2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Second-order Properties of Noisy Distributed Gradient Descent

Lei Qin, Michael Cantoni, and Ye Pu

***Abstract*— We study a fixed step-size distributed gradient descent algorithm for solving optimization problems in which the objective is a finite sum of smooth but possibly non-convex functions. Random perturbations of the gradient descent directions are introduced at each step to actively evade saddle points. Under certain regularity conditions, and with a suitable step-size, it is established that each agent converges to a neighborhood of a local minimizer; the size of the neighborhood depends on the step-size and a probabilistic confidence parameter. A numerical example is presented to illustrate the effectiveness of the random perturbations in terms of escaping saddle points in fewer iterations than without the perturbations.**

***Index Terms*— Non-convex optimization; consensus-based distributed optimisation; first-order methods; random perturbations; evading saddle points**

I. Introduction

We consider the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \triangleq \min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} f_i(\mathbf{x}), \tag{1}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is *smooth* but possibly *non-convex*, and $\mathbf{x} \in \mathbb{R}^n$ is the decision vector. The aim is to employ m agents to iteratively solve the optimization problem in (1), over an undirected and connected network graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Each agent $i \in \mathcal{V} := \{1, \dots, m\}$ only knows the corresponding function f_i and its gradient. The pair of agents $(i, j) \in \mathcal{V} \times \mathcal{V}$ is able to directly exchange information if and only if $(i, j) \in \mathcal{E}$. Collaborative distributed optimization over a network is of significant interest in the contexts of control, learning and estimation, particularly in large-scale system scenarios, such as unmanned vehicle systems [1], electric power systems [2], transit systems [3], and wireless sensor networks [4].

In optimization, two primary classes of distributed methods can be identified: dual decomposition methods and consensus-based methods. Dual decomposition methods minimize an augmented Lagrangian based on agreement-enforcing constraints through iterative primal-dual updates [5]. The distributed dual decomposition algorithm in [6] involves agents alternating between updating their primal and dual variables and communicating with their neighbors. In [7], it is established that distributed alternating direction method of multipliers (ADMM) exhibits linear convergence rates in *strongly convex* settings. Consensus-based methods, originating from the models in [8], seek to eliminate disagreements through local iterate exchange and weighted averaging. The distributed (sub)gradient methods proposed in [9] and [10] use this concept to solve problem (1) with convex f_i. In the case of diminishing step-size, each agent converges to an optimizer [10]; with constant step-size, convergence is typically faster, but only to the vicinity of an optimizer [9].

This paper investigates a perturbed variant of **D**istributed **G**radient **D**escent (**DGD**) [9], using a constant step-size. In fixed step-size **DGD**, the update for each agent $i \in \mathcal{V}$ at iteration k is given by

$$\hat{\mathbf{x}}_i^{k+1} = \sum_{j=1}^{m} \mathbf{W}_{ij}\hat{\mathbf{x}}_j^k - \alpha\nabla f_i(\hat{\mathbf{x}}_i^k), \tag{2}$$

where $\alpha > 0$ is the constant step-size, ∇f_i is the gradient of f_i, $\hat{\mathbf{x}}_i \in \mathbb{R}^n$ is the local copy of the decision vector $\mathbf{x}$ at agent $i \in \mathcal{V}$, and $\mathbf{W}_{ij}$ is the scalar entry in the i-th row and j-th column of a given mixing matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$. The mixing matrix is consistent with the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ in the sense that $\mathbf{W}_{ii} > 0$ for all $i \in \mathcal{V}$, $\mathbf{W}_{ij} > 0$ if $(i, j) \in \mathcal{E}$, and $\mathbf{W}_{ij} = 0$ otherwise. The convergence rates of **DGD** in *strongly convex* settings are examined in [11] and [12]. Several techniques like Nesterov momentum [13], inexact proximal-gradient method [14], and gradient tracking [15], [16] are employed in *(strongly) convex* settings to enhance convergence rates, handle non-smooth functions, and achieve exact consensus with a constant step-size, respectively.

In *non-convex* settings, gradient descent methods may struggle due to saddle points. Use of the Hessian can be beneficial, but may be computationally expensive for large problems. In [17], fixed step-size **DGD** is shown to retain the property of convergence to a neighborhood of a consensus stationary solution under some regularity assumptions. Further, it is shown in [18] that fixed step-size **DGD** converges almost surely to a neighborhood of second-order stationary solutions. However, random initialization is required to avoid the zero measure manifold of saddle point attraction, and moreover, the underlying analysis lacks techniques for actively escaping saddle points.

Standard gradient descent methods can take exponential time to escape saddle points [19]. It is shown that adding noise (random perturbations) to descent directions is effective for escaping saddle points in [20] and [21]. However, these works are limited to centralized algorithms. In [22], it is shown that distributed stochastic gradient descent converges to local minima almost surely when diminishing step sizes are used. For constant step-size, the diffusion strategy with stochastic gradient in [23] and [24] only returns approximate

This work was supported by a Melbourne Research Scholarship and the Australian Research Council (DE220101527 and DP210103272).

L. Qin, M. Cantoni, and Y. Pu are with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville VIC 3010, Australia leqin@student.unimelb.edu.au, {cantoni, ye.pu}@unimelb.edu.au.

second-order stationary points, rather than outcomes that lie in a neighborhood of a local minimizer with controllable size.

In this paper, the main contribution is an analysis of a fixed step-size noisy distributed gradient descent (**NDGD**) algorithm for solving the optimization problem in (1). To this end, we expand upon and combine ideas from [20] and [21] on centralized stochastic gradient descent, and from [11], [17], [18] on unperturbed **DGD**. In particular, random perturbations are added to the gradient descent directions at each step to actively evade saddle points. It is established that under certain regularity conditions, and with a suitable step-size, each agent converges to a neighborhood of a local minimizer. We determine a probabilistic upper bound for the distance between the iterate at each agent and the set of local minimizers after a sufficient number of iterations. A numerical example is presented to illustrate the effectiveness of the algorithm in terms of escaping from the vicinity of a saddle point in fewer iterations than the standard (i.e., unperturbed) fixed step-size **DGD**.

A. Notation

Let $\mathbf{I}_n$ denote the $n \times n$ identity matrix, $\mathbf{1}_n$ denote the n-vector with all entries equal to 1, and $\mathbf{A}_{ij}$ denote the entry in the i-th row and j-th column of the matrix $\mathbf{A}$. For a square symmetric matrix $\mathbf{B}$, we use $\lambda_{\min}(\mathbf{B})$, $\lambda_{\max}(\mathbf{B})$ and $\|\mathbf{B}\|$ to denote its minimum eigenvalue, maximum eigenvalue and spectral norm, respectively. The Kronecker product is denoted by $\otimes$. The distance from the point $\mathbf{x} \in \mathbb{R}^n$ to a given set $\mathcal{Y} \subseteq \mathbb{R}^n$ is denoted by $\text{dist}(\mathbf{x}, \mathcal{Y}) := \inf_{\mathbf{y} \in \mathcal{Y}} \|\mathbf{x} - \mathbf{y}\|$. We say that a point $\mathbf{x}$ is δ-close to a point $\mathbf{y}$ (resp., a set $\mathcal{Y}$) if $\text{dist}(\mathbf{x}, \mathbf{y}) \leq \delta$ (resp., $\text{dist}(\mathbf{x}, \mathcal{Y}) \leq \delta$). We use the Bachmann–Landau (asymptotic) notations including $\mathcal{O}(g(x,y))$, $\Omega(g(x,y))$ and $\Theta(g(x,y))$ to hide dependence on variables other than x and y.

II. Problem Setup and Supporting Results

In this section, we present a reformulation of the optimization problem defined in (1) and provide a list of assumptions used in subsequent analysis. We then provide some supporting lemmas (see Lemma 2.3.1-2.3.4), which will be used to establish relevant properties of the local minimizers of f (see Theorem 2.1).

A. Problem Setup

By introducing additional local variables, the optimization problem in (1) can be reformulated as

$$\begin{aligned} \min_{\hat{\mathbf{x}} \in (\mathbb{R}^n)^m} F(\hat{\mathbf{x}}) \triangleq \min_{\hat{\mathbf{x}} \in (\mathbb{R}^n)^m} \sum_{i=1}^m f_i(\hat{\mathbf{x}}_i), \\ \text{s.t. } \hat{\mathbf{x}}_i = \hat{\mathbf{x}}_j \text{ for all } (i,j) \in \mathcal{E}, \end{aligned} \tag{3}$$

where $\hat{\mathbf{x}}_i \in \mathbb{R}^n$ is the local copy of the decision vector $\mathbf{x}$ at agent $i \in \mathcal{V}$, and $\hat{\mathbf{x}} = [\hat{\mathbf{x}}_1^T, \cdots, \hat{\mathbf{x}}_m^T]^T \in (\mathbb{R}^n)^m$.

Definition 2.1 *For differentiable function h, a point $\mathbf{x}$ is said to be first-order stationary if $\|\nabla h(\mathbf{x})\| = 0$.*

Definition 2.2 *For twice differentiable function h, a first-order stationary point $\mathbf{x}$ is: (i) a local minimizer, if $\nabla^2 h(\mathbf{x}) \succ 0$; (ii) a local maximizer, if $\nabla^2 h(\mathbf{x}) \prec 0$; and (iii) a saddle point if $\lambda_{\min}(\nabla^2 h(\mathbf{x})) < 0$ and $\lambda_{\max}(\nabla^2 h(\mathbf{x})) > 0$.*

Assumption 2.1 (Local regularity) *The function f in (1) is such that for all first-order stationary points $\mathbf{x}$, either $\lambda_{\min}(\nabla^2 f(\mathbf{x})) > 0$ (i.e., $\mathbf{x}$ is a local minimizer), or $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < 0$ (i.e., $\mathbf{x}$ is a saddle point or a maximizer).*

Assumption 2.2 (Lipschitz gradient) *Each objective f_i has $L_{f_i}^g$-Lipschitz continuous gradient, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and each $i \in \mathcal{V}$, $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L_{f_i}^g \|\mathbf{x} - \mathbf{y}\|$.*

Assumption 2.3 (Lipschitz Hessian) *Each objective f_i has $L_{f_i}^H$-Lipschitz continuous Hessian, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and each $i \in \mathcal{V}$, $\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f_i(\mathbf{y})\| \leq L_{f_i}^H \|\mathbf{x} - \mathbf{y}\|$.*

If Assumption 2.2 holds, then F defined in (3) has L_F^g-Lipschitz continuous gradient with $L_F^g = \max_i\{L_{f_i}^g\}$. Further, if Assumption 2.3 holds, then F has L_F^H-Lipschitz continuous Hessian with $L_F^H = \max_i\{L_{f_i}^H\}$.

Assumption 2.4 (Coercivity and properness) *Each local objective f_i is coercive (i.e., its sublevel set is compact) and proper (i.e., not everywhere infinite).*

B. Distributed Gradient Descent

Assumption 2.5 (Network) *The undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is connected.*

The **DGD** algorithm in (2), with constant step-size $\alpha > 0$, can be written in a matrix/vector form as

$$\hat{\mathbf{x}}^{k+1} = \hat{\mathbf{W}}\hat{\mathbf{x}}^k - \alpha \nabla F(\hat{\mathbf{x}}^k), \tag{4}$$

where $\hat{\mathbf{W}} := \mathbf{W} \otimes \mathbf{I}_n$. Note that from this point on, the mixing matrix $\mathbf{W}$ is taken to be symmetric, doubly stochastic and strictly diagonally dominant, i.e., $\mathbf{W}_{ii} > \sum_{j \neq i} \mathbf{W}_{ij}$ for all $i \in \mathcal{V}$. Thus, $\mathbf{W}$ is positive definite by the Gershgorin circle theorem. As proposed in some early works, including [11], [17], [18], we can analyze the convergence properties using an auxiliary function. Let the Q_α denote the auxiliary function,

$$\begin{aligned} Q_\alpha(\hat{\mathbf{x}}) &= \sum_{i=1}^m f_i(\hat{\mathbf{x}}_i) + \frac{1}{2\alpha} \sum_{i=1}^m \sum_{j=1}^m (\mathbf{I}_m - \mathbf{W})_{ij} (\hat{\mathbf{x}}_i)^T (\hat{\mathbf{x}}_j) \\ &= F(\hat{\mathbf{x}}) + \frac{1}{2\alpha} \|\hat{\mathbf{x}}\|^2_{\mathbf{I}_{mn} - \hat{\mathbf{W}}}, \end{aligned} \tag{5}$$

consisting of the objective function in (3) and a quadratic penalty, which depends on the step-size and the mixing matrix. We use $\hat{\mathbf{x}}^*$ to denote a local minimizer of Q_α. Note that the **DGD** update (4) applied to (3) can be interpreted as an instance of the standard gradient descent algorithm applied to (5), i.e.,

$$\hat{\mathbf{x}}^{k+1} = \hat{\mathbf{x}}^k - \alpha \nabla Q_\alpha(\hat{\mathbf{x}}^k). \tag{6}$$

Thus, iteratively running (4) and (6) from the same initialization yields the same sequence of iterates.

If Assumption 2.2 holds, then Q_α defined in (5) has $L^g_{Q_\alpha}$-Lipschitz continuous gradient with $L^g_{Q_\alpha} = L^g_F + \alpha^{-1}(1 - \lambda_{\min}(\mathbf{W})) = \max_i\{L^g_{f_i}\} + \alpha^{-1}(1 - \lambda_{\min}(\mathbf{W}))$. We have that $1-\lambda_{\min}(\mathbf{W}) \geq 0$ because the spectrum of a symmetric, positive definite and doubly stochastic matrix is contained in the interval $(0, 1]$ by the Perron–Frobenius theorem, with 1 being the only largest eigenvalue (the Perron root). Further, if Assumption 2.3 holds, then Q_α has $L^H_{Q_\alpha}$-Lipschitz continuous Hessian with $L^H_{Q_\alpha} = \max_i\{L^H_{f_i}\} = L^H_F$.

C. Relationships between Local minimizers of f and Q_α

In this section, we show that $\hat{\mathbf{x}}^*_i$, the component of a local minimizer $\hat{\mathbf{x}}^*$ of Q_α associated with agent i, can be made arbitrarily close to the set of local minimizers $\mathbf{x}^*$ of f by choosing sufficiently small $\alpha > 0$ (see Theorem 2.1). This expands upon the outcomes in [18] and is further used to prove the main theorem (see Theorem 3.1). Full proofs of intermediate lemmas are provided in the extended version of this work [25].

Let $\mathcal{X}^*_f$ and $\hat{\mathcal{X}}^*_{Q_\alpha}$ denote the set of local minimizers of f and Q_α, respectively:

$$\begin{aligned} \mathcal{X}^*_f &:= \{\mathbf{x} \in \mathbb{R}^n : \nabla f(\mathbf{x}) = 0,\ \nabla^2 f(\mathbf{x}) \succ 0\}, \\ \hat{\mathcal{X}}^*_{Q_\alpha} &:= \{\hat{\mathbf{x}} \in (\mathbb{R}^n)^m : \nabla Q_\alpha(\hat{\mathbf{x}}) = 0,\ \nabla^2 Q_\alpha(\hat{\mathbf{x}}) \succ 0\}. \end{aligned} \tag{7}$$

Lemma 2.3.1 *Let Assumption 2.5 hold. Given $\alpha > 0$, let $\hat{\mathbf{x}}^*$ be a local minimizer of Q_α. Then, for each $i \in \mathcal{V}$,*

$$\|\hat{\mathbf{x}}^*_i - \bar{\mathbf{x}}^*\| \leq \alpha \cdot \frac{\|\nabla F(\hat{\mathbf{x}}^*)\|}{1 - \lambda_2},$$

where $\bar{\mathbf{x}}^* = \frac{1}{m}(\mathbf{1}_m \otimes \mathbf{I}_n)^T \hat{\mathbf{x}}^*$, and $0 < \lambda_2 < 1$ is the second-largest eigenvalue value of $\mathbf{W}$.

Lemma 2.3.2 *Let Assumptions 2.2, 2.5 hold. Given $\alpha > 0$, let $\hat{\mathbf{x}}^*$ be a local minimizer of Q_α. Then*

$$\|\nabla f(\bar{\mathbf{x}}^*)\| \leq \alpha \cdot L^g_F \frac{m\sqrt{m}\|\nabla F(\hat{\mathbf{x}}^*)\|}{1 - \lambda_2},$$

where $\bar{\mathbf{x}}^* = \frac{1}{m}(\mathbf{1}_m \otimes \mathbf{I}_n)^T \hat{\mathbf{x}}^*$ and $0 < \lambda_2 < 1$ is the second-largest eigenvalue value of $\mathbf{W}$.

Lemma 2.3.3 *Let Assumptions 2.3, 2.5 hold. Given $\alpha > 0$, let $\hat{\mathbf{x}}^*$ be a local minimizer of Q_α. Then*

$$\lambda_{\min}(\nabla^2 f(\bar{\mathbf{x}}^*)) \geq -\alpha \cdot L^H_F \frac{m^2\|\nabla F(\hat{\mathbf{x}}^*)\|}{1 - \lambda_2},$$

where $\bar{\mathbf{x}}^* = \frac{1}{m}(\mathbf{1}_m \otimes \mathbf{I}_n)^T \hat{\mathbf{x}}^*$ and $0 < \lambda_2 < 1$ is the second-largest eigenvalue value of $\mathbf{W}$.

Lemma 2.3.4 *Let Assumptions 2.1, 2.2, 2.3 hold. Then, for any given compact set $\mathcal{X} \subset \mathbb{R}^n$,*

$$\lim_{\alpha \downarrow 0}(\sup\{\text{dist}(\mathbf{x}, \mathcal{X}^*_f) : \mathbf{x} \in \mathcal{X}^\alpha_f \cap \mathcal{X}\}) = 0,$$

where $\mathcal{X}^\alpha_f := \{\mathbf{x} : \|\nabla f(\mathbf{x})\| \leq \alpha \cdot c_1, \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\alpha \cdot c_2\}$, with

$$c_1 = L^g_F \frac{m\sqrt{m}\|\nabla F(\hat{\mathbf{x}}^*)\|}{1 - \lambda_2},\ c_2 = L^H_F \frac{m^2\|\nabla F(\hat{\mathbf{x}}^*)\|}{1 - \lambda_2}.$$

By combining Lemmas 2.3.1 through 2.3.4, the following theorem can be established.

Theorem 2.1 *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5 hold. Given $\Delta_1 > 0$, there exists threshold $\bar{\alpha}(\Delta_1) > 0$ such that, if $0 < \alpha \leq \bar{\alpha}(\Delta_1)$, and $\hat{\mathbf{x}}^*$ is a local minimizer of Q_α, then $\text{dist}(\hat{\mathbf{x}}^*_i, \mathcal{X}^*_f) \leq \Delta_1$ for each $i \in \mathcal{V}$.*

Proof: Given $\alpha > 0$, by the triangle inequality, $\text{dist}(\hat{\mathbf{x}}^*_i, \mathcal{X}^*_f) \leq \|\hat{\mathbf{x}}^*_i - \bar{\mathbf{x}}^*\| + \text{dist}(\bar{\mathbf{x}}^*, \mathcal{X}^*_f)$, where $\bar{\mathbf{x}}^* = \frac{1}{m}(\mathbf{1}_m \otimes \mathbf{I}_n)^T \hat{\mathbf{x}}^*$. By coercivity and properness of each f_i (see Assumption 2.4), F is coercive and proper. Therefore, $\hat{\mathcal{X}}^*_{Q_\alpha}$ is bounded, and there exists an upper bound $G > 0$ such that for all $\hat{\mathbf{x}}^* \in \hat{\mathcal{X}}^*_{Q_\alpha}$, $\|\nabla F(\hat{\mathbf{x}}^*)\| \leq G$. By Lemma 2.3.1, if

$$0 < \alpha \leq \bar{\alpha}_1(\Delta_1) := \frac{\Delta_1(1 - \lambda_2)}{2G}$$

and $\hat{\mathbf{x}}^* \in \hat{\mathcal{X}}^*_{Q_\alpha}$ defined in (7), then $\|\hat{\mathbf{x}}^*_i - \bar{\mathbf{x}}^*\| \leq \Delta_1/2$ holds for each $i \in \mathcal{V}$. Now, note that $\bar{\mathbf{x}}^* \in \mathcal{X}^\alpha_f$ in view of Lemmas 2.3.2 and 2.3.3, with $\mathcal{X}^\alpha_f$ as defined in Lemma 2.3.4. As such, by application of Lemma 2.3.4 with $\mathcal{X} = \{\bar{\mathbf{x}}^*\}$, there exists $\bar{\alpha}_2(\Delta_1) > 0$ such that if $0 < \alpha \leq \bar{\alpha}_2(\Delta_1)$, then $\text{dist}(\bar{\mathbf{x}}^*, \mathcal{X}^*_f) \leq \Delta_1/2$ holds. Therefore, if

$$0 < \alpha \leq \bar{\alpha}(\Delta_1) := \min\{\bar{\alpha}_1(\Delta_1), \bar{\alpha}_2(\Delta_1)\},$$

then $\text{dist}(\hat{\mathbf{x}}^*_i, \mathcal{X}^*_f) \leq \Delta_1$ as claimed. ∎

III. Method and Main Result

The **N**oisy **D**istributed **G**radient **D**escent (**NDGD**) method is formulated in Algorithm 1, as a variant of fixed step-size **DGD**. The key innovation is the addition of random perturbations to the distributed gradient descent directions at each iteration. The required properties of the noise ξ^k_i in Algorithm 1 are presented in Theorem 3.1, which establishes the second-order properties of the **NDGD** algorithm.

Algorithm 1 **N**oisy **D**istributed **G**radient **D**escent (**NDGD**)

Initialization;
for $k = 0, 1, \cdots$ **do**
  for $i = 0, 1, \cdots, m$ **do**
    Sample i.i.d ξ^k_i;
    $\hat{\mathbf{x}}^{k+1}_i = \sum_{j=1}^m \mathbf{W}_{ij}\hat{\mathbf{x}}^k_j - \alpha(\nabla f_i(\hat{\mathbf{x}}^k_i) + \xi^k_i)$;
  end for
end for

As per (7), recall that $\mathcal{X}^*_f$ and $\hat{\mathcal{X}}^*_{Q_\alpha}$ denote the set of local minimizers of f and Q_α, respectively. Further, given $\epsilon > 0$, $\gamma > 0$, $\mu > 0$, $\delta > 0$, and $\alpha > 0$, define

$$\begin{aligned} \mathcal{L}^1_{\alpha,\epsilon} &:= \{\hat{\mathbf{x}} : \|\nabla F(\hat{\mathbf{x}}) + \alpha^{-1}(\mathbf{I}_{mn} - \hat{\mathbf{W}})\hat{\mathbf{x}}\| \geq \epsilon\}, \\ \mathcal{L}^2_{\alpha,\gamma} &:= \{\hat{\mathbf{x}} : \lambda_{\min}(\nabla^2 F(\hat{\mathbf{x}})) \leq -\gamma - \alpha^{-1}\}, \\ \mathcal{L}^3_{\alpha,\mu,\delta} &:= \{\hat{\mathbf{x}} : \lambda_{\min}(\nabla^2 F(\hat{\mathbf{x}})) \geq \mu,\ \text{dist}(\hat{\mathbf{x}}, \hat{\mathcal{X}}') \leq \delta\}, \end{aligned} \tag{8}$$

where $\hat{\mathcal{X}}' := \{\hat{\mathbf{x}} : \|\nabla F(\hat{\mathbf{x}}) + \alpha^{-1}(\mathbf{I}_{mn} - \hat{\mathbf{W}})\hat{\mathbf{x}}\| = 0,\ \lambda_{\min}(\nabla^2 F(\hat{\mathbf{x}})) > 0\}$. With this, the main result of the paper is formulated below; a proof is given in Section IV. Note that the focus of the result relates to the role of the given step-size α and confidence parameter ζ; as such, the

factors with polynomial dependence on all other parameters (including Δ_1, ϵ, γ, μ, δ and σ) are hidden.

Theorem 3.1 *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5 hold, and given $\Delta_1 > 0$ and $0 < \zeta < 1$, suppose the following:*

1) *There exist $\epsilon > 0$, $\gamma \in (0, L_F^g]$, $\mu \in (0, L_F^g]$, $\delta > 0$, and $\alpha \in (0, \hat{\alpha}(\Delta_1, \zeta)]$, such that $\mathcal{L}^1_{\alpha,\epsilon} \cup \mathcal{L}^2_{\alpha,\gamma} \cup \mathcal{L}^3_{\alpha,\mu,\delta} = (\mathbb{R}^n)^m$, where*

$$\hat{\alpha}(\Delta_1, \zeta) := \min\{\bar{\alpha}(\Delta_1), \frac{\sqrt{2}-1}{L_F^g}, \frac{\lambda_{\min}(\mathbf{W})}{L_F^g \cdot \max\{1, \log(\zeta^{-1})\}}\} > 0$$

with $\bar{\alpha}(\Delta_1)$ as per Theorem 2.1;

2) *The random perturbation ξ_i^k at step $k > 0$ is i.i.d. and zero mean with variance $\sigma^2 \leq \sigma^2_{\max}(\epsilon) := (\lambda_{\min}(\mathbf{W})\epsilon^2)/(mn)$;*
3) *The generated sequence $\{Q_\alpha(\hat{\mathbf{x}}^k)\}$ is bounded.*

Then, with probability at least $1 - \zeta$, after $K(\alpha, \zeta) = \mathcal{O}(\alpha^{-2} \log \zeta^{-1})$ iterations, Algorithm 1 reaches a point $\hat{\mathbf{x}}^{K(\alpha,\zeta)} \in (\mathbb{R}^n)^m$ that is $\Delta_2(\alpha, \zeta)$-close to $\mathcal{X}^*_{Q_\alpha}$, where $\Delta_2(\alpha, \zeta) = \mathcal{O}(\sqrt{\alpha \log(\alpha^{-1}\zeta^{-1})})$. Moreover, $\hat{\mathbf{x}}^* = \inf_{\hat{\mathbf{x}} \in \mathcal{X}^*_{Q_\alpha}} \|\hat{\mathbf{x}}_i^{K(\alpha,\zeta)} - \hat{\mathbf{x}}\|$ is such that $\hat{\mathbf{x}}_i^*$ is Δ_1-close to $\mathcal{X}_f^*$, whereby for all $i \in \mathcal{V}$,

$$\operatorname{dist}(\hat{\mathbf{x}}_i^{K(\alpha,\zeta)}, \mathcal{X}_f^*) \leq \Delta_1 + \Delta_2(\alpha, \zeta).$$

Remark 1 *Intuitively, condition 1) in Theorem 3.1 requires all points where the gradient of Q_α is small to either result in sufficient descent or reside within a neighborhood of a local minimizer, where local strong convexity holds.*

Remark 2 *For condition 2) in Theorem 3.1, one way to generate the required i.i.d. noise is to sample ξ_i^k uniformly from an n-dimensional sphere with the radius r. This ensures $\mathbb{E}(\xi_i^k) = \mathbf{0}$, $\mathbb{E}(\xi_i^k(\xi_i^k)^T) = (r^2/n)\mathbf{I}_n$, and $\|\xi_i^k\| \leq r$ for all $i \in \mathcal{V}$ and $k \in \mathbb{N}$. By choosing $r^2 \leq n\sigma^2_{\max}(\epsilon)$, we have $\mathbb{E}(\xi^k(\xi^k)^T) = (r^2/n)\mathbf{I}_{mn} \preceq \sigma^2_{\max}(\epsilon)\mathbf{I}_{mn}$.*

Second-order guarantees of **DGD** have been studied in [18] and [22] based on the almost sure non-convergence to saddle points with random initialization. In this paper, we propose to use random perturbations to actively evade saddle points. The second-order guarantees of **NDGD** stated in Theorem 3.1 do not require any additional initialization conditions. Second-order guarantees of the stochastic variant of **DGD** have been studied in [23] and [24], although they only show the convergence to an approximate second-order stationary point. Here, an upper bound is given for the distance between the iterate at each agent and the set of local minimizers after a sufficient number of iterations.

IV. Proof of Theorem 3.1

A proof of Theorem 3.1 is provided in this section. First, we consider the behavior of **NDGD** for the following three different cases, in line with the development of the related result in [20] for centralized gradient descent: i) large in norm $\nabla Q_\alpha(\hat{\mathbf{x}}^k)$ (see Lemma 4.1.1); ii) sufficiently negative $\lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}}^k))$ (see Lemma 4.1.2); and iii) $\hat{\mathbf{x}}^k$ in a neighborhood of the local minimizers of Q_α with local strong convexity (see Lemma 4.1.3). Combining the outcome of this with Theorem 2.1, we then prove that with probability at least $1 - \zeta$, after $K(\alpha, \zeta)$ iterations, the state $\hat{\mathbf{x}}_i^k$ of each agent $i \in \mathcal{V}$ in the **NDGD** algorithm is $\Delta_1 + \Delta_2(\alpha, \zeta)$-close to some local minimizer of f.

A. Behavior of NDGD for three different cases

The following lemmas rely on the proofs of Lemma 16 and Lemma 17 in [20]. Detailed proofs can be found in the extended version of this work [25]. Given $\epsilon > 0$, $\gamma > 0$, $\mu > 0$, $\delta > 0$, and $\alpha > 0$, define

$$\begin{aligned} \mathcal{I}^1_{\alpha,\epsilon} &:= \{\hat{\mathbf{x}} : \|\nabla Q_\alpha(\hat{\mathbf{x}})\| \geq \epsilon\}, \\ \mathcal{I}^2_{\alpha,\gamma} &:= \{\hat{\mathbf{x}} : \Lambda_\alpha(\hat{\mathbf{x}}) \leq -\gamma\}, \\ \mathcal{I}^3_{\alpha,\mu,\delta} &:= \{\hat{\mathbf{x}} : \Lambda_\alpha(\hat{\mathbf{x}}) \geq \mu,\ \operatorname{dist}(\hat{\mathbf{x}}, \hat{\mathcal{X}}^*_{Q_\alpha}) \leq \delta\}, \end{aligned} \tag{9}$$

where $\Lambda_\alpha(\hat{\mathbf{x}}) = \lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}}))$.

Lemma 4.1.1 *Let Assumption 2.2 hold. Given $\epsilon > 0$, suppose the random perturbation ξ_i^k in Algorithm 1 is i.i.d. and zero mean with variance $\sigma^2 \leq \sigma^2_{\max}(\epsilon) := (\lambda_{\min}(\mathbf{W})\epsilon^2)/(mn)$. Then, given $0 < \alpha \leq 1/L_F^g$, for any $\hat{\mathbf{x}}^k$ such that $\|\nabla Q_\alpha(\hat{\mathbf{x}}^k)\| \geq \epsilon$, after one iteration,*

$$\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{k+1}) \mid \hat{\mathbf{x}}^k] - Q_\alpha(\hat{\mathbf{x}}^k) \leq -l_1(\alpha),$$

where $l_1(\alpha) = \Omega(\alpha)$.

Lemma 4.1.2 *Let Assumptions 2.2, 2.3 hold. Let $\gamma \in (0, L_F^g]$. Given $\epsilon > 0$, suppose the random perturbation ξ_i^k in Algorithm 1 is i.i.d. and zero mean with variance $\sigma^2 \leq \sigma^2_{\max}(\epsilon) := (\lambda_{\min}(\mathbf{W})\epsilon^2)/(mn)$. Further, given $0 < \alpha \leq (\sqrt{2}-1)/L_F^g$, suppose that the generated sequence $\{Q_\alpha(\hat{\mathbf{x}}^k)\}$ is bounded. Then, for any $\hat{\mathbf{x}}^k$ with $\|\nabla Q_\alpha(\hat{\mathbf{x}}^k)\| < \epsilon$ and $\lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}})) \leq -\gamma$, there exists a number of steps $T(\hat{\mathbf{x}}^k) > 0$ such that*

$$\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{k+T(\hat{\mathbf{x}}^k)}) \mid \hat{\mathbf{x}}^k] - Q_\alpha(\hat{\mathbf{x}}^k) \leq -l_2(\alpha),$$

where $l_2(\alpha) = \Omega(\alpha)$. The number of steps $T(\hat{\mathbf{x}}^k)$ has a fixed upper bound $T_{\max}(\alpha)$ that is independent of $\hat{\mathbf{x}}^k$, i.e., $T(\hat{\mathbf{x}}^k) \leq T_{\max}(\alpha) = \mathcal{O}(\alpha^{-1})$ for all $\hat{\mathbf{x}}^k$.

Lemma 4.1.3 *Let Assumptions 2.2 hold. Let $\mu \in (0, L_F^g]$. Given $\epsilon > 0$, suppose the random perturbation ξ_i^k in Algorithm 1 is i.i.d. and zero mean with variance $\sigma^2 \leq \sigma^2_{\max}(\epsilon) := (\lambda_{\min}(\mathbf{W})\epsilon^2)/(mn)$. Further, given $\delta > 0$, $0 < \alpha \leq (\lambda_{\min}(\mathbf{W}))/(L_F^g \cdot \max\{1, \log(\zeta^{-1})\})$ and local minimizer $\hat{\mathbf{x}}^* \in \mathcal{X}^*_{Q_\alpha}$, suppose $\lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}})) \geq \mu$ for all $\hat{\mathbf{x}}$ such that $\|\hat{\mathbf{x}} - \hat{\mathbf{x}}^*\| < \delta$. Then, there exists $\delta_1(\alpha) = \mathcal{O}(\sqrt{\alpha}) \in [0, \delta)$ such that, for any $\hat{\mathbf{x}}^k$ with $\|\hat{\mathbf{x}}^k - \hat{\mathbf{x}}^*\| < \delta_1(\alpha)$, with probability at least $1 - \zeta/2$,*

$$\|\hat{\mathbf{x}}^{k+s} - \hat{\mathbf{x}}^*\| \leq \Delta_2(\alpha, \zeta)$$

for all $s \leq K(\alpha, \zeta) = \mathcal{O}(\alpha^{-2} \log(\zeta^{-1}))$, where $\Delta_2(\alpha, \zeta) = \mathcal{O}(\sqrt{\alpha \log(\alpha^{-1}\zeta^{-1})}) < \delta$.

Intuitively, the above results show that when the norm of $\nabla Q_\alpha(\hat{\mathbf{x}}^k)$ is large enough (see Lemma 4.1.1), the expectation of the function value decreases by a certain amount after one

iteration. For $\hat{\mathbf{x}}^k$ with small gradient and sufficiently negative $\lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}}^k))$ (see Lemma 4.1.2), there exists upper bound $T_{\max}(\alpha)$ such that the expectation of the function value decreases by a certain amount after $T \leq T_{\max}(\alpha)$ iterations. Finally, when the iterate $\hat{\mathbf{x}}^k$ is close enough to a local minimizer (see Lemma 4.1.3), with high probability subsequent iterates do not leave the neighborhood.

B. Main proof

Proof: The main proof includes two steps: i) it is shown that three sets defined in (9) cover all possible points with respect to Q_α; ii) it is shown that the upper bound of the decrease in Q_α can be used to derive a lower bound for the probability that the $K(\alpha,\zeta)$-th update at each agent is close to a local minimizer of f.

Step 1. By the supposition in Theorem 3.1, given $\Delta_1 > 0$ and $0 < \zeta < 1$, there exist $\epsilon > 0$, $0 < \gamma \leq L_F^g$, $0 < \mu \leq L_F^g$, $\delta > 0$, and $0 < \alpha \leq \hat{\alpha}(\Delta_1, \zeta)$ such that $\mathcal{L}^1_{\alpha,\epsilon} \cup \mathcal{L}^2_{\alpha,\gamma} \cup \mathcal{L}^3_{\alpha,\mu,\delta} = (\mathbb{R}^n)^m$, with respect to (8), and thus $\mathcal{L}^3_{\alpha,\mu,\delta} \supseteq (\mathcal{L}^1_{\alpha,\epsilon} \cup \mathcal{L}^2_{\alpha,\gamma})^c$, where the superscript c denote set complement. If $\hat{\mathbf{x}} \in \mathcal{L}^1_{\alpha,\epsilon}$, $\|\nabla Q_\alpha(\hat{\mathbf{x}})\| = \|\nabla F(\hat{\mathbf{x}}) + \alpha^{-1}(\mathbf{I}_{mn} - \hat{\mathbf{W}})\hat{\mathbf{x}}\| \geq \epsilon$; if $\hat{\mathbf{x}} \in \mathcal{L}^2_{\alpha,\gamma}$, then by Weyl's inequality, $\lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}})) = \lambda_{\min}(\nabla^2 F(\hat{\mathbf{x}}) + \alpha^{-1}(\mathbf{I}_{mn} - \hat{\mathbf{W}})) \leq -\gamma$; if $\hat{\mathbf{x}} \in \mathcal{L}^3_{\alpha,\mu,\delta}$, then again by Weyl's inequality, $\lambda_{\min}(\nabla^2 Q_\alpha(\hat{\mathbf{x}})) = \lambda_{\min}(\nabla^2 F(\hat{\mathbf{x}}) + \alpha^{-1}(\mathbf{I}_{mn} - \hat{\mathbf{W}})) \geq \mu$ and $\operatorname{dist}(\hat{\mathbf{x}}, \hat{\mathcal{X}}^*_{Q_\alpha}) \leq \delta$. Therefore, $\mathcal{L}^1_{\alpha,\epsilon} \subseteq \mathcal{I}^1_{\alpha,\epsilon}$, $\mathcal{L}^2_{\alpha,\gamma} \subseteq \mathcal{I}^2_{\alpha,\gamma}$, $\mathcal{L}^3_{\alpha,\mu,\delta} \subseteq \mathcal{I}^3_{\alpha,\mu,\delta}$, whereby $\mathcal{I}^1_{\alpha,\epsilon} \cup \mathcal{I}^2_{\alpha,\gamma} \cup \mathcal{I}^3_{\alpha,\mu,\delta} = (\mathbb{R}^n)^m$, $\mathcal{I}^3_{\alpha,\mu,\delta} \supseteq (\mathcal{I}^1_{\alpha,\epsilon} \cup \mathcal{I}^2_{\alpha,\gamma})^c$.

Step 2. Define stochastic process $\{\kappa_i\} \subset \mathbb{N}$ as

$$\kappa_i := \begin{cases} 0, & i = 0 \\ \kappa_{i-1} + 1, & \hat{\mathbf{x}}^{\kappa_{i-1}} \in \mathcal{I}^1_{\alpha,\epsilon} \cup \mathcal{I}^3_{\alpha,\mu,\delta} \\ \kappa_{i-1} + T(\hat{\mathbf{x}}^{\kappa_{i-1}}), & \hat{\mathbf{x}}^{\kappa_{i-1}} \in \mathcal{I}^2_{\alpha,\gamma} \end{cases}, \quad (10)$$

where $T(\hat{\mathbf{x}}) \leq T_{\max}(\alpha) = \tilde{\mathcal{O}}(\alpha^{-1})$ for all $\hat{\mathbf{x}}$ as per Lemma 4.1.2. By Lemma 4.1.1 and Lemma 4.1.2, Q_α decreases by a certain amount after a certain number of iterations for $\hat{\mathbf{x}} \in \mathcal{I}^1_{\alpha,\epsilon}$, and $\hat{\mathbf{x}} \in \mathcal{I}^2_{\alpha,\gamma}$, respectively, as follows

$$\begin{aligned} &\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}}) - Q_\alpha(\hat{\mathbf{x}}^{\kappa_i}) \mid \hat{\mathbf{x}}^{\kappa_i} \in \mathcal{I}^1_{\alpha,\epsilon}] \leq -l_1(\alpha), \\ &\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}}) - Q_\alpha(\hat{\mathbf{x}}^{\kappa_i}) \mid \hat{\mathbf{x}}^{\kappa_i} \in \mathcal{I}^2_{\alpha,\gamma}] \leq -l_2(\alpha), \end{aligned} \quad (11)$$

where $l_1(\alpha) = \Omega(\alpha)$ and $l_2(\alpha) = \Omega(\alpha)$ are defined in Lemma 4.1.1 and Lemma 4.1.2.

Defining event $\mathcal{E}_i := \{(\exists j \leq i)\ \hat{\mathbf{x}}^{\kappa_j} \in \mathcal{L}^3_{\alpha,\mu,\delta}\}$, by law of total expectation,

$$\begin{aligned} &\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}}) - Q_\alpha(\hat{\mathbf{x}}^{\kappa_i})] \\ &\quad = \mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}}) - Q_\alpha(\hat{\mathbf{x}}^{\kappa_i}) \mid \mathcal{E}_i] \cdot \mathbb{P}[\mathcal{E}_i] \\ &\qquad + \mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}}) - Q_\alpha(\hat{\mathbf{x}}^{\kappa_i}) \mid \mathcal{E}_i^c] \cdot \mathbb{P}[\mathcal{E}_i^c]. \end{aligned}$$

Combining (10) and (11) gives

$$\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}}) - Q_\alpha(\hat{\mathbf{x}}^{\kappa_i}) \mid \mathcal{E}_i^c] \leq -l(\alpha) \cdot \Delta\kappa_i,$$

where $l(\alpha) = \min\{l_1(\alpha), l_2(\alpha)/T_{\max}(\alpha)\} = \Omega(\alpha^2)$ and $\Delta\kappa_i = \kappa_{i+1} - \kappa_i$. Since $\mathbb{P}[\mathcal{E}_{i-1}] \leq \mathbb{P}[\mathcal{E}_i]$, we obtain

$$\begin{aligned} &\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}})] - \mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_i})] \\ &\quad \leq \mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_i}) \mid \mathcal{E}_i] \cdot (\mathbb{P}[\mathcal{E}_i] - \mathbb{P}[\mathcal{E}_{i-1}]) - l(\alpha) \cdot \Delta\kappa_i. \end{aligned}$$

Since the generated sequence $\{Q_\alpha(\hat{\mathbf{x}}^k)\}$ is assumed bounded, there exists $b > 0$ such that $\|Q_\alpha(\hat{\mathbf{x}}^k)\| \leq b$ for all $k = 0, 1, \cdots$. As such,

$$\begin{aligned} &\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_{i+1}})] - \mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_i})] \\ &\qquad \leq b \cdot (\mathbb{P}[\mathcal{E}_i] - \mathbb{P}[\mathcal{E}_{i-1}]) - l(\alpha) \cdot \Delta\kappa_i \cdot \mathbb{P}[\mathcal{E}_i^c]. \end{aligned}$$

Summing both sides of the inequality over i gives

$$\begin{aligned} &\mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_i})] - \mathbb{E}[Q_\alpha(\hat{\mathbf{x}}^{\kappa_1})] \\ &\quad \leq b \cdot (\mathbb{P}[\mathcal{E}_{i-1}] - \mathbb{P}[\mathcal{E}_0]) - l(\alpha) \cdot (\kappa_i - \kappa_1) \cdot \mathbb{P}[\mathcal{E}_i^c]. \end{aligned}$$

Since $\|Q_\alpha(\hat{\mathbf{x}}^k)\| \leq b$ for all $k = 0, 1, \cdots$, it follows that $-2b \leq b - l(\alpha) \cdot (\kappa_i - \kappa_1) \cdot \mathbb{P}[\mathcal{E}_i^c]$, which leads to

$$\mathbb{P}[\mathcal{E}_i^c] \leq \frac{3b}{l(\alpha)(\kappa_i - \kappa_1)}.$$

Therefore, if $\kappa_i - \kappa_1$ grows larger than $6b/l(\alpha)$, then $\mathbb{P}[\mathcal{E}_i^c] \leq 1/2$. Since $\kappa_1 \leq T_{\max}(\alpha) = \mathcal{O}(\alpha^{-1})$, after $K'(\alpha) = 6b/l(\alpha) + T_{\max}(\alpha) = \mathcal{O}(\alpha^{-2})$ steps, $\{\hat{\mathbf{x}}^k\}$ must enter $\mathcal{L}^3_{\alpha,\mu,\delta}$ at least once with probability at least $1/2$. Therefore, by repeating this step $\log \zeta^{-1}$ times, the probability of $\{\hat{\mathbf{x}}^k\}$ entering $\mathcal{L}^3_{\alpha,\mu,\delta}$ at least once is lower bounded:

$$\mathbb{P}[\{(\exists k \leq K(\alpha,\zeta))\ \hat{\mathbf{x}}^k \in \mathcal{L}^3_{\alpha,\mu,\delta}\}] \geq 1 - \frac{\zeta}{2},$$

where $K(\alpha,\zeta) = \mathcal{O}(\alpha^{-2} \log \zeta^{-1})$. Combining this with Lemma 4.1.3, we have that, after $K(\alpha,\zeta)$ iterations, Algorithm 1 produces a point that is $\Delta_2(\alpha,\zeta)$-close to $\hat{\mathcal{X}}^*_{Q_\alpha}$ with probability at least $1 - \zeta$, where $\Delta_2(\alpha,\zeta) = \mathcal{O}(\sqrt{\alpha \log(\alpha^{-1}\zeta^{-1})})$. For given $\Delta_1 > 0$, since $\alpha \leq \bar{\alpha}(\Delta_1)$ satisfies requirements of Theorem 2.1, $\hat{\mathbf{x}}^* = \inf_{\hat{\mathbf{x}} \in \mathcal{X}^*_{Q_\alpha}} \|\hat{\mathbf{x}}_i^{K(\alpha,\zeta)} - \hat{\mathbf{x}}\|$ is such that $\hat{\mathbf{x}}_i^*$ is Δ_1-close to $\mathcal{X}_f^*$. To summarize, we have for $i \in \mathcal{V}$,

$$\operatorname{dist}(\hat{\mathbf{x}}_i^{K(\alpha,\zeta)}, \mathcal{X}_f^*) \leq \Delta_1 + \Delta_2(\alpha,\zeta)$$

as claimed. ∎

V. Numerical Example

Consider the following non-convex optimization problem over $\mathbf{x} = (x_1, x_2)$:

$$\min_{\mathbf{x} \in \mathbb{R}^2} f(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^2} \sum_{i=1}^5 f_i(\mathbf{x}) = x_1^4 - x_1^2 + x_2^4 + x_2^2,$$

where $f_1(\mathbf{x}) = 0.25x_1^4 - x_1^2 - x_2^2$, $f_2(\mathbf{x}) = 0.25x_1^4 + 0.5x_2^4 + 1.5x_2^2$, $f_3(\mathbf{x}) = -x_1^2 + x_2^2$, $f_4(\mathbf{x}) = 0.5x_1^4 - 0.5x_2^2$, and $f_5(\mathbf{x}) = x_1^2 + 0.5x_2^4$. The mixing matrix is taken to be

$$\mathbf{W} = \begin{bmatrix} 0.6 & 0 & 0.2 & 0 & 0.2 \\ 0 & 0.6 & 0 & 0.2 & 0.2 \\ 0.2 & 0 & 0.6 & 0.2 & 0 \\ 0 & 0.2 & 0.2 & 0.6 & 0 \\ 0.2 & 0.2 & 0 & 0 & 0.6 \end{bmatrix}.$$

It can be verified that $\mathbf{x} = (0,0)$ is a saddle point of f, and that $\mathbf{x} = (-\frac{\sqrt{2}}{2}, 0)$ and $\mathbf{x} = (\frac{\sqrt{2}}{2}, 0)$ are two local minimizers. We compare the performance of **DGD** and

NDGD with constant step-size $\alpha = 0.005$, both initialized from $\mathbf{x}^0 = (10^{-6}, 10^{-6})$ (i.e., close to a saddle point). For **NDGD**, we generate the i.i.d noise from a sphere of radius 0.5 according to Remark 2.

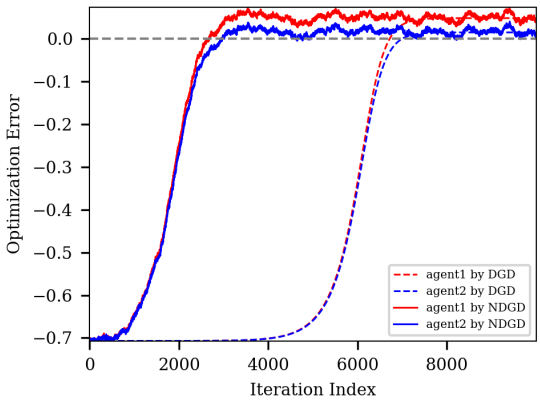

Fig. 1: Evolution of two agents' optimization errors by **DGD** and **NDGD**.

From Fig. 1, although not trapped forever, it does take **DGD** about 5000 iterations to escape the vicinity of the saddle point and converge to the neighborhood of a local minimizer. In contrast, **NDGD** escapes the vicinity of the saddle point after about 1000 iterations and converges to the neighborhood of a local minimizer. The effectiveness of **NDGD** over **DGD** is evident through this example.

VI. Conclusion

A fixed step-size noisy distributed gradient descent (**NDGD**) algorithm is formulated for solving optimization problems in which the objective is a finite sum of smooth but possibly non-convex functions. Random perturbations are added to the gradient descent at each step to actively evade saddle points. Under certain regularity conditions, and with a suitable step-size, each agent converges (in probability with specified confidence) to a neighborhood of a local minimizer. In particular, we determine a probabilistic upper bound on the distance between the iterate at each agent, and the set of local minimizers, after a sufficient number of iterations.

The potential applications of the **NDGD** algorithm are vast and varied, including multi-agent systems control, federated learning and sensor networks location estimation, particularly in large-scale network scenarios. Further exploration of different approaches to introducing random perturbations, and analysis of convergence rate performance can be pursued in future work.

References

[1] X. Dong, Y. Hua, Y. Zhou, Z. Ren, and Y. Zhong, "Theory and experiment on formation-containment control of multiple multirotor unmanned aerial vehicle systems," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 1, pp. 229–240, 2018.

[2] Z. Qiu, G. Deconinck, and R. Belmans, "A literature survey of optimal power flow problems in the electricity market context," in *2009 IEEE/PES Power Systems Conference and Exposition*. IEEE, 2009, pp. 1–6.

[3] W. Gao, J. Gao, K. Ozbay, and Z.-P. Jiang, "Reinforcement-learning-based cooperative adaptive cruise control of buses in the lincoln tunnel corridor with time-varying topology," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3796–3805, 2019.

[4] A. Nedić and J. Liu, "Distributed optimization for control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 77–103, 2018.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[6] H. Terelius, U. Topcu, and R. M. Murray, "Decentralized multi-agent optimization via dual decomposition," *IFAC proceedings volumes*, vol. 44, no. 1, pp. 11 245–11 251, 2011.

[7] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.

[8] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.

[9] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[10] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[11] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[12] K. I. Tsianos and M. G. Rabbat, "Distributed strongly convex optimization," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 593–600.

[13] 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.

[14] A. I. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 601–608.

[15] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[16] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[17] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on signal processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[18] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 3029–3068, 2020.

[19] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[20] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on learning theory*. PMLR, 2015, pp. 797–842.

[21] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1724–1732.

[22] B. Swenson, R. Murray, H. V. Poor, and S. Kar, "Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 14 751–14 812, 2022.

[23] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part i: Agreement at a linear rate," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1242–1256, 2021.

[24] ——, "Distributed learning in non-convex environments—part ii: Polynomial escape from saddle-points," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1257–1270, 2021.

[25] L. Qin, M. Cantoni, and Y. Pu, "Second-order properties of noisy distributed gradient descent," *arXiv preprint arXiv:2303.17165*, 2023.