards the mean which, in turn, behave approximately like stochastic gradient descent for (1).

Continuous-time DGD is the continuous-time counterpart of D-SGD. More precisely, if one takes an expected value in (2) with respect to the gradient noise (which will be assumed to be zero mean below), then (2) is a discretization of (3). (This will be seen formally in Section II-B). Continuous-time processes are generally easier to study than discrete-time processes—this is the central idea behind the ODE approach to stochastic approximation analysis [45]. Accordingly, we will analyze (2) by first characterizing the continuous-time dynamics (3) and then extending these results to (2). This will be reflected in our presentation of the main results below.
Our main results for these two processes will be presented in the next two sections. Briefly, the main results will be the following: Under mild assumptions, the following holds for both (2) and (3):

1. Agents’ states reach consensus asymptotically.
2. Agents’ states converge to the set of critical points of $f$.
3. Agents’ states avoid “nondegenerate” saddle points of $f$.
4. If all saddle points are nondegenerate, then agents’ states converge to the set of local minima of $f$.

We remark that in the case of D-SGD, these results will be shown to hold with probability 1.

A. Continuous-Time DGD: Main Results

We now present our main convergence results for continuous-time DGD (3). We will make the following assumptions.

**Assumption 1.** The graph $G = (V, E)$ is undirected and connected.

**Assumption 2.** $f_n : \mathbb{R}^d \to \mathbb{R}$ is of class $C^2$.

**Assumption 3.** $\nabla f_n$ is Lipschitz continuous.

**Assumption 4.** $f_n$ is coercive.

**Assumption 5.** $\alpha_t = \Theta(t^{-\tau_\alpha})$ and $\beta_t = \Theta(t^{-\tau_\beta})$, with $0 \leq \tau_\beta < \tau_\alpha \leq 1$, $\alpha_t, \beta_t \neq 0$.

Assumption 1 ensures that information can disseminate freely throughout the network. Assumption 2 ensures that our notion of a nondegenerate saddle point will be well defined, while Assumptions 2–3 ensure that the ODE (3) is well defined. Assumption 4 is a mild assumption used to ensure that a minimizer of (1) exists and that solutions to (3) remain in a compact set. The format assumed for the weights in Assumption 5 simplifies the analysis and ensures that consensus will be reached (since $\beta_t$ decays more slowly than $\alpha_t$).

Our first main result shows that agents reach consensus and converge to the set of critical points of $f$.

**Theorem 5** (Convergence to Critical Points). Suppose $\{x_n(t)\}_{n=1}^N$ is a solution to (3) with arbitrary initial condition and suppose that Assumptions 1–5 hold. Then for each $n = 1, \ldots, N$ we have

(i) $\lim_{t \to \infty} \|x_n(t) - x_\ell(t)\| = 0$, for all $\ell = 1, \ldots, N$.

(ii) $x_n(t)$ converges to the set of critical points of $f$.

Next, we will consider convergence to local minima, or more precisely, nonconvergence to saddle points. We say that $x^* \in \mathbb{R}^d$ is a saddle point of $f$ if $\nabla f(x^*) = 0$ and $x^*$ is neither a local maximum or minimum. We will consider saddle points satisfying the following notion of regularity.

**Definition 6** (Nondegenerate or Regular Saddle Point). A saddle point $x^*$ of $f$ will be said to be nondegenerate (or regular) if the Hessian $\nabla^2 f(x^*)$ is invertible.

The term nondegenerate is standard for this concept in the optimization literature. However, since we will deal with nonconvergence to these points, we will generally prefer to use the term “regular” to avoid frequent use of double negatives.

**Remark 7.** We note that some literature considers the behavior of gradient algorithms near so-called “strict” saddle points, i.e., saddle points where the Hessian has at least one negative eigenvalue [1]. On the other hand, at a regular saddle point, all eigenvalues of the Hessian are nonzero and there exists at least one positive and one negative eigenvalue. Thus, the assumption of a regular saddle point is stronger than that of a strict saddle point. We have chosen to focus on regular saddle points in order to simplify
the analysis. We believe that the results below also hold at strict saddle points, but we have not attempted to prove this here.

We will also require the following assumption regarding the structure of the set of local functions \( \{f_n\}_{n=1}^N \). The assumption is quite mild, but is somewhat technical.

**Assumption 6 (Continuity of Eigenvectors).** Let \( x = (x_n)_{n=1}^N \in \mathbb{R}^{Nd} \) and let \( \tilde{f}(x) := (f_n(x_n))_{n=1}^N \). Suppose that \( x^* \in \mathbb{R}^d \) is a saddle point of \( f \). Let \( \tilde{x} := (x^*, \ldots, x^*) \in \mathbb{R}^{Nd} \) be the \( N \)-fold repetition of \( x^* \). Assume that the eigenvectors of \( \nabla^2 \tilde{f}(x) \) are continuous at \( \tilde{x} \) in the sense that, for each \( x \) near \( \tilde{x} \), there exists an orthonormal matrix \( U(x) \) that diagonalizes \( \nabla^2 \tilde{f}(x) \) such that \( x \mapsto U(x) \) is continuous at \( \tilde{x} \).

We emphasize that the above assumption is relatively innocuous and is needed only to rule out certain highly pathological cases, e.g., see Example 24. We expect that the assumption will be satisfied by most functions encountered in practice. Moreover, the assumption is, in fact, is guaranteed to hold under more familiar (and less technical) conditions. For example, the assumption always holds when each eigenvalue of \( \nabla^2 f(x^*) \) (with \( f \) as in (1)) is unique or if each \( f_n \) is analytic [46]. However, we have chosen to state Assumption 6 in its present form in order to keep it as unrestricted as possible.

The next result shows that DGD typically avoids regular saddle points.

**Theorem 8 (Nonconvergence to Saddle Points).** Suppose \( \{x_n(t)\}_{n=1}^N \) is a solution to (3) with arbitrary initial condition. Let \( x^* \) be a regular saddle point of \( f \) and suppose that Assumptions 7–9 hold. Then the set of initial conditions in \( \mathbb{R}^{Nd} \) from which \( x_n(t) \) may converge to \( x^* \) for some \( n \) has Lebesgue-measure zero.

We note that in this paper we will in fact prove a more general result than given in the above theorem. In particular, the above result follows from Theorem 25 where we establish the existence of a smooth stable manifold for DGD near nondegenerate saddle points. However, to keep the presentation accessible we only present the simplified version of the result here.

We emphasize that while Theorem 8 and more generally, Theorem 25 are, in a sense, intermediate results in our analysis of (2), these results are significant in and of themselves, and of independent interest. In particular, the stable manifold theorem plays a critical role in dynamical systems theory, and, in particular, optimization. Theorem 25 establishes the existence of a stable manifold for distributed GD. Using similar techniques, it may be possible to establish stable manifold theorems for more general or varied distributed optimization processes (e.g., [32]), and thereby characterize saddle point nonconvergence more generally. We consider this to be an important direction of future research.

**B. Distributed SGD: Main Results**

We now state our main results for D-SGD (2). We will make a few additional assumptions beyond those already made for continuous-time DGD above.

**Assumption 7.** The discrete-time weight sequences satisfy \( \alpha_k = \Theta(k^{-\tau_{\alpha}}) \) and \( \beta_k = \Theta(k^{-\tau_{\beta}}) \) with \( 0 \leq \tau_{\beta} < \tau_{\alpha} \leq 1, \alpha_k, \beta_k \neq 0 \).

**Assumption 8.** \( \mathbb{E}(\xi_n(k) | F_{k-1}) = 0 \) and \( |\xi_n(k)| < B \) for some \( B > 0 \) for all \( n \in \{1, \ldots, N\} \) and \( k \geq 1 \).

Assumption 7 pertains to the discrete-time consensus and gradient weights while Assumption 8 is a relatively standard assumption for stochastic gradient descent type algorithms, ensuring that noise is zero-mean and higher order moments are bounded.

Our first main result concerning (2) is that agents reach both consensus and the set of critical points.
Theorem 9 (Convergence to Critical Points). Let \( \{x_n(k)\}_{n=1}^N \) be a distributed SGD process satisfying (2). Suppose Assumptions 1–4 and 7–8 hold. Then, given a fixed initial condition \( x_0 \), for each \( n = 1, \ldots, N \) the following hold with probability 1:

(i) Agents achieve consensus in the sense that \( \lim_{k \to \infty} \|x_n(k) - x_\ell(k)\| = 0 \) for all \( \ell = 1, \ldots, N \).

(ii) \( x_n(k) \) converges to the set of critical points of \( f \).

Our next result considers nonconvergence to saddle points. Before stating this result, we present two additional assumptions. The first assumption introduces a minimum excitation condition on the noise so that the D-SGD process cannot get stuck in “bad” sets, while the second assumption assumes \( f \) is smoother than previously assumed. In the following we let \( \bar{\xi}(k) = \frac{1}{N} \sum_{n=1}^N \xi_n(k) \) denote the mean noise.

Assumption 9 (Minimum Excitation). For some constant \( c_1 > 0 \), \( \mathbb{E}((\bar{\xi}(k) \cdot \theta)^+ | F_{k-1}) \geq c_1 \) for every unit vector \( \theta \in \mathbb{R}^d \).

Assumption 10. \( f : \mathbb{R}^d \to \mathbb{R} \) is of class \( C^3 \).

Remark 10. We note that Assumption 10 is a stronger smoothness assumption than required for our previous results (see Assumption 2). We also note that Assumption 9 (ii) deals with the average noise, not the noise of individual agents. The noise of individual agents may be degenerate (or even zero) so long as the average noise is not. For example, it could be the case that only a single agent injects noise, or that each agent injects degenerate noise along a different dimension of \( \mathbb{R}^d \).

Theorem 9 showed that D-SGD reaches critical points. The next result refines Theorem 9 by showing that the critical point reached by D-SGD cannot be a regular saddle point.

Theorem 11 (Nonconvergence to Saddle Points). Suppose \( \{x_n(k)\}_{n=1}^N \) is a D-SGD process (2) and \( x^* \) is a regular saddle point of \( f \). Suppose Assumptions 1–3, and 6–10 hold with \( \tau_\alpha \in (\frac{1}{2}, 1] \). Then, for each \( n = 1, \ldots, N \),

\[ \mathbb{P}(x_n(k) \to x^*) = 0. \]

Finally, as an immediate consequence of Theorem 11 we get the following result.

Theorem 12 (Convergence to Local Minima). Suppose \( \{x_n(k)\}_{n=1}^N \) is a D-SGD process (2). Suppose Assumptions 1–4 and 7–8 hold with \( \tau_\alpha \in (\frac{1}{2}, 1] \), and that every saddle point of \( f \) is regular. Then for each \( n = 1, \ldots, N \), \( x_n(k) \) converges to the set of local minima of \( f \), with probability 1.

C. Outline of Proof Strategy

Our approach to proving the above results will be as follows. First, we observe that the problem of minimizing (1) in a distributed setting is a special case of a general subspace-constrained optimization problem. Rather than focus only on the particular (and restrictively narrow) setting of DGD, we will study the broader problem of subspace-constrained optimization, and we will prove our main results about DGD as direct corollaries to results in this general framework.

We wish to reassure readers that the move to a more general framework will not come at the cost of a more complex presentation. In fact, we consider the effect to be the opposite—the general framework allows us to dispense with cumbersome notation associated with consensus processes and focus precisely on the simple general problem where our results apply. Proofs are simplified and intuition is more transparent.

In Section III we introduce the general subspace-constrained optimization framework. In this general framework, we introduce both a general optimization problem, and dynamics for solving this optimization problem that generalize (2)–(3).
In Section [V] we prove convergence to critical points for both continuous- and discrete-time dynamics. This proves Theorems 5 and 9. The proof of convergence to critical points uses relatively standard ODE-based stochastic approximation techniques [45].

In Section [V] we address the problem of nonconvergence to saddle points for continuous-time dynamics. In particular, we prove the existence of a stable manifold for continuous-time dynamics. This proves Theorem 8. The proof relies on a modified Perron-Lyapunov technique [47].

Finally, in Section [VI] we treat the problem of nonconvergence to saddle points for discrete-time dynamics. This section will prove Theorem 11. Our method of proof here relies critically on the stable manifold established in the previous section. In particular, we will show that the discrete-time dynamics are repelled from the continuous-time stable manifold.

D. Notation

Throughout the paper, we use bold face letters, e.g., $\mathbf{x}(t)$ to refer to continuous-time processes where $t \in \mathbb{R}$ is the time index, and we use non-bold letters, e.g., $x(k)$ to refer to discrete-time processes, where $k$ is a positive integer. When we say that a function $g : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is $C^k$ (or $g \in C^k$), $k \geq 1$ we mean that it is $k$-times continuously differentiable. If $g$ is $C^1$, we use the notation $D[g, x]$ to denote the derivative of $g$ at the point $x$. Treating $D[g, x] : \mathbb{R}^m \rightarrow \mathbb{R}^n$ as a linear operator, we use the notation $D[g, x](y)$, where $y \in \mathbb{R}^m$ to indicate the action of $D[g, x]$ on $y$. When the meaning is clear from the context, we will sometimes use the shorthand $Dg(x)$ to denote $D[g, x]$, and treat $D[g, x]$ as $n \times m$ matrices. When $g \in C^2$, then we use $D^2[g, x]$ to indicate the second derivative of $g$ at $x$, where $D^2[g, x] : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ is a bilinear operator, and we use the notation $D^2[g, x](y, z)$ to indicate its action on inputs $y, z \in \mathbb{R}^m$. Moreover, in the case that $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is $C^2$, we often use the standard notation $\nabla g$ and $\nabla^2 g$ to refer to the gradient and Hessian of $g$ respectively. Given functions $g, h : \mathbb{R} \rightarrow \mathbb{R}$ we say that $g(t) = \Theta(h(t))$ if there exist constants $a, b > 0$ such that $ah(t) \leq g(t) \leq bh(t)$ for all $t$ sufficiently large.

We will use $\| \cdot \|$ to denote the standard Euclidean norm. Given a set $\mathcal{C} \subset \mathbb{R}^d$ and point $x \in \mathbb{R}^d$, $d(x, \mathcal{C}) := \inf_{y \in \mathcal{C}} \| x - y \|$, and when we say $x(k) \rightarrow \mathcal{C}$ as $k \rightarrow \infty$, we mean that $\lim_{k \rightarrow \infty} d(x(k), \mathcal{C}) = 0$. Given $a, b \in \mathbb{R}$, $a \wedge b$ is the minimum of $a$ and $b$. $A \otimes B$ indicates the Kronecker product of matrices $A$ and $B$ of compatible dimension. Given a matrix $A \in \mathbb{R}^{d \times d}$, $\text{diag}(A)$ is the $d$-dimensional vector containing the diagonal entries of $A$. In an abuse of notation, given a vector $v \in \mathbb{R}^d$, we also use $\text{diag}(v)$ to denote the $d \times d$ diagonal matrix with entries of $v$ on the diagonal. $D_v h(x)$ denotes directional derivative in direction $v$.

Given a graph $G = (V, E)$, the set of vertices $V = \{1, \ldots, N\}$ will be used to denote the set of agents and an edge $(i, j) \in E$ will denote the ability of two agents to exchange information. In this paper we will assume $G$ is undirected. We let $\Omega_n$ denote the set of neighbors of agent $n$, i.e., the set of agents connected to $n$ by an edge (not including agent $n$), and we let $d_n = |\Omega_n|$. The graph Laplacian is given by the $N \times N$ matrix $L = D - A$, where $D = \text{diag}(d_1, \ldots, d_N)$ is the degree matrix and $A = (a_{ij})$ is the adjacency matrix defined by $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ otherwise. Further details on spectral graph theory can be found in [48].

Suppose that $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is of class $C^1$, and consider the general gradient-descent differential equation

$$\dot{x} = -\nabla F(x),$$

(4)

where $x : \mathbb{R} \rightarrow \mathbb{R}^d$ and \dot{x} denotes $\frac{d}{dt}x(t)$. We say $x$ is a solution to (4) with initial condition $x_0$ at time $t_0$ if $x$ is $C^1$, satisfies $x(t_0) = x_0$, and satisfies (4) for all $t \geq t_0$.

We will consider recursive stochastic processes $\{y(k)\}$ of the form $y(k+1) = y(k) + G(y(k), \xi(k+1), k)$, where $G : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ denotes a noise term and $k$ denotes the iteration number. We use $\mathcal{F}_k = \sigma(\{x(j), \xi(j)\}_{j=1}^k)$ to denote the filtration representing the information available at iteration $k$. 
A list of shorthand symbols commonly used throughout the paper (e.g., \( N = \) number of agents) can be found in the appendix.

III. GENERALIZED SETUP: SUBSPACE-CONSTRAINED OPTIMIZATION

The problem of minimizing (1) in a distributed setting is equivalent to the subspace-constrained optimization problem

\[
\min_{x_n \in \mathbb{R}^d} \sum_{n=1}^{N} f_n(x_n). \tag{5}
\]

Rather than focus on the particular problem (5), we will study the broader class of all subspace-constrained optimization problems, of which (5) is a special case. There are several advantages to taking this approach. Notation is simplified, proofs and intuition are more transparent, and results apply in a more general context.

In this section, we set up a simple subspace-constrained optimization problem (generalizing (5)) that will be considered in the remainder of the paper. We then set up optimization dynamics generalizing (2)–(3) for addressing the general subspace-constrained optimization problem.

The general optimization problem and dynamics will be set up in Section III-B. However, before setting up the general framework, we will first briefly discuss time-changes that yield equivalent but more convenient dynamical systems in Section III-A. After a time-change, the algorithms (2)–(3) will admit a clean and intuitive interpretation in terms of gradient descent on a penalty function. This interpretation will become clear in Section III-B.

A. Time Changes

The ODE (3) may be expressed compactly as

\[ \dot{x} = \beta_t (L \otimes I_N)x - \alpha_t (\nabla f_n(x_n))_{n=1}^N, \]

where we assume \( \alpha_t = o(\beta_t) \). At times we will find it convenient to study this ODE under a time change. In particular, assuming \( \alpha_t \neq 0 \) for \( t \geq 0 \), set \( S(t) = \int_0^t \alpha_r \, dr \) and let \( T(\tau) \) denote the inverse of \( S(t) \) so that \( T(S(t)) = t \). Letting \( y(\tau) = x(T(\tau)) \) we have

\[ \dot{y}(\tau) = \gamma_\tau (L \otimes I_N)y(\tau) - (\nabla f_n(y(\tau)))_{n=1}^N, \tag{6} \]

where \( \gamma_\tau = \frac{\beta_T(\tau)}{\alpha_T(\tau)} \to \infty \) as \( \tau \to \infty \). Likewise, if we set \( S(t) = \int_0^t \beta_r \, dr \) and let \( T(\tau) \) denote the inverse of \( S(t) \) we have

\[ \dot{y}(\tau) = (L \otimes I_N)y(\tau) - \gamma_\tau (\nabla f_n(y(\tau)))_{n=1}^N, \tag{7} \]

where \( \gamma_\tau = \frac{\alpha_T(\tau)}{\beta_T(\tau)} \to 0 \) as \( \tau \to \infty \). Thus, processes of the form (6) or (7), with \( \gamma_t \to \infty \) or \( \gamma_t \to 0 \) respectively, generalize dynamics of the form (3). When convenient we will study (6) or (7) (with associated potential \( \gamma_\tau \)) in lieu of (3).
B. Subspace-Constrained Optimization Framework

Consider the optimization problem

\[
\begin{align*}
\min_{x \in \mathbb{R}^M} & \quad h(x) \\
\text{subject to} & \quad x^\top Q x = 0,
\end{align*}
\]  

(P.1)

where \( h : \mathbb{R}^M \to \mathbb{R} \) is a \( C^2 \) function and \( Q \in \mathbb{R}^{M \times M} \) is a positive semidefinite matrix. For ease of notation we will denote the constraint set by

\[
C := \{ x \in \mathbb{R}^M : x^\top Q x = 0 \}.
\]  

(8)

Since \( Q \) is positive semidefinite, \( C \) is simply the nullspace of \( Q \). In this paper we will focus on algorithms for computing local minima of (P.1).

Continuous-Time Dynamics. Consider the following continuous-time dynamical system for solving (P.1):

\[
\dot{x} = -\nabla h(x) - \gamma_t Q x,
\]  

(9)

where the weight \( \gamma_t \to \infty \). Note that these may be viewed as the gradient descent dynamics associated with the (time-varying) function \( x \mapsto h(x) + \gamma_t x^\top Q x \), i.e.,

\[
\dot{x} = -\nabla x \left( h(x) + \gamma_t x^\top Q x \right).
\]

The term \( \gamma_t x^\top Q x \) may be thought of as a quadratic penalty term that punishes deviations from \( C \) with increasing severity as \( t \to \infty \).

Under Assumptions 1–5, the DGD dynamics (7) are a special case of this general framework in which we let the dimension be given by \( M = N d \), the state \( x \in \mathbb{R}^{N d} \) is given by the vectorization of all agents’ states \( x = \{ x_n \}_{n=1}^N \), the objective function is given by \( h(x) = \sum_{n=1}^N f_n(x_n) \), and the penalty term is generated by the matrix \( Q = (L \otimes I_N) \), where \( I_N \) is the \( N \times N \) identity matrix and \( L \) denotes the graph Laplacian of \( G \) given in Assumption 1. In this setup, the constraint set \( C \) is the consensus subspace, which is given by the nullspace of \((L \otimes I_N)\).

Discrete-Time Dynamics. Consider the following discrete-time dynamics for solving (P.1):

\[
x(k+1) = x(k) + \alpha_k \left( -\nabla h(x(k)) - \gamma_k Q x(k) + \xi(k+1) \right),
\]

(10)

where \( \gamma_k \to \infty \), \( \alpha_k \gamma_k \to 0 \), and \( \alpha_k = \Theta(\tau_\alpha) \), \( \tau_\alpha \in (1/2, 1] \), and \( \{ \xi(k) \} \) represents noise given by a martingale difference sequence. The recursion (10) may be viewed as a perturbed discretization of the ODE (9) with (diminishing) step size \( \alpha_k \) in the sense that the expected update satisfies

\[
\mathbb{E}(x(k+1)|\mathcal{F}_k) = x(k) + \alpha_k \left( -\nabla h(x(k)) - \gamma_k Q x(k) \right).
\]

Using the same reasoning as in the continuous-time case, the discrete-time DGD process (2) may be seen as a special case of (10).
C. Interpreting Results: From General Framework to DGD Framework

All of the results through the remainder of the paper will be proved in the context of the problem (P.1) and optimization dynamics (9)–(10). We take a moment now to emphasize how the main results of the paper, discussed in Section II, fit into this general framework.

Let $h|_C : C \rightarrow \mathbb{R}$ denote the restriction of $h$ to $C$.

**Definition 13.** We say that a point $x^* \in C$ is a critical point of $h|_C$ if the directional derivative $D_v h|_C(x^*) = 0$ for all directions $v \in C$.

The notion of a regular saddle point of $h|_C$ is defined similarly. The main results of the paper (Sections II-A–II-B) deal with convergence to consensus and (non)convergence of the network-averaged process to certain critical points of (1). In the context of the general framework, convergence to consensus corresponds to convergence to the constraint set $C$, and critical points of $f$ correspond to critical points of $h|_C$.

In order to show the main results in Sections II-A–II-B we will show that the general dynamics satisfy the following:

1. (9)–(10) converge to $C$,
2. (9)–(10) converge to critical points of $h|_C$,
3. (9)–(10) do not converge to regular saddle points of $h|_C$,
4. if all saddle points of $h|_C$ are regular, then (9)–(10) converge to local minima of $h|_C$,

where appropriate modifications (e.g., with probability 1) are made where necessary. More precise statements will be given as results are presented.

IV. CONVERGENCE TO CRITICAL POINTS

In this section we show that the processes (9) and (10) converge to critical points of $h|_C$ (i.e., we prove Theorems 5 and 9).

We begin by presenting several assumptions. We emphasize that the assumptions made through the remainder of the paper are distinct from all assumptions made thus far in that these later assumptions apply to the general subspace-constrained optimization framework of (P.1) and (9) and (10). We will make sufficiently broad assumptions so that our DGD algorithms are a special case of the general framework.

**Assumption 11.** $h$ is of class $C^1$.

We note that some results (e.g., demonstrating the existence of a stable manifold) will require $h$ to be smoother than $C^1$. We will assume stronger smoothness conditions later as needed.

**Assumption 12.** $h$ is coercive.

**Assumption 13.** $h$ has Lipschitz gradient.

**Assumption 14.** $Q \in \mathbb{R}^{M \times M}$ is positive semidefinite with at least one zero eigenvalue.

For the discrete-time dynamics (10) we assume the following format for $\alpha_k$ and $\gamma_k$.

**Assumption 15.** $\alpha_k = \Theta(k^{-\tau_\alpha})$ and $\gamma_k = \Theta(k^{\tau_\gamma})$ where $1/2 < \tau_\alpha \leq 1$, $0 < \tau_\gamma \leq \tau_\alpha$, $\alpha_k, \gamma_k \neq 0$.

Finally, we will assume that the noise process $\{\xi(k)\}$ in (10) satisfies the following assumption.

**Assumption 16.**

(i) $\mathbb{E}(\xi(k)|\mathcal{F}_{k-1}) = 0$ and $|\xi(k)| < B$ for some $B > 0$, for all $k \geq 1$. 

(ii) For some constant $c_1 > 0$ there holds

$$
\mathbb{E}((\xi(k) \cdot \theta)^+ | \mathcal{F}_{k-1}) \geq c_1,
$$

for every unit vector $\theta \in \mathcal{C}$.

In the particular context of D-SGD, this assumption is equivalent to the noise-relevant parts of Assumptions 8 and 9. Note that the second part of the assumption only requires noise to be nondegenerate tangential to the constraint set.

We will prove the following two results that imply Theorems 5 and 9.

**Theorem 14.** Let $x$ be a solution to (9) and suppose that Assumptions 11–14 hold, and that $\gamma_t \to \infty$. Then,

(i) $x(t) \to C$ as $t \to \infty$.

(ii) $x(t)$ converges to the set of critical points of $h\vert_C$ as $t \to \infty$.

**Theorem 15.** Let $\{x(k)\}$ be a solution to (10) and suppose that Assumptions 11–16 hold. Then,

(i) $x(k) \to C$ as $k \to \infty$.

(ii) $x(k)$ converges to the set of critical points of $h\vert_C$ as $k \to \infty$.

The proofs of Theorems 14–15 will be given in Sections IV-B–IV-C below. The proofs of these theorems will rely on techniques from the theory of stochastic approximation and perturbed differential equations. Before proceeding to the proofs, we will first briefly review relevant tools from the literature in the next section.

### A. Intermediate Results

Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ is of class $C^1$ and is coercive, and consider the general ODE (4). We will now consider the notion of a perturbed solution to (4). The class of perturbed solutions includes discrete-time (possibly stochastic) interpolated solutions to (4) as a special case. The following results show that perturbed solutions to (4) are asymptotically equivalent to classical solutions to (4). These results will allow us to study (9) and (10) by treating solutions of these systems as perturbed solutions of a simpler ODE.

**Continuous-Time Dynamics.** We will consider the following notion of a perturbed solution.

**Definition 16** (Perturbed Solution). A continuous function $y : [0, \infty) \to \mathbb{R}^d$ will be called a perturbed solution to (4) if:

1) $y$ is absolutely continuous,

2) There exists a function $r : [0, \infty) \to \mathbb{R}^d$ such that $r(t) \to 0$ as $t \to \infty$ and

$$
\frac{dy(t)}{dt} = \nabla F(y(t)) + r(t)
$$

for almost every $t > 0$.

The following result shows that the set of limit points of perturbed solutions to (4) coincide with the set of limit points of classical solutions to (4).

**Theorem 17.** Suppose $y$ is a perturbed solution to (4). Assume that $F \in C^1$ and is coercive. Then $y(t)$ converges to the set of critical points of $F$.

The theorem follows from Theorem 3.6 and Proposition 3.27 in [49].
**Discrete-Time Dynamics:** Consider a recursion of the form

\[ x(k + 1) = x(k) + \gamma_{k+1}(\nabla F(x(k)) + U_{k+1}), \]  

(11)

where \( U_{k+1} \) represents a (possibly random) perturbation and \( \gamma_k \) represents a step size. Note that this is a discretization of the of the ODE (4) (possibly subject to some perturbation \( U_k \)). The following result allows us to study the asymptotic behavior of (11) in terms of limit sets of (4).

**Theorem 18.** Suppose that \( F \in C^1 \) and \( F \) is coercive. Suppose also that the following holds with probability 1:

\[
\lim_{n \to \infty} \sup \left\{ \left\| \sum_{i=n}^{k-1} \gamma_{i+1} U_{i+1} \right\| : k = n + 1, \ldots, m(\tau_n + T) \right\} = 0
\]

where \( m(t) = \sup \{ k \geq 0 : t \geq \tau_k \} \), \( \tau_0 = 0 \) and \( \tau_k = \sum_{i=1}^n \gamma_i, k \geq 1 \). Then \( x(k) \) converges to the set of critical points of \( F \), almost surely.

The theorem is an amalgam of Theorem 3.6, Proposition 1.3, and Corollary 3.28 in [45].

Finally, in studying convergence of discrete-time processes, the following result from [50] will be useful.

**Lemma 19 (Lemma 4.1 in [50]).** Let \( \{z_k\} \) be an \( \mathbb{R}^+ \) valued sequence satisfying

\[ z_{k+1} \leq (1 - r_1(k))z_k + r_2(k), \]

where \( \{r_1(k)\} \) and \( \{r_2(k)\} \) are deterministic sequences with

\[
\frac{a_1}{(k+1)^\delta_1} \leq r_1(k) \leq 1 \quad \text{and} \quad r_2(k) \leq \frac{a_2}{(k+1)^\delta_2},
\]

and \( a_1, a_2 > 0, 0 \leq \delta_1 < 1, \delta_2 > 0 \). Then, if \( \delta_1 < \delta_2, (k+1)^\delta_z z_k \to 0 \) as \( k \to \infty \) for all \( 0 \leq \delta_0 < \delta_2 - \delta_1 \).

**B. Continuous-Time Dynamics**

We now prove Theorem 14. The theorem follows immediately from Lemmas 20 and 21 below.

**Lemma 20.** Let \( x \) be a solution to (9) and suppose that Assumptions 11–14 hold, and that \( \beta_t \to \infty \). Then \( x(t) \to \overline{C} \).

In the proof of the lemma we will use the following conventions. Without loss of generality, assume the coordinate system is rotated so that the constraint space is given by

\[ C = \{ x \in \mathbb{R}^M : x_{d+1} = \cdots = x_M = 0 \}, \]  

(12)

where we let \( d = \dim(C) \). Consistent with this form for \( C \), assume \( Q \) is diagonal with

\[ Q = \text{diag}(\tilde{Q}), \]  

(13)

where \( \tilde{Q} \in \mathbb{R}^{d \times d} \) is diagonal with all positive entries and here \( 0 \in \mathbb{R}^{(M-d) \times (M-d)} \) denotes the zero matrix. Let \( x(t) \) be decomposed as

\[ x(t) = \begin{pmatrix} \tilde{x}(t) \\ x^\perp(t) \end{pmatrix}, \]  

(14)

where \( \tilde{x}(t) \in \mathbb{R}^d \) and \( x^\perp(t) \in \mathbb{R}^{M-d} \).

We now prove Lemma 20.
Proof. Let $C$ and $Q$ be as given in (12)-(13) and let $x(t)$ be decomposed as (14). After a time change (e.g., Section III-A) (9) is equivalent to
\[
\dot{x} = \gamma_t \nabla h(x) - Qx,
\]
where $\gamma_t \to 0$. From this we have
\[
\dot{x}^\perp(t) = -\hat{Q}x^\perp(t) + \gamma_t r(x(t)),
\]
where $r(x(t)) = [-\nabla h(x(t))]_{i=1}^d$, where the notation $[.]^d_{i=1}$ indicates extracting the appropriate subvector. By Assumption 12 $x(t)$ remains in a compact set. Thus by Assumption 13 we have that $\gamma_t r(x(t)) \to 0$ as $t \to \infty$. Thus, $x^\perp(t)$ is a perturbed solution to the ODE
\[
\dot{y} = -\hat{Q}y,
\]
where $y : [0, \infty) \to \mathbb{R}^{M-d}$. Since $x \mapsto \hat{Q}x$ is the gradient of $x^\top \hat{Q}x$, by Theorem 17 we see that $x^\perp(t) \to 0$ as $t \to \infty$. \hfill \square

Lemma 21. Let $x$ be a solution to (9) and suppose that Assumptions 11-12 hold, and that $\gamma_t \to \infty$. Then $x(t)$ converges to the set of critical points of $h|_C$.

Proof. Let $C$ and $Q$ be as given in (12)-(13) and let $x(t)$ be decomposed as (14). For convenience, let \( \hat{h} : \mathbb{R}^d \to \mathbb{R} \) be the restriction of $h$ to $C$, defined as follows: for $y \in \mathbb{R}^d$, let
\[
\hat{h}(y) := h(y, 0),
\]
where here $0 \in \mathbb{R}^{M-d}$.

By Lemma 20, it is sufficient to show that $\dot{x}(t)$ converges to the critical points set of $\hat{h}$. Given the assumed coordinate system and the structure of $Q$, for $x$ satisfying (9) we have
\[
\dot{x}(t) = -[\nabla h(x(t))]_{i=1}^d = -\nabla \hat{h}(\check{x}(t)) + r(t),
\]
where $r(t) = -\left(\nabla h(x(t)) - \nabla \hat{h}(\check{x}(t))\right)$. By Assumption 13 and Lemma 20 we have $r(t) \to 0$ as $t \to \infty$. Recalling Definition 16 solutions of (15) may be viewed as perturbed solutions of the ODE
\[
\dot{z} = -\nabla \hat{h}(z),
\]
where $z : [0, \infty) \to \mathbb{R}^d$. Since $r(t) \to 0$ as $t \to \infty$, by Theorem 17 we see that solutions to (15) converge to the critical points set of $\hat{h}$. \hfill \square

C. Discrete-Time Dynamics

We now prove Theorem 15. The theorem follows immediately from Lemmas 22 and 23 below.

Lemma 22. Let $\{x(k)\}$ be a solution to (10) and suppose that Assumptions 11-14 hold, and that $\gamma_k \to \infty$. Then $x(k) \to C$ as $k \to \infty$ with probability 1.

Proof. Let $C$ and $Q$ be as given in (12)-(13) and let $x(k)$ be decomposed as
\[
x(k) = \begin{pmatrix} \check{x}(k) \\ \check{\perp}(k) \end{pmatrix},
\]
where $\check{x}(k) \in \mathbb{R}^d$ and $\check{\perp}(k) \in \mathbb{R}^{M-d}$. By (10) we have
\[
x^\perp(k + 1) = x^\perp(k) - \alpha_k \gamma_k \hat{Q} x^\perp(k) + \alpha_k r(x(k)) + \alpha_k \xi^\perp(k + 1),
\]
(17)
where \( r(x(k)) = -[\nabla h(x(k))]_{i=1}^{M} \), and \( \xi^\perp(k) = [\xi_i(k)]_{k=i+1}^{M} \). By Assumption 12, \( x(k) \) remains in a bounded set with probability 1. Thus, by Assumption 13 there exists a \( K > 0 \) such that \( \sup_{k \geq 1} |r(x(k))| < K \) with probability 1.

By Assumption 16 we may choose the previous \( K \) sufficiently large so that we also have \( \|\xi^\perp(k)\| < K \) for all \( k \). Letting \( \lambda_{\min} \) be the smallest eigenvalue of \( Q \), from (17) we have

\[
\|x^\perp(k + 1)\| = (1 - \alpha_k \gamma_k \lambda_{\min}) \|x^\perp(k)\| + \alpha_k 2K
\]

Invoking the step size characterization in Assumption 15 by Lemma 19 we have \( \|x^\perp(k)\| \to 0 \). \( \square \)

**Lemma 23.** Let \( \{x(k)\} \) be a solution to (10) and suppose that Assumptions 17-14 hold. Then \( x(k) \) converges to the set of critical points of \( h|_C \), almost surely.

**Proof.** Let \( C \) and \( Q \) be as given in (12)-(13) and let \( x(k) \) be decomposed as in (16). Let \( \hat{h} : \mathbb{R}^d \to \mathbb{R} \) represent the restriction of \( h \) to \( C \) as given in Lemma 21. Then we have

\[
\hat{x}(k + 1) = \hat{x}(k) + \alpha_k \left( -[\nabla h(x(k))]_{i=1}^{d} + \xi(k + 1) \right)
\]

where \( r(x(k)) = -[\nabla h(x(k))]_{i=1}^{d} + \hat{\xi}(x(k)) \) and \( \xi(k + 1) = [\xi_i(k + 1)]_{i=1}^{d} \). By Lemma 22 we see that \( \hat{x}^\perp(k) \to 0 \) as \( k \to \infty \), almost surely. By Assumption 13 this implies that \( r(x(k)) \) is bounded and, in particular, that \( \alpha_k r(x(k)) \to 0 \) as \( k \to \infty \), almost surely.

Thus, the process \( \{\hat{x}(k)\} \) fits the template of Theorem 18 with \( F(x) = -\nabla \hat{h}(x) \), \( V = \hat{h} \) and \( \Lambda \) equal to the set of critical points of \( \hat{h} \). Hence, by Theorem 18 \( x(k) \) converges to the set of critical points of \( \hat{h} \).

\( \square \)

V. **Continuous-Time Dynamics: Stable-Manifold Theorem**

In this section we will establish the stable-manifold theorem for continuous-time DGD. More precisely, we will prove a stable manifold theorem for the general ODE (9) which will imply Theorem 8. We will make the following assumptions.

**Assumption 17.** \( h \) is of class \( C^2 \).

**Assumption 18.** \( Q \) is positive semidefinite. Moreover, zero is an eigenvalue of \( Q \) with geometric multiplicity \( \geq 2 \).

Note that these assumptions will supersede (in the sense that they are stronger and will be used in place of) Assumptions 11 and 14 respectively. Assumption 17 will be needed to ensure that the notion of a regular saddle point of \( h \) is well defined. Assumption 18 ensures that the constraint space \( C \) has dimension at least 2 so that critical points of \( h|_C \) may be saddle points.

We will assume that the eigenvectors of \( f \) are continuous near saddle points in the following sense.

**Assumption 19.** Let \( x^* \) be a saddle point of \( h \). Assume that the eigenvectors of \( \nabla^2 h(x) \) are continuous near \( x^* \) in the sense that for each \( x \) near \( x^* \), there exists an orthonormal matrix \( U(x) \) that diagonalizes \( \nabla^2 h(x) \) such that \( x \mapsto U(x) \) is continuous at \( x^* \).

**Example 24.** Consider the following matrix-valued function (46), Example 5.3; let \( T : [0, \infty) \to \mathbb{R}^{2 \times 2} \), where

\[
T(t) := I + e^{\frac{t}{\tau}} \begin{pmatrix} \cos \left( \frac{t}{\tau} \right) & \sin \left( \frac{t}{\tau} \right) \\ \sin \left( \frac{t}{\tau} \right) & -\cos \left( \frac{t}{\tau} \right) \end{pmatrix}, \quad T(0) := I,
\]

16
The function $T$ is continuous ($C^\infty$ in fact) but there do not exist eigenvectors of $T'(t)$ that are continuous at 0. This example illustrates the pathological cases which can arise and necessitate Assumption 19 (and earlier, Assumption 3). A function $h : \mathbb{R}^M \to \mathbb{R}$ such that $T(t)$ is the Hessian of $h$ can be constructed using the Whitney extension theorem.

The following theorem demonstrates the existence of a stable manifold near nondegenerate saddle points.

**Theorem 25.** Suppose that Assumptions 13 and 17–19 hold. Assume the weight function $t \mapsto \gamma_t$ is $C^1$ and satisfies $\gamma_t \to \infty$. Suppose $x^*$ is a nondegenerate saddle point of $h|_C$ and let $p$ denote the number of positive eigenvalues of $\nabla^2 h|_C(x^*)$. Then there exists a $C^1$ manifold $S \subset [0, \infty) \times \mathbb{R}^M$ with dimension $M - p + 1$ such that the following holds: A solution $x$ to (9) converges to $x^*$ if and only if $x$ is initialized on $S$, i.e., $x(t_0) = x_0$, with $(t_0, x_0) \in S$.

Since (10) is a generalization of (2), and since $S$ is $C^1$ (and hence a measure zero set in $\mathbb{R}^M$), Theorem 25 implies Theorem 8. We remark that coercivity (Assumption 12) is not needed for existence of the stable manifold.

The remainder of the section is organized as follows. Section V-A gives a proof of Theorem 25. In Section V-B we consider a slight refinement of Theorem 25 where we show that if $h \in C^4$ (rather than $C^2$), then $S$ is of class $C^2$ (rather than $C^1$). This refined result will be useful in the study of discrete-time algorithms considered in the subsequent section.

A. Proof of Theorem 25

We will break the proof of Theorem 25 into two main parts. Lemma 27 demonstrates existence of the stable manifold, but does not show smoothness. Lemma 28 shows that the manifold is smooth. Lemmas 27, 28 together immediately prove Theorem 25.

We begin with the following preliminary lemma.

**Lemma 26.** Suppose Assumptions 14, 19 hold and suppose that 0 is a critical point of $h|_C(0)$. There exists a function $g : [0, \infty) \to \mathbb{R}^M$ such that (i) $\nabla h(g(\gamma)) - g(\gamma)^T Q = 0$ for all $\gamma$ sufficiently large and (ii) $g(\gamma) \to 0$ as $\gamma \to \infty$. Moreover, the arc length of $\{g(\gamma) : \gamma \geq \gamma_0\}$ is finite, where $\gamma_0$ is a sufficiently large constant, i.e.,

$$\int_{\gamma_0}^\infty |g'(s)| ds < \infty. \quad (18)$$

The idea of the lemma is that, since 0 is a critical point of the restricted function $h|_C$, there is a critical point of the penalized function $h(x) + \gamma x^T Q x$ near 0, for $\gamma$ large. Moreover, as $\gamma \to \infty$, this critical point of the penalized function converges to zero. The proof of the lemma is given below.

**Proof.** The lemma follows by repeated application of the implicit function theorem. Let

$$d = \dim C$$

and without loss of generality, assume that the constraint set is given by $C = \text{span}\{e_1, \ldots, e_d\}$, i.e., the span of the first $d$ canonical vectors. Let $x \in \mathbb{R}^M$ be decomposed as $x = (x_c, x_{nc})$, where $x_c \in \mathbb{R}^d$ refers to the ‘constraint’ component and $x_{nc} \in \mathbb{R}^{M-d}$ refers to the ‘not constraint’ component of $x$. Let $G_c : \mathbb{R}^M \to \mathbb{R}^d$ be given by

$$G_c(x_c, x_{nc}) := D_{x_c} \left( h(x_c, x_{nc}) + x^T Q x \right) = D_{x_c} h(x_c, x_{nc}),$$
where the second line follows from the fact that, by construction, $Q$ is null in directions along the constraint set. Observe that $G_c$ is $C^1$ and $G_c(0, 0) = 0$. Recalling that $\nabla^2 h|_c(0)$ is invertible (i.e., $D^2_{xx} h(x_c, y_c)(x_c, y_c) = (0, 0)$ is invertible), the implicit function theorem implies that there exists a function $x^c: \mathbb{R}^{M-d} \to \mathbb{R}^d$ such that

$$G_c(x^c(x_{nc}), x_{nc}) = 0$$

for $x_{nc}$ in a neighborhood of zero.

Given that $C = \text{span}\{e_1, \ldots, e_d\}$, the matrix $Q$ takes the form $Q = \text{diag}(0, Q_{nc})$, where $0 \in \mathbb{R}^{d \times d}$ is the zero matrix, and $Q_{nc} \in \mathbb{R}^{(M-d) \times (M-d)}$ is positive definite.

For $\tau \geq 0$, let $G_{nc}: \mathbb{R}^M \to \mathbb{R}^{M-m}$ be given by

$$G_{nc}(\tau, x_{nc}) := \tau D_{x_{nc}} h(x^c(x_{nc}), x_{nc}) + x_{nc}^T Q_{nc},$$

where, in an abuse of notation, by $D_{x_{nc}} h(x^c(x_{nc}), x_{nc})$ we mean $D_{x_{nc}} h(x^c(x_{nc}), x_{nc})$. Note that $G_{nc}$ is $C^1$, $G_{nc}(0, 0) = 0$, and $D_{x_{nc}} G_{nc}(\tau, x_{nc})|_{(\tau, x_{nc})=(0,0)} = Q_{nc}$, which is invertible. By the implicit function theorem there exists a function $x_{nc}(\tau)$ such that $G_{nc}(\tau, x_{nc}(\tau)) = 0$ for $\tau$ near zero.

For $\gamma > 0$ sufficiently large let $g(\gamma) := (x^c(x_{nc}(\gamma)), x_{nc}(\gamma))$. By construction, for all $\gamma$ sufficiently large, $\frac{1}{\gamma} \nabla h(x) + x^T Q = 0$, or equivalently, $\nabla h(x) + \gamma x^T Q = 0$ for $x = g(\gamma)$.

The integrability claim (18) follows by noting that $\tau \to \hat{x}(\tau) := (x^c(x_{nc}(\tau)), x_{nc}(\tau))$ is $C^1$ (by our use of the implicit function theorem), and the integral (18) is equivalent to $\int_0^{\tau_1} |D_{\tau} \hat{x}(\tau)| d\tau$ for some finite $\tau_1$. Since $\hat{x}$ is $C^1$, the integral is finite.

The next lemma establishes the existence of a stable manifold. The proof technique relies on an adaptation of the classic Perron-Lyapunov method [47] tailored to the particular nonautonomous dynamical system (9).

**Lemma 27.** Suppose Assumptions [13] and [17-19] hold. Let $h$, $\gamma_t$, and $x^*$ be as in Theorem 25. Then for all $t_0$ sufficiently large there exists a manifold $S \subset [0, \infty) \times \mathbb{R}^M$ with dimension $M - p + 1$ such that the following holds: A solution $x$ to (9) converges to $x^*$ if and only if $x$ is initialized on $S$, i.e., $x(t_0) = x_0$ with $(t_0, x_0) \in S$.

**Proof.** 1. (Recenter) Without loss of generality we will assume that $x^* = 0$. By Lemma 26 there exists a function $g \in C^1([0, \infty), \mathbb{R}^M)$ such that, for each $\gamma \geq 0$ sufficiently large, $g(\gamma)$ is a critical point of the penalized function $h(x) + \gamma x^T Q x$ and $g(\gamma) \to 0$ as $\gamma \to \infty$.

Letting $y(t) = x(t) - g(\gamma t)$ we see that $x$ is a solution to (9) if and only if $y$ is a solution to

$$\dot{y} = \nabla_x h(y + g(\gamma t)) - \gamma t Q(y + g(\gamma t)) - g'(\gamma t) \gamma t,$$  

(19)

where we use the notation $g'(\gamma)$ to denote $Dg(\gamma)$. For $t \geq 0$ let

$$A(t) := \nabla^2_x (h(x) + \gamma t x^T Q x) \big|_{x=g(\gamma t)}$$

(20)

and let

$$F(y, t) := -\nabla_x h(y + g(\gamma t)) - \gamma t Q(y + g(\gamma t)) - A(t)y$$

(21)

so that we may express (19) as

$$\dot{y}(t) = A(t)y(t) + F(y(t), t) - g'(\gamma t) \gamma t.$$  

(22)

2. (Diagonalize) For each $t \geq 0$, let $U(t)$ be a unitary matrix that diagonalizes $A(t)$, so that

$$A(t) := U(t)^{-1} A(t) U(t)^{-1},$$

(23)

In this context, the notation $\nabla$ and $D$ are both used refer to the gradient and are used interchangeably.
where \( \Lambda(t) \) is diagonal. Since \( \gamma_t \in C^1 \) we may construct \( U(t) \) as a differentiable function of \( t \), by Assumption 19 we may construct \( U(t) \) as a differentiable function with \( U(t) \) that converges to some fixed matrix as \( t \to \infty \) (or, equivalently, as \( g(\gamma_t) \to 0 \)). Changing coordinates again, let \( z(t) = U(t)y(t) \) so that \( y \) is a solution to (22) if and only if \( z \) is a solution to

\[
\dot{z}(t) = U(t)\dot{y}(t) + \dot{U}(t)y(t)
\]

\[
= U(t) \left( A(t)U(t)^T z(t) + F(U(t)^T z(t), t) - g'(\gamma_t)\gamma_t \right) + \dot{U}(t)U(t)^T z(t)
\]

Letting

\[ \tilde{F}(z, t) := U(t)F(U(t)^T z(t), t) + \dot{U}(t)U(t)z, \] (24)

the above is equivalent to

\[
\dot{z}(t) = \Lambda(t)z(t) + \tilde{F}(z(t), t) - U(t)g'(\gamma_t)\gamma_t. \] (25)

Note that \( F(0, t) = 0 \) and \( F(y, t) = o(|y|^2) \) for \( t \geq 0 \). Consequently, for any \( \epsilon > 0 \) there exists an \( r > 0 \) and a \( T \geq 0 \) such that for all \( t \geq T \) we have

\[
|\tilde{F}(z, t) - \tilde{F}(\bar{z}, t)| \leq \epsilon |z - \bar{z}|, \quad \forall z, \bar{z} \in B_r(0) \] (26)

3. (Compute Stable Solutions) Let \( \lambda_1(t), \ldots, \lambda_M(t) \) denote the eigenvalues of \( \Lambda(t) \). Without loss of generality, we may assume that the eigenvalues are ordered so each \( \lambda_i(t) \) varies smoothly in \( t \) (see Theorem II.5.1 in [46]). Let

\[
B = \nabla^2 h |c(0). \]

and let \( \lambda_1, \ldots, \lambda_d \) denote the eigenvalues of \( B \). By Lemma 45 in the appendix, for each eigenvalue \( \lambda_i \) of \( B \), there exists an eigenvalue \( \lambda_i(t) \) of \( \Lambda(t) \) such that \( \lambda_i(t) \to \lambda_i \). Moreover, for each remaining eigenvalue of \( \Lambda(t) \) there holds \( \lambda_i(t) \to \infty \). Given the limits established for each \( \lambda_i(t) \), there exists a \( T \) sufficiently large such that for each \( i \) the sign of \( \lambda_i(t) \) remains constant for \( t \geq T \).

Without loss of generality assume that the coordinates are ordered so that the first \( n_s < M \) diagonal entries of \( \Lambda(t) \) are negative and the remaining \( M - n_s \) diagonal entries are positive for all \( t \) sufficiently large. (The notation \( n_s \) is indicative of number of “stable” eigenvalues.) Let \( \Lambda(t) \) be decomposed as

\[
\Lambda(t) = \begin{pmatrix} \Lambda^s(t) & 0 \\ 0 & \Lambda^u(t) \end{pmatrix} \] (27)

where \( \Lambda^s(t) \in \mathbb{R}^{n_s \times n_s} \) and \( \Lambda^u(t) \in \mathbb{R}^{(M-n_s) \times (M-n_s)} \) denote the ‘stable’ and ‘unstable’ diagonal submatrices respectively. Let

\[
V^s(t_2, t_1) := \begin{pmatrix} e^{\int_{t_1}^{t_2} \Lambda^s(\tau) \, d\tau} & 0 \\ 0 & 0 \end{pmatrix}, \quad V^u(t_2, t_1) := \begin{pmatrix} 0 & 0 \\ 0 & e^{\int_{t_1}^{t_2} \Lambda^u(\tau) \, d\tau} \end{pmatrix}. \] (28)

By construction we have \( \limsup_{t \to \infty} \lambda_j(t) < 0 \), \( j = 1, \ldots, k \). Hence, we may choose an \( \alpha > 0 \) such that \( \lambda_j(t) < -\alpha < 0 \) for \( j = 1, \ldots, k \) and all \( t \) sufficiently large. We may also choose constants \( \sigma > 0 \) and \( K > 0 \) such that the following estimates hold

\[
\|V^s(t_2, t_1)\| \leq Ke^{-\alpha(t_2-t_1)}, \quad t_2 \geq t_1 \]

\[
\|V^u(t_2, t_1)\| \leq Ke^{\sigma(t_2-t_1)}, \quad t_2 \leq t_1. \] (29)
Let \( t_0 \in \mathbb{R} \) and suppose \( a^s \in \mathbb{R}^k \) and consider the integral equation

\[
\mathbf{u}(t, a^s) = V^s(t, t_0) \left( a^s \right)
+ \int_{t_0}^{t} V^s(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau
- \int_{t}^{\infty} V^u(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau,
\]

where \( \mathbf{u} : [t_0, \infty) \times \mathbb{R}^n \to \mathbb{R}^M \).

Suppose \( \varepsilon < \frac{\gamma}{4K} \) and let \( r \) and \( T \) be chosen so that (26) holds for all \( t \geq t_0 \geq T \). Using standard successive approximation techniques (see, e.g., [51]), it is straightforward to verify that (30) has a unique solution for all \( a^s \) sufficiently small and \( t_0 \) sufficiently large, and that the solution satisfies

\[
|\mathbf{u}(t, a^s)| \leq 2K|a^s|e^{-\alpha(t-t_0)}.
\]

If \( t \mapsto \mathbf{u}(t, a^s) \) is continuous and solves (30) then, \( \mathbf{u}(t, a^s) \) is differentiable in \( t \) and solves (25) with componentwise initialization \( \mathbf{u}_i(t_0, a^s) = a^s_i \) for \( i = 1, \ldots, k \). To be more precise, suppose \( \mathbf{u}(t, a_s) \) solves (30) given \( a^s \). Note that

\[
\int_{t_0}^{\infty} V^u(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau
= \int_{t_0}^{\infty} V^u(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau
- \int_{t}^{\infty} V^u(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau.
\]

This holds since the integrals in question are all finite (and, in particular, due to (26) and (31)). Reformulating (30) we have

\[
\mathbf{u}(t, a^s) = V^s(t, t_0)x_0 + V^u(t_0, t)x_0
+ \int_{t_0}^{t} V^s(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau
+ \int_{t}^{\infty} V^u(t, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau,
\]

where \( x^s_0 = a^s \) and \( x^u_0 = \left[ \int_{t_0}^{t} V^u(t_0, \tau) \left( \tilde{F}(\mathbf{u}(\tau, a^s), \tau) - U(\tau)g'(\gamma_\tau)\dot{\gamma}_\tau \right) d\tau \right]^u \), where we use the notation \([\cdot]^u \) to indicate extraction of the “unstable” (last \( M-n_s \)) components of the vector. Differentiating (32) it is clear that \( \mathbf{u}(t, a_s) \) solves (25).

4. (Construct Stable Manifold) We now construct the stable set \( S \) corresponding to the ODE (25). For each \( z^s_0 \in B_\varepsilon(0) \subset \mathbb{R}^k \) let \( \mathbf{u}(.; z^s_0) \) be the (unique) solution to (30). For each \( t \in [T, \infty) \) define the component map \( \psi_j : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R} \) by

\[
\psi_j(t, z^s_0) := u_j(t, z^s_0), \quad j = k+1, \ldots, M,
\]

and let \( \psi = (\psi_j)_{j=k+1}^M \). The stable manifold (with respect to (25)) is given by

\[
S := \{ (t, z^s_0, \psi(t, z^s_0)), t \geq T, z^s_0 \in \mathbb{R}^k \cap B_\varepsilon(0) \}.
\]
By construction, for any initialization \((t_0, z_0^a, z_0^b) \in S\), the corresponding solution \(z\) of (25) with \(z(t_0) = (z_0^a, z_0^b)\) satisfies \(z(t) \to 0\). Moreover, by Lemma 44 we see that \(S\) contains all stable initializations \((t_0, z_0)\). That is, if \(z\) is a solution to (25) with \(z(t_0) = z_0\) and \(z(t) \to 0\), then \((t_0, z_0) \in S\).

Having constructed \(S\) (the stable manifold for (25)) the stable manifold for (9), denoted here by \(\hat{S}\), is obtained by an appropriate change of coordinates, \(\hat{S} := \{(t, x) \in \mathbb{R} \times \mathbb{R}^M : U(t)(x - g(\gamma_i)) \in S\}\).

Finally, the fact that \(S\) is a \(C^1\) manifold will be shown in the following lemma.

**Lemma 28.** Assume the hypotheses of Theorem 25 hold. Then the stable manifold \(S\) is of class \(C^1\).

**Proof.** Let \(u(t, a^z)\) be the solution to (30) with stable initialization \(a^z\) at time \(t\). We will begin by establishing the existence of derivatives of \(u\).

Fix a coordinate \(i \in \{1, \ldots, n_s\}\). We will compute the vector of partial derivatives \(\left(\frac{\partial u_j(t, a^z)}{\partial a^z_i}\right)_{j=n_s+1}^{M}\).

Define the integral equation

\[
\dot{z}(t, a^z) = V^s(t, t_0)e_i + \int_{t_0}^{t} V^s(t, \tau)D_x \tilde{F}(u(\tau, a^z), \tau)z(\tau, a) \, d\tau
\]

\[
- \int_{t}^{\infty} V^u(t, \tau)D_x \tilde{F}(u(\tau, a^z), \tau)z(\tau, a) \, d\tau.
\]

Note that, since \(u(\tau, a^z) \to 0\) as \(\tau \to \infty\), using (26) we see that \(\|D_x \tilde{F}(u(\tau, a^z), \tau)\|\) may be taken to be arbitrarily small by taking \(\tau \to \infty\). Again using standard successive approximation techniques (see [21]), we see that there exists a unique solution to (34) for all \(a^z\) sufficiently small and \(t_0\) sufficiently large, and moreover, the solution satisfies

\[
|z(t, a^z)| \leq 2K|a^z|e^{-\alpha(t-t_0)}
\]

for \(t \geq t_0\).

We now confirm that \(z(t, a^z)\) of (34) is in fact equal to \(\left(\frac{\partial u_j(t, a^z)}{\partial a^z_i}\right)_{j=n_s+1}^{M}\). This will be accomplished using standard techniques (see, e.g., [21] Ch. 13). Let

\[
q(t, a, h) := \frac{1}{h}(u(t, a + he_i) - u(t, a)).
\]

Using (30) we have

\[
q(t, a^z, h) = V^s(t, t_0)e_i + \int_{t_0}^{t} V^s(t, \tau)[D_x \tilde{F}(u(\tau, a^z), \tau)q(t, a^z, h) + \Delta],
\]

where \(\Delta = \frac{1}{h} \left[\tilde{F}(u(\tau, a^z + he_i), \tau) - \tilde{F}(u(\tau, a^z), \tau)\right] - D_x \tilde{F}(u(\tau, a^z), \tau)q(t, a^z, h)\).

Let \(K\) and \(\sigma\) be as in (29). Using (26) we see that for any \(\eta > 0\) we may choose a sufficiently small neighborhood of the origin such that \(|\Delta| < 2K\eta\) for all \(a^z\) in the neighborhood. Let \(\varepsilon > 0\) be such that \(2K^2\varepsilon < \frac{1}{2}\). Using (36) and (34) and letting \(m(h) = \sup_{t \geq t_0} \|z(t, a^z) - q(t, a^z, h)\|\) we have

\[
m(h) \leq \varepsilon \int_{t_0}^{t} e^{-\sigma(t-\tau)}(m(h) + \|\Delta\|) \, d\tau - \varepsilon \int_{t}^{\infty} e^{\sigma(t-\tau)}(m(h) + \|\Delta\|) \, d\tau
\]

\[
\leq K\varepsilon m(h) + 2K^2\eta \frac{2}{\sigma},
\]

which implies that \(m(h) \leq \frac{8K^2\varepsilon}{\sigma}\). Letting \(\eta \to 0\) as \(h \to 0\) we see that \(m(h) \to 0\) as \(h \to 0\), and hence \(z\) is the desired derivative. 

\[\square\]
B. Refinement of Theorem 25: The $h \in C^3$ case.

The following lemma will be instrumental in the study of discrete-time algorithms in Section VII. The lemma shows that if we assume $h$ is one degree smoother than assumed in Theorem 25 then the stable manifold will also be one degree smoother. In the following section, the lemma will be required to guarantee that $\eta$ (to be defined in 52) possesses the properties outlined in Lemma 40.

We make the following Assumption (which will be used in place of previous smoothness assumptions, Assumptions 11 and 17 where necessary).

**Assumption 20.** $h \in C^3$.

**Lemma 29.** Suppose that the hypotheses of Theorem 25 hold. Suppose, moreover, that the stronger smoothness condition of Assumption 20 holds. Then the stable manifold $S$ is $C^2$, uniformly in $t$. That is, the functions $\psi_i$ used to define $S$ in (33) are $C^2$ and the second derivative of each $\psi_i$ is bounded uniformly in $t$.

**Proof.** Let $u(t, a^s)$ be the solution to (30) given $a^s \in \mathbb{R}^n_a$. Fix $i \in \{1, \ldots, k\}$ and let $z(t, a^s)$ be the resultant solution to (34). Let $j \in \{1, \ldots, n_x\}$. We will compute the vector of partial derivatives \((\frac{\partial u_i(t, a^s)}{\partial a^s_j})^M_{t=n+1}\). Consider the integral equation

$$w(t, a^s) = \int_{t_0}^{t} V^s(t, \tau) \left[ D^2_x \hat{F}(u(\tau, a^s), \tau) z^2(\tau, a^s) + D_x \hat{F}(u(\tau, a^s), \tau) w(\tau, a^s) \right] d\tau \quad \text{(37)}$$

Again using standard successive approximation techniques, we see that for $a^s$ sufficiently small and $t_0$ sufficiently large, there exists a unique solution to (37).

Using the same reasoning used to show that $z(t, a^s) = (\frac{u_i(t, a^s)}{\partial a^s_j})^M_{j=n+1}$ in the proof of Lemma 28 it follows that $w(t, a^s) = (\frac{\partial u_i(t, a^s)}{\partial a^s_j})^M_{j=n+1} = (\frac{\partial u_i(t, a^s)}{\partial a^s_j})^M_{j=n+1}$.

We now show that $w(t, a^s)$ is uniformly bounded in $t$. Using (37) we have

$$\|w(t, a^s)\| \leq \int_{t_0}^{t} \|V^s(t, \tau)\| \|D^2_x \hat{F}(u(\tau, a^s), \tau)\| \|z^2(\tau, a^s)\| d\tau$$

$$+ \int_{t_0}^{t} \|V^s(t, \tau)\| \|D_x \hat{F}(u(\tau, a^s), \tau)\| \|w(\tau, a^s)\| d\tau$$

$$+ \int_{t}^{\infty} \|V^u(t, \tau)\| \|D^2_x \hat{F}(u(\tau, a^s), \tau)\| \|z^2(\tau, a^s)\| d\tau$$

$$+ \int_{t}^{\infty} \|V^u(t, \tau)\| \|D_x \hat{F}(u(\tau, a^s), \tau)\| \|w(\tau, a^s)\| d\tau.$$
Recalling (35), we may bound the $w$-independent components of (38) as

$$
\int_{t_0}^{t} \| V^s(t, \tau) \| D_2^2 \tilde{F}(u(\tau, a^s), \tau) \| z^2(\tau, a^s) \| d\tau \\
+ \int_{t}^{\infty} \| V^u(t, \tau) \| D_2^x \tilde{F}(u(\tau, a^s), \tau) \| z^2(\tau, a^s) \| d\tau \\
\leq C \int_{t_0}^{t} e^{-(\sigma+\alpha)(t-\tau)} \, d\tau + C \int_{t}^{\infty} e^{\sigma(t-\tau)} \, d\tau \leq C^2 \frac{2}{\sigma},
$$

for some constant $C > 0$.

Let $\varepsilon > 0$ be such that $\frac{K\varepsilon}{\sigma} < \frac{1}{2}$. Let $t_0$ be such that $D_2^2 \tilde{F}(u(t, a^s), t) < \varepsilon$ for all $t \geq t_0$ and all $a^s$ sufficiently small. Returning to (38) we have

$$
\| w(t, a^s) \| \leq \frac{2C}{\sigma} + K\varepsilon \int_{t_0}^{t} e^{-(\alpha+\sigma)(t-\tau)} \| w(\tau, a^s) \| d\tau + K\varepsilon \int_{t}^{\infty} e^{\sigma(t-\tau)} \| w(\tau, a^s) \| d\tau.
$$

Letting $M = \sup_{t \geq t_0} \| w(t, a^s) \|$ the above yields

$$
M \leq \frac{2C}{\sigma} + \frac{2K\varepsilon M}{\sigma}.
$$

Since $\frac{2K\varepsilon M}{\sigma} < \frac{1}{2}$, we get $M < \frac{4C}{\sigma}$, which concludes the proof.

VI. DISCRETE-TIME STOCHASTIC GRADIENT DESCENT: NON-CONVERGENCE TO SADDLE POINTS

In this section we will show that D-SGD (2) does not converge to saddle points, i.e., we will prove Theorem 11. We will prove the result by considering the generalization of D-SGD given in (10). The main result of the section is Theorem 30.

The following theorem shows that (10) does not typically converge to saddle points.

**Theorem 30.** Let $\{x(k)\}_{k \geq 1}$, satisfy (10). Suppose Assumptions 13, 15 and 18–20 hold. Suppose further that $x^*$ is a regular saddle point of $h|_{C}$. Then,

$$
P(\lim_{k \to \infty} x(k) = x^*) = 0.
$$

In the above theorem, note that we do not require the gradient of $h$ to be Lipschitz or that $h$ be coercive. Neither of these properties are required for avoidance of saddle points (they are required to reach consensus and ensure convergence to local minima). Since (10) is a generalization of (2), Theorem 30 implies Theorem 11.

The remainder of the section will focus on proving Theorem 30. The proof of the theorem is broken into two parts. First, Section VI-A considers the relationship of the discrete-time process (10) to the stable manifold established in Theorem 25 (for continuous-time dynamics) and establishes a key inequality. Second, Section VI-B carries out the stochastic analysis of (10) required to prove Theorem 30.

We remark that our approach to proving Theorem 30 is based off of the techniques developed in [16]. We note, however, that [16] studies autonomous dynamical systems. The dynamics (10) (and 9) are non-autonomous, and the approach used in [16] requires substantial modification to address the non-autonomous setting.
A. Distance to the Stable Manifold and a Key Inequality

In this section we will characterize the manner in which state-time pairs \((x, t)\) are repelled from the stable manifold \(S\) under the dynamics \((10)\). For this purpose, we will define a suitable function \(\eta\) that gauges the distance of particles from the stable manifold (see \((52)\)). The main result of this section will be a key inequality that characterizes the way in which the mean dynamics of \((10)\) push particles away from the stable manifold (see Lemma \(40\) and in particular, Property 4).

To this end, without loss of generality we assume \(x^* = 0\) and let \(C = \text{span}\{e_1, e_2, \ldots, e_{\text{dim}C}\}\), where \(e_i\) denotes the \(i\)-th canonical vector in \(R^{M}\). Let the constraint space \(C\) be decomposed as
\[
C = E_s + E_u,
\]
where
\[
E_s = \{ x \in C : \nabla^2 h(0)x = \lambda x, \lambda < 0 \} \quad \text{and} \quad E_u = \{ x \in C : \nabla^2 h(0)x = \lambda x, \lambda \geq 0 \}.
\]
Here, \(E_s\) and \(E_u\) correspond to the stable eigenspaces of the gradient-descent system restricted to \(C\) and linearized about the origin.

The (nonautonomous) stable manifold \(S\) constructed in Section \([V]\) may be represented locally as a function \(\psi : E_s \times C^2 \times [0, \infty) \to E_u\). More precisely, the stable manifold may be represented locally as
\[
S = \{ (x, t) : x_u = \psi(x_s, x_{nc}, t), t \geq t_0, x_s \in B_\delta(0), x_{nc} \in B_\delta(0) \},
\]
for some \(\delta > 0\).

For convenience, we now construct a map which flattens out the stable manifold. Namely, we define
\[
\Phi(x, t) := \begin{pmatrix}
x_u \\
x_s \\
x_{nc}
\end{pmatrix} - \begin{pmatrix}
\psi(x_s, x_{nc}, t) \\
0 \\
0
\end{pmatrix}.
\]
This function locally maps the stable manifold \(S\) to the subspace \(\{(y, t) : y_i = 0 \text{ for } i = 1 \ldots n_u\}\), where \(n_u\) is the number of unstable coordinates.

Next, we notice that
\[
D_x \Phi(x, t) = \begin{pmatrix}
I_{n_u} & D_{x_s, x_{nc}} \psi \\
0 & I_{M - n_u}
\end{pmatrix}.
\]
Since \(D\psi(0, t) = 0\) (see e.g. equation \((35)\)) and since \(\psi\) is \(C^1\) in \(x\) uniformly in time (cf. Lemma 21), we may use the inverse function theorem to establish that there exist a \(C^1\) function \(x \mapsto \Phi^{-1}(\cdot, t)\) in some ball \(B(0, r)\), for any time \(t\). We emphasize that \(\Phi^{-1}(\cdot, t)\) inverts the first argument given a time \(t\).

Now, suppose that \(x(t)\) satisfies the ODE \(\dot{x} = F(x, t)\), with \(F\) as given in \((21)\). If we let \(y(t) = \Phi(x(t), t)\), then, by construction of \(\Phi\), the space \(U := \{(y, t) : y_i = 0 \text{ for } i = 1 \ldots n_u\}\) is invariant for \(y\). A chain rule computation shows that \(y\) satisfies the ODE
\[
\dot{y} = G(y, t) := D_x [\Phi, \Phi^{-1}(y, t)] F(\Phi^{-1}(y, t), t) + D_t [\Phi, (\Phi^{-1}(y, t), t)].
\]
In particular, note that \(U\) is the stable manifold for the above ODE.

Our first result in this section will be to show that \(\Phi(\cdot, t)\) converges to a limit as \(t \to \infty\). Equivalently, this may be thought of as showing that “time slices” of the stable manifold converge to a limit as \(t \to \infty\). To this end, consider the (autonomous) ODE
\[
\dot{x} = -\nabla h|_C(x),
\]
(42)
with \( x : [0, \infty) \to \mathbb{R}^{\dim C} \). By the classical stable manifold theorem [47], there exists a stable manifold \( S^* \) for (42), associated with a rest point at the origin. Let \( \psi^* : \mathbb{R}^{\dim C - n_u} \to \mathbb{R}^{n_u} \) be the function defining \( S^* \), i.e.,

\[
S^* = \{ x \in \mathbb{R}^{\dim C} : x_u = \psi^*(x) \}.
\]

Let \( \Phi^* : \mathbb{R}^{\dim C} \to \mathbb{R}^{\dim C} \) be given by

\[
\Phi^*(x) := \begin{pmatrix} x_u \\ x_s \end{pmatrix} - \begin{pmatrix} \psi^*(x) \\ 0 \end{pmatrix},
\]

where here 0 denotes the zero vector in \( \mathbb{R}^{\dim C - n_u} \). Here, \( \Phi^* \) serves an analogous role to \( \Phi \) in (39), straightening out the stable manifold of the autonomous system into the stable eigenspace of the autonomous system.

The following result strengthens the conclusion of Lemma 20 slightly, so that we obtain uniform convergence to \( C \) for initializations in a neighborhood of the origin.

**Lemma 31** (Uniform convergence to \( C \)). Suppose Assumptions 11–14 hold. For any neighborhood \( N \) of 0, and any \( \varepsilon > 0 \) there exists a \( \bar{t} > 0 \) such that for any solution \( x(t) \) of (9) with initial condition \( x(t_0) = x_0 \in N \), \( t_0 \geq 0 \), there holds \( d(x(t), C) \leq \varepsilon \) for all \( t \geq \bar{t} \).

**Proof.** Without loss of generality, let \( C \) be as given in (12) and let \( x(t) \) be decomposed as

\[
x(t) = \begin{pmatrix} \bar{x}(t) \\ x^-(t) \end{pmatrix},
\]

By Assumption 12 there exists a compact set \( K \ni 0 \) that is invariant under (9). Without loss of generality, assume that \( N \subset K \). Let \( \mathcal{M} = \sup_{x \in K} \| \nabla h(x) \| \). Let \( \lambda_{\min} \) be the smallest positive eigenvalue of \( Q \). Choose \( t_1 \) so that \( \gamma t \lambda_{\min} \varepsilon > 2\mathcal{M} \) for all \( t \geq t_1 \). Then for all \( t \geq t_1 \) and all \( x \) with \( \| \bar{x} \| > \varepsilon \) we have

\[
\frac{d}{dt} \| \bar{x}(t) \| \leq \mathcal{M} - \gamma t \lambda_{\min} \varepsilon \leq -\mathcal{M}.
\]

Using Gronwall’s inequality we get

\[
\| \bar{x}(t) \| \leq \max \{ e^{-\mathcal{M}(t-t_1)}, \varepsilon \}
\]

for all \( t \geq t_1 \). Letting \( \bar{t} = (\log(\varepsilon) + \mathcal{M} t_1)/\mathcal{M} \) now yields the desired result. \( \Box \)

The following lemma shows that \( x \mapsto \Phi(x,t) \) has a limit as \( t \to \infty \).

**Lemma 32.** Suppose Assumptions 12, 13 and 17–19 hold and that 0 is a regular saddle point of \( h|_C \). Then for all \( x_{nc} \) in a neighborhood of 0, \( \Phi(\cdot, x_{nc}, t) \to \Phi^*(\cdot) \) pointwise as \( t \to \infty \).

**Proof.** Let \( U(t) \in \mathbb{R}^{M \times M} \) be the diagonalization of \( A(t) \). Let

\[
B := \nabla^2 h|_C(0).
\]

and let \( U \in \mathbb{R}^{\dim C \times \dim C} \) be a matrix that diagonalizes \( B \) so \( B = U A U^\top \), where \( A = \text{diag} (\Lambda^s, \Lambda^u) \), and where \( \Lambda^s \in \mathbb{R}^{n_u \times n_u} \) has negative diagonal entries and \( \Lambda^u \in \mathbb{R}^{(M-n_u) \times (M-n_u)} \) has positive diagonal entries. Thus far, we have assumed coordinates to be rotated so that \( C = \text{span} \{ e_1, \ldots, e_{\dim C} \} \). Without loss of generality, we will now assume a rotation of coordinates within \( C \); namely, we will assume that \( U = I \).

25
Analogous to (28), define
\[ \hat{V}^s(t, t_1) := \left( e^{\Lambda^s(t_2-t_1)} 0 \right), \quad \hat{V}^u(t, t_1) := \left( 0 e^{\Lambda^u(t_2-t_1)} \right). \]

Finally, let \( \hat{F}(x) := -\nabla h|_C(x) - Bx \). Solutions to the following equation define the “classical” stable manifold of the \((C\text{-restricted})\) gradient system (42) (cf. (30)):
\[ w(t, a^s) = \hat{V}^s(t, t_0) \left( \begin{array}{c} a^c_0 \\ 0 \end{array} \right) + \int_{t_0}^{t} \hat{V}^s(t, \tau) \hat{F}(w(\tau, a^c_0)) d\tau - \int_{t}^{\infty} \hat{V}^u(t, \tau) \hat{F}(w(\tau, a^c_0)) d\tau, \quad (44) \]
where \( a^c_0 \in \mathbb{R}^{\dim C - n_u} \). Let \( T : \mathbb{R}^M \times [0, \infty) \to \mathbb{R}^M \), given by
\[ T(u, t) = U(t)u + g(\gamma_t) \]
denote the coordinate transformation we use to recenter and diagonalize in the computation of the nonautonomous stable manifold in Lemma 27. Note that \( T^{-1}(u, t) = U^T(t)(u - g(\gamma_t)) \) is well defined and that we have \( U(t) \to I \) and \( g(\gamma_t) \to 0 \) as \( t \to \infty \).

Given \( a^s \in \mathbb{R}^{n_s} \) (where the integer \( n_s = M - n_u \), as in the proof of Lemma 27), let \( u(t, a^s) \) be the solution to (30). Let \( a^s \) be decomposed as \( a^s = (a^c_0, a^c_0) \in \mathbb{R}^{\dim C}, a^c_0 \in \mathbb{R}^{M - \dim C} \), and let \( w(t, a^c_0) \) be the solution to (44) given \( a^c_0 \).

Let \( P_C := (I_{\dim C} 0) \in \mathbb{R}^{\dim C \times M} \) be the matrix that projects onto \( C = \text{span}\{e_1, \ldots, e_{\dim C}\} \) and observe that \( P_C U(t) \to U = I_{\dim C} \). The statement of the lemma is equivalent to showing that
\[ \lim_{t_0 \to \infty} \sup_{t \geq t_0} \|P_C T^{-1}\left( \begin{array}{c} u(t, a^s) \\ 0 \end{array} \right) - w(t, a^c_0)\| = 0, \quad (45) \]
where \( a^s = P_C T((a_0, 0), t_0) = U(t)^T((a^s, 0) - g(t_0)) \). We note that in the above equation, dependence on \( t_0 \) is implicit in the definitions of \( u \) and \( w \).

Explicitly expanding \( T^{-1}(u(t, a^s), t) \) and using (30) we have
\[ T^{-1}(u(t, a^s), t) = U^T(t)V^s(t, t_0) \left( \begin{array}{c} \tilde{a}^s \\ 0 \end{array} \right) + \int_{t_0}^{t} U^T(t)V^s(t, \tau) \hat{F}(u(\tau, \tilde{a}^s), \tau) d\tau - \int_{t}^{\infty} U^T(t)V^u(t, \tau) \hat{F}(u(\tau, \tilde{a}^s), \tau) d\tau \]
\[ + U^T(t) \int_{t_0}^{t} V^s(t, \tau)U^T(\tau)g'(\gamma_{\tau})\gamma_{\tau} d\tau - U^T(t) \int_{t}^{\infty} V^u(t, \tau)U^T(\tau)g'(\gamma_{\tau})\gamma_{\tau} d\tau - U^T(t)g(\gamma_t) \]

Using this expansion along with (44) and a triangle-inequality decomposition we obtain the bound
\[ |P_C T^{-1}(u(t, a^s), t) - w(t, a^c_0)| \leq (a) + (b) + (c) + (d), \quad (46) \]
where
\[ (a) = |P_C U^T(t)V^s(t, t_0) \left( \begin{array}{c} \tilde{a}^s \\ 0 \end{array} \right) - V^s(t, t_0) \left( \begin{array}{c} a^c_0 \\ 0 \end{array} \right) \|
\]
\[ (b) = |\int_{t_0}^{t} P_C U^T(t)V^s(t, \tau) \hat{F}(u(\tau, \tilde{a}^s), t) d\tau - \int_{t_0}^{t} \hat{V}^s(t, \tau) \hat{F}(w(t, a^c_0)) d\tau| \]
\[ (c) = |\int_{t}^{\infty} P_C U^T(t)V^u(t, \tau) \hat{F}(u(\tau, \tilde{a}^s), t) d\tau - \int_{t}^{\infty} \hat{V}^u(t, \tau) \hat{F}(w(t, a^c_0)) d\tau| \]
\[ (d) = \text{equation (49)}. \]
We will bound each of these in turn. Beginning with (a), let \( \Lambda(t) \) be as defined in (23). Let the smallest elements of \( \Lambda(t) \) be ordered as \( \lambda_1(t) \leq \cdots \leq \lambda_{\dim C}(t) \), and likewise for elements of \( \Lambda \). By Lemma 45, we see that \( \lambda_i(t) \to \lambda_i, \ i = 1, \ldots, \dim C \). Thus,
\[
\lim_{t_0 \to \infty} \sup_{t \geq t_0} |e^{\lambda_i(t-t_0)} - e^{\int_{t_0}^t \lambda_i(s)\,ds}| = 0
\]
for all \( i = 1, \ldots, \dim C \). Since \( P_C U^T(t) \to I_{\dim C} \) this implies
\[
\lim_{t_0 \to \infty} \sup_{t \geq t_0} \left| P_C U^T(t) V^s(t,t_0) \begin{pmatrix} \tilde{a}^s_0 \\ 0 \end{pmatrix} - \tilde{V}^s(t,t_0) \begin{pmatrix} \tilde{a}^s_0 \\ 0 \end{pmatrix} \right| = 0,
\]
which implies that (a) converges to zero as \( t_0 \to \infty \).

We now consider (b). Suppressing some arguments, we have
\[
\left| \int_{t_0}^t P_C U^T(t) V^s \tilde{F} \, dt - \int_{t_0}^t \tilde{V}^s \tilde{F} \, d\tau \right| \leq \int_{t_0}^t |\tilde{V}^s(\tilde{F} - \hat{F})| \, d\tau + \int_{t_0}^t (P_C U^T(t) V^s - \tilde{V}^s)(\hat{F}) \, d\tau
\]
(48)
We will now bound the right hand side of (48), beginning with the first term. By construction, we have \( \tilde{F}(0,t) = 0 \) and \( \hat{F}(0) = 0 \). Moreover, by construction \( \tilde{F}(\cdot,t) \) and \( \hat{F} \) are uniformly Lipschitz in a neighborhood of 0. By (31) (and a similar argument for \( w \)), we have \( u(t,a_s) \leq ce^{-\alpha(t-t_0)} \) and \( w(t,a_s) \leq ce^{-\alpha(t-t_0)} \) for some constant \( c > 0 \) independent of \( t_0 \). Hence,
\[
|P_C U^T(t) \tilde{F}(u(t,a_s),t) - \tilde{F}(w(t,a_s))| \leq 2ce^{-\alpha(t-t_0)},
\]
and we have
\[
\lim_{t_0 \to \infty} \sup_{t \geq t_0} \int_{t_0}^t |\tilde{V}^s(t,\tau)(P_C U(t) \tilde{F}(u(t,a_s),\tau) - \tilde{F}(w(t,a_s)))| \, d\tau = 0.
\]
By (47), we have that \( |P_C U^T(t) V^s(t,t_0) - \tilde{V}^s(t,t_0)| \leq c \) for some constant \( c > 0 \) for all \( t_0 \) sufficiently large and all \( t \geq t_0 \). Again using the facts that \( \tilde{F}(0,t) = 0 \) for all \( t \), \( \tilde{F}(\cdot,t) \) is uniformly Lipschitz in a ball about zero, and \( |u(t,a_s)| \leq e^{-\alpha(t-t_0)} \), we get that
\[
\lim_{t_0 \to \infty} \sup_{t \geq t_0} \int_{t_0}^t (P_C U^T(t) V^s(t,\tau) - \tilde{V}^s(t,\tau))\tilde{F}(u(t,a_s),\tau) \, d\tau = 0.
\]
The terms in (c) can be handled using similar reasoning to (b) to yield
\[
\lim_{t_0 \to \infty} \sup_{t \geq t_0} \int_{t_0}^\infty |\tilde{V}^u(t,\tau)(P_C U^T(t) \tilde{F}(u(t,a_s),\tau) - \tilde{F}(w(t,a_s)))| \, d\tau = 0
\]
and
\[
\lim_{t_0 \to \infty} \sup_{t \geq t_0} \int_{t_0}^\infty (P_C U^T(t) V^u(t,\tau) - \tilde{V}^u(t,\tau))\tilde{F}(u(t,a_s),\tau) \, d\tau = 0.
\]
Finally, handling (d), we note that, as \( \int_t^\infty |g'(\gamma_\tau)\gamma_\tau| \, d\tau < \infty \) by Lemma 26 and as \( V^s \) and \( V^u \) are bounded, we can deduce that
\[
U^T(t) \int_{t_0}^t V^s(t,\tau)U(\tau)g'(\gamma_\tau)\gamma_\tau \, d\tau - U^T(t) \int_{t_0}^\infty V^u(t,\tau)U(\tau)g'(\gamma_\tau)\gamma_\tau \, d\tau + g(\gamma_t) \to 0.
\]
(49)
This accounts for all terms on the right hand side of (46), thus giving (45).

\[\square\]

**Assumption 21.** \( \gamma_t \) takes the form \( \gamma_t = t^r \) for some \( r > 0 \).
Remark 33. Note that under Assumption 27 the following technical condition is satisfied: \( \gamma_t \) satisfies
\[
\int_{t_0}^t \gamma_t e^{-\int_{\tau}^t \gamma_s \, ds} e^{-\alpha(t-t_0)} \, d\tau \to 0 \quad \text{as} \quad t \to \infty
\]
where \( \alpha > 0 \). We make Assumption 27 to ensure that this condition is satisfied in the subsequent lemma, however, the assumption is consistent with Assumption 3 and the time transformation used after (7). To see that the condition holds, note that under Assumption 21 there will exist a \( \kappa > 0 \) so that for any \( t > \tau > t_0 \) satisfying \( t - \tau > \kappa \) we have that \( e^{-(t-\tau)} \geq e^{-\int_{\tau}^t \gamma_s \, ds} \). We may then estimate
\[
\begin{align*}
\gamma_t \int_{t_0}^t e^{-\int_{\tau}^t \gamma_s \, ds} e^{-\alpha(t-t_0)} \, d\tau & \leq \int_{t_0}^t e^{-\alpha(t-t_0)} \, d\tau \\
& \quad + C \gamma_t \int_{t_0}^t e^{-(t-\tau)-\alpha(t-t_0)} \, d\tau \\
& \leq Ct \alpha^{-1} e^{-\alpha t} \to 0,
\end{align*}
\]
for some \( C > 0 \). Hence in this case Assumption 27 holds.

Lemma 34. Suppose Assumptions 12, 17, 19 and 21 hold and that 0 is a regular saddle point of \( \bar{h}\big|_{\mathcal{C}} \).
Then \( D_t \Phi(0,t) \to 0 \) and \( D_{xt} \Phi(0,t) \to 0 \) as \( t \to \infty \).

Proof. Recalling that \( \Phi \) is defined in (39), the result is equivalent to \( \frac{\partial}{\partial y} \psi(t,0) \to 0 \) and \( \frac{\partial^2}{\partial x \partial y} \psi(t,0) \to 0 \), where \( \psi \) is defined componentwise in (33) and where \( u(t,z) \) denotes a solution to (30) with stable initialization \( z^* \).

Thus, the claim holds if \( \frac{d}{dt} u(t,0) \to 0 \) and \( \frac{\partial^2}{\partial x \partial t} u(t,0) \to 0 \), where \( u \) satisfies
\[
\begin{align*}
u(t, a^*) = & V^s(t, t_0) \begin{pmatrix} a^* \\ 0 \end{pmatrix} \\
& + \int_{t_0}^t V^s(t, \tau) \left( \tilde{F}(u(\tau, a^*), \tau) - U(\tau)g'(\gamma_\tau)\gamma_\tau \right) \, d\tau \\
& \quad - \int_{t}^{\infty} V^u(t, \tau) \left( \tilde{F}(u(\tau, a^*), \tau) - U(\tau)g'(\gamma_\tau)\gamma_\tau \right) \, d\tau,
\end{align*}
\]
We begin by estimating \( \frac{d}{dt} u(t,0) \to 0 \). In this case, the first term is zero (as \( a_x = 0 \)). For the second term, taking a derivative in \( t \) we obtain
\[
\begin{align*}
\tilde{F}(u(\tau, 0), t) - U(t)g'(\gamma_\tau)\gamma_\tau + \int_{t_0}^t \Lambda(t)V^s(t, \tau) \left( \tilde{F}(u(\tau, a^*), \tau) - U(\tau)g'(\gamma_\tau)\gamma_\tau \right) \, d\tau.
\end{align*}
\]
The first term here clearly goes to zero as \( u(0,t) \to 0 \). For the second term, using (31) we can bound
\[
|\tilde{F}(u(\tau, a^*)) - U(\tau)g'(\gamma_\tau)\gamma_\tau| \leq Ce^{-\alpha(t-t_0)},
\]
for some \( C > 0 \), and hence the entire second term is bounded by
\[
C \int_{t_0}^t \Lambda(t)V^s(t, \tau)e^{-\alpha(t-t_0)} \, d\tau,
\]
for some \( C > 0 \). Note that for \( \lambda_i(t) \) in the stable block of \( \Lambda(t) \) (see 27), either \( \lambda_i(t) \) converges to a limit (by Lemma 45) or \( \lambda_i(t) \to \infty \) at rate \( \lambda_i(t) = \Theta(\gamma_t) \). By Assumption 21 and Remark 33 this also goes to zero. The third term is bounded similarly, with the simplification that elements in the unstable block of \( \Lambda(t) \) are actually bounded (i.e., the converge to a finite limit by Lemma 45).

The \( D_{xt} \) terms are handled in a completely analogous way. \( \square \)
Let
\[ T_t := D_x G(0, t). \] (50)

The following three lemmas characterize the key properties of \( T_t \).

**Lemma 35.** Suppose that \( A \) is a symmetric, \( m \times m \) matrix and for some \( \|x\| = 1 \) we have that \((A - \lambda I)x = y \) and \( \|y\|_2 = \epsilon \). Then \( d(\lambda, \sigma(A)) \leq C\epsilon \), where \( C > 0 \) is a constant depending only on \( A \). Suppose, moreover, that there is a set of \( k \) orthogonal vectors \( \{x_i\} \) satisfying \((A - \lambda I)x_i = y_i \) with \( \|y_i\| \leq \epsilon \). Then as long as \( \epsilon \) is small enough, \( A \) has at least \( k \) mutually orthogonal eigenvectors whose eigenvalues satisfy \( |\lambda - \lambda_i| \leq C\epsilon \).

**Proof.** Let \( \lambda_i \) and \( v_i \) be the eigenvalues/eigenvectors of \( A \) satisfying \( \|v\| = 1 \). Because \( A \) is symmetric it possesses orthogonal eigenvectors, and we may write
\[
(A - \lambda I)x = \sum_i (\lambda_i - \lambda)(v_i \cdot x)v_i = y.
\]

In turn, we have that, for all \( i \), \( v_i \cdot x = \frac{y \cdot v_i}{\lambda_i - \lambda} \). As \( |y \cdot v_i| \leq \epsilon \) and \( v_i \cdot x \) must be greater than \( 1/\sqrt{m} \) for at least one \( i \), we then have that there exists an \( i \) such that \( |\lambda_i - \lambda| \leq C\epsilon \).

Let \( P_{x,e} \) be the projection onto the eigenspace associated with all the eigenvalues within distance \( C\epsilon \) of \( \lambda \). We have that \( |P_{x,e} x_i| \sim 1 - \epsilon \) for each of the \( x_i \). As the \( x_i \) are orthonormal, we can infer that \( P_{x_i} \cdot P_{x_j} \sim \epsilon \) for \( i \neq j \). Thus for \( \epsilon \) small enough one has that the \( P_{x_i} \) are linearly independent, and hence the projection has rank at least \( k \), which completes the proof.

**Lemma 36.** (Spectral gap of \( T_t \)) Suppose Assumptions [12] [13] [17] [19] and [21] hold and that 0 is a regular saddle point of \( h|_C \). The following two properties hold:
(i) For all \( t \) sufficiently large, \( T_t \) has precisely \( M - n_u \) negative eigenvalues and \( n_u \) positive eigenvalues.
(ii) There exists a \( t^* \) such that \( \inf \{ \lambda \in \sigma(T_t) : \lambda > 0, t \geq t^* \} > 0 \).

**Proof.** By Equation (41) we have \( D_x \Phi(0, t) = I \). By Lemma [32] we have \( \Phi^{-1}(0, t) \to 0 \) as \( t \to \infty \). By Lemma [34] we have \( D_t \Phi(0, t) \to 0 \) as \( t \to \infty \).

From (41) and (50) we see that
\[
T_t = D_x \left( [D_x \Phi(0, t)]F(\Phi^{-1}(0, t), t) - D_t \Phi(\Phi^{-1}(0, t), t) \right)
= D_x^2 \Phi(0, t)F(\Phi^{-1}(0, t), t) + D_x \Phi(0, t)D_x F(\Phi^{-1}(0, t), t)D_x \Phi^{-1}(0, t) + D_{xt} \Phi(\Phi^{-1}(0, t), t)D_x \Phi^{-1}(0, t).
\]

By Lemmas [32] and [34] we obtain that
\[
T_t = D_x F(\Phi^{-1}(0, t), t) + o(1) = -D_x^2 h(0) - \gamma_t Q + o(1).
\]

Let
\[
-D_x^2 h(0) - \gamma_t Q = \begin{pmatrix} A & B \\ C & D \end{pmatrix} - \gamma_t \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}
\]

Let \( e \) be a unit vector in the off-constraint space. Then \((T_t + \gamma_t I)e = (Be, 0) + o(1)\). Dividing these matrices by \( \gamma \) and applying the previous lemma gives that, for large \( t \), \( T_t \) has at least \( \text{rank}(Q) \) linearly independent eigenvectors with eigenvalue given by \(-\gamma_t + O(1)\).

Similarly, if we let \( e \) be a unit length eigenvector of \( A \) with eigenvalue \( \lambda \), then we have
\[
(T_t - \lambda I) \begin{pmatrix} e \\ \frac{C}{\sqrt{-\gamma - \lambda}} e \end{pmatrix} = \frac{1}{-\gamma - \lambda} \begin{pmatrix} B C e \\ -D C e \end{pmatrix} + o(1).
\]

29
Applying Lemma 35 to these approximate eigenvectors, and using the fact that \( A \) has a spectral gap (since we assume that 0 is a regular saddle point of \( h|_C \)) then completes the proof.

**Lemma 37.** Suppose Assumptions [12][13][17][19] and [21] hold and that 0 is a regular saddle point of \( h|_C \). Then for all \( t \) sufficiently large, \( T_t \) has the block diagonal form

\[
T_t = \begin{pmatrix} P_t & 0 \\ 0 & Q_t \end{pmatrix},
\]

where \( P_t \) is positive definite and \( Q_t \) is negative definite.

**Proof.** First, note that, if an eigenvector of \( T_t \) has positive eigenvalue, then it must lie in \( U = \{ x \in \mathbb{R}^M : x_{n_u+1} = \cdots = x_M = 0 \} \). If this were false, then the space \( U \) would not be stable under (41). By Lemma 36, for \( t \) sufficiently large, \( T_t \) has precisely \( n_u \) positive eigenvalues and \( M-n_u \) negative eigenvalues. Let the eigenstructure be arranged so that \( \lambda_1, \ldots, \lambda_{n_u} \) are positive and the remaining eigenvalues are negative for \( t \) sufficiently large. The corresponding eigenvectors of \( v_1, \ldots, v_M \) of \( T_t \) are divided into two sets so that \( \text{span}\{v_1, \ldots, v_{n_u}\} = \text{span}\{e_1, \ldots, e_{n_u}\} = U \) and \( \text{span}\{v_{n_u+1}, \ldots, v_M\} = \text{span}\{e_{n_u+1}, \ldots, e_M\} = U^\perp \).

Letting \( V = [v_1, \ldots, v_M] \) be the matrix formed by taking the eigenvectors as columns, by orthogonality \( V \) has block diagonal structure

\[
V = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix},
\]

and for \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_M) \), we have

\[
T_t = V \Lambda V^T = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \begin{pmatrix} V_1^T & 0 \\ 0 & V_2^T \end{pmatrix} = \begin{pmatrix} V_1 \Lambda_1 V_1^T & 0 \\ 0 & V_2 \Lambda_2 V_2^T \end{pmatrix}.
\]

We now define a function that gives a convenient notion of distance to the stable manifold. Let

\[
d^2(x) := \sum_{i=1}^{n_u} x_i^2,
\]

and

\[
\eta(x, t) := d(\Phi(x, t)).
\]

The following lemma characterizes the manner in which taking a step in (9) pushes away from the set \( U = \{ x \in \mathbb{R}^M : x_1 = \cdots = x_{n_u} \} \) (i.e., the straightened-out version of \( S \)).

**Remark 38.** (Use of \( h \in C^3 \) assumption) We remark that the following lemma is the only point in the paper at which Assumption [20] (\( h \in C^3 \)) and Lemma 29 are used. All other uses of Assumption 20 in the paper propagate from this Lemma.

**Lemma 39.** Suppose Assumptions [12][13][17][20] hold and that 0 is a regular saddle point of \( h|_C \).
Then there exists a constant \( c > 0 \) and a \( \delta > 0 \) such that

\[
d^2(x + \varepsilon G(x, t)) \geq (1 + \varepsilon c)^2 d^2(x)
\]

for all \( \varepsilon \in [0, 1] \), \( x \in B_\delta(0) \), and all \( t \) sufficiently large.
Proof. Recall that \( \mathcal{U} = \{ x : x_1 = \cdots = x_{n_u} = 0 \} \) is an invariant set for (11); i.e., \( G(\mathcal{U}, t) \subset \mathcal{U} \) for all \( t \geq 0 \). This implies that \( G_i(x, t) = 0 \) for \( x \in \mathcal{U}, i \in \{ n_u + 1, \ldots, M \} \), and \( t \geq 0 \) which implies that \( \frac{\partial^2 G_i(0, t)}{\partial x_j \partial x_k} = 0 \) if \( i, j \in \{ n_u + 1, \ldots, M \} \). Thus, by Taylor’s theorem, for \( i = n_u + 1, \ldots, M \) we have

\[
G_i(x, t) = (T_x)_i + \sum_{j \in [M]} \sum_{k \in [n_u]} \frac{\partial^2 G_i(0, t)}{\partial x_j \partial x_k} x_j x_k + R_i(x, t),
\]

where, given an integer \( n \), we use the notation \( [n] := \{ 1, \ldots, n \} \), and where \( R_i(x, t) \) denotes the remainder term, and we note that \( G \in C^2 \) since \( F \in C^2 \). By Lemma 29 we have that \( \| D_x^2 G_i(0, t) \| \leq C \) for some \( C > 0 \), for all \( t \geq t_0 \) and all \( x \) in a neighborhood of zero. Expressing the Taylor remainder \( R_i(x, t) \) in integral form \( R_i(x, t) = \int_0^t (x - v) D_x^2 G_i(v, t) \, dv \) we see that the bounded second derivative implies that

\[
R_i(x, t) = O(x_j x_k)
\]

for \( j \in [n_u], k \in [M] \).

We now compute

\[
d^2(x + \varepsilon G(x, t)) = \sum_{i=1}^r (x + \varepsilon G(x, t))^2_i
\]

\[
= \sum_{i=1}^r ((x + \varepsilon T_i x)_i + \varepsilon R_i(x, t))^2_i
\]

\[
= \sum_{i=1}^r (x + \varepsilon T_i x)^2_i + \sum_{i=1}^r (2(x + \varepsilon T_i x)_i \varepsilon R_i(x, t) + \varepsilon^2 R_i(x, t)^2_i)
\]

\[
= R(x, t)
\]

and we define \( R(x, t) \) as in the last line above. By (53) and the definition of \( R(x, t) \) we see that

\[
R(x, t) = o(x_j x_k),
\]

for \( j \in [n_u], k \in [M] \).

We now focus on estimating the first term on the right hand side (last line) of (54). For \( t \) sufficiently large, \( T_i \) has the block diagonal structure indicated in Lemma 37. Let \( P_i \) be the positive definite block. Let \( \lambda^*_t \) denote smallest eigenvalue of \( T_i \) at time \( t \) and note that by Lemma 36 there exists a time \( t^* \) such that \( \lambda := \inf_{t \geq t^*} \lambda^*_t > 0 \). Thus we see that

\[
\sum_{i=1}^r (x + T_i x)^2_i = (x_u + \varepsilon P_t x_u)^T (x_u + \varepsilon P_t x_u) = x_u^T (I + \varepsilon P_t^T (I + \varepsilon P_t) x_u \geq (1 + \varepsilon \lambda)^2 d^2(x),
\]

for all \( t \) sufficiently large. By (55) we may choose a constant \( c_1 \in (0, \lambda) \), and choose a \( \delta > 0 \) such that

\[
R(x, t) \leq \frac{1}{2} c_1 d^2(x)
\]

holds for all \( x \in B_\delta(0) \). Letting \( c = \lambda - c_1 \) and returning to (54) this implies that

\[
d^2(x + \varepsilon G(x, t)) \geq (1 + \varepsilon c)^2 d^2(x)
\]

for all \( x \in B_\delta(0) \) and \( t \) sufficiently large.

The following lemma characterizes the properties of \( d \) and \( \eta \) defined in (51) and (52). More to the point, the lemma characterizes the relationship between steps of (10) and the stable manifold \( S \), in particular,
showing that (10) is repelled from \( S \). The properties demonstrated in this lemma will be used in the following section to prove Theorem 30.

**Lemma 40.** Suppose Assumptions 12–13 and 17–20 hold and that 0 is a regular saddle point of \( h \mid C \).

Then \( d(\cdot) \) and \( \eta(\cdot, \cdot) \) have the following properties.

1. \( d(cx) = cd(x) \) for all \( c > 0 \)
2. \( d(\cdot) \) is convex
3. \( d(\cdot) \) is Lipschitz
4. There exist constants \( c_3, c_2 > 0 \) and a \( \delta > 0 \) such that
   \[
   \eta(x + \varepsilon F(x, t), t + \varepsilon) \geq (1 + c_2 \varepsilon)\eta(x, t) - c_3 \varepsilon^2
   \]
   for \( \varepsilon \in [0, 1] \), \( x \in B_\delta(0) \), and \( t \) sufficiently large.
5. For \( \eta(x, t) \neq 0 \), \( x \) in a neighborhood of 0 and all \( t \geq 0 \), we have
   \[
   D[\eta, (x, t)](F(x, t), 1) > 0.
   \]

**Remark 41** (\( S \) as the repelling object). In this paper we have discussed stable manifolds for continuous-time systems. Discrete-time systems (with constant step size) also possess stable manifolds [22], and these manifold generally differ from their continuous-time counterparts (when the discrete-time processes is obtained by discretization of a continuous-time process). It is important to note that, in each case, the associated stable manifold is precisely the set that the process (be it discrete- or continuous-time) is repelled from. Note that in Lemma 40 we study the continuous-time stable manifold \( S \) as a repelling object for a discretization of (9) with step size \( \varepsilon \). Because we are using the “wrong” stable manifold, these dynamics are not perfectly repelled from \( S \); this is captured by the error term at the end of property 4 above, indicating that arbitrarily close to \( S \), the discretization may not step away from \( S \). However, as \( \varepsilon \to 0 \), \( S \) approximates the stable manifold of the discretized system with higher fidelity, and this error term goes to zero. Since Theorem 30 considers a discretization of (9) with decaying step size (i.e., (10)), \( S \) is asymptotically repelling for these dynamics.

We now prove Lemma 40.

**Proof.** Properties 1–2 follow readily from the definition of \( d \). To see that Property 3 holds, note that \( d(\cdot) \) satisfies the triangle inequality \( d(x + y) \leq d(x) + d(y) \) and \( d \) is Lipschitz at the origin in the sense that \( d(x) \leq K\|x\| \). Hence, \( d(x) - d(y) \leq K(\|x\| - \|y\|) \).

Property 4 is proved as follows. For convenience, let \( H(x, t) \) be the associated autonomous vector field

\[
H(x, t) = \begin{pmatrix} F(x, t) \\ 1 \end{pmatrix}
\]

Letting \( v = (x, t) \) and making the obvious modifications to the definitions of \( \eta \) and \( d \) we have

\[
\eta(x + \varepsilon F(x, t), t + \varepsilon) = \eta(v + \varepsilon H(v))
= d(\Phi(v + \varepsilon H(v)))
= d(\Phi(v) + \varepsilon D[\Phi, v]H(v) + O(\varepsilon^2))
= d(\Phi(x, t) + \varepsilon D_x[\Phi, F(x, t)]F(x, t) + \varepsilon D_t\Phi(x, t)) + O(\varepsilon^2).
\]
For convenience, let \( y = \Phi(x, t) \). Continuing from above we have
\[
\begin{align*}
&= d(y + \varepsilon D_x [\Phi, (x, t)] F(\Phi^{-1}(y), t) + \varepsilon D_t \Phi(\Phi^{-1}(y), t)) + O(\varepsilon^2) \\
&= d(y + \varepsilon G(y, t)) + O(\varepsilon^2) \\
&\geq (1 + c_2 \varepsilon)^{1/2} d(y) + O(\varepsilon^2) \\
&= (1 + c_2 \varepsilon)^{1/2} d(\Phi(x, t)) + O(\varepsilon^2) \\
&\geq (1 + c_2 \varepsilon)^{1/2} \eta(x, t) - c_3 \varepsilon^2,
\end{align*}
\]
for some \( c_3, c_2 > 0 \), where we apply Lemma 39 to get to the third line.

Finally, Property 5 follows by taking \( \varepsilon \to 0 \) in Property 4. \( \square \)

**B. Stochastic Analysis**

We now prove Theorem 30. Our analysis strategy will rely on the observation observe that (10) is a discretization of the continuous-time process (9). As a consequence, we will see that solutions to (10) are asymptotically repelled from the stable manifold of (9).

To be more precise, suppose that the hypotheses of Theorem 30 hold. Note that the stable manifold established in Theorem 25 depends not only on \( h \) and \( Q \), but also on the (continuous-time) weight parameters \( \alpha_t, \beta_t \) and \( \gamma_t \). In order to construct appropriate continuous-time weight parameters given discrete-time weights \( \alpha_k, \beta_k \) and \( \gamma_k \), let \( t \mapsto \gamma_t \) be constructed as a smooth interpolation of the given \( \gamma_k \) so that \( \gamma_t \) and \( \gamma_k \) coincide when \( t = k, k \in \{1, 2, \ldots\} \) and \( \gamma_t \in C^1 \). Let \( \alpha_t \) be constructed likewise. Finally, let \( \beta_t = \alpha_t \gamma_t \). Let \( S \) be the stable manifold associated with the process (9) at the given saddle point, given these (continuous-time) weight functions. We will see that solutions to (10) are repelled from \( S \), thus constructed.

Our analysis follows a similar approach to [16], Section 4. Let \( d(\cdot) \) and \( \eta(\cdot, \cdot) \) be as in (51) and (52). Let
\[
S_k := \eta(x(k), k),
\]
let \( X_k := S_k - S_{k-1} \), and let \( F_k := \sigma \left( \{ x(j), \xi(j) \}_{j=1}^k \right) \), \( k \geq 1 \). Here, \( S_k \) represents the distance of the S-DGD process, \( x(k) \), from the stable manifold at iteration \( k \), and \( X_k \) represents the incremental process. To show Theorem 30 it is sufficient to show that \( \mathbb{P}(S_0 \neq 0) = 1 \).

Intuitively, the proof of Theorem 30 may be broken down into two parts. First, the isotropic nature of the noise sequence \( \{ \xi(k) \} \) (see Assumption 16 (ii)) ensures that \( S_k \) will eventually wander far from zero (Lemma 42 below). Second, due to the instability of \( S \) under the vector field (9), \( S_k \) has a positive drift so that, if \( S_k \) wanders far from 0, it is unlikely to return (Lemma 43 below). These ideas are formalized in Lemmas 42, 43. Together, these two lemmas immediately prove Theorem 30.

The following lemma shows that \( S_k \) is likely to wander far from zero.

**Lemma 42.** Suppose that the hypotheses of Theorem 30 hold. Then there exists a constant \( c_4 > 0 \) such that for all \( k \) sufficiently large,
\[
\mathbb{P} \left( \sup_{j \geq k} S_j > c_4 k^{1/2 - \tau} | F_k \right) \geq 1/2.
\]

**Proof.** The proof follows the same general strategy as the proof of Lemma 1 in [16], but adapted to the nonautonomous case. Without loss of generality, we will assume that the saddle point of interest lies at \( x^* = 0 \).
Let
\[ \tau := \inf \{ j \geq k : S_j > c_4 k^{1/2 - \tau_\alpha} \}, \]
be a stopping time indicating the first time (after time \( k \)) that \( S_j \) attains the value \( c_4 k^{1/2 - \tau_\alpha} \), where \( \tau_\alpha \) is the decay rate of \( \alpha_k \) assumed in Assumption \( \xi \). We will prove the result by considering the growth of the second moment of \( E(S_j^2 | F_k) \). To that end, we begin by estimating the incremental growth
\[
E(S_{\tau \wedge (m+1)}^2 | F_k) - E(S_{\tau \wedge m}^2 | F_k) = E(1_{\tau > m}(2X_{m+1}S_m + X_{m+1}^2) | F_k)
\]
\[
= E(E(1_{\tau > m}2X_{m+1}S_m | F_m) | F_k) + E(E(1_{\tau > m}X_{m+1}^2 | F_m) | F_k).
\]
(56)

We will estimate both of the terms on the right hand side above, beginning with the term \( E(1_{\tau > m}2X_{m+1}S_m | F_m) \).

Note that \( 1_{\tau > m} \) and \( S_m \) are \( F_m \)-measurable and so may be pulled out of the conditional expectation.

For \( \ell \geq 1 \) let
\[ \zeta_\ell := \sum_{j=1}^{\ell} \alpha_j. \]
The process (10) may be thought of as a noisy Euler interpolation of the ODE (9) with a decaying step size given by \( \alpha_k \), so that \( \zeta_k \) represents the time elapsed at iteration \( k \).

We have the following estimate
\[
E(X_{m+1} | F_m) = E(\eta(x(m+1), \zeta_{m+1}) - \eta(x(m), \zeta_m) | F_m)
\]
\[
= E(d(\Phi(x(m+1), \zeta_{m+1}) | F_m) - S_m
\]
\[
\geq d(E(\Phi(x(m+1), \zeta_{m+1} | F_m) - S_m
\]
\[
= d\left( \left( \Phi(x(m), \zeta_{m+1}) + D_x[\Phi, (x(m), \zeta_{m+1})](x(m+1) - x(m))
\right.
\]
\[
\left. + O(|x(m+1) - x(m)|^2 | F_m) \right) \right) - S_m
\]
(58)

\[
= d\left( \Phi(x(m), \zeta_{m+1}) + D_x[\Phi, (x(m), \zeta_{m+1})]E(x(m+1) - x(m) | F_m) \right)
\]
\[
+ O(E(|x(m+1) - x(m)|^2 | F_m)) - S_m
\]
(59)

Line (57) follows by the convexity of \( d \). In (58) we use the first-order Taylor approximation of \( \Phi(\cdot, t) \) at \( x(m) \) and the fact that \( D^2 \Phi \) is uniformly bounded in \( t \) (Lemma \( \xi \)). Line (59) follows from Assumption \( \xi \), the update equation (10), and the assumption that \( \alpha_k = \Theta(k^{-\tau_\alpha}) \). Line (60) follows using the Taylor approximation of \( \Phi(\cdot, \zeta_{m+1}) \) and again using the fact that \( D^2 \Phi \) is uniformly bounded in \( t \) (Lemma \( \xi \)).

Finally, (61) follows from Property 4 of Lemma \( \xi \).

Thus we get
\[
E(2X_{m+1}S_m | F_m) \geq c_2 m^{-\tau} S_m^2 + c_3 m^{-2\tau} S_m.
\]
(62)
We now estimate the second term on the right hand side of (56). At \( x = 0 \), where \( d \) is not differentiable, we will take \( D[d, 0] \) to be the particular subgradient of \( d \) given by \( D[d, 0] = \lim_{\delta \to 0} D[d, \delta \hat{x}] \) where \( \hat{x} = (1_{n_u}, 0_{m-n_u}) \) so that, by (51), we have \( D[d, 0] = (1_{n_u}, 0_{m-n_u}) \). Similarly, at points where \( \Phi(x, t) = 0 \), we define \( D[\eta, (x, t)] \) in terms of the previously mentioned definition of \( D[d, 0] \).

By the convexity of \( d \) and smoothness of \( \Phi \) we have
\[
X_{m+1} = \eta(x(m+1), \zeta_{m+1}) - \eta(x(m), \zeta_m)
\]
\[
= d(\Phi(x(m+1), \zeta_{m+1}))-d(\Phi(x(m), \zeta_m))
\]
\[
\geq D[d, (\Phi(x(m), \zeta_m))](\Phi(x(m+1), \zeta_{m+1})-\Phi(x(m), \zeta_m))
\]
\[
= D[d, (\Phi(x(m), \zeta_m))](D[\Phi(x(m), \zeta_m)])[x(m+1)-x(m)]
\]
\[
= \alpha_m D[\eta, (x(m), \zeta_m)](F(x(m), \zeta_m)+\xi(m+1)) + O(m^{-2\tau_0})
\]
where in the last line we use the fact that \( \zeta_{m+1}-\zeta_m = \alpha_m \) and again use the fact that \( D^2 \Phi \) is uniformly bounded in \( t \) (Lemma 29). Note that there exists a constant \( c_5 > 0 \) such that, for all \( x \) in a neighborhood of zero, all \( t \) sufficiently large,
\[
\|D[\eta, x]\| \geq c_5.
\]
This holds since \( \Phi \in C^1 \) (by (39) and Lemma 29), \( D_t \Phi(0, t) \to 0 \) as \( t \to \infty \) (by Lemma 34), and \( D_x \Phi(0, t) \to I \) as \( t \to \infty \), and we use the aforementioned convention for \( D[d, 0] \). From here we get
\[
\mathbb{E}(X_{m+1}^+|\mathcal{F}_m) \geq \alpha_m \mathbb{E}
\]
\[
\left(\left(D[\eta, (x(m), \zeta_m)](F(x(m), \zeta_m)+\xi(m+1))\right)^+|\mathcal{F}_m\right) + O(m^{-2\tau_0})
\]
\[
\geq \alpha_m \mathbb{E}
\]
\[
\left(\left(D[\eta, (x(m), \zeta_m)](\xi(m+1))\right)^+|\mathcal{F}_m\right) + O(m^{-2\tau_0})
\]
\[
\geq \alpha_m c_5 \mathbb{E}
\]
\[
\left(\left(D[\eta, x]|D[\eta, x]|(\xi(m+1))\right)^+|\mathcal{F}_m\right) + O(m^{-2\tau_0})
\]
\[
\geq c_4 m^{-\tau_0} + O(m^{-2\tau_0}),
\]
where the second line follows from Lemma 40 property 5, the third line follows from (63) and the fourth line follows from Assumption 16 (ii).

From here, the proof proceeds identical to the autonomous case treated in [16], Lemma 1. We see that
\[
\mathbb{E}(2X_{m+1} S_m + X_{m+1}^2|\mathcal{F}_m) \geq \frac{c}{m^{2\tau_0}}
\]
for some \( c > 0 \). Substituting this back into (56) gives
\[
\mathbb{E}\left(S_{\tau \wedge (m+1)}^2|\mathcal{F}_k\right) - \mathbb{E}\left(S_{\tau \wedge m}^2|\mathcal{F}_k\right) \geq \mathbb{E}\left(1_{\tau > m} \frac{c}{m^{2\tau_0}}|\mathcal{F}_k\right)
\]
\[
\geq \frac{c}{m^{2\tau_0}} \mathbb{P}(\tau = \infty|\mathcal{F}_k).
\]
By induction we have

\[ \mathbb{E}(S^2_{m+\tau} | \mathcal{F}_k) \geq S^2_k + c \mathbb{P}(\tau = \infty | \mathcal{F}_k) \sum_{j=0}^{m-1} \frac{1}{j^{2r_n}} \]

\[ \geq c \mathbb{P}(\tau = \infty | \mathcal{F}_k) \left( \frac{1}{k^{2r_n-1}} - \frac{1}{m^{2r_n-1}} \right). \]

But, by Assumption \[7\] we have \( \alpha_k \leq c k^{-r_n} \) for some \( c > 0 \), so (using the definition of \( \tau \)) we have

\[ \mathbb{E}(S^2_{\tau \wedge m} | \mathcal{F}_k) \leq c_4 k^{2r_n} + c k^{-r_n}. \] For all \( k \) sufficiently large, \( c_4 k^{2r_n} > c k^{-r_n} \), and so \( S_{\tau \wedge m} < 2 c_4 k^{2r_n}. \) Hence,

\[ \frac{4c_4^2}{k^{2r_n-1}} \geq c \mathbb{P}(\tau = \infty | \mathcal{F}_k) \left( \frac{1}{k^{2r_n-1}} - \frac{1}{m^{2r_n-1}} \right). \]

Letting \( m \to \infty \) we get that \( \mathbb{P}(\tau = \infty | \mathcal{F}_k) \) is bounded by a constant times \( c_4^2. \) This can be made smaller than \( \frac{1}{2} \) by choosing \( c_4 \) small enough, in which case we have

\[ \mathbb{P}(\sup_{j \geq k} S_j \geq c_4 k^{1/2-r_n} | \mathcal{F}_k) = 1 - \mathbb{P}(\tau = \infty | \mathcal{F}_k) \geq 1/2. \]

\[ \square \]

The next lemma shows that if \( S_k \) wanders sufficiently far from 0, it may not return.

**Lemma 43.** Suppose that the hypotheses of Theorem \[30\] hold. Then there exists a constant \( a > 0 \) such that

\[ \mathbb{P} \left( \inf_{j \geq k} S_j \geq \frac{c_4}{2} k^{1/2-r_n} \mid \mathcal{F}_k, \ S_k \geq c_4 k^{1/2-r_n} \right) > a. \]

The proof of this lemma is identical to the proof of the autonomous case found in Lemma 2 in \[16\]. Informally, the lemma follows from the observation that for \( m \geq k \) we have

\[ \mathbb{E}(X_{m+1} | \mathcal{F}_m) > 0, \]

so that \( S_k \) has positive drift. (The above follows from the estimate derived in \[57\]–\[62\] and the fact that we condition on the event \( S_k \geq c_4 k^{1/2-r_n} \) in Lemma \[43\].

Theorem \[30\] now follows immediately from Lemmas \[42\] and \[43\].

**APPENDIX**

**Lemma 44** (\( \mathcal{S} \) contains all stable initializations). Let \( \varepsilon, r, \) and \( T \) be chosen as in the construction of \( \mathcal{S} \). Let \( a^i \in \mathbb{R}^K \), with \( |a^i| < r/3 \), let \( t_0 \geq T \) and suppose that \( z \) is a solution to \[25\] with \( z_i(t_0, a^i) = a^i_i, \)

\( i = 1, \ldots, k \). If \( z(t, a^i) \to 0 \) as \( t \to \infty \) then \( (t_0, y_0) \in \mathcal{S} \).

**Proof.** By variation of constants we see that

\[ z(t) := V^s(t, t_0)z(t_0) + V^u(t, t_0)c \]

\[ + \int_{t_0}^{t} V^s(t, \tau) \left( \tilde{F}(z(\tau), \tau) - U(\tau)g'(\tau)\tilde{\gamma}_r \right) d\tau \]

\[ - \int_{t}^{\infty} V^u(t, \tau) \left( \tilde{F}(z(\tau)) - U(\tau)g'(\tau)\tilde{\gamma}_r \right) d\tau, \]

where \( c = z(t_0) + \int_{t_0}^{\infty} V^u(t_0, \tau) \left( \tilde{F}(z(\tau)) - U(\tau)g'(\tau)\tilde{\gamma}_r \right) d\tau. \) Note that integral in \( c \) converges by \[28\] and the fact that \( \int_{t_0}^{\infty} U(\tau)g'(\tau)\tilde{\gamma}_r d\tau < \infty \). Every term on the right hand side of \[64\] is uniformly
bounded in $t$, except possibly the term $V^u(t, t_0)c$. In particular, if $c_j \neq 0$, $j > k$, then $|V^u(t, t_0)c| \to \infty$. Since the left hand side of (64) is bounded uniformly in time, it follows that the right hand side is likewise bounded and thus all $c_j$, $j > k$ must be zero and hence $V^u(t, t_0)c = 0$.

This implies that $u(t, a^s) = z$ is a solution to the integral equation (30) given $a^s$. In the proof of Lemma 30 we saw that $u(t, a^s)$ is the unique continuous solution of (30) given $a^s$. By the definitions of $S$ and $\psi$ we thus see that $(t_0, z_0) \in S$.

The following Lemma characterizes the asymptotic properties of the linearization of (9) near saddle points.

**Lemma 45.** Let $A(t)$ be given by (20) and let $b$ be given by (43). Let $\{\lambda_1(t), \ldots, \lambda_M(t)\}$ and $\{\lambda_1, \ldots, \lambda_d\}$ denote the eigenvalues of $A(t)$ and $B$ respectively, and assume that $\lambda_i(t) \leq \lambda_j(t)$, $i < j$, and likewise for the $(\lambda_i)_{i=1}^d$. Then $\lambda_i(t) \to \lambda_i$, $i = 1, \ldots, d$, and $\lambda_i(t) \to \infty$, $i = d+1, \ldots, M$.

**Proof.** This follows by the continuity of eigenvectors and eigenvalues as a function of matrices which holds under Assumptions 17 and 19 (see e.g. [46], p. 110).

## Summary of Some Common Notational Conventions.

- $f$ = sum function \([1]\)
- $N$ = number of agents
- $d$ = dimension of domain of $f$
- $M$ = dimension of ambient space in general setup (see Section III)
- $C$ = constraint subspace (see (3))
- $d = \dim C$ (always consistent with above usage)
- $h : \mathbb{R}^M \to \mathbb{R}$ is general objective function (see (P.1))
- $Q \in \mathbb{R}^{M \times M}$ is quadratic penalty function matrix (see (P.1))
- $x^*$ = critical point of interest
- $n_u$ = number of positive (i.e., “unstable”) eigenvalues of $\nabla^2 h|_C(x^*)$
- $n_s = M - n_u$ = dimension of stable eigenspace of $\nabla^2_x(h(x^*) + \gamma x^*\top Qx^*)$, for $\gamma > 0$ large (see above (27))
- $c_1$ = constant in Assumption 16
- $c_2, c_3$ = constants in Lemma 40
- $c_4$ = constant in Lemma 42

## References

[1] R. Ge, F. Huang, C. Jin, and Y. Yuan, “Escaping from saddle points online stochastic gradient for tensor decomposition,” in *Conference on Learning Theory*, 2015, pp. 797–842.

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.

[3] Y. Chi, Y. M. Lu, and Y. Chen, “Nonconvex optimization meets low-rank matrix factorization: An overview,” *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.

[4] Y. Chen, Y. Chi, J. Fan, and C. Ma, “Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval,” *Mathematical Programming*, vol. 176, no. 1–2, pp. 5–37, 2019.

[5] P. Jain and P. Kar, “Non-convex optimization for machine learning,” *Foundations and Trends® in Machine Learning*, vol. 10, no. 3–4, pp. 142–336, 2017.

[6] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[7] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[8] “Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy,” *White House, Washington, DC*, pp. 1–62, 2012.
H. Sun and M. Hong, “Distributed non-convex first-order optimization and information processing: Lower complexity,” in Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, vol. 54, 2017.

T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, “Distributed learning in the non-convex world: From batch to streaming data, and beyond,” arXiv preprint arXiv:2001.04786, 2020.

X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, “Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent,” in Proceedings of Advances in Neural Information Processing Systems, 2017, pp. 5330–5340.

C. Lee, C. H. Lim, and S. J. Wright, “A distributed quasi-newton algorithm for empirical risk minimization with nonsmooth regularization,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1646–1655.

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 Proceedings of Advances in Neural Information Processing Systems, 2017, pp. 1067–1077.

C. Jin, P. Netrapalli, R. Ge, S. Kakade, and M. Jordan, “On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points,” 2019, arxiv preprint arXiv:1902.04811.

R. Pemantle, “Nonconvergence to unstable points in urn models and stochastic approximations,” The Annals of Probability, vol. 18, no. 2, pp. 698–712, 1990.

A. Daneshmand, G. Scutari, and V. Kungurtsev, “Second-order guarantees of distributed gradient algorithms,” arXiv preprint arXiv:1809.08694, 2018.

M. Hong, J. D. Lee, and M. Razaviyayn, “Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization,” in Proceedings of the International Conference on Machine Learning, 2018.

W. Su, S. Boyd, and E. Candès, “A differential equation for modeling nesterovs accelerated gradient method: Theory and insights,” in Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 2510–2518.

W. Krichene, A. Bayen, and P. L. Bartlett, “Accelerated mirror descent in continuous and discrete time,” in Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 2845–2853.

E. A. Coddington and N. Levinson, Theory of Ordinary Differential Equations. Tata McGraw-Hill Education, 1955.

M. Shub, Global Stability of Dynamical Systems. Springer Science & Business Media, 2013.

A. Nedic and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” IEEE Transactions on Automatic Control, vol. 54, no. 1, p. 48, 2009.

M. Rabbat and R. Nowak, “Distributed optimization in sensor networks,” in Proceedings of the International Symposium on Information Processing in Sensor Networks, 2004, pp. 20–27.

D. Jakovetić, J. Xavier, and J. M. Moura, “Fast distributed gradient methods,” IEEE Transactions on Automatic Control, vol. 59, no. 5, pp. 1131–1146, 2014.

A. Nedić and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” IEEE Transactions on Automatic Control, vol. 60, no. 3, pp. 601–615, 2014.

J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Püschel, “D-ADMM: A communication-efficient distributed algorithm for separable optimization,” IEEE Transactions on Signal Processing, vol. 61, no. 10, pp. 2718–2723, 2013.

P. Bianchi and J. Jakubowicz, “Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization,” IEEE Transactions on Automatic Control, vol. 58, no. 2, pp. 391–405, 2012.

P. Di Lorenzo and G. Scutari, “Next: In-network nonconvex optimization,” IEEE Transactions on Signal and Information Processing over Networks, vol. 2, no. 2, pp. 120–136, 2016.

Y. Sun, G. Scutari, and D. Palomar, “Distributed nonconvex multiagent optimization over time-varying networks,” in Proceedings of Asilomar Conference on Signals, Systems and Computers, 2016, pp. 788–794.

G. Scutari and Y. Sun, “Distributed nonconvex constrained optimization over time-varying digraphs,” Mathematical Programming, vol. 176, no. 1-2, pp. 497–544, 2019.

T. Tatarenko and B. Touri, “Non-convex distributed optimization,” IEEE Transactions on Automatic Control, vol. 62, no. 8, pp. 3744–3757, 2017.

Y. Tian, Y. Sun, and G. Scutari, “Asy-sonata: Achieving linear convergence in distributed asynchronous multiagent optimization,” in 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton). IEEE, 2018, pp. 543–551.

H. Sun and M. Hong, “Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms,” IEEE Transactions on Signal Processing, vol. 67, no. 22, pp. 5912–5928, 2019.

A. Daneshmand, G. Scutari, and V. Kungurtsev, “Second-order guarantees of gradient algorithms over networks,” in Proceedings of Allerton Conference on Communication, Control, and Computing, 2018, pp. 359–365.

S. Vlaski and A. H. Sayed, “Distributed learning in non-convex environments–Part I: Agreement at a linear rate,” 2019, online: https://arxiv.org/pdf/1907.01848.pdf.
[37] ——, “Distributed learning in non-convex environments–Part II: Polynomial escape from saddle-points,” 2019, online: https://arxiv.org/pdf/1907.01849.pdf.

[38] ——, “Second-order guarantees of stochastic gradient descent in non-convex optimization,” arXiv preprint arXiv:1908.07023, 2019.

[39] B. Swenson, R. Murray, H. V. Poor, and S. Kar, “Distributed gradient descent: Nonconvergence to saddle points and the stable-manifold theorem,” in Proceedings of Allerton Conference on Communication, Control, and Computing, 2019, pp. 595–601.

[40] B. Swenson, S. Kar, H. V. Poor, and J. M. F. Moura, “Annealing for distributed global optimization,” 2019, to appear in Proceedings of IEEE Conference on Decision and Control, arXiv preprint arXiv:1903.07258.

[41] ——, “Distributed global optimization by annealing,” 2019, to appear in Proceedings of IEEE International Workshop on Computational Advances in in Multisensor Adaptive Processing, arXiv preprint arXiv:1907.08802.

[42] B. Swenson, A. Sridhar, and H. V. Poor, “On distributed stochastic gradient algorithms for global optimization,” 2019, to appear in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, arXiv preprint arXiv:1910.09587.

[43] V. Kungurtsev, “Stochastic gradient Langevin dynamics on a distributed network,” arXiv preprint arXiv:2001.00665, 2020.

[44] A. G. Dimakis, S. Kar, J. M. Moura, M. G. Rabbat, and A. Scaglione, “Gossip algorithms for distributed signal processing,” Proceedings of the IEEE, vol. 98, no. 11, pp. 1847–1864, 2010.

[45] M. Benaim, “A dynamical system approach to stochastic approximations,” SIAM Journal on Control and Optimization, vol. 34, no. 2, pp. 437–472, 1996.

[46] T. Kato, Perturbation Theory for Linear Operators. Springer Science & Business Media, 2013, vol. 132.

[47] C. Chicone, Ordinary Differential Equations with Applications. Springer Science & Business Media, 2006, vol. 34.

[48] F. R. K. Chung, Spectral Graph Theory. American Mathematical Society, 1997, no. 92.

[49] M. Benaim, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions,” SIAM Journal Control and Optimization, vol. 44, no. 1, pp. 328–348, 2005.

[50] S. Kar, J. M. Moura, and H. V. Poor, “Distributed linear parameter estimation: Asymptotically efficient adaptive strategies,” SIAM Journal on Control and Optimization, vol. 51, no. 3, pp. 2200–2229, 2013.

[51] L. Perko, Differential Equations and Dynamical Systems. Springer Science & Business Media, 2013, vol. 7.