The Geometric Effects of Distributing Constrained Nonconvex Optimization Problems

Qiuwei Li and Xinshuo Yang
Colorado School of Mines
{qiuli, xinshuoyang}@mines.edu

Zhihui Zhu
Johns Hopkins University
zzhu29@jhu.edu

Gongguo Tang and Michael B. Wakin
Colorado School of Mines
{gtang, mwakin}@mines.edu

***Abstract*—A variety of nonconvex machine learning problems have recently been shown to have benign geometric landscapes, in which there are no spurious local minima and all saddle points are strict saddles at which the Hessian has at least one negative eigenvalue. For such problems, a variety of algorithms can converge to global minimizers. We present a general result relating the geometry of a centralized problem to its distributed extension; our result is new in considering the scenario where the centralized problem obeys a manifold constraint such as when the variables are normalized to the sphere. We show that the first/second-order stationary points of the centralized and distributed problems are one-to-one correspondent, implying that the distributed problem—in spite of its additional variables and constraints—can inherit the benign geometry of its centralized counterpart. We apply this result to show that the distributed matrix eigenvalue problem, multichannel blind deconvolution problem, and dictionary learning problem all enjoy benign geometric landscapes.**

I. Introduction

A variety of nonconvex machine learning problems have recently been shown to have benign geometric landscapes, in which there are no spurious local minima and all saddle points are strict saddles at which the Hessian has at least one negative eigenvalue [1]–[11]. For such problems a variety of iterative algorithms—such as gradient descent with a random initialization—can exploit negative curvature directions to escape from strict saddle points and thus provably converge to a global minimizer [12].

In some scenarios, the size of a machine learning problem demands that computations or storage be *distributed* across some network [13], [14]. Distributing an optimization problem typically involves creating extra variables—local to each node in the network—and using algorithms, regularizers, or constraints that encourage consensus among these variables. Unfortunately, these changes from the original centralized problem have the potential to affect its geometric landscape.

Example 1. *As a simple example, consider the centralized problem of the form*

$$\operatorname*{minimize}_{x} c(x) \doteq \frac{1}{2}(x^2-1)^2,$$

which has three critical points: $x=\pm 1$ are global minima and $x=0$ is a local maxima. Now let us consider two variants of distributed formulations. The first involves two variables and an objective function that coincides with $c(x)$ when these two variables obey a consensus constraint:

$$\operatorname*{minimize}_{x_1,x_2} d(x_1,x_2) \doteq c(x_1)+c(x_2) \quad \text{s.t.} \quad x_1=x_2.$$

The second involves regularizing the sum of objective functions with a quadratic penalty that encourages consensus: for some $\lambda>0$,

$$\operatorname*{minimize}_{x_1,x_2} d_\lambda(x_1,x_2) \doteq c(x_1)+c(x_2)+\lambda(x_1-x_2)^2.$$

To see how the geometric landscapes of these two distributed variants relate to the original geometric landscape, we plot the centralized landscape (i.e., $c(x)$), the constrained distributed landscape (i.e., $d(x_1,x_2)$ constrained on $x_1=x_2$), and the quadratic regularized landscape (i.e., $d_\lambda(x_1,x_2)$ in Figure 1. We see that the constrained distributed formulation (see (b)) preserves the centralized geometric landscape (see (a)) in the consensus space *where $x_1=x_2$, while the quadratic regularized formulation (see (c)) does not preserve the original geometric landscape, as spurious local minima are introduced at* non-consensus *points, i.e., where $x_1\neq x_2$. Note, however, that since $d_\lambda(x_1,x_2)=d(x_1,x_2)=c(x)$ on the consensus space, $d_\lambda(x_1,x_2)$ still shares the same geometric landscape as $c(x)$ or $d(x_1,x_2)$ when $x_1=x_2=x$.*

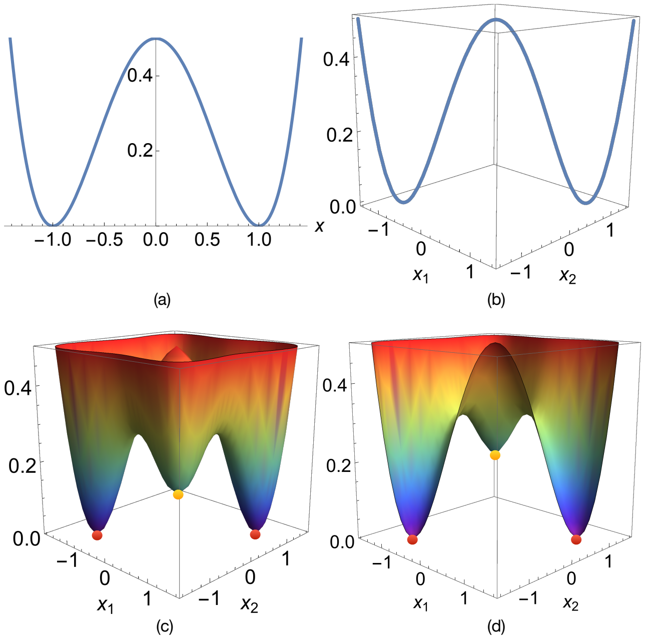

Fig. 1. (a) The landscape of $c(x)$; (b) The landscape of $d(x_1,x_2)$ constrained on $x_1=x_2$; (c) The landscape of $d_\lambda(x_1,x_2)$ for $\lambda=0.1$; (d) The sliced view on the plane $x_1=x_2$ of the landscape in (c). For convenience, we marked the global minima of $d(x_1,x_2)$ using the red balls and the local minima using the orange balls.

978-1-7281-5549-4/19/$31.00 ©2019 IEEE

The preceding example demonstrates that using a quadratic penalty for consensus can result in the introduction of spurious local minima that are not present in the original centralized problem. We have recently shown, however, that in certain problems such quadratic penalties do not affect the ability to converge to globally optimal consensus points [15].

We have also recently proved that when equality constraints are used to enforce consensus, there is a certain equivalence between the geometry of the distributed problem and its centralized counterpart [16]. In particular, we established one-to-one correspondences of the first-order critical points, second-order critical points, and strict saddle points between the two problems. This is in spite of the fact that critical points have a distinctly different definition (in terms of the Lagrangian) for the constrained distributed problem.

However, our previous result [16] applied only to the case where the centralized problem was originally unconstrained. In this paper, we present an analogous result for settings where the centralized problem obeys a manifold constraint such as when the variables are normalized to the sphere. We discuss several applications covered by this result and present numerical experiments on a distributed dictionary learning problem.

II. Main Result

Consider a centralized optimization problem of the form

$$\mathcal{P}_C : \underset{\mathbf{x}}{\text{minimize}}\, c(\mathbf{x}) \quad \text{s.t.} \quad \begin{cases} q_i(\mathbf{x}) = 0, & \forall i \in \mathcal{E} \\ q_i(\mathbf{x}) \geq 0, & \forall i \in \mathcal{I}, \end{cases}$$

where the objective $c(\mathbf{x})$ and constraint functions q_i are twice differentiable. Example equality constraints (such as in dictionary learning—see Section III) include $\|\mathbf{x}\|_2^2 - 1 = 0$. As further discussed in Section III, such problems may have a benign geometry, where all second-order stationary points correspond to global minima, and a variety of algorithms have been proposed that can converge to global minimizers of problems with such a benign geometry [12], [17], [18].

Second-order stationary points are defined as follows.

Definition II.1 ([19])**.** *Denote the Lagrange function of the constrained program $\mathcal{P}_C$ as*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\lambda}) = c(\mathbf{x}) - \sum_{i \in \mathcal{E}} \boldsymbol{\nu}(i) q_i(\mathbf{x}) - \sum_{i \in \mathcal{I}} \boldsymbol{\lambda}(i) q_i(\mathbf{x}).$$

Assume that LICQ holds at all feasible points.[1,2] *Then:*

1) $\mathbf{x}^\star$ *is a first-order stationary point if there exist* $\boldsymbol{\nu}^\star, \boldsymbol{\lambda}^\star$ *such that*

$$\begin{cases} \nabla c(\mathbf{x}^\star) - \sum_{i \in \mathcal{E}} \boldsymbol{\nu}^\star(i) \nabla q_i(\mathbf{x}^\star) \\ \quad - \sum_{i \in \mathcal{I}} \boldsymbol{\lambda}^\star(i) \nabla q_i(\mathbf{x}^\star) = \mathbf{0} \\ q_i(\mathbf{x}^\star) = 0, \ \forall i \in \mathcal{E} \\ q_i(\mathbf{x}^\star) \geq 0, \ \forall i \in \mathcal{I} \\ \boldsymbol{\lambda}^\star \geq 0 \\ \boldsymbol{\lambda}^\star(i) \cdot q_i(\mathbf{x}^\star) = 0, \ \forall i \in \mathcal{I} \end{cases}$$

2) $\mathbf{x}^\star$ *is a second-order stationary point of* $\mathcal{P}_C$ *if it is a first-order stationary point and satisfies*

$$\mathbf{d}^\top \nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}^\star, \boldsymbol{\nu}^\star, \boldsymbol{\lambda}^\star) \mathbf{d} \geq 0,$$
$$\forall \begin{cases} \nabla q_i(\mathbf{x}^\star)^\top \mathbf{d} = 0, & \forall i \in \mathcal{E} \\ \nabla q_i(\mathbf{x}^\star)^\top \mathbf{d} = 0, & \forall i \in \mathcal{A}(\mathbf{x}^\star) \cap \mathcal{I}, \lambda_i^\star > 0 \\ \nabla q_i(\mathbf{x}^\star)^\top \mathbf{d} \geq 0, & \forall i \in \mathcal{A}(\mathbf{x}^\star) \cap \mathcal{I}, \lambda_i^\star = 0 \end{cases}$$

where $\mathcal{A}(\mathbf{x}^\star) := \{i : q_i(\mathbf{x}^\star) = 0\}$ *denotes the set of active constraints.*

To accommodate distributed formulations of such problems, we define the following extension:

$$\mathcal{P}_D : \underset{\mathbf{x},\mathbf{y}}{\text{minimize}}\, d(\mathbf{x}, \mathbf{y}) \quad \text{s.t.} \quad \begin{cases} \mathbf{Ax} + \mathbf{By} = \mathbf{b} \\ q_i(\mathbf{x}) = 0, & \forall i \in \mathcal{E} \\ q_i(\mathbf{x}) \geq 0, & \forall i \in \mathcal{I}, \end{cases}$$

where $d(\mathbf{x}, \mathbf{y})$ is twice differentiable and satisfies $d(\mathbf{x}, \mathbf{y}) = c(\mathbf{x})$ when $\mathbf{Ax} + \mathbf{By} = \mathbf{b}$, and $\mathbf{B}$ is a square and invertible matrix. Such a problem formulation is quite general but accommodates in particular the following scenario. Suppose the centralized objective function $c(\mathbf{x})$ decouples into a sum of J terms:

$$c(\mathbf{x}) = \sum_{j=1}^{J} c_j(\mathbf{x}).$$

Each of these objective functions might be available only at some node j in a network, and so we wish to distribute the optimization problem, defining local variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_J$ at nodes $1, 2, \dots, J$, respectively, and minimizing $\sum_{j=1}^{J} c_j(\mathbf{x}_j)$ subject to the constraints $\mathbf{x} = \mathbf{x}_j$ for $j = 1, 2, \dots, J$. In such a scenario, $\mathbf{x}$ might be handled by a central node with a star topology connectivity to the other nodes.[3] We see that such a distributed problem is an instantiation of $\mathcal{P}_D$ by defining

$$\mathbf{y} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_J \end{bmatrix}, \mathbf{A} = \begin{bmatrix} -\mathbf{I} \\ \vdots \\ -\mathbf{I} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \quad (1)$$

and $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{J} c_j(\mathbf{x}_j)$. We note also that Problem $\mathcal{P}_D$ inherits all constraints q_i from Problem $\mathcal{P}_C$, imposing them on $\mathbf{x}$, and by proxy, all $\mathbf{x}_j$.

Our main result establishes a geometric equivalence between the landscapes of problem $\mathcal{P}_C$ and problem $\mathcal{P}_D$. We omit the proof of this theorem due to space limitations.

Theorem II.1. *Assume that LICQ holds for all feasible points of* $\mathcal{P}_C$*. Then* $\mathbf{x}^\star$ *is a first/second-order stationary point of Problem* $\mathcal{P}_C$ *iff* $(\mathbf{x}^\star, \mathbf{B}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{x}^\star))$ *is a first/second-order stationary point of Problem* $\mathcal{P}_D$*.*

If LICQ holds at a point $\mathbf{x}$ for Problem $\mathcal{P}_C$, then LICQ also holds at any $(\mathbf{x}, \mathbf{y})$ (such that $\mathbf{Ax}+\mathbf{By} = \mathbf{b}$) for Problem $\mathcal{P}_D$ (see, e.g., (2)). Thus, LICQ holds for all feasible points of $\mathcal{P}_D$ since it holds for all feasible points of $\mathcal{P}_C$. With this, the second-order stationary points of Problem $\mathcal{P}_D$ are defined

[1] We say that LICQ is satisfied if the gradients of equality/active inequality constraints are linearly independent.

[2] Generally LICQ is required to hold at $\mathbf{x}^\star$, but we assume LICQ holds for all feasible points for convenience.

[3] As we will describe in a future paper, other constraint topologies can also be accommodated by relaxing the requirement that $\mathbf{B}$ is a square matrix.

analogously to those of Problem $\mathcal{P}_C$ (see Definition II.1), accounting for the additional equality constraints $\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}=\mathbf{b}$.

Theorem II.1 implies that once the centralized problem $\mathcal{P}_C$ has a benign geometry, this benign geometry will be inherited by the distributed formulation Problem $\mathcal{P}_D$. In particular, any second-order stationary point of $\mathcal{P}_D$ is a global minimum if $\mathcal{P}_C$ has no spurious local minima and obeys the strict saddle property. When the distributed problem has the form of (1), every second-order stationary point of Problem $\mathcal{P}_D$ will have the form $(\mathbf{x}^\star, \mathbf{y}^\star)$, where $\mathbf{y}^\star = \begin{bmatrix}\mathbf{x}^{\star \mathrm{T}} & \cdots & \mathbf{x}^{\star \mathrm{T}}\end{bmatrix}^{\mathrm{T}}$ with $\mathbf{x}^\star$ beging a global minimizer of $\mathcal{P}_C$. We expand on such scenarios in the next section.

III. Applications and Demonstrations

A. Distributed symmetric matrix eigenvalue problem

For a symmetric matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ with a unique smallest eigenvalue $\lambda_{\min}$, consider the following centralized problem

$$\mathcal{P}_C : \underset{\mathbf{x}}{\operatorname{minimize}}\, \mathbf{x}^\top \mathbf{Q}\mathbf{x} \;\text{ s. t. }\; \|\mathbf{x}\|_2^2 = 1.$$

Theorem III.1 ([20, Section 4.6.2])**.** *The eigenvector $\mathbf{x}^\star$ corresponding to $\lambda_{\min}$ is the only second-order stationary point of $\mathcal{P}_C$.*

In the case that $\mathbf{Q} = \sum_{i=1}^J \mathbf{Q}_j$ which arises in applications like covariance estimation where $\mathbf{Q}_j = \mathbf{z}_j\mathbf{z}_j^\top$, we consider the distributed formulation as

$$\mathcal{P}_D : \underset{\mathbf{x},\mathbf{y}}{\operatorname{minimize}} \sum_{j=1}^J \mathbf{x}_j \mathbf{Q}_j \mathbf{x}_j \quad \text{s.t.} \quad \begin{cases}\|\mathbf{x}\|_2^2 = 1\\ \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}=\mathbf{b},\end{cases}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{y}, \mathbf{b}$ are defined in (1) and $\mathbf{Q} = \sum_{j=1}^J \mathbf{Q}_j$.[4]

Corollary III.1. *$(\mathbf{x}^\star, \cdots, \mathbf{x}^\star)$ is the only second-order stationary point of $\mathcal{P}_D$ on the consensus space.*

Proof: First of all, we verify the LICQ condition in the feasible region of $\mathcal{P}_D$. Towards this end, we build up the following matrix with the columns as the gradients of the active equality/inequality constraints of problem $\mathcal{P}_D$:

$$\begin{bmatrix}\mathbf{A}^\top & 2\mathbf{x}\\ \mathbf{B}^\top & \mathbf{0}\end{bmatrix} = \begin{bmatrix}-\mathbf{I} & \cdots & -\mathbf{I} & 2\mathbf{x}\\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0} & \vdots\\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0}\end{bmatrix}, \tag{2}$$

which is of full column-rank since $\mathbf{x} \neq \mathbf{0}$. Therefore, the LICQ condition is satisfied in the feasible region. Then, the proof follows by combining Theorem II.1 and Theorem III.1. ■

B. Multichannel blind deconvolution

Consider the multichannel blind deconvolution model as

$$\mathbf{y}_j = \mathbf{s}_j^\star \circledast \mathbf{x}^\star = \mathbf{C}_{\mathbf{s}_j}\mathbf{x}^\star, \quad j = 1, \ldots, J,$$

where $\mathbf{y}_j \in \mathbb{R}^n$ are measurements, $\mathbf{s}_j^\star \in \mathbb{R}^n$ are unknown target channels of sparsity θ, $\mathbf{x}^\star \in \mathbb{R}^n$ is an unknown target signal, and $\mathbf{C}_\mathbf{a}$ represents the circulant matrix whose first column is $\mathbf{a}$. Consider the optimization problem

$$\mathcal{P}_C : \underset{\mathbf{x}}{\operatorname{minimize}} - \sum_{j=1}^J \|\mathbf{C}_{\mathbf{y}_j}\mathbf{R}\mathbf{x}\|_4^4 \quad \text{s.t.} \quad \|\mathbf{x}\|_2 = 1$$

with $\mathbf{R} \doteq \left(\frac{1}{\theta n J}\sum_{k=1}^J \mathbf{C}_{\mathbf{y}_k}^\top \mathbf{C}_{\mathbf{y}_k}\right)^{-1/2}$.

Theorem III.2 (Informal, [18])**.** *Under certain conditions on the number of measurements m, sparsity θ, and incoherence of $\mathbf{x}^\star$ (see [18] for the details), any second-order stationary point of $\mathcal{P}_C$ is close to a sign shifted version of $\mathbf{x}^\star$.*

Consider its distributed formulation as

$$\mathcal{P}_D : \underset{\mathbf{x},\mathbf{y}}{\operatorname{minimize}} - \sum_{i=1}^J \|\mathbf{C}_{\mathbf{y}_j}\mathbf{R}\mathbf{x}_j\|_4^4 \quad \text{s.t.} \quad \begin{cases}\|\mathbf{x}\|_2^2 = 1\\ \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}=\mathbf{b},\end{cases}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{y}, \mathbf{b}$ are defined in (1).

Corollary III.2. *Under the same assumption as in Theorem III.2, if $(\mathbf{x}, \cdots, \mathbf{x})$ is a second-order stationary point of $\mathcal{P}_D$ on the consensus space, then $\mathbf{x}$ is close to a sign shifted version of $\mathbf{x}^\star$.*

C. Distributed dictionary learning

Given m signal samples from $\mathbb{R}^n$, i.e. a data matrix $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_p] \in \mathbb{R}^{n\times m}$, the goal of the complete dictionary learning problem is to seek an orthogonal matrix $\mathbf{X}^\star = [\mathbf{x}_1^\star \;\cdots\; \mathbf{x}_n^\star] \in \mathcal{O}_n$ such that $\mathbf{Y} \approx \mathbf{X}^\star\mathbf{S}^\star$ and the coefficient matrix $\mathbf{S}^\star$ is sparse. Inspired by [21], we formulate the dictionary learning problem as a (centralized) nonconvex smooth optimization problem on the sphere:

$$\mathcal{P}_C : \underset{\mathbf{x}}{\operatorname{minimize}} \frac{1}{m} c\left(\mathbf{x}; \mathbf{Y}\right) \doteq \frac{1}{m}\sum_{k=1}^m h_\mu\left(\mathbf{x}^T\mathbf{y}_k\right) \;\text{ s.t. }\; \|\mathbf{x}\|_2^2 = 1,$$

where $h_\mu(z) = \mu \log\left(\cosh\left(z/\mu\right)\right)$ is a convex smooth approximation to $|\cdot|$ and μ is a smoothing parameter.

Theorem III.3 (Informal, [22])**.** *Under certain conditions on the number of samples m and sparsity of $\mathbf{S}^\star$(see [22] for the details), any second-order stationary point of $\mathcal{P}_C$ is close to $\{\pm\mathbf{x}_i^\star\}_{i=1}^n$.*

Without loss of generality, we partition the columns of $\mathbf{Y}$ as

$$\mathbf{Y} = \left[\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_J\right],$$

where for $j \in \{1, 2, \ldots, J\}$, matrix $\mathbf{Y}_j$ (which is stored at node j) has size $n \times m_j$ with $m = \sum_{j=1}^J m_j$. We then obtain the following distributed problem

$$\mathcal{P}_D : \underset{\mathbf{x},\mathbf{y}}{\operatorname{minimize}} \frac{1}{m}\sum_{j=1}^J c(\mathbf{x}_j; \mathbf{Y}_j) \quad \text{s.t.} \quad \begin{cases}\|\mathbf{x}\|_2^2 = 1,\\ \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{y}=\mathbf{b},\end{cases} \tag{3}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{y}, \mathbf{b}$ are defined in (1).

Corollary III.3. *Under the same assumption as in Theorem III.3, if $(\mathbf{x}, \cdots, \mathbf{x})$ is a second-order stationary point of $\mathcal{P}_D$ on the consensus space, then $\mathbf{x}$ is close to $\{\pm\mathbf{x}_i^\star\}_{i=1}^n$.*

[4] The data matrix is often formed as sum of matrices, e.g., correlation matrix.

As a demonstration, we solve (3) using a distributed Riemannian Gradient Descent (DRGD) with the update

$$\begin{aligned}\mathbf{x}_j(k+1) =& \mathcal{P}_{\mathbb{S}^{n-1}}\left(\mathbf{x}_j(k) + \mathcal{P}_{\tau(\mathbf{x}_j(k))}\left(\sum_{i\neq j}\omega_{ji}\mathbf{x}_i(k)\right)\right. \\ &\left. -\eta(k)\,\mathcal{P}_{\tau(\mathbf{x}_j(k))}\left(\frac{1}{m}\nabla c\left(\mathbf{x}_j(k);\mathbf{Y}_j\right)\right)\right) \quad (4)\end{aligned}$$

where $\eta(k)$ is the stepsize, $\tau(\mathbf{x}_j(k))$ denotes the tangent space of $\mathbb{S}^{n-1}$ at $\mathbf{x}_j(k)$, $\mathcal{P}_{\tau(\mathbf{x}_j(k))} = I - \mathbf{x}_j(k)(\mathbf{x}_j(k))^T$ and $\mathcal{P}_{\mathbb{S}^{n-1}}$ are the orthogonal projectors onto $\tau(\mathbf{x}_j(k))$ and $\mathbb{S}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ respectively, and ω_{ji} are a set of symmetric nonnegative weights, and ω_{ji} is positive if and only if nodes i and j are neighbors in the network. We make the common assumption that $\sum_{i=1}^{J}\omega_{ji} = 1 \quad \forall j \in [J]$. We note that DRGD update (4) with a constant stepsize $\eta(k) = \eta$ is equivalent to running Riemannian Gradient Descent (RGD) with stepsize η on the following centralized problem:

$$\begin{aligned}&\underset{\mathbf{x}_1,\cdots,\mathbf{x}_J}{\text{minimize}}\ \frac{1}{m}\sum_{j=1}^{J} c(\mathbf{x}_j;\mathbf{Y}_j) + \sum_{i,j=1}^{J}\frac{\omega_{ji}}{4\eta}\|\mathbf{x}_j - \mathbf{x}_i\|_2^2 \\ &\text{s.t. } \mathbf{x}_j \in \mathbb{S}^{n-1},\ \forall j \in [J].\end{aligned}$$

In our experiment, we set the true dictionary $\mathbf{X}^\star$ to be the identity matrix so that $\mathbf{Y} = \mathbf{S}^\star$, and the goal is to find the standard basis vectors $\{\pm\mathbf{e}_1, \pm\mathbf{e}_2, \ldots, \pm\mathbf{e}_n\}$. The case of a general orthogonal $\mathbf{X}^\star$ can be reduced to the identity case via an orthogonal rotation. The entries of the coefficient matrix $\mathbf{S}^\star$ are sampled from the i.i.d. Bernoulli-Gaussian distribution with parameter $\theta = 0.3$: each entry $s^\star_{ij}$ is independently drawn from a standard Gaussian with probability θ and is zero otherwise. We select $J = 5$, $n \in \{5, 10, 15, 20\}$ and for each n we choose $m = 10n^p$ where $p \in \{1, 1.5, 2, 2.5\}$. For each pair of (m, n), we generate 10 problem instances, corresponding to re-sampling the coefficient matrix 10 times. To solve the dictionary learning problem, we use the DRGD update (4) with a diminishing stepsize $\eta(k) = 1/\sqrt{k}$. For each problem instance, we perform $5n\log(n)$ runs with independent initial guesses $\mathbf{x}_1(0) = \cdots = \mathbf{x}_J(0)$ on $\mathbb{S}^{n-1}$. We consider a column of the true dictionary $\mathbf{x}^\star$ to be recovered at step k if $\|\mathbf{x}^\star - \bar{\mathbf{x}}(k)\| < 10^{-1}$, where $\bar{\mathbf{x}}(k) = \sum_{j=1}^{J}\mathbf{x}_j(k) \Big/ \left\|\sum_{j=1}^{J}\mathbf{x}_j(k)\right\|$. In each instance, we claim that the dictionary $\mathbf{X}^\star$ is successfully recovered if all columns of the dictionary $\mathbf{X}^\star$ are recovered within the $5n\log(n)$ runs. We plot the empirical success rate of recovering the true dictionary $\mathbf{X}^\star$ over the 10 instances in Figure 2. We observe that DRGD can successfully recover all the basis vectors when $p \geq 2$. In Figure 3, we perform a single run (to recover one column of $\mathbf{X}^\star$) for all $n \in \{5, 10, 15, 20\}$, $m = 10n^2$ and $J = 5$ and report the consensus error $\max_{j=1,\cdots,J}\|\mathbf{x}_j(k) - \bar{\mathbf{x}}(k)\|$ as a function of k. As can be seen, the consensus error gradually decreases.

These experiments demonstrate that the DRGD algorithm can find globally optimal solutions with near consensus on the distributed dictionary learning problem. We attribute this to the benign geometry that the distributed problem inherits from its centralized counterpart. A theoretical analysis of the DRGD performance is underway.

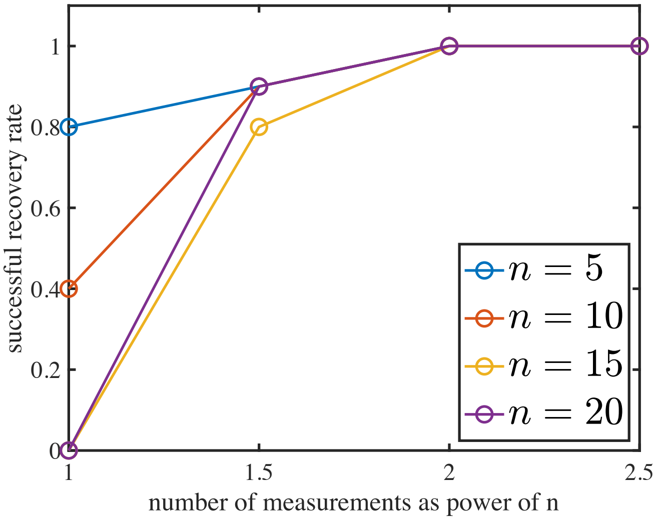

Fig. 2. Empirical success rate of identity matrix recovery of the distributed Riemannian gradient descent with $5n\log n$ runs averaged over 10 instances

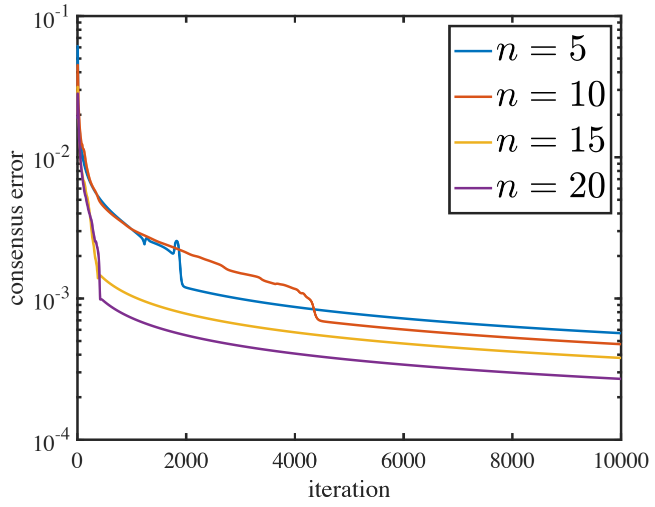

Fig. 3. Consensus error $\max_{j=1,\cdots,J}\|\mathbf{x}_j(k) - \bar{\mathbf{x}}(k)\|$ as a function of k for a single run with random initialization.

IV. Conclusion

Nonconvex optimization problems can be distributed while preserving both their constraints and their benign geometric landscape. Algorithms with provable convergence to second-order stationary points of problem $\mathcal{P}_D$ are an open question, but the empirical results using DRGD for distributed dictionary learning are encouraging.

Acknowledgment

This work was supported by the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020.

References

[1] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *arXiv preprint arXiv:1809.09573*, 2018.

[2] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14–31, July 2018.

[3] P. Jain and P. Kar, *Non-Convex Optimization for Machine Learning*, ser. Foundations and Trends in Machine Learning Series. Now Publishers, 2017. [Online]. Available: https://books.google.com/books?id=buuLtAEACAAJ

[4] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.

[5] ——, "The global optimization geometry of low-rank matrix optimization," *arXiv preprint arXiv:1703.01256*, 2017.

[6] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 51–96, 2018.

[7] X. Li, Z. Zhu, A. M.-C. So, and R. Vidal, "Nonconvex robust low-rank matrix recovery," *arXiv preprint arXiv:1809.09237*, 2018.

[8] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *International Conference on Machine Learning*, 2017, pp. 1233–1242.

[9] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[10] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[11] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao, "Symmetry, saddle points, and global geometry of nonconvex matrix factorization," *arXiv preprint arXiv:1612.09296*, 2016.

[12] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid saddle points," *arXiv preprint arXiv:1710.07406*, 2017.

[13] L. Gu, D. Zeng, P. Li, and S. Guo, "Cost minimization for big data processing in geo-distributed data centers," *IEEE transactions on Emerging topics in Computing*, vol. 2, no. 3, pp. 314–323, 2014.

[14] G. Scutari, F. Facchinei, L. Lampariello, S. Sardellitti, and P. Song, "Parallel and distributed methods for constrained nonconvex optimization-part II: Applications in communications and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1945–1960, 2017.

[15] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in distributed low-rank matrix factorization," *arXiv preprint arXiv:1811.03129*, 2018.

[16] Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, "The geometry of equality-constrained global consensus problems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7928–7932.

[17] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 885–914, 2016.

[18] Y. Li and Y. Bresler, "Global geometry of multichannel sparse blind deconvolution on the sphere," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 1132–1143.

[19] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[20] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[21] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?" *arxiv:1510.06096*, 2015. [Online]. Available: http://arxiv.org/abs/1510.06096

[22] ——, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2016.