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

Igor Molybog
University of California, Berkeley
igormolybog@berkeley.edu

Somayeh Sojoudi
University of California, Berkeley
sojoudi@berkeley.edu

Javad Lavaei
University of California, Berkeley
lavaei@berkeley.edu

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 contains no spurious local minima. This assumption is too strong for real-world applications where the amount of data cannot be sufficiently high. In this paper, we show that the incoherence assumption can be significantly relaxed by exploiting the underlying structure of the sensing operator. In particular, we study sparse operators that have a low-dimensional representation. We develop general necessary and sufficient conditions for no spurious solutions under incoherence and structure assumed simultaneously. Using that, we prove that sparsity and structure together could make the incoherence assumption almost redundant and offer a case study on data analytics for electric power systems for which the original incoherence assumption is not satisfied.

1 Introduction

Even under ideal conditions of no noise and zero approximation error, many highly-efficient machine learning techniques involve solving potentially hard or intractable computational problems while learning the data. In practice, they are approached with heuristic optimization algorithms, based on relaxations or greedy principals. The lack of guarantees on their output 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, transport, energy production and distribution. Commonly, non-convexity is the main obstacle for guaranteed learning of continuous parameters.

It is well known that even relatively simple in formulation non-convex problems can be $\mathcal{NP}$-hard [1]. As a consequence of complicated geometrical structure, a non-convex function may contain exponential number of saddle points and spurious local minima, and local search algorithms may become trapped in them. Other heuristics, like generic convex relaxations, may require work in an unrealistically large dimensional space to guarantee exactness of their solution. Nevertheless, empirical observations show positive results regarding application of these approaches to the most

*This work was supported by grants from ARO, ONR, AFOSR, and NSF

practically important instances. This provokes a large branch of research trying to explain the 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 machine learning applications. The most commonly applied class of such algorithms are the *local search methods*, which we are going to study. For a twice continuously differentiable objective function $f : \mathbb{R}^{n\times r} \to \mathbb{R}$ that reaches its global minimum f^*, if the point x attains $f(x) = f^*$, then we call it a *globally minimum*; otherwise, we call it a *spurious point.* The point x is said to be a *local minimum* if $f(x) \leq f(x')$ holds for all x' within a local neighborhood of x. If x is a local minimum, then it must satisfy the first- and second-order *necessary* condition for local optimality:

$$\nabla f(x) = 0, \tag{1a}$$

$$\nabla^2 f(x) \succeq 0. \tag{1b}$$

Conversely, a point x satisfying (1a) is called a *first-order critical point,* while a point satisfying (1) is called a *second-order critical point.* We also call it a *solution* since the local search algorithms are guaranteed to converge to a first- or a second-order critical point, and not necessarily a local minimum; for further details on gradient methods see [2]–[5], and [6]–[9] for the details on trust-region methods.

Analysis of the landscape of the objective functions around a global optimum may lead to an optimality guarantee for the local search algorithms initialized sufficiently close to it [10]–[15]. This approach is easy to take, but good initializations are highly problem-specific and difficult to generalize. Global analysis of the landscape is harder, but potentially more rewarding. It may help to prove global convergence of a local search method for an arbitrary initialization.

Both local and global convergence guarantees were developed to justify local search methods in the applications like dictionary learning [16], basic non-convex M-estimators [17], shallow [18] and deep [19] artificial neural networks with different activation [20] and loss [21] functions, phase retrieval [22], [23] and more general matrix sensing [24], [25]. 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 [24], [26]–[30]. Given a 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

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

Recent work has generally found certain incoherence assumption on the sensing operator to be sufficient for the matrix sensing problem to be "easy". 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 if there are constants $\delta_r \in [0, 1)$ and $\gamma > 0$ such that*

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

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

Most of the results proving absence of spurious local minima by using this notion ([2], [24], [26], [31]–[35]) are based on a norm-preserving argument: the problem turns out to be a low-dimension embedding of a canonical problem known to contain no spurious local minima. While the approach is widely applicable in its scope, it turns out to be restrictive in the problem data and does not

provide a way to analyze necessary conditions.
In contrast, [30], [36] introduced an approach of finding a certificate that a particular point cannot be a spurious local minimum for any instance of the problem

$$\left\{ f_{z,\mathcal{A}}(x) \to \min_{x \in \mathbb{R}^{n\times r}} \;:\; \mathcal{A} \text{ satisfies } \delta_{2r}\text{-RIP, } z \in \mathbb{R}^{n\times r} \right\}, \qquad (\text{Problem}^{\text{RIP}})$$

where n and r are implied to be fixed, and m can vary, or $m = n^2$ without loss of generality. They improve the bound established in [36] for the rank-1 case and show that it cannot be improved in general. The following theorem summarizes these results:

Theorem 1 ([24], [33], [36])**.**

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

Non-existence of a spurious second-order stationary point effectively means that any algorithm that converges to a second-order critical point is guaranteed to recover zz^T exactly. One example of such an algorithm is the stochastic gradient descent (SGD) which is known to avoid saddle or even spurious local minimal points [37], and widely used in machine learning [38], [39].
Theorem 1 disclosures the limits on the guarantees that the notion of RIP can provide with. However, linear maps from applications like Power System analysis typically have the RIP constant higher than 0.9, and the non-convex matrix sensing still manages to work on some of the instances. This distance between theory and practice motivates the following question.
What is the alternative property practical problems are satisfying that makes them easy to be solved via simple local search?
We propose and test a possible answer to this question. In Section 2 we motivate it with real-worlds and synthetic examples, providing more insights into why this question should be asked. Section 3 provides with the formal definitions and the general framework that we introduce in our intent to answer the main question. Section 4 applies the introduced framework to particular problems, showing both positive and negative theoretical results, testing the hypothesis. In Section 5 we present numerical results of successful application of the framework to the real-world problems appearing in Power Systems analysis. All the proofs, technical details and lemmas are collected in the Appendix.

Notation

$\mathbb{R}^n$ and $\mathbb{R}^{n\times r}$ denote the sets of real n-dimensional vectors and $n \times r$ matrices, respectively. $\mathbb{S}^n$ denotes the sets 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 size. 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 columns of the matrix in the form of a vector. The *matricization* operator $\text{mat}(\cdot)$ is the inverse of $\text{vec}\,(\cdot)$.
For a linear operator $\mathcal{L} : \mathbb{R}^{n\times r} \to \mathbb{R}^m$, the adjoint operator is denoted with $\mathcal{L}^T : \mathbb{R}^m \to \mathbb{R}^{n\times r}$. 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. Bold letters are reserved for matrix representations of corresponding linear operators.
Sparsity pattern S of a set of matrices $\mathcal{X} \subset \mathbb{R}^{m\times n}$ is a subset of $\{1, \ldots, \max\{n, m\}\}^2$ such that

$(i,j) \in S$ if and only if there is $X \in \mathcal{X}$ such that $X_{ij} \neq 0$. Given a sparsity pattern S, define its matrix representation $\mathbf{S} \in \mathbb{S}^{m \times n}$ such that

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

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

$$P = \text{orth}(A) \qquad \Longleftrightarrow \qquad PP^T A = A, \qquad P^T P = I_{\text{rank}\,(A)}.$$

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

2 Motivating Example

In this section we motivate our developments by considering an example from the area of electric power grid control. The state of a power system can be modeled with 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. Since the voltages cannot be measured directly, this crucial information has to be inferred from other measurable parameters, such as the power that is generated and consumed on buses or transmitted through the lines. The power network can be modeled with a number of parameters grouped into the admittance matrix $Y \in \mathbb{C}^{n \times n}$. The state estimation problem consists in recovering the target voltages vector $v \in \mathbb{C}^n$ from the available measurements. In the noiseless scenario, they are given with m real numbers of the form

$$v^* M_i v, \quad \forall\, i \in \{1, \dots, m\}, \tag{3}$$

where $M_i = M_i(Y) \in \mathbb{C}^{n \times n}$ are sparse Hermitian matrices representing power-flow measurements. The sparsity pattern of the measurement matrices is determined by the topology of the network, while the nonzero values are known functions of the entries of Y. While the total number of nonzero elements in matrices M_i exceeds the total number 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. [40].

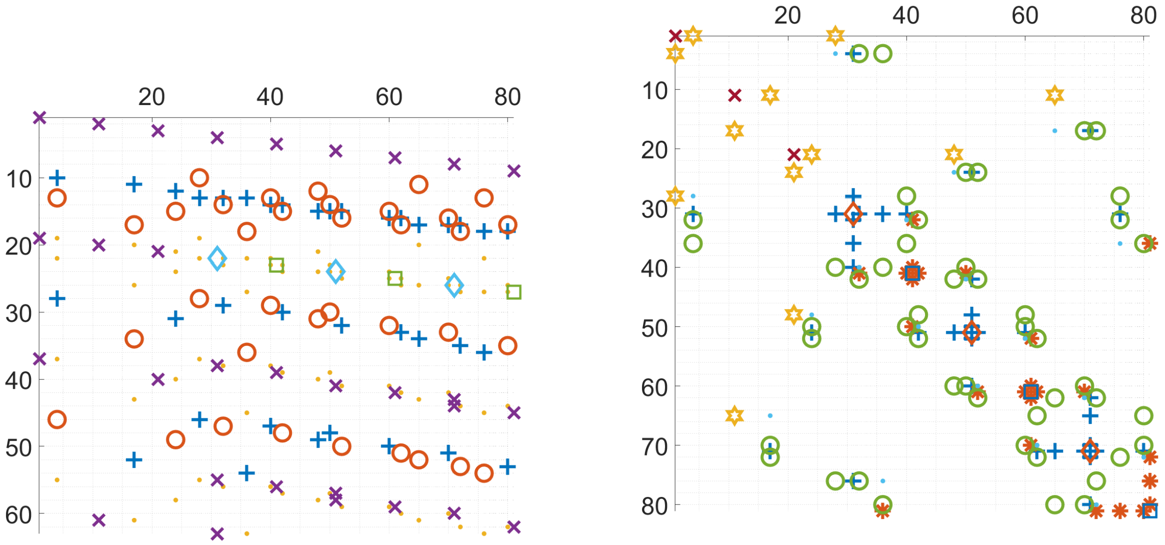
To put the problem into the form of low-rank matrix recovery, introduce a sparse matrix $\mathbf{A} = \mathbf{A}(Y) \in \mathbb{C}^{m \times n^2}$ with i-th row equal to $\text{vec}(M_i)^T$. The measurements vector can be written as $\mathbf{A}\text{vec}(vv^T)$. To find v from the measurements, we consider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

$$\underset{x \in \mathbb{C}^n}{\text{minimize}} \|\mathbf{A}\text{vec}\,(xx^T - vv^T)\|^2 \tag{4}$$

In practice, it is usually solved with local search methods, which converge to a second-order critical point at best. Since $f(x) = \|\mathbf{A}\text{vec}\,(xx^T - vv^T)\|_F^2 = \langle xx^T - vv^T, \mathbf{A}^T\mathbf{A}\text{vec}\,(xx^T - vv^T)\rangle$, the set of critical points in 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. Let's consider an example of a question that might arise in this setting

Example 1. *After applying a local search algorithm to the problem* $\min_{x \in \mathbb{R}^n} \|\mathbf{A}\text{vec}\,(xx^T - yy^T)\|$ *with* $\mathbf{H} = \mathbf{A}^T\mathbf{A}$ *of the form*

$$\mathbf{H} = \begin{bmatrix} 30 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 30 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$



(a)

(b)

Figure 1: Examples of structure patterns of operators $\mathcal{A}$ and $\mathcal{H}$ in power system applications. The positions of repeated nonzero entries of a matrix are marked with the same markers.

is it reasonable to expect that for an arbitrary starting point, the end point $x_\star$ *is such that* $x_\star x_\star^T = zz^T$*?*

While not immediately obvious, due to the structure of the operator, the answer to this question is positive, although the RIP constant of the underlying operator is bigger than 0.59, which does not allow to apply the Theorem 1. Problems arising in power systems analysis are based on operators that also possess a specific structure. An example of a structure of matrix $\mathbf{A}$ is given on Fig. 1a, and the structure of corresponding $\mathbf{H}$ is described on Fig. 1b. The respective power network will be considered in more details in Section 5. 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 stationary point of (4) while $xx^T \neq vv^T$. Absence of these points proves that a local search method recovers v exactly, certifying safety of its use. Formally, the answer can be provided by the following problem having optimal solution equal to zero:

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

However, it is an $\mathcal{NP}$-hard in general and doesn't have a scalable solution. Even if we solved it, the sensing operator $\mathcal{A}$ could change over time without changing its structure, but we had to solve it over again. One way around it is to develop a sufficient condition for $\mathbf{H}$ satisfying a particular property. RIP is one example of a property, but, as we mentioned, it does not completely explain the empirical results.

3 Introduce kernel structure

Consider a linear operator $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ with the matrix representation $\mathbf{A} \in \mathbb{R}^{m \times n^2}$, a sparsity pattern $S_{\mathcal{A}}$ and a low-dimensional structure captured by $\mathbf{A} = \mathbf{A}(w), w \in \mathbb{R}^d, d \ll m$. We assume

that $\|\mathcal{A}(xx^T - zz^T)\| = 0$ if and only if $xx^T = zz^T$, which translates into $\mathbf{A}$ satisfying rank-$2r$ RIP with some constant. The motivating example in Section 2 could be stated entirely with real-valued vectors and matrices, so this construction includes it as a partial case.
We define the nonconvex objective

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

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 attained. To emphasize dependence on parameters, we will write them in the subscript. 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 matrix representation $\mathbf{H} = \mathbf{A}^T\mathbf{A}$ and sparsity pattern $S_{\mathcal{H}}$, which is well understood. 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 pretty low in applications like power systems. Besides sparsity, $\mathcal{H}$ inherits the low-dimensional structure from $\mathcal{A}$, which can be expressed by $\mathbf{H} = \mathbf{H}(w)$. This structure can be further relaxed to a linear one. Space of matrices that obey a sparsity pattern is also linear, so we further consider a linear constraint on the kernel operator.

Definition 2 (Kernel Structure Property)**.** *The linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is said to satisfy $\mathcal{T}$-KSP if 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 $\mathcal{A}$.

In this paper, we assume $\mathcal{T}$ to consist of sparsity operator $\mathcal{S}(\mathbf{H}) = \mathbf{S} \circ \mathbf{H}$ and a low-dimensional embedding operator $\mathcal{W}$, which substitutes the constraint $\mathbf{H} = \mathbf{H}(w)$, so that $\mathcal{T} = (\mathcal{S}, \mathcal{W})$.
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 pick any of them for our benefit in the following discussion.
After fixing the kernel structure of the sensing operators, we can state the problem we are now dealing with:

$$\left\{ f_{z,\mathcal{A}}(x) \to \min_{x\in\mathcal{X}} \;:\; \mathcal{A} \text{ satisfies } \delta_{2r}\text{-RIP, and } \mathcal{T}\text{-KSP, } z \in \mathcal{Z} \right\}, \qquad (\text{Problem}^{\text{KSP+RIP}})$$

The class of problems in the form (Problem$^{\text{RIP}}$) is a subclass of problems (Problem$^{\text{KSP+RIP}}$) with $\mathcal{T}$ being the trivial operator. For (Problem$^{\text{KSP+RIP}}$) we come up with both necessary and sufficient conditions of having no spurious second order stationary point, and consequently a spurious local minima. They are given with the following theorems.

Theorem 2 (Necessary condition)**.** *For $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^{n\times r}$, there are no spurious second-order stationary points in any instance of the* (Problem$^{\text{KSP+RIP}}$) *only if*

$$\delta_{2r} < \min_{\substack{x\in\mathcal{X}, z\in\mathcal{Z} \\ xx^T \neq zz^T}} \text{LMI}^T(x, z; \mathcal{T}) \tag{5}$$

Here LMI^T is a convex programming problem that can be put in the form

$$LMI^T(x,z;\mathcal{T}) = \underset{y\in\mathbb{R}^{n\times r}, V\succeq 0, \lambda\in\mathbb{R}^t}{\text{maximize}} \frac{\sum_{i=1}^{d}(-\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{T}^T(\lambda)))_+}{\sum_{i=1}^{d}(+\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{T}^T(\lambda)))_+} \tag{6}$$

with $\mathcal{L}_{x,z}(\mathcal{H}) = \nabla f_{z,\mathcal{H}}(x)$; $\mathcal{M}_{x,z}(\mathcal{H}) = \nabla^2 f_{z,\mathcal{H}}(x)$. 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 with the exact form stated in the appendix.

Theorem 3 (Sufficient condition)**.** *For $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^{n\times r}$, there are no spurious second-order stationary points in any instance of the* (Problem$^{\text{KSP+RIP}}$) *if*

$$\delta_{2r} < \min_{\substack{x\in\mathcal{X}, z\in\mathcal{Z}\\ xx^T\neq zz^T}} \text{LMI}_P(x,z;\mathcal{T}) \tag{7}$$

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

where $\mathcal{P}$ is the linear operator from $\mathbb{R}^{rank([x\ z])^2}$ to $\mathbb{R}^{m^2}$ which is represented by the matrix $\mathbf{P} = \text{orth}([x\ z]) \otimes \text{orth}([x\ z])$. As follows from the results of [36], the necessary and sufficient conditions stated in the theorems 2 and 3 coincide for the partial case $\mathcal{X} = \mathcal{Z} = \mathbb{R}^{n\times r}$ and the trivial structure operator. Further in the paper, we also discuss another structure for which the necessary condition also plays the role of sufficient.

4 Theoretical applications

In this section we discuss applications and limitations of the notions and theory presented above. We work with examples of (Problem$^{\text{KSP+RIP}}$) to investigate the spurious solutions of their instances.

4.1 Sparse structure and normalization

In this subsection we are concerned with the question on how much sparsity alone ($\mathcal{T} \equiv \mathcal{S}$) can impact the best bound on RIP that certifies global convergence. By Theorem 1, in rank-1 case, this bound is equal to $\frac{1}{2}$. After shrinking the problem by enforcing sparsity, it is natural to expect that the bound goes up, becoming less restrictive. However, this is not the case with sparsity.
Let $n = 2, r = 1$, 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 2.

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

The x is spurious for $f_{z,\mathbf{A}}$, since $f_{z,\mathbf{A}}(x) = 24$, however, 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 stationary point. 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 a 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 biggest δ_2 for this instance is equal to 1/2, which coincides with the upper bound for unstructured problems. Somewhat counterintuitively, the tight bound established in [30], [36] holds even when a very restrictive sparsity pattern of the kernel operator is enforced. In general, tighter sparsity constraint entails a less restrictive bound on incoherence.

Proposition 1. *If sparsity pattern S has a sub pattern S' meaning that $S' \subset S$, then* $\mathrm{LMI}_P(x, z; \mathcal{W}, \mathcal{S}') \leq \mathrm{LMI}_P(x, z; \mathcal{W}, \mathcal{S})$ *for any $x, y \in \mathbb{R}^{n\times r}$. 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, bound on sparsity of the kernel operator can only push the necessary bound on RIP higher up. Consequently, Example 2 proves 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 $n = 2, r = 1$, but there is a straightforward extension to an arbitrary n by adding 0 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 2, after normalization, $\mathbf{A}$ would take the form $\mathbf{A} = diag(1, 1, 1, 1)$. The corresponding instance of the problem is known to have no spurious stationary points. Thus, normalization helps to improve the isometry property of the sensing operator and removed all spurious second-order stationary points out of the corresponding instance of the problem.

4.2 Sparsity + implicit structure

Unlike sparsity alone, sparsity pattern coupled with a certain type of low-dimensional structure assumed simultaneously can push the bound on RIP arbitrary close to 1. Assume $r = 1$. We illustrate this with the following proposition.

Proposition 2. *For the kernel structure operator such that*

- *sparsity structure: $\mathbf{H} = diag(H_{11}, \ldots, H_{nn}) \in \mathbb{S}^{n^2}$*
- *low-dimensional inherent structure: $H_{ii} = H_{jj}$, $i, j \in \{1, \ldots, n\}$,*

for any $z \in \mathbb{R}^n$, $\delta_2 \in [0, 1)$, it holds that every instance of the (Problem$^{\text{KSP+RIP}}$) *has no spurious second-order stationary points over $\mathbb{R}^n$.*

Under no structure assumption, the sharp upper bound on the RIP of the sensing operator is equal to $\frac{1}{2}$, while this result states that an assumption on KSP can make any upper bound on RIP redundant. It also gives an example of the problem where the necessary and sufficient conditions, given by Theorems 2 and 3, coincide.

5 Numerical results

Our main motivation and aim is the explanation of numerical success of non-convex matrix recovery in problems with structured sensing operator. Now we show how the general theory developed in Section 3 can be applied to a real-world example of such a problem, namely the power system state estimation discussed in Section 2.

In this section, we work with networks provided in the package MATPOWER 7.0b1 [41]. Keeping the structure of a network, we set the parameters of the lines equal to each other for simplicity ($r_s = 0, x_s = 1, b_c = 0, \tau = 1, \theta_{\text{shift}} = 0$).

We focus our attention on the networks from `case9` and `case14` of MATPOWER package. For `case9`, the number of buses is $n = 9$, and there are $m = 63$ possible power measurements that can be collected. We generate the sensing operator $\mathcal{A}_0^9$ to represent them all as discussed in Section 2. Although the matrix $\mathbf{A}_0^9$ has complex entries, the corresponding kernel operator $\mathcal{H}_0^9$ is represented with a real matrix due to the properties of the measurements. Both matrices $\mathbf{A}_0^9$ and $\mathbf{H}_0^9$ are visualized on Figure 1. Repetition of the non-zero entries of $\mathbf{H}_0^9$ can be considered as a form of low-dimensional structure, along with sparsity. Based on that, we form the linear operator $\mathcal{T}_0^9$. All the matrices in its kernel subspace are rank deficient. In this case, Theorem 2 can only provide us with the trivial upper bound on the RIP: $\delta < 1$. However, this operator will allow us to use Theorem 3 to bound the RIP constant from below, which turns out to be productive. Although the power system state estimation aims to find a complex vector, it is straightforward to verify that $\langle a + ib, \mathbf{H}(a + ib)\rangle = \langle a, \mathbf{H}a\rangle + \langle b, -\mathbf{H}b\rangle$ for any real vectors a and b. Therefore, it is enough to consider the problem in (7) over $\mathcal{X} = \mathcal{Z} = \mathbb{R}^n$ to make the claim regarding $\mathcal{X} = \mathcal{Z} = \mathbb{C}^n$.

In general, optimization problems (5) and (7) are hard to solve, moreover, we found that the landscape of the objective is mostly flat. Thus, we use Monte-Carlo simulations in order to obtain a numerical estimation of their solution.

A numerical test on $160,000$ trials on independent x, z uniformly distributed over $[-1, 1]^9$, showed no values of LMI_P smaller than 1. By Theorem 3, this gives us an idea that **with no regard for the RIP constant, there must be no spurious local min for any operator of the given structure. In particular, for $\mathcal{A}_0^9$ since, as we discuss further, it likely satisfies RIP with a constant strictly smaller than one.** This is analogous to the statement of Proposition 2, but for a real-world system.

In practice, some of the power measurements might be unavailable, which makes the problem harder. We further look for the limits of applicability of our method, applying it to these harder problems. In the next computational experiment, we randomly choose a nested sequence of subsets $\{S_\ell \subset 1..m\}_{\ell=1}^{11}$ of power flow measurements to be ignored. Each subset is of cardinality $|S_\ell| = \ell$, which means that the corresponding sensing operator $\mathcal{A}_\ell^9$ is represented by a matrix of the size $(63 - \ell) \times 81$. To get robust and statistically justified results, we approximate the minimum of LMI_P over the Monte-Carlo sample with an α-quantile, where $\alpha = 0.0003$ is chosen for convenience. For each value of $\ell \in 0..11$, we generate $30,000$ samples of x, z from the standard Gaussian distribution and evaluate $\text{LMI}_P(x, z; \mathcal{T}_\ell^9)$, where $\mathcal{T}_\ell^9$ is formed for $\mathcal{H}_\ell^9 = \mathcal{A}_\ell^{9T}\mathcal{A}_\ell^9$ by the same principle as $\mathcal{T}_0^9$ is formed from $\mathcal{H}_0^9$. The resulting estimation of the α-quantile is shown on figure 2a with the red solid line. The shaded region around corresponds to 95% confidence interval. The blue dashed line corresponds to the Monte-Carlo estimation of the best rank-2 restricted isometry constant, feasible for $\mathcal{A}_\ell^9$. It is estimated based on $10,000,000$ couples of Gaussian-generated vectors that formed rank-2 matrices. Maximal and minimal singular values were estimated as 0.9995 and 0.0005-quantiles, and the shaded region around corresponds to the 95%-confidence interval for the estimation. Since Monte-Carlo approach provides us with estimations on extreme quantiles for the given distribution of the samples, it is reliable enough approximation for our purposes, given the large sample size we use.

While it might appear to be constant, the blue curve slowly grows and reaches $\delta_2 = 1$ for larger ℓ. The red curve naturally goes down, reflecting increasing hardness of the problem. Besides $\ell = 0$, the red curve stays above the blue one for $\ell = 1$, going under afterwards. Thus, our approach justifies the application of the local search method to problems with sensing operators possessing structure similar to one of $\mathcal{A}_0^9$ or $\mathcal{A}_1^9$, and fails if we continue removing measurements. Since the condition we are applying is just sufficient, it does not mean that other operators necessarily generate problem

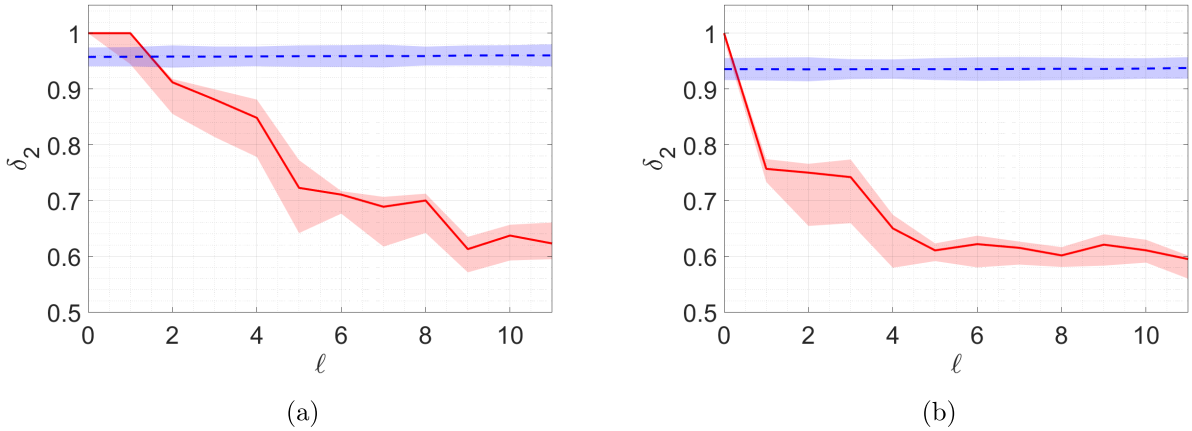


Figure 2: RIP bound, sufficient for global convergence (solid) and the RIP constant of the sensing operator (dashed) on the number of ignored measurements in case9 (a) and case14 (b)

with spurious solutions, but it can be possible to prove or disprove this fact by using other structure operators $\mathcal{T}$.

In its turn, `case14` has $n = 14$ buses, and $m = 98$ possible measurements. Similarly, we choose a nested sequence of subsets of measurements $\{S_\ell \subset 1..m\}_{\ell=1}^{11}$ to be excluded and form corresponding $\mathbf{A}_\ell^{14}$ and $\mathbf{H}_\ell^{14}$ for $\ell = 0..11$. Just like in `case9`, we define $\mathcal{T}_\ell^{14}$ such that each matrix in its kernel is forced to have repeated entries if corresponding entries were repeated in $\mathbf{H}_\ell^{14}$. Figure 2b presents a similar picture to the one for `case9`. It is obtained by the same method with all constants preserved. Important difference of `case9` is that already for $\ell = 0$ the estimation has nondegenerate confidence interval, although it is very small (~ 0.01). The picture allows us to make positive conclusion regarding structure of $\mathbf{A}_0^{14}$, but does not provides us with guarantees for larger ℓ.

Note that the proposed method, based on Theorem 3, allows to develop a guarantee of absence of spurious solution that is strictly better than the previous bound $\delta_2 = \frac{1}{2}$ by [36] for every ℓ, although it is just a sufficient condition. All the presented simulations were done with MATLAB modeling toolbox CVX [42], [43] with SDPT3 [44], [45] as the underlying solver. All optimization algorithms parameters are left at their default values.

References

[1] 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, ISSN: 1573-2916. DOI: 10.1007/BF00120662. [Online]. Available: https://doi.org/10.1007/BF00120662.

[2] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points — online stochastic gradient for tensor decomposition*, 2015. arXiv: 1503.02101 [cs.LG].

[3] 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.

[4] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 1724–1732.

[5] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, "Gradient descent can take exponential time to escape saddle points," in *Advances in neural information processing systems*, 2017, pp. 1067–1077.

[6] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods.* Siam, 2000, vol. 1.

[7] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[8] C. Cartis, N. I. Gould, and P. L. Toint, "Complexity bounds for second-order optimality in unconstrained optimization," *Journal of Complexity*, vol. 28, no. 1, pp. 93–108, 2012.

[9] N. Boumal, P.-A. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," *IMA Journal of Numerical Analysis*, vol. 39, no. 1, pp. 1–33, 2018.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] C. Yun, S. Sra, and A. Jadbabaie, "Global optimality conditions for deep neural networks," *arXiv preprint arXiv:1707.02444*, 2017.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] R. Y. Zhang, S. Sojoudi, and J. Lavaei, "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery," *arXiv preprint arXiv:1901.01631*, 2019.

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

[38] 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.

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

[40] 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.

[41] 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.

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

[43] ——, “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., `http://stanford.edu/~boyd/graph_dcp.html`, Springer-Verlag Limited, 2008, pp. 95–110.

[44] 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.

[45] 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.

Appendix

Proof of Theorems 2 and 3

Following [36], we will construct the proof by nonexistence of a counterexample. Specifically, given $\mathcal{T}$, for a point x and a parameter value z, we hope to find a value $\delta_{2r}^{x,z}$ that makes true the claim

There exists $\mathcal{A}$ that satisfies $\mathcal{T}$-KSP and δ_{2r}-RIP such that x
is a second-order stationary point of $f_{z,\mathcal{A}}$ if an only if $\delta_{2r} > \delta_{2r}^{x,z}$.

With that in mind, we construct the function $\delta(x, z)$ by the following optimization procedure:

$$\begin{aligned}
\delta(x,z) \quad \equiv \quad & \underset{\delta_{2r}\in\mathbb{R},\mathcal{A}}{\text{minimum}} && \delta_{2r} \\
& \text{subject to} && \mathcal{L}_{x,z}(\mathcal{A}^T\mathcal{A}) = 0 \\
& && \mathcal{M}_{x,z}(\mathcal{A}^T\mathcal{A}) \succeq 0 \\
& && \mathcal{T}(\mathcal{A}^T\mathcal{A}) = 0 \\
& && \mathcal{A} \text{ satisfies } \delta_{2r}\text{-RIP.}
\end{aligned}$$

Here the first two constraints represent the requirement that x is a second-order stationary point of $f_{z,\mathcal{A}}$, the third constraint takes care of the KS property, and the last one of the RI property. It is straightforward to verify that this function takes the value of the desired $\delta_{2r}^{x,z}$. Minimization of $\delta_{2r}^{x,z}$ over $\{x \in \mathcal{X}, z \in \mathcal{Z} : xx^T \neq zz^T\}$ gives $\delta_{2r}^\star$ such that the (Problem$^{\text{KSP+RIP}}$) with δ_{2r} has an instance with a spurious second-order stationary point if and only if $\delta_{2r} > \delta_{2r}^\star$.
Suppose we were able to find $\overline{\delta_{2r}^{x,z}}$ and $\underline{\delta_{2r}^{x,z}}$ such that $\underline{\delta_{2r}^{x,z}} \leq \delta_{2r}^{x,z} \leq \overline{\delta_{2r}^{x,z}}$ for all $x \in \mathcal{X}, z \in \mathcal{Z}$, then

$$\underline{\delta^\star} = \min_{\substack{x\in\mathcal{X},z\in\mathcal{Z}\\ xx^T\neq zz^T}} \underline{\delta_{2r}^{x,z}} \leq \min_{\substack{x\in\mathcal{X},z\in\mathcal{Z}\\ xx^T\neq zz^T}} \delta_{2r}^{x,z} \leq \min_{\substack{x\in\mathcal{X},z\in\mathcal{Z}\\ xx^T\neq zz^T}} \overline{\delta_{2r}^{x,z}} = \overline{\delta^\star}.$$

This inequality shows that $\delta_{2r} \geq \underline{\delta_{2r}^\star}$ is a sufficient, and $\delta_{2r} \leq \overline{\delta_{2r}^\star}$ is a necessary conditions for absence of spurious second order stationary point in the instances of the problem (Problem$^{\text{KSP+RIP}}$).

Now we will show that we can take $\mathrm{LMI}_P(x,z;\mathcal{T}) = \underline{\delta_{2r}^{x,z}}$ and $\mathrm{LMI}^T(x,z;\mathcal{T}) = \overline{\delta_{2r}^{x,z}}$. To do that, we formulate two problems that define a function LMI and re-define LMI_P:

$$\begin{aligned} \mathrm{LMI}(x,z) \equiv \underset{\delta\in\mathbb{R},\mathcal{H}}{\text{minimum}} \quad & \delta \\ \text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = 0 && \text{(8a)} \\ & \mathcal{M}_{x,z}(\mathcal{H}) \succeq 0 && \text{(8b)} \\ & \mathcal{T}(\mathcal{H}) = 0 && \text{(8c)} \\ & (1-\delta)I \preceq \mathcal{H} \preceq (1+\delta)I && \text{(8d)} \end{aligned}$$

$$\begin{aligned} \mathrm{LMI}_P(x,z) \equiv \underset{\delta\in\mathbb{R},\mathcal{H}}{\text{minimum}} \quad & \delta \\ \text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = 0 && \text{(9a)} \\ & \mathcal{M}_{x,z}(\mathcal{H}) \succeq 0 && \text{(9b)} \\ & \mathcal{T}(\mathcal{H}) = 0 && \text{(9c)} \\ & (1-\delta)\mathcal{P}^T\mathcal{P} \preceq \mathcal{P}^T\mathcal{H}\mathcal{P} \preceq (1+\delta)\mathcal{P}^T\mathcal{P} && \text{(9d)} \end{aligned}$$

where $\mathcal{P}$ is the linear operator from $\mathbb{R}^{rank([x\ z])^2}$ to $\mathbb{R}^{m^2}$ which is represented by the matrix $\mathbf{P} = \mathrm{orth}([x\ z]) \otimes \mathrm{orth}([x\ z])$. It does not depend on $\mathcal{H}$, and, given that both $\mathcal{L}_{x,z}$ and $\mathcal{M}_{x,z}$ are linear, both problems are convex, semidefinite conic programs (SDP).

Claim 1.

$$\mathrm{LMI}_P(x,z) \le \delta(x,z) \le LMI(x,z)$$

Proof. Notice that for $P = \mathrm{orth}([x,z])$, the sequence of inclusions holds

$$\{PYP^T : Y \in \mathbb{S}^{rank([x\ z])}\} \subseteq \{X \in \mathbb{S}^n : \mathrm{rank}\,(X) \le 2r\} \subseteq \mathbb{S}^n. \tag{10}$$

Denote $(\mathcal{H}^*, \delta^*)$ the minimizer of the problem corresponding to $LMI(x,z)$. By (8d), for every $X \in \mathbb{S}^n$ it holds that

$$(1-\delta^*)\|X\|_F^2 \le \langle X, \mathcal{H}^*(X)\rangle = \|\mathcal{A}^*(X)\|^2 \le (1+\delta^*)\|X\|_F^2$$

where operator $\mathcal{A}^*$ is such that $\mathcal{H}^* = \mathcal{A}^{*T}\mathcal{A}^*$ exists because $\mathcal{H}^* \succeq 0$. If the inequality holds for all $X \in \mathbb{S}^n$, it must hold for $rank(X) \le 2r$, as noticed by (10). Thus, we conclude that the pair $(\mathcal{A}^*, \delta^*)$ is feasible for the problem defining $\delta(x,z)$. This proves the upper bound.
Similarly, if $(\mathcal{A}_*, \delta_*)$ is the minimizer of the problem defining $\delta(x,z)$, then by (10), the pair $(\mathcal{A}_*^T\mathcal{A}_*, \delta_*)$ is feasible for the problem defining $\mathrm{LMI}_P(x,z)$. It can be verified after rewriting (9d) in the form

$$(1-\delta)\|PYP^T\|_F^2 \le \langle PYP^T, \mathcal{A}_*^T\mathcal{A}_*(PYP^T)\rangle = \|\mathcal{A}(PYP^T)\|^2 \le (1+\delta)\|PYP^T\|_F^2$$

for all $Y \in \mathbb{S}^{rank(x,z)}$. It's important to notice that this same argument will work with an arbitrary choice of $P \in \mathbb{R}^{n\times d}$ with $d \le 2r$. □

The claim finishes the proof of Theorem 3. To finish the proof of Theorem 2, formulate the dual of the problem defining LMI ::

$$\begin{aligned} \overline{\mathrm{LMI}}^T(x,z) = \underset{y,\lambda,U_1\succeq 0,U_2\succeq 0,V\succeq 0}{\text{maximize}} \quad & \mathrm{tr}[U_1 - U_2] && \text{(11a)} \\ \text{subject to} \quad & \mathrm{tr}[U_1 + U_2] = 1, && \text{(11b)} \\ & \mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{T}^T(\lambda) = U_1 - U_2 && \text{(11c)} \end{aligned}$$

Notice that the problem (6) is the exact reformulation of (11) (see e.g. [36, Lemma 14]), and $\overline{\mathrm{LMI}}^T(x,z) = \mathrm{LMI}^T(x,z)$. Both primal and dual problems are bounded and the dual is strictly feasible. To show the second, introduce

$$\mathbf{e} = \mathrm{vec}\,(xx^T - zz^T), \qquad \mathbf{X}\mathrm{vec}\,(u) = \mathrm{vec}\,(xu^T + ux^T) \qquad \forall u \in \mathbb{R}^{n\times r},$$

and write the operators explicitly:

$$\begin{aligned}
\mathcal{L}_{x,z} &: \mathbb{S}^{n^2} \to \mathbb{R}^{n\times r} & \mathcal{L}_{x,z}(\mathbf{H}) &= 2\cdot \mathbf{X}^T\mathbf{He}, \\
\mathcal{L}_{x,z}^T &: \mathbb{R}^{n\times r} \to \mathbb{S}^{n^2} & \mathcal{L}_{x,z}^T(y) &= \mathbf{e}y^T\mathbf{X}^T + \mathbf{X}y\mathbf{e}^T \\
\mathcal{M}_{x,z} &: \mathbb{S}^{n^2} \to \mathbb{S}^{nr} & \mathcal{M}_{x,z}(\mathbf{H}) &= 2\cdot[I_r \otimes \mathrm{mat}(\mathbf{He})] + \mathbf{X}^T\mathbf{HX}, \\
\mathcal{M}_{x,z}^T &: \mathbb{S}^{nr} \to \mathbb{S}^{n^2} & \mathcal{M}_{x,z}^T(V) &= \mathrm{vec}\,(V)\mathbf{e}^T + \mathbf{e}\mathrm{vec}\,(V)^T + \mathbf{X}V\mathbf{X}^T.
\end{aligned}$$

Strictly feasible point of (11) has the components $y = 0$, $\lambda = 0$, $V = \varepsilon I$, $U_1 = \eta I - \varepsilon W$ and $U_2 = \eta I + \varepsilon W$ where $2\eta = n^{-2}$, $2W = r[\mathrm{vec}\,(I)\mathbf{e}^T + \mathbf{e}\mathrm{vec}\,(I)^T] - \mathbf{XX}^T$, and ε is sufficiently small to ensure that both U_1 and U_2 are PSD. Consequently, Slater's condition and strong duality holds and we have $\mathrm{LMI}^T(x.z) = \mathrm{LMI}(x,z)$.

Ellipsoid norm

Throughout this subsection we treat the structured case of a block diagonal matrix $\mathbf{H} = diag(H_{11},\dots,H_{nn}) \in \mathbb{S}^{n^2}$ such that $H_{ii} = H_{jj}$ for any $i,j \in \{1,\dots,n\}$, as it appears in Theorem 2. Let's denote $H_{11} = \ldots = H_{nn} = M = QQ$. Then $\mathcal{A}(X) = QX$ and $h(x) = \|Q(xx^T - zz^T)\|_F^2 = f_{z,\mathbf{H}}(x)$ for this particular choice of $\mathbf{H}$.

First of all, we claim that the kernel operator with this structure has to be defined with a full-rank matrix Q in order to satisfy RIP for any r.

Claim 2. For any $\delta_r \in [0,1)$, the matrix $Q \in \mathbb{S}^n$ satisfies

$$(1-\delta_r)\|X\|_F^2 \le \|QX\|_F^2 \le (1+\delta_r)\|X\|_F^2 \text{ for any } X : \mathrm{rank}\,(X) \le r$$

only if $\mathrm{rank}\,(Q) = n$

Proof. Suppose $u \in \mathrm{Ker}(Q), u \neq 0$. Take $X = uu^T$ and observe that $QX = 0$, which contradicts that $(1-\delta_r)\|X\|_F^2 \le \|QX\|_F^2$. □

This allows us further consider $M \succ 0$ exclusively. By expanding $h(x+u)$:

$$h(x+u) = h(x) + \mathrm{Tr}(2x^T\left((xx^T - zz^T)M + M(xx^T - zz^T)\right)u) \quad + \\ \mathrm{Tr}\left(u^T\left((xx^T - zz^T)M + M(xx^T - zz^T)\right)u + (xu^T + ux^T)M(xu^T + ux^T))\right) + o(|u|^2)$$

we arrive with a more specified expression for the second-order necessary conditions for local optimality:

$$\langle \nabla h(x), u\rangle = 2\langle Q(xx^T - zz^T), Q(xu^T + ux^T)\rangle = 0 \qquad \forall u \in \mathbb{R}^{n\times r} \tag{12a}$$

$$\langle \nabla^2 h(x)u, u\rangle = 2\langle Q(xx^T - zz^T), uu^T\rangle + \|Q(xu^T - ux^T)\|_F^2 \ge 0 \qquad \forall u \in \mathbb{R}^{n\times r} \tag{12b}$$

The following lemma provides with a certificate that a point cannot be spurious stationary.

Lemma 1. *Given $z \in \mathbb{R}^{n\times r}$, a point $x \in \mathbb{R}^{n\times r}$ is not a first-order stationary point of the function h for any $M \succ 0$ if and only if there is $\lambda \in \mathbb{R}^{n\times r}$ such that*

$$0 \neq \mathrm{Sym}\left[(x\lambda^T + \lambda x^T)(xx^T - zz^T)\right] \succeq 0$$

Proof. Rephrase the first order condition (12a):

$$\Big((xx^T - zz^T)M + M(xx^\top - zz^\top)\Big)\, z = 0 \tag{13}$$

If for any $M \succ 0$ the equation (13) does not hold, then x cannot be a stationary point for given z and any $M \succ 0$. Consequently, the problem

$$\begin{array}{lll} \underset{M\in\mathbb{S}^n,\alpha\in\mathbb{R}}{\text{minimize}} & -\alpha & \\ \text{subject to} & \Big((yy^\top - xx^\top)M + M(yy^\top - xx^\top)\Big)\, y = 0 & \text{(14a)} \\ & M - \alpha\, I \succeq 0, & \text{(14b)} \end{array}$$

is bounded by 0 if and only if the equation (13) does not hold for arbitrary $M \succ 0$. If x can be stationary for some $M \succ 0$, then it is unbounded.

The problem (14) is an SDP problem with zero duality gap, since $M = 0$, $\alpha = -1$ is a strictly feasible primal point. Introduce the dual variables $\lambda \in \mathbb{R}^{n\times r}$ for the equality constraint (14a), $G \in \mathbb{S}^n$ for the PSD constraint (14b). The dual problem:

$$\max_{\lambda\in\mathbb{R}^{n\times r},G\succeq 0}\ \min_{M\in\mathbb{S}^n,\alpha\in\mathbb{R}} \mathrm{Tr}\Big[\Big((y\lambda^T + \lambda y^T)(yy^T - xx^T) +$$
$$+ (yy^T - xx^T)(y\lambda^\top + \lambda y^\top) - G\Big) M\Big] + \alpha(\mathrm{Tr}(G) - 1)$$

The inner optimization problem has a finite solution if and only if

$$\begin{cases} G &= (y\lambda^\top + \lambda y^\top)(yy^\top - xx^\top) + (yy^\top - xx^\top)(y\lambda^\top + \lambda y^\top) \\ \mathrm{Tr}(G) &= 1 \end{cases}$$

Rephrase the dual problem:

$$\begin{array}{lrcl} \underset{\lambda\in\mathbb{R}^{n\times r},G\in\mathbb{S}^n}{\text{maximize}} & 0 & & \\ \text{subject to} & G & = & \mathrm{Sym}[(y\lambda^\top + \lambda y^\top)(yy^\top - xx^\top)], \\ & \mathrm{Tr}(G) & = & 1, \\ & G & \succeq & 0 \end{array}$$

It is feasible if and only if the primal problem (14) is bounded. Consequently, it is feasible if and only if the point $x \in \mathbb{R}^{n\times r}$ is not a stationary point of the function h for any $M \succ 0$.

To get rid of the condition on trace, just notice that a PSD matrix has nonnegative trace, which is equal to zero if and only if the matrix is the zero matrix. All the constraints are homogeneous in G, so the trace can always be normalized to 1. Thus, the dual feasibility is equivalent to the condition $0 \neq G \succeq 0$. This concludes the proof. □

Lemma 2. *Given $z \in \mathbb{R}^{n\times r}$, a point $x \in \mathbb{R}^{n\times r}$ is not a first-order stationary point of the function h for any $M \succ 0$ if there are $T_1 \in \mathbb{R}^{r\times r}$ and $T_2 \in \mathbb{S}^r$ such that for the matrix $T = \begin{bmatrix} 0 & T_1 \\ -T_1^T & T_2 \end{bmatrix}$ it holds that*

$$0 \neq [-z \;\; x]\left(T^T P + PT\right)\begin{bmatrix} -z^T \\ x^T \end{bmatrix} \succeq 0 \tag{15}$$

where $P = \begin{bmatrix} z^T \\ x^T \end{bmatrix} [z \;\; x]$.

Proof. Suppose that there exists T from the statement of the lemma. Notice that

$$xT_1^T z^T + \frac{1}{2}xT_2x^T + zT_1x^T + \frac{1}{2}xT_2x^T =$$

$$[-z \;\; x]\begin{bmatrix} 0 & -T_1 \\ T_1^T & T_2 \end{bmatrix}\begin{bmatrix} z^T \\ x^T \end{bmatrix} = [z \;\; x]\begin{bmatrix} 0 & T_1 \\ -T_1^T & T_2 \end{bmatrix}\begin{bmatrix} -z^T \\ x^T \end{bmatrix}$$

$$xx^T - zz^T = [z \;\; x]\begin{bmatrix} -z^T \\ x^T \end{bmatrix} = [-z \;\; x]\begin{bmatrix} z^T \\ x^T \end{bmatrix}$$

Use it to expand the condition (15) in order to obtain

$$0 \neq \mathrm{Sym}[(x(zT_1 + \frac{1}{2}xT_2)^T + (zT_1 + \frac{1}{2}xT_2)x^T)(xx^T - zz^T)] \succeq 0,$$

Conclusion immediately follows by applying Lemma 1 with $\lambda = zT_1 + \frac{1}{2}xT_2$. □

Proof of Proposition 2

We start by proving that any point except for 0 and $\pm z$ cannot be a first-order stationary point of the function h.
Assume $x \notin \{0, \pm z\}$. By Lemma 2, it is sufficient to prove that there are α and β in $\mathbb{R}$ such that the matrix $T = \begin{bmatrix} 0 & \alpha \\ -\alpha & \beta \end{bmatrix}$ satisfies

$$0 \neq G = [-z \;\; x]\left(T^TP + PT\right)\begin{bmatrix} -z^T \\ x^T \end{bmatrix} \succeq 0$$

where $P = \begin{bmatrix} z^T \\ x^T \end{bmatrix}[z \;\; x]$. Consider three situations of x and z :
<u>Case 1</u> $x = \gamma z$:

$$G = [-z \;\; x]\left(T^TP + PT\right)\begin{bmatrix} -z^T \\ x^T \end{bmatrix} = 2\gamma(\gamma^2 - 1)(2\alpha + \beta\gamma)zz^Tzz^T$$

For $\alpha = \gamma(\gamma^2 - 1)$ and $\beta = 0$, it holds that $G = \left(2\gamma(\gamma^2-1)zz^T\right)^2 \succeq 0$. The matrix is nonzero for $x \notin \{0, \pm z\}$.
<u>Case 2</u> $z^Tx = 0$
The matrix P takes the form $P = \begin{bmatrix} \|z\|^2 & 0 \\ 0 & \|x\|^2 \end{bmatrix}$, so for $\alpha = 0$ and $\beta = 1$,

$$G = 2\|x\|^2xx^T \succeq 0$$

The matrix is nonzero for $x \neq 0$.
<u>Case 3</u> $0 < (z^Tx)^2 < \|z\|_2^2\|x\|_2^2$
By scaling, we can assume without loss of generality that $z^Tx = 1$, thus $P = \begin{bmatrix} \|z\|_2^2 & 1 \\ 1 & \|x\|_2^2 \end{bmatrix}$. It is sufficient to show that $T^TP + PT \succ 0$ to guarantee G to be nonzero PSD. To show this, use Sylvester's criterion. The upper left corner of this matrix is equal to -2α.

$$\det(T^TP + PT) =(-(\|x\|_2^2 - \|y\|_2^2)^2 - 4)\alpha^2 - 2(\|x\|_2^2 + \|y\|_2^2)\alpha\beta - \beta^2$$

For $\alpha = -1$, discriminant of this quadratic polynomial with respect to β is equal to $D = 16(\|z\|_2^2\|x\|_2^2 - 1)$. By the strict Cauchy-–Schwarz inequality in the assumption of the case, $D > 0$. Thus, there exists β such that the matrix is positive definite. Conclude that none of $x \notin \{0, \pm z\}$ satisfies the first order condition (1a).
Assume $x = 0$. The quadratic form on the Hessian at this point

$$\langle \nabla^2 h(0)u, u\rangle = -2\langle Qzz^T, uu^T\rangle$$

takes a negative value on $u = z$. Thus, it does not satisfy the second-order condition (1b).
The points $x = \pm z$ are not spurious points, which concludes the proof.

5.1 Proof of Proposition 1

Let $\mathbf{S}$ and $\mathbf{S}'$ be the matrix representation of $\mathcal{S}$ and $\mathcal{S}'$ accordingly. $S' \subset S$ means that $S = S' \cup S^\Delta$ and $\mathbf{S}' = \mathbf{S} + \mathbf{S}^\Delta$. It's straightforward to verify that for any $\mathbf{R} \in \mathbb{S}^{n^2}$ there exists $\mathbf{T} \in \mathbb{S}^{n^2}$ such that $\mathbf{S} \circ \mathbf{T} + \mathbf{S}^\Delta \circ \mathbf{R} = \mathbf{S}' \circ \mathbf{T}$. The opposite is also true: for any $\mathbf{T} \in \mathbb{S}^{n^2}$ there exists $\mathbf{R} \in \mathbb{S}^{n^2}$ such that $\mathbf{S} \circ \mathbf{T} + \mathbf{S}^\Delta \circ \mathbf{R} = \mathbf{S}' \circ \mathbf{T}$.

$$\begin{aligned}
&\mathrm{LMI}_P(x, z; \mathcal{W}, \mathcal{S}') = \\
&\underset{y\in\mathbb{R}^{n\times r}, V\succeq 0, u\in\mathbb{R}^l, \mathbf{T}\in\mathbb{S}^{n^2}}{\text{minimize}} \frac{\sum_{i=1}^d(-\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{W}^T(u) + \mathbf{S}'\circ\mathbf{T}))_+}{\sum_{i=1}^d(+\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{W}^T(u) + \mathbf{S}'\circ\mathbf{T}))_+} = \\
&\underset{y\in\mathbb{R}^{n\times r}, V\succeq 0, u\in\mathbb{R}^l, \mathbf{T}\in\mathbb{S}^{n^2}, \mathbf{R}\in\mathbb{S}^{n^2}}{\text{minimize}} \frac{\sum_{i=1}^d(-\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{W}^T(u) + \mathbf{S}\circ\mathbf{T} + \mathbf{S}^\Delta\circ\mathbf{R}))_+}{\sum_{i=1}^d(+\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{W}^T(u) + \mathbf{S}\circ\mathbf{T} + \mathbf{S}^\Delta\circ\mathbf{R}))_+} \le \\
&\underset{y\in\mathbb{R}^{n\times r}, V\succeq 0, u\in\mathbb{R}^l, \mathbf{T}\in\mathbb{S}^{n^2}}{\text{minimize}} \frac{\sum_{i=1}^d(-\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{W}^T(u) + \mathbf{S}\circ\mathbf{T} + \mathbf{S}^\Delta\circ\mathbf{0}))_+}{\sum_{i=1}^d(+\lambda_i(\mathcal{L}_{x,z}^T(y) - \mathcal{M}_{x,z}^T(V) + \mathcal{W}^T(u) + \mathbf{S}\circ\mathbf{T} + \mathbf{S}^\Delta\circ\mathbf{0}))_+} = \\
&\mathrm{LMI}_P(x, z; \mathcal{W}, \mathcal{S})
\end{aligned}$$