No Spurious Solutions in Non-convex Matrix Sensing: Structure Compensates for Isometry

Igor Molybog, Somayeh Sojoudi and Javad Lavaei

Abstract— **The paper is concerned with the theoretical explanation of the recent empirical success of solving the low-rank matrix sensing problem via nonconvex optimization. It is known that under an incoherence assumption (namely, RIP) on the sensing operator, the optimization problem has no spurious local minima. This assumption is too strong for real-world applications where the amount of data cannot be sufficiently high. We develop the notion of Kernel Structure Property (KSP), which can be used alone or combined with RIP in this context. KSP explains how the inherent structure of an operator contributes to the non-existence of spurious local minima. As a special case, we study sparse sensing operators that have a low-dimensional representation. Using KSP, we obtain novel necessary and sufficient conditions for no spurious solutions in matrix sensing and demonstrate their usefulness in analytical and numerical studies.**

I. Introduction

Even under the ideal condition of no noise and zero approximation error, many highly-efficient machine learning techniques involve solving potentially hard or intractable computational problems while learning from data. In practice, they are tackled by heuristic optimization algorithms, based on relaxations or greedy principals. The lack of guarantees on their performance limits their use in applications with significant cost of an error, impacting our ability to implement progressive data analysis techniques in crucial social and economic systems, such as healthcare, transportation, and energy production and distribution. Commonly, non-convexity is the main obstacle for a guaranteed learning of continuous parameters.

It is well known that many fundamental problems with a natural non-convex formulation can be $\mathcal{NP}$-hard [2]. Sophisticated techniques for addressing this issue, like generic convex relaxations, may require working in an unrealistically high-dimensional space to guarantee exactness of the solution. As a consequence of complicated geometrical structures, a non-convex function may contain an exponential number of saddle points and spurious local minima, and therefore local search algorithms may become trapped in such points. Nevertheless, empirical observations show positive results regarding the application of these approaches to several practically important instances. This has led to a large branch of research that aims to explain the success of experimental results in order to understand the boundaries of applicability of the existing algorithms and develop new ones. A recent direction in non-convex optimization consists in studying how simple algorithms can solve potentially hard problems arising in data analysis applications. The most commonly applied class of such algorithms is based on *local search*, which will be the focus of this work.

Consider searching over some given domain $\mathcal{X}$. For a twice continuously differentiable objective function $f : \mathcal{X} \to \mathbb{R}$ that attains its global infimum f^*, if the point x attains $f(x) = f^*$, then we call it a *global minimizer*. The point x is said to be a *local minimizer* if $f(x) \leq f(x')$ holds for all x' within a local neighborhood of x. If x is a local minimizer, then it must satisfy the first- and second-order *necessary* optimality conditions. Conversely, a point x satisfying only the first-order condition is called a *first-order critical point,* while a point satisfying both of the conditions is called a *second-order critical point*. We also call it a *solution*. We call a solution *spurious* if it is not a global minimum. In this work, we study how existence of a spurious solution depends on the size/volume of the domain as well as the underlying structure of the problem.

The analysis of the landscape of the objective function around a global optimum may lead to an optimality guarantee for local search algorithms initialized sufficiently close to the solution [3], [4], [5], [6], [7], [8]. Finding a good initialization scheme is highly problem-specific and difficult to generalize. Global analysis of the landscape is harder, but potentially more rewarding.

Both local and global convergence guarantees have been developed to justify the success of local search methods in various applications like dictionary learning [9], basic non-convex M-estimators [10], shallow [11] and deep [12] artificial neural networks with different activation [13] and loss [14] functions, phase retrieval [15], [16], [17] and more general matrix sensing problems [18], [19]. Particularly, significant progress has been made towards understanding different variants of *low-rank matrix recovery,* although explanations of the simplest version called *matrix sensing* are still under active development [20], [21], [22], [18], [23], [24]. Given a linear sensing operator $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ and a ground truth matrix $z \in \mathbb{R}^{n\times r}$ $(r < n)$, an instance of the rank-r matrix sensing problem consists in minimizing over $\mathbb{R}^{n\times r}$ the nonconvex function

$$f_{z,\mathcal{A}}(x) = \|\mathcal{A}(xx^T - zz^T)\|_2^2 = \|\mathcal{A}(xx^T) - b\|_2^2, \qquad (1)$$

where $b = \mathcal{A}(zz^T)$. Recent work has generally found a certain assumption on the sensing operator to be sufficient for the matrix sensing problem to be "computationally easy

University of California at Berkeley, CA 94720, USA `igormolybog@berkeley.edu`, `sojoudi@berkeley.edu`, `lavaei@berkeley.edu`

This work was supported by grants from ARO, AFOSR, ONR and NSF. All of the proofs can be found in the technical report [1].

to solve". Precisely, this assumption works with the notion of RIP.

Definition 1 (Restricted Isometry Property)**.** *The linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is said to satisfy δ_r-RIP for some constant $\delta_r \in [0, 1)$ if there is $\gamma > 0$ such that*

$$(1 - \delta_r)\|X\|_F^2 \leq \gamma\|\mathcal{A}(X)\|_2^2 \leq (1 + \delta_r)\|X\|_F^2$$

holds for all $X \in \mathbb{S}^n$ satisfying $\operatorname{rank}(X) \leq r$.

The existing results proving absence of spurious local minima using this notion (such as [25], [26], [27], [28], [29], [18], [30], [20]) are based on a norm-preserving argument: the problem turns out to be a low-dimensional embedding of a canonical problem known to contain no spurious local minima. While the approach is widely applicable in its scope, it requires fairly strong assumptions on the data. In contrast, [24], [31] introduced a technique to find a certificate to guarantee that any given point cannot be a spurious local minimum of the problem of minimizing $f_{z,\mathcal{A}}$ over $\mathbb{R}^{n\times r}$, where $z \in \mathbb{R}^{n\times r}$ and $\mathcal{A}$ satisfies δ_{2r}-RIP. Since $f_{z,\mathcal{A}}$ depends on z and $\mathcal{A}$, this introduces a class of optimization problems defined as

$$\left\{ \underset{x\in\mathbb{R}^{n\times r}}{\operatorname{minimize}} \; f_{z,\mathcal{A}}(x) \;\middle|\; \mathcal{A} \text{ satisfies } \delta_{2r}\text{-RIP, } z \in \mathbb{R}^{n\times r} \right\}. \tag{Problem$^{\text{RIP}}$}$$

(Problem$^{\text{RIP}}$) consists of infinitely many instances of an optimization problem, each corresponding to some point z in $\mathbb{R}^{n\times r}$ and some operator $\mathcal{A}$ satisfying δ_{2r}-RIP. The state-of-the-art results for (Problem$^{\text{RIP}}$) are stated below.

Theorem 1 ([28], [18], [31])**.** *The following statements hold:*

- *If $\delta_{2r} < 1/5$, no instance of* (Problem$^{\text{RIP}}$) *has a spurious second-order critical point.*
- *If $r = 1$ and $\delta_2 < 1/2$, then no instance of* (Problem$^{\text{RIP}}$) *has a spurious second-order critical point.*
- *If $r = 1$ and $\delta_2 \geq 1/2$, then there exists an instance of* (Problem$^{\text{RIP}}$) *with a spurious second-order critical point.*

Non-existence of a spurious second-order critical point effectively means that any algorithm that converges to a second-order critical point is guaranteed to recover zz^T exactly. Examples of such algorithms include variants of the stochastic gradient descent (SGD) that is known to avoid saddle or even spurious local minimum points under certain assumptions [32], and widely used in machine learning [33], [34]. Besides SGD, many local search methods have been shown to be convergent to a second-order critical point with high probability under mild conditions, including the classical gradient descent [35], alternating minimizations [36] and Newton's method [37]. In this paper, we present guarantees on the global optimality of the second-order critical points, which means that our results can be combined with any of the algorithms mentioned above to guarantee the global convergence.

Theorem 1 discloses the limits on the guarantees that the notion of RIP can provide. However, linear maps in applications related to physical systems, such as power system analysis, typically have no RIP constant smaller than 0.9, and yet the non-convex matrix sensing still manages to work on those instances. This gap between theory and practice motivates the following question.

What is the alternative property practical problems satisfy that makes them easy to solve via simple local search?

We address this problem by developing a theoretical framework that precisely characterizes when a structured linear map has no spurious solution. It allows us to relax the bounds on δ_{2r} in Theorem 1. More precisely, we obtain different theoretical bounds on the RIP constant δ_{2r} that guarantee the absence of a spurious second-order critical point by leveraging the underlying structure of a given mapping. Since the existing methods have not studied this fundamental problem, we compare our bounds with the baseline $\delta_{2r} < 1/2$, which is the best known bound in the literature. In Section II, we motivate the need for a new notion replacing or improving RIP with real-world examples. Section III introduces some formal definitions and develops a mathematical framework to analyze spurious solutions and relate them to the underlying sparsity and structure of the problem, using techniques in conic optimization. Sections IV and V give the theory behind this notion and examples of its application. In Section VI, we present numerical results of the application of the developed theory to a real-world problem appearing in power systems analysis. Concluding remarks are given in Section VII. Some of the proofs, technical details and lemmas can be found in the technical report [1].

Notation

$\mathbb{C}^n$, $\mathbb{R}^n$ and $\mathbb{R}^{n\times r}$ denote the sets of complex and real n-dimensional vectors, and $n \times r$ matrices, respectively. $\mathbb{S}^n$ denotes the set of $n \times n$ symmetric matrices. $\text{Tr}(A)$, $\|A\|_F$ and $\langle A, B\rangle$ are the trace of a square matrix A, its Frobenius norm, and the Frobenius inner product of matrices A and B of compatible sizes. For a square matrix A, we define the symetric part $\text{Sym}(A) = (A + A^T)/2$. For a symmetric matrix A, its null space is denoted with $Ker(A)$. For square matrices $A_1, A_2, \ldots, A_n$, the matrix $diag(A_1, \ldots, A_n)$ is block-diagonal, with A_i's on the block diagonal. The notation $A \circ B$ refers to the Hadamard (entrywise) multiplication, and $A \otimes B$ refers to the Kronecker product of matrices. The vectorization operator $\text{vec} : \mathbb{R}^{n\times r} \to \mathbb{R}^{nr}$ stacks the columns of a matrix into a vector. The *matricization* operator $\text{mat}(\cdot)$ is the inverse of $\text{vec}(\cdot)$. Let $\succeq$ denote the positive semidefinite sign.

For a linear operator $\mathcal{L} : \mathbb{R}^{n\times r} \to \mathbb{R}^m$, the adjoint operator is denoted by $\mathcal{L}^T : \mathbb{R}^m \to \mathbb{R}^{n\times r}$. The matrix $\mathbf{L} \in \mathbb{R}^{m\times nr}$ such that $\mathcal{L}(x) = \mathbf{L}\text{vec}(x)$ is called the *matrix representation* of the linear operator $\mathcal{L}$. Bold letters are reserved for matrix representations of corresponding linear operators.

Sparsity pattern S of a set of matrices $\mathrm{M} \subset \mathbb{R}^{m\times n}$ is a subset of $\{1, \ldots, \max\{n, m\}\}^2$ such that $(i, j) \in S$ if and only if there is $X \in \mathrm{M}$ with the property that $X_{ij} \neq 0$.

Given a sparsity pattern S, define its matrix representation $\mathbf{S} \in \mathbb{S}^{m \times n}$ as

$$S_{ij} = \begin{cases} 0 & \text{if } (i,j) \in S, \\ 1 & \text{if } (i,j) \notin S, \end{cases}$$

The *orthogonal basis* of a given $m \times n$ matrix A (with $m \geq n$) is a matrix $P = \operatorname{orth}(A) \in \mathbb{R}^{m \times rank(A)}$ consisting of $\operatorname{rank}(A)$ orthonormal columns that span $\operatorname{range}(A)$:

$$P = \operatorname{orth}(A) \quad \iff \quad PP^T A = A,\ P^T P = I_{\operatorname{rank}(A)}.$$

Positive part means $(\cdot)_+ = \max\{0, \cdot\}$, and eigenvalues in an arbitrary order are denoted by $\lambda_i(\cdot)$.

II. Motivating example

In this section, we motivate this work by offering a case study on data analytics for energy systems. The state of a power system can be modeled by a vector of complex voltages on the nodes (buses) of the network. Monitoring the state of a power system is obviously a necessary requirement for its efficient and safe operation. This crucial information should be inferred from some measurable parameters, such as the power that is generated and consumed at each bus or transmitted through a line. The power network can be modeled by a number of parameters grouped into the admittance matrix $Y \in \mathbb{C}^{n \times n}$. The state estimation problem consists in recovering the unknown voltage vector $v \in \mathbb{C}^n$ from the available measurements. In the noiseless scenario, these measurements are m real numbers of the form

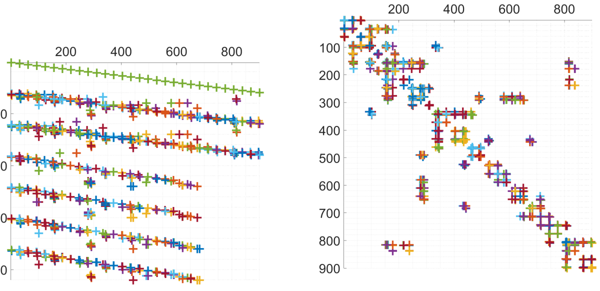
$$v^* M_i v, \quad \forall\, i \in \{1, \ldots, m\}, \tag{2}$$

where $M_i = M_i(Y) \in \mathbb{C}^{n \times n}$ are sparse Hermitian matrices representing power-flow and power-injection as well as voltage magnitudes measurements. The sparsity pattern of the measurement matrices is determined by the topology of the network, while its nonzero entries are certain known functions of the entries of Y. Since the total number of nonzero elements in matrices M_i exceeds the total number of parameters contained in Y, we can think of $Y \to \{M_i\}_{i=1}^m$ as an embedding from a low-dimensional space. For a detailed discussion on the problem formulation and approaches to its solution see e.g. [38].

To formulate the problem as a low-rank matrix recovery, we introduce a sparse matrix $\mathbf{A} = \mathbf{A}(Y) \in \mathbb{C}^{m \times n^2}$ with i-th row equal to $\operatorname{vec}(M_i)^T$. The measurement vector can be written as $\mathbf{A}\operatorname{vec}(vv^T)$. To find v from the measurements, one can solve the non-convex optimization problem:

$$\underset{x \in \mathbb{C}^n}{\text{minimize}}\ \|\mathbf{A}\operatorname{vec}(xx^T - vv^T)\|^2 \tag{3}$$

In practice, this non-convex optimization problem is usually solved via local search methods, which converge to a second-order critical point at best. Since $f(x) = \|\mathbf{A}\operatorname{vec}(xx^T - vv^T)\|_F^2 = \langle xx^T - vv^T, \mathbf{A}^T\mathbf{A}\operatorname{vec}(xx^T - vv^T)\rangle$, the set of critical points for the problem is defined by the linear map represented with the matrix $\mathbf{H} = \mathbf{A}^T\mathbf{A}$, which thus is the key subject of the study. Problems arising in power systems analysis are based on operators that possess a specific structure. An example of a structure for the matrix $\mathbf{A}$ is given in Fig. 1a, and the structure of the corresponding $\mathbf{H}$ is described in Fig. 1b. The respective power network will be considered in more details in Section VI. As discussed previously, given $\mathbf{H}$, it is practically important to know if there exist $v, x \in \mathbb{C}^n$ such that x is a critical point of (3) while $xx^T \neq vv^T$. Absence of these points proves that a local search method recovers v exactly, certifying safety of its use. It is equivalent to the following problem having its optimal objective value equal to zero:



(a)

(b)

Fig. 1: Examples of the structure patterns of operators $\mathcal{A}$ (left plot) and $\mathcal{H}$ (right plot) in power system applications. The positions of the identical nonzero entries of a matrix are marked with the same markers.

$$\begin{aligned} \underset{v,\ x \in \mathbb{C}^n}{\text{maximize}} \quad & \|\mathcal{A}(xx^T - vv^T)\|^2 \\ \text{subject to} \quad & \nabla_x f(x, v) = 0 \\ & \nabla_x^2 f(x, v) \succeq 0 \end{aligned}$$

However, this is an $\mathcal{NP}$-hard problem in general and cannot be solved efficiently. Even if we solved it, the sensing operator $\mathcal{A}$ could change over time without changing its structure, and therefore any conclusion made for a specific problem cannot be generalized to other ones that should be solved for real-world problems where data analysis is to be performed periodically. One way to circumvent this issue is to develop a sufficient condition for all mappings $\mathbf{H}$ with the same structure.

III. Introducing Kernel Structure

Consider a linear operator $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ with the matrix representation $\mathbf{A} \in \mathbb{R}^{m \times n^2}$ and a sparsity pattern $S_{\mathcal{A}}$. Assume that there is a set of hidden parameters $\xi \in \mathbb{R}^d, d \ll m$, such that $\mathbf{A}$ is the image of ξ in the space of a much higher dimension. In this way, $\mathcal{A}$ has a low-dimensional structure beyond sparsity, which is captured by $\mathbf{A} = \mathbf{A}(\xi)$ and $\mathbf{A}(0) = \mathbf{0}$. The motivating example in Section II is a special case of this construction since it could be stated entirely with the real vectors and matrices of a bigger size than the complex ones. We define the nonconvex objective

$$f : \mathbb{R}^{n \times r} \to \mathbb{R} \quad \text{such that} \quad f(x) = \|\mathcal{A}(xx^T - zz^T)\|^2$$

parametrized by $\mathcal{A}$ and $z \in \mathbb{R}^{n \times r}$. Its value is always nonnegative by construction, and the global minimum 0 is attainable. To emphasize the dependence on certain parameters, we will write them in the subscript. To align the minimization problem with the problem of reconstructing zz^T, we need to introduce a regularity assumption:

Assumption 1. *The $2r$-RIP constant δ_{2r} of $\mathcal{A}$ exists (and by definition is strictly smaller than 1).*

Note that we do not assume any particular value for the RIP constant here. We will rely on Assumption 1 throughout the paper. This assumption implies that for all $x, z \in \mathbb{R}^{n\times r}$:

$$\|\mathcal{A}(xx^T - zz^T)\| = 0 \text{ if and only if } xx^T = zz^T$$

Another way to express the objective is

$$f(x) = \langle xx^T - zz^T, \mathcal{H}(xx^T - zz^T)\rangle.$$

Here, $\mathcal{H} = \mathcal{A}^T\mathcal{A}$ is the linear *kernel* operator that has the matrix representation $\mathbf{H} = \mathbf{A}^T\mathbf{A}$ and sparsity pattern $S_{\mathcal{H}}$. Namely, $(i,j) \in S_{\mathcal{H}}$ if and only if there exists k such that $(k,i) \in S_{\mathcal{A}}$ and $(k,j) \in S_{\mathcal{A}}$. Sparsity of $\mathcal{H}$ is controlled by the out-degree of the graph represented by $S_{\mathcal{A}}$, and tends to be low in applications like power systems. $S_{\mathcal{H}}$ is represented by a matrix $\mathbf{S}$, so that $S_{\mathcal{H}}$-sparse operators are exclusively those satisfying the linear equation $\mathcal{S}(\mathbf{H}) = \mathbf{S} \circ \mathbf{H} = \mathbf{0}$. Besides sparsity, $\mathcal{H}$ inherits the low-dimensional structure from $\mathcal{A}$, which can be captured by $\mathbf{H} = \mathbf{A}(\xi)^T\mathbf{A}(\xi) = \mathbf{H}(\xi)$ where $\xi \in \mathbb{R}^d$. This dependence can be locally approximated in the hidden parameter space with a linear one. More precisely, suppose that there is a linear operator $\mathcal{W}$ defined over $\mathbb{S}^{n^2}$ such that $\mathcal{W}(\mathbf{H}(\xi)) \approx 0$ for the values of ξ under consideration. Thus, from now on we focus exclusively on low-dimensional structures of the form $\mathcal{W}(\mathbf{H}) = 0$. Together, the sparsity operator $\mathcal{S}$ and the low-dimensional structure operator $\mathcal{W}$ form the combined structure operator $\mathcal{T} = (\mathcal{S}, \mathcal{W})$ that accumulates the structure of the kernel operator.

Definition 2 (Kernel Structure Property or KSP)**.** *The linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is said to satisfy $\mathcal{T}$-KSP if it satisfies Assumption 1 and there is a linear structure operator $\mathcal{T} : \mathbb{S}^{n^2} \to \mathbb{R}^t$ such that*

$$\mathcal{T}(\mathbf{A}^T\mathbf{A}) = 0$$

where $\mathbf{A}$ is the matrix representation of $\mathcal{A}$.

Notice that a particular sensing operator $\mathcal{A}$ can be kernel structured with respect to an entire family of structure operators, and we can possibly select any of them for our benefit in the following section.

IV. Using KSP

After fixing the kernel structure of the sensing operators, we can state the problem under study as follows:

$$\left\{ \underset{x\in\mathbb{R}^{n\times r}}{\text{minimize}}\ f_{z,\mathcal{A}}(x) \;\middle|\; \mathcal{A} \text{ satisfies } \substack{\text{Assumption 1} \\ \text{and } \mathcal{T}\text{-KSP}}, z \in \mathbb{R}^{n\times r} \right\}, \tag{Problem$^{\text{KSP}}$}$$

Note that (Problem$^{\text{KSP}}$) consists of infinitely many instances of an optimization problem, each corresponding to some point $z \in \mathbb{R}^{n\times r}$ and some operator $\mathcal{A}$ satisfying $\mathcal{T}$-KSP.

If X is regarded as an input and the operator $\mathcal{A}$ is regarded as a system with its output being $\mathcal{A}(X)$, the RIP constant aims at characterizing the input-output behavior of the system. This input-output relationship can also be controlled by imposing the following constraint on the matrix $\mathcal{H}$:

$$(1-\delta)\mathcal{I} \preceq \mathcal{H} \preceq (1+\delta)\mathcal{I},$$

where $\mathcal{I}$ is the identity operator. More precisely, the above inequality guarantees that the operator $\mathcal{A}$ has an RIP constant less than or equal to δ. Inspired by this observation, we introduce the function $\mathbb{O}(x, z; \mathcal{T})$ to be the optimal objective value of the convex optimization problem:

$$\begin{aligned} \underset{\delta\in\mathbb{R},\mathcal{H}}{\text{minimum}} \quad & \delta \\ \text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = 0 && \text{(4a)} \\ & \mathcal{M}_{x,z}(\mathcal{H}) \succeq 0 && \text{(4b)} \\ & \mathcal{T}(\mathcal{H}) = 0 && \text{(4c)} \\ & (1-\delta)\mathcal{I} \preceq \mathcal{H} \preceq (1+\delta)\mathcal{I} && \text{(4d)} \end{aligned}$$

where $\mathcal{L}_{x,z}(\mathcal{H}) = \nabla f_{z,\mathcal{H}}(x)$ and $\mathcal{M}_{x,z}(\mathcal{H}) = \nabla^2 f_{z,\mathcal{H}}(x)$. This optimization is performed over all operators $\mathcal{H}$ satisfying the KSP. We will later show that the function $\mathbb{O}$ sets an upper bound on the δ_{2r} such that none of the functions $f_{z;\mathcal{A}}$ with $\mathcal{A}$ satisfying $\mathcal{T}$-KSP and δ_{2r}-RIP has a spurious second-order critical point at x.

Since $f_{z,\mathcal{H}}(x)$ is linear in $\mathcal{H}$, the operators $\mathcal{L}_{x,z}$ and $\mathcal{M}_{x,z}$ are both linear. Thus, the problem defining the function $\mathbb{O}$ is convex.

To relax the δ_{2r}-RIP condition, we consider those operators that have a bounded effect on a linear subspace of limited-rank inputs. Indeed, for any $2r$ linearly independent vectors, the linear span of them is a linear subspace of the manifold of the $2r$-rank matrices. Thus, for any linear operator $\mathcal{P}$ from a $2r$-dimensional (or lower) vector space to $\mathbb{R}^{n^2}$, the following condition on $\mathcal{H}$ holds if $\mathcal{A}$ satisfies δ-RIP:

$$(1-\delta)\mathcal{P}^T\mathcal{P} \preceq \mathcal{P}^T\mathcal{H}\mathcal{P} \preceq (1+\delta)\mathcal{P}^T\mathcal{P}. \tag{5}$$

Based on this observation, we define the function $\mathbb{O}_P(x, z; \mathcal{T})$ as the optimal objective value of the following convex optimization problem:

$$\begin{aligned} \underset{\delta\in\mathbb{R},\mathcal{H}}{\text{minimum}} \quad & \delta \\ \text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = 0 && \text{(6a)} \\ & \mathcal{M}_{x,z}(\mathcal{H}) \succeq 0 && \text{(6b)} \\ & \mathcal{T}(\mathcal{H}) = 0 && \text{(6c)} \\ & (1-\delta)\mathcal{P}^T\mathcal{P} \preceq \mathcal{P}^T\mathcal{H}\mathcal{P} \preceq (1+\delta)\mathcal{P}^T\mathcal{P} && \text{(6d)} \end{aligned}$$

where $\mathcal{P}$ is the linear operator from $\mathbb{R}^{rank([x\ z])^2}$ to $\mathbb{R}^{n^2}$ that is represented by the matrix $\mathbf{P} = \text{orth}([x\ z]) \otimes \text{orth}([x\ z])$. Note that (6) is obtained from (4) by replacing its constraint (4d) with the milder condition (5). We will show that the function $\mathbb{O}_P$ sets a lower bound on the δ_{2r} such that none of the functions $f_{z;\mathcal{A}}$ with $\mathcal{A}$ satisfying $\mathcal{T}$-KSP and δ_{2r}-RIP has a spurious second-order critical point at x.

Now, we are ready to state one of the main results of this paper.

Theorem 2 (KSP necessary and sufficient conditions)**.** *For all instances of* (Problem$^{\text{KSP}}$)*, there are no spurious second-order critical points if*

$$\mathbb{O}_P(x, z; \mathcal{T}) \equiv 1 \textit{ over } \mathbb{R}^{n\times r} \times \mathbb{R}^{n\times r} \setminus \{xx^T = zz^T\} \quad (7)$$

and only if

$$\mathbb{O}(x, z; \mathcal{T}) \equiv 1 \textit{ over } \mathbb{R}^{n\times r} \times \mathbb{R}^{n\times r} \setminus \{xx^T = zz^T\} \quad (8)$$

To elaborate on implications and practicality of the result, we present its application for a specific structure of the sensing operator below.

A. Ellipsoid norm: Rank 1

In this subsection, we prove a special case of Theorem 2 for the ellipsoid norm objective function. This proof first provides useful intuition behind the proof of the general case and then simplifies the conditions of Theorem 2 to show that they always hold for a specific class of operators.

Consider the ellipsoid norm of $xx^T - zz^T$ given by a full-rank matrix $Q \in \mathbb{R}^{n\times n}$, denoted with h :

$$h(x) = \|Q(xx^T - zz^T)\|_F^2 = f_{z,\mathbf{A}}(x)$$

With no loss of generality, assume that $Q \in \mathbb{S}^n$ since $h(\cdot)$ really depends only on Q^TQ. The function can be implemented with a block-diagonal sensing operator matrix $\mathbf{A} = diag(Q, \ldots, Q) \in \mathbb{S}^{n^2}$, which generates a block-diagonal kernel matrix $\mathbf{H} = diag(QQ, \ldots, QQ)$. Thus, the kernel matrix is a block-diagonal matrix $\mathbf{H} = diag(H_{11}, \ldots, H_{nn}) \in \mathbb{S}^{n^2}$ with blocks of size $n \times n$ equal to each other; in other words, $H_{ii} = H_{jj}$ for all $i, j \in \{1, \ldots, n\}$. This generates a kernel structure. By applying the theory introduced above, we obtain the following result for the rank-one case.

Proposition 1. *Consider a kernel structure operator* $\mathcal{T} = (\mathcal{S}, \mathcal{W})$ *such that*

- $\mathcal{S}(\mathbf{H}) = \mathbf{0}$ *iff* $\mathbf{H} = diag(H_{11}, \ldots, H_{nn})$
- $\mathcal{W}(\mathbf{H}) = \mathbf{0}$ *iff* $H_{ii} = H_{jj}$, $i, j \in \{1, \ldots, n\}$,

Then, no instance of the (Problem$^{\text{KSP}}$) *has a spurious second-order critical point over* $\mathbb{R}^n$*.*

The proposition implies that the function $h(x)$ can never have a spurious solution for rank-1 arguments.

V. Combining KSP with RIP

After fixing the kernel structure of the sensing operators and the RIP constant, we can state the problem under study in this section as follows:

$$\left\{ \underset{x\in\mathbb{R}^{n\times r}}{\text{minimize}} f_{z,\mathcal{A}}(x) \;\middle|\; \begin{array}{c}\mathcal{A} \text{ satisfies } \delta_{2r}\text{-RIP and } \mathcal{T}\text{-KSP},\\ z \in \mathbb{R}^{n\times r}\end{array} \right\}, \quad (\text{Problem}^{\text{KSP+RIP}})$$

Note that (Problem$^{\text{KSP+RIP}}$) consists in minimization of a class of functions $f_{z,\mathcal{A}}$ that correspond to some point $z \in \mathbb{R}^{n\times r}$ and some operator $\mathcal{A}$ that satisfies $\mathcal{T}$-KSP and δ_{2r}-RIP simultaniously. This is a generalization of both (Problem$^{\text{RIP}}$) and (Problem$^{\text{KSP}}$). For (Problem$^{\text{KSP+RIP}}$), we provide necessary and sufficient conditions for having no spurious second-order critical point, and consequently no spurious local minimum.

Theorem 3 (KSP+RIP necessary and sufficient conditions)**.** *For all instances of* (Problem$^{\text{KSP+RIP}}$)*, there are no spurious second-order critical points if*

$$\delta_{2r} < \min_{\substack{x\in\mathbb{B}_R(\omega), z\in\mathbb{B}_R(\omega)\\ xx^T \neq zz^T}} \mathbb{O}_P(x, z; \mathcal{T}) \quad (9)$$

and only if

$$\delta_{2r} < \min_{\substack{x\in\mathbb{B}_R(\omega), z\in\mathbb{B}_R(\omega)\\ xx^T \neq zz^T}} \mathbb{O}(x, z; \mathcal{T}) \quad (10)$$

Following from the results of [31], the necessary and sufficient conditions coincide for the trivial structure operator $\mathcal{T} \equiv 0$.

A. Sparse structure and normalization

Due to Theorem 1 for the rank-1 case, the instances of (Problem$^{\text{KSP+RIP}}$) have no spurious solutions with $\mathcal{T} \equiv 0$ as long as δ_2 is upper bounded by $\frac{1}{2}$. In this subsection, we are concerned with the question of how much sparsity can impact the best bound on RIP that certifies global convergence. Formally, we set $\mathcal{W} \equiv 0$ and $\mathcal{T} \equiv \mathcal{S}$ and find a tighter upper bound for δ_2. After enforcing sparsity, it is natural to expect that the bound grows and becomes less restrictive. However, this turns out not to be the case.

Let $n = 2$ and $r = 1$, and consider the smallest sparsity pattern possible for $\mathcal{H} = \mathcal{A}^T\mathcal{A} \succ 0$. It consists exclusively of elements (i, i), and thus enforces $\mathbf{H}$ to be diagonal. Consider the point x with respect to the instance of the problem given by z and $\mathbf{A}$ as in the example below:

Example 1. Assume that

$$x = (1, 1); \quad z = (\sqrt{2}, -\sqrt{2}); \quad \mathbf{A} = diag(\sqrt{3}, 1, 1, \sqrt{3})$$

Then, x is spurious for $f_{z,\mathbf{A}}$ since it satisfies the second-order necessary conditions:

$$\nabla f_{z,\mathbf{A}}(x) = 0, \quad \nabla^2 f_{z,\mathbf{A}}(x) = 16 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \succeq 0$$

which makes it a spurious second-order critical point (note that $xx^T \neq zz^T$). Notice that $\mathcal{H} = \mathcal{A}^T\mathcal{A}$ is indeed diagonal. Moreover, for all $X \in \mathbb{S}^2$, the operator $\mathcal{A}$ satisfies the tight bound $\|X\|_F^2 \leq \|\mathcal{A}(X)\|^2 = \| \begin{bmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{bmatrix} \circ X\| \leq 3\|X\|_F^2$. Therefore, the largest number δ_2 for this instance is equal to $1/2$, which coincides with the upper bound for unstructured problems. Somewhat counter-intuitively, the tight bound established in [24], [31] holds even when a very restrictive sparsity pattern of the kernel operator is enforced. Nevertheless, for an arbitrary low-dimensional structure $\mathcal{W}$, a tighter sparsity constraint entails a less restrictive bound on incoherence as discussed below.

Proposition 2. *If the sparsity pattern S has a sub-pattern S' meaning that $S' \subset S$, then $\mathbb{O}(x, z; \mathcal{W}, \mathcal{S}') \leq \mathbb{O}(x, z; \mathcal{W}, \mathcal{S})$ for all $x, y \in \mathbb{R}^{n\times r}$. Thus, the necessary bound on incoherence for $\mathbf{H}$ with S' is not more restrictive than the bound for $\mathbf{H}$ with S.*

In other words, a more restricting assumption on the sparsity of the kernel operator can only push the upper bound on the RIP constant higher up. Consequently, Example 1 shows that there is no sparsity pattern of cardinality > 3 that can itself compensate the lack of isometry. Note that the example is given for the case $n = 2$, but there is a straightforward extension to an arbitrary n by adding zero components to x and z. It is common in practice to normalize the rows of the sensing matrix before proceeding to recovery. In the context of power systems, it is expressed as $x^T M_i x \to \frac{x^T M_i x}{\|M_i\|_F}$. For Example 1, after normalization, $\mathbf{A}$ turns into the identity. The corresponding instance of the problem is known to have no spurious critical points. This illustrates how normalization helps to improve the isometry property of the sensing operator and removes the spurious second-order critical points out of the corresponding instance of the problem. Normalization in this case can be regarded as inducing structure on top of sparsity.

VI. Numerical results

In this section, we present numerical studies of the matrix recovery problems for structured sensing operators. One objective is to demonstrate how the analytical framework developed in Section III can be applied to evaluate the hardness of a real-world problem, namely the power system state estimation discussed in Sections II and V. We calculate a numerical estimation of the minimal value of the function $\mathbb{O}_P$ for different structural patterns $\mathcal{T}$. After that, we estimate the best RIP constant of a sensing operator satisfying the KSP, which is sufficient to guarantee the absence of spurious local second-order critical points. We call this constant the *sufficient best RIP constant* or just the *sufficient RIP*. There is a clear connection between the hardness of a matrix sensing problem and the sufficient RIP, and the study we conduct eventually aims to find out the role of structure in non-convex optimization.

In general, the optimization problem (9) is non-convex. Thus, we propose to use Bayesian optimization [39] in order to obtain a numerical estimation of its solution. We have empirically observed that Bayesian optimization tends to obtain the same optimal solution to this problem much faster than random shooting or cross-entropy.

Recall that the structure operator is defined by two operators stack together: $\mathcal{T} = (\mathcal{S}, \mathcal{W})$. Here, $\mathcal{W}$ captures the underlying structure that is not captured by the sparsity operator $\mathcal{S}$. We will consider a particular form of this operator. Given the matrix representation $\mathbf{H}$ of the kernel operator, denote the unique nonzero values in this matrix with the scalars $h_1, \ldots, h_{d_\mathcal{W}}$. It means that $\mathbf{H}$ is representable in the form $\mathbf{H} = h_1\mathbf{E}_1 + \ldots + h_{d_\mathcal{W}}\mathbf{E}_{d_\mathcal{W}}$, where $\mathbf{E}_i$ is a matrix of the same size as $\mathbf{H}$, with 0 and 1 entries. The operator $\mathcal{W}$ that we use in this section is any operator that has the subspace $\{\beta_1\mathbf{E}_1 + \ldots + \beta_{d_\mathcal{W}}\mathbf{E}_{d_\mathcal{W}} : \beta_1, \ldots \beta_{d_\mathcal{W}} \in \mathbb{R}\}$ as its kernel. This is inspired by the kernel operators for real-world power systems.

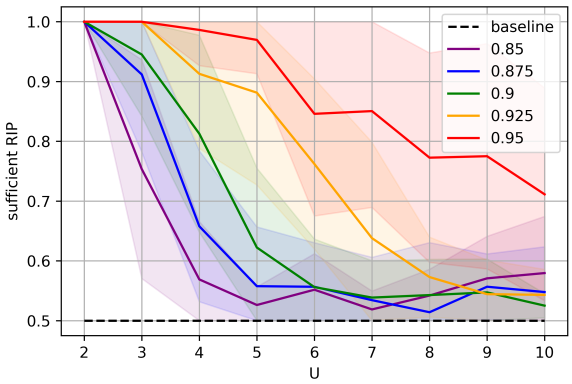


Fig. 2: The average sufficient best RIP constant obtained from the developed analytic framework (Theorem 3) for random structures generated from the distribution $RS(p_0, U)$ (each colored line stands for one specific value of p_0), compared with the baseline method from Theorem 1 (shown as black and dashed). Shaded area represents the standard deviation window.

A. Synthetic data

We generate a class of randomly generated structures by introducing a distribution $RS(p_0, U)$ over the space of structure operators. First, we generate the measurement structure matrix $\mathbf{A}_{\text{st}}$ such that each of its components takes value 0 with probability p_0 and any of the values $1, \ldots, U$ with the equal probability of $\frac{1-p_0}{U}$. We then form the kernel structure matrix as $\mathbf{H}_{\text{st}} = \mathbf{A}_{\text{st}}^T\mathbf{A}_{\text{st}}$ and construct the sparsity operator $\mathcal{S}$ and the extra structure operator $\mathcal{W}$ as discussed previously. The obtained structure operator $\mathcal{T}$ is such that the operator represented with $\mathbf{A}_{\text{st}}$ satisfies the $\mathcal{T}$-KSP. Note that the average sparsity of $\mathbf{A}_{\text{st}}$ is p_0 and the number of unique nonzero values is U with high probability, which implies that p_0 is the parameter for the amount of sparsity structure in the problem, and U is the parameter for the amount of additional structure.

Figure 2 depicts the estimated sufficient RIP for random problems with different values for the sparsity (p_0) and the unique counter (U). Observe that the sparsity and the additional structure (the number of unique nonzero values in the measurement matrix in this particular case) both have a significant impact on the sufficient RIP. Note that higher p_0 means more sparsity and lower U means more extra structure. Although it was observed theoretically that sparsity alone cannot guarantee the increase in the sufficient best RIP constant, it appears to be an important characteristic when combined with the additional structure.

Even for structures with a considerably low sparsity (0.85), the tight extra structure ($U = 2$) has the sufficient best RIP of

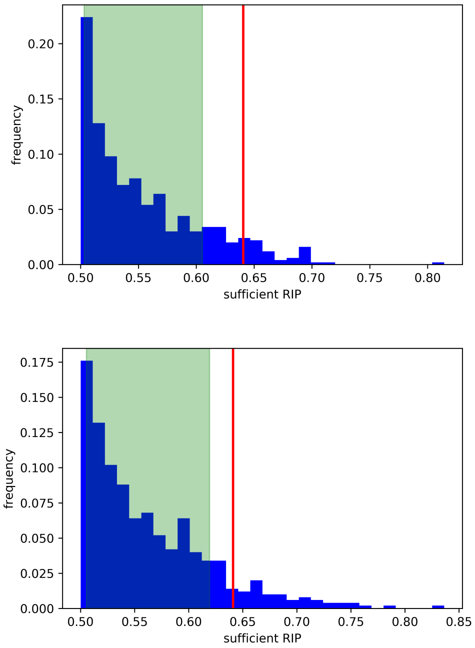


Fig. 3: The distribution of the sufficient best RIP constant obtained for random structure operators (blue) and the sufficient best RIP constant obtained for the structure operator from real-world data (red) for the 9-bus (top) and 14-bus (bottom) power systems. Green shaded region depicts the standard deviation window of the sufficient RIP for the synthetic data.

1, which is a counter-intuitive result. The the sufficient RIP seems to decay exponentially as we relax extra structure by increasing U, but with different bases for different p_0. This behaviour coincides with the one predicted in Proposition 2. If the goal is to make the RIP higher than a certain threshold, the amount of extra structure needed to achieve this reduces dramatically with the increase of the sparsity structure.

The key takeaway from this experiment is that our method captures the structural properties of a given mapping and shows that it significantly affects the sufficient RIP, which leads to certifying the absence of spurious solutions under far less restrictive requirements (by improving the previous bound 0.5 for arbitrary mappings).

B. Power systems data

In this subsection, we estimate the sufficient best RIP bound for the matrix sensing formulation of different cases of the power system state estimation problem using Theorem 3. After that, we compare the bounds against the sufficient best RIP of random problems generated from RS with the same p_0 and U parameters as in power systems. .

We focus our attention on two networks named `case9` and `case14`, which are provided in the MATPOWER package. For `case9`, the $\mathbf{A}$ and $\mathbf{H}$ matrices are visualized in Figure 1. The sufficient best RIP constant in both cases leads to the bound 0.64, which is substantially better that the previously best known bound 0.5, although our bound does not yet fully explain the success of the non-convex optimization approach for the power system state estimation (as discussed below, one may exploit more structures in the Y matrix to tighten the bound).

Figure 3 shows the distribution of the sufficient RIP for structures both randomly generated and originated from real-world data. The distribution parameters are set in a way that sparsity and extra structure parameters in both cases are likely to coincide. Observe that the sufficient RIP values for power systems are regularly larger than the ones for the synthetic data. Consequently, the original structure has considerably different properties from the randomly generated structures, which gives us the opportunity to build better bounds for the real-world system than for a randomly generated system, by finding a more detailed structure. For example, our model just used the number of unique nonzero values, but one may also model the relationship between the nonzero elements.

The above simulations were based on the networks provided in the package MATPOWER 7.0b1 [40]. All of the presented simulations were done using the MATLAB bayesopt toolbox for non-convex optimization and MATLAB modeling toolbox CVX [41], [42] with SDPT3 [43], [44] as the underlying convex solver.

VII. Conclusion

In this work, we study the optimization landscape of the non-convex matrix sensing problem that is known to have many local minima in the worst case. Since the existing results are related to the notion of restricted isometry property (RIP) that cannot directly capture the underlying structure of a given problem, they can hardly be applied to real-world problems where the amount of data is not exorbitantly high. To address this issue, we develop the notion of kernel structure property to obtain necessary and sufficient conditions for the inexistence of spurious local solution of any class of matrix sensing problems over a given search space. This notion precisely captures the underlying sparsity and structure of the problem, based on tools in conic optimization.We simplify the conditions for a certain class of problems to show their satisfaction and apply them to data analytics for both power systems and a class of randomly generated structured systems.

References

[1] I. Molybog, S. Sojoudi, and J. Lavaei, “Role of sparsity and structure in the optimization landscape of non-convex matrix sensing,” 2020, https://lavaei.ieor.berkeley.edu/KSP_2019_1.pdf.

[2] P. M. Pardalos and S. A. Vavasis, “Quadratic programming with one negative eigenvalue is np-hard,” *Journal of Global Optimization*, vol. 1, no. 1, pp. 15–22, Mar 1991.

[3] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” *IEEE transactions on information theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[4] ——, “Matrix completion from noisy entries,” *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2057–2078, 2010.

[5] P. Jain, P. Netrapalli, and S. Sanghavi, “Low-rank matrix completion using alternating minimization,” in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 665–674.

[6] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 109–117.

[7] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix estimation," in *Advances in Neural Information Processing Systems*, 2015, pp. 559–567.

[8] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[9] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli, "Learning sparsely used overcomplete dictionaries via alternating minimization," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2775–2799, 2016.

[10] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for non-convex losses," *arXiv preprint arXiv:1607.06534*, 2016.

[11] M. Soltanolkotabi, "Learning relus via gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 2007–2017.

[12] C. Yun, S. Sra, and A. Jadbabaie, "Global optimality conditions for deep neural networks," in *International Conference on Learning Representations*, 2018.

[13] D. Li, T. Ding, and R. Sun, "Over-parameterized deep neural networks have no strict local minima for any continuous activations," *arXiv preprint arXiv:1812.11039*, 2018.

[14] M. Nouiehed and M. Razaviyayn, "Learning deep models: Critical points and local openness," *arXiv preprint arXiv:1803.02968*, 2018.

[15] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Mathematical Programming*, pp. 1–33, 2018.

[16] N. Vaswani, S. Nayer, and Y. C. Eldar, "Low-rank phase retrieval," *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 4059–4074, 2017.

[17] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[18] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1233–1242.

[19] C. Josz, Y. Ouyang, R. Zhang, J. Lavaei, and S. Sojoudi, "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization," in *Advances in neural information processing systems*, 2018, pp. 2441–2449.

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

[21] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan, "Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization," *arXiv preprint arXiv:1902.07698*, 2019.

[22] X. Li, J. Lu, R. Arora, J. Haupt, H. Liu, Z. Wang, and T. Zhao, "Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization," *IEEE Transactions on Information Theory*, 2019.

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

[24] R. Zhang, C. Josz, S. Sojoudi, and J. Lavaei, "How much restricted isometry is needed in nonconvex matrix recovery?" in *Advances in neural information processing systems*, 2018, pp. 5591–5602.

[25] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points — online stochastic gradient for tensor decomposition," 2015.

[26] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery using nonconvex optimization," in *International Conference on Machine Learning*, 2015, pp. 2351–2360.

[27] ——, "A geometric analysis of phase retrieval," *Foundations of Computational Mathematics*, vol. 18, no. 5, pp. 1131–1198, 2018.

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

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

[30] 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.

[31] R. Y. Zhang, S. Sojoudi, and J. Lavaei, "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery," *Journal of Machine Learning Research*, vol. 20, no. 114, pp. 1–34, 2019.

[32] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann, "Escaping saddles with stochastic gradients," *arXiv preprint arXiv:1803.05999*, 2018.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[34] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in neural information processing systems*, 2008, pp. 161–168.

[35] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on learning theory*, 2016, pp. 1246–1257.

[36] Q. Li, Z. Zhu, and G. Tang, "Alternating minimizations converge to second-order optimal solutions," in *International Conference on Machine Learning*, 2019, pp. 3935–3943.

[37] S. Paternain, A. Mokhtari, and A. Ribeiro, "A newton-based method for nonconvex optimization with fast evasion of saddle points," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 343–368, 2019.

[38] Y. Zhang, R. Madani, and J. Lavaei, "Conic relaxations for power system state estimation with line measurements," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1193–1205, 2018.

[39] P. I. Frazier, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

[40] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on power systems*, vol. 26, no. 1, pp. 12–19, 2011.

[41] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[42] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[43] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, "Sdpt3—a matlab software package for semidefinite programming, version 1.3," *Optimization methods and software*, vol. 11, no. 1-4, pp. 545–581, 1999.

[44] R. H. Tütüncü, K.-C. Toh, and M. J. Todd, "Solving semidefinite-quadratic-linear programs using sdpt3," *Mathematical programming*, vol. 95, no. 2, pp. 189–217, 2003.