THE GEOMETRY OF EQUALITY-CONSTRAINED GLOBAL CONSENSUS PROBLEMS

Qiuwei Li, Zhihui Zhu, Gongguo Tang, and Michael B. Wakin

Department of Electrical Engineering, Colorado School of Mines, Golden, CO, USA

ABSTRACT

A variety of unconstrained nonconvex optimization problems have been shown to have benign geometric landscapes that satisfy the strict saddle property and have no spurious local minima. We present a general result relating the geometry of an unconstrained centralized problem to its equality-constrained distributed extension. It follows that many global consensus problems inherit the benign geometry of their original centralized counterpart. Taking advantage of this fact, we demonstrate the favorable performance of the Gradient ADMM algorithm on a distributed low-rank matrix approximation problem.

***Index Terms*—** Constrained nonconvex optimization, global consensus, low-rank matrix approximation

1. INTRODUCTION

With an abundance of data, the scale of machine learning problems continues to grow. Consequently, nonconvex optimization problems have received growing attention as alternatives to convex approaches for solving machine learning problems [1–4]. Algorithms for solving nonconvex problems can offer reduced memory usage and computational complexity compared to their convex counterparts, see, e.g. [5, 6]. However, the potential for undesirable features in the nonconvex landscape (spurious local minima [7–9], degenerate saddle points [9, 10], etc.) raises questions about these algorithms' convergence to optimal points.

Recent research has shown, though, that many machine learning problems—including a variety of low-rank matrix optimization problems—actually have a benign nonconvex landscape in which there are no spurious local minima and all saddle points are strict (non-degenerate) saddles at which the Hessian has at least one negative eigenvalue [2–4, 11–18]. 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 [19].

This work was supported by the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020. The authors gratefully acknowledge Waheed Bajwa, Haroon Raja, Clement Royer, and Stephen Wright for helpful discussions.

To date, however, most of the results establishing benign geometric landscapes have been limited to *unconstrained* nonconvex problems [11–18, 20]. Meanwhile, constraints can be important to consider, particularly when the size of a machine learning problem demands that computations or storage be *distributed* across some network [21, 22]. One way to ensure consensus among optimization variables in a distributed problem is via equality constraints across the network nodes. As one transitions from a centralized problem to a distributed one, a question arises of whether the distributed problem inherits the benign geometry of the centralized problem. Since there is a general lack of geometric analysis for constrained nonconvex problems, this question is essentially open.

In Section 2, we present a general result relating the geometry of a centralized problem to its distributed extension. This result establishes 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 constrained optimization problems. In Section 3, we highlight one application of this theorem, in establishing an equivalence between geometric landscapes for broad classes of centralized problems and their distributed formulations as global consensus problems. We show that under certain conditions, every second-order critical point of the distributed problem corresponds to a global minimizer of the centralized problem. In Section 4, we discuss algorithmic aspects for solving equality-constrained distributed optimization problems. The recent GADMM algorithm [23] can be guaranteed under certain conditions to converge to a second-order critical point of an equality-constrained distributed optimization problem. Our theory establishes conditions under which this point will correspond to a global minimizer of the original centralized optimization problem. This guarantee is stronger than what appear in the literature for distributed gradient descent (DGD), a popular alternative algorithm for solving consensus problems. Existing DGD results show convergence either to stationary points (which are global if the problem is convex) [24–26], or to an arbitrarily small neighborhood of a second-order critical point with an appropriately small stepsize [27]. As a case study, in Section 5, we illustrate the performance of GADMM on a distributed low-rank matrix approximation in factored form.

2. RELATING UNCONSTRAINED GEOMETRY TO CONSTRAINED GEOMETRY

We present a general theorem that establishes an equivalence between the landscape of two types of optimization problems: one that is unconstrained, and one that involves additional variables but is constrained to an affine subspace, along which it has a certain equivalence to the first problem.

Theorem 1. *Consider two problems:*

- *Problem UC (unconstrained centralized):*

$$\min_{\boldsymbol{x}} c(\boldsymbol{x})$$

- *Problem ECD (equality-constrained distributed):*

$$\min_{\boldsymbol{x},\boldsymbol{y}} d(\boldsymbol{x},\boldsymbol{y}) \text{ subject to } \boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{y}=\boldsymbol{b}$$

where $d(\boldsymbol{x},\boldsymbol{y})$ satisfies $d(\boldsymbol{x},\boldsymbol{y}) = c(\boldsymbol{x})$ when $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{y}=\boldsymbol{b}$, and $\boldsymbol{B}$ is a square and invertible matrix.

Then $\boldsymbol{x}$ is a [first-order/second-order/strict saddle] critical point of Problem UC iff $(\boldsymbol{x}, \boldsymbol{B}^{-1}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}))$ is a [first-order/second-order/strict saddle] point of Problem ECD.

Theorem 1 is proved in Appendix A, where the precise notions of [first-order/second-order/strict saddle] point are defined for both Problem UC and Problem ECD. Critical points of Problem ECD are defined in terms of the Lagrangian function for $d(\boldsymbol{x},\boldsymbol{y})$. This theorem has applications outside of distributed optimization, but we adopt the terminology "centralized" and "distributed" in the theorem above because the latter problem involves additional optimization variables beyond those in the first, and we focus on applications in distributed optimization in this paper.

3. GEOMETRY OF GLOBAL CONSENSUS

Consider any unconstrained centralized optimization problem of the form

$$\underset{\boldsymbol{w},\{\boldsymbol{z}_j\}}{\text{minimize}} \left(\sum_{j=1}^{J} f_j(\boldsymbol{w},\boldsymbol{z}_j)\right) + g(\boldsymbol{w}), \tag{1}$$

where first term in the objective function decouples into a sum of objectives f_j. One can distribute this problem across a network of $J+1$ nodes in a "star topology",[1] where J agents are connected to a central node. The resulting problem is known as a *global consensus problem* (see [23, (3)]) and can be posed as follows[2]:

$$\underset{\boldsymbol{w},\{\boldsymbol{z}_j\},\{\boldsymbol{w}^j\}}{\text{minimize}} \left(\sum_{j=1}^{J} f_j(\boldsymbol{w}^j,\boldsymbol{z}_j)\right) + g(\boldsymbol{w}) \text{ s.t. } \boldsymbol{w}^j=\boldsymbol{w}\ \forall j. \tag{2}$$

[1] We remark that our results can also be applied to other network topologies, such as the series topology where $\boldsymbol{w}^j=\boldsymbol{w}^{j+1}$, $\forall j$ and $\boldsymbol{w}^J=\boldsymbol{w}$.

[2] Strictly speaking, our problem (2) is more general than [23, (3)] as (2) involves local variables $\{\boldsymbol{z}_j\}$ which are not constrained to be equal.

Here, $\boldsymbol{w}$ is the optimization variable at the central node, and $\boldsymbol{w}^j$ and $\boldsymbol{z}_j$ are the optimization variables at node j.

Unfortunately, relatively little is currently known about the geometric landscape of equality-constrained machine learning problems in the form of (2): Do they have spurious local minima? Do they satisfy the strict saddle property, or could they have degenerate saddle points?

However, insight into the geometry of problem (2) can be gained by applying Theorem 1. Problem (1) can be expressed in the form of Problem UC by taking[3] $\boldsymbol{x}=[\boldsymbol{w};\boldsymbol{z}]$ with $\boldsymbol{z}=[\boldsymbol{z}_1;\cdots;\boldsymbol{z}_J]$ and $c(\boldsymbol{x})=\sum_{j=1}^{J} f_j(\boldsymbol{w},\boldsymbol{z}_j)+g(\boldsymbol{w})$, while problem (2) can be expressed in the form of Problem ECD by taking $\boldsymbol{x}=[\boldsymbol{w};\boldsymbol{z}]$, $\boldsymbol{y}=[\boldsymbol{w}^1;\cdots;\boldsymbol{w}^J]$, $d(\boldsymbol{x},\boldsymbol{y})=\sum_{j=1}^{J} f_j(\boldsymbol{w}^j,\boldsymbol{z}_j)+g(\boldsymbol{w})$,

$$\boldsymbol{A}=\begin{bmatrix}-\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0}\\ \vdots & \mathbf{0} & \ddots & \mathbf{0}\\ -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix},\ \boldsymbol{B}=\begin{bmatrix}\mathbf{I} & \cdots & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{0}\\ \mathbf{0} & \cdots & \mathbf{I}\end{bmatrix},\ \boldsymbol{b}=\begin{bmatrix}\mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}. \tag{3}$$

We note that $\boldsymbol{B}$ (the identity matrix) is square and invertible. Under the constraint that $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{y}=\boldsymbol{b}$, which requires all $\boldsymbol{w}^j=\boldsymbol{w}$, we see that $d(\boldsymbol{x},\boldsymbol{y})=c(\boldsymbol{x})$. By applying Theorem 1, we obtain the following result.

Corollary 1. *$[\boldsymbol{w};\boldsymbol{z}]$ is a [first-order/second-order/strict saddle] critical point of problem* (1) *iff $([\boldsymbol{w};\boldsymbol{z}],[\boldsymbol{w};\cdots;\boldsymbol{w}])$ is a [first-order/second-order/strict saddle] point of problem* (2)*. Moreover, if problem* (1) *satisfies the strict saddle property and has no spurious local minima, then for every second-order critical point $([\boldsymbol{w};\boldsymbol{z}],[\boldsymbol{w};\cdots;\boldsymbol{w}])$ of problem* (2)*, $[\boldsymbol{w};\boldsymbol{z}]$ is a global minimizer of problem* (1)*.*

Corollary 1 allows one to borrow centralized geometric analysis for problem (1) to understand the landscape of the equality-constrained distributed problem (2).

4. GRADIENT ADMM (GADMM) ALGORITHM

We briefly discuss algorithmic aspects for solving equality-constrained distributed optimization problems. The recent Gradient ADMM (GADMM) algorithm [23] can be guaranteed under certain conditions to converge to a second-order critical point of an equality-constrained distributed optimization problem. Corollary 1 establishes conditions under which this point will correspond to a global minimizer of the original centralized optimization problem.

As outlined in [23, (38)], GADMM is intended for problems that can be expressed as[4]

$$\underset{\boldsymbol{x},\boldsymbol{y}}{\text{minimize}}\ f(\boldsymbol{x})+g(\boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{y}=\boldsymbol{b}. \tag{4}$$

[3] To simplify notation, we use $[\boldsymbol{p};\boldsymbol{q}]$ to represent $[\boldsymbol{p}^{\mathrm{T}}\ \boldsymbol{q}^{\mathrm{T}}]^{\mathrm{T}}$.

[4] The notations f and g are interchanged with respect to what appears in [23, (38)].

The global consensus problem (2) is of this form; to see this, let $\boldsymbol{x} = [\boldsymbol{w}^1; \cdots; \boldsymbol{w}^J; \boldsymbol{z}]$, $\boldsymbol{y} = \boldsymbol{w}$, $f(\boldsymbol{x}) = \sum_{j=1}^J f_j(\boldsymbol{w}^j, \boldsymbol{z}_j)$, $g(\boldsymbol{y}) = g(\boldsymbol{w})$,

$$\boldsymbol{A} = \begin{bmatrix} -\mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{5}$$

In [23, Section 3.1], it is shown how GADMM can be applied to the global consensus problem (2), with the resulting iterations

$$\begin{aligned} \boldsymbol{w}^j(k+1) &= \boldsymbol{w}^j(k) - \frac{1}{\beta}\big(\nabla f_j(\boldsymbol{w}^j(k), \boldsymbol{z}_j(k)) \\ &\qquad + \boldsymbol{\lambda}_j(k) + \rho(\boldsymbol{w}^j(k) - \boldsymbol{w}(k))\big), \\ \boldsymbol{z}_j(k+1) &= \boldsymbol{z}_j(k) - \frac{1}{\beta}\nabla f_j(\boldsymbol{w}^j(k), \boldsymbol{z}_j(k)), \\ \boldsymbol{w}(k+1) &= \boldsymbol{w}(k) - \frac{1}{\beta}\big(\nabla g(\boldsymbol{w}_k) \\ &\qquad - \sum_{j=1}^J (\boldsymbol{\lambda}_j(k) + \rho(\boldsymbol{w}^j(k+1) - \boldsymbol{w}(k))\big), \\ \boldsymbol{\lambda}_j(k+1) &= \boldsymbol{\lambda}_j(k) + \rho(\boldsymbol{w}^j(k) - \boldsymbol{w}(k)). \end{aligned} \tag{6}$$

These iterations require communication only between the central node and each of the nodes $1, 2, \ldots, J$. We note that this is the reason that we utilize (5) instead of (3) when applying GADMM for solving (2) since the resulting algorithm (6) is more suitable for distributed implementation. On the other hand, the form (3) is mainly utilized for analyzing the landscape of (2) by invoking Theorem 1.

For the global consensus problem, under assumptions B1–B5 in [23], with the proper selection of parameters β and ρ, and with random initialization of $\boldsymbol{w}(0)$, $\{\boldsymbol{w}^j(0)\}$, $\{\boldsymbol{z}_j(0)\}$, and $\{\boldsymbol{\lambda}_j(0)\}$ it is shown [23, Theorem 3.1] that with probability one, GADMM will converge to a second-order critical point of (2). According to Corollary 1, when problem (1) satisfies the strict saddle property and has no spurious local minima, this second-order critical point of (2) corresponds to a global minimizer of problem (1).

5. APPLICATION TO DISTRIBUTED LOW-RANK MATRIX APPROXIMATION

We now discuss our results in the context of distributed low-rank matrix approximation. Consider first the prototypical problem of finding, for a given a data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, a low-rank approximation by solving

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times m}}{\text{minimize}} \ \|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + \mu\|\boldsymbol{X}\|_*. \tag{7}$$

Here, the nuclear norm penalty promotes low-rank structure in the approximation $\boldsymbol{X}$. Problem (7) is an unconstrained convex optimization problem in the matrix variable $\boldsymbol{X}$. It is natural to consider solving problem (7) in factored form, where we replace the optimization variable $\boldsymbol{X}$ with $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times r}$ are tall matrices, and r is a parameter that must be set in advance (typically on the order of the rank r' expected of the optimal solution to (7)). Under this reparameterization, (7) becomes

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r}, \boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \ \|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{Y}\|_F^2 + \mu\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\|_*. \tag{8}$$

One can solve this problem using local search algorithms such as gradient descent. Such algorithms do not require expensive SVDs, nor do they require explicit storage of the matrix $\boldsymbol{X}$.

Unfortunately, problem (8) is nonconvex in the optimization variables $(\boldsymbol{U}, \boldsymbol{V})$. We have studied [13] the geometric landscape of problem (8) with a minor modification to the objective function:

$$\underset{\boldsymbol{U},\boldsymbol{V}}{\text{minimize}} \ \|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{Y}\|_F^2 + \frac{\mu}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right). \tag{9}$$

Despite the change of the objective function, the global minimizers remain unchanged. That is, any $(\boldsymbol{U}, \boldsymbol{V})$ that minimize (9) are also a global minimizer of (8).

We have shown [13] that every critical point of problem (9) is either a global minimum or a strict saddle point. This implies that local search algorithms such as gradient descent can be applied to problem (9) and will converge to a global minimum of (9). As previously noted, this then coincides with a global minimum of the original objective function, (8). This favorable geometry of problem (9) holds under the condition that there exists a global minimizer of (7) having rank r' and that $r \geq r'$.

One can generalize the unconstrained centralized problem (9) to an equality-constrained distributed problem similar to the global consensus problems outlined in Section 3. Suppose the columns of the data matrix $\boldsymbol{Y}$ are distributed among J nodes/sensors. Without loss of generality, partition the columns of $\boldsymbol{Y}$ as $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \cdots & \boldsymbol{Y}_J\end{bmatrix}$ where for $j \in \{1, 2, \ldots, J\}$, matrix $\boldsymbol{Y}_j$ (which is stored at node j) has size $n \times m_j$, and where $m = \sum_{j=1}^J m_j$. Partitioning $\boldsymbol{V}$ similarly as $\boldsymbol{V} = \begin{bmatrix}\boldsymbol{V}_1^{\mathrm{T}} & \boldsymbol{V}_2^{\mathrm{T}} & \cdots & \boldsymbol{V}_J^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$, where $\boldsymbol{V}_j$ has size $m_j \times r$, we can write $\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{Y}\|_F^2 = \sum_{j=1}^J \|\boldsymbol{U}\boldsymbol{V}_j^{\mathrm{T}} - \boldsymbol{Y}_j\|_F^2$. We use this fact to plug in for the term $\|\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} - \boldsymbol{Y}\|_F^2$ which appears in (9).

Suppose we introduce in problem (9) the optimization variables $\boldsymbol{U}^1, \ldots, \boldsymbol{U}^J \in \mathbb{R}^{n\times r}$ (all the same size as $\boldsymbol{U}$) and add an equality constraint to enforce consensus among these variables. We obtain the equality-constrained optimization problem

$$\begin{aligned} \underset{\boldsymbol{U},\{\boldsymbol{V}_j\},\{\boldsymbol{U}^J\}}{\text{minimize}} & \left(\sum_{j=1}^J \|\boldsymbol{U}^j\boldsymbol{V}_j^{\mathrm{T}} - \boldsymbol{Y}_j\|_F^2 + \frac{\mu}{2}\|\boldsymbol{V}_j\|_F^2\right) \\ & + \frac{\mu}{2}\|\boldsymbol{U}\|_F^2 \ \text{subject to} \ \boldsymbol{U}^j = \boldsymbol{U}, \ \forall\, j, \end{aligned} \tag{10}$$

which has the form of global consensus problem appearing in (2) by taking $\boldsymbol{w} = \mathrm{vec}(\boldsymbol{U}), \boldsymbol{z}_j = \mathrm{vec}(\boldsymbol{V}_j), \boldsymbol{w}^j = \mathrm{vec}(\boldsymbol{U}^j)$, and defining $f_j(\boldsymbol{w}^j, \boldsymbol{z}_j)$, $g(\boldsymbol{w})$ in the natural resulting way. By applying Corollary 1, we obtain the following result.

Corollary 2. *$\boldsymbol{x}_{UC} = [vec(\boldsymbol{U}); vec(\boldsymbol{V}_1); \cdots; vec(\boldsymbol{V}_J)]$ is a [first-order/second-order/strict saddle] critical point of problem* (9) *iff $\boldsymbol{x}_{ECD} = (\boldsymbol{x}_{UC}, [vec(\boldsymbol{U}); \cdots; vec(\boldsymbol{U})])$ is a [first-order/second-order/strict saddle] critical point of problem* (10)*. Moreover, under the condition that there exists a global minimizer of* (7) *having rank r' and that $r \geq r'$, for every second-order critical point $\boldsymbol{x}_{ECD}$ of problem* (10)*, $\boldsymbol{x}_{UC}$ is a global minimizer of problem* (9)*.*

We apply GADMM to solve (10). [23, Theorem 3.1] shows that, under suitable conditions, GADMM is guaranteed to converge to a second-order critical point of (10). Although we do not confirm those conditions for the matrix factorization problem (10), we use numerical simulations to illustrate the ability of GADMM to reach solutions that correspond to global minimizers of the centralized problem (9).

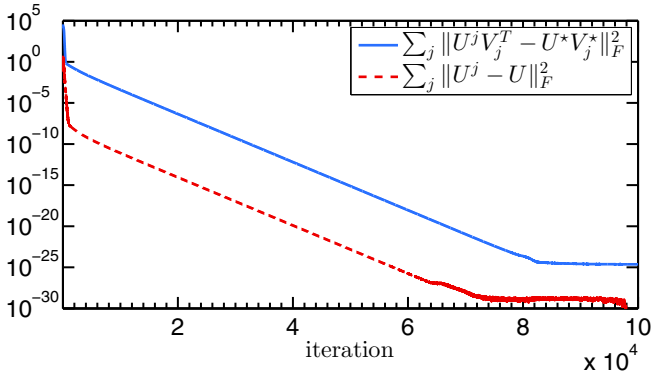

Fig. 1. Solving (10) by using GADMM (6).

To set up the experiments, we first generate the rank-r ground truth matrix $\boldsymbol{Y}^\# = [\boldsymbol{Y}_1^\# \ \cdots \ \boldsymbol{Y}_J^\#] \in \mathbb{R}^{n \times Jm}$ ($m = \sum_{j=1}^J m_j$) where $r = 2$, $n = 50$, $J = 10$, $m_j = 20 \ \forall j$ by multiplying two standard Gaussian matrices (i.e., each entry i.i.d. from $\mathcal{N}(0,1)$) of size $n \times r$ and $r \times m$, respectively. Then adding a noise matrix $\boldsymbol{N} \in \mathbb{R}^{n \times m}$ with each entry i.i.d. drawn from $\mathcal{N}(0, \sigma_Z^2)$ with $\sigma_Z = 0.1$, we get the noisy observation $\boldsymbol{Y} = \boldsymbol{Y}^\# + \boldsymbol{N}$. In this case, the signal-to-noise ratio can be computed as $\mathrm{SNR} = 10\log_{10}\left(\mathbb{E}\left[\|\boldsymbol{Y}^\#\|_F^2\right]/\mathbb{E}\left[\|\boldsymbol{N}\|_F^2\right]\right) = 10\log_{10}(\frac{r}{\sigma_Z^2}) = 23\,\mathrm{dB}$.

To estimate the ground truth, we then solve (10) with $\mu = 1$ by using GADMM (6) with $\rho = 10$, $\beta = 1000$ and a random initialization. To verify our main results (cf. Corollary 2), we plot the optimality distance $\sum_{j=1}^J \|\boldsymbol{U}^j \boldsymbol{V}_j^{\mathrm{T}} - \boldsymbol{U}^\star \boldsymbol{V}_j^\star\|_F^2$ and consensus error $\sum_{j=1}^J \|\boldsymbol{U}^j - \boldsymbol{U}\|_F^2$ as a function of the number of iterations, where $(\boldsymbol{U}^\star, [\boldsymbol{V}_1^\star \ \cdots \ \boldsymbol{V}_J^\star])$ is a global minimizer of problem (9). Figure 1 shows that the GADMM achieves both global optimum and exact consensus.

A. PROOF OF THEOREM 1

Proof. The first-order critical points $\boldsymbol{x}$ of Problem UC are those that satisfy

$$\nabla_{\boldsymbol{x}} c(\boldsymbol{x}) = 0. \tag{11}$$

The second-order critical points of Problem UC additionally satisfy

$$\nabla_{\boldsymbol{x}}^2 c(\boldsymbol{x}) \succeq 0, \tag{12}$$

and a first-order critical point is a strict saddle if it does not satisfy (12).

The critical points $(\boldsymbol{x}, \boldsymbol{y})$ of Problem ECD are defined through the Lagrangian function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) = d(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{b})$. The first-order critical points $(\boldsymbol{x}, \boldsymbol{y})$ of Problem ECD are those that satisfy the first-order optimality condition: $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{b}$ and there exists $\boldsymbol{\lambda}$ such that

$$\nabla_{[\boldsymbol{x}; \boldsymbol{y}]} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) = 0. \tag{13}$$

The second-order critical points of Problem ECD additionally satisfy the second-order optimality condition:

$$[\nabla_{[\boldsymbol{x}; \boldsymbol{y}]}^2 \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda})](\boldsymbol{v}, \boldsymbol{v}) \geq 0 \ \forall \boldsymbol{v} \in \mathcal{T}, \tag{14}$$

where

$$\mathcal{T} = \{\boldsymbol{v} = [\boldsymbol{v}_x; \boldsymbol{v}_y] : \boldsymbol{A}\boldsymbol{v}_x + \boldsymbol{B}\boldsymbol{v}_y = 0\} = \begin{bmatrix} \mathbb{R}^n \\ -\boldsymbol{B}^{-1}\boldsymbol{A}(\mathbb{R}^n) \end{bmatrix} \tag{15}$$

is the tangent plane of the constraint set $\mathcal{F} = \{\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{b}\}$, where we have used the nonsingularity of $\boldsymbol{B}$. A first-order critical point is a strict saddle if it does not satisfy (14).

For convenience, define

$$h(\boldsymbol{x}, \boldsymbol{y}) := d(\boldsymbol{x}, \boldsymbol{y}) - c(\boldsymbol{x}), \tag{16}$$

and note that $h(\boldsymbol{x}, \boldsymbol{y}) = 0$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}$. Note that $h(\boldsymbol{x}, \boldsymbol{y})$ has zero directional derivative and zero Hessian curvature along the tangent plane of $\mathcal{F}$. That is

$$\nabla_{[\boldsymbol{x}; \boldsymbol{y}]} h(\boldsymbol{x}, \boldsymbol{y})^{\mathrm{T}} \boldsymbol{v} = 0 \text{ and } [\nabla_{[\boldsymbol{x}; \boldsymbol{y}]}^2 h(\boldsymbol{x}, \boldsymbol{y})](\boldsymbol{v}, \boldsymbol{v}) = 0 \tag{17}$$

for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}$ and $\boldsymbol{v} \in \mathcal{T}$.

For any $\boldsymbol{x}$, let $\boldsymbol{y} = \boldsymbol{B}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$ and note that $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}$. Moreover, (13) holds iff $\nabla_{[\boldsymbol{x}; \boldsymbol{y}]} d(\boldsymbol{x}, \boldsymbol{y}) = [\boldsymbol{A} \ \boldsymbol{B}]^{\mathrm{T}} \boldsymbol{\lambda}$, which holds iff $[\nabla_{\boldsymbol{x}} c(\boldsymbol{x}); 0] + \nabla_{[\boldsymbol{x}; \boldsymbol{y}]} h(\boldsymbol{x}, \boldsymbol{y}) \perp \mathcal{T}$ (due to (15) and (16)), which holds iff $[\nabla_{\boldsymbol{x}} c(\boldsymbol{x}); 0] \perp \mathcal{T}$ (due to (17)), which holds iff (11) holds (due to (15)). Similarly, we have that (14) holds iff $[\nabla_{[\boldsymbol{x}; \boldsymbol{y}]}^2 c(\boldsymbol{x})](\boldsymbol{v}, \boldsymbol{v}) \geq 0 \ \forall \boldsymbol{v} \in \mathcal{T}$ (due to (16) and (17)), which holds iff (12) holds (due to (15)). This completes the proof of the three types of equivalence between a critical point $\boldsymbol{x}$ of Problem UC and a critical point $(\boldsymbol{x}, \boldsymbol{B}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}))$ of Problem ECD. □

B. REFERENCES

[1] Mark A Davenport and Justin Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

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

[3] 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.

[4] P. Jain and P. Kar, *Non-Convex Optimization for Machine Learning*, Foundations and Trends in Machine Learning Series. Now Publishers, 2017.

[5] Samuel Burer and Renato DC Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[6] Jasson DM Rennie and Nathan Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 713–719.

[7] Katta G Murty and Santosh N Kabadi, "Some np-complete problems in quadratic and nonlinear programming," *Mathematical programming*, vol. 39, no. 2, pp. 117–129, 1987.

[8] Eduardo D Sontag and Héctor J Sussmann, "Backpropagation can give rise to spurious local minima even for networks without hidden layers," *Complex Systems*, vol. 3, no. 1, pp. 91–106, 1989.

[9] Sean R. Eddy, "Profile hidden markov models.," *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, 1998.

[10] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Advances in neural information processing systems*, 2014, pp. 2933–2941.

[11] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.

[12] Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin, "The global optimization geometry of low-rank matrix optimization," *arXiv preprint arXiv:1703.01256*, 2017.

[13] Qiuwei Li, Zhihui Zhu, and Gongguo Tang, "The non-convex geometry of low-rank matrix optimization," *Information and Inference: A Journal of the IMA*, p. iay003, 2018.

[14] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal, "Nonconvex robust low-rank matrix recovery," *arXiv preprint arXiv:1809.09237*, 2018.

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

[16] Rong Ge, Jason D Lee, and Tengyu Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[17] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[18] Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao, "Symmetry, saddle points, and global geometry of nonconvex matrix factorization," *arXiv preprint*, 2016.

[19] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht, "First-order methods almost always avoid saddle points," *arXiv preprint arXiv:1710.07406*, 2017.

[20] Qiuwei Li, Zhihui Zhu, and Gongguo Tang, "Geometry of factored nuclear norm regularization," *arxiv:1704.01265*, 2017.

[21] Lin Gu, Deze Zeng, Peng Li, and Song 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.

[22] Gesualdo Scutari, Francisco Facchinei, Lorenzo Lampariello, Stefania Sardellitti, and Peiran 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.

[23] Mingyi Hong, Jason D Lee, and Meisam Razaviyayn, "Gradient primal-dual algorithm converges to second-order stationary solutions for nonconvex distributed optimization," *arXiv preprint arXiv:1802.08941*, 2018.

[24] Annie I-An Chen, *Fast distributed first-order methods*, Ph.D. thesis, Massachusetts Institute of Technology, 2012.

[25] Dušan Jakovetić, Joao Xavier, and José MF Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[26] Jinshan Zeng and Wotao Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[27] Amir Daneshmand, Gesualdo Scutari, and Vyacheslav Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *arXiv preprint arXiv:1809.08694*, 2018.