GLOBAL OPTIMALITY IN LOW-RANK MATRIX OPTIMIZATION

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

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

ABSTRACT

This paper considers the minimization of a general objective function $f(\boldsymbol{X})$ over the set of non-square $n \times m$ matrices where the optimal solution $\boldsymbol{X}^\star$ is low-rank. To reduce the computational burden, we factorize the variable $\boldsymbol{X}$ into a product of two smaller matrices and optimize over these two matrices instead of $\boldsymbol{X}$. We analyze the global geometry for a general and yet well-conditioned objective function $f(\boldsymbol{X})$ whose restricted strong convexity and restricted strong smoothness constants are comparable. In particular, we show that the reformulated objective function has no spurious local minima and obeys the strict saddle property. These geometric properties imply that a number of iterative optimization algorithms (such as gradient descent) can provably solve the factored problem with global convergence.

Index Terms— Low-rank matrix optimization, matrix sensing, nonconvex optimization, optimization geometry, strict saddle

1. INTRODUCTION

Consider the minimization of a general objective function $f(\boldsymbol{X})$ over all $n \times m$ matrices:

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n\times m}}{\text{minimize}}\, f(\boldsymbol{X}), \tag{1}$$

which we suppose admits a low-rank solution $\boldsymbol{X}^\star \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(\boldsymbol{X}^\star) = r^\star$. Low-rank matrix optimizations of the form (1) appear in a wide variety of applications, including quantum tomography [1], projection matrix design for compressive sensing [2], collaborative filtering [3], low-rank matrix recovery from compressive measurements [4], and matrix completion [5]. In order to find a low-rank solution, the nuclear norm is widely used in matrix inverse problems [4] arising in machine learning [6], signal processing [7], and control [8]. Although nuclear norm minimization enjoys strong statistical guarantees [5], its computational complexity is very high (as most algorithms require performing an expensive singular value decomposition (SVD) in each iteration), prohibiting it from scaling to practical problems.

This work was supported by NSF grant CCF-1409261, NSF grant CCF-1464205, and NSF CAREER grant CCF-1149225. Email: {zzhu, qiuli, gtang, mwakin}@mines.edu.

To relieve the computational bottleneck, recent studies propose to factorize the matrix variable into the Burer-Monteiro type decomposition [9, 10] with $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, and optimize over the $n \times r$ and $m \times r$ ($r \geq r^\star$) matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. With this parameterization of $\boldsymbol{X}$, we can recast (1) into the following program:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\, h(\boldsymbol{U},\boldsymbol{V}) := f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}). \tag{2}$$

The bilinear nature of the parameterization renders the objective function of (2) nonconvex even when $f(\boldsymbol{X})$ is a convex function. Hence, the objective function in (2) can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or "bad" saddle points that prevent a number of iterative algorithms from converging to the global solution. By analyzing the landscape of nonconvex functions, several recent works have shown that with an exact factorization ($r = r^\star$), the factored objective function $h(\boldsymbol{U},\boldsymbol{V})$ in matrix inverse problems has no spurious local minima [11–17].

We generalize this line of work by focusing on a general objective function $f(\boldsymbol{X})$ in the optimization (1), not necessarily a quadratic loss function coming from a matrix inverse problem. We provide a geometric analysis for the factored program (2) and show that all the critical points of the the objective function are well-behaved. Our characterization of the geometry of the objective function ensures a number of iterative optimization algorithms converge to a global minimum.

This paper continues in Section 2 with formal definitions for strict saddles and the strict saddle property. We present the main results in Section 3 and their implications in matrix recovery in Section 4.

2. PRELIMINARIES

2.1. Notation

To begin, we first briefly introduce some notation used throughout the paper. The symbols $\mathbf{I}$ and $\mathbf{0}$ respectively represent the identity matrix and zero matrix with appropriate sizes. The set of $r \times r$ orthonormal matrices is denoted by $\mathcal{O}_r := \{\boldsymbol{R} \in \mathbb{R}^{r\times r} : \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \mathbf{I}\}$. If a function $h(\boldsymbol{U},\boldsymbol{V})$ has two arguments, $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times r}$, we occasionally use the notation $h(\boldsymbol{W})$ when we put these two arguments into a new one as $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$. For a scalar func-

tion $f(\boldsymbol{Z})$ with a matrix variable $\boldsymbol{Z} \in \mathbb{R}^{n\times m}$, its gradient is an $n \times m$ matrix whose (i,j)-th entry is $[\nabla f(\boldsymbol{Z})]_{ij} = \frac{\partial f(\boldsymbol{Z})}{\partial \boldsymbol{Z}_{ij}}$ for all $i \in \{1,2,\ldots,n\}, j \in \{1,2,\ldots,m\}$. The Hessian of $f(\boldsymbol{Z})$ can be represented with a bilinear form defined via $[\nabla^2 f(\boldsymbol{Z})](\boldsymbol{A},\boldsymbol{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\boldsymbol{Z})}{\partial \boldsymbol{Z}_{ij}\partial \boldsymbol{Z}_{kl}} \boldsymbol{A}_{ij}\boldsymbol{B}_{kl}$ for any $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{n\times m}$.

2.2. Strict Saddle Property

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable objective function. We begin with the notion of strict saddles and the strict saddle property.

Definition 1 (Strict saddles)**.** *A critical point $\boldsymbol{x}$ is a strict saddle if the Hessian matrix evaluated at this point has a strictly negative eigenvalue, i.e., $\lambda_{\min}(\nabla^2 h(\boldsymbol{x})) < 0$.*

Definition 2 (Strict saddle property [18])**.** *A twice differentiable function satisfies the strict saddle property if each critical point either corresponds to a local minimum or is a strict saddle.*

Intuitively, the strict saddle property requires a function to have a directional negative curvature at all the critical points but local minima. This property allows a number of iterative algorithms such as noisy gradient descent [18] and the trust region method [19] to further decrease the function value at all the strict saddles and thus converge to a local minimum.

Theorem 1. *[18, 20, 21] (informal) For a twice continuously differentiable objective function satisfying the strict saddle property, a number of iterative optimization algorithms (such as gradient descent and the the trust region method) can find a local minimum.*

3. PROBLEM FORMULATION AND MAIN RESULTS

3.1. Problem Formulation

This paper considers the problem (1) of minimizing a general function $f(\boldsymbol{X})$ admitting a low-rank solution $\boldsymbol{X}^\star$ with $\mathrm{rank}(\boldsymbol{X}^\star) = r^\star \le r$. We factorize the variable $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{n\times r}, \boldsymbol{V} \in \mathbb{R}^{m\times r}$ and transform (1) into its factored counterpart (2). Throughout the paper, $\boldsymbol{X}$, $\boldsymbol{W}$ and $\widehat{\boldsymbol{W}}$ are matrices depending on $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix},\ \widehat{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{U}\\ -\boldsymbol{V}\end{bmatrix},\ \boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}.$$

Although the new variable $\boldsymbol{W}$ has much smaller size than $\boldsymbol{X}$ when $r \ll \min\{n,m\}$, the objective function in the factored problem (2) may have a much more complicated landscape due to the bilinear form about $\boldsymbol{U}$ and $\boldsymbol{V}$. The reformulated objective function $h(\boldsymbol{U},\boldsymbol{V})$ could introduce spurious local minima or degenerate saddle points even when $f(\boldsymbol{X})$ is convex. Our goal is to guarantee that this does not happen.

Let $\boldsymbol{X}^\star = \boldsymbol{Q}_{\boldsymbol{U}^\star}\boldsymbol{\Sigma}^\star\boldsymbol{Q}_{\boldsymbol{V}^\star}^{\mathrm{T}}$ denote an SVD of $\boldsymbol{X}^\star$, where $\boldsymbol{Q}_{\boldsymbol{U}^\star} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{Q}_{\boldsymbol{V}^\star} \in \mathbb{R}^{m\times r}$ are orthonormal matrices of appropriate sizes, and $\boldsymbol{\Sigma}^\star \in \mathbb{R}^{r\times r}$ is a diagonal matrix with non-negative diagonals (but with some zero diagonals if $r > r^\star = \mathrm{rank}(\boldsymbol{X}^\star)$). We denote

$$\boldsymbol{U}^\star = \boldsymbol{Q}_{\boldsymbol{U}^\star}\boldsymbol{\Sigma}^{\star 1/2}, \quad \boldsymbol{V}^\star = \boldsymbol{Q}_{\boldsymbol{V}^\star}\boldsymbol{\Sigma}^{\star 1/2},$$

where $\boldsymbol{X}^\star = \boldsymbol{U}^\star\boldsymbol{V}^{\star\mathrm{T}}$ forms a balanced factorization of $\boldsymbol{X}^\star$ since $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ have the same singular values. Throughout the paper, we utilize the following two ways to stack $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ together:

$$\boldsymbol{W}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ \boldsymbol{V}^\star\end{bmatrix}, \quad \widehat{\boldsymbol{W}}^\star = \begin{bmatrix}\boldsymbol{U}^\star\\ -\boldsymbol{V}^\star\end{bmatrix}.$$

Before moving on, we note that for any solution $(\boldsymbol{U},\boldsymbol{V})$ to (2), $(\boldsymbol{U}\boldsymbol{\Psi},\boldsymbol{V}\boldsymbol{\Phi})$ is also a solution to (2) for any $\boldsymbol{\Psi},\boldsymbol{\Phi} \in \mathbb{R}^{r\times r}$ such that $\boldsymbol{U}\boldsymbol{\Psi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$. In order to address this ambiguity (i.e., to reduce the search space of $\boldsymbol{W}$ for (2)), we utilize the trick in [13, 22, 23] by introducing a regularizer

$$g(\boldsymbol{U},\boldsymbol{V}) = \frac{\mu}{4}\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F^2 \tag{3}$$

and solving the following problem

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\ \rho(\boldsymbol{U},\boldsymbol{V}) := f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}) + g(\boldsymbol{U},\boldsymbol{V}), \tag{4}$$

where $\mu > 0$ controls the weight for the term $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\|_F^2$, which will be discussed soon.

We remark that $\boldsymbol{W}^\star$ is still a global minimizer to the factored problem (4) since $f(\boldsymbol{X})$ and $g(\boldsymbol{W})$ achieve their global minimum at $\boldsymbol{X}^\star$ and $\boldsymbol{W}^\star$, respectively. The regularizer $g(\boldsymbol{W})$ is applied to force the difference between the two Gram matrices of $\boldsymbol{U}$ and $\boldsymbol{V}$ to be as small as possible. The global minimum of $g(\boldsymbol{W})$ is 0, which is achieved when $\boldsymbol{U}$ and $\boldsymbol{V}$ have the same Gram matrices, i.e., when $\boldsymbol{W}$ belongs to

$$\mathcal{E} := \left\{\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} : \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{0}\right\}. \tag{5}$$

Informally, we can view (4) as finding a point from $\mathcal{E}$ that also minimizes $f(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}})$. This is formally established in Theorem 2.

3.2. Main Results

Before presenting our main results, we lay out the necessary assumptions on the objective function $f(\boldsymbol{X})$. As is known, without any assumptions on the problem, even minimizing traditional quadratic objective functions is challenging. For this purpose, we focus on the model where $f(\boldsymbol{X})$ is $(2r,4r)$-restricted strongly convex and smooth, i.e., for any $n \times m$ matrices $\boldsymbol{X},\boldsymbol{G}$ with $\mathrm{rank}(\boldsymbol{X}) \le 2r$ and $\mathrm{rank}(\boldsymbol{G}) \le 4r$, the Hessian of $f(\boldsymbol{X})$ satisfies

$$\alpha\|\boldsymbol{G}\|_F^2 \le [\nabla^2 f(\boldsymbol{X})](\boldsymbol{G},\boldsymbol{G}) \le \beta\|\boldsymbol{G}\|_F^2 \tag{6}$$

for some positive α and β. A similar assumption is also utilized in [23, Conditions 5.3 and 5.4]. With this assumption on $f(\boldsymbol{X})$, we now summarize our main results.

Our main argument is that the objective function $\rho(\boldsymbol{W})$ has no spurious local minima and satisfies the strict saddle property. This is equivalent to categorizing all the critical points into two types: 1) the global minima which correspond to the global solution of the original convex problem (1) and 2) strict saddles such that the Hessian matrix $\nabla^2\rho(\boldsymbol{W})$ evaluated at these points has a strictly negative eigenvalue. We formally establish this in the following theorem, whose proof is omitted due to space limitations.

Theorem 2. *For any $\mu > 0$, each critical point $\boldsymbol{W} = \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}$ of $\rho(\boldsymbol{W})$ defined in* (4) *satisfies*

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} - \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{0}. \tag{7}$$

Furthermore, suppose the function $f(\boldsymbol{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6) *with positive constants α and β satisfying $\frac{\beta}{\alpha} \leq 1.5$. Set $\mu \leq \frac{\alpha}{16}$ for the factored problem* (4)*. Then $\rho(\boldsymbol{W})$ has no spurious local minimum, i.e., any local minimum of $\rho(\boldsymbol{W})$ is a global minimum corresponding to the global solution of the original convex problem* (1)*: $\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{X}^\star$. In addition, $\rho(\boldsymbol{W})$ obeys the strict saddle property that any critical point not being a local minimum is a strict saddle with*

$$\lambda_{\min}\left(\nabla^2\left(\rho(\boldsymbol{W})\right)\right) \leq \begin{cases} -0.08\alpha\sigma_r(\boldsymbol{X}^\star), & r = r^\star \\ -0.05\alpha\cdot\min\left\{\sigma^2_{r^c}(\boldsymbol{W}), 2\sigma_{r^\star}(\boldsymbol{X}^\star)\right\}, & r > r^\star \\ -0.1\alpha\sigma_{r^\star}(\boldsymbol{X}^\star), & r_c = 0, \end{cases} \tag{8}$$

where $r^c \leq r$ is the rank of $\boldsymbol{W}$, $\lambda_{\min}(\cdot)$ represents the smallest eigenvalue, and $\sigma_\ell(\cdot)$ denotes the ℓ-th largest singular value.

Remark 1. The above result implies that we can recover the rank-$r^\star$ global minimizer $\boldsymbol{X}^\star$ of (1) by many iterative algorithms (such as the trust region method [24] and stochastic gradient descent [18]) even from a random initialization. This is because 1) as guaranteed by Theorem 1, the strict saddle property ensures local search algorithms converge to a local minimum, and 2) there are no spurious local minima.

Remark 2. Since our main result only requires the $(2r, 4r)$-restricted strong convexity and smoothness property (6), aside from low-rank matrix recovery [25], it can also be applied to many other low-rank matrix optimization problems [26] which do not necessarily involve quadratic loss functions. Typical examples include robust PCA [27], 1-bit matrix completion [28] and Possion principal component analysis (PCA) [29].

Remark 3. Equation (7) shows that any critical point $\boldsymbol{W}$ belongs to $\mathcal{E}$ for the objective function in the factored problem (4) with any positive μ. This demonstrates the reason for adding the regularizer $g(\boldsymbol{U}, \boldsymbol{V})$. Thus, any iterative optimization algorithm converging to some critical point of $\rho(\boldsymbol{W})$ results in a solution within $\mathcal{E}$.

Remark 4. For any critical point $\boldsymbol{W} \in \mathbb{R}^{(n+m)\times r}$ but not being a local minimum, the right hand side of (8) is strictly negative, implying $\boldsymbol{W}$ is a strict saddle. We also note that Theorem 2 not only covers exact parameterization where $r = r^\star$, but also includes over-parameterization where $r > r^\star$.

Remark 5. The constants appearing in Theorem 2 are not optimized. We use $\mu \leq \frac{1}{16}\alpha$ simply to include $\mu = \frac{1}{16}$ which is utilized for the matrix sensing problem in [22, p.3]. If the ratio between the restricted strong convexity and smoothness constants $\frac{\beta}{\alpha} \leq 1.4$, then we can show $\rho(\boldsymbol{W})$ has no spurious local minima and obeys the strict saddle property for any $\mu \leq \frac{1}{4}\alpha$ (where $\mu = \frac{1}{4}$ is utilized for the matrix sensing problem in [13, p.3]). In all cases, a smaller μ yields a more negative constant in (8). This implies that when the restricted strong convexity constant α is not provided a priori, one can always choose a small μ to ensure the strict saddle property holds, and hence guarantee the global convergence of many iterative optimization algorithms. We finally note that the regularizer $g(\boldsymbol{W})$ is added mostly to avoid bad iterates $\boldsymbol{W}$. If we apply local search algorithms with a random initialization, then it is possible to drop the regularizer (i.e., set $\mu = 0$) for practical implementation; see the experiments in Section 4.1.

4. STYLIZED APPLICATION: MATRIX RECOVERY

We first consider the implication of Theorem 2 in the matrix sensing problem where

$$f(\boldsymbol{X}) = \frac{1}{2}\left\|\mathcal{A}\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)\right\|_2^2.$$

Here $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ is a known measurement operator satisfying the following restricted isometry property.

Definition 3. *(Restricted Isometry Property (RIP) [4, p.11]) The map $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ satisfies the r-RIP with constant δ_r if*

$$(1-\delta_r)\left\|\boldsymbol{X}\right\|_F^2 \leq \left\|\mathcal{A}(\boldsymbol{X})\right\|^2 \leq (1+\delta_r)\left\|\boldsymbol{X}\right\|_F^2 \tag{9}$$

holds for any $n \times m$ matrix $\boldsymbol{X}$ with $\mathrm{rank}(\boldsymbol{X}) \leq r$.

Note that in this case, the Hessian quadrature form $\nabla^2 f(\boldsymbol{X})[\boldsymbol{Y}, \boldsymbol{Y}]$ for any $n \times m$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$\nabla^2 f(\boldsymbol{X})[\boldsymbol{Y}, \boldsymbol{Y}] = \left\|\mathcal{A}(\boldsymbol{Y})\right\|^2.$$

If $\mathcal{A}$ satisfies the $4r$-restricted isometry property with constant δ_{4r}, then $f(\boldsymbol{X})$ satisfies the $(2r, 4r)$-restricted strong convexity and smoothness condition (6) with constants $\alpha = 1 - \delta_{4r}$ and $\beta = 1 - \delta_{4r}$ since

$$(1-\delta_{4r})\left\|\boldsymbol{Y}\right\|_F^2 \leq \left\|\mathcal{A}(\boldsymbol{Y})\right\|^2 \leq (1+\delta_{4r})\left\|\boldsymbol{Y}\right\|_F^2$$

for any rank-$4r$ matrix $\boldsymbol{Y}$. Now applying Theorem 2, we can characterize the geometry for the following matrix sensing problem with the factorization approach:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r},\boldsymbol{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \frac{1}{2}\|\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}-\boldsymbol{X}^{\star})\|_2^2+g(\boldsymbol{U},\boldsymbol{V}), \quad (10)$$

where $g(\boldsymbol{U},\boldsymbol{V})$ is the added regularizer defined in (3).

Corollary 1. *Suppose $\mathcal{A}$ satisfies the $4r$-RIP with constant $\delta_{4r}\le\frac{1}{5}$, and set $\mu\le\frac{1-\delta_{4r}}{16}$. Then the objective function in* (10) *has no spurious local minima and satisfies the strict saddle property.*

This result follows directly from Theorem 2 by noting that $\frac{\beta}{\alpha}=\frac{1+\delta_{4r}}{1-\delta_{4r}}\le 1.5$ if $\delta_{4r}\le\frac{1}{5}$. We remark that Park et al. [13, Theorem 4.3] provided a similar geometric result for (10). Compared to their result which requires $\delta_{4r}\le\frac{1}{100}$, our result has a much weaker requirement on the RIP of the measurement operator.

4.1. Experiments

In this section, we present some experiments to illustrate the performance of the factorization approach for matrix completion[1] where we want to recover a low-rank matrix $\boldsymbol{X}^{\star}$ from incomplete measurements $\{X^{\star}_{ij}\}_{(i,j)\in\Omega}$, where $\Omega\subset[n]\times[m]$. We compare the performance of the matrix factorization approach with SVP [30], the convex approach, and singular value thresholding[2] (SVT) [31]. Let $\mathcal{P}_{\Omega}$ denote the projection onto the index set Ω. The convex approach (denoted by CVX) attempts to use the nuclear norm as a convex relaxation of the rankness and solves

$$\underset{\boldsymbol{X}}{\text{minimize}}\ \|\boldsymbol{X}\|_*,\ \text{subject to}\ \ \mathcal{P}_{\Omega}(\boldsymbol{X})=\mathcal{P}_{\Omega}(\boldsymbol{X}^{\star}). \quad (11)$$

In the first set of experiments, we set $n=m=100$ and vary the rank r from 1 to 30. We generate a rank-r random matrix and randomly obtain p entries, i.e., $|\Omega|=p$. Figure 1 displays the phase transition for matrix factorization (with $\mu=0$) solved by gradient descent with a random initialization, the SVP [30], the singular value thresholding (SVT) [31], and the convex approach. As can been seen, the matrix factorization approach has a similar phase transition to SVP, and is slightly better than SVT and the convex approach in terms of the number of measurements needed for successful recovery. Similar phase transitions for matrix factorization are also observed for different $\mu>0$.

In the second set of experiments, we set $r=5$ and $p=3r(2n-r)$ (3 times the number of degrees of freedom within a rank-r $n\times n$ matrix), and vary n from 40 to 5120. We compare the time needed for the four approaches in Figure 2; our matrix factorization approach is much faster than the other methods. The time savings for the matrix factorization approach comes from avoiding the SVD, which is needed both for SVT and SVP in each iteration. We also observe that the convex approach has highest computational complexity and is not scalable (which is the reason that we only present its time for n up to 640).

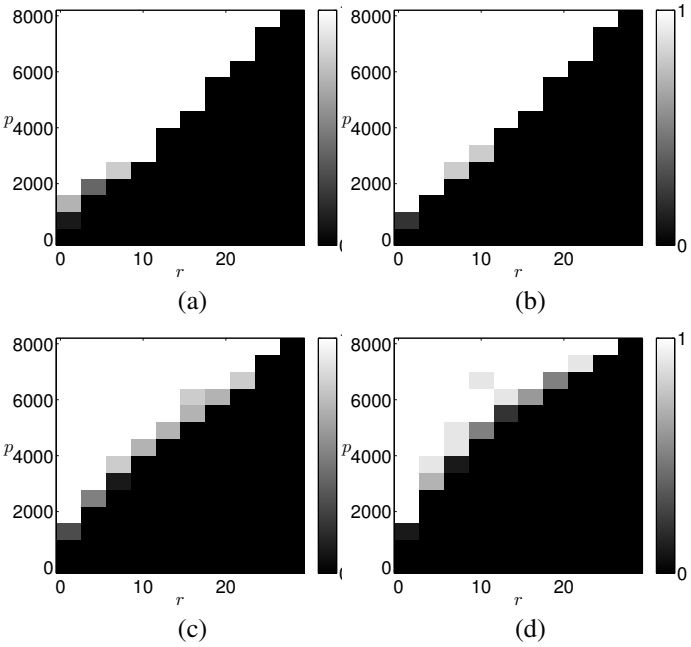

Fig. 1. Rate of success for matrix completion by (a) the matrix factorization approach with gradient descent; (b) SVP [30]; (c) solving the convex problem (11); (d) SVT [30]. 10 Monte Carlo trials are carried out and for each trial, we claim matrix recovery to be successful if the relative reconstruction error satisfies $\frac{\|\boldsymbol{X}^{\star}-\widehat{\boldsymbol{X}}\|_F}{\|\boldsymbol{X}^{\star}\|_F}\le 10^{-4}$.

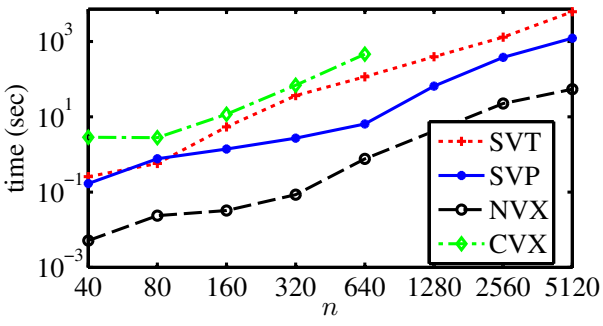

Fig. 2. Average computation time needed for different algorithms solving matrix completion.

[1]Though $\mathcal{P}_{\Omega}$ does not satisfy the r-RIP (9) for all low-rank matrices $\boldsymbol{X}$, it satisfies the RIP when restricted to low-rank incoherent matrices [30, Theorem 4.2]. Thus, if the iterates of local search algorithms remain incoherent (which is experimentally observed), then Theorem 2 guarantees the global convergence of the matrix factorization approach with these algorithms.

[2]Software available at http://svt.stanford.edu/

5. REFERENCES

[1] S. Aaronson, "The learnability of quantum states," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 463, pp. 3089–3114, The Royal Society, 2007.

[2] T. Hong, H. Bai, S. Li, and Z. Zhu, "An efficient algorithm for designing projection matrix in compressive sensing based on alternating optimization," *Signal Processing*, vol. 125, pp. 9–20, 2016.

[3] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 1329–1336, 2004.

[4] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[5] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[6] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick, "Large-scale image classification with trace-norm regularization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3386–3393, IEEE, 2012.

[7] M. A. Davenport and J. 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.

[8] K. Mohan and M. Fazel, "Reweighted nuclear norm minimization with application to system identification," in *Proceedings of the 2010 American Control Conference*, pp. 2953–2959, IEEE, 2010.

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

[10] S. Burer and R. D. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Mathematical Programming*, vol. 103, no. 3, pp. 427–444, 2005.

[11] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," *arXiv preprint arXiv:1605.07221*, 2016.

[12] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," *arXiv preprint arXiv:1605.07272*, 2016.

[13] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach," *arXiv preprint arXiv:1609.03240*, 2016.

[14] Q. Li and G. Tang, "The nonconvex geometry of low-rank matrix optimizations with general objective functions," *arXiv preprint arXiv:1611.03060*, 2016.

[15] 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.

[16] Q. Li, Z. Zhu, and G. Tang, "Geometry of factored nuclear norm regularization," *arXiv preprint arXiv:1704.01265*, 2017.

[17] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "The global optimization geometry of nonsymmetric matrix factorization and sensing," *arXiv preprint arXiv:1703.01256*, 2017.

[18] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.

[19] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*. SIAM, 2000.

[20] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?," *arXiv preprint arXiv:1510.06096*, 2015.

[21] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent converges to minimizers," *University of California, Berkeley*, vol. 1050, p. 16, 2016.

[22] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via Procrustes flow," *arXiv preprint arXiv:1507.03566*, 2015.

[23] L. Wang, X. Zhang, and Q. Gu, "A unified computational and statistical framework for nonconvex low-rank matrix estimation," *arXiv preprint arXiv:1610.05275*, 2016.

[24] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv preprint arXiv:1602.06664*, 2016.

[25] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[26] M. Udell, C. Horn, R. Zadeh, and S. Boyd, "Generalized low rank models," *arXiv preprint arXiv:1410.0342*, 2014.

[27] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[28] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, vol. 3, no. 3, pp. 189–223, 2014.

[29] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, "Poisson noise reduction with non-local PCA," *Journal of Mathematical Imaging and Vision*, vol. 48, no. 2, pp. 279–294, 2014.

[30] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Information Processing Systems*, pp. 937–945, 2010.

[31] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.