No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis

Rong Ge [1] **Chi Jin** [2] **Yi Zheng** [1]

Abstract

In this paper we develop a new framework that captures the common landscape underlying the common non-convex low-rank matrix problems including matrix sensing, matrix completion and robust PCA. In particular, we show for all above problems (including asymmetric cases): 1) all local minima are also globally optimal; 2) no high-order saddle points exists. These results explain why simple algorithms such as stochastic gradient descent have global converge, and efficiently optimize these non-convex objective functions in practice. Our framework connects and simplifies the existing analyses on optimization landscapes for matrix sensing and symmetric matrix completion. The framework naturally leads to new results for asymmetric matrix completion and robust PCA.

1. Introduction

Non-convex optimization is one of the most powerful tools in machine learning. Many popular approaches, from traditional ones such as matrix factorization (Hotelling, 1933) to modern deep learning (Bengio, 2009) rely on optimizing non-convex functions. In practice, these functions are optimized using simple algorithms such as alternating minimization or gradient descent. Why such simple algorithms work is still a mystery for many important problems.

One way to understand the success of non-convex optimization is to study the optimization landscape: for the objective function, where are the possible locations of global optima, local optima and saddle points. Recently, a line of works showed that several natural problems including tensor decomposition (Ge et al., 2015), dictionary learning (Sun et al., 2015a), matrix sensing (Bhojanapalli et al., 2016; Park et al., 2016) and matrix completion (Ge et al., 2016) have well-behaved optimization landscape: all local optima are also globally optimal. Combined with recent results (e.g. Ge et al. (2015); Carmon et al. (2016); Agarwal et al. (2016); Jin et al. (2017)) that are guaranteed to find a local minimum for many non-convex functions, such problems can be efficiently solved by basic optimization algorithms such as stochastic gradient descent.

In this paper we focus on optimization problems that look for low rank matrices using partial or corrupted observations. Such problems are studied extensively (Fazel, 2002; Rennie & Srebro, 2005; Candès & Recht, 2009) and has many applications in recommendation systems (Koren, 2009), see survey by Davenport & Romberg (2016). These optimization problems can be formalized as follows:

$$\min_{\mathbf{M}\in\mathbb{R}^{d_1\times d_2}} \quad f(\mathbf{M}), \qquad (1)$$
$$s.t. \quad \text{rank}(\mathbf{M}) = r.$$

Here $\mathbf{M}$ is an $d_1 \times d_2$ matrix and f is a convex function of $\mathbf{M}$. The non-convexity of this problem stems from the low rank constraint. Several interesting problems, such as matrix sensing (Recht et al., 2010), matrix completion (Candès & Recht, 2009) and robust PCA (Candès et al., 2011) can all be framed as optimization problems of this form(see Section 3).

In practice, Burer & Monteiro (2003) heuristic is often used – replace $\mathbf{M}$ with an explicit low rank representation $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{d_1\times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2\times r}$. The new optimization problem becomes

$$\min_{\mathbf{U}\in\mathbb{R}^{d_1\times r},\mathbf{V}\in\mathbb{R}^{d_2\times r}} f(\mathbf{U}\mathbf{V}^\top) + Q(\mathbf{U},\mathbf{V}). \qquad (2)$$

Here $Q(\mathbf{U},\mathbf{V})$ is a (optional) regularizer. Despite the objective being non-convex, for all the problems mentioned above, simple iterative updates from random or even arbitrary initial point find the optimal solution in practice. It is then natural to ask: **Can we characterize the similarities between the optimization landscape of these problems?** We show this is indeed possible:

Theorem 1 (informal)**.** *The objective function of matrix sensing, matrix completion and robust PCA have similar*

Authors listed alphabetically. [1]Duke University, Durham NC [2]UC Berkeley, Berkeley CA. Correspondence to: Rong Ge <rongge@cs.duke.edu>, Chi Jin <chijin@cs.berkeley.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

optimization landscape. In particular, for all these problems, 1) all local minima are also globally optimal; 2) any saddle point has at least one strictly negative eigenvalue in its Hessian.

More precise theorem statements appear in Section 3. Note that there were several cases (matrix sensing (Bhojanapalli et al., 2016; Park et al., 2016), symmetric matrix completion (Ge et al., 2016)) where similar results on the optimization landscape were known. However the techniques in previous works are tailored to the specific problems and hard to generalize. Our framework captures and simplifies all these previous results, and also gives new results on asymmetric matrix completion and robust PCA.

The key observation in our analysis is that for matrix sensing, matrix completion, and robust PCA (when fixing sparse estimate), function f (in Equation (1)) is a quadratic function over the matrix $\mathbf{M}$. Hence the Hessian $\mathcal{H}$ of f with respect to $\mathbf{M}$ is a constant. More importantly, the Hessian $\mathcal{H}$ in all above problems has similar properties (that it approximately preserves norm, similar to the RIP properties used in matrix sensing (Recht et al., 2010)), which allows their optimization landscapes to be characterized in a unified way. Specifically, our framework gives principled way of defining a *direction of improvement* for all points that are not globally optimal.

Another crucial property of our framework is the interaction between the regularizer and the Hessian $\mathcal{H}$. Intuitively, the regularizer makes sure the solution is in a nice region $\mathcal{B}$ (e.g. set of incoherent matrices for matrix completion), and only within $\mathcal{B}$ the Hessian has the norm preserving property. On the other hand, regularizer should not be too large to severely distort the landscape. This interaction is crucial for matrix completion, and is also very useful in handling noise and perturbations. In Section 4, we discuss ideas required to apply this framework to matrix sensing, matrix completion and robust PCA.

Using this framework, we also give a way to *reduce* asymmetric matrix problems to symmetric PSD problems (where the desired matrix is of the form $\mathbf{U}\mathbf{U}^\top$). See Section 5 for more details.

In addition to the results of no spurious local minima, our framework also implies that any saddle point has at least one strictly negative eigenvalue in its Hessian. Formally, we proved all above problems satisfy a robust version of this claim — strict saddle property (see Definition 2), which is one of crucial sufficient conditions to admit efficient optimization algorithms, and thus following corollary:

Corollary 2 (informal)**.** *For matrix sensing, matrix completion and robust PCA, simple local search algorithms can find the desired low rank matrix $\mathbf{U}\mathbf{V}^\top = \mathbf{M}^\star$ from an arbitrary starting point in polynomial time with high probability.*

Several algorithms, including many variants of gradient descent Ge et al. (2015); Carmon et al. (2016); Agarwal et al. (2016); Jin et al. (2017) are known to converge to a local optimum for strict-saddle functions, and hence can be applied to the problems discussed in this paper. There are some technicalities in the exact guarantees, which we defer to supplementary material.

For simplicity, we present most results in the noiseless setting, but our results can also be generalized to handle noise. See supplementary material for details.

1.1. Related Works

The landscape of low rank matrix problems have recently received a lot of attention. Ge et al. (2016) showed symmetric matrix completion has no spurious local minimum. At the same time, Bhojanapalli et al. (2016) proved similar result for symmetric matrix sensing. Park et al. (2016) extended the matrix sensing result to asymmetric case. All of these works guarantee global convergence to the correct solution.

There has been a lot of work on the local convergence analysis for various algorithms and problems. For matrix sensing or matrix completion, the works (Keshavan et al., 2010a;b; Hardt & Wootters, 2014; Hardt, 2014; Jain et al., 2013; Chen & Wainwright, 2015; Sun & Luo, 2015; Zhao et al., 2015; Zheng & Lafferty, 2016; Tu et al., 2015) showed that given a good enough initialization, many simple local search algorithms, including gradient descent and alternating least squares, succeed. Particularly, several works (e.g. Sun & Luo (2015); Zheng & Lafferty (2016)) accomplished this by showing a geometric property which is very similar to strong convexity holds in the neighborhood of optimal solution. For robust PCA, there are also many analysis for local convergence (Lin et al., 2010; Netrapalli et al., 2014; Yi et al., 2016; Zhang et al., 2017).

Several works also try to unify the analysis for similar problems. Bhojanapalli et al. (2015) gave a framework for local analysis for these low rank problems. Belkin et al. (2014) showed a framework of learning basis functions, which generalizes tensor decompositions. Their techniques imply the optimization landscape for all such problems are very similar. For problems looking for a symmetric PSD matrix, Li & Tang (2016) showed for objective similar to (2) (but in the symmetric setting), restricted smoothness/strong convexity on the function f suffices for local analysis. However, their framework does not address the interaction between regularizer and the function f, hence cannot be directly applied to problems such as matrix completion or robust PCA.

Organization We will first introduce notations and basic optimality conditions in Section 2. Then Section 3 introduces the problems and our results. For simplicity, we present our framework for the symmetric case in Section 4, and briefly discuss how to reduce asymmetric problem to symmetric problem in Section 5. For clean presentation, many proofs are deferred to supplementary material.

2. Preliminaries

In this section we introduce notations and basic optimality conditions.

2.1. Notations

We use bold letters for matrices and vectors. For a vector $\mathbf{v}$ we use $\|\mathbf{v}\|$ to denote its ℓ_2 norm. For a matrix $\mathbf{M}$ we use $\|\mathbf{M}\|$ to denote its spectral norm, and $\|\mathbf{M}\|_F$ to denote its Frobenius norm. For vectors we use $\langle \mathbf{u}, \mathbf{v}\rangle$ to denote inner-product, and for matrices we use $\langle \mathbf{M}, \mathbf{N}\rangle = \sum_{i,j} \mathbf{M}_{ij}\mathbf{N}_{ij}$ to denote the trace of $\mathbf{M}\mathbf{N}^\top$. We will always use $\mathbf{M}^\star$ to denote the optimal low rank solution. Further, we use $\sigma_1^\star$ to denote its largest singular value, $\sigma_r^\star$ to denote its r-th singular value and $\kappa^\star = \sigma_1^\star/\sigma_r^\star$ be the condition number.

We use ∇f to denote the gradient and $\nabla^2 f$ to denote its Hessian. Since function f can often be applied to both $\mathbf{M}$ (as in (1)) and $\mathbf{U}, \mathbf{V}$ (as in (2)), we use $\nabla f(\mathbf{M})$ to denote gradient with respect to $\mathbf{M}$ and $\nabla f(\mathbf{U}, \mathbf{V})$ to denote gradient with respect to $\mathbf{U}, \mathbf{V}$. Similar notation is used for Hessian. The Hessian $\nabla^2 f(\mathbf{M})$ is a crucial object in our framework. It can be interpreted as a linear operator on matrices. This linear operator can be viewed as a $d_1 d_2 \times d_1 d_2$ matrix (or $\binom{d+1}{2} \times \binom{d+1}{2}$ matrix in the symmetric case) that applies to the vectorized version of matrices. We use the notation $\mathbf{M} : \mathcal{H} : \mathbf{N}$ to denote the quadratic form $\langle \mathbf{M}, \mathcal{H}(\mathbf{N})\rangle$. Similarly, the Hessian of objective (2) is a linear operator on a pair of matrices $\mathbf{U}, \mathbf{V}$, which we usually denote as $\nabla^2 f(\mathbf{U}, \mathbf{V})$.

2.2. Optimality Conditions

Local Optimality Suppose we are optimizing a function $f(\mathbf{x})$ with no constraints on $\mathbf{x}$. In order for a point $\mathbf{x}$ to be a local minimum, it must satisfy the first and second order necessary conditions. That is, we must have $\nabla f(\mathbf{x}) = 0$ and $\nabla^2 f(\mathbf{x}) \succeq 0$.

Definition 1 (Optimality Condition)**.** Suppose $\mathbf{x}$ is a **local minimum** of $f(\mathbf{x})$, then we have

$$\nabla f(\mathbf{x}) = 0, \quad \nabla^2 f(\mathbf{x}) \succeq 0.$$

Intuitively, if one of these conditions is violated, then it is possible to find a direction that decreases the function value. (Ge et al., 2015) characterized the following *strict-saddle* property, which is a quantitative version of the optimality conditions, and can lead to efficient algorithms to find local minima.

Definition 2. We say function $f(\cdot)$ is (θ, γ, ζ)-**strict saddle**. That is, for any $\mathbf{x}$, at least one of followings holds:

1. $\|\nabla f(\mathbf{x})\| \geq \theta$.
2. $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \leq -\gamma$.
3. $\mathbf{x}$ is ζ-close to $\mathcal{X}^\star$ – the set of local minima.

Intuitively, this definition says for any point $\mathbf{x}$, it either violates one of the optimality conditions significantly (first two cases), or is close to a local minima. Note that ζ and θ are often closely related. For a function with strict-saddle property, it is possible to efficiently find a point near a local minimum.

Local vs. Global However, of course finding a local minimum is not sufficient in many case. In this paper we are also going to prove that all local minima are also globally optimal, and they correspond to the desired solutions.

3. Low Rank Problems and Our Results

In this section we introduce matrix sensing, matrix completion and robust PCA. For each problem we give the results obtained by our framework. The proof ideas are illustrated later in Sections 4 and 5.

3.1. Matrix Sensing

Matrix sensing (Recht et al., 2010) is a generalization of compressed sensing (Candes et al., 2006). In the matrix sensing problem, there is an unknown low rank matrix $\mathbf{M}^\star \in \mathbb{R}^{d_1 \times d_2}$. We make linear observations on this matrix: let $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_m \in \mathbb{R}^{d_1 \times d_2}$ be m sensing matrices, the algorithm is given $\{\mathbf{A}_i\}$'s and the corresponding $b_i = \langle \mathbf{A}_i, \mathbf{M}^\star\rangle$. The goal is now to find the unknown matrix $\mathbf{M}^\star$. In order to find $\mathbf{M}^\star$, we need to solve the following nonconvex optimization problem

$$\min_{\mathbf{M}\in\mathbb{R}^{d_1\times d_2}, \text{rank}(\mathbf{M})=r} \quad f(\mathbf{M}) = \frac{1}{2m}\sum_{i=1}^m (\langle \mathbf{M}, \mathbf{A}_i\rangle - b_i)^2.$$

We can transform this constraint problem to an unconstraint problem by expressing $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$. We also need an additional regularizer (common for all asymmetric problems):

$$\min_{\mathbf{U},\mathbf{V}} \quad \frac{1}{2m}\sum_{i=1}^m (\langle \mathbf{U}\mathbf{V}^\top, \mathbf{A}_i\rangle - b_i)^2 + \frac{1}{8}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2. \tag{3}$$

The regularizer has been widely used in previous works (Zheng & Lafferty, 2016; Park et al., 2016). In Section 5 we show how this regularizer can be viewed as a way to deal with the additional invariants in asymmetric case, and reduce the asymmetric case to the symmetric case. A crucial concept in standard sensing literature is Restrict Isometry Property (RIP), which is defined as follows:

Definition 3. A group of sensing matrices $\{\mathbf{A}_1, .., \mathbf{A}_m\}$ satisfies the (r, δ)-RIP condition, if for every matrix $\mathbf{M}$ of rank at most r,

$$(1-\delta)\|\mathbf{M}\|_F^2 \leq \frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{A}_i, \mathbf{M}\rangle^2 \leq (1+\delta)\|\mathbf{M}\|_F^2.$$

Intuitively, RIP says operator $\frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{A}_i, \cdot\rangle^2$ approximately perserve norms for all low rank matrices. When the sensing matrices are chosen to be i.i.d. matrices with independent Gaussian entries, if $m \geq c(d_1+d_2)r$ for large enough constant c, the sensing matrices satisfy the $(2r, \frac{1}{20})$-RIP condition (Candes & Plan, 2011). Using our framework we can show:

Theorem 3. *When measurements $\{\mathbf{A}_i\}$ satisfy $(2r, \frac{1}{20})$-RIP, for matrix sensing objective (3) we have 1) all local minima satisfy $\mathbf{U}\mathbf{V}^\top = \mathbf{M}^\star$ 2) the function is $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon}{\sigma_r^\star}))$-strict saddle.*

This in particular says 1) no spurious local minima existsl; 2) whenever at some point $(\mathbf{U}, \mathbf{V})$ so that the gradient is small and the Hessian does not have significant negative eigenvalue, then the distance to global optimal (see Definition 6 and Definition 7) is guaranteed to be small. Such a point can be found efficiently (see supplementary material).

3.2. Matrix Completion

Matrix completion is a popular technique in recommendation systems and collaborative filtering (Koren, 2009; Rennie & Srebro, 2005). In this problem, again we have an unknown low rank matrix $\mathbf{M}^\star$. We observe each entry of the matrix $\mathbf{M}^\star$ independently with probability p. Let $\Omega \subset [d_1] \times [d_2]$ be a set of observed entries. For any matrix $\mathbf{M}$, we use $\mathbf{M}_\Omega$ to denote the matrix whose entries outside of Ω are set to 0. That is, $[\mathbf{M}_\Omega]_{i,j} = \mathbf{M}_{i,j}$ if $(i,j) \in \Omega$, and $[\mathbf{M}_\Omega]_{i,j} = 0$ otherwise. We further use $\|\mathbf{M}\|_\Omega$ to denote $\|\mathbf{M}_\Omega\|_F$. Matrix completion can be viewed as a special case of matrix sensing, where the sensing matrices only have one nonzero entry. However such matrices do not satisfy the RIP condition.

In order to solve matrix completion, we try to optimize the following:

$$\min_{\mathbf{M}\in\mathbb{R}^{d_1\times d_2}, \mathrm{rank}(\mathbf{M})=r} \quad \frac{1}{2p}\|\mathbf{M}-\mathbf{M}^\star\|_\Omega^2.$$

A well-known problem in matrix completion is that when the true matrix $\mathbf{M}^\star$ is very sparse, then we are very likely to observe only 0 entries, and has no chance to learn the other entries of $\mathbf{M}^\star$. To avoid this case, previous works have assumed following *incoherence* condition:

Definition 4. A rank r matrix $\mathbf{M} \in \mathbb{R}^{d_1\times d_2}$ is μ-incoherent, if for the rank-r SVD $\mathbf{X}\mathbf{D}\mathbf{Y}^\top$ of $\mathbf{M}$, we have for all $i \in [d_1], j \in [d_2]$

$$\|\mathbf{e}_i^\top \mathbf{X}\| \leq \sqrt{\mu r/d_2}, \quad \|\mathbf{e}_j^\top \mathbf{Y}\| \leq \sqrt{\mu r/d_1}.$$

We assume the unknown optimal low rank matrix $\mathbf{M}^\star$ is μ-incoherent.

In the non-convex program, we try to make sure the decomposition $\mathbf{U}\mathbf{V}^\top$ is also incoherent by adding a regularizer $Q(\mathbf{U}, \mathbf{V}) = \lambda_1 \sum_{i=1}^{d_1}(\|\mathbf{e}_i^\top \mathbf{U}\| - \alpha_1)_+^4 + \lambda_2 \sum_{j=1}^{d_2}(\|\mathbf{e}_j^\top \mathbf{V}\| - \alpha_2)_+^4$. Here $\lambda_1, \lambda_2, \alpha_1, \alpha_2$ are parameters that we choose later, $(x)_+ = \max\{x, 0\}$. Using this regularizer, we can now transform the objective function to the unconstraint form

$$\min_{\mathbf{U},\mathbf{V}} \quad \frac{1}{2p}\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\|_\Omega^2 + \frac{1}{8}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 + Q(\mathbf{U}, \mathbf{V}). \tag{4}$$

Using the framework, we can show following:

Theorem 4. *Let $d = \max\{d_1, d_2\}$, when sample rate $p \geq \Omega(\frac{\mu^4 r^6 (\kappa^\star)^6 \log d}{\min\{d_1, d_2\}})$, choose $\alpha_1^2 = \Theta(\frac{\mu r \sigma_1^\star}{d_1}), \alpha_2^2 = \Theta(\frac{\mu r \sigma_1^\star}{d_2})$ and $\lambda_1 = \Theta(\frac{d_1}{\mu r \kappa^\star}), \lambda_2 = \Theta(\frac{d_2}{\mu r \kappa^\star})$. With probability at least $1 - 1/poly(d)$, for Objective Function (4) we have 1) all local minima satisfy $\mathbf{U}\mathbf{V}^\top = \mathbf{M}^\star$ 2) The objective is $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon}{\sigma_r^\star}))$-strict saddle for polynomially small ϵ.*

3.3. Robust PCA

Robust PCA (Candès et al., 2011) is a generalization to the standard Principled Component Analysis. In Robust PCA, we are given an observation matrix $\mathbf{M}_o$, which is an true underlying matrix $\mathbf{M}^\star$ corrupted by a sparse noise $\mathbf{S}^\star$ ($\mathbf{M}_o = \mathbf{M}^\star + \mathbf{S}^\star$). In some sense the goal is to decompose the matrix $\mathbf{M}$ into these two components. There are many models on how many entries can be perturbed, and how they are distributed. In this paper we work in the setting where $\mathbf{M}^\star$ is μ-incoherent, and the rows/columns of $\mathbf{S}^\star$ can have at most α-fraction non-zero entries.

In order to express robust PCA as an optimization problem, we need constraints on both $\mathbf{M}$ and $\mathbf{S}$:

$$\min \ \frac{1}{2}\|\mathbf{M}+\mathbf{S}-\mathbf{M}_o\|_F^2. \quad s.t. \ \mathrm{rank}(\mathbf{M}) \leq r, \mathbf{S} \text{ is sparse.} \tag{5}$$

There can be several ways to specify the sparsity of $\mathbf{S}$. In this paper we restrict attention to the set $\mathcal{S}_\alpha$ which is the set of matrices that have at most α-fraction non-zero entries in each column/row, and entries have absolute value at most $2\frac{\mu r\sigma_1^\star}{\sqrt{d_1 d_2}}$.

Assuming the true sparse matrix $\mathbf{S}^\star$ is in $\mathcal{S}_\alpha$. Note that the infinite norm requirement on $\mathbf{S}^\star$ is without loss of generality, because by incoherence $\mathbf{M}^\star$ cannot have entries with absolute value more than $\frac{\mu r\sigma_1^\star}{\sqrt{d_1 d_2}}$. Any entry larger than that is obviously in the support of $\mathbf{S}^\star$ and can be truncated.

In objective function, we allow $\mathbf{S}$ to be γ times denser (in $\mathcal{S}_{\gamma\alpha}$) where γ is a parameter we choose later. Now the constraint optimization problem can be tranformed to the unconstraint problem

$$\min_{\mathbf{U},\mathbf{V}} \ f(\mathbf{U},\mathbf{V}) + \frac{1}{8}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2, \tag{6}$$
$$f(\mathbf{U},\mathbf{V}) := \min_{\mathbf{S}\in\mathcal{S}_{\gamma\alpha}} \frac{1}{2}\|\mathbf{U}\mathbf{V}^\top + \mathbf{S} - \mathbf{M}_o\|_F^2.$$

Of course, we can also think of this as a joint minimization problem of $\mathbf{U},\mathbf{V},\mathbf{S}$. However we choose to present it this way in order to allow extension of the strict-saddle condition. Since $f(\mathbf{U},\mathbf{V})$ is not twice-differetiable w.r.t $\mathbf{U},\mathbf{V}$, it does not admit Hessian matrix, so we use the following generalized version of strict-saddle

Definition 5. We say function $f(\cdot)$ is (θ,γ,ζ)-**pseudo strict saddle** if for any $\mathbf{x}$, at least one of followings holds:

1. $\|\nabla f(\mathbf{x})\| \geq \theta$.
2. $\exists g_{\mathbf{x}}(\cdot)$ so that $\forall \mathbf{y}, g_{\mathbf{x}}(\mathbf{y}) \geq f(\mathbf{y})$; $g_{\mathbf{x}}(\mathbf{x}) = f(\mathbf{x})$; $\lambda_{\min}(\nabla^2 g_{\mathbf{x}}(\mathbf{x})) \leq -\gamma$.
3. $\mathbf{x}$ is ζ-close to $\mathcal{X}^\star$ – the set of local minima.

Note that in this definition, the upperbound in 2 can be viewed as similar to the idea of subgradient. For functions with non-differentiable points, subgradient is defined so that it still offers a lowerbound for the function. In our case this is very similar – although Hessian is not defined, we can use a smooth function that upperbounds the current function (upper-bound is required for minimization). In the case of robust PCA the upperbound is obtained by a fixed $\mathbf{S}$. Using this formalization we can prove

Theorem 5. *There is an absolute constant $c > 0$, if $\gamma > c$, and $\gamma\alpha \cdot \mu r \cdot (\kappa^\star)^5 \leq \frac{1}{c}$ holds, for objective function Eq.(6) we have 1) all local minima satisfies $\mathbf{U}\mathbf{V}^\top = \mathbf{M}^\star$; 2) objective function is $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon\sqrt{\kappa^\star}}{\sigma_r^\star}))$-pseudo strict saddle for polynomially small ϵ.*

4. Framework for Symmetric Positive Definite Problems

In this section we describe our framework in the simpler setting where the desired matrix is positive semidefinite. In particular, suppose the true matrix $\mathbf{M}^\star$ we are looking for can be written as $\mathbf{M}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\top$ where $\mathbf{U}^\star \in \mathbb{R}^{d\times r}$. For objective functions that is quadratic over $\mathbf{M}$, we denote its Hessian as $\mathcal{H}$ and we can write the objective as

$$\min_{\mathbf{M}\in\mathbb{R}^{d\times d}_{\text{sym}},\text{rank}(\mathbf{M})=r} \ \frac{1}{2}(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H} : (\mathbf{M}-\mathbf{M}^\star), \tag{7}$$

We call this objective function $f(\mathbf{M})$. Via Burer-Monteiro factorization, the corresponding unconstraint optimization problem, with regularization Q can be written as

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times r}} \frac{1}{2}(\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star) : \mathcal{H} : (\mathbf{U}\mathbf{U}^\top - \mathbf{M}^\star) + Q(\mathbf{U}). \tag{8}$$

In this section, we also denote $f(\mathbf{U})$ as objective function with respect to parameter $\mathbf{U}$, abuse the notation of $f(\mathbf{M})$ previously defined over $\mathbf{M}$.

Direction of Improvement The optimality condition (Definition 1) implies if the gradient is non-zero, or if we can find a negative direction of the Hessian (that is a direction $\mathbf{v}$, so that $\mathbf{v}^\top\nabla^2 f(\mathbf{x})\mathbf{v} < 0$), then the point is not a local minimum. A common technique in characterizing the optimization landscape is therefore trying to explicitly find this negative direction. We call this the direction of improvement. Different works (Bhojanapalli et al., 2016; Ge et al., 2016) have chosen very different directions of improvement.

In our framework, we show it suffices to choose a single direction Δ as the direction of improvement. Intuitively, this direction should bring us close to the true solution $\mathbf{U}^\star$ from the current point $\mathbf{U}$. Due to rotational symmetry ($\mathbf{U}$ and $\mathbf{U}\mathbf{R}$ behave the same for the objective if $\mathbf{R}$ is a rotation matrix), we need to carefully define the difference between $\mathbf{U}$ and $\mathbf{U}^\star$.

Definition 6. Given matrices $\mathbf{U},\mathbf{U}^\star \in \mathbb{R}^{d\times r}$, define their difference $\Delta = \mathbf{U} - \mathbf{U}^\star\mathbf{R}$, where $\mathbf{R} \in \mathbb{R}^{r\times r}$ is chosen as $\mathbf{R} = \text{argmin}_{\mathbf{Z}^\top\mathbf{Z}=\mathbf{Z}\mathbf{Z}^\top=\mathbf{I}} \|\mathbf{U} - \mathbf{U}^\star\mathbf{Z}\|_F^2$.

Note that this definition tries to "align" $\mathbf{U}$ and $\mathbf{U}^\star$ before taking their difference, and therefore is invariant under rotations. In particular, this definition has the nice property that as long as $\mathbf{M} = \mathbf{U}\mathbf{U}^\top$ is close to $\mathbf{M}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\top$, we have Δ is small (we defer the proof to Appendix):

Lemma 6. *Given matrices $\mathbf{U},\mathbf{U}^\star \in \mathbb{R}^{d\times r}$, let $\mathbf{M} = \mathbf{U}\mathbf{U}^\top$ and $\mathbf{M}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\top$, and let Δ be defined as in Definition 6, then we have $\|\Delta\Delta^\top\|_F^2 \leq 2\|\mathbf{M}-\mathbf{M}^\star\|_F^2$, and $\sigma_r^\star\|\Delta\|_F^2 \leq \frac{1}{2(\sqrt{2}-1)}\|\mathbf{M}-\mathbf{M}^\star\|_F^2$.*

Now we can state the main Lemma:

Lemma 7 (Main). *For the objective (8), let Δ be defined as in Definition 6 and $\mathbf{M} = \mathbf{U}\mathbf{U}^\top$. Then, for any $\mathbf{U} \in \mathbb{R}^{d\times r}$, we have*

$$\begin{aligned}&\Delta : \nabla^2 f(\mathbf{U}) : \Delta = \Delta\Delta^\top : \mathcal{H} : \Delta\Delta^\top \\ &- 3(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H} : (\mathbf{M}-\mathbf{M}^\star) \\ &+ 4\langle \nabla f(\mathbf{U}), \Delta\rangle + [\Delta : \nabla^2 Q(\mathbf{U}) : \Delta - 4\langle \nabla Q(\mathbf{U}), \Delta\rangle]\end{aligned} \tag{9}$$

To see why this lemma is useful, let us look at the simplest case where $Q(\mathbf{U}) = 0$ and $\mathcal{H}$ is identity. In this case, if gradient is zero, by Eq. (9)

$$\Delta : \nabla^2 f(\mathbf{U}) : \Delta = \|\Delta\Delta\|_F^2 - 3\|\mathbf{M}-\mathbf{M}^\star\|_F^2$$

By Lemma 6 this is no more than $-\|\mathbf{M}-\mathbf{M}^\star\|_F^2$. Therefore, all stationary point with $\mathbf{M} \neq \mathbf{M}^*$ must be saddle points, and we immediately conclude all local minimum satisfies $\mathbf{U}\mathbf{U}^\top = \mathbf{M}^\star$!

Interaction with Regularizer For problems such as matrix completion, the Hessian $\mathcal{H}$ does not preserve the norm for all low rank matrices. In these cases we need to use additional regularizer. In particular, conceptually we need the following steps:

1. Show that the regularizer Q ensures for any $\mathbf{U}$ such that $\nabla f(\mathbf{U}) = 0$, $\mathbf{U} \in \mathcal{B}$ for some set $\mathcal{B}$.
2. Show that whenever $\mathbf{U} \in \mathcal{B}$, the Hessian operator $\mathcal{H}$ behaves similarly as identity: for some $c > 0$ we have: $\Delta\Delta^\top : \mathcal{H} : \Delta\Delta^\top - 3(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H} : (\mathbf{M}-\mathbf{M}^\star) < -c\|\Delta\|_F^2$.
3. Show that the regularizer does not contribute a large positive term to $\Delta : \nabla^2 f(\mathbf{U}) : \Delta$. This means we show an upperbound for $4\langle \nabla f(\mathbf{U}), \Delta\rangle + [\Delta : \nabla^2 Q(\mathbf{U}) : \Delta - 4\langle \nabla Q(\mathbf{U}), \Delta\rangle]$.

Interestingly, these steps are not just useful for handling regularizers. Any deviation to the original model (such as noise, or if the optimal matrix is not exactly low rank) can be viewed as an additional "regularizer" function $Q(\mathbf{U})$ and argued in the same framework. See supplementary material for more details.

4.1. Matrix Sensing

Matrix sensing is the ideal setting for this framework. For symmetric matrix sensing, the objective function is

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r}} \frac{1}{2m}\sum_{i=1}^m (\langle \mathbf{A}_i, \mathbf{U}\mathbf{U}^\top\rangle - b_i)^2. \tag{10}$$

Recall that matrices $\{\mathbf{A}_i : i = 1, 2, ..., m\}$ are known sensing matrices, and $b_i = \langle \mathbf{A}_i, \mathbf{M}^\star\rangle$ is the result of i-th observation. The intended solution is the unknown low rank matrix $\mathbf{M}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\top$. For any low rank matrix $\mathbf{M}$, the Hessian operator satisfies

$$\mathbf{M} : \mathcal{H} : \mathbf{M} = \sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{M}\rangle^2.$$

Therefore if the sensing matrices satisfy the RIP property (Definition 3), the Hessian operator is close to identity for all low rank matrices! In the symmetric case there is no regularizer, so the landscape for symmetric matrix sensing follows immediately from our main Lemma 7.

Theorem 8. *When measurement $\{\mathbf{A}_i\}$ satisfies $(2r, \frac{1}{10})$-RIP, for matrix sensing objective (10) we have 1) all local minima $\mathbf{U}$ satisfy $\mathbf{U}\mathbf{U}^\top = \mathbf{M}^\star$; 2) the function is $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon}{\sigma_r^\star}))$-strict saddle.*

Proof. For point $\mathbf{U}$ with small gradient satisfying $\|\nabla f(\mathbf{U})\|_{\mathrm{F}} \le \epsilon$, by $(2r, \delta_{2r})$-RIP property:

$$\begin{aligned}\Delta : \nabla^2 f(\mathbf{U}) : \Delta \le &(1+\delta_{2r})\|\Delta\Delta^\top\|_{\mathrm{F}}^2 \\ &- 3(1-\delta_{2r})\|\mathbf{M}-\mathbf{M}^\star\|_{\mathrm{F}}^2 + 4\epsilon\|\Delta\|_{\mathrm{F}} \\ \le &-(1-5\delta_{2r})\|\mathbf{M}-\mathbf{M}^\star\|_{\mathrm{F}}^2 + 4\epsilon\|\Delta\|_{\mathrm{F}} \\ \le &-0.4\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2 + 4\epsilon\|\Delta\|_{\mathrm{F}}\end{aligned}$$

The second last inequality is due to Lemma 6 that $\|\Delta\Delta^\top\|_{\mathrm{F}}^2 \le 2\|\mathbf{M}-\mathbf{M}^\star\|_{\mathrm{F}}^2$, and last inequality is due to $\delta_{2r} = \frac{1}{10}$ and second part of Lemma 6. This means if $\mathbf{U}$ is not close to $\mathbf{U}^\star$, that is, if $\|\Delta\|_{\mathrm{F}} \ge \frac{20\epsilon}{\sigma_r^\star}$, we have $\Delta : \nabla^2 f(\mathbf{U}) : \Delta \le -0.2\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2$. This proves $(\epsilon, 0.2\sigma_r^\star, \frac{20\epsilon}{\sigma_r^\star})$-strict saddle property. Take $\epsilon = 0$, we know all stationary points with $\|\Delta\|_{\mathrm{F}} \neq 0$ are saddle points. This means all local minima are global minima (satisfying $\mathbf{U}\mathbf{U}^\top = \mathbf{M}^\star$), which finishes the proof. □

4.2. Matrix Completion

For matrix completion, we need to ensure the incoherence condition (Definition 4). In order to do that, we add a regularizer $Q(\mathbf{U})$ that penalize the objective function when some row of $\mathbf{U}$ is too large. We choose the same regularizer as (Ge et al., 2016): $Q(\mathbf{U}) = \lambda\sum_{i=1}^d (\|\mathbf{U}_i\| - \alpha)_+^4$. The objective is then

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r}} \frac{1}{2p}\|\mathbf{M}^\star - \mathbf{U}\mathbf{U}^\top\|_\Omega^2 + Q(\mathbf{U}). \tag{11}$$

Using our framework, we first need to show that the regularizer ensures all rows of $\mathbf{U}$ are small (step 1).

Lemma 9. *There exists an absolute constant c, when sample rate $p \ge \Omega(\frac{\mu r}{d}\log d)$, $\alpha^2 = \Theta(\frac{\mu r\sigma_1^\star}{d})$ and $\lambda =$*

$\Theta(\frac{d}{\mu r\kappa^\star})$, *we have for any points* $\mathbf{U}$ *with* $\|\nabla f(\mathbf{U})\|_F \le \epsilon$ *for polynomially small* ϵ*, with probability at least* $1 - 1/poly(d)$*:*

$$\max_i \|\mathbf{e}_i^\top \mathbf{U}\|^2 \le O\left(\frac{(\mu r)^{1.5}\kappa^\star\sigma_1^\star}{d}\right)$$

This is a slightly stronger version of Lemma 4.7 in (Ge et al., 2016). Next we show under this regularizer, we can still select the direction Δ, and the first part of Equation (9) is significantly negative when Δ is large (step 2):

Lemma 10. *When sample rate* $p \ge \Omega(\frac{\mu^3 r^4(\kappa^\star)^4 \log d}{d})$*, by choosing* $\alpha^2 = \Theta(\frac{\mu r\sigma_1^\star}{d})$ *and* $\lambda = \Theta(\frac{d}{\mu r\kappa^\star})$ *with probability at least* $1 - 1/poly(d)$*, for all* $\mathbf{U}$ *with* $\|\nabla f(\mathbf{U})\|_F \le \epsilon$ *for polynomially small* ϵ *we have*

$$\Delta\Delta^\top : \mathcal{H} : \Delta\Delta^\top - 3(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H} : (\mathbf{M}-\mathbf{M}^\star) \le -0.3\sigma_r^\star\|\Delta\|_F^2$$

This lemma follows from several standard concentration inequalities, and is made possible because of the incoherence bound we proved in the previous lemma.

Finally we show the additional regularizer related term in Equation (9) is bounded (step 3).

Lemma 11. *By choosing* $\alpha^2 = \Theta(\frac{\mu r\sigma_1^\star}{d})$ *and* $\lambda\alpha^2 \le O(\sigma_r^\star)$*, we have:*

$$\frac{1}{4}[\Delta : \nabla^2 Q(\mathbf{U}) : \Delta - 4\langle\nabla Q(\mathbf{U}), \Delta\rangle] \le 0.1\sigma_r^\star\|\Delta\|_F^2$$

Combining these three lemmas, it is easy to see

Theorem 12. *When sample rate* $p \ge \Omega(\frac{\mu^3 r^4(\kappa^\star)^4 \log d}{d})$*, by choosing* $\alpha^2 = \Theta(\frac{\mu r\sigma_1^\star}{d})$ *and* $\lambda = \Theta(\frac{d}{\mu r\kappa^\star})$*. Then with probability at least* $1 - 1/poly(d)$*, for matrix completion objective* (11) *we have 1) all local minima satisfy* $\mathbf{U}\mathbf{U}^\top = \mathbf{M}^\star$ *2) the function is* $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon}{\sigma_r^\star}))$*-strict saddle for polynomially small* ϵ*.*

Notice that our proof is different from (Ge et al., 2016), as we focus on the direction Δ for both first and second order conditions while they need to select different directions for the Hessian. The framework allowed us to get a simpler proof, generalize to asymmetric case and also improved the dependencies on rank.

4.3. Robust PCA

In the robust PCA problem, for any given matrix $\mathbf{M}$ the objective function try to find the optimal sparse perturbation $\mathbf{S}$. In the symmetric PSD case, recall we observe $\mathbf{M}_o = \mathbf{M}^\star + \mathbf{S}^\star$, we define the set $\mathbf{S}_\alpha$ to be the set of matrices whose rows/columns have at most α-fraction nonzero entries, and entries are bounded by $2\frac{\mu r\sigma_1^\star}{d}$. Note the projection onto set $\mathcal{S}_\alpha$ be computed in polynomial time (using a max flow algorithm).

We assume $\mathbf{S}^\star \in \mathcal{S}_\alpha$, the objective can be written as

$$\min_{\mathbf{U}} f(\mathbf{U}), \text{where} f(\mathbf{U}) := \min_{\mathbf{S}\in\mathcal{S}_{\gamma\alpha}} \frac{1}{2}\|\mathbf{U}\mathbf{U}^\top + \mathbf{S} - \mathbf{M}_o\|_F^2. \tag{12}$$

Here γ is a slack parameter that we choose later.

Note that now the objective function $f(\mathbf{U})$ is not quadratic, so we cannot use the framework directly. However, if we fix $\mathbf{S}$, then $f_{\mathbf{S}}(\mathbf{U}) := \frac{1}{2}\|\mathbf{U}\mathbf{U}^\top + \mathbf{S} - \mathbf{M}_o\|_F^2$ is a quadratic function with Hessian equal to identity. We can still apply our framework to this function. In this case, since the Hessian is identity for all matrices, we can skip the first step. The problem becomes a matrix factorization problem:

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\mathbf{A} - \mathbf{U}\mathbf{U}^\top\|_F^2. \tag{13}$$

The difference here is that the matrix $\mathbf{A}$ (which is $\mathbf{M}^\star + \mathbf{S}^\star - \mathbf{S}$) is not equal to $\mathbf{M}^\star$ and is in general not low rank. We can use the framework to analyze this problem (and treat the residue $\mathbf{A} - \mathbf{M}^\star$ as the "regularizer" $Q(\mathbf{U})$).

Lemma 13. *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be a symmetric PSD matrix, and matrix factorization objective to be:*

$$f(\mathbf{U}) = \|\mathbf{U}\mathbf{U}^\top - \mathbf{A}\|_F^2$$

where $\sigma_r(\mathbf{A}) \ge 15\sigma_{r+1}(\mathbf{A})$*. then 1) all local minima satisfies* $\mathbf{U}\mathbf{U}^\top = \mathcal{P}_r(\mathbf{A})$ *(best rank-r approximation), 2) objective is* $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon}{\sigma_r^\star}))$*-strict saddle.*

To deal with the case $\mathbf{S}$ not fixed (but as minimizer of Eq.(12)), we let $\mathbf{U}^\dagger(\mathbf{U}^\dagger)^\top$ be the best rank r-approximation of $\mathbf{M}^\star + \mathbf{S}^\star - \mathbf{S}$. The next lemma shows when $\mathbf{U}$ is close to $\mathbf{U}^\dagger$ up to some rotation, $\mathbf{U}$ will actually be already close to $\mathbf{U}^\star$ up to some rotation.

Lemma 14. *There is an absolute constant c, assume* $\gamma > c$*, and* $\gamma\alpha \cdot \mu r \cdot (\kappa^\star)^5 \le \frac{1}{c}$*. Let* $\mathbf{U}^\dagger(\mathbf{U}^\dagger)^\top$ *be the best rank r-approximation of* $\mathbf{M}^\star + \mathbf{S}^\star - \mathbf{S}$*, where* $\mathbf{S}$ *is the minimizer as in Eq.*(12)*. Assume* $\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{R}\mathbf{R}^\top=\mathbf{I}} \|\mathbf{U} - \mathbf{U}^\dagger\mathbf{R}\|_F \le \epsilon$*. Let* Δ *be defined as in Definition 6, then* $\|\Delta\|_F \le O(\epsilon\sqrt{\kappa^\star})$ *for polynomially small* ϵ*.*

The proof of Lemma 14 is inspired by Yi et al. (2016) and uses the property of the optimally chosen sparse set $\mathbf{S}$. Combining these two lemmas we get our main result:

Theorem 15. *There is an absolute constant c, if* $\gamma > c$*, and* $\gamma\alpha \cdot \mu r \cdot (\kappa^\star)^5 \le \frac{1}{c}$ *holds, for objective function Eq.*(12) *we have 1) all local minima satisfies* $\mathbf{U}\mathbf{U}^\top = \mathbf{M}^\star$*; 2) objective function is* $(\epsilon, \Omega(\sigma_r^\star), O(\frac{\epsilon\sqrt{\kappa^\star}}{\sigma_r^\star}))$*-pseudo strict saddle for polynomially small* ϵ*.*

5. Handling Asymmetric Matrices

In this section we show how to reduce problems on asymmetric matrices to problems on symmetric PSD matrices.

Let $\mathbf{M}^\star = \mathbf{U}^\star\mathbf{V}^{\star\top}$, and $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, and objective function:

$$f(\mathbf{U},\mathbf{V}) = 2(\mathbf{M}-\mathbf{M}^\star) : \mathcal{H}_0 : (\mathbf{M}-\mathbf{M}^\star) + \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2 + Q_0(\mathbf{U},\mathbf{V}) \quad (14)$$

Note this is a scaled version of objectives introduced in Sec.3 (multiplied by 4), and scaling will not change the property of local minima, global minima and saddle points.

We view the problem as if it is trying to find a $(d_1+d_2)\times r$ matrix, whose first d_1 rows are equal to $\mathbf{U}$, and last d_2 rows are equal to $\mathbf{V}$.

Definition 7. Suppose $\mathbf{M}^\star$ is the optimal solution, and its SVD is $\mathbf{X}^\star\mathbf{D}^\star\mathbf{Y}^{\star\top}$. Let $\mathbf{U}^\star = \mathbf{X}^\star(\mathbf{D}^\star)^{\frac{1}{2}}$, $\mathbf{V}^\star = \mathbf{Y}^\star(\mathbf{D}^\star)^{\frac{1}{2}}$, $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$ is the current point, we reduce the problem into a symmetric case using following notations.

$$\mathbf{W} = \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix}, \mathbf{W}^\star = \begin{pmatrix}\mathbf{U}^\star\\ \mathbf{V}^\star\end{pmatrix}, \mathbf{N} = \mathbf{W}\mathbf{W}^\top, \mathbf{N}^\star = \mathbf{W}^\star\mathbf{W}^{\star\top} \quad (15)$$

Further, Δ is defined to be the difference between $\mathbf{W}$ and $\mathbf{W}^\star$ up to rotation as in Definition 6.

We will also transform the Hessian operators to operate on $(d_1+d_2)\times r$ matrices. In particular, define Hessian $\mathcal{H}_1, \mathcal{G}$ such that for all $\mathbf{W}$ we have:

$$\mathbf{N} : \mathcal{H}_1 : \mathbf{N} = \mathbf{M} : \mathcal{H}_0 : \mathbf{M}$$
$$\mathbf{N} : \mathcal{G} : \mathbf{N} = \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2$$

Now, let $Q(\mathbf{W}) = Q(\mathbf{U},\mathbf{V})$, and we can rewrite the objective function $f(\mathbf{W})$ as

$$\frac{1}{2}\left[(\mathbf{N}-\mathbf{N}^\star) : 4\mathcal{H}_1 : (\mathbf{N}-\mathbf{N}^\star) + \mathbf{N} : \mathcal{G} : \mathbf{N}\right] + Q(\mathbf{W}) \quad (16)$$

We know $\mathcal{H}_0$ perserves the norm of low rank matrices $\mathbf{M}$. To reduce asymmetric problems to symmetric problem, intuitively, we also hope $\mathcal{H}_0$ to approximately preserve the norm of $\mathbf{N}$. However this is impossible as by definition, $\mathcal{H}_0$ only acts on $\mathbf{M}$, which is the *off-diagonal* blocks of $\mathbf{N}$. We can expect $\mathbf{N} : \mathcal{H}_0 : \mathbf{N}$ to be close to the norm of $\mathbf{U}\mathbf{V}^\top$, but for all matrices $\mathbf{U},\mathbf{V}$ with the same $\mathbf{U}\mathbf{V}^\top$, the matrix $\mathbf{N}$ can have very different norms. The easiest example is to consider $\mathbf{U} = \mathrm{diag}(1/\epsilon,\epsilon)$ and $\mathbf{V} = \mathrm{diag}(\epsilon,1/\epsilon)$: while $\mathbf{U}\mathbf{V}^\top = \mathbf{I}$ no matter what ϵ is, the norm of $\mathbf{N}$ is of order $1/\epsilon^2$ and can change drastically. The regularizer is exactly there to handle this case: the Hessian $\mathcal{G}$ of the regularizer will be related to the norm of the diagonal components, therefore allowing the full Hessian $\mathcal{H} = 4\mathcal{H}_1 + \mathcal{G}$ to still be approximately identity.

Now we can formalize the reduction as the following main Lemma:

Lemma 16. *For the objective (16), let $\Delta, \mathbf{N}, \mathbf{N}^\star$ be defined as in Definition 7. Then, for any $\mathbf{W} \in \mathbb{R}^{(d_1+d_2)\times r}$, we have*

$$\Delta : \nabla^2 f(\mathbf{W}) : \Delta \le \Delta\Delta^\top : \mathcal{H} : \Delta\Delta^\top - 3(\mathbf{N}-\mathbf{N}^\star) : \mathcal{H} : (\mathbf{N}-\mathbf{N}^\star) + 4\langle\nabla f(\mathbf{W}),\Delta\rangle + [\Delta : \nabla^2 Q(\mathbf{W}) : \Delta - 4\langle\nabla Q(\mathbf{W}),\Delta\rangle] \quad (17)$$

where $\mathcal{H} = 4\mathcal{H}_1 + \mathcal{G}$. Further, if $\mathcal{H}_0$ satisfies $\mathbf{M} : \mathcal{H}_0 : \mathbf{M} \in (1\pm\delta)\|\mathbf{M}\|_F^2$ for some matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, let $\mathbf{W}$ and $\mathbf{N}$ be defined as in (15), then $\mathbf{N} : \mathcal{H} : \mathbf{N} \in (1\pm 2\delta)\|\mathbf{N}\|_F^2$.

Intuitively, this lemma shows the same direction of improvement works as before, and the regularizer is exactly what it requires to maintain the norm-preserving property of the Hessian.

The proofs are deferred to supplementary material.

6. Conclusions

In this paper we give a framework that explains the recent success in understanding optimization landscape for low rank matrix problems. Our framework connects and simplifies the existing proofs, and generalizes to new settings such as asymmetric matrix completion and robust PCA. The key observation is when the Hessian operator preserves the norm of certain matrices, one can use the same directions of improvement to prove similar optimization landscape. We show the regularizer $\frac{1}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2$ is exactly what it requires to maintain this norm preserving property in the asymmetric case.Our analysis also allows the interaction between regularizer and Hessian to handle difficult settings such as.

For low rank matrix problems, there are generalizations such as weighted matrix factorization(Li et al., 2016) and 1-bit matrix sensing(Davenport et al., 2014) where the Hessian operator may behave differently as the settings we can analyze. How to characterize the optimization landscape in these settings is still an open problem.

In order to get general ways of understanding optimization landscapes for more generally, there are still many open problems. In particular, how can we decide whether two problems are similar enough to share the same optimization landscape? A minimum requirement is that the non-convex problem should have the same *symmetry* structure – the set of equivalent global optimum should be the same. In this work, we show if the problems come from convex objective functions with similar Hessian properties, then they have the same optimization landscape. We hope this serves as a first step towards general tools for understanding optimization landscape for groups of problems.

References

Agarwal, Naman, Allen-Zhu, Zeyuan, Bullins, Brian, Hazan, Elad, and Ma, Tengyu. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Belkin, Mikhail, Rademacher, Luis, and Voss, James. Basis learning as an algorithmic primitive. *arXiv preprint arXiv:1411.1420*, 2014.

Bengio, Yoshua. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

Bhojanapalli, Srinadh, Kyrillidis, Anastasios, and Sanghavi, Sujay. Dropping convexity for faster semi-definite optimization. *arXiv:1509.03917*, 2015.

Bhojanapalli, Srinadh, Neyshabur, Behnam, and Srebro, Nathan. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

Burer, Samuel and Monteiro, Renato DC. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Candes, Emmanuel J and Plan, Yaniv. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Candès, Emmanuel J and Recht, Benjamin. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Candes, Emmanuel J, Romberg, Justin K, and Tao, Terence. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

Candès, Emmanuel J, Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Carmon, Yair, Duchi, John C, Hinder, Oliver, and Sidford, Aaron. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

Chen, Yudong and Wainwright, Martin J. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Davenport, Mark A and Romberg, Justin. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

Davenport, Mark A, Plan, Yaniv, van den Berg, Ewout, and Wootters, Mary. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

Fazel, Maryam. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

Ge, Rong, Huang, Furong, Jin, Chi, and Yuan, Yang. Escaping from saddle points—online stochastic gradient for tensor decomposition. *arXiv:1503.02101*, 2015.

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

Hardt, Moritz. Understanding alternating minimization for matrix completion. In *FOCS 2014*. IEEE, 2014.

Hardt, Moritz and Wootters, Mary. Fast matrix completion without the condition number. In *COLT 2014*, pp. 638–678, 2014.

Hotelling, Harold. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Jin, Chi, Ge, Rong, Netrapalli, Praneeth, Kakade, Sham M, and Jordan, Michael I. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

Keshavan, Raghunandan H, Montanari, Andrea, and Oh, Sewoong. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010a.

Keshavan, Raghunandan H, Montanari, Andrea, and Oh, Sewoong. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 11:2057–2078, 2010b.

Koren, Yehuda. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 81, 2009.

Li, Qiuwei and Tang, Gongguo. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv preprint arXiv:1611.03060*, 2016.

Li, Yuanzhi, Liang, Yingyu, and Risteski, Andrej. Recovery guarantee of weighted low-rank approximation via alternating minimization. *arXiv preprint arXiv:1602.02262*, 2016.

Lin, Zhouchen, Chen, Minming, and Ma, Yi. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

Nesterov, Yurii and Polyak, Boris T. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Netrapalli, Praneeth, Niranjan, UN, Sanghavi, Sujay, Anandkumar, Animashree, and Jain, Prateek. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

Park, Dohyung, Kyrillidis, Anastasios, Caramanis, Constantine, and Sanghavi, Sujay. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

Recht, Benjamin, Fazel, Maryam, and Parrilo, Pablo A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

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

Sun, Ju, Qu, Qing, and Wright, John. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *arXiv:1511.03607*, 2015a.

Sun, Ju, Qu, Qing, and Wright, John. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015b.

Sun, Ruoyu and Luo, Zhi-Quan. Guaranteed matrix completion via nonconvex factorization. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pp. 270–289. IEEE, 2015.

Tu, Stephen, Boczar, Ross, Soltanolkotabi, Mahdi, and Recht, Benjamin. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Yi, Xinyang, Park, Dohyung, Chen, Yudong, and Caramanis, Constantine. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pp. 4152–4160, 2016.

Zhang, Xiao, Wang, Lingxiao, and Gu, Quanquan. A nonconvex free lunch for low-rank plus sparse matrix recovery. *arXiv preprint arXiv:1702.06525*, 2017.

Zhao, Tuo, Wang, Zhaoran, and Liu, Han. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pp. 559–567, 2015.

Zheng, Qinqing and Lafferty, John. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.