The complexity of non-convex and conic optimization problems in data science applications

by

Igor Molybog

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering & Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Javad Lavaei, Chair
Professor Anant Sahai
Professor Alper Atamturk

Spring 2022

The complexity of non-convex and conic optimization problems in data science applications

Abstract

The complexity of non-convex and conic optimization problems in data science applications

by

Igor Molybog

Doctor of Philosophy in Engineering - Industrial Engineering & Operations Research

University of California, Berkeley

Professor Javad Lavaei, Chair

Mathematical optimization is the cornerstone for the data-driven design and exploitation of various large-scale cyber-physical and information systems. The generic off-the-shelf tools for guaranteed global optimization used today often have a prohibitively high computational complexity on practical instances, preventing their convergence within a reasonable time. Instead, highly-scalable heuristics are utilized with no guarantee of global optimality of their result, which poses a problem in the analysis of complex safety-critical systems. Thus, specialized highly-scalable algorithms that come with a guarantee of their performance are required. These can be constructed by providing guarantees and limits of applicability for the well-performing heuristics. In this dissertation, we study how to leverage system-specific characteristics to analyze computational heuristics leading to guarantees of global optimality of their outputs. A prominent example of a large-scale safety-critical system that we consider throughout the first part of the dissertation is the sensor network within the measurement system of an electrical power grid. It is known that the associated generic problem of recovery of the state of a power system is computationally complex yet critical. The three main ways to reduce the complexity of computation associated with a sensor network are to increase the density of sensors, place them strategically, or enhance their security and accuracy to limit the set of possible errors and faults of measurement. All three of these paths are investigated within the first part of this dissertation. We provide numerical quantification to the computation complexity of inverse problems depending on the total number of measurements, the properties of the graph structure of the measurement network, and the signal-to-noise ratio measured as the ratio between the numbers of good and bad measurements. We address both the computational complexity and the matter of constructing a suitable, efficient algorithm in each of the three scenarios. Our results offer implications that should be considered at multiple stages throughout the life cycle of a system, starting from the design of the sensing mechanisms to ensure robustness and security of operation both in normal conditions and in situations of an emergency such as during a cyber-attack.

In the second part of this dissertation, we present a pioneering work considering an automatic approach toward selecting the most suitable efficient heuristic algorithm for an optimization problem at hand. We focus on a machine learning technique that can be adopted for heuristic selection and investigate its properties and the guarantees of its performance from the perspective of mathematical optimization. We conclude that adopting this machine learning technique in its original form would not result in guaranteed optimal heuristic selection even in a benign scenario, propose modifications and outline future work toward the automatic selection of heuristics.

To my parents Oleg and Lena

Contents

List of Figures

List of Tables

Acknowledgments

Here, I would like to express my sincere gratitude to everyone who has helped me make this dissertation possible.

In the first place, I would like to thank my advisor, Professor Javad Lavaei, for his guidance, inspiration, and motivation throughout my Ph.D. years. I am grateful to him for being a hands-on mentor, for making his research vision and expertise always available to me, enriching the content of my research and this dissertation in particular. I am deeply indebted to Javad for providing so many research opportunities, promoting my skill set, and giving me a chance to grow as an independent researcher.

I am also indebted to my wonderful collaborators. Ramtin Madani has been deeply involved in my first research projects, giving me a hand when I needed it the most and playing a significant role in forming my perspective on science and research. I am grateful to Somayeh Sojoudi for making available her research intuition, great ideas on new topics, and her kind help throughout my Ph.D. years.

I want to express my sincere gratitude to Alper Atamturk, who I always look up to as a researcher and a teacher, for his immense help and support. I would like to thank Anant Sahai, Anil Aswani, and Paul Grigas for providing invaluable feedback regarding my academic work. I was fortunate enough to work with Ilan Adler, and I am grateful to him for sharing his wisdom and perspective on teaching and research.

My collaboration with DeepMind Alberta and especially Nolan Bard have significantly broadened my vision of AI research, impacted my later study reported in the second part of this dissertation, and helped me determine my career preferences.

I am happy to have worked in a brilliant research group and a department with students and post-docs of outstanding proficiency and personality. My gratitude goes to Richard Y. Zhang, Cedric Josz, Salar Fattahi, who I look up to like my academic elder brothers, and Han Feng and SangWoo Park, who were always there for me to discuss research or to share in an adventure. I want to thank my co-authors, Reza Mohammadi-Ghazi and Ming Jin, as well as the peers from my home IEOR department Mahbod Olfat, Georgios Patsakis, Pedro Hespanhol, Yuhao Ding, Julie Mulvaney-Kemp, Baturalp Yalcin, Donghao Ying, and Haixiang Zhang.

I am grateful to Maxim Zubkov, Tahsin Saffat, Yonah Borns-Weil, and the rest of my peers from the department of Mathematics for accepting me into their circles, making their department a second home for me, and for giving me both the nerdy and the fun hours whenever I needed them.

I am most grateful to the entire community of Russian Speaking Student Association at UC Berkeley, including Nikita Samarin, Alexey Sinyashin, Georgy Grigorev, Sasha Avdoshkin, Sasha Tsigler, and Vlad Kozii, for becoming like a family to me over the years and for helping me in keeping my work-life balance healthy. A special shoutout to Boris Sobolev for the unforgettable pandemic adventures and beyond.

It is impossible to put in words my sincere gratitude towards my Parents Lena and Oleg, my brother Yaroslav and the rest of my family, whose warmth of love and support has been

keeping me striving through all these years. Thank my dear friends Denis Pylypenko, Anton Nikolaev, Amelia Rowland, Dmytro Rzhemosky, Andrey Ryazanov, Yuliia Lut, and others for their tremendous contribution to my life and personality over the years of my work.

Chapter 1

Introduction

Even under the ideal condition of no noise and zero approximation error, many highly-efficient techniques for data analysis involve solving potentially challenging or intractable computational problems while learning from data. They are tackled by heuristic optimization algorithms based on relaxations or greedy principles in practice. The lack of guarantees on their performance limits their use in applications with a high 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.

This dissertation focuses on addressing the challenge of designing efficient, highly scalable, and at the same time, provably safe and reliable algorithms for the solution to the big data challenges of the twenty-first century. The rapid development of automation in heavy industry, agriculture, transportation, energy, and everyday life implies the growing interconnectivity and complexity of our society's systems: sensor and communication networks, infrastructure, governmental organizations, etc. This trend leads to new opportunities, massively increasing the amount of data available for processing by the decision-makers on the one hand. On the other hand, this introduces new correlations, meaning higher risks for errors and failures to quickly propagate and amplify through the systems, causing significant disruptions (traffic jams in transportation, goods shortages in supply chains) or even disasters (blackouts in power systems, pandemics in healthcare and public policy). The situation demands conceptually new computational tools to be developed for robust analytics and control of the systems on an increasingly large scale, driving the probability of an error to zero at the cost of large-scale computation. The work presented here aims at confronting this grave challenge.

In the following sections of this chapter, we first formally introduce the core subject of our study: optimization problems. Then, we motivate our work and provide a general introductory overview of the practical problems considered in this dissertation, along with a brief summary of our contributions. Next, we provide the relevant publications, notations, and definitions that are used throughout the dissertation. We conclude this chapter by presenting the Semidefinite Affine Rank Feasibility problem that will be the core subject throughout the first part of this dissertation.

1.1 Comprehensive Framework of Optimization Problem

The questions of numerical optimization and its computation complexity are in the core of this dissertation. An instance of a minimization problem can be written in the form

$$\text{minimize}_{x\in\mathcal{X}}\, f(x)$$

where

- $\mathcal{X}$ is the domain of the instance, which in this dissertation is supposed to be a set of multivariate decision variables $x \in \mathcal{X} \subseteq \mathbb{R}^n$. For example, when x is the vector of voltage magnitude and phases on the nodes of a power system, $\mathcal{X}$ consists of the voltage regimes that are safe to operate the power system in. In case x represents a control policy for a dynamical system, the set $\mathcal{X}$ restricts the space of policies only to those stabilizing the system. In the first part of this dissertation, the set $\mathcal{X}$ is normally assumed to be the space of solutions of a system of inequalities without loss of generality.
- $f : \mathcal{X} \to \mathbb{R}$ is the objective function of the instance. For example, it may return the operational cost of a particular power system, the probability of misclassification by a computer vision algorithm on a given dataset or it may capture some notion of robustness in a dynamical system. A pair $(\mathcal{X}, f)$ uniquely defines the instance of a minimization problem.

An optimization problem is a set of instances. We will consider problems where instances can be parametrized with $\theta \in \Theta$ by assuming the domain to be dependent on the parameter $\mathcal{X} = \mathcal{X}(\theta)$ as well as the objective function $f(x) = f(x;\theta)$. Here the space of parameters Θ is an infinite-dimensional vector space consisting of vectors with finite support, i.e. for each $\theta \in \Theta$ there is $p \in \mathbb{Z}_+$ such that $\theta \in \mathbb{R}^p$. The number of bits necessary for encoding of θ is called its size. Using this notation, a minimization problem (P) can be written formally as

$$(P) \qquad \{\text{minimize}_{x\in\mathcal{X}(\theta)}\, f(x;\theta) \;\big|\; \theta \in \Theta\}$$

If the parametrizations of the domain sets X and objective functions f is fixed, the vector θ uniquely defines a minimization instance and the set Θ uniquely defines a minimization problem. For example, for the general problem of Optimal Power Flow (OPF) the parameter set Θ contains an encoding for any of the possible admittance matrices of electrical grids. For the problem of OPF for the electrical grids with tree topology, the set Θ contains only the encodings for the admittance matrices of tree-like electrical grids. When the instances of two problems (P_1) and (P_2) have the same parametrizations of the domain sets $\mathcal{X} = \mathcal{X}(\theta)$ and the objective functions $f(x) = f(x;\theta)$ and the problems' parameter sets are nested $\Theta_1 \subset \Theta_2$ we say that (P_1) is a *sub-problem* of (P_2). In applications, the parameter θ is the available data.

A class of optimization problems is a set of problems $(\mathcal{X}(\theta), f(x,\theta), \Theta)$. An important class of optimization problems is the class of polynomial-time solvable (or tractable) problems which consists of the problems $\{\text{minimize}_{x\in\mathcal{X}(\theta)} f(x;\theta) \;\big|\; \theta \in \Theta\}$ such that there exists an algorithm $a : \Theta \to \mathbb{R}^n$ which runs in polynomial time of the size of the input θ and returns the solution to $\text{minimize}_{x\in\mathcal{X}(\theta)} f(x;\theta)$. Here we assume that both $\mathcal{X}(\theta)$ and $f(x,\theta)$ have a compact description compared to the size of θ. There are plenty of practical problems that belong to this class and even more of those conjectured to be in the complement of it.

1.2 Motivation and the Summary of Contributions

Unfortunately, the most powerful problems that are useful in the modeling of a wide range of processes and phenomena in the real world are believed not to be polynomial-time solvable. For example, the mixed-integer linear programming problem (MIP), often used to model electrical generation, production, and scheduling does not have a known efficient algorithm for its solution. In practice, however, the modeling often does not require the full power of MIP and can be formulated as a simpler sub-problem. Thus, the logistics and scheduling tasks can often be modeled through tractable sub-problems of MIP: the linear programming problem (LP) and the minimum cost network flow (MCNF). Thus, this dissertation is devoted to the study of $\mathcal{NP}$-hard optimization problems and their polynomial-time solvable sub-problems that arise in the modeling of interconnected systems and statistical learning tasks. We focus on the three main reasons for a practical problem to be tractable: the amount of data, the strength of the available signal compared to the noise, and the presence of sparsity structure in the problem. In the first part, we quantitatively characterize their effect on the complexity of the practical problems and propose and evaluate efficient algorithms for their solution. We work on the vast space of tasks that can be formulated in the form of a Semidefinite Affine Rank Feasibility problem. In the second part, we propose a general paradigm to automatically recognize and solve tractable practical problems and study a prominent method that is aligned with it.

- **Chapter 2**

  In many engineering applications increasing the amount of data makes a statistical inference problem easier to solve, even in the absence of noise. In some cases, the transition from an intractable to a tractable problem occurs smoothly, while in others, it is a leap. Since different problem instances require different numbers of variables to be estimated, it is fair to directly compare only those having the same dimension n of the domain $\mathcal{X}(\theta) \in \mathbb{R}^{n(\theta)}$. The data in our formalism is enclosed in θ, and thus, the amount of data now can be measured through the relation between the size of the parameter vector θ and the dimension of the domain n. To quantify the impact of the amount of data, we would need to determine the complexity of the problem composed of the instances for which the size of the parameter vector is larger than a certain threshold that depends on n. In Chapter 2, we consider solving the feasibility problem

via its convex relaxation as it is a vital tool in analyzing a wide variety of systems. Convex relaxations are typically used as a heuristic for the solution of optimal power flow problems and their variations, mixed-integer programming formulations arising in logistics, scheduling, and production planning, and the approximation of max-cut problems in integrated circuit design. In the past, profound results from algebraic geometry and measure theory allowed the development of a systematic approach to the convexification of general-purpose polynomial optimization, namely the Lasserre hierarchy. However, this method has found limited applications in practice, in part, due to the complicated analysis of computational complexity under a high data volume regime. In Chapter 2, we provide an in-depth analysis of an alternative systematic method of convexification. In particular, we consider semidefinite programming relaxations of the problem with systematically constructed objective functions and studied their properties. In particular, we proposed an analytical bound on the number of relaxations that are sufficient to solve in order to obtain a solution of a generic instance of the semidefinite affine rank feasibility problem or prove that there is no solution to it. This bound is followed by a heuristic algorithm based on semidefinite relaxation and an experimental testament of its performance on a large sample of synthetic data. As one of the main contributions of this paper, we were able to observe the information-computation tradeoff in this problem and captured how the increasing amount of data causes the problem to go from $\mathcal{NP}$-hard to tractable in the process of a phase transition.

- **Chapter 3**

  The cybersecurity of electrical grids is considered to be one of the critical problems of the 21st century. The aftermath of a power grid emergency took billions of dollars to repair in the past (e.g., major blackouts of 1977, 2003, and 2019 in Northern America). With the advancement of automation and the transition to smart grids, the stakes become even higher. Among the most dangerous threats to a power system is the failure of a control unit, which can lead to inefficient or even self-destructing modes of operation. While it is possible to impose high-security standards for the central controller, enforcing the protection of all of the periphery infrastructure, such as electrical current measurement devices, is often infeasible, making manipulating the measurements a potential threat. This leaves vulnerability in the system's defense against a hacker attack, as even a perfectly functional control algorithm could be tricked into self-adversarial behavior given corrupted data. This vulnerability cannot be offset by the standard provably robust data processing methods as they are developed for linear systems, which is often not an acceptable model for the power grid. In Chapter 3, we model the event that an unknown part of a power grid is being attacked by an adversary infusing noise into the measurement of the power flow to impact the decision-making process that follows the solution of the state estimation problem. The amount of noise in this setting can be measured by the number of measurements affected, while the signal is proportional to the total number of measurements as well as the quality of the prior guess of the solution if one is available. The principal question in this setting

is the signal-to-noise ratio that can be tolerated by a tractable problem without producing an error in the solution. To measure it quantitatively, we first develop the first tractable algorithms to learn the states of a nonlinear (quadratic) system subject to sparse adversarial noise, which is well-suited for power flow analysis. Then, we analyze the precision of the recovered state in instances with different relationships between the number of corrupted measurements, the distance between the prior guess and the actual solution, and the total number of measurements. It is important to note that even without noise, quadratic regression is an $\mathcal{NP}$-hard problem, and introducing adversarial noise into the system makes it even harder in general. The proposed algorithms are constructed to tolerate sparse errors of arbitrary magnitude. They are based on penalized semidefinite and second-order convex relaxations. Under the assumption of the data being Gaussian, we were able to theoretically prove the exactness of these relaxations by adapting the technique of primal-dual witness. The practical efficacy of the developed methods was demonstrated on synthetic data and the data of the power system state estimation (PSSE) problem for the European power grid (with over 1300 buses) under injected sparse adversarial noise combined with dense white noise.

- **Chapter 4**

  Acquiring more correct measurements seems to help to make the hard statistical estimation problems easier to solve. However, sometimes the number of measurements is strictly bounded and cannot be increased for fundamental reasons. For example, this is the case for many applications possessing a network structure. Luckily, such structure itself may be leveraged to identify a relevant sub-problem and device a tractable algorithm for its solution. A bright example is the min-cost network flow, which is a tractable sub-problem of the mixed-integer linear programming problem. What is the structure induced by power systems which allows local search algorithms to solve the problems of their analysis efficiently? Chapter 4 describes an answer to that question. There, we used tools from conic optimization to develop necessary and sufficient conditions for the inexistence of spurious local solutions in the non-convex matrix sensing formulation. We developed the novel notion of kernel structure property and conditions that are likely to be satisfied in instances arising from the analysis of sparse networks, such as power and transportation systems. We used our approach to analytically prove that the low-rank Q-norm minimization problem does not possess spurious local minima and also experimentally demonstrated that one should not expect spurious local solutions in problems that possess low-dimensional structures.

- **Chapter 5**

  The mappings learned from data have advanced significantly in the areas of computer vision, natural language processing, and robotics. Through an iterative procedure of an optimization algorithm or heuristic, a machine learning system selects parameters for the mapping between the input and the output of an algorithm. Machine's ability to tune to the thin structure of every individual problem, obtaining a better result using a

minimal amount of time and resources, is remarkable and makes it a preferred method of algorithm design. Before machine learning techniques became widely available, the tasks of algorithm design for these areas were carried out by field researchers designing pre-processing procedures and feature maps for individual pattern recognition problems. These tasks are very similar to the problem we have been solving manually throughout the first part of this dissertation, which brings up a question: How can we make a machine learning system that learns the structure of an optimization problem from a dataset of instances, and picks the algorithm for it? A related question has already been addressed by the researchers inside the machine learning community, where the optimization problem of interest is the problem of training a model. The corresponding domain is called meta-learning, and the dominant technique there is the Model-Agnostic Meta Learning (MAML). In Chapter 5 we study this technique from the perspective of designing optimization algorithms that converge to the globally optimal solution. We show that the original form of MAML is guaranteed to succeed in selecting a correct algorithm if the instances of the optimization problem of interest have a benign landscape and do not differ much from one another. For other scenarios, we conclude that MAML does not fit the task, propose a modification to the scheme of the MAML algorithm and summarise further directions of research.

1.3 Related Publications

- **Chapter 2**

  Main paper:

  - Igor Molybog and Javad Lavaei. "On Sampling Complexity of the Semidefinite AffineRank Feasibility Problem". In:Proceedings of the AAAI Conference on Artificial Intelligence. 2019, pp. 1568–1575.

- **Chapter 3**

  Main paper:

  - Igor Molybog, Ramtin Madani, and Javad Lavaei. "Conic Optimization for Quadratic Regression Under Sparse Noise", *Journal of Machine Learning Research (JMLR)*, 2020

  Related paper:

  - Igor Molybog, Ramtin Madani, and Javad Lavaei, "Conic Optimization for Robust Quadratic Regression: Deterministic Bounds and Statistical Analysis", *Conference on Decision and Control (CDC)*, 2018

- **Chapter 4**

  Main paper:

  - Igor Molybog, Somayeh Sojoudi, and Javad Lavaei. "Role of sparsity and structure in the optimization landscape of non-convex matrix sensing", *Mathematical Programming*, 2020

  Related paper:

  - Igor Molybog, Somayeh Sojoudi, and Javad Lavaei. "No Spurious Solutions in Non-convex Matrix Sensing: Structure Compensates for Isometry", *American Control Conference (ACC)*, 2021

- **Chapter 5**

  Main paper:

  - Igor Molybog and Javad Lavaei, "When Does MAML Objective Have Benign Landscape?", *Conference on Control Technology and Applications (CCTA)*, 2021

1.4 Definitions and Notation

Notation

Sets: The cardinality of a set S is indicated as $|S|$. For a sequence S, the symbol $\{\alpha_i\}_{i\in S}$ denotes a sequence indexed by S. Whenever the index set is obvious from the context, we drop the subscript for simplicity of the notation.

Vectors: The symbols $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers, while $\mathbb{R}^n$, $\mathbb{R}^n_+$ and $\mathbb{C}^n$ denote the sets of real, real nonnegative and complex n-dimensional vectors, respectively. In any finite-dimensional vector space, 1 is a vector whose entries are all equal to 1, while 0 is the zero vector and I represents the identity operator or its matrix. With v_i we denote the i-th component of a vector v, and denote with v^i the i-th vector in the sequence of vectors $\{v^i\}_{i\in A}$ or the i-th column or row in the matrix V, which will be understood from the context. Thus, e^i stands for the i-th column of the identity matrix I of an appropriate dimension. Given a vector v, the symbols $\|v\|_\infty$ $\|v\|_2$, $\|v\|_1$ and $\|v\|_0$ denote the maximum norm, the Euclidian norm, the Manhattan norm and the cardinality of the support of v, i.e., the number of its nonzero elements. Given a vector $v \in \mathbb{C}^n$ and an index set $S \subset \{1, \dots, n\}$, the vector a_S is defined to be a subvector of a obtained by collecting together the entries of a with indexes from S.

Matrices: Let $\mathbb{R}^{n\times k}$ and $\mathbb{C}^{n\times k}$ denote the sets of real and complex $n \times k$ matrices. The symbols $\mathbb{H}^n$ and $\mathbb{S}^n$ denote the sets of $n \times n$ Hermitian and symmetric matrices. Let the symbol $\mathbb{L}^{n;k} = \{V \in \mathbb{R}^{n\times k} \,|\, V_{ij} = 0 \text{ if } i < j\}$ denote the set of lower triangular matrices. $M \succeq 0$ means that the matrix M is symmetric and positive semidefinite, while $M \succeq_{\mathcal{C}} 0$ means that M belongs to the cone $\mathcal{C}$. Let $\mathbb{S}^{n;r}_+$ be the set of $n \times n$ positive semidefinite matrices of

rank r and let $\mathbb{T}^{n;r} \subset \mathbb{S}_+^{n;r}$ be the set of matrices in $\mathbb{S}_+^{n;r}$ whose nonzero eigenvalues are all equal to 1 (Stiefel manifold).

Given a matrix $M \in \mathbb{C}^{n\times k}$, its singular values are denoted as $\sigma_1(M), \ldots, \sigma_{\min\{n,k\}}(M)$ in descending order. $\sigma_{min}(M) = \sigma_{\min\{n,k\}}(M)$ and $\sigma_{max}(M) = \sigma_1(M)$ are the smallest and the largest singular values. $[M]_{ij}$ and M_{ij} denote the (i,j)-component of M. Given a set $S \subset \{1, \ldots, k\}$, the matrix M_S is defined to be a matrix obtained by adjoining the columns of M with indexes in S. Given two index sets A and B, the submatrix $M_{[A,B]}$ contains the entries of M located on the intersection of rows indexed by A and columns indexed by B. The symbols $\|M\|_1$, $\|M\|_\infty$, $\|M\|_2$, and $\|M\|_F$ denote the maximum absolute column sum, maximum absolute row sum, maximum singular value, and Frobenius norm of M, respectively. The transpose, conjugate transpose and Moore-Penrose pseudoinverse of a matrix M are shown as $M^\top$, M^* and M^+. The matrix vectorization operator $\mathrm{vec} : \mathbb{C}^{n\times k} \to \mathbb{C}^{nk}$ stacks the columns of a matrix into a vector. $\mathrm{vecnd} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n^2-n}$ is a non-diagonal vectorization operator that puts non-diagonal entries of the argument matrix into the form of a vector. $\mathrm{veclt} : \mathbb{C}^{n\times k} \to \mathbb{C}^{nk-\frac{k(k-1)}{2}}$ is a lower triangular vectorization operator that puts the elements on and below the diagonal of the argument matrix into the form of a vector. For any two compatible matrices A and B, their Frobenius inner product is $\langle A, B\rangle$, their Hadamard (entrywise) product is $A \circ B$ and their Kronecker product is $A \otimes B$.

Given a square matrix $M \in \mathbb{C}^{n\times n}$, its symmetric part is denoted with $\mathrm{Sym}(M) = (M + M^\top)/2$ and its trace is denoted with $\mathrm{tr}(M)$. For square matrices $M_1, M_2, \ldots, M_n$, the matrix $diag(M_1, \ldots, M_n)$ is block-diagonal with M_i on the diagonal. The eigenvalues of a matrix $M \in \mathbb{H}^n$ are denoted as $\lambda_1(M), \ldots, \lambda_n(M)$ in descending order. $\lambda_{min}(M)$ is the eigenvalue of M with the smallest absolute value.

Operators: For a linear operator $\mathcal{L} : \mathbb{R}^{n\times k} \to \mathbb{R}^m$, the adjoint operator is denoted by $\mathcal{L}^T : \mathbb{R}^m \to \mathbb{R}^{n\times k}$. The matrix $\mathbf{L} \in \mathbb{R}^{m\times nk}$ such that for any $x \in \mathbb{R}^{n\times k}$ it holds that $\mathcal{L}(x) = \mathbf{L}\mathrm{vec}\,(x)$ is called the *matrix representation* of the linear operator $\mathcal{L}$. Bold letters are reserved for matrix representations of corresponding linear operators. The null space of an operator or a matrix that represents it is denoted with $\mathrm{Ker}(\mathcal{L})$ or $\mathrm{Ker}(L)$.

Probability: Given a probability space, let $\mathbb{P}[E]$ be the probability of an event E in the space, $\mathbb{E}[x]$ the mathematical expectation of a random variable x. The notation $x \sim \mathcal{N}(\alpha, \beta)$ means that x is a normally distributed random variable with the parameters α and β. A multivariate normal distribution with mean vector μ and covariance matrix Σ is denoted as $\mathcal{N}(\mu, \Sigma)$.

Functions: Given the sequences $a(n)$ and $b(n)$, the notation $a(n) \sim \mathcal{O}(b(n))$ means that there exists a number $C \geq 0$ such that $a(n) \leq Cb(n)$ for all $n \geq 1$. Positive part is denoted with $(\cdot)_+ = \max\{0, \cdot\}$

Special definitions: *Sparsity pattern* S of a set of matrices $\mathcal{M} \subset \mathbb{R}^{n\times k}$ is a subset of $\{1,\dots,n\}\times\{1,\dots,k\}$ such that $(i,j)\in S$ if and only if there is $M\in\mathcal{M}$ such that $M_{ij}\neq 0$. Given a sparsity pattern S, define its matrix representation $S\in\mathbb{R}^{n\times k}$ as

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

and its operator representation $\mathcal{S}(M) = S\circ M$.

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

$$P=\operatorname{orth}(M) \quad\Longleftrightarrow\quad PP^TM=M,\ P^TP=I_{\operatorname{rank}(M)}.$$

For a set $S=\{a_1,\dots,a_{|S|}\}\subseteq\{1,\dots,n\}$ define a permutation $\pi_A=(1,a_1),\dots(k,a_k)$ of $\{1,\dots,n\}$ and its matrix $\Pi_{n;S}$, i.e. the matrix that permutes the entries of a vector it acts upon according to the permutation $\pi_{n,S}$.

For $\omega\in\mathbb{R}^{n\times k}$ and $R\in\mathbb{R}\cup\{+\infty\}$, we define $\mathbb{B}_R(\omega)=\{a\in\mathbb{R}^{n\times k} \ : \ \|a-\omega\|_F\leq R\}$, $\bar{\mathbb{B}}_R(\omega)=\{a\in\mathbb{R}^{n\times k} \ : \ \|a-\omega\|_F< R\}$ and $\partial\mathbb{B}_R(\omega)=\{a\in\mathbb{R}^{n\times k} \ : \ \|a-\omega\|_F= R\}$. It follows from the definition that $\partial\mathbb{B}_{+\infty}(\omega)=\emptyset$.

General Definitions

In this dissertation, we consider a point $x\in\mathbb{R}^n$ to be a solution for the instance $\operatorname{minimize}_{x\in\mathcal{X}} f(x)$, saying that it is a feasible solution (feasible point) if $x\in\mathcal{X}$. The value of a solution is $f(x)$ if defined. We distinguish between a locally optimal solution (local solution) $x\in\mathcal{X}$ such that $f(x)\leq f(x')$ for all x in the neighborhood $U_x\subset\mathcal{X}$ of x, and a globally optimal solution (global solution) $x\in\mathcal{X}$ such that $f(x)\leq f(x')$ for all $x\in\mathcal{X}$. We denote the value of a globally optimal solution with $\min_{x\in\mathcal{X}} f(x)$ if the set of globally optimal solutions of an instance denoted with $\operatorname{Arg\,min}_{x\in\mathcal{X}} f(x)$ is not empty. A solution that is localy optimal but not globally optimal is called spurious.

Suppose $\mathcal{X}$ is the set of points $x\in\mathbb{R}^n$ satisfying a system of inequalities of the form $f_i(x)\leq 0$ for $i\in\{1,\dots,m\}$. If the objective function f and all of the constraint functions f_i are continuously differentiable at a feasible solution $x'\in\mathcal{X}$ and the following conditions hold for some nonnegative vector $\mu\in\mathbb{R}^m_+$

$$\begin{cases} \nabla f(x') + \sum_{i=1}^m \mu_i \nabla f_i(x') = 0 \\ \sum_{i=1}^m \mu_i f_i(x') = 0 \end{cases}$$

then the solution x' is called a first-order stationary point. In case for the same x' and μ it also holds that

$$s^\top \left[\nabla^2 f(x') + \sum_{i=1}^m \mu_i \nabla^2 f_i(x')\right] s \geq 0$$

is satisfied by every nonzero $s \in \mathbb{R}^n, s \neq 0$ such that $s^\top \nabla f_i(x') = 0$ for all $i \in \{1, \ldots, m\}$ then x' is called a second-order stationary point. It is important within the scope of this dissertation that local search methods commonly have guaranteed convergence to a first- or second-order stationary point. It is also known that a locally optimal solution of an optimization problem has to be a stationary point if a constraint qualification condition is held at the solution. When considering non-convex landscapes, we will focus on smooth problems where constraint qualifications hold everywhere over the domain, therefore we will assume that the sets of first-order critical points, second order critical points, local minima and global minima are nested.

Affine Rank Feasibility Problem

The first three chapters are concerned with the problem of recovery of a real $n \times n$ symmetric and positive semidefinite matrix X of rank k that satisfies m linear specifications. In case there is no such a matrix up to a given accuracy, we expect to receive the corresponding information within a finite time. We refer to this problem as the *Semidefinite Affine Rank Feasibility (SARF)*, which can be formally written as:

$$\begin{aligned} &\text{find} && X \in \mathbb{S}^n \\ &\text{subject to} && \langle M_r, X \rangle = y_r, \quad r = 1, \ldots, m, && \text{(1.1a)} \\ & && X \succeq 0, && \text{(1.1b)} \\ & && \operatorname{rank}\{X\} = k. && \text{(1.1c)} \end{aligned}$$

where $M_1, \ldots M_m \in \mathbb{S}^n$ are some known symmetric matrices, $y_1, \ldots, y_m \in \mathbb{R}$ are some scalars that parametrize the problem, and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product.

The above problem is closely related to the *Affine Rank Feasibility (ARF)* problem, which can be obtained by dropping the constraint (1.1b). It is clear that a solution to the Semidefinite Affine Rank Feasibility problem must also be a solution to its ARF relaxation. In addition, every ARF problem can be equivalently converted to a SARF problem [34, 68].

Both of the above feasibility problems encompass several practically important cases. For example, the Phase Retrieval [18] and Quantum State Tomography [59] problems can be treated as special cases of the Semidefinite Affine Rank Feasibility problem with rank-one matrices M_r and $k = 1$. In Power Systems Engineering, the State Estimation problem consists in recovering complex voltages based on measured power flows and select voltage magnitudes for a power grid, which can be formulated as the Semidefinite Affine Rank-one Feasibility problem [121]. In this case, the matrices M_r depend on the topology of the power system and change a few times a day, while the measurements y_r change every 5-15 minutes based on the data updated by system operators. This defines a class of problems with a fixed set of matrices M_r's. In wireless communication systems, the problem of Feasible Downlink Beamforming [75] can be formulated as a block-diagonally shaped Affine Rank-one Feasibility problem. The most commonly studied problem with the target rank being not necessarily equal to 1 is the Low-rank Matrix Completion problem, for which each matrix M_r has exactly one nonzero element in its upper/lower triangular part [19].

While the *Affine Rank Minimization (ARM)* [89] is probably the most studied rank-constrained problem, the solution of any Affine Rank Minimization problem can also be obtained from the solutions of $\mathcal{O}(n)$ Affine Rank Feasibility problems of the same size. The same is also true for those problems with a positive semidefinite constraint, so the results of this chapter can be extended to rank minimization problems as well. From the point of view of computational complexity, the Affine Rank Minimization is an $\mathcal{NP}$-hard problem, since it contains Cardinality Minimization as a sub-problem [78]. Moreover, the work by Marianna, Laurent, Varvitsiotis, et al. [70] establishes the result that the rank-constrained completion of a semidefinite matrix with all diagonal entries equal to 1 is $\mathcal{NP}$-hard as soon as $k \geq 2$. Hence, the Semidefinite Affine Rank Feasibility is also $\mathcal{NP}$-hard.

Nevertheless, the above problems have been extensively studied over the past decade, and practical approaches have been developed in various special cases. A major line of research assumes that the matrices M_r form a linear operator that satisfies a Restricted Isometry Property. For example, those matrices whose elements are independently sampled from the standard Gaussian distribution satisfy this property with high probability. Classical papers with this assumption adopt Nuclear Norm Minimization [90, 19, 89, 15, 59], while more recent developments deploy the minimization of other surrogate functions [28]. In particular, the minimization of a rank function has received a lot of attention for applications related to artificial intelligence. For example, Xu, Lin, and Zha [112] proposes a surrogate of the Schatten-p norm as a spectral regularization, while Zhang and Zhang [117] attacks the problem of rank-constrained distance matrices. For the Low-rank Matrix Completion problem, a number of nonconvex techniques have been developed recently. More importantly, Bhojanapalli, Neyshabur, and Srebro [12], Ge, Lee, and Ma [42], Ge, Jin, and Zheng [41], and Zhang et al. [119] show that the classical l_2-norm regression has no spurious local minimum under a Restricted Isometry Property. In addition, Josz et al. [53] prove a similar result for nonsmooth problems, including l_1-norm regression. One possible option to avoid the Restricted Isometry Property assumption is to require the distribution of the true solution X^* to belong to a certain ensemble [111].

Given $\phi \in \Phi$ and an arbitrary $\varepsilon > 0$, it follows from the Stone-Weierstrass theorem that there exists a polynomial $p_a : \mathbb{R}^n \to \mathbb{R}$ that uniformly approximates f on $\mathcal{X}$ with the precision error of ε. This way, given the data $\theta = \{(y_i, \phi_i)\}_{i=1}^m$, there exists a nonlinear regression model

$$y_i = p_{\phi_i}(x) + \hat{\varepsilon}_i, \qquad \forall i \in \{1, ..., m\}$$

where each function $p_{\phi_i}(x)$ is a polynomial and $\hat{\varepsilon}_i$ is the difference between $p_{\phi_i}(x)$ and $f(x, \phi_i)$ that is bounded from above by ε. Notice that $\hat{\varepsilon}$ is dense noise of a small value that we do not consider in this paper since its presence just shifts the solutions recovered using our methods by a small value that can be naturally bounded (this corresponds to the sensitivity analysis of conic optimization). On the other hand, each polynomial equation can be converted to a quadratic equation by introducing new variables and adding new quadratic equations [93]. As an example, the polynomial equation $1 = x^4 - x^3 + x$ can be written as $1 = z^2 - xz + x$ with the additional variable z and measurement equation $0 = z - x^2$ (note that the number

of variables and constraints increases in a logarithmic fashion in terms of the degree of the polynomial). This discussion implies that every nonlinear regression could be approximated up to any arbitrary precision with a quadratic regression where the augmented model of the system is quadratic. For this reason, the focus of this paper is only on quadratic regression.

As a far more general case of phase retrieval, a quadratic regression problem with the variable x can be modeled as $f(x; A_i) = x^* A_i x$. The state estimation problem for power systems belongs to the above model due to the quadratic laws of physics (i.e., the quadratic relationship between voltage and power), where each matrix A_i has rank 1 or 2. Robust regression in power systems is referred to as *bad data detection*. This problem was first studied in 1971 by [73], and there are many recent signs of progress on this topic [31, 109, 69].

In the context of the electric power grid, the regression problem is known as state estimation, where the goal is to find the operating point of the system based on the voltage signals measured at buses and power signals measured over lines and at buses [1, 69, 121].

Part I

The Effect of Data, Signal and Structure on the Complexity of Semidefinite Affine Rank Feasibility

Chapter 2

Sampling Complexity of the Noiseless Problem

In this chapter, we study a generic semidefinite affine rank feasibility (SARF) problem, which consists in finding a positive semidefinite matrix of a given rank from its linear measurements. We aim to quantify the dependence between the complexity of the problem and the number m of linear measurements available in it. We measure the complexity with respect to the algorithm based on solving a number of semidefinite programming problems of the dimension n of the desired matrix. The complexity is higher the more semidefinite problems are required to solve within the algorithm. In this chapter, we propose the first nontrivial analytical upper bound on this number which depends on the number m of generic linear measurements in the SARF. Our theoretical results suggest that the problem transits to a polynomial-time solvable state when it has at least an order of n^2 of measurements.

Besides analytical bound, we propose a randomized version of the algorithm and study its performance on a large sample of synthetic data. We obtain the approximate characteristic for the point of "phase transition" of a uniformly sampled problem. It turns out that for the uniform distribution over instances, the point of transition behaves linearly with n, which suggests that our theoretical bound can be improved if the problem parameters θ are assumed to come from a specific distribution.

2.1 Introduction

Methods of solution of the SARF sub-problems vary depending on the assumed noise model. Besides Gaussian additive noise [50, 74], a particularly interesting case is the presence of sparse noise, which has been considered by Candès et al. [20], Klopp, Lounici, and Tsybakov [57], and Akhriev, Marecek, and Simonetto [3] and will be addressed in the next chapter. However, even under the absence of any noise, SARF remains a difficult problem if the amount of information is not sufficient.

The key question studied in this chapter is the question of sampling complexity, which

can be interpreted as the question on the amount of information contained in the linear measurements (1.1a). This can be measured in terms of the number of measurements m and/or the sampling strategy of selecting M_r's such that the problem becomes polynomial-time solvable. The known results on this topic are related to the Matrix Completion problem [19, 108]. The present chapter develops the first result in the literature that studies the sampling complexity of a general Semidefinite Affine Rank Feasibility problem.

This work builds upon an important contribution into the Low-rank Matrix Recovery that has been made by Madani et al. [68], where conic relaxations of linear matrix inequalities (LMIs) are proposed and it is shown that there is a low-rank solution for any sparse LMI with an upper bound on the rank being a function of certain graph-theoretic parameters such as the treewidth. Note that any LMI Rank Feasibility problem is equivalent to a Semidefinite Affine Rank Feasibility problem. From this perspective, there is another application of Semidefinite Affine Rank Feasibility problem, which can be found in reducing the complexity of large-scale semidefinite programs [40, 5].

2.2 Problem Formulation

To make the computational complexity analysis of Semidefinite Affine Rank Feasibility problem (1.1) meaningful, we consider a feasibility problem specified by the measurement matrices $\{M_1, \ldots, M_m\}$ and the measurements $y_1, \ldots, y_m$. This defines infinitely many feasibility instances, where every rank-k positive semidefinite matrix X is a solution to some feasibility problem in this class (by appropriately designing $y_1, \ldots, y_m$). The analysis of this problem is partially motivated by data analytics for electric power systems, where the matrices M_r are designed based on the parameters of the infrastructure that are considered to be fixed (as long as there is no network reconfiguration) while the measurements $y_1, \ldots, y_m$ used by power operators change every 5-15 minutes. In this regard, we have infinitely many feasibility problem instances with the same matrices $M_1, \ldots, M_m$.

The results of this chapter are based on the idea of constructing a semidefinite programming (SDP) of the form

$$\begin{aligned} \underset{X \in \mathbb{S}^n}{\text{minimize}} \quad & \langle M, X \rangle && \text{(2.1a)} \\ \text{subject to} \quad & \langle M_r, X \rangle = y_r, \quad r = 1, \ldots, m, && \text{(2.1b)} \\ & X \succeq 0, && \text{(2.1c)} \end{aligned}$$

corresponding to the Semidefinite Affine Rank Feasibility problem (1.1). Here, M is the matrix of an orthogonal projection onto an $(n-k)$–dimensional linear subspace. In this case, the convex problem (2.1) is called an **SDP relaxation with the parameters** (M, y). A solution $X^\star$ of the Semidefinite Affine Rank Feasibility problem (1.1) is said to be **recoverable** through the SDP relaxation (2.1) with M if it is the unique optimal solution of (2.1) with the parameters $(M, [\langle M_r, X^\star \rangle]_{r=1}^m)$.

The **recovery region** of the matrix M associated with the class $\{M_1, \ldots, M_m\}$ is the set of all rank-k positive semidefinite matrices that are recoverable for their corresponding feasibility problems via the SDP relaxation with the objective function $\langle M, X \rangle$. Figure 2.1 depicts a two-dimensional slice of the recovery region of a single matrix M for a randomly generated problem (see the section "Numerical Results" for more details). It can be observed that the recovery region is not convex in general, but there is a ball with a positive radius in the space of matrices such that every matrix in the ball is recoverable via an SDP relaxation with the single objective matrix M for the corresponding Affine Rank Feasibility problem. This observation was first noticed and formalized in Theorem 1 of Ashraphijuo, Madani, and Lavaei [6].

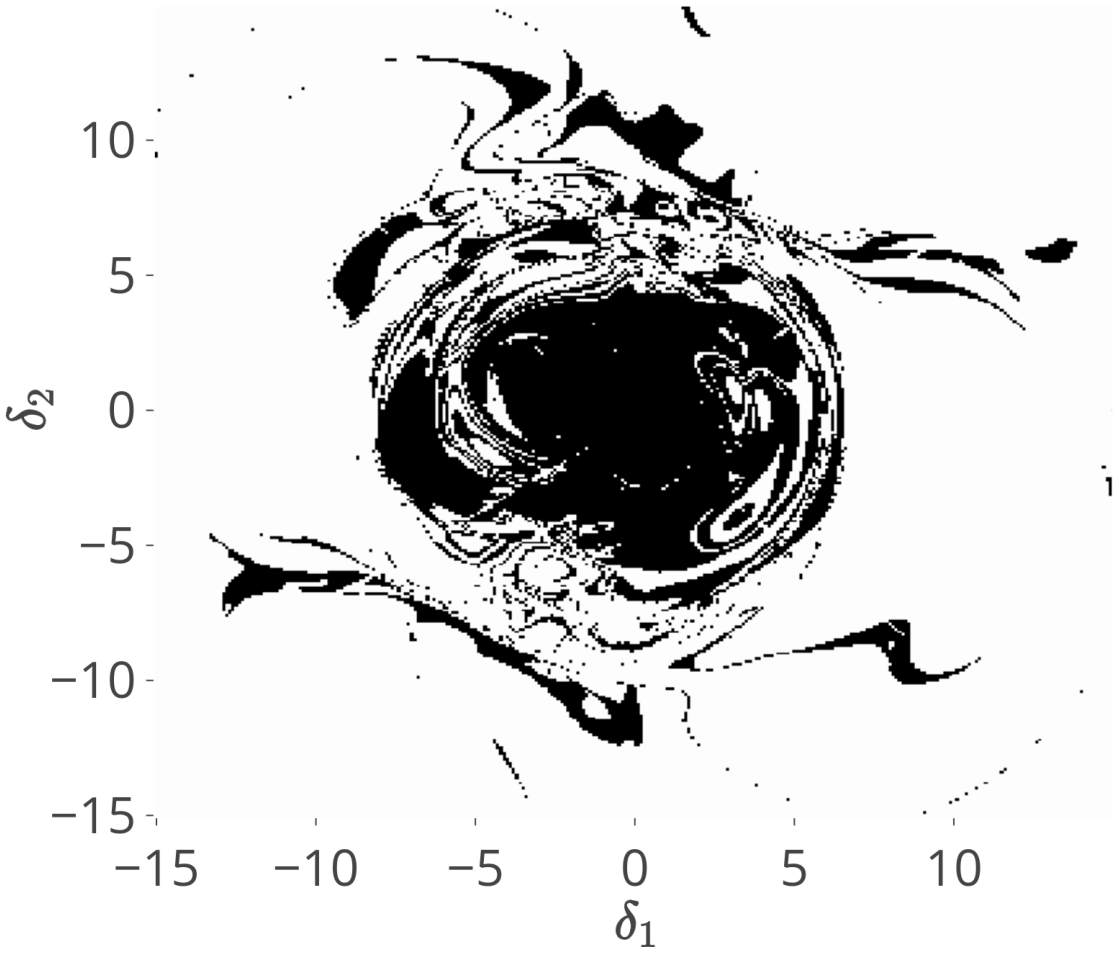


Figure 2.1: A slice of the recovery region of a single matrix. The x and y axes represent δ_1 and δ_2, respectively (for definition, see "Implementation" in the section "Numerical Results").

One of the main results of Ashraphijuo, Madani, and Lavaei [6] is that under $m > k(2n - k)$, there are a finite number of SDPs with specially designed objectives such that

every instance of the Semidefinite Affine Rank Feasibility problem (1.1) in the infinite class defined by $\{M_1, \ldots, M_m\}$ is recoverable via one of those SDPs. In other words, there is a finite list of matrices M defining a set of SDPs that can be used to solve any of the infinity many feasibility problems sharing the same model. The practical application of this technique has been demonstrated by Ashraphijuo, Madani, and Lavaei [7] for solving a set of polynomial equations. The paper by Madani, Lavaei, and Baldick [66] applies the above technique to the power flow / state estimation problem for power systems, where a small number of SDPs have solved many real-world feasibility problems and outperformed the existing methods.

The main objective of this chapter is to find an upper bound on the number of SDP problems (2.1) to be solved in order to guarantee obtaining a solution of (1.1). Since we do not make any assumption on the structure of the matrices $\{M_1, \ldots, M_m\}$, the upper bound cannot be a polynomial function but the methodology pursued in this chapter could be used to study specialized problems (such as those appearing in power systems) to obtain a tighter upper bound for structured problems. Although we do not make any assumption on the uniqueness of the solution of (1.1), it is well known that the solution is unique when m is large enough.

2.3 Main Results

Remember that veclt is a lower triangular vectorization operator and form a matrix $H \in \mathbb{R}^{(nk-\frac{k(k-1)}{2})\times m}$ with the vectors $\text{veclt}(M_r)$ as its columns. To set up a uniform bound on the size of a region that can be recovered through a single objective M, define the following function of the measurements matrices:

$$r(H) \quad = \quad \min_{\substack{Y\in\mathbb{L}^{n;k}:YY^\top\in\mathbb{T}^{n;k};\\ A\subseteq\{1..n\}:|A|=k}} \sigma_{min}([\text{veclt}(\Pi_{n;A}M_1\Pi_{n;A}^\top Y)\ldots\text{veclt}(\Pi_{n;A}M_m\Pi_{n;A}^\top Y)])$$

Note that, according to Ashraphijuo, Madani, and Lavaei [6], it holds that $r(H) > 0$ if $m > k(2n-k)$, for a generic choice of $\{M_r\}_{r=1}^m$. This implies that the inequality holds true for almost every choice of $\{M_r \in \mathbb{S}^n\}_{r=1}^m$. A formal definition comes next.

Definition 1. *A property (Q) is said to hold for every generically chosen member of a topological space if there exists an open dense subset of it whose members all satisfy (Q).*

The main analytic result of this chapter will be stated below, which sets up a bound on the sufficient number of SDP relaxations.

Theorem 1. *Given an arbitrary positive number κ, there are constants $C_1 = C_1(k, \{M_r\})$ and $C_2 = C_2(k, \{M_r\}, \kappa, \|y\|_2) \in \mathbb{R}$ and at most*

$$\min\left\{C_1^{k(n-k)}, C_2^{\frac{n(n+1)}{2}-m+1}\right\}$$

SDP relaxations of the form (2.1) such that any solution (satisfying $\frac{\sigma_1(\cdot)}{\sigma_k(\cdot)} \leq \kappa$) of a generic instance of the Semidefinite Affine Rank Feasibility problem (1.1) in the class of infinitely many feasibility problems defined by $\{M_1, ..., M_m\}$ can be obtained via one of those SDPs.

Proof. This is a direct corollary of Theorems 2 and 3 in the section "Proofs". □

It can be observed that for a fixed k the bound in Theorem 1 is at most exponential in terms of the dimension of the problem for a generic choice of the matrices M_r, while all known finite-time algorithms for the general ARF problem have at least doubly exponential running times [89]. Moreover, the bound has a dependence on m, which implies the following corollary:

Corollary 1. *For every polynomial $p(\cdot)$, if the number of measurements obeys the inequality*

$$m \geq \frac{n(n+1)}{2} - \log p(n) \sim \mathcal{O}(n^2),$$

then a generic problem becomes polynomial-time solvable with any fixed sensing constant κ and up to an arbitrary nonzero precision error.

This captures the fact that the problem is expected to be easy when m is large because the feasible region of (2.1) for $m = \frac{n(n+1)}{2}$ collapses into at most a single point that should be the solution to the SARF problem (1.1).

The above result is the first one that studies the notion of "phase transition of complexity" for the generic rank-constrained feasibility problem. This notion has already been studied in the literature for other $\mathcal{NP}$-hard problems [27]. We will continue this topic during the discussion of Numerical Results.

2.4 Proofs of the main result

The dual problem of (2.1) can be written in the form:

$$\underset{u \in \mathbb{R}^m}{\text{minimize}} \qquad y^{\mathrm{T}} u \tag{2.2a}$$

$$\text{subject to} \qquad M + \sum_{i=1}^{m} u_i M_i \succeq 0 \tag{2.2b}$$

Since the primal and dual feasible matrices must be positive semidefinite, the Karush-Kuhn-Tucker (KKT) conditions impose the relationship:

$$\left(M + \sum_{r=1}^{m} u_r M_r \right) X = 0. \tag{2.3}$$

Without loss of generality, we can assume that the vectors $\mathrm{veclt}(M_r)$ form an orthonormal basis. To support this, suppose that the original problem (1.1) is associated with the measurement matrices $\hat{M}_r$ ($\langle \hat{M}_r, X\rangle = \hat{y}_r$). One can apply the Gram-Schmidt orthogonalization process to the system of vectors $\mathrm{veclt}(\hat{M}_r)$ to obtain an orthonormal system of vectors that would generate symmetric matrices M_r in the obvious way. Note that M_r has a linear dependence on $\hat{M}_r$ and, therefore, $y_r = \langle M_r, X\rangle$ can be computed by applying the same linear transformations to $\hat{y}_r$. Furthermore, now we can refer to m as to the maximum number of linearly independent matrices M_r.

The next lemma states that each matrix of a potential solution can be represented through a projection matrix onto the span of its eigenvectors.

Lemma 1. *For any orthogonal matrix U, if $X = UU^\top$ is not a singular point of $[\langle M_r, \cdot\rangle]_{r=1}^m$ and recoverable with M, then $X' = U\Lambda U^\top$ is also recoverable with M for any positive definite diagonal matrix Λ.*

Proof. Strong duality holds for the pair of problems (2.1)-(2.2) due to Lemma 4 by Ashraphijuo, Madani, and Lavaei [6]. Let (X, u) be a primal-dual optimal pair of the SDP relaxation (2.1) with the parameters $(M, [\langle M_r, X\rangle]_{r=1}^m)$. Then, the equation (2.3) is still satisfied after being multiplied by X' on the right. It takes the form $(M + \sum_{r=1}^m u_r M_r)\, X' = 0$, which implies that Complementary Slackness holds for the SDP relaxation (2.1) with the parameters $(M, [\langle M_r, X'\rangle]_{r=1}^m)$ together with primal and dual feasibility. □

Due to Lemma 1, it is enough to only consider the case when a solution $X^\star$ of (1.1) to be found belongs to $\mathbb{T}^{n;k}$. We assume $X^\star \in \mathbb{T}^{n;k}$ henceforth, except for the proof of Theorem 2.

Consider a matrix $X \in \mathbb{T}^{n;k}$, the set of indexes $A \subseteq \{1, \dots n\}$ ($|A| = k$) of linearly independent columns of X, and its Cholesky embedding $\mathcal{C}_{n;A} : \mathbb{S}_+^{n;A} \to \mathbb{L}^{n;k}$ defined as

$$\mathcal{C}_{n;A}(X) \triangleq \left[\ L^{\mathrm{T}},\quad L^{-1}X_{[A,\{1,\dots,n\}\setminus A]}\ \right]^{\mathrm{T}}$$

where $LL^{\mathrm{T}} = X_{[A,A]}$ is the Cholesky decomposition of $X_{[A,A]}$. Using the Guttman rank additivity formula, it is possible to show that $X^\star = \Pi_{n;A}^\top \mathcal{C}_{n;A}(X^\star)\mathcal{C}_{n;A}(X^\star)^\top \Pi_{n;A}$. We will rewrite this factorization in the form

$$X^\star = CC^\top \tag{2.4}$$

Let us introduce

$$J = [\mathrm{veclt}(\Pi_{n;A} M_1 C) \quad \dots \quad \mathrm{veclt}(\Pi_{n;A} M_m C)]^\top,$$

It follows directly from the form of the pushforward function of the mapping that J has full column rank if and only if $X^\star$ is not a singular point of the mapping $[\langle M_r, \cdot\rangle]_{r=1}^m$ [6]. Consider the dual vector

$$u^\star = -(J^+)^{\mathrm{T}}\mathrm{veclt}(MC)$$

By multiplying both sides of this equation by $J^\top$, it is easy to observe that if J has full column rank, then $u^\star$ solves the equation (2.3). Let R be a matrix whose columns form an orthonormal basis in the space that is orthogonal to the span of the columns of C.

Lemma 2. *Assume that $X^\star$ is not a singular point of $[\langle M_r, \cdot\rangle]_{r=1}^m$. The pair $(X^\star, u^\star)$ is the primal-dual optimal pair for the SDP problem (2.1) with the parameters $(M, [\langle M_r, X^\star\rangle]_{r=1}^m)$ if and only if $\|R^\top F R\|_2 \leq 1$ for the matrix $F \in \mathbb{S}^n$ defined through the equation*

$$\mathrm{veclt}(F) = H(J^+)^\top \mathrm{veclt}(MC)$$

Proof. Primal feasibility of $X^\star$ is obvious. Complementary slackness (2.3) is satisfied by the construction of $u^\star$. Now show the dual feasibility of $u^\star$:

$$M + \sum_{r=1}^m u_r^\star M_r = M + \sum_{i\geq j} E_{ij} f_{ij}$$

where E_{ij} is a matrix with the only nonzero entries equal to 1 in the (i,j) and (j,i) locations (or just (i,i)), while $f_{ij} = m_{ij}^\top u^\star$ with $m_{ij} = [M_{ij}^1 \quad \dots \quad M_{ij}^m]^\top$. Note that $m_{ij}^\top = \mathrm{veclt}(E_{ij})^\top H$ and

$$f_{ij} = -\mathrm{veclt}(E_{ij})^\top H (J^+)^\top \mathrm{veclt}(MC)$$

It follows from (2.3) that the dual matrix has k zero eigenvalues in the subspace of the span of $X^\star$. We study the minimum eigenvalue of the matrix in the rest of the space:

$$\begin{aligned}
&\lambda_{\min}(M + \sum_{r=1}^m u_r^\star M_r) = \\
1 - \max_{v:\|v\|_2=1; C^\top v = 0} &2\sum_{i\geq j} v_i v_j \mathrm{veclt}(E_{ij})^\top \mathrm{veclt}(F) - \\
&\sum_i v_i^2 \mathrm{veclt}(E_{ii})^\top \mathrm{veclt}(F) = \\
&1 - \max_{v:\|v\|_2=1; C^\top v = 0} \sum_{i,j=1}^n [vv^\top]_{ij} F_{ij} = \\
&1 - \max_{\phi\in\mathbb{R}^{n-k}:\|\phi\|_2=1} \mathrm{trace}(\phi^\top R^\top F R \phi),
\end{aligned}$$

which is greater than or equal to zero if and only if $\|R^\top F R\|_2 \leq 1$. □

Note that if $m = \frac{n(n+1)}{2}$, then $\mathrm{rank}(H) = \frac{n(n+1)}{2}$, and it is possible (e.g., using a Kronecker product representation) to show that $F = \Pi_{n,A}^\top MCC^+$, so $\|RFR^\top\|_2 \leq 1$, (this is expected since the only feasible point of the SDP relaxation should be the solution to the SARF problem).

Lemma 3. *If $MX = 0$, then*

$$\|MC\|_F \leq \sqrt{k}\|X - X^\star\|_2$$

Proof.

$$\begin{aligned}\|MC\|_F = \sqrt{\mathrm{trace}(MX^\star M)} = \sqrt{\mathrm{trace}(MX^\star X^\star M)} = \\ \|MX^\star\|_F \leq \sqrt{k}\|MX^\star\|_2 = \sqrt{k}\|M(X^\star - X)\|_2 \leq \\ \sqrt{k}\|M\|_2\|X^\star - X\|_2 \leq \|X^\star - X\|_2\sqrt{k}\end{aligned}$$

□

The previous lemma and the definition of $r(H)$ lead to the result stated below.

Corollary 2. *If $MX = 0$, then then the matrices R and F in Lemma 2 satisfy the inequality*

$$\|RFR^\top\|_2 \leq \frac{\sqrt{k}}{r(H)}\|X^\star - X\|_2$$

The above result will be used to prove Theorem 1. In a normed vector space $(V, \|\cdot\|)$, let the symbol $\mathbb{B}_{V,\|\cdot\|}(r)$ stand for the closed r-ball centred at zero

$$\mathbb{B}_{V,\|\cdot\|}(r) = \{v \in V \ : \ \|v\| \leq r\}$$

. Define $\Phi : (\mathbb{S}^n, \|\cdot\|_F) \to (\mathbb{S}^n, \|\cdot\|_F)$ such that $\Phi(X) = XX^+$ and $\Xi_{n;k}(t) = \{X \in \mathbb{B}_{\mathbb{S}^n,\|\cdot\|_2}(1) \cap \mathbb{S}_+^{n;k} : \sigma_k(X) \geq t\}$, where $\sigma_k(\cdot)$ is the k-th largest singular value of a symmetric matrix.

Lemma 4. *For every $t > 0$, the operator Φ is Lipschitz over $\Xi_{n;k}(t)$ with the constant $L = \frac{1}{t}$*

Proof. It is known that the projection of $X \in (\mathbb{S}^n, \|\cdot\|_F)$ onto $B_{\mathbb{S}^n,\|\cdot\|_2}(1)$ is given by the matrix X' which can be obtained from X by replacing with 1 all eigenvalues that are greater than 1. Thus, XX^+ can be viewed as the $\|\cdot\|_F$-projection of $X \in \Xi_{n;k}(1)$ onto the convex set $B_{\mathbb{S}^n,\|\cdot\|_2}(1)$. Consequently, Φ is Lipschitz with the constant 1 over $\Xi_{n;k}(1)$. Consider $X, Y \in \Xi_{n;k}(t)$:

$$\begin{aligned}\|\Phi(X) - \Phi(Y)\|_F = \\ \|t^{-1}X\left(t^{-1}X\right)^+ - t^{-1}Y\left(t^{-1}Y\right)^+\|_F \leq \\ \|t^{-1}X - t^{-1}Y\|_F = t^{-1}\|X - Y\|_F\end{aligned}$$

□

Note that $\Xi_{n;1}(1) \supset \mathbb{T}^{n;1}$, so the function Φ has the Lipschitz constant 1 over the entire $\mathbb{T}^{n;1}$.

Theorem 2. *The number*

$$\left(\frac{3\kappa\sqrt{k}}{r(H)}\frac{\max\{\|y\|_2^2,1\}}{\min\{\|y\|_2^2,1\}}\right)^{\frac{n(n+1)}{2}-m+1}$$

is an upper bound on the number of SDP relaxations needed to find a solution with the property $\frac{\sigma_1(\cdot)}{\sigma_k(\cdot)} \leq \kappa$ *for every instance of the Semidefinite Affine Rank Feasibility problem defined by* $\{M_1, ..., M_m\}$.

Proof. By Lemma 1, Lemma 2 and Corollary 2, given $A \in \mathbb{T}^{n;k}$, every $X^\star \in \mathbb{S}_+^{n;k}$ with the property $\|(X^\star)^+X^\star - A\|_2 \leq \frac{r(H)}{\sqrt{k}}$ is recoverable through M such that $MA = 0$. Now, we aim to compute the covering number for the set of all possible solutions to the problem. Define

$$X(x) = x_0\sum_{r=1}^{m} y_r M_r + \sum_{r=1}^{\frac{n(n+1)}{2}-m} x_r K_r$$

and

$$\mathcal{S} = \{X(x)\big|x \in \mathbb{R}^{\frac{n(n+1)}{2}-m+1}\},$$

where $\{K_r \in \mathbb{S}^n\}$ are normalized as vectors that are orthogonal to $\{M_r\}$ and to each other. $X(\{x_0 = 1\})$ includes the set of all possible solutions to the SARF. By considering $X^\star = X(x^\star)$ as the solution to be found, $\frac{X^\star}{\|X^\star\|_2} = X(\frac{x^\star}{\|X^\star\|_2})$ belongs to $\mathcal{S}$. In light of Lemma 1, if $X(\frac{x^\star}{\|X^\star\|_2})$ belongs to the recovery region of M, then $X^\star$ belongs to it as well. Therefore, it is enough to cover $\mathcal{S} \cap B_{\mathbb{S}^n,\|\cdot\|_F}(\sqrt{k}) \cap \Xi_{n;k}(\frac{\sigma_k(X^\star)}{\|X^\star\|_2})$ with recovery regions to guarantee that $X^\star$ lies in one of them.

Lemma 4 yields that $\|X(\frac{x^\star}{\|X^\star\|_2})X(\frac{x^\star}{\|X^\star\|_2})^+ - A\|_2 \leq \frac{\|X^\star\|_2}{\sigma_k(X^\star)}\|X(\frac{x^\star}{\|X^\star\|_2}) - A\|_F$. Therefore, it is sufficient to cover $\mathcal{S} \cap B_{\mathbb{S}^n,\|\cdot\|_F}(\sqrt{k}) \cap \Xi_{n;k}(\frac{\sigma_k(X^\star)}{\|X^\star\|_2})$ with the $\|\cdot\|_F$-balls of radius $\frac{r(H)\sigma_k(X^\star)}{\sqrt{k}\|X^\star\|_2}$. After noticing

$$\|X(x)\|_F^2 = x_0^2(\|y\|_2^2 - 1) + \|x\|_2^2 \geq \|x\|_2^2 \min\{\|y\|_2^2, 1\},$$
$$\|X(x)\|_F^2 \leq \|x\|_2^2 \max\{\|y\|_2^2, 1\},$$

we conclude

$$\mathcal{S} \cap B_{\mathbb{S}^n,\|\cdot\|_F}(R) \subset X\left(\{\|x\|_2 \leq \frac{R}{\sqrt{\min\{\|y\|_2^2,1\}}}\}\right)$$

$$X\left(\{\|x\|_2 \leq \frac{r}{\sqrt{\max\{\|y\|_2^2,1\}}}\}\right) \subset \mathcal{S} \cap B_{\mathbb{S}^n,\|\cdot\|_F}(r).$$

It is known that the covering number of the ball $B_{\mathbb{R}^{\frac{n(n+1)}{2}-m+1},\|\cdot\|_2}(R)$ with balls of radius r obeys the bound $\left(\frac{3}{r}R\right)^{\frac{n(n+1)}{2}-m+1}$ [82]. Applying the function $X(\cdot)$ to this cover, one can

obtain that $\mathcal{S} \cap B_{\mathbb{S}^n,\|\cdot\|_F}(\sqrt{k}) \cap \Xi_{n;k}(\frac{\sigma_k(X^\star)}{\|X^\star\|_2})$ belongs to the union of

$$\left(\frac{3\sqrt{\max\{\|y\|_2^2, 1\}}}{\frac{r(H)\sigma_k(X^\star)}{\sqrt{k}\|X^\star\|_2}} \frac{\sqrt{k}}{\sqrt{\min\{\|y\|_2^2, 1\}}} \right)^{\frac{n(n+1)}{2}-m+1}$$

balls having radius $\frac{r(H)\sigma_k(X^\star)}{\sqrt{k}\|X^\star\|_2}$ in Frobenius norm. This concludes the proof. □

Theorem 3. *There is an absolute constant C such that*

$$\left(\frac{C\sqrt{k}}{r(H)} \right)^{k(n-k)}$$

is an upper bound on the number of SDP relaxations needed to solve every instance of the Semidefinite Affine Rank Feasibility problem in the class defined by $\{M_1, ..., M_m\}$.

Proof. Due to Proposition 8 by Szarek [99], there is an absolute constant C such that the covering number of the Grassmann manifold $\mathcal{G}_{n;k}$ obeys the inequality

$$M_{\mathcal{G}_{n;k}}(\varepsilon) \leq \left(\frac{C\ diam(\mathcal{G}_{n;k})}{\varepsilon} \right)^{k(n-k)}.$$

Similarly to the proof of Theorem 2, the diameter of the Grassmann manifold for the l_2-induced norm is equal to 1. Therefore, the proof is completed by noting that $\varepsilon = \frac{r(H)}{\sqrt{k}}$. □

2.5 Numerical results

In this section, we present and study a randomized algorithm for solving the SARF problem via an SDP relaxation that is based on the theoretical results of this chapter. Algorithm 1 iteratively solves SDP relaxations of the problem with randomly sampled objective matrices. Under the assumption that no prior information is available about the unknown solution, we sample M uniformly since it belongs to the compact set $\mathbb{T}^{n;k}$ that is isomorphic to the Grassmann manifold $\mathcal{G}_{n;k}$.

We present experimental results on the performance of Algorithm 1 on a large set of synthetic data. The main goal of these experiments is to study the dependence between the probability of success of the convex (SDP) procedure and the number of linearly independent measurements in the problem. For a number of values of m and k, we sample a random problem with the data $\{M_r\}_{r=1}^m$ and a random solution $X^\star$ (from which we design $y_1, ..., y_m$), and aim to solve it with Algorithm 1. After a successful ending, we sample another dataset together with a solution and then start over. After constructing and solving 300 SDP relaxations for a particular value of m, we proceed to the next value. The details on the sampling strategy are given below in the Implementation paragraph, and the results are

Algorithm 1 Heuristic algorithm for solving the Semidefinite Affine Rank Feasibility problem (1.1)

Require: $\{M_r, y_r\}_{r=1}^m$
initialization
for all $t \in \{1, \ldots, T\}$ **do**
  Sample a random M
  X = solution of (2.1) with M
  if $rank(X) = k$ **then**
    return X
  end if
end for

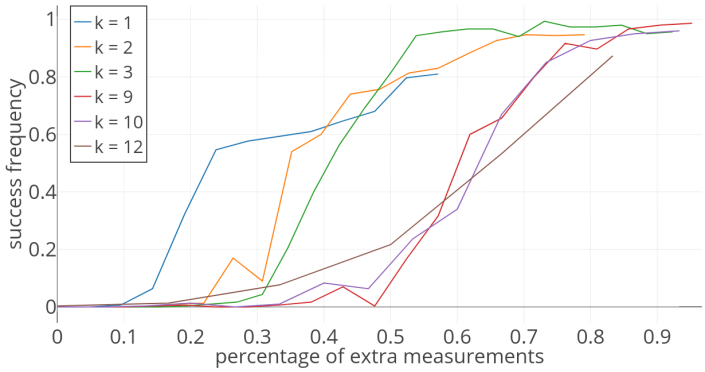
summarized in Figure 2.2. It can be observed that it is easy to design an SDP that recovers the true solution even for those values of m that are much smaller than $n(n+1)/2$. Notice that the frequency of recovery decreases linearly at first, and then turns to exponential at a certain point, which appears to be a constant loosely related to n but closely connected to the value of k. This shows the existence and characterizes the behavior of the point of "phase transition" of the problem from easy to hard, at least with respect to the considered algorithm.

Implementation To build the region for the example in Figure 1, we randomly sample $M_r \in \mathbb{S}^n$ and the matrix $X \in \mathbb{S}_+^{n;k}$ following the procedure to be explained later. Afterwards, we select M in such a way that $MX = 0$. Let $Q_0 \in \mathbb{R}^{n \times k}$ be the matrix with orthonormal columns such that $X = Q_0 Q_0^\top$. In this notation, y_r is obtained as follows: $y_r = \langle M_r, Q(\delta_1, \delta_2) Q(\delta_1, \delta_2)^\top \rangle$, where $Q(\delta_1, \delta_2) = Q_0 + \delta_1 E_{00} + \delta_2 E_{01}$. We solve problem (2.1) with the parameters (M, y) for different values of δ_1 and δ_2, compare the result to the matrix $Q(\delta_1, \delta_2) Q(\delta_1, \delta_2)^\top$, and mark the points where they almost coincide.

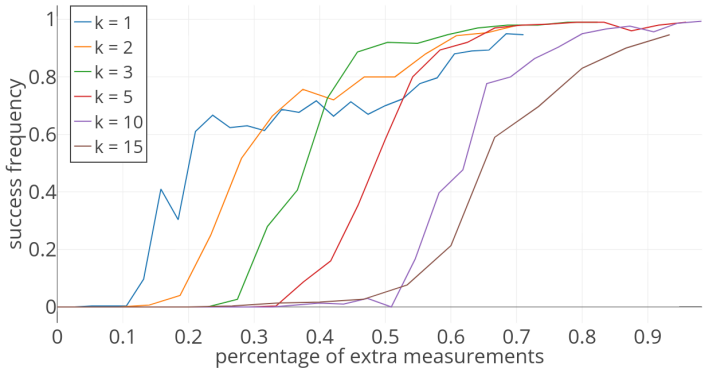
Now, let us turn to the generation procedures. $u(\{1, \ldots, n\})$ denotes the uniform distribution over the set $\{1, \ldots, n\}$; the uniform distribution over all $n \times n$ orthogonal matrices, named $O(n)$ Haar distribution, is denoted as Haar$(O(n))$.

- Generating $\{M_r\}$: We denote the uniformly distributed subset of indexes as $\gamma \sim u(\{1, \ldots, \frac{n(n+1)}{2}\})$, where $|\gamma| = m$. For sampling a random matrix, we obtain $H' \sim$ Haar$(O(\frac{n(n+1)}{2}))$ and subsample $H = H'_{[\{1,\ldots,\frac{n(n+1)}{2}\},\gamma]}$. Afterwards, M_r is the only symmetric matrix such that $\text{veclt}(M_r) = H_{[\{1,\ldots,\frac{n(n+1)}{2}\},\{r\}]}$.
- Generating $X^\star$: Similarly, we set up indexes $\alpha \sim u(\{1, \ldots, n\})$, where $|\alpha| = k$. Then, we obtain $X' \sim$ Haar$(O(n))$ and set $Q = X'_{[\{1,\ldots,n\},\alpha]}$. In this notation, $X^\star = QQ^\top$

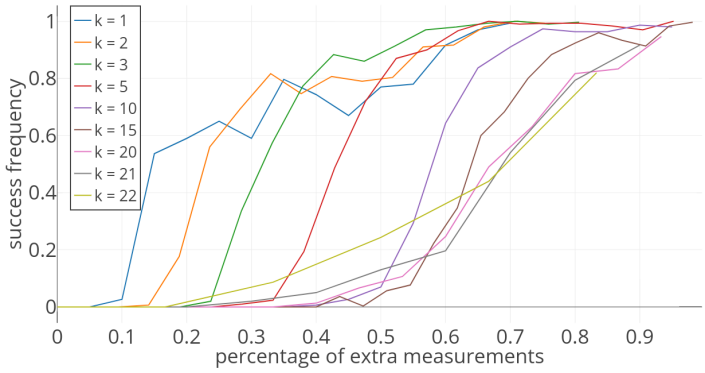
For more statistically significant results, we also use the same scheme to design a random objective matrix:



$n = 15$



$n = 20$



$n = 25$

Figure 2.2: These plots show the frequency of recovery for synthetic data. The x axis is the percentage (normalized) of total number of extra measurements available. This means that 0 corresponds to $m = m_{min} = nk - k(k-1)/2$, and 1 corresponds to $m = m_{max} = n(n+1)/2$. The y axis shows the probability of successful recoveries.

- Generating M : We obtain $\beta \sim u(\{1, \ldots, n\})$, $|\beta| = n - k$ and $K' \sim \text{Haar}(O(n))$. Similarly to the previous cases, set $K = K'_{[\{1,\ldots,n\},\beta]}$ and $M = KK^\top$.

The experiments have been scripted in Python with the use of CVXOPT as the mathematical optimization library.

Chapter 3

Complexity of the Problem Under Sparse Noise

In this chapter, we focus our attention on the semidefinite affine rank-1 feasibility problem, which can be equivalently thought of as the problem of quadratic regression. In this problem, the goal is to find the unknown state (numerical parameters) of a system modeled by a set of equations that are quadratic in the state. We study the setting when a subset of equations of a fixed cardinality is subject to errors of arbitrary, possibly adversarially selected, magnitudes. Our aim is to determine the conditions on the instances θ that guarantee tractability of the problem of the exact state recovery. Understanding these conditions could clear the path towards designing cyber-physical systems that, by design, limit the scope of instances of the inference problem to the tractable ones, even in conditions of a cyber-attack, and therefore are robust to cybersecurity threats of data augmentation.

To measure complexity, we develop two algorithm schemes that address the quadratic regression problem. Both of them are based on conic optimization and are able to utilize a prior guess of the solution if one is available. We develop a set of conditions on the properties of the data (the coefficients of the equations), the quality of the prior guess, the size of the support of the noise vector, and the hyperparameters of the algorithms from each of the schemes. Under these conditions, polynomial-time recovery of the unknown state is possible with the corresponding algorithms. We show that there exists a trade-off between the time complexity of an algorithm schema and its tolerance to errors in the measurements. This makes one method better at enforcing cybersecurity and the other better at scaling for larger systems. The key feature of the obtained conditions consists in bounds on the number of bad measurements each method can tolerate without producing a nonzero estimation error. It is proved that the proposed methods allow up to half of the total number of measurements to be grossly erroneous if the prior guess is close enough to the true solution. We also propose a surrogate iterative method for quadratic regression which is based on conic programming and does not rely on the prior guess. It offers another level of the trade-off between robustness and the complexity of the algorithm. The efficacy of the developed methods is demonstrated in four empirical experiments on learning dynamical systems and power network states under

both sparse and dense errors in the measurements.

3.1 Introduction

Nonlinear regression aims to find the parameters of a given model based on observational data. One may assume the existence of a potentially nonlinear continuous function $f(x;\phi) : \mathcal{X} \times \Phi \to \mathbb{R}$ defined over the set of models $x \in \mathcal{X}$ and inputs $\phi \in \Phi$, where the goal is to estimate the true model given a set of imperfect measurements y_i's:

$$y_i = f(z;\phi^i) + \eta_i, \qquad \forall i \in \{1, ..., m\} \tag{3.1}$$

In this formulation, the unknown error vector η could be the measurement noise with modest values. However, a more drastic scenario corresponds to the case where the vector η is sparse and its nonzero entries are allowed to be arbitrarily large. Under this circumstance, *a priori* information about the probability distribution of the sparse vector η may be available, in addition to an upper bound on the cardinality of η. This important problem is referred to as *robust regression* and appears in real-world situations when some observations, named outliers, are completely wrong in an unpredictable way. This could occur during an image acquisition with several corrupted pixels, or result from communication issues during data transmission in sensor networks. Such problems arise in different domains of applications and have been studied in the literature. The problem of errors in estimation of the state of an electric power grid is of a special importance. Outliers in this case are associated with faulty sensors, cyber attacks, or regional data manipulation to impact the electricity market [52, 69].

There are several classical works on robust regression and outliers detection. The book by Rousseeuw and Leroy [91] offers an overview of many fundamental results in this area dating back to 1887 when Edgeworth proposed the least-absolute-value regression estimator. Modern techniques for handling sparse errors of arbitrary magnitudes vary with respect to different features: statistical properties of the error, class of the regression model $f(x;\phi)$, set of possible true models, type of theoretical guarantees, and characteristics of the adversary model generating errors [20, 77, 10, 116, 57]. There is a plethora of papers on this topic for the well-known linear regression problem [17, 110, 95, 22, 11]. In this case, the function $f(x;\phi)$ is linear in the model vector x, and can be written as $\phi^* x$. Nevertheless, there are far less results known for nonlinear regression. This is due to the fact that linear regression amounts to a system of linear equations with a cubic solution complexity if the measurements are error-free, whereas nonlinear regression is NP-hard and its complexity further increases with the inclusion of premeditated errors. However, very special cases of nonlinear regression have been extensively studied in the literature. In particular, the robust phase retrieval problem that can be formulated with $f(x;\phi_i) = |\phi_i^* x|^2$ has received considerable attention [116, 48, 21]. Another special case is the trace regression problem that has been studied by Hamidi and Bayati [47] under a low-rank assumption on the unknown matrix solution. However, this has not yet been studied under adversarial sparse additive errors. The mathematical

framework provided in the current chapter addresses the trace regression problem under a low-rank assumption and sparse adversarial noise.

The existing approaches for robust regression include the analysis of the unconstrained case [17, 95, 10, 11, 53], the constrained scenario with conditions on the sparsity of the solution vector z [110, 77, 79, 71], and more sophisticated scenarios in the context of matrix completion [20, 23, 57, 119]. Motivated by applications in inverse covariance estimation [106], the papers by Xu, Caramanis, and Mannor [113], Yang and Xu [114], and McWilliams et al. [71] consider sparse noise in the input vector ϕ_i as opposed to the additive error considered in the present chapter. The work of [17] is based on l_1-minimization, whereas [77] solve an extended Lasso formulation defined as the minimization of $\|y - Ax + \nu\|_2^2 + \mu_1\|x\|_1 + \mu_2\|\nu\|_1$. The work by Dalalyan and Chen [29] proposes to solve a second-order cone programming (SOCP) for robust linear regression, which is related to the current chapter with a focus on robust nonlinear regression. In contrast to the above-mentioned papers that aim to develop a single optimization problem to estimate the solution of a regression, there are iterative-based methods as well. For instance, [22, 10, 11] propose iterative algorithms via hard thresholding.

Due to the diversity in the problem formulation and approaches described in different papers, it is difficult to compare the existing results since there is no single dominant method. However, the most common measures of performance for robust regression algorithms are the traditional algorithmic complexity and the permissible number of gross measurements $\|\eta\|_0$ compared to the total number of measurements m. In this chapter, the objective is to design a polynomial-time algorithm, in contrast with potentially exponential-time approaches [104], with guaranteed convergence under technical assumptions. As far as the robustness of an algorithm is concerned, the existing works often provide probabilistic guarantees on the recoverability of the original parameter vector z for linear Gaussian stochastic systems under various assumptions on the relationship between $\|\eta\|_0$ and m. In this case, the ratio $\frac{\|\eta\|_0}{m}$, named breakdown point, is limited by a constant and could even approach 1 if the unknown solution z is sparse.

Contributions and Organization

The main objective of this chapter is to analyze a robust regression problem for an arbitrary quadratic model that includes power system state estimation and phase retrieval as special cases. The focus is on the calculation of the maximum number of bad measurements that does not compromise the exact reconstruction of the model vector z. In Section 3.2, we formally state the problem. In Section 3.3, we propose two conic optimization methods and study their properties. In particular, we obtain conditions that guarantee the exact reconstruction of z. In Section 3.4, we develop the main results of this chapter. Under certain technical assumptions, we discover the dependence between the number of perfect measurements and the maximum admissible number of wrong measurements. After that, we consider a stochastic setting based on Gaussian distributions. In this case, we show that the number of bad measurements can safely be on the order of the square root of the total number of measurements, and moreover the breakpoint approaches 1/2 if there is enough

prior information. To provide a broader range of possible approaches to the problem, Section 3.5 designs an alternative iterative-based method. Numerical results are presented in Section 3.6, which includes a case study on a European power grid.

3.2 Problem Formulation and Preliminaries

The quadratic regression under sparse noise aims to find a vector z in $\mathbb{R}^n$ or $\mathbb{C}^n$ such that

$$y_r = z^* M_r z + \eta_r, \quad \forall r \in \{1, \dots, m\}, \tag{3.2}$$

where

- $y_1, \dots, y_m$ are some known real-valued measurements.
- $\eta_1, \dots, \eta_m$ are unknown but sparsely occurring real-valued noise of arbitrary magnitudes.
- $M_1, \dots M_m$ are some known $n \times n$ Hermitian matrices.

The regression problem could have two solutions $\pm z$ in the real-valued case, which increases to infinitely many in the form of $z \times e^{\sqrt{-1}\psi}$ in the complex case. To avoid this ambiguity, the objective of this work is to find the matrix zz^* rather than z since this matrix is invariant under the rotation of z. At the same time, the recovery of z from zz^* is a simple task that can be accomplished using the spectral decomposition. If m is large enough, then zz^* is expected to be unique. In this chapter we aim to recover any solution zz^* in case there are multiple ones. In Problem (3.2), each measurement equation could have a linear term in addition to its purely quadratic function $x^* M_r x$. By introducing one additional variable u such that $u^2 = 1$, one can multiply the linear terms with u to make them quadratic [68]. As a result, Problem (3.2) is a general quadratic regression problem.

Let $\hat{z}$ be an initial guess for the unknown solution z. We refer to this as *prior knowledge*. We do not make any assumption about the gap between $\hat{z}$ and z, and develop different methods that can be run independent of this gap. However, the goal is to show that as this gap becomes smaller, the performance of these methods increases. More precisely, we define a measure to quantify the amount of information in the prior knowledge and use it to study the to-be-developed techniques. Note that it is easy to deduce prior knowledge for many real-world systems. For example, we will show that the physics of power systems naturally provide such useful knowledge.

Consider the complex-valued case and write $x = a + \sqrt{-1}b \in \mathbb{C}^n$ and $M = A + \sqrt{-1}B \in \mathbb{H}^n$, where $a, b \in \mathbb{R}^n$, $A \in \mathbb{S}^n$ and $B = -B^\top \in \mathbb{R}^{n \times n}$. It is straightforward to verify that:

$$x^* M x = a^\top A a + 2b^\top B a + b^\top A b = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Notice that the matrices in the right-hand side of the above equation are real-valued. As a result, we will only develop the theoretical results of this work in the real-valued case $z \in \mathbb{R}^n$ and $M_r \in \mathbb{S}^n$ since they can be easily carried over to the complex-valved case. However, we will offer a case study on power systems where the unknown state is a complex vector.

In the regression problem under sparse noise, the vector η is assumed to be sparse. To distinguish between error-free and erroneous measurements, we partition the set of measurements into two subsets of *good* and *bad* measurements:

$$\mathcal{G} = \{r \in \{1, \dots, m\} | \eta_r = 0\}, \quad \mathcal{B} = \{1, \dots, m\} \backslash \mathcal{G}$$

To streamline the derivation of the analytical results of this chapter, we assume that $\mathcal{G} = \{1, \dots, |\mathcal{G}|\}$ and $\mathcal{B} = \{|\mathcal{G}|+1, \dots, m\}$. However, the algorithms to be designed are completely oblivious to the type of each measurement and its membership in either $\mathcal{G}$ or $\mathcal{B}$.

The objective of this chapter is to develop efficient algorithms for finding z precisely as long as η is sufficiently sparse. This statement will be formalized in the next sections.

3.3 Conic Optimization Methods

Consider a variable matrix X playing the role of $xx^\top$. This matrix is positive semidefinite and has rank 1. By dropping the rank constraint, one can cast the quadratic regression as a linear matrix regression. Motivated by this relaxation, consider the optimization problem

$$\begin{aligned} \underset{X \in \mathbb{S}^n,\ \nu \in \mathbb{R}^m}{\text{minimize}} \quad & \langle X, M \rangle + \mu \|\nu\|_1 \\ \text{subject to} \quad & \langle X, M_r \rangle + \nu_r = y_r, \quad \forall r \in \{1, \dots, m\} && \text{(3.3a)} \\ & X = X^\top \succeq_{\mathcal{C}} 0 && \text{(3.3b)} \end{aligned}$$

where the notation $\succeq_{\mathcal{C}}$ is the generalized inequality sign with respect to $\mathcal{C}$, which is either the cone of symmetric positive semidefinite (PSD) matrices or the 2×2 principal sub-matrices PSD cone [see 86]. The above cones are formally defined in Subsection 3.3.

The problem definition involves a matrix M that is to be designed based on the prior knowledge $\hat{z} \in \mathbb{R}^n$ in such a way that the term $\langle W, M \rangle$ in the objective function promotes a low-rank structure on W. The construction of M will be studied later in the chapter. We refer to Problem (3.3) as *penalized conic program*, but call it with more specific names in two special cases: (i) *penalized semidefinite program (SDP)* if $\mathcal{C}$ is the cone of PSD matrices, (ii) *penalized second-order cone program (SOCP)* if $\mathcal{C}$ is the cone of matrices with all 2×2 principal sub-matrices being PSD. The penalized conic program is a convex problem and can be solved in polynomial time up to any given accuracy.

A popular approach to solving rank minimization problems is via an approximation technique that replaces the non-convex objective function with the nuclear norm of the unknown matrix [20]. We exploit a different approach for three main reasons:

- The nuclear norm minimization is rooted in the fact that the nuclear norm is a convex envelope of the rank over a certain ball, but the connection between nuclear norm and rank fades away when the ball is intersected with the hyperplanes given by (3.3a).
- In many practical applications, some prior knowledge about the unknown state is available. However, the nuclear norm minimization cannot incorporate such information to improve the search for the unknown solution. This is in contrary to the standard numerical algorithms for optimization that allow the initialization of the process for finding an optimal solution. Therefore, one would expect to have a new learning method for quadratic regression that exploits prior knowledge about the solution.
- The minimization of the trace is meaningless in many applications where the trace of all feasible matrices W is automatically in a narrow bound. In this case, the trace cannot be used to distinguish low-rank solutions from high-rank solutions. This naturally occurs in power systems, for which the trace is almost fixed since voltage magnitudes are always close to nominal values (e.g., 110 volts) [67]

The method to be developed in this chapter addresses the above issues via a major generalization of the nuclear norm minimization. In particular, if $\hat{z} = 0$, then the proposed approach is equivalent to the nuclear norm minimization. We refer to $\hat{z}$ as prior knowledge, and aim to show how the amount of information in the prior knowledge—measured in terms of the closeness between $\hat{z}$ and the unknown solution—affects the performance of the penalized conic program and the estimation error.

In the following two subsections, we will introduce the functions κ and ξ, matrices $\bar{J}$ and $\tilde{J}$, and vectors $\bar{d}$ and $\tilde{d}$, and then study the problem of designing M based on the prior knowledge $\hat{z}$.

Penalized Semidefinite Programming

Consider the penalized SDP that corresponds to Problem (3.3) with $\mathcal{C}$ equal to the PSD cone. Given a matrix $X \in \mathbb{S}^n$, define $\kappa(X)$ to be the sum of the two smallest eigenvalues of X, i.e.,

$$\kappa(X) := \lambda_n(X) + \lambda_{n-1}(X)$$

Let the matrix M in the objective function of the penalized SDP be chosen to have the following properties:

$$\begin{aligned} &M\hat{z} = 0, \\ &\operatorname{rank}(M) \geq n - 1 \\ &\kappa(M) > 0 \end{aligned}$$

If there is no prior knowledge available, one can select $\hat{z}$ to be zero and then choose M as I. This will correspond to the famous nuclear norm minimization. As will become evident later in the chapter, the linear term $\langle X, M \rangle$ with the above-mentioned matrix M penalizes the deviation of X from $\hat{z}\hat{z}^\top$. Since $\hat{z}\hat{z}^\top$ is low-rank, the inclusion of this linear

term automatically takes care of both prior knowledge and low-rank promotion. There are infinitely many choices for M, and it is not important which one to select as far as the analysis of this chapter is concerned. One natural choice for M is the matrix of orthogonal projection onto the hyperplane that is orthogonal to $\hat{z}$. This particular matrix is computationally cheap to construct. However, if more than one initial guess is available, it is beneficial to design the matrix M via an optimization problem that attempts to minimize the violation of the above conditions for all initial values of $\hat{z}$.

Observe that the dual of Problem (3.3) can be obtained as:

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^m}{\text{maximize}} \quad & -y^\top \lambda \\ \text{subject to} \quad & M + \sum_{r=1}^{m} \lambda_r M_r \succeq 0 && \text{(3.4a)} \\ & \|\lambda\|_\infty \leq \mu && \text{(3.4b)} \end{aligned}$$

where $\succeq 0$ is the positive semidefinite sign. Define the matrix $\bar{J}$ and the vector $\bar{d}$ as:

$$\bar{J} = [M_1 z \ \ldots \ M_m z] \tag{3.5a}$$
$$\bar{d} = Mz \tag{3.5b}$$

where z is the solution of the original problem (3.2). The matrix $\bar{J}$ captures the coherence between the model vector z and the measurement matrices M_r. At its turn, the vector $\bar{d}$ measures the alignment of the solution z and the prior knowledge $\hat{z}$. Note that $\bar{J}$ and $\bar{d}$ are both completely noise-agnostic. The regularity property of the matrix $\bar{J}$ and the norm of the vector $\bar{d}$ play important roles in guaranteeing the correct recovery of z. A preliminary result is provided below, which will later be used to study the penalized SDP.

Lemma 5. *Assume that there exists an index $r \in \{1, \ldots, m\}$ such that $\hat{z}^\top M_r \hat{z} \neq 0$ and*

$$\mu > \|\bar{J}_{\mathcal{G}}^{+} \left(\bar{d} - \mu \bar{J}_{\mathcal{B}} \operatorname{sign}(\eta_{\mathcal{B}})\right) \|_\infty \tag{3.6a}$$
$$\frac{\kappa(M)}{2 \max\limits_r \|M_r\|_2} > \|\bar{J}_{\mathcal{G}}^{+} \left(\bar{d} - \mu \bar{J}_{\mathcal{B}} \operatorname{sign}(\eta_{\mathcal{B}})\right) \|_1 + \mu |\mathcal{B}| \tag{3.6b}$$

Then, $(zz^\top, \eta)$ is the unique solution of the penalized SDP. Moreover, $\hat{\lambda} = \begin{bmatrix} \hat{\lambda}_{\mathcal{G}} \\ \hat{\lambda}_{\mathcal{B}} \end{bmatrix}$ defined as

$$\begin{aligned} \hat{\lambda}_{\mathcal{B}} &= -\mu \operatorname{sign}(\eta_{\mathcal{B}}) \\ \hat{\lambda}_{\mathcal{G}} &= -\bar{J}_{\mathcal{G}}^{+} \left(\bar{d} + \bar{J}_{\mathcal{B}} \hat{\lambda}_{\mathcal{B}}\right) \end{aligned}$$

is a dual solution.

Proof. The proof is provided in Appendix. □

The conditions given in Lemma 5 will be refined and further studied in Section 3.4 to uncover useful properties of the penalized SDP.

Penalized Second-Order Cone Programming

Although penalized SDP is a convex optimization, its memory and time complexities make it less appealing for large-scale problems [14]. These complexities can be significantly reduced if the union of the 0-1 sparsity patterns of the matrices $M, M_1, ..., M_m$ is a sparse matrix itself [40]. This requires a natural sparsity in the measurement matrices M_r and also the design of a sparse matrix M, which is not always possible. As an alternative, one can break down the complexity of the penalized SDP by replacing its constraint $X \succeq 0$ with second-order conic constraints. Although penalized SDP offers better recovery guarantees than penalized SOCP, the latter has a significantly lower computational complexity and can be efficiently solved for large-scale problems using interior-point methods [4]. In this part, we study the penalized SOCP as a counterpart of penalized SDP. This optimization problem is obtained by building the cone $\mathcal{C}$ based on the 2×2 principal submatrices of W, as explained below.

Definition 2 ($2\mathcal{PSM}$)**.** *A matrix $X \in \mathbb{S}^n$ belongs to the 2×2 principal sub-matrices PSD cone if each 2×2 principal sub-matrix of X is positive semidefinite, i.e.,*

$$[e^i \quad e^j]^\top X [e^i \quad e^j] \succeq 0, \qquad \forall\ i < j$$

Since the $2\mathcal{PSM}$ cone is not self-dual, we introduce the scaled diagonally dominant cone below.

Definition 3 ($\mathcal{SDD}$)**.** *A matrix $X \in \mathbb{R}^{n \times n}$ belongs to the scaled diagonally dominant cone if there exists a set of 2×2 positive semidefinite matrices $\{X^{ij}\}_{i<j}^{j \leq n}$ such that*

$$\sum_{j=2}^{n} \sum_{i=1}^{j-1} [e^i \quad e^j] X^{ij} [e^i \quad e^j]^\top = X$$

The notation $\{X^{ij}\}_{i<j}^{j \leq n}$ in the above definition means $\{X^{ij} \big| j = 2, \ldots, n, i = 1, \ldots, j-1\}$. The next lemma explains the connection between $2\mathcal{PSM}$ and $\mathcal{SDD}$ cones.

Lemma 6 ([86])**.** *The dual of the 2×2 principal sub-matrices PSD cone is the scaled diagonally dominant cone of the same dimension.*

In what follows, we will define and describe certain properties of a linear space of diagonal decompositions of matrices. These definitions are somewhat more tedious than the ones in the previous subsection, but they serve the same aim: they formally define the matrix M and the counterparts of $\bar{J}$ and $\bar{d}$ for the penalized SOCP.

Definition 4. *The sequence* $\{A^{ij} \in \mathbb{S}^2\}_{i<j}$ *is said to be a diagonal decomposition (or just decomposition) of* $A \in \mathbb{S}^n$ *if*

$$A = \sum_{j=2}^{n} \sum_{i=1}^{j-1} [e^i, e^j] A^{ij} [e^i, e^j]^\top$$

A decomposition that consists of PSD matrices is a certificate that a matrix belongs to the $\mathcal{SSD}$ cone. Similarly to the function κ defined for the penalized SDP, we introduce the function $\chi(\{X^{ij}\}_{i<j})$ as follows:

$$\chi\left(\{X^{ij}\}_{i<j}\right) := \min_{i<j} \operatorname{tr}(X^{ij}) = \min_{i<j} \left(\lambda_1(X^{ij}) + \lambda_2(X^{ij})\right)$$

Consider a sequence $\{M^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ such that

$$\begin{cases} \chi(\{M^{ij}\}_{i<j}) > 0 \\ M^{ij} [\hat{z}_i \;\; \hat{z}_j]^\top = 0 \;\; \text{for all } i < j \end{cases}$$

Define the corresponding penalized SOCP as Problem (3.3) with $\mathcal{C}$ equal to the $2\mathcal{PSM}$ cone and

$$M = \sum_{i<j} [e^i \;\; e^j] M^{ij} [e^i \;\; e^j]^\top. \tag{3.7}$$

Since $M^{ij} \succeq 0$, the matrix M belongs to the $\mathcal{SDD}$ cone. Similarly to the penalized SDP, there is an infinite number of possible matrices M, one of which can naturally be obtained by selecting M^{ij} to be the orthogonal projection onto the line orthogonal to $[\hat{z}_i \;\; \hat{z}_j]^\top$.

The dual of the penalized SOCP takes the form:

$$\begin{aligned} \max_{\lambda, H, H^{ij}} \quad & -y^\top \lambda \\ \text{subject to} \quad & M + \sum_{r=1}^{m} \lambda_r M_r = H && \text{(3.8a)} \\ & \sum_{i<j} [e^i \;\; e^j] H^{ij} [e^i \;\; e^j]^\top = H && \text{(3.8b)} \\ & H^{ij} \succeq 0 && \text{(3.8c)} \\ & \|\lambda\|_\infty \leq \mu && \text{(3.8d)} \end{aligned}$$

where the variables are $\lambda \in \mathbb{R}^m$, $H \in \mathbb{S}^n$ and $\{H^{ij}\}_{i<j}^{j\leq n} \subset \mathbb{S}^2$. Now, it is easy to observe that each conic constraint $[e^i \;\; e^j]^\top X [e^i \;\; e^j] \succeq 0$ in Problem (3.3) corresponds to the dual variable matrix H^{ij}. Hence, the complementary slackness condition can be written as

$$\langle [e^i \;\; e^j]^\top X [e^i \;\; e^j], H^{ij} \rangle = 0, \qquad \text{for all } i < j \leq n$$

Define G to be a symmetric matrix such that $M^{ij}[z_i \;\; z_j]^\top = [G_{ij} \;\; G_{ji}]^\top$ for all $i < j \in \{1,\dots,n\}$ and $G_{ii} = 0$ for all $i \in \{1,\dots,n\}$. Furthermore, for every $r \in \{1,\dots m\}$, define R^r as a matrix with the properties:

$$\begin{cases} \sum_{j=1}^{n} R^r_{ij} = M^r_{ii} \;\; \text{for all } i \in \{1,\dots,n\} \\ R^r_{ii} = 0 \;\; \text{for all } i \in \{1,\dots,n\} \end{cases}$$

One simple example of this matrix is a matrix with the (i,j)-entry equal to $R^r_{ij} = \frac{M^r_{ii}}{n-1}$ for $i \neq j$. Given $r \in \{1,\dots,m\}$, define $G^r \in \mathbb{R}^{n\times n}$ as a matrix with the components $G^r_{ij} = z_j M^r_{ij} + z_i R^r_{ij}$ and $G^r_{ii} = 0$ for all $i,j \in \{1,\dots,n\}$. Similarly to (3.5), define:

$$\tilde{J} = \begin{bmatrix} \text{vecnd}\,(G^1) & \dots & \text{vecnd}\,(G^m) \end{bmatrix} \tag{3.9a}$$

$$\tilde{d} = \text{vecnd}\,(G) \tag{3.9b}$$

where the vectorization operator vecnd : $\mathbb{R}^{n\times n} \to \mathbb{R}^{n^2-n}$ puts all elements excluding the diagonal of its matrix argument into the form of a vector. Similarly to the penalized SDP case, here $\tilde{J}$ captures the coherence between the data and the true model, while $\tilde{d}$ captures the correlation between the true model and the prior knowledge.

The counterpart of Lemma 5 is stated below for the penalized SOCP.

Lemma 7. *Assume that the components of the initial guess are nonzero (i.e., $\hat{z}_i \neq 0$ for all $i \in \{1,\dots,n\}$) and that there exists an index $r \in \{1,\dots,m\}$ such that $\hat{z}^* M_r \hat{z} \neq 0$ and*

$$\mu > \|\tilde{J}^+_{\mathcal{G}} \left(\tilde{d} - \mu \tilde{J}_{\mathcal{B}}\, \text{sign}(\eta_{\mathcal{B}})\right)\|_\infty \tag{3.10a}$$

$$\frac{\chi(\{M^{ij}\}_{i<j})}{\max\limits_{r,\; i<j} |\text{tr}(\{M^{ij}_r\}_{i<j})|} > \|\tilde{J}^+_{\mathcal{G}} \left(\tilde{d} - \mu \tilde{J}_{\mathcal{B}}\, \text{sign}(\eta_{\mathcal{B}})\right)\|_1 + \mu|\mathcal{B}| \tag{3.10b}$$

Then, $(zz^\top, \eta)$ is the unique solution of the penalized SOCP. Moreover, $\hat{\lambda} = \begin{bmatrix} \hat{\lambda}_{\mathcal{G}} \\ \hat{\lambda}_{\mathcal{B}} \end{bmatrix}$ defined as

$$\hat{\lambda}_{\mathcal{B}} = -\,\mu\, \text{sign}(\eta_{\mathcal{B}})$$
$$\hat{\lambda}_{\mathcal{G}} = -\,\tilde{J}^+_{\mathcal{G}} \left(\tilde{d} + \tilde{J}_{\mathcal{B}} \hat{\lambda}_{\mathcal{B}}\right)$$

can be completed to a dual optimal solution.

Proof. The proof is provided in Appendix. □

We need to mention that while there is some freedom in the choice of $\hat{\lambda}_{\mathcal{G}}$ in the proof of Lemma 7, the dual variables $\hat{\lambda}_{\mathcal{B}}$ associated with the bad measurements are inflexible. This is elaborated below.

Lemma 8. *$\hat{\lambda}_{\mathcal{B}} = -\mu \operatorname{sign}(\eta_{\mathcal{B}})$ is the only possible choice for the optimal dual variables if the optimal primal variables are $(zz^\top, \eta)$.*

Proof. The proof is provided in Appendix. □

3.4 Main Results

In this section, we develop the key theoretical results on the Robust Quadratic Regression solution via the conic methods presented in the preceding section. The common structure of the conditions in Lemmas 5 and 7 allows us to derive results providing guarantees for both the SDP and the SOCP approaches simultaneously. To do so, we will use the universal notations J and d to denote $\bar{J}$ and $\bar{d}$ (defined in Subsection 3.3) in the penalized SDP case and to denote $\tilde{J}$ and $\tilde{d}$ (defined in Subsection 3.3) in the penalized SOCP case. Define

$$\alpha_{\text{SDP}} = \frac{\kappa(M)}{2\|\bar{d}\|_2 \max\limits_r \|M_r\|_2} \quad \text{or} \quad \alpha_{\text{SOCP}} = \frac{\chi(\{M^{ij}\}_{i<j})}{\|\tilde{d}\|_2 \max\limits_{r,\ i<j} |\text{tr}(M_r^{ij})|}$$

For the particular matrices M constructed in Section 3.3 using the projection operator, both $\kappa(M)$ and $\chi(\{M^{ij}\}_{i<j})$ are equal to 1. In addition, one can normalize the equations in (3.2) before solving the problem via a rescaling so that $\|M_r\| = 1$, in which case the terms with M_r in the definitions of α_{SDP} and α_{SOCP} can be eliminated (or bounded by a constant). Therefore, we can write that $\alpha_{\text{SDP}} \propto |\langle \frac{z}{\|z\|}, \frac{\hat{z}}{\|\hat{z}\|} \rangle|^{-1}$ and $\alpha_{\text{SOCP}} \propto (\sum_{i<j} |\langle [\frac{z_i}{\|z\|} \quad \frac{z_j}{\|z\|}], [\frac{\hat{z}_i}{\|\hat{z}\|} \quad \frac{\hat{z}_j}{\|\hat{z}\|}] \rangle|)^{-1}$, which imply that these parameters measure the amount of information in the prior knowledge. Henceforth, we use the shorthand notation α to denote α_{SDP} or α_{SOCP} depending on whether the penalized SDP or SOCP is analyzed. The same notation is used for l that takes one of the following values:

$$l_{\text{SDP}} = n; \qquad l_{\text{SOCP}} = n^2 - n$$

Deterministic Bound

In this subsection, we establish a uniform bound on the number of bad measurements that a penalized conic relaxation can tolerate. To do so, we make use of two matrix properties introduced in [11].

Definition 5 (*SSC property*)**.** *A matrix $X \in \mathbb{R}^{l \times m}$ is said to satisfy the Subset Strong Convexity (SSC) Property at level p with constant $\gamma_p > 0$ if*

$$\gamma_p \leq \min_{|S|=p} \sqrt{\lambda_{min}(X_S X_S^\top)}$$

Definition 6 (*SSS property*)**.** *A matrix $X \in \mathbb{R}^{l \times m}$ is said to satisfy the Subset Strong Smoothness (SSS) Property at level p with constant $\Gamma_p > 0$ if*

$$\max_{|S|=p} \sqrt{\lambda_{max}(X_S X_S^\top)} \leq \Gamma_p$$

Note that the notation $|S| = p$ in the above definition specifies the index set of any p columns of the matrix X. The ratio of the constants γ_p and Γ_{m-p} can be interpreted as a uniform condition number at level p.

Theorem 4. *Consider Problem* (3.2)*, and let* $N = |\mathcal{B}|$ *denote the cardinality of the support of the noise vector* η*. Without any future assumption on the noise, consider the corresponding penalized conic problem. Consider arbitrary constants* $\bar{\alpha}$*,* γ *and* Γ *such that*

$$\bar{\alpha} > \frac{\Big(\sqrt{N}\frac{\Gamma}{\gamma} + (1 - \frac{\Gamma}{\gamma})\Big)\sqrt{m-N} + N}{\gamma - \Gamma},$$

- *If the exact solution* z *of* (3.2)*, the prior knowledge* $\hat{z}$ *and the measurement matrices* M_r *are such that* $\bar{J}$ *satisfies the SSC property at level* $m - N = |\mathcal{G}|$ *with the constant* γ *and the SSS property at level* $N = |\mathcal{B}|$ *with the constant* Γ*, then there exists a constant* μ *for which* $(zz^\top, \eta)$ *is the unique solution of the penalized SDP problem if* $\alpha_{SDP} \geq \bar{\alpha}$*.*

- *If the exact solution* z *of* (3.2)*, the prior knowledge* $\hat{z}$ *and the measurement matrices* M_r *are such that* $\tilde{J}$ *satisfies the SSC property at level* $m - N = |\mathcal{G}|$ *with the constant* γ *and the SSS property at level* $N = |\mathcal{B}|$ *with the constant* Γ*, then there exists a constant* μ *for which* $(zz^\top, \eta)$ *is the unique solution of the penalized SOCP problem if* $\alpha_{SOCP} \geq \bar{\alpha}$*.*

Proof. The proof follows from Lemmas 5 and 7, together with Lemma 9 to be stated later in the chapter. □

Theorem 4 implies that the penalized conic relaxations are exact and the corresponding instances of the $\mathcal{NP}$-hard problem (3.2) can be solved in polynomial time, provided that they satisfy a certain condition. This condition is not restrictive as long as the amount of information in the prior knowledge is not too low.

Notice that aside from the cardinalities of the sets $\mathcal{G}$ and $\mathcal{B}$, Theorem 10 imposes no condition on the noise values. Therefore, the guarantee provided by this Theorem is established for the "worst-case scenario" when the adversary is adaptive and strategically selects the indexes of the error vector η based on the true solution z to have the most impact. Theorem 4 in the chapter is based on two basic assumptions:

- Incoherence of the good measurements and dominance of good measurements over bad ones: This is implied by the terms of the form $\gamma - \Gamma$ in the inequality bound;

- The amount of information in the prior knowledge: This is implied by the lower bound $\alpha \geq \bar{\alpha}$.

Although Theorem 4 just states the existence of the hyperparameter μ, we will identify an interval for this parameter below.

Lemma 9. *Let J be a matrix in $\mathbb{R}^{l\times m}$ that satisfies the SSC and SSS properties on levels $|\mathcal{G}|$ and $|\mathcal{B}|$ with the respective constants $\gamma_{|\mathcal{G}|}$ and $\Gamma_{|\mathcal{B}|}$ ($\gamma_{|\mathcal{G}|} > \Gamma_{|\mathcal{B}|}$). Moreover, let d be a vector in $\mathbb{R}^l$ and λ be a vector in $\mathbb{R}^m$ such that $\lambda_{\mathcal{B}} = \mu \cdot s$, where μ is a scalar and the entries of s are only $+1$ or -1. If*

$$\alpha\gamma_{|\mathcal{G}|}(\gamma_{|\mathcal{G}|} - \Gamma_{|\mathcal{B}|}) - |\mathcal{B}|\gamma_{|\mathcal{G}|} > \Big(\sqrt{|\mathcal{B}|}\Gamma_{|\mathcal{B}|} + (\gamma_{|\mathcal{G}|} - \Gamma_{|\mathcal{B}|})\Big)\sqrt{|\mathcal{G}|}$$

then the interval

$$\left[\frac{\|d\|_2}{\gamma_{|\mathcal{G}|} - \Gamma_{|\mathcal{B}|}}, \frac{(\alpha\gamma_{|\mathcal{G}|} - \sqrt{|\mathcal{G}|})\|d\|_2}{\sqrt{|\mathcal{B}||\mathcal{G}|}\Gamma_{|\mathcal{B}|} + |\mathcal{B}|\gamma_{|\mathcal{G}|}}\right] \tag{3.11}$$

is not empty and the system of inequalities

$$\begin{cases} \mu > \|\lambda_{\mathcal{G}}\|_\infty \\ \alpha\|d\|_2 > \|\lambda_{\mathcal{G}}\|_1 + \mu|\mathcal{B}| \end{cases}$$

is satisfied with $\lambda_{\mathcal{G}} = -J_{\mathcal{G}}^{+}(J_{\mathcal{B}}\lambda_{\mathcal{B}} + d)$ for every μ in the interval (3.11).

Proof. The proof directly follows from Definitions 5 and 6, as well as Lemma 19 proved in Appendix. □

Lemma 9 provides an interval for the hyperparameter μ. The length of this interval and its location depend on the solution z and the amount of information in the measurements, but they are independent of the noise values. This is consistent with the existing results for the precedents of the quadratic regression problem, such as Lasso [105] and Graphical Lasso [88]. In such problems, the existence of this interval with unknown endpoints is enough for developing iterative methods, such as bisection techniques, to repeatedly solve the problem and update μ based on measuring the quality of estimation at each run of the optimization. This fits within the realm of model selection, where one can use information-theoretic methods such as the Akaike criterion. In Section 3.6, we will verify that the simple idea of trying multiple values for μ with different orders of magnitude performs well on real data.

Stochastic Bound

In the preceding section, we developed theoretical results on the correct recovery of the state of the problem and the number of permissible bad measurements. Unlike the existing results that focus on particular types of quadratic regression problems, these results apply to any arbitrary set of matrices M_r's. This generality of the results has made the conditions somewhat sophisticated. In what follows, we will simplify the results and provide some intuition under a stochastic setting.

Definition 7. *A matrix X is called standard Gaussian over $\mathbb{R}$ if its entries are independent and identically distributed random variables with a standard normal distribution.*

The data in this subsection is assumed to be stochastic, and therefore the associated theoretical results should be stated in a probabilistic sense. We select $\delta \in (0,1)$, and define $\varepsilon^* = \arg\min_{\varepsilon>0} 2\sqrt{6}e \cdot \frac{\sqrt{l \log \frac{3}{\varepsilon}+\log \frac{2}{\delta}}}{1-2\varepsilon}$ and $\tau_\delta = 2\sqrt{6}e \cdot \frac{\sqrt{l \log \frac{3}{\varepsilon^*}+\log \frac{2}{\delta}}}{1-2\varepsilon^*}$ where e is the Euler's number.

Theorem 5. *Consider a random instance of Problem* (3.2) *where the measurement matrices M_r and the exact solution z are random and distributed such that J (either $\bar{J}$ or $\tilde{J}$) is a standard Gaussian matrix. $\mathcal{B}$ consists of N elements selected uniformly on random from the measurement index set $\{1,2,...,m\}$. Consider an arbitrary constant $\delta \in (0,1)$. Introduce shortcut notation: $a = \sqrt{\sqrt{m-N}-\tau_\delta}$; $b = \sqrt{\sqrt{N}+\tau_\delta}$ and $c = \sqrt{\sqrt{m-N}+\tau_\delta}$, let $\bar{\alpha}$ be a constant satisfying*

$$\bar{\alpha} > \frac{(m-N)^{\frac{1}{4}} - N^{\frac{1}{4}}\frac{b}{a} + \frac{N}{a}\left[\frac{b}{N^{\frac{1}{4}}} + \frac{c}{(m-N)^{\frac{1}{4}}}\right]}{a - N^{\frac{1}{4}}(m-N)^{-\frac{1}{4}}b}$$

- *If the exact solution z of* (3.2) *and the prior knowledge $\hat{z}$ are such that $\alpha_{SDP} \geq \bar{\alpha}$, then with probability at least $(1-\delta)^2$ there exists a constant μ for which $(zz^\top, \eta)$ is the unique solution of the penalized SDP problem.*

- *If the exact solution z of* (3.2) *and the prior knowledge $\hat{z}$ are such that $\alpha_{SOCP} \geq \bar{\alpha}$, then with probability at least $(1-\delta)^2$ there exists a constant μ for which $(zz^\top, \eta)$ is the unique solution of the penalized SOCP problem.*

Proof. The proof follows from Lemmas 5 and 7, together with Lemma 10 to be stated later in the chapter. □

Unlike the results of the previous subsection, the stochastic bounds given above are established for a "random scenario" when the adversary is oblivious and selects the indexes of the nonzero components of the error vector η on random with a uniform distribution. Nevertheless, we still consider the noise values to be completely arbitrary and possibly engineered to have the most negative impact on the regression problem.

It is important to discuss when J becomes a Gaussian matrix in order to use the stochastic bounds in Theorem 5. The easiest scenario corresponds to the case where the true solution z is a deterministic vector while M_r's are stochastic matrices. For example, $[M_r]_{ij}$ with the distribution $\mathcal{N}(0, \frac{1}{nz_j^2})$ makes $[M_r z]$ a standard normal vector, independent of any other column vector in the matrix $\bar{J}$. Likewise, an example of the data distribution for the SOCP case is $[M_r]_{ii} \sim \mathcal{N}(0, n-1)$ and $[M_r]_{ij} \sim \mathcal{N}(0, (\frac{z_i}{z_j})^2)$ when $i \neq j$. Indeed, $R_{ij} \sim \mathcal{N}(0,1)$ whenever $i \neq j$ will make $\tilde{J}$ a standard Gaussian matrix.

The major difference between Theorem 5 and Theorem 4 is the elimination of the SSC property conditions. The simplification of the deterministic bounds was carried out for a Gaussian setting, but the developed techniques could be used to study other distributions as well. We will identify an interval for the hyperparameter μ below.

Lemma 10. *Let J be a matrix in $\mathbb{R}^{l\times m}$ that is sampled from a normal standard Gaussian distribution. Moreover, let d be a vector in $\mathbb{R}^l$ and λ be a vector in $\mathbb{R}^m$ such that $\lambda_{\mathcal{B}} = \mu \cdot s$, where μ is a scalar and the entries of s are only $+1$ or -1. Consider arbitrary numbers $\delta \in (0,1)$ and $\epsilon > 0$. Denote $\tau_{\delta,\varepsilon} = \frac{\sqrt{cl+c'\log\frac{2}{\delta}}}{1-2\varepsilon}$, where $c = 24e^2\log\frac{3}{\varepsilon}$ and $c' = 24e^2$. If*

$$\sqrt{|\mathcal{G}|} > \sqrt{|\mathcal{B}|}\frac{\sqrt{1+\Delta_{|\mathcal{B}|}}}{\sqrt{1-\Delta_{|\mathcal{G}|}}} + \frac{|\mathcal{B}|}{\alpha\sqrt{1-\Delta_{|\mathcal{G}|}}-1}\frac{\sqrt{1+\Delta_{|\mathcal{B}|}}+\sqrt{1+\Delta_{|\mathcal{G}|}}}{\sqrt{1-\Delta_{|\mathcal{G}|}}}, \tag{3.12}$$

where $\Delta_t \geq \frac{\tau_{\delta,\varepsilon}}{\sqrt{t}}$ for $t = |\mathcal{B}|$ and $|\mathcal{G}|$, then with probability at least $(1-\delta)^2$ the interval

$$\left[\frac{\|d\|_2}{\sqrt{|\mathcal{G}|(1-\Delta_{|\mathcal{G}|})}-\sqrt{|\mathcal{B}|(1+\Delta_{|\mathcal{B}|})}}, \frac{(\alpha\sqrt{(1-\Delta_{|\mathcal{G}|})}-1)\|d\|_2}{|\mathcal{B}|(\sqrt{(1+\Delta_{|\mathcal{B}|})}+\sqrt{(1+\Delta_{|\mathcal{G}|})})}\right] \tag{3.13}$$

is not empty and the system of inequalities

$$\begin{cases} \mu > \|\lambda_{\mathcal{G}}\|_\infty \\ \alpha\|d\|_2 > \|\lambda_{\mathcal{G}}\|_1 + \mu|\mathcal{B}| \end{cases}$$

is satisfied with $\lambda_{\mathcal{G}} = -J_{\mathcal{G}}^{+}(J_{\mathcal{B}}\lambda_{\mathcal{B}} + d)$ for every μ in the interval (3.13).

Proof. The proof is provided in Appendix. □

As explained after Lemma 9, the existence of the unknown interval given in (3.13) enables the design of iterative techniques to adaptively find a suitable value for μ. We will illustrate the insensitivity of the conic programs to the exact value of μ in Section 3.6.

It can be verified that to satisfy the condition (3.12), the number of good measurements $|\mathcal{G}|$ must grow quadratically in the number of bad measurements $|\mathcal{B}|$ for both the penalized SDP and the penalized SOCP relaxations. Nevertheless, in the case when $\alpha \to \infty$, this condition on $|\mathcal{G}|$ and $|\mathcal{B}|$ can be reduced to the simple inequality

$$|\mathcal{G}|(1 - \frac{\tau_{\delta,\varepsilon}}{\sqrt{|\mathcal{G}|}}) > |\mathcal{B}|(1 + \frac{\tau_{\delta,\varepsilon}}{\sqrt{|\mathcal{B}|}})$$

or equivalently

$$\begin{cases} |\mathcal{G}| > \tau_{\delta,\varepsilon}^2 \\ |\mathcal{B}| < |\mathcal{G}| + \tau_{\delta,\varepsilon}^2 - 2\tau_{\delta,\varepsilon}\sqrt{|\mathcal{G}|} \end{cases} \tag{3.14}$$

The above inequalities imply that the number of bad measurements $|\mathcal{B}|$ is allowed to increase from $\mathcal{O}(\sqrt{|\mathcal{G}|})$ to $\mathcal{O}(|\mathcal{G}|)$ as the amount of information in the prior knowledge increases.

Numerical studies show that the function $\tau_{\delta,\varepsilon}$ is expected to be fairly flat with respect to ε for practically important values of the parameters. For illustration purposes, consider $l = 100$ and $\delta = 0.05$. In this setting, $\varepsilon = \varepsilon^* = 0.05514$ is the minimum of $\tau_{\delta,\varepsilon}$ and

$$\tau_{\delta,\varepsilon^*}^2 \simeq 893.7l + 223.6\log\frac{2}{\delta} \simeq \tau_\delta^2$$

which demonstrates the asymptotic behavior of the function. Since $l_{SDP} = n$ but $l_{SOCP} = n^2 - n$, it can be concluded that the guarantee for the SDP approach works whenever the number of measurements is on order of the size of the problem, while the guarantee for the SOCP approach requires a higher number of measurements. This gives rise to a salient difference between the penalized SDP and the penalized SOCP: the SDP approach offers a higher performance over the SOCP approach but is computationally more expensive. Another important difference between the SDP and SOCP approaches—coming from the nature of the problem itself—is rooted in the definition of the coefficient α. The amount of prior knowledge needed for the SDP approach to work is less than or equal to that needed for the SOCP approach.

3.5 Robust Least-Squares Regression

Taking a step back from the penalized convex program, note that the literature on regression under sparse noise utilizes other methods along with convex relaxation techniques. To consider an alternative baseline, in this section we focus our attention on the development of an iterative technique inspired by Bhatia et al. [11]. This new method is most useful when no prior knowledge about the unknown solution is available. To build an iterative algorithm for solving Problem (3.2), consider the optimization

$$\begin{aligned} \underset{X\in\mathbb{S}^n,\ \nu\in\mathbb{R}^m}{\text{minimize}} \quad & \frac{1}{2}\sum_{r=1}^{m}\left(\langle X, M_r\rangle + \nu_r - y_r\right)^2 \\ \text{subject to} \quad & X \succeq_{\mathcal{C}} 0 \\ & \|\nu\|_0 \leq k \end{aligned} \tag{3.15}$$

where k is a parameter. This problem is nonconvex due to a cardinality constraint.

Definition 8. *Define $HT_k : \mathbb{R}^m \to \mathbb{R}^m$ as a hard thresholding operator such that*

$$[HT_k(y)]_i = \begin{cases} y_i & \text{if } |z_i| \text{ is among the } k \text{ largest-in-magnitude entries of } z \\ 0 & \text{otherwise,} \end{cases}$$

where $[HT_k(y)]_i$ denotes the i^{th} entry of $HT_k(y)$.
Consider the function

$$f(\nu) := \min_{X\succeq_{\mathcal{C}}0} \frac{1}{2}\sum_{r=1}^{m}\left(\langle X, M_r\rangle - (y_r - \nu_r)\right)^2$$

and let $\hat{W}(\nu)$ denote a solution to this problem. We propose a Hard Thresholding method for solving the quadratic regression problem, which consists of the iterative scheme

$$\nu^{t+1} = HT_k(\nu^t - d(\nu^t))$$

where

$$d(\nu) = \frac{1}{2}\nabla_\nu \left(\sum_{r=1}^{m} \left(\langle X, M_r \rangle - (y_r - \nu_r) \right)^2 \right) \Big|_{X=\hat{W}(\nu)}$$

(the symbol ∇_ν stands for the gradient with respect to ν). By Lemma 3.3.1 in [9], if $\hat{W}(\nu)$ is a continuously differentiable mapping, then $\nabla f(\nu) = d(\nu)$. Inspired by this fact, one may informally regard $d(\nu)$ as the gradient of the optimal value of the optimization problem (3.15) without its cardinality constraint. Define $w = \text{vec}\,(X)$, $\hat{w}(\nu) = \text{vec}\,(\hat{W}(\nu))$, $a_r = \text{vec}\,(M_r)$ for $r = 1, ..., m$, and $A = [a_1 \ \dots \ a_m]^\top$. It can be verified that

$$d(\nu) = A\hat{w}(\nu) - y + \nu$$

which implies that

$$HT_k(\nu - d(\nu)) = HT_k(y - A \cdot \text{vec}\,(\hat{W}(\nu)))$$

Based on this formula, we propose a conic hard thresholding method in Algorithm 1. Unlike

Algorithm 2 Conic Hard Thresholding

Input: Covariates A, responses y, corruption index k, tolerance ε, and cone $\mathcal{C}$

Initialization :

$\nu^0 \leftarrow 0$, $t \leftarrow 0$;

LOOP Process

while $\|\nu^t - \nu^{t-1}\| > \varepsilon$ **do**

  $\hat{W}^t = \arg\min_{X \succeq_{\mathcal{C}} 0} \sum_{r=1}^{m} \left(\langle X, M_r \rangle - (y_r - \nu_r^t) \right)^2$;

  $\nu^{t+1} = HT_k(y - A \cdot \text{vec}\,(\hat{W}^t))$;

  $t \leftarrow t + 1$;

end while

return $\hat{W}^{t+1}$

the penalized SDP and penalized SOCP methods, Algorithm 1 does not rely on any prior knowledge. Instead of the penalty terms in the objective, it solves a sequence of conic programs to identify the set of bad measurements through a thresholding technique. In the regime where $m \geq \frac{n(n+1)}{2}$, this algorithm with a high computational complexity can be further relaxed by letting the cone $\mathcal{C}$ be the set of symmetric matrices. We refer to this as **Algorithm 2**, where the condition $W \succeq_{\mathcal{C}} 0$ is reduced to $W = W^\top$. Note that Algorithm 2 is not effective if $m < n(n+1)/2$ because the number of measurements becomes less than the number of scalar variables in W. On the other hand, as m grows, the feasibility constraint $W \succeq_{\mathcal{C}} 0$ would more likely be satisfied for free (since the feasible set shrinks) and Algorithm 1 would perform similarly to Algorithm 2. Inspired by this property, we analyze the asymptotic behavior of Algorithm 2 for Gaussian systems below.

Lemma 11. *Suppose that* $|\mathcal{B}| < \frac{m}{20000}$, $m \geq n^2$, *and* M_r *is a random normal Gaussian matrix for* $r = 1, ..., m$. *For every* $\epsilon > 0$, *Algorithm 2 recovers a matrix* W *such that* $\|W - zz^\top\|_2 \leq \epsilon$ *within* $\mathcal{O}(\log(\frac{\|\eta\|_2}{\epsilon}) + \log(\frac{2m}{n^2+n}))$ *iterations.*

Proof. It follows from Theorem 4 by [11]. □

Let W^* be any solution obtained by Algorithm 2. Then, one can use its eigenvalue decomposition to find a vector u such that $u = \arg\min_{v \in \mathbb{C}^n} \|vv^\top - W\|_2$. Therefore,

$$\begin{aligned} \|uu^\top - zz^\top\|_2 &= \|(uu^\top - W^*) - (zz^\top - W^*)\|_2 \\ &\leq \|uu^\top - W^*\|_2 + \|zz^\top - W^*\|_2 \leq 2\varepsilon \end{aligned} \tag{3.16}$$

This means that Algorithm 2 can be used to find an approximate solution u with any arbitrary precision for the robust regression problem for Gaussian systems with a large number of measurements and yet it allows up to a constant fraction of measurements to be completely wrong. Comparing this with the guarantee $O(|\mathcal{B}|) = O(|\mathcal{G}|^{\frac{1}{2}})$ for the penalized conic methods, given by Theorem 5, it can be concluded that Algorithm 1 (or 2) is asymptotically more robust to outliers than the penalized conic program since it solves a sequence of optimization problems iteratively as opposed to a single one. This leads to another level of tradeoff between the complexity of an estimation method and its robustness level.

The theoretical analyses of this work were all on a regression model subject to a sparse error vector. However, the results can be slightly modified to account for modest noise values in addition to sparse errors. The bounds derived in this work remain the same, but the solutions found by the penalized conic problem and Algorithm 1 would no longer match the true regression solution being sought (as expected, due to a corruption in all equations). The mismatch error is a function of the modest noise values. The details are omitted for brevity; however, the this scenario will later be analyzed in numerical examples.

3.6 Experiments

In this section, we study the numerical properties of the penalized conic methods and the conic hard thresholding Algorithm 1. The simulation results in Section 3.6 and 3.6 are on physical systems for which we use the realistic prior knowledge that can be inferred from the physics of the problem (without having access to any information about the solution). In addition, we do not use any prior knowledge for the simulations in Sections 3.6 and 3.6, and select M to simply be a diagonal matrix.

Synthetic Data

Remember that, following [65], we define the representation of a sparsity pattern of an arbitrary matrix $X \in \mathbb{S}^n$ to be a binary matrix $N \in \mathbb{S}^n$ whose (i, j)-entry is equal to 1 if and

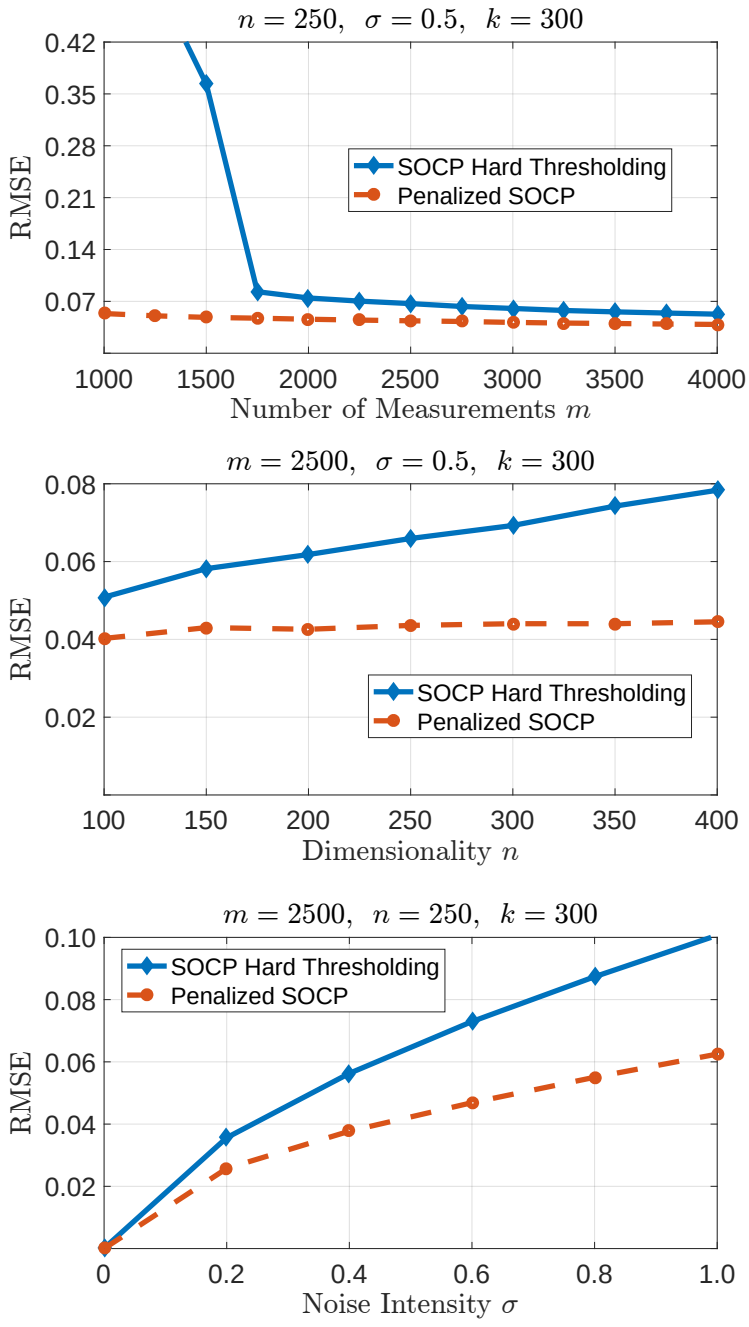


Figure 3.1: Estimation error as a function of: (a) the number of data points m, (b) the dimensionality n, (c) the standard deviation σ of additive white noise.

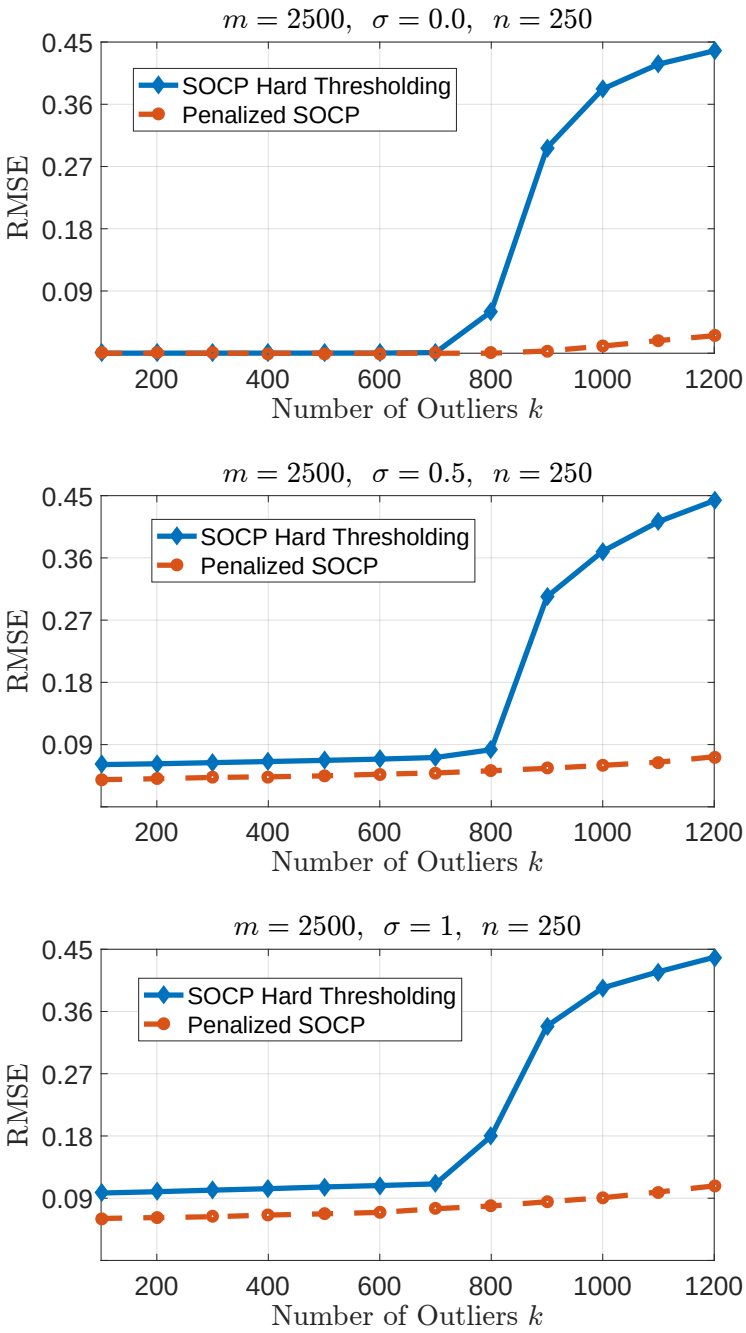


Figure 3.2: Estimation error as a function of the number of bad measurements k for different magnitudes of additive dense Gaussian noise.

only if $X_{ij} \neq 0$. Define the set of all of the matrices with the same sparsity pattern

$$\mathcal{S}(N) \triangleq \{X \in \mathbb{S}^n | X \circ N = X\}$$

We conduct some experiments on synthetically generated quadratic regression data sets with corruptions. The true model vector z is chosen to be a random unit-norm vector, while the input matrices M_r's are chosen from $\mathcal{S}(N)$ according to a common random sparsity pattern N. The nonzero entries of M_r's are generated from a normal standard distribution. The matrix N has all diagonal elements and $3n$ off-diagonal elements nonzero. The off-diagonal positions are selected uniformly. The measurements to be corrupted are chosen uniformly at random and the value of each corruption is generated uniformly from the interval $[10, 20]$. The measurements are then generated as $y_r = z^* M_r z + \eta_r + \omega_r$, where in addition to the sparse error vector η there is a random dense noise vector ω whose entries are Gaussian with zero mean and standard deviation σ. All reported results are averaged over 10 random trials.

By assuming that no prior information about the solution z is available, we set the matrix M to be the identity matrix. The parameter μ is chosen as 10^{-2}. Regarding Algorithm 1, the parameter k is selected as the true number of corrupted measurements, the tolerance ε is set to 10^{-3}, and the algorithm is terminated early if the number of conic iterations exceeds 50. In both of the methods, $\mathcal{C}$ is considered to be the $2\mathcal{PSM}$ cone. Hence, we refer to these methods as penalized SOCP and SOCP hard thresholding. Due to the sparsity in the data, the SOCP formulation can be simplified by only imposing those 2×2 constraints in (2) that correspond to the members of $\{(i, j) \mid N_{ij} = 1\}$.

We measure the performance of each algorithm using the root mean squared error (RMSE) defined as $\frac{\|z-x\|_2}{\sqrt{n}}$, where x is the output of the algorithm and z is the correct solution. Figure 3.1 shows the RSME in three different plots as a function of the number of data points m, the dimensionality n, and the additive white noise standard deviation σ. Figure 3.2 depicts the RSME as a function of the number of bad measurements k for different magnitudes of additive dense Gaussian noise. It can be observed that both the penalized conic problem and the conic hard thresholding algorithm exhibit an exact recovery property for systems with up to 700 randomly corrupted measurements out of 2500 measurements in the absence of dense Gaussian noise. The same behavior is observed in the presence of dense Gaussian noise of different magnitudes: the error of the penalized SOCP solution grows gradually, while the error of the hard thresholding algorithm has a jump at around 800 bad measurements. These simulations support the statement that up to a constant fraction of measurements could be completely wrong, and yet the unknown regression solution is found precisely.

Although the theoretical analysis provided in this chapter favors Algorithm 1 over the penalized conic problem, our empirical analysis shows that the penalized SOCP method has a better performance than the hard thresholding algorithm uniformly in the number of measurements, dimensionality, noise magnitude and the number of outliers. To explain this observation, note that the derived theoretical bounds correspond to the worst-case scenario and are more conservative for an average scenario. Moreover, the implementation of Algo-

rithm 1 in this section has limited the number of iterations to 50, while Theorem 11 requires the number of iterations to grow with respect to the amount of corruption.

The results of this part are produced using the standard MOSEK v7. SOCP-solving procedure, run in MATLAB on a 12-core 2.2GHz machine with 256GB RAM. The CPU time for each round of solving SOCP ranges from 3s (for $n = 250$, $m = 2500$) to 30s (for $n = 400$, $m = 2500$).

State Estimation for Power Systems

In this subsection, we present empirical results for the penalized conic problem with a PSD cone $\mathcal{C}$ tested on the real data for the power flow state estimation with outliers. As discussed in [69], this problem can be formulated as robust quadratic regression. The experiment is run on the PEGASE 1354-bus European system borrowed from the MATPOWER package [38, 54]. This system has 1354 nodes and the objective is to estimate the nodal voltages based on voltage magnitude and power measurements of the form $y_r = z^* M_r z + \eta_r + \omega_r$, where ω is a dense additive noise whose r^{th} entry is Gaussian with mean zero and the standard deviation equal to σ times the true value of the corresponding voltage/power parameter. The dimension of the complex vector x is 1354, which leads to 2708 real variables in the problem. In this model, the measurements are voltage magnitude squares, active and reactive nodal power injections, and active and reactive power flows from both sides of every line of the power system. This amounts to $3n + 4t = 12026$ measurements, where $t = 1991$ denotes the number of lines in the system. Note that the quadratic regression problem is complex-valued in this case.

The penalty parameter μ of the penalized conic problem is set to 10^2 and the matrix M is chosen as $-Y + \gamma I$, where Y is the susceptance matrix of the system and γ is the smallest positive number that makes M positive semidefinite. This choice of M corresponds to $\hat{z}$ being equal the eigenvector of $-Y$ associated with its smallest eigenvalue. This eigenvector provides a combination of voltages that results in the minimum amount of reactive power loss, as shown by Madani, Lavaei, and Baldick [66]. Hence, by using this particular M, we implicitly assume that the ground truth vector of voltages does not create a large amount of reactive power loss, which is a physical feature of real-world power systems. Since the penalized SDP problem is large-scale, we employ a tree decomposition technique to leverage the sparsity of the problem to solve it more efficiently [64]. The width of the tree decomposition used to reduce the complexity is equal to 12. We do not report any results on Algorithm 1 because it requires solving large-scale SDPs successively and this could be time-consuming. Moreover, the number of measurements is not high enough to use Algorithm 2, and, therefore, we will not test this method either.

The numerical results are reported in Figure 3.3. Remarkably, if the dense Gaussian noise is non-existent, the conic problem recovers the solution precisely as long as the number of bad measurements is less than 150 (note that $\sqrt{m} \simeq 109$). Note that power systems are sparse networks, their models are far from Gaussian, but the bounds from Theorem 5 are still valid in this numerical example.

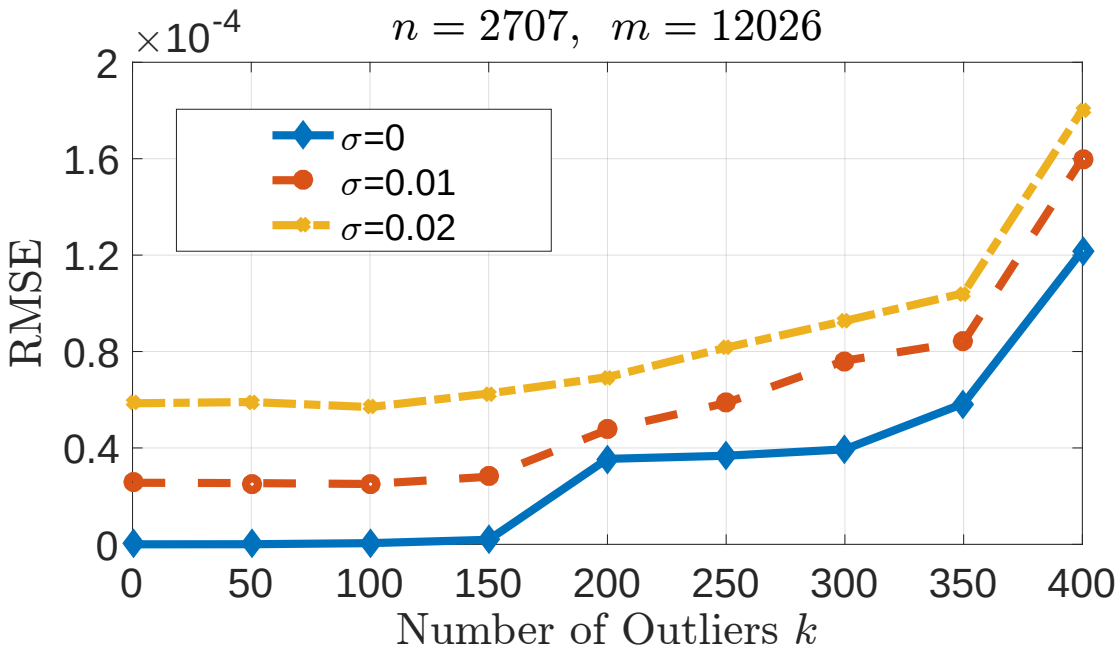


Figure 3.3: This plot shows the RMSE with respect to the number of corrupted measurements k for the PEGASE 1354-bus system.

Dynamic State Estimation

In this subsection, we demonstrate the usefulness of the proposed mathematical technique for solving sequential decision-making problems. We again consider data analytics for power systems, but leverage the fact that state estimation is solved regularly due to the time-varying and stochastic demands requested by millions of consumers. At each time instance, we use the estimated state of the system at the previous time as prior knowledge for inferring the current state of the system. We use the IEEE 300-bus benchmark system from the MATPOWER package and simulate two hours of its evolution under varying nodal active and reactive powers to reflect the changes in supply and demand. The net nodal powers are the only time-varying measurements of this system (each net power is the difference between the generation and the consumption at the node). To make the analysis realistic, we simulate both continuous changes and sudden jumps in the time-varying nodal powers. The continuous changes are modeled by a Wiener random process, while the jump values and locations are sampled from a uniform distribution that affect each time-varying measurement (curve) 5 times over the considered interval on average. The evolution of some of these measurements is depicted in Fig. 3.4.

At each time step of the simulation, happening every 2 minutes, we solve the state estimation under sparse noise via the penalized SOCP method described in Section 3.3. We let the number of corrupted measurements be 20% of the total number of measurements

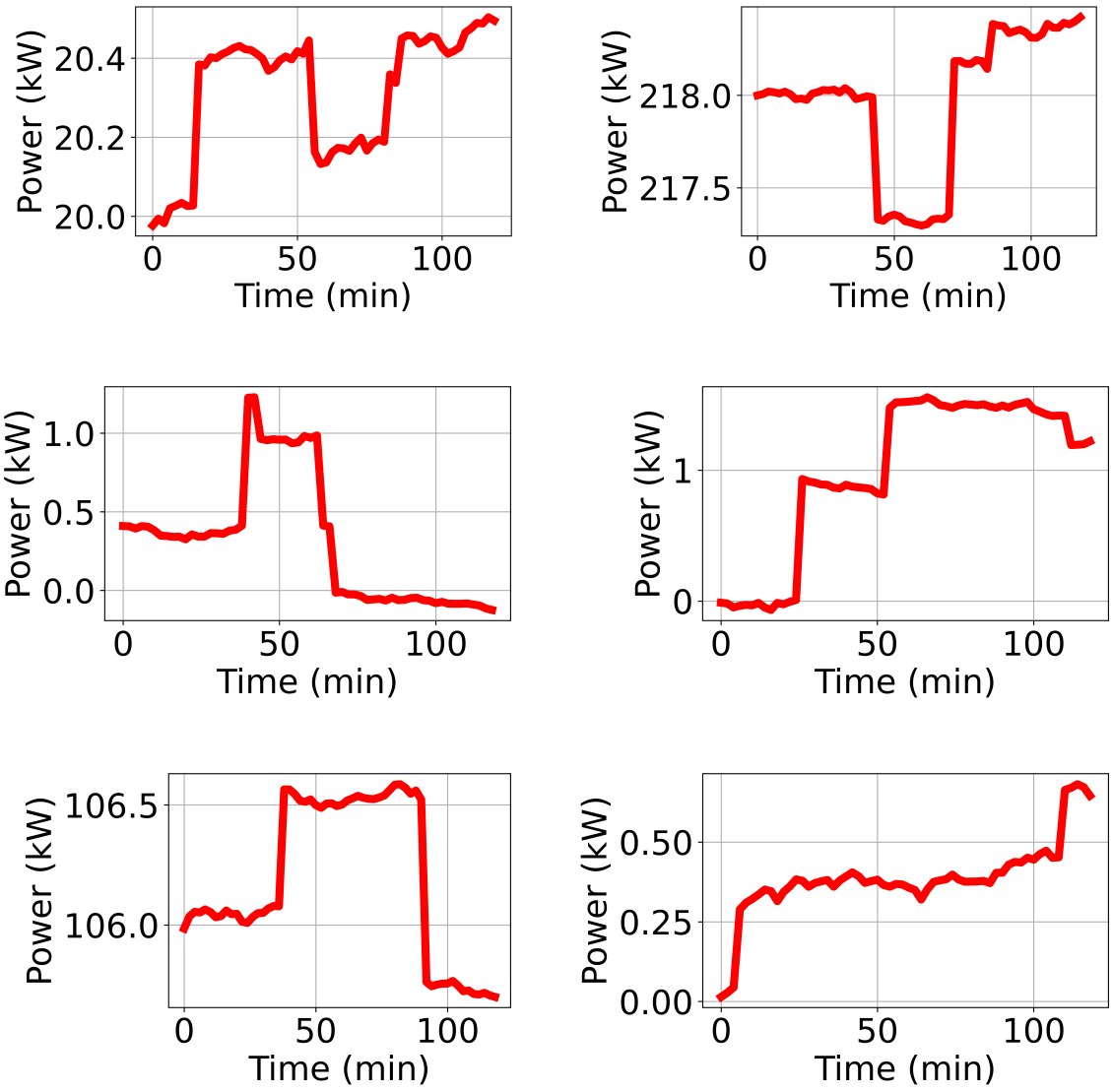


Figure 3.4: Net active (top) and reactive (bottom) powers at buses 2 (left), 50 (middle) and 93 (right) over the period of simulation.

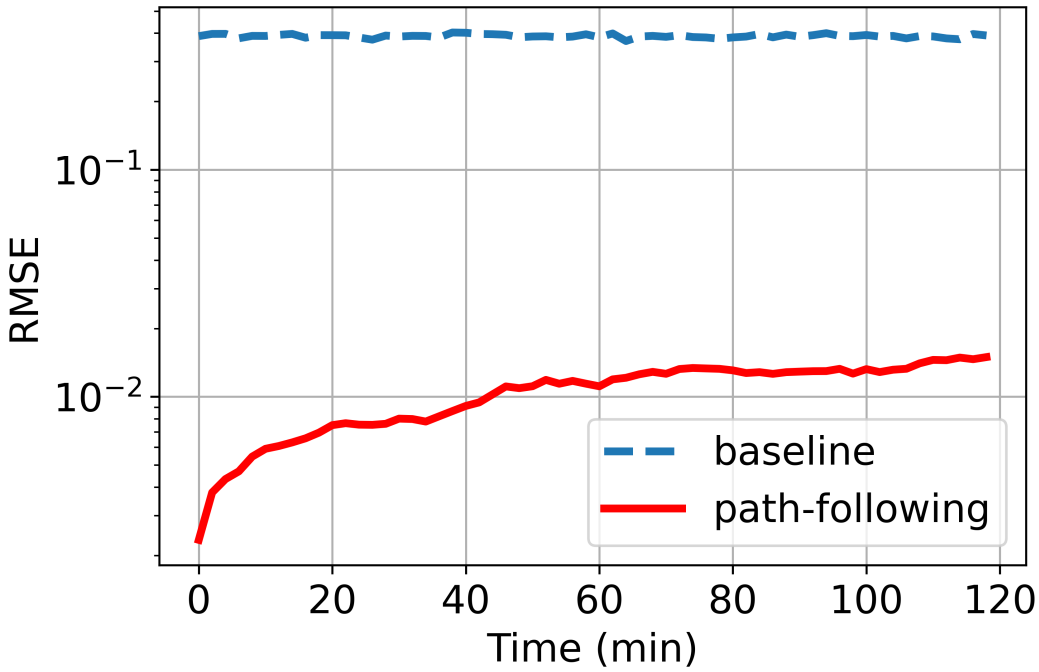


Figure 3.5: This plot shows the root mean squared error over the time period of the simulation. The dashed line denotes the error obtained by applying the SOCP penalized method with the objective matrix M constructed from the matrix Y as in Section 3.6. The solid line denotes the error obtained by applying the same method, but using a dynamic method for designing M through the path-following approach.

$m = 3444$. In the first time step, we construct the matrix M according to the formula (3.7) based on the true state of the system, set in accordance with the IEEE 300-bus system data set. At each subsequent time instance, we construct the matrix M based on the solution obtained in the previous time step. We refer to this procedure as the path-following experiment.

As a baseline for comparison, we also study a different strategy where we apply the penalized SOCP method at each time step with the objective matrix M constructed from the matrix Y as in Section 3.6, without using the prior knowledge in the solution of the previous time instance. Both the baseline and path-following experiments were conducted 5 times to produce an average result. The values of the parameter μ were chosen prior to the experiment as $5 \cdot 10^{-1}$ for the baseline and $5 \cdot 10^{-4}$ for the path-following part. They were chosen experimentally from the set $\{5, 5 \cdot 10^{-1}, 5 \cdot 10^{-2}, 5 \cdot 10^{-3}, 5 \cdot 10^{-4}, 5 \cdot 10^{-5}\}$.

Figure 3.5 demonstrates that the errors produced in the path-following experiment are smaller than the errors produced in the baseline experiment by an order of magnitude. Given

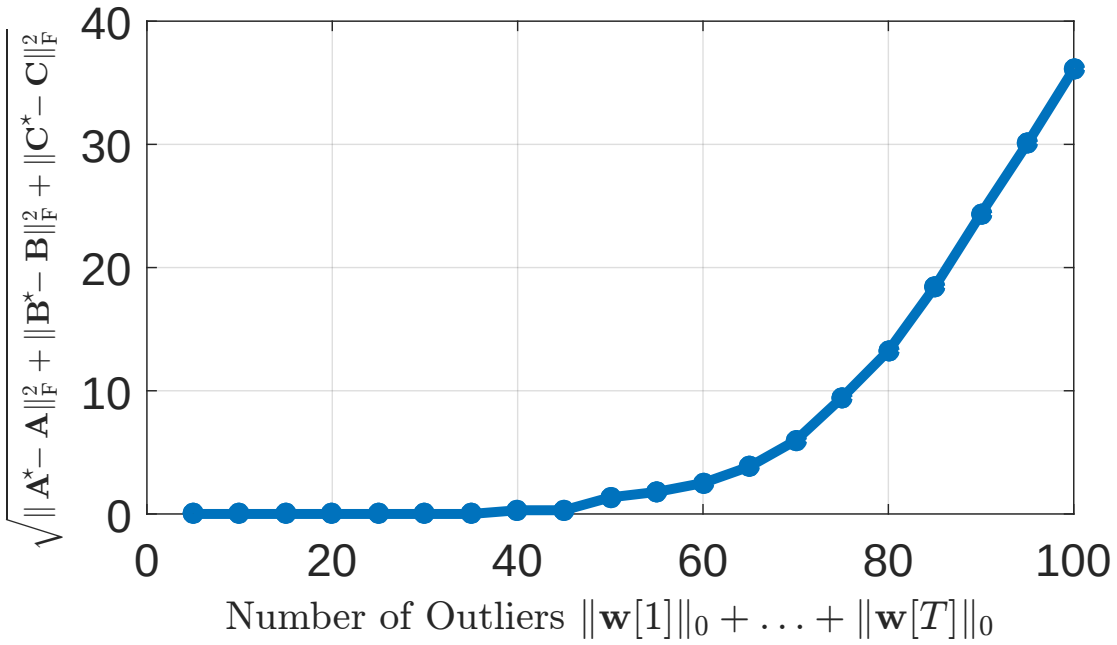


Figure 3.6: This plot shows the average estimation error of 15 random ground truth realizations with respect to the number of corrupted observations.

that the only difference between these two approaches is in the construction of the objective matrix, one can conclude that the solution of the problem in each time step can serve as useful prior knowledge for the next time step.

There is also a notable uptrend of the bottom curve, which reflects the error accumulation during the path-following experiment. This is due to the fact that the prior knowledge at each time is considered to be the state estimated at the previous time, rather than the true state of the system at the previous time. As a result, if the previous estimated state has some error, it affects learning the current state and this error accumulates over time. However, since the model (e.g., topology) of a power network changes on a slow time scale (e.g., every few hours), there is a reset in the process that eliminates the error.

Linear System Identification

Following [33], this case study is concerned with the problem of identifying the parameters of a linear dynamical system, given limited observation and non-uniform snapshots of the state vector. Consider a discrete-time linear system described by the equations

$$z[\tau+1] = A^\star z[\tau] + B^\star u[\tau] \qquad \tau = 1, 2, \ldots, T-1, \tag{3.17a}$$

$$y[\tau] = C^\star x^\star[\tau] + w^\star[\tau] \qquad \tau = 1, 2, \ldots, T, \tag{3.17b}$$

where

- $\{z[\tau] \in \mathbb{R}^n\}_{\tau=1}^T$ are the state vectors that are known at times $\tau \in \{\tau_1, \ldots, \tau_o\}$,
- $\{u[\tau] \in \mathbb{R}^m\}_{\tau=1}^T$ and $\{y[\tau] \in \mathbb{R}^k\}_{\tau=1}^T$ are the known control and observation vectors, respectively,
- $A^\star \in \mathbb{R}^{n\times n}$, $B^\star \in \mathbb{R}^{n\times m}$ and $C^\star \in \mathbb{R}^{k\times n}$ are fixed unknown matrices, and
- $\{w^\star[\tau] \in \mathbb{R}^k\}_{\tau=1}^T$ are the vectors of sparsely occurring observation errors that are unknown.

We propose to determine the triplet $(A^\star, B^\star, C^\star)$ by solving the following system of quadratic equations:

$$
\begin{aligned}
0 &= e \times x[\tau+1] - (e \times B)u[\tau] - Ax[\tau] && \tau = 1, 2, \ldots, T-1, && \text{(3.18a)}\\
y[\tau] &= Cx[\tau] + w[\tau] && \tau = 1, 2, \ldots, T, && \text{(3.18b)}\\
x[\tau] &= e \times x[\tau] && \tau = \tau_1, \tau_2, \ldots, \tau_o, && \text{(3.18c)}\\
1 &= e^2, && && \text{(3.18d)}
\end{aligned}
$$

with the unknown vector

$$
x \triangleq [e,\ x[1]^\top,\ x[2]^\top,\ \ldots,\ x[T]^\top, \operatorname{vec}\{A\}^\top, \operatorname{vec}\{B\}^\top, \operatorname{vec}\{C\}^\top]^\top \tag{3.19}
$$

and the noise estimation vectors $\{w[\tau] \in \mathbb{R}^k\}_{\tau=1}^T$. The auxiliary variable e is added to make the system of equations homogeneous, similar to the canonical quadratic regression problem (3.2). In order to solve the system of equations (3.18), we formulate the penalized SDP problem (3.3) by introducing the matrix variable X accounting for $xx^\top$. In this experiment, we use the objective function

$$
\langle M, X\rangle + \eta \times \sum_{\tau=1}^{T} \|w[\tau]\|_1 \tag{3.20}
$$

where $M = \operatorname{diag}\{[0,\ 0.001\times 1_{1\times nT},\ 1_{1\times n^2},\ 0_{1\times nm},\ 1_{1\times nk}]\}$ and $\eta = 0.1$.

We consider system identification problems with $n = 5$, $m = 2$, $k = 3$, and $T = 50$ time epochs. We assume that, for every $\tau \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, the state vector $z[\tau]$ is unknown. The elements of the ground truth matrices $A^\star \in \mathbb{R}^{5\times 5}$, $B^\star \in \mathbb{R}^{5\times 2}$, $C^\star \in \mathbb{R}^{3\times 5}$ and the control vectors $\{u[\tau]\}_{\tau=1}^T$, as well as the initial state $z[1]$ have independent Gaussian distribution with zero mean and variance $\frac{1}{3}$. Unstable ground truth matrices A with an eigenvalue outside of the unit circle are excluded. For various values of ρ, we randomly choose ρ elements of $\{y[\tau]\}_{\tau=1}^T$ and corrupt them by adding observation errors chosen uniformly from the interval $[10, 20]$. Figure 3.6, demonstrates the average estimation error for 15 trials. As shown in Figure 3.6, with up to 35 corrupted observations, the triplet $(A^\star, B^\star, C^\star)$ can be recovered with zero error. Exploiting the sparsity of the problems [76], each round of penalized SDP has been solved within 5 minutes.

Appendix

Proof of Lemma 5

The following lemma studies Slater's condition for the dual problem (3.4).

Lemma 12. *If there exists an index $r \in \{1, \dots, m\}$ such that $\hat{z}^\top M_r \hat{z} \neq 0$, then the interior of the feasible region of the problem (3.4) is not empty and strong duality holds for the penalized SDP.*

Proof. Choose $c \in \{-1, +1\}$ such that $c\hat{z}^\top M_r \hat{z} > 0$. To construct a strictly feasible point for Problem (3.4), it is enough to consider $\lambda = ue^r$, where $u > 0$ is a constant that is smaller than μ and $M + cuM_{\hat{r}} \succ 0$. Such a constant exists due to Lemma 3.2.1 in [8]. □

Proof. **of Lemma 5** Strong duality of the penalized SDP follows from Lemma 12. We aim to prove that under such a choice of $\hat{\lambda}$, the matrix $M + \sum \hat{\lambda}_r M_r$ is a PSD matrix. The complementary slackness condition:

$$\langle zz^\top, M + \sum_{r=1}^{m} \lambda_r M_r \rangle = 0$$

or equivalently

$$(M + \sum_{r=1}^{m} \lambda_r M_r) z = 0. \tag{3.21}$$

It is straightforward to verify that the condition (3.21) is satisfied for $\lambda = \hat{\lambda}$. Therefore, $\operatorname{rank}(M + \sum_{r=1}^{m} \hat{\lambda}_r M_r) \leq n - 1$. In light of Corollary 4.3.39 in [49], that $\kappa(\cdot)$ is a concave function. Now, it follows from condition (3.6b) that

$$\kappa(M + \sum_{r=1}^{m} \hat{\lambda}_r M_r) \geq \kappa(M) + \sum_{r=1}^{m} \kappa(\hat{\lambda}_r M_r) \geq \kappa(M) - 2 \sum_{r=1}^{m} |\hat{\lambda}_r| \, \|M_r\|_2 \quad > 0$$

which, combined with (3.6b), yields that $\operatorname{rank}(M + \sum_{r=1}^{m} \hat{\lambda}_r M_r) \geq n - 1$. Dual feasibility for $\hat{\lambda}_{\mathcal{G}}$ follows from condition (3.21), the above inequality, definition of κ and condition (3.6a). On the other hand, primal feasibility is satisfied for $(zz^\top, \eta)$. Therefore, $(zz^\top, \eta)$ and $\hat{\lambda}$ is a primal-dual optimal pair for the problem. This completes the proof. □

Proof of Lemma 7

Lemma 13. *The sequence $\{A^{ij} \in \mathbb{S}^2\}_{i<j}$ is a decomposition of A if and only if:*

$$\begin{cases} [A^{ij}]_{21} = [A^{ij}]_{12} = A_{ij} = A_{ji} \\ \sum_{i=2}^{n} \sum_{j=1}^{i-1} [A^{ji}]_{22} + \sum_{j=2}^{n} \sum_{i=1}^{j-1} [A^{ij}]_{11} = A_{ii} \end{cases}$$

Proof. The proof is based on basis linear algebra and is omitted for brevity. □

We define linear operations over decompositions below.

Definition 9. *Given the sequences* $\{A^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and* $\{B^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$*, define the sum:*

$$\{A^{ij}\}_{i<j}^{j\leq n} + \{B^{ij}\}_{i<j}^{j\leq n} := \{A^{ij} + B^{ij}\}_{i<j}^{j\leq n}$$

Definition 10. *For a sequence* $\{A^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and a scalar* $c \in \mathbb{R}$*, define the multiplication:*

$$c\{A^{ij}\}_{i<j}^{j\leq n} := \{cA^{ij}\}_{i<j}^{j\leq n}$$

In the following statements, we sometimes omit the indexes of decompositions when they are obvious from the context.

Lemma 14. *If* $\{A^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and* $\{B^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *are decompositions of* A *and* B *respectively, then* $\{A^{ij} + B^{ij}\}$ *is a decomposition of* $A + B$ *and* $c\{A^{ij}\}$ *is a decomposition of* cA*, for all* $c \in \mathbb{R}$*.*

Proof. To prove the first part, one can write:

$$\sum_{i<j} [e^i \;\; e^j] \left(A^{ij} + B^{ij}\right) [e^i \;\; e^j]^\top =$$

$$\sum_{i<j} [e^i \;\; e^j] A^{ij} [e^i \;\; e^j]^\top + \sum_{i<j} [e^i \;\; e^j] B^{ij} [e^i \;\; e^j]^\top = A + B$$

Moreover,

$$\sum_{i<j} [e^i \;\; e^j] cA^{ij} [e^i \;\; e^j]^\top = c \sum_{i<j} [e^i \;\; e^j] A^{ij} [e^i \;\; e^j]^\top = cA$$

This proves the second part of the lemma. □

Recall that κ is a concave function, and an analogous property of χ will be stated below.

Lemma 15. *Given the sequences* $\{A^{ij}\} = \{A^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and* $\{B^{ij}\} = \{B^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *as well as* $c \in \mathbb{R}$*, the following properties hold:*

$$\chi(\{A^{ij}\} + \{B^{ij}\}) \geq \chi(\{A^{ij}\}) + \chi(\{B^{ij}\})$$
$$\chi(c\{A^{ij}\}) \geq -|c| \max_{i'<j'} \; |\mathrm{tr}(A^{i'j'})|$$

Proof. Introduce

$$(i', j') = \arg\min_{i<j} \; \mathrm{tr}(A^{ij});$$
$$(i'', j'') = \arg\min_{i<j} \; \mathrm{tr}(B^{ij});$$
$$(i^*, j^*) = \arg\min_{i<j} \; \mathrm{tr}(A^{ij} + B^{ij});$$

The proof of the first inequality follows from the following expression:

$$\chi(\{A^{ij}\}+\{B^{ij}\}) \geq \operatorname{tr}(A^{i^*j^*}+B^{i^*j^*}) = \operatorname{tr}(A^{i^*j^*})+\operatorname{tr}(B^{i^*j^*}) \geq \\ \geq \operatorname{tr}(A^{i'j'})+\operatorname{tr}(B^{i''j''}) = \chi(\{A^{ij}\})+\chi(\{B^{ij}\})$$

For the second inequality, one can write

$$\chi(c\{A^{ij}\}) = \min_{i<j}\ \operatorname{tr}(cA^{ij}) = \min_{i<j}\ c\ \operatorname{tr}(A^{ij}) \geq -\max_{i<j}\ |c\ \operatorname{tr}(A^{ij})| \geq -|c|\max_{i<j}\ |\operatorname{tr}(A^{ij})|$$

This completes the proof. □

Lemma 16. *If the components of the initial guess are nonzero ($\hat{z}_i \neq 0$ for all $i \in \{1,\dots,n\}$) and there exists an index $r \in \{1,\dots,m\}$ such that $\hat{z}^* M_r \hat{z} \neq 0$, then the interior of the feasible region of Problem* (3.8) *is not empty, and strong duality holds for the penalized SOCP.*

Proof. Recall that $\hat{z}$ is an initial guess for the solution z and M a matrix in the objective function constructed based on $\hat{z}$. We choose $c \in \{-1,+1\}$ such that $c\hat{z}^\top M_r \hat{z} > 0$, and select $\lambda = uce^r$. It is desirable to show that if u is a sufficiently small positive number, then $M + ucM_r$ belongs to the interior of the $\mathcal{SDD}$ cone, i.e., it can be written as

$$M + ucM_r = \sum_{i<j}[e^i\ e^j]H^{ij}[e^i\ e^j]^\top,$$

where each H^{ij} is a 2×2 symmetric positive-definite matrix. By construction, the matrix M can be written as

$$M = \sum_{i<j}[e^i\ e^j]M^{ij}[e^i\ e^j]^\top$$

where each M^{ij} is a 2×2 symmetric positive semidefinite matrix that has rank 1 and $[\hat{z}_i, \hat{z}_j]$ belongs to the null space of M^{ij}. Now, we need to find a decomposition $\{B^{ij}\}_{i<j}$ of $F := cM_r$ such that $M^{ij} + uB^{ij}$ becomes positive definite if u is small. Since the null space of M^{ij} is one dimensional, it suffices to show that $[\hat{z}_i\ \ \hat{z}_j]B^{ij}[\hat{z}_i\ \ \hat{z}_j]^\top > 0$ (due to Lemma 3.2.1 in [8]). To this end, consider the following decomposition:

$$\begin{aligned}
[B^{ij}]_{11} &= (d_i - d_j - F_{ij})\tfrac{\hat{z}_j}{\hat{z}_i} + \tfrac{s}{n-1}\\
[B^{ij}]_{12} &= F_{ij}\\
[B^{ij}]_{21} &= F_{ji}\\
[B^{ij}]_{22} &= (d_j - d_i - F_{ji})\tfrac{\hat{z}_i}{\hat{z}_j} + \tfrac{s}{n-1}
\end{aligned}$$

where $d_i = \frac{\hat{z}^\top F e^i - s\hat{z}_i}{1^\top \hat{z}}$ for every $i \in \{1,\dots,n\}$ and $s = \frac{\hat{z}^\top F \hat{z}}{\hat{z}^\top \hat{z}}$. To complete the proof, it suffices to show that

$$\begin{cases} F = \sum_{i,j}[e^i\ e^j]B^{ij}[e^i\ e^j]^\top \\ [\hat{z}_i\ \hat{z}_j]B^{ij}[\hat{z}_i\ \hat{z}_j]^\top > 0 \end{cases}$$

Which according to lemma 13 is equivalent to the following three conditions satisfied simultaneously:

$$B_{12}^{ij} = F_{ij}, \quad B_{21}^{ij} = F_{ji}, \quad \forall i < j \tag{3.22}$$

$$B_{11}^{ij}\hat{z}_i^2 + B_{22}^{ij}\hat{z}_j^2 > -(F_{ij} + F_{ji})\hat{z}_i\hat{z}_j \quad \forall\ i < j \tag{3.23}$$

$$\sum_{j<i} B_{22}^{ji} + \sum_{j>i} B_{11}^{ij} = F_{ii}, \quad \forall i \tag{3.24}$$

Condition (3.22) is straightforward to verify. To verify (3.23), notice that

$$\begin{aligned} B_{11}^{ij}\hat{z}_i^2 \ &+\ B_{22}^{ij}\hat{z}_j^2 \\ &=\ ((d_i - d_j - F_{ij})\frac{\hat{z}_j}{\hat{z}_i} + \frac{s}{n-1})\hat{z}_i^2 + ((d_j - d_i - F_{ji})\frac{\hat{z}_i}{\hat{z}_j} + \frac{s}{n-1})\hat{z}_j^2 \\ &=\ (-F_{ij} - F_{ji})\hat{z}_i\hat{z}_j + s\frac{\hat{z}_i^2 + \hat{z}_j^2}{n-1} > -(F_{ij} + F_{ji})\hat{z}_i\hat{z}_j \end{aligned}$$

To analyze (3.24), one can write:

$$\begin{aligned} \sum_{j<i} B_{22}^{ji} \ &+\ \sum_{j>i} B_{11}^{ij} \\ &=\ \sum_{j<i}((d_i - d_j - F_{ij})\tfrac{\hat{z}_j}{\hat{z}_i} + \tfrac{s}{n-1}) + \sum_{j>i}((d_i - d_j - F_{ij})\tfrac{\hat{z}_j}{\hat{z}_i} + \tfrac{s}{n-1}) \\ &=\ \sum_{j\neq i}(d_i - d_j - F_{ij})\tfrac{\hat{z}_j}{\hat{z}_i} + s \\ &=\ \tfrac{d_i}{\hat{z}_i}(\sum_j \hat{z}_j - \hat{z}_i) - \tfrac{1}{\hat{z}_i}(\sum_j d_j\hat{z}_j - d_i\hat{z}_i) - \tfrac{1}{\hat{z}_i}(\sum_j F_{ij}\hat{z}_j - F_{ii}\hat{z}_i) + s \\ &=\ \tfrac{d_i}{\hat{z}_i}\sum_j \hat{z}_j - \tfrac{1}{\hat{z}_i}\sum_j d_j\hat{z}_j - \tfrac{1}{\hat{z}_i}\sum_j F_{ij}\hat{z}_j + s + F_{ii} \\ &=\ \tfrac{d_i}{\hat{z}_i}1^\top\hat{z} - \tfrac{1}{\hat{z}_i}\sum_j d_j\hat{z}_j - \tfrac{1}{\hat{z}_i}[e^i]^\top F\hat{z} + s + F_{ii} \\ &=\ \tfrac{d_i}{\hat{z}_i}1^\top\hat{z} - \tfrac{1}{\hat{z}_i 1^\top\hat{z}}\sum_j[\hat{z}^\top Fe^j - \hat{z}_j s]\hat{z}_j - \tfrac{1}{\hat{z}_i}[e^i]^\top F\hat{z} + s + F_{ii} \\ &=\ \tfrac{d_i}{\hat{z}_i}1^\top\hat{z} - \tfrac{1}{\hat{z}_i 1^\top\hat{z}}[\hat{z}^\top F\hat{z} - s\hat{z}^\top\hat{z}] - \tfrac{1}{\hat{z}_i}[e^i]^\top F\hat{z} + s + F_{ii} \\ &=\ \tfrac{d_i}{\hat{z}_i}1^\top\hat{z} - \tfrac{1}{\hat{z}_i}[e^i]^\top F\hat{z} + s + F_{ii} \\ &=\ F_{ii}. \end{aligned}$$

As a result, if u is small, then $\|uce^r\|_\infty \leq \mu$ and $A^{ij} + uB^{ij}$ is positive definite. Therefore $M + uF$ belongs to the interior of the $\mathcal{SDD}$ cone. □

Using the notation from Section 3.3, define

$$M_r^{ij} := \begin{bmatrix} R_{ij}^r & M_{ij}^r \\ M_{ji}^r & R_{ji}^r \end{bmatrix}$$

and state the following lemma.

Lemma 17. *The sequence* $\{M_r^{ij}\}_{i<j}^{j\leq n}$ *is a decomposition of* M_r.

Proof. It is straightforward to verify that

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}[M_r^{ji}]_{22} + \sum_{j=2}^{n}\sum_{i=1}^{j-1}[M_r^{ij}]_{11} = \sum_{j=1}^{n} R_{ij}^r = M_{ii}^r$$

The rest of the proof follows from Lemma 13. □

Lemma 18. $\{M^{ij}\}_{i<j} + \sum_{r=1}^{m}\lambda_r\{M_r^{ij}\}_{i<j}$ *is a decomposition of* $M + \sum_{r=1}^{m}\lambda_r M_r$

Proof. The proof follows immediately from Lemmas 14 and 17. □

Proof. **of Lemma 7** Strong duality of the penalized SOCP follows from Lemma 16. In sight of Lemma 18, it is desirable to show that under $\lambda = \hat{\lambda}$ each matrix $M^{ij} + \sum \hat{\lambda}_r M_r^{ij}$ is a PSD matrix. The complementary slackness condition can be written as

$$\langle [z_i \;\; z_j][z_i \;\; z_j]^\top, M^{ij} + \sum \lambda_r M_r^{ij}\rangle = 0$$

or, given $M^{ij} + \sum_{r=1}^{m}\hat{\lambda}_r M_r^{ij} \succeq 0$, equivalently,

$$(M^{ij} + \sum \lambda_r M_r^{ij}) \begin{bmatrix} z_i \\ z_j \end{bmatrix} = 0. \tag{3.25}$$

The condition (3.25) combined with $\chi(\{M^{ij}\}_{i<j} + \sum_{r=1}^{m}\hat{\lambda}_r\{M_r^{ij}\}_{i<j}) > 0$ yields

$$M^{ij} + \sum_{r=1}^{m}\hat{\lambda}_r M_r^{ij} \succeq 0$$

for all $i, j \in \{1, \ldots, n\}$, and thus $M + \sum_{r=1}^{m}\hat{\lambda}_r M_r \in \mathcal{SDD}$ (by Lemma 14). To satisfy the condition (3.25), λ must be such that:

$$\sum_{r=1}^{m}\lambda_r \begin{bmatrix} R_{ij}^r & M_{ij}^r \\ M_{ji}^r & R_{ji}^r \end{bmatrix} \begin{bmatrix} z_i \\ z_j \end{bmatrix} = -\begin{bmatrix} G_{ij} \\ G_{ji} \end{bmatrix} \quad \forall\, i < j$$

or equivalently

$$\sum_{r\in\mathcal{G}\cup\mathcal{B}}\lambda_r R_{ij}^r z_i = -\sum_{r\in\mathcal{G}\cup\mathcal{B}}\lambda_r M_{ij}^r z_j - G_{ij} \quad \forall\, i \neq j$$

Use the definitions given in (3.9) and rewrite this as

$$\tilde{J}_{\mathcal{G}}\lambda_{\mathcal{G}} = -(\tilde{J}_{\mathcal{B}}\hat{\lambda}_{\mathcal{B}} + \tilde{d})$$

One solution to the above system is

$$\hat{\lambda}_{\mathcal{G}} = -\tilde{J}_{\mathcal{G}}^{+}(\tilde{J}_{\mathcal{B}}\hat{\lambda}_{\mathcal{B}} + \tilde{d})$$

To conclude with dual feasibility, it is sufficient to show that

$$\chi(\{M^{ij}\}_{i<j} + \sum_{r=1}^{m} \hat{\lambda}_r\{M_r^{ij}\}_{i<j}) > 0,$$

which is guaranteed by condition (3.10b) and Lemma 15, and $\|\hat{\lambda}\|_\infty \leq \mu$ which is guaranteed by condition (3.10a). On the other hand, primal feasibility is satisfied for $(zz^\top, \eta)$. Therefore, (zz^*, η) and $(\hat{\lambda}, \{M^{ij}\} + \sum_{r=1}^{m} \hat{\lambda}_r\{M_r^{ij}\})$ is a primal-dual optimal pair for the problem. This completes the proof. □

Proof. **of Lemma 8** Consider strong duality:

$$\begin{aligned}
\langle zz^\top, M\rangle + \mu\|\eta\|_1 = -y^\top\hat{\lambda} &\iff \\
z^\top Mz + \mu\|\eta_{\mathcal{B}}\|_1 = -\sum_{r=1}^{m} z^\top\hat{\lambda}_r M_r z - \eta_{\mathcal{B}}^\top\hat{\lambda}_{\mathcal{B}} &\iff \\
z^\top\left(M + \sum_{r=1}^{m}\hat{\lambda}_r M_r\right)z = -\left(\mu\|\eta_{\mathcal{B}}\|_1 + \eta_{\mathcal{B}}^\top\hat{\lambda}_{\mathcal{B}}\right)&
\end{aligned}$$

By complementary slackness condition, we have

$$\begin{aligned}
z^\top(M + \sum_{r=1}^{m}\hat{\lambda}_r M_r)z = & \\
&= z^\top\left\{\sum_{i<j}[e^i \;\; e^j](M^{ij} + \sum\hat{\lambda}_r M_r^{ij})[e^i \;\; e^j]^\top\right\}z \\
&= \sum_{i<j}[z_i \;\; z_j](M^{ij} + \sum\hat{\lambda}_r M_r^{ij})\begin{bmatrix} z_i \\ z_j \end{bmatrix} = 0
\end{aligned}$$

Subject to the constraint $\|\hat{\lambda}\|_\infty < \mu$, the only solution of $\mu\|\eta_{\mathcal{B}}\|_1 + \eta_{\mathcal{B}}^\top\hat{\lambda}_{\mathcal{B}} = 0$ is $\hat{\lambda}_{\mathcal{B}} = -\mu\,\mathrm{sign}(\eta_{\mathcal{B}})$ □

Proof of Lemma 10

The next lemma will help prove some key results of this Chapter.

Lemma 19. *Let J be a matrix in $\mathbb{R}^{l\times m}$, d be a vector in $\mathbb{R}^l$ and λ be a vector in $\mathbb{R}^m$ such that $\lambda_{\mathcal{B}} = \mu \cdot s$, where μ is a scalar and s consists of $+1$ or -1. If*

$$\sigma_{min}(J_{\mathcal{G}}) > \sigma_{max}(J_{\mathcal{B}})$$

and

$$(\sigma_{min}(J_{\mathcal{G}}) - \sigma_{max}(J_{\mathcal{B}}))(\alpha\sigma_{min}(J_{\mathcal{G}}) - \sqrt{|\mathcal{G}|}) > \sqrt{|\mathcal{B}|}\sigma_{max}(J_{\mathcal{B}})\sqrt{|\mathcal{G}|} + |\mathcal{B}|\sigma_{min}(J_{\mathcal{G}}) \tag{3.26}$$

then the interval

$$\left[\frac{\|d\|_2}{\sigma_{min}(J_{\mathcal{G}}) - \sigma_{max}(J_{\mathcal{B}})}, \frac{(\alpha\sigma_{min}(J_{\mathcal{G}}) - \sqrt{|\mathcal{G}|})\|d\|_2}{\sqrt{|\mathcal{B}||\mathcal{G}|}\sigma_{max}(J_{\mathcal{B}}) + |\mathcal{B}|\sigma_{min}(J_{\mathcal{G}})}\right] \tag{3.27}$$

is not empty and the system of inequalities

$$\begin{cases} \mu > \|\lambda_{\mathcal{G}}\|_\infty \\ \alpha\|d\|_2 > \|\lambda_{\mathcal{G}}\|_1 + \mu|\mathcal{B}| \end{cases} \tag{3.28}$$

is satisfied by $\lambda_{\mathcal{G}} = -J_{\mathcal{G}}^+(J_{\mathcal{B}}\lambda_{\mathcal{B}} + b)$ for every μ in the interval given in (3.27).

Proof. Set $\lambda_{\mathcal{G}} = -J_{\mathcal{G}}^+(J_{\mathcal{B}}\lambda_{\mathcal{B}} + d)$ and check the set of values of μ under which the system (3.28) is satisfied. It can be shown that $\|\lambda_{\mathcal{B}}\|_\infty = \mu$; $\|\lambda_{\mathcal{B}}\|_2 = \mu\sqrt{|\mathcal{B}|}$. One can use several auxiliary inequalities:

1. $\|J_{\mathcal{G}}^+ d\|_1 \leq \sqrt{|\mathcal{G}|}\|J_{\mathcal{G}}^+ d\|_2 \leq \sqrt{|\mathcal{G}|}\|J_{\mathcal{G}}^+\|_2\|d\|_2$
2. $\|J_{\mathcal{G}}^+ d\|_\infty \leq \|J_{\mathcal{G}}^+ d\|_2 \leq \|J_{\mathcal{G}}^+\|_2\|d\|_2$
3. $\|J_{\mathcal{G}}^+ J_{\mathcal{B}}\lambda_{\mathcal{B}}\|_1 \leq \sqrt{|\mathcal{G}|}\|J_{\mathcal{G}}^+ J_{\mathcal{B}}\lambda_{\mathcal{B}}\|_2 \leq \sqrt{|\mathcal{G}|}\|J_{\mathcal{G}}^+\|_2\|J_{\mathcal{B}}\lambda_{\mathcal{B}}\|_2 \leq \mu\sqrt{|\mathcal{G}||\mathcal{B}|}\|J_{\mathcal{G}}^+\|_2\|J_{\mathcal{B}}\|_2$
4. $\|J_{\mathcal{G}}^+ J_{\mathcal{B}}\lambda_{\mathcal{B}}\|_\infty \leq \|J_{\mathcal{G}}^+ J_{\mathcal{B}}\|_\infty\|\lambda_{\mathcal{B}}\|_\infty \leq \mu\|J_{\mathcal{G}}^+ J_{\mathcal{B}}\|_2 \leq \mu\|J_{\mathcal{G}}^+\|_2\|J_{\mathcal{B}}\|_2$

It is desirable to show that

$$\begin{cases} \mu > \|J_{\mathcal{G}}^+(J_{\mathcal{B}}\lambda_{\mathcal{B}} + d)\|_\infty \\ \alpha\|b\|_2 > \|J_{\mathcal{G}}^+(J_{\mathcal{B}}\lambda_{\mathcal{B}} + d)\|_1 + \mu|\mathcal{B}| \end{cases} \tag{3.29}$$

One can use $\|J_{\mathcal{G}}^+(J_{\mathcal{B}}\lambda_{\mathcal{B}} + d)\| \leq \|J_{\mathcal{G}}^+ d\| + \|J_{\mathcal{G}}^+ J_{\mathcal{B}}\lambda_{\mathcal{B}}\|$ and relax the inequalities in (3.29) by applying the auxiliary inequalities:

$$\begin{cases} \mu > \|J_{\mathcal{G}}^+\|_2\|d\|_2 + \mu\|J_{\mathcal{G}}^+\|_2\|J_{\mathcal{B}}\|_2 \\ \alpha\|d\|_2 > \sqrt{|\mathcal{G}|}\|J_{\mathcal{G}}^+\|_2\|d\|_2 + \mu\sqrt{|\mathcal{G}|}\|J_{\mathcal{G}}^+\|_2\sqrt{|\mathcal{B}|}\|J_{\mathcal{B}}\|_2 + \mu|\mathcal{B}| \end{cases}$$

Using $\|J_B\|_2 = \sigma_{max}(J_{\mathcal{B}})$ and $\|J_{\mathcal{G}}^+\|_2 = \sigma_{min}(J_{\mathcal{G}})^{-1}$, it yields that

$$\begin{cases} \mu(1 - \frac{\sigma_{max}(J_{\mathcal{B}})}{\sigma_{min}(J_{\mathcal{G}})}) > \frac{\|d\|_2}{\sigma_{min}(J_{\mathcal{G}})} \\ \alpha\|d\|_2 > \frac{\sqrt{|\mathcal{G}|}}{\sigma_{min}(J_{\mathcal{G}})}\|d\|_2 + \mu\left(\frac{\sqrt{|\mathcal{G}|}}{\sigma_{min}(J_{\mathcal{G}})}\sqrt{|\mathcal{B}|}\sigma_{max}(J_{\mathcal{B}}) + |\mathcal{B}|\right) \end{cases}$$

One can express the bounds on μ as

$$\begin{cases} \mu > \frac{\|d\|_2}{\sigma_{min}(J_\mathcal{G}) - \sigma_{max}(J_\mathcal{B})} \\ \mu < \frac{\alpha\sigma_{min}(J_\mathcal{G})\|d\|_2 - \sqrt{|\mathcal{G}|}\|b\|_2}{\sqrt{|\mathcal{B}|}\sigma_{max}(J_\mathcal{B})\sqrt{|\mathcal{G}|} + |\mathcal{B}|\sigma_{min}(J_\mathcal{G})} \end{cases}$$

This gives rise to a condition to guarantee that the interval is not empty:

$$\frac{(\alpha\sigma_{min}(J_\mathcal{G}) - \sqrt{|\mathcal{G}|})\|d\|_2}{\sqrt{|\mathcal{B}|}\sigma_{max}(J_\mathcal{B})\sqrt{|\mathcal{G}|} + |\mathcal{B}|\sigma_{min}(J_\mathcal{G})} > \frac{\|d\|_2}{\sigma_{min}(J_\mathcal{G}) - \sigma_{max}(J_\mathcal{B})}$$

The above inequality holds by (3.26). This concludes the proof. □

Proof. **of Lemma 10** Note that the inequality (3.12) is stronger than

$$\sqrt{|\mathcal{G}|(1 - \Delta_{|\mathcal{G}|})} > \sqrt{|\mathcal{B}|(1 + \Delta_{|\mathcal{B}|})}$$

In light of Lemma 14 in [10], any randomly sampled Gaussian matrix $X \in \mathbb{R}^{l \times m}$ satisfies the inequalities

$$\lambda_{max}(XX^\top) \leq m + (1 - 2\varepsilon)^{-1}\sqrt{cml + c'm\log\tfrac{2}{\delta}}$$
$$\lambda_{min}(XX^\top) \geq m - (1 - 2\varepsilon)^{-1}\sqrt{cml + c'm\log\tfrac{2}{\delta}}$$

with probability at least $1 - \delta$ for every $\varepsilon > 0$, where $c = 24e^2\log\frac{3}{\varepsilon}$ and $c' = 24e^2$. This implies that the relations

$$\sigma_{min}(J_\mathcal{G}) \in [\sqrt{|\mathcal{G}|(1 - \Delta_{|\mathcal{G}|})}, \sqrt{|\mathcal{G}|(1 + \Delta_{|\mathcal{G}|})}]$$

and

$$\sigma_{max}(J_\mathcal{B}) \in [\sqrt{|\mathcal{B}|(1 - \Delta_{|\mathcal{B}|})}, \sqrt{|\mathcal{B}|(1 + \Delta_{|\mathcal{B}|})}]$$

are each satisfied with the probability $1-\delta$, and both are met simultaneously with probability at least $(1-\delta)^2$. By tightening the bounds in Lemma 19 with these limits on singular values, it is straightforward to verify the statement of the theorem. □

Chapter 4

Complexity of Linearly Structured Problem

In this chapter, we come back to studying a generic semidefinite affine rank feasibility (SARF) problem, which consists in finding a positive semidefinite matrix of a given rank from its linear measurements. Here we assume that instances θ belong to a low-dimensional manifold of a known structure and investigate the conditions on this manifold that would impose tractability on the corresponding SARF sub-problem. In this chapter, we measure the complexity of SARS sub-problems with respect to an approach based on local optimization of a non-convex objective, called non-convex matrix sensing. It is popular in the applications where the low-dimensionality assumption on Θ holds, such as sensor networks data processing. It is known that the optimization landscape of the non-convex matrix sensing problem may have many stationary points to which a local search method can converge. Thus, we connect the tractability of the problem and the inexistence of spurious second-order stationary points. The existing results on the properties of the landscape of non-convex matrix sensing are related to the notion of restricted isometry property (RIP). The RIP-based conditions are likely to hold for data coming from a Gaussian distribution but are too strong in the context of real-world applications where the amount of data is not exorbitantly high, especially those with inherent sparsity coming from the graph and network structure. To relax these conditions, we develop the notion of Kernel Structure Property (KSP) and use it to formulate the necessary and sufficient conditions for the inexistence of spurious local solutions in matrix sensing problems. We demonstrate the applicability and use of the novel results in analytical and numerical studies on data analytics for power systems.

4.1 Introduction

Non-convexity is the main obstacle for a guaranteed learning of optimal continuous parameters of any cyber-physical system. It is well known that many fundamental problems with a natural non-convex formulation are $\mathcal{NP}$-hard [83]. Sophisticated techniques for address-

ing this issue, like generic convex relaxations, may require working in an unrealistically high-dimensional space to guarantee the 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 any of those points. Nevertheless, empirical observations show positive results regarding the application of these approaches to several practically important instances. This provokes a major branch of research that aims to explain the success of experimental results in order to understand the boundaries of the 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 chapter. In some cases, prior information about the location of the solution is available, which significantly reduces the complexity of the search.

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 [55, 56, 51, 124, 123, 98]. 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, such as dictionary learning [2], basic non-convex M-estimators [72], shallow [94] and deep [115] artificial neural networks with different activation and loss functions [61, 81], phase retrieval [24, 103, 16] and more general matrix sensing problems [41, 53]. Particularly, significant progress has been made towards understanding different variants of the *low-rank matrix recovery,* although explanations of the simplest version called *matrix sensing* are still under active development [126, 25, 63, 41, 26, 119]. 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

$$f_{z,\mathcal{A}}(x) = \|\mathcal{A}(xx^\top - zz^\top)\|_2^2 = \|\mathcal{A}(xx^\top) - b\|_2^2, \tag{4.1}$$

where $b = \mathcal{A}(zz^\top)$. We consider this function over a general set $\mathcal{X} \subseteq \mathbb{R}^{n\times r}$, although in this section we assume $\mathcal{X} = \mathbb{R}^{n\times r}$. Recent work has generally found a certain condition on the sensing operator to be sufficient for the matrix sensing problem to be "computationally easy to solve". Precisely, this condition works with the notion of Restricted Isometry Property (RIP).

Definition 11 (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*

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

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

The existing results proving absence of spurious local minima using this notion (such as [43, 97, 96, 12, 42, 41, 84, 126]) 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, [119, 120] 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 the set $\mathcal{X} \subseteq \mathbb{R}^{n\times r}$, where $z \in \mathcal{Z} \subseteq \mathbb{R}^{n\times r}$ and $\mathcal{A}$ satisfies δ_{2r}-RIP. Note that two different sets $\mathcal{X}$ and $\mathcal{Z}$ are involved here. Since $f_{z,\mathcal{A}}$ is parametrized with $\theta = (z, \mathcal{A})$, this introduces a problem defined as

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

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

Theorem 6 ([12, 41, 120])**.** *By taking $\mathcal{X} = \mathcal{Z} = \mathbb{R}^{n\times r}$, 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 $zz^\top$ exactly. Examples of such algorithms include variants of the stochastic gradient descent (SGD) that are known to avoid saddle or even spurious local minimum points under certain assumptions [30], and widely used in machine learning [58, 13]. Besides SGD, many local search methods have been shown to be convergent to second-order critical points with high probability under mild conditions, including the classical gradient descent [60], alternating minimizations [62] and Newton's method [85]. In this chapter, 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 global convergence.

Theorem 6 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. The 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?

This question was studied earlier for special cases of matrix sensing, namely phase retrieval [96] and matrix completion [42]. In case of the phase retrieval problem, the alternative property consists in the particular distribution of the measurements operator. Regarding the matrix completion problem, the assumption includes conditions on the properties of the matrix being recovered along with conditions on the measurement operator itself. We address the above mentioned question by developing a new notion that deals with the measurement operator and precisely captures when a structured matrix recovery problem has no spurious solution over an arbitrary ball. We focus the analysis over a given ball since local search methods tend to search over a neighborhood rather than the entire space, based on prior knowledge. In Section 4.2, we motivate the need for a new notion replacing or improving RIP with real-world examples. Section 4.3 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 4.4 and 4.5 give the theory behind this notion and examples of its application. In Section 4.6, we present numerical results on the application of the developed theory in a real-world problem appearing in power systems analysis. Some of the proofs, technical details and lemmas are collected in the appendix.

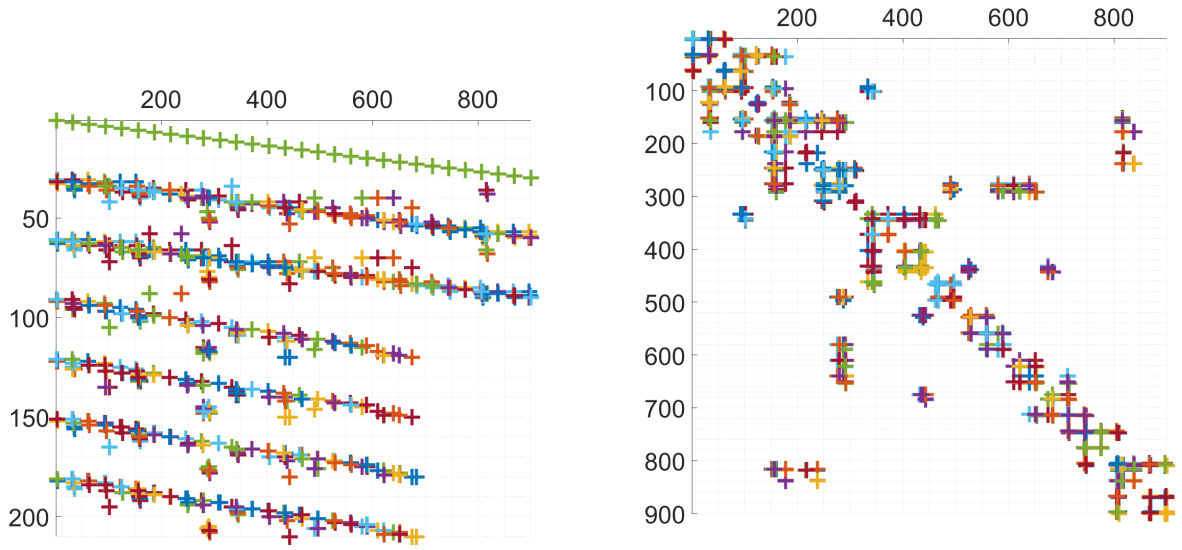
4.2 Motivating example

In this section, we motivate this work by offering a case study on data analytics for energy systems. Remember that 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

$$v^* M_i v, \quad \forall\, i \in \{1, \dots, m\}, \tag{4.2}$$

where $M_i \in \mathbb{C}^{n\times n}$ are sparse Hermitian matrices that are obtained from Y and model power flow, power injection, and voltage magnitude 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 the matrices M_i exceeds the total number of parameters contained in Y, one can regard the mapping $Y \to \{M_i\}_{i=1}^m$ as an embedding from a low-dimensional space. For a detailed discussion on the problem formulation and approaches to its solution, please refer to e.g. [122].

To formulate the problem as a low-rank matrix recovery, we introduce a sparse matrix $\mathbf{A} \in \mathbb{C}^{m\times n^2}$ with its i-th row equal to $\mathrm{vec}(M_i)^\top$. The measurement vector can be written as



(a) (b)

Figure 4.1: 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.

$\mathbf{A}\text{vec}(vv^\top)$. To find v from the measurements, one may solve the non-convex optimization problem:

$$\underset{x\in\mathbb{C}^n,\|x-\omega\|_F\leq R}{\text{minimize}} \ \|\mathbf{A}\text{vec}\,(xx^\top - vv^\top)\|^2 \tag{4.3}$$

where $\omega \in \mathbb{C}^n$ and $R \in \mathbb{R} \cup \{+\infty\}$ are some parameters determined by the prior knowledge about the solution v. In practice, this non-convex optimization problem is usually solved via local search methods, which converges to a second-order critical point at best. Since $f_v(x) = \|\mathbf{A}\text{vec}\,(xx^\top - vv^\top)\|_F^2 = \langle xx^\top - vv^\top, \mathbf{A}^\top\mathbf{A}\text{vec}\,(xx^\top - vv^\top)\rangle$, the set of critical points of the problem is defined by the linear map represented with the matrix $\mathbf{H} = \mathbf{A}^\top\mathbf{A}$, 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. 4.1a, and the structure of the corresponding $\mathbf{H}$ is described in Fig. 4.1b. The respective power network will be considered in more details in Section 4.6. 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 (4.3) while $xx^\top \neq vv^\top$. Absence of these points proves that a local search method recovers v exactly, certifying safety of its use. For example, in case of unconstrained optimization ($R = +\infty$), the answer is affirmative if the optimal objective value of the

following problem is equal to zero:

$$\begin{array}{ll} \underset{v,\ x\in\mathbb{C}^n}{\text{minimize}} & \|\mathcal{A}(xx^\top - vv^\top)\|^2 \\ \text{subject to} & \nabla f_v(x) = 0 \\ & \nabla^2 f_v(x) \succeq 0 \end{array}$$

where ∇ and ∇^2 denote the gradient and Hessian operators. However, this is an $\mathcal{NP}$-hard problem in general and cannot be solved efficiently. Even if we were able to solve 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 instances of the problem for real-world applications where the 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.

4.3 Introducing Kernel Structure

Consider a set of linear operators $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ with the matrix representations $\mathbf{A} \in \mathbb{R}^{m\times n^2}$ and a sparsity pattern $S_\mathcal{A}$. Assume that there is a set of hidden parameters $\xi \in \mathbb{R}^d$ with $d \ll m$ such that $\mathbf{A}$ is the image of ω in the space of a much higher dimension. In this way, $\mathcal{A}$ has a low-dimensional non-sparsity structure beyond sparsity, which is captured by $\mathbf{A} = \mathbf{A}(\xi)$. Without loss of generality, we assume $\mathbf{A}(0) = \mathbf{0}$. The motivating complex-valued example in Section 4.2 is a special case of this construction since it could be stated entirely with only real-valued vectors and matrices of a bigger size. We define the nonconvex objective

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

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 $zz^\top$, we need to introduce a regularity assumption:

Assumption 1. *For all $x, z \in \mathbb{R}^{n\times r}$:*

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

We will rely on Assumption 1 throughout the chapter. Note that this assumption is weaker than the assumption of existence of a $2r$-RIP constant, although it is stronger than the assumption of existence of a r-RIP constant. Another way to express the objective is

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

Here, $\mathcal{H} = \mathcal{A}^\top\mathcal{A}$ is the linear *kernel* operator that has the matrix representation $\mathbf{H} = \mathbf{A}^\top\mathbf{A}$ 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 S, so that $S_{\mathcal{H}}$-sparse operators are exclusively those satisfying the linear equation $\mathcal{S}(\mathbf{H}) = S \circ \mathbf{H} = \mathbf{0}$. Besides sparsity, $\mathcal{H}$ inherits the low-dimensional non-sparsity structure from $\mathcal{A}$, which can be captured by $\mathbf{H} = \mathbf{A}(\xi)^\top \mathbf{A}(\xi) = \mathbf{H}(\xi)$ 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 12 (Kernel Structure Property or KSP)**.** *Given a linear structure operator $\mathcal{T} : \mathbb{S}^{n^2} \to \mathbb{R}^t$, 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*

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

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.

4.4 Analysis based on KSP

Given a kernel structure $\mathcal{T}$, $\omega \in \mathbb{R}^{n \times r}$ and $R \in \mathbb{R} \cup \{+\infty\}$, we can state the problem under study as follows:

$$\left\{ \underset{x \in \mathbb{B}_R(\omega)}{\text{minimize}} \; f_{z,\mathcal{A}}(x) \;\middle|\; \mathcal{A} \text{ satisfies Assumption 1 and } \mathcal{T}\text{-KSP}, \, z \in \mathbb{B}_R(\omega) \right\}, \quad (\text{Problem}^{\text{KSP}})$$

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

The RIP constant is designed to characterize the input-output behavior of the system represented with $\mathcal{A}$. 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 $\mathcal{A}$ will have an RIP constant less than or equal to δ. Inspired by this observation, we

define the function $\mathbb{O}(x, z; \mathcal{T})$ to be the optimal objective value of the convex optimization problem:

$$\begin{aligned}
\underset{\delta\in\mathbb{R},\mathcal{H}}{\text{minimum}} \quad & \delta \\
\text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = 0 && \text{(4.4a)}\\
& \mathcal{M}_{x,z}(\mathcal{H}) \succeq 0 && \text{(4.4b)}\\
& \mathcal{T}(\mathcal{H}) = 0 && \text{(4.4c)}\\
& (1-\delta)\mathcal{I} \preceq \mathcal{H} \preceq (1+\delta)\mathcal{I} && \text{(4.4d)}
\end{aligned}$$

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 $\mathcal{T}$-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, provided that x is not on the boundary of the optimization domain $\mathbb{B}_R(\omega)$.

For completeness and for further reference within this chapter, we calculate the analytic forms of the first- and second-order derivatives of $f_{z,\mathcal{H}}$ below. Introduce a vector e and a matrix X such that for all $u \in \mathbb{R}^{n\times r}$ it holds that

$$e = \operatorname{vec}(xx^T - zz^T), \qquad\qquad X\operatorname{vec}(u) = \operatorname{vec}(xu^T + ux^T).$$

We write the operators $\mathcal{L}$, $\mathcal{M}$ and their transpose operators in closed form:

$$\begin{aligned}
\mathcal{L}_{x,z} &: \mathbb{S}^{n^2} \to \mathbb{R}^{n\times r} && \mathcal{L}_{x,z}(\mathbf{H}) = 2\cdot X^T\mathbf{H}e,\\
\mathcal{L}^T_{x,z} &: \mathbb{R}^{n\times r} \to \mathbb{S}^{n^2} && \mathcal{L}^T_{x,z}(y) = ey^TX^T + Xye^T,\\
\mathcal{M}_{x,z} &: \mathbb{S}^{n^2} \to \mathbb{S}^{nr} && \mathcal{M}_{x,z}(\mathbf{H}) = [I_r \otimes \left(\operatorname{mat}(\mathbf{H}e) + \operatorname{mat}(\mathbf{H}e)^\top\right)] + X^T\mathbf{H}X,\\
\mathcal{M}^T_{x,z} &: \mathbb{S}^{nr} \to \mathbb{S}^{n^2} && \mathcal{M}^T_{x,z}(V) = \operatorname{vec}(V)e^T + e\ \operatorname{vec}(V)^T + XVX^T.
\end{aligned}$$

where mat is the inverse operator to vec . 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 operators. Thus, the problem defining the function $\mathbb{O}$ is convex.

The function $\mathbb{O}$ is useful only for analysis of those points x that are located strictly inside the domain $\mathbb{B}_R(\omega)$. This is due to the constraints (4.4a) and (4.4b) are meant to be optimality conditions for a point inside the domain $\mathbb{B}_R(\omega)$. For a point x that lies on the boundary, we define the corresponding function $\mathbb{O}^{\partial\mathbb{B}}(x, z; \mathcal{T}, \omega)$ as the optimal objective value of the convex optimization problem:

$$\begin{aligned}
\underset{\delta,\mu\geq 0,\mathcal{H}}{\text{minimum}} \quad & \delta \\
\text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = -\mu(x-\omega) && \text{(4.5a)}\\
& P_{x-\omega}\mathcal{M}_{x,z}(\mathcal{H})P^\top_{x-\omega} \succeq 0 && \text{(4.5b)}\\
& \mathcal{T}(\mathcal{H}) = 0 && \text{(4.5c)}\\
& (1-\delta)\mathcal{I} \preceq \mathcal{H} \preceq (1+\delta)\mathcal{I} && \text{(4.5d)}
\end{aligned}$$

where $P_{x-\omega} \in \mathbb{R}^{(nr-1)\times nr}$ is the matrix of orthogonal projection onto the subspace orthogonal to $x-\omega$. The role of the function $\mathbb{O}^{\partial\mathbb{B}}$ is the same as of $\mathbb{O}$ but is designed to analyse only those values of x such that $\|x-\omega\| = R$. Note that (4.5a) and (4.5b) are the necessary optimal conditions for a solution on the boundary of $\mathbb{B}_R(\omega)$.

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:

$$(1-\delta)\mathcal{P}^\top\mathcal{P} \preceq \mathcal{P}^\top\mathcal{H}\mathcal{P} \preceq (1+\delta)\mathcal{P}^\top\mathcal{P}. \tag{4.6}$$

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:

$$\begin{aligned}
\underset{\delta\in\mathbb{R},\mathcal{H}}{\text{minimum}} \quad & \delta \\
\text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = 0 && \text{(4.7a)}\\
& \mathcal{M}_{x,z}(\mathcal{H}) \succeq 0 && \text{(4.7b)}\\
& \mathcal{T}(\mathcal{H}) = 0 && \text{(4.7c)}\\
& (1-\delta)\mathcal{P}^\top\mathcal{P} \preceq \mathcal{P}^\top\mathcal{H}\mathcal{P} \preceq (1+\delta)\mathcal{P}^\top\mathcal{P} && \text{(4.7d)}
\end{aligned}$$

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 (4.7) is obtained from (4.4) by replacing its constraint (4.4d) with the milder condition (4.6). 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, provided that x is not on the boundary of the optimization domain $\mathbb{B}_R(\omega)$.

Similarly to $\mathbb{O}^{\partial\mathbb{B}}$, the function $\mathbb{O}_P^{\partial\mathbb{B}}(x,z;\mathcal{T},\omega)$ is defined as the optimal objective value of the convex optimization problem:

$$\begin{aligned}
\underset{\delta,\mu\geq 0,\mathcal{H}}{\text{minimum}} \quad & \delta \\
\text{subject to} \quad & \mathcal{L}_{x,z}(\mathcal{H}) = -\mu(x-\omega) \\
& P_{x-\omega}\mathcal{M}_{x,z}(\mathcal{H})P_{x-\omega}^\top \succeq 0 \\
& \mathcal{T}(\mathcal{H}) = 0 \\
& (1-\delta)\mathcal{P}^\top\mathcal{P} \preceq \mathcal{P}^\top\mathcal{H}\mathcal{P} \preceq (1+\delta)\mathcal{P}^\top\mathcal{P}
\end{aligned}$$

which is designed to lower bound the constant 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, provided that x is on the boundary of $\mathbb{B}_R(\omega)$.

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

Theorem 7 (KSP necessary and sufficient conditions)**.** *For all instances of* (Problem$^{\text{KSP}}$)*, there are no spurious second-order critical points if*

$$\begin{cases} \mathbb{O}_P(x, z; \mathcal{T}) \equiv 1 \text{ over } \mathbb{B}_R(\omega) \times \mathbb{B}_R(\omega) \setminus \{xx^\top = zz^\top\} \\ \mathbb{O}_P^{\partial\mathbb{B}}(x, z; \mathcal{T}, \omega) \equiv 1 \text{ over } \partial\mathbb{B}_R(\omega) \times \mathbb{B}_R(\omega) \setminus \{xx^\top = zz^\top\} \end{cases} \tag{4.8}$$

and only if

$$\begin{cases} \mathbb{O}(x, z; \mathcal{T}) \equiv 1 \text{ over } \mathbb{B}_R(\omega) \times \mathbb{B}_R(\omega) \setminus \{xx^\top = zz^\top\} \\ \mathbb{O}^{\partial\mathbb{B}}(x, z; \mathcal{T}, \omega) \equiv 1 \text{ over } \partial\mathbb{B}_R(\omega) \times \mathbb{B}_R(\omega) \setminus \{xx^\top = zz^\top\} \end{cases} \tag{4.9}$$

This theorem is formally proven in the appendix. To elaborate on implications and practicality of the result, we present its application for a specific structure of the sensing operator below.

Ellipsoid norm: Rank 1

In this subsection, we prove a special case of Theorem 7 for the ellipsoid norm objective function and $R = +\infty$. This proof first provides useful intuition behind the proof of the general case and then simplifies the conditions of Theorem 7 to show that they always hold for a specific class of operators.

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

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

With no loss of generality, assume that $Q \in \mathbb{S}^n$ since $h(\cdot)$ really depends only on $Q^\top Q$. 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* $H_{ii} = H_{jj}$ *for all* $i, j \in \{1, \ldots, n\}$,

Then, no instance of the (Problem$^{\text{KSP}}$) *with $R = \infty$ 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. To prove this result, first notice that, Assumption 1 and the following lemma combined imply that H_{ii}, and, consequently, its decomposition $H_{ii} = QQ$, are full-rank matrices.

Lemma 20. *Given a constant $\delta_r \in [0,1)$, the matrix $Q \in \mathbb{S}^n$ satisfies*

$$(1-\delta_r)\|X\|_F^2 \leq \|QX\|_F^2 \leq (1+\delta_r)\|X\|_F^2$$

for every X such that $\operatorname{rank}(X) \leq r$ *only if* $\operatorname{rank}(Q) = n$

Proof. By contradiction, suppose that $u \in \operatorname{Ker}(Q)$ and $u \neq 0$. Take $X = uu^\top$ and observe that $QX = 0$, which contradicts that $(1-\delta_r)\|X\|_F^2 \leq \|QX\|_F^2$. □

The following lemma provides a version the conditions (4.8) and (4.9) combined for this particular structure operator.

Lemma 21. *Given $z \in \mathbb{R}^{n\times r}$, a point $x \in \mathbb{R}^{n\times r}$ is not a first-order critical point of the function $h(\cdot)$ for an arbitrary full-rank matrix Q if and only if there is $\lambda \in \mathbb{R}^{n\times r}$ such that*

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

Proof. By expanding $h(x+u)$ as

$$\begin{aligned} h(x+u) = h(x) + \operatorname{tr}(2x^\top \left((xx^\top - zz^\top)M + M(xx^\top - zz^\top)\right) u) \quad + \\ \operatorname{tr}\left(u^\top \left((xx^\top - zz^\top)M + M(xx^\top - zz^\top)\right) u + (xu^\top + ux^\top)M(xu^\top + ux^\top)\right)) \quad + o(|u|^2) \end{aligned}$$

one can arrive at a more specified expression for the second-order necessary conditions for local optimality:

$$\langle \nabla h(x), u\rangle = 2\langle Q(xx^\top - zz^\top), Q(xu^\top + ux^\top)\rangle = 0 \tag{4.10a}$$

$$\langle \nabla^2 h(x)u, u\rangle = 2\langle QQ(xx^\top - zz^\top), uu^\top\rangle + \|Q(xu^\top - ux^\top)\|_F^2 \geq 0 \tag{4.10b}$$

for all $u \in \mathbb{R}^{n\times r}$. We re-arrange the first-order condition (4.10a):

$$\left((xx^\top - zz^\top)M + M(xx^\top - zz^\top)\right) z = 0 \tag{4.11}$$

If the equation (4.11) does not hold for some $M \succ 0$, then x cannot be a critical point for that z and M. Consequently, the problem

$$\begin{aligned} \underset{M\in\mathbb{S}^n,\alpha\in\mathbb{R}}{\text{minimize}} \quad & -\alpha \\ \text{subject to} \quad & \left((yy^\top - xx^\top)M + M(yy^\top - xx^\top)\right) y = 0 && \text{(4.12a)} \\ & M - \alpha\, I \succeq 0, && \text{(4.12b)} \end{aligned}$$

is bounded from below by 0 if and only if the equation (4.11) does not hold for arbitrary $M \succ 0$. If x is critical for some $M \succ 0$, then it is unbounded.

The problem (4.12) is a semidefinite program with a zero duality gap, since $M = 0$ and $\alpha = -1$ constitute a strictly feasible primal point. We introduce the dual variable $\lambda \in \mathbb{R}^{n\times r}$

for the equality constraint (4.12a) and the dual variable $G \in \mathbb{S}^n$ for the positive semi-definite (PSD) constraint (4.12b). The dual problem can be written as

$$\max_{\lambda\in\mathbb{R}^{n\times r}, G\succeq 0} \min_{M\in\mathbb{S}^n,\alpha\in\mathbb{R}} \operatorname{tr}[\big(2\operatorname{Sym}[(y\lambda^\top + \lambda y^\top)(yy^\top - xx^\top)] - G\big)\, M] + \alpha(\operatorname{tr}(G) - 1)$$

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

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

The dual problem can be expressed as

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

This is feasible if and only if the primal problem (4.12) is bounded. Consequently, it is feasible if and only if the point $x \in \mathbb{R}^{n\times r}$ is not a critical point of the function h for all $M \succ 0$.

To eliminate the condition on the trace, notice that a PSD matrix has a nonnegative trace that is equal to zero if and only if the matrix is the zero matrix. Since the constraints are homogeneous in G, 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. □

This condition will be relaxed further for simplicity below.

Lemma 22. *Given $z \in \mathbb{R}^{n\times r}$, a point $x \in \mathbb{R}^{n\times r}$ is not a first-order critical point of the function $h(\cdot)$ for an arbitrary full-rank matrix Q if there are $T_1 \in \mathbb{R}^{r\times r}$ and $T_2 \in \mathbb{S}^r$ such that the matrix $T = \begin{bmatrix} 0 & T_1 \\ -T_1^\top & T_2 \end{bmatrix}$ satisfies the relations*

$$0 \neq [-z \;\; x]\left(T^\top P + PT\right)\begin{bmatrix} -z^\top \\ x^\top \end{bmatrix} \succeq 0 \tag{4.13}$$

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

Proof. Suppose that there exists T satisfying the condition of the lemma. Notice that

$$xT_1^\top z^\top + \frac{1}{2}xT_2x^\top + zT_1x^\top + \frac{1}{2}xT_2x^\top = [-z \;\; x]\begin{bmatrix} 0 & -T_1 \\ T_1^\top & T_2 \end{bmatrix}\begin{bmatrix} z^\top \\ x^\top \end{bmatrix} = [z \;\; x]\begin{bmatrix} 0 & T_1 \\ -T_1^\top & T_2 \end{bmatrix}\begin{bmatrix} -z^\top \\ x^\top \end{bmatrix}$$

and

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

We use the above formulas to expand the condition (4.13) and obtain

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

The proof follows immediately from applying Lemma 21 with $\lambda = zT_1 + \frac{1}{2}xT_2$. □

To prove Proposition 1, we check the previous condition for all pairs of z and x.

Proof of Proposition 1. We start by proving that no point other than 0 and $\pm z$ can be a first-order critical point of the function h. Assume that $x \notin \{0, \pm z\}$. By Lemma 22, 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

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

where $P = \begin{bmatrix} z^\top \\ x^\top \end{bmatrix} [z \;\; x]$. Consider three scenarios for x and z :
Case 1 $x = \gamma z$: One can write

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

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

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

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

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

For $\alpha = -1$, the 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 is strictly greater than 0. Thus, there exists β such that the matrix is positive definite. This implies that none of $x \notin \{0, \pm z\}$ satisfies the first-order necessary condition of local optimality for an unconstrained problem. Assume that $x = 0$. The quadratic form on the Hessian at this point

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

takes a negative value at $u = z$. Thus, it does not satisfy the second-order necessary condition of local optimality for an unconstrained problem. The points $x = \pm z$ are not spurious points, which concludes the proof. □

Ellipsoid norm: Higher ranks

The function $h(\cdot)$ defined over $\mathbb{R}^{n\times r}$ is significantly harder to study analytically if $r > 1$. An empirical analysis of Theorem 7 allows us to make a conjecture.

Conjecture 1. *For the kernel structure operator introduced in Proposition 1, no instance of the* (Problem$^{\text{KSP}}$) *with* $\mathcal{Z} = \mathbb{R}^{n\times r}$ *has a spurious second-order critical points over* $\mathbb{R}^{n\times r}$ *for an arbitrary* r*.*

This conjecture is based on the evaluation of $\mathbb{O}(x, z; \mathcal{T})$ at 72000 pairs of points $x, z \in \mathbb{R}^{8\times 3}$ randomly sampled from a standard Gaussian distribution. All of them have the optimal value 1. However, if we consider first-order critical points as well, then it is straightforward to find a counterexample. After dropping the constraint on $\mathcal{M}_{x,z}$ in the formulation of $\mathbb{O}(x, z; \mathcal{T})$ and $\mathbb{O}_P(x, z; \mathcal{T})$, one can formulate a statement similar to Theorem 7 tailored to first-order solutions. The following proposition presents a corollary of the result.

Proposition 2. *For the kernel structure operator introduced in Proposition 1, for every* $n \geq 8$ *and* $r > 1$*, there is* $z \in \mathbb{R}^{n\times r}$ *such that* (Problem$^{\text{KSP}}$) *has a spurious saddle point.*

Proof. First, we prove it for $n = 8$ and $r = 2$ by a counterexample. Consider the two points

$$x = \begin{bmatrix} 0 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 0 \\ -1 & 1 \\ -1 & 1 \\ 0 & 0 \end{bmatrix}, \qquad z = \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 1 & -1 \\ -1 & 0 \\ 1 & 0 \\ 1 & -1 \\ 1 & -1 \\ -1 & -1 \end{bmatrix}$$

and find a matrix $\mathcal{H}$ that solves $\mathbb{O}(x, z; \mathcal{T})$ without the constraint on $\mathcal{M}_{x,z}$. This will result in $\mathcal{H}$ such that $\mathcal{M}_{x,z}(\mathcal{H})$ has both negative and positive eigenvalues. For larger values of n and r, one can fill up the extra entries with zeros and the proof carries over. □

The code for reproducing the result is available on-line[1]. It took 482 tosses to generate

[1] `github.com/igormolybog/matrix-sense-global`

the counterexample of the matrices containing only ± 1 or 0 as their components. We used the uniform distribution over those matrices to generate the tosses.

DC power systems with acyclic topology

In the previous subsection, we studied one particular structure for the operator $\mathcal{A}$. Now, we analyze a real-world problem to highlight the role of the KSP. Recall that the power system discussed in Section 4.2 was an AC network for which the voltages were complex numbers. To simplify the computation, we analyze a DC system in this section, where all voltages are real-valued [44]. Assume that there are n nodes, associated with the unknown real-valued voltages $\tilde{x}_1, ..., \tilde{x}_n$. The power is measured at each node $i \in \{1, ..., n\}$ and is denoted as $\tilde{p}_i$, which can be calculated according to the formula:

$$p_i(\tilde{x}) = \sum_{j \in N(i)} \tilde{x}_i(\tilde{x}_i - \tilde{x}_j)\frac{1}{r_{ij}} = \tilde{p}_i,$$

where $N(i) \subset \{1, \ldots, n\}$ is the set of nodes adjacent to node i and $r_{ij} = r_{ji} > 0$ is the resistance of the line between nodes i and j. The least-squares formulation of the voltage recovery problem consists in minimization over the set $v \in \mathbb{B}_R(\mathbf{1})$ of

$$f(x) = \sum_{i=1}^{n} (p_i(x) - \tilde{p}_i)^2$$

which is a special case of the function (4.1). Let R be a number such that $2x_i > x_n$ for all $i \in \{1, \ldots, n-1\}$. In this subsection, we will demonstrate the application of our results on a specific topology of the network, although as discussed later on, our conclusion applies to any acyclic topology.

Suppose that the network possesses a star topology, meaning that each node i among $\{1, \ldots, n-1\}$ is connected to only one node — namely, node n — and no others. This means that $N(i) = \{n\}$ if $i \neq n$ and $N(n) = \{1, \ldots, n-1\}$. As a result, the power measurements in this particular case can be written as

$$\begin{aligned} p_i(x) &= x_i(x_i - x_n)\tfrac{1}{r_{in}}, \qquad i \in \{1, ..., n-1\} \\ p_n(x) &= \sum_{j=1}^{n-1} x_n(x_n - x_j)\tfrac{1}{r_{jn}} \end{aligned}$$

which generate a particular structure for the sensing operator. Solving $a_i \mathrm{vec}\,(xx^\top - \tilde{x}\tilde{x}^\top) = p_i$ for a_i, we conclude that the rows $a_1, \ldots, a_n$ of $\mathbf{A}$ can be written as

$$\begin{aligned} a_i &= \xi_i \mathrm{vec}\,(E_{ii} - E_{ni}), \qquad i \in \{1, ..., n-1\} \\ a_n &= -\mathrm{vec}\,(\xi_n E_{nn} + \sum_{j=1}^{n-1} \xi_j E_{jn}) \end{aligned}$$

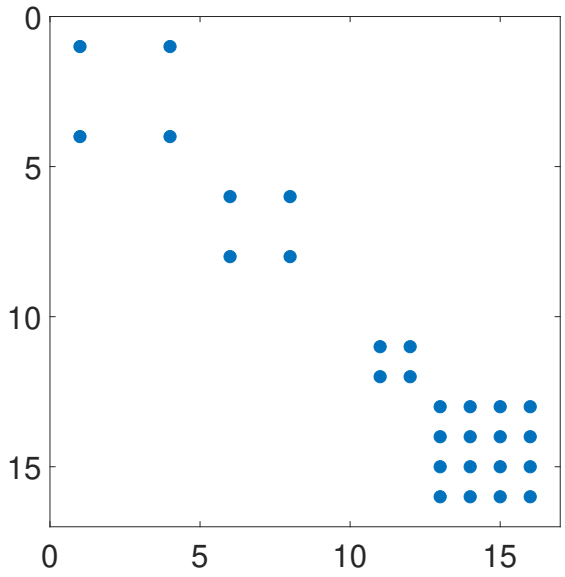


Figure 4.2: Sparsity pattern of the matrix $\mathbf{H}$ corresponding to a DC power system with a star topology consisting of four buses.

where E_{ij} is an $n \times n$ matrix with (i,j)-th entry equal to 1 and all other entries equal to zero, and where $\xi_i = \frac{1}{r_{in}} = \frac{1}{r_{ni}} > 0$ for $i \neq n$ and $\xi_n = -\sum_{j=1}^{n-1} \xi_j$. The corresponding kernel matrix is given by

$$\mathbf{H}_{star}(\xi) := \mathbf{H} = \mathbf{A}^\top \mathbf{A} = \sum_{i=1}^{n} a_i^\top a_i$$

As a result, $\mathbf{H}$ has a structured sparsity pattern: it is a block-diagonal matrix with n blocks $M_1, \ldots, M_n \in \mathbb{S}^n$ such that the first $n-1$ blocks each have only four nonzero entries:

$$M_i = \xi_i^2 [E_{ii} - E_{in} - E_{ni} + E_{nn}] = \xi_i^2 [e_n - e_i][e_n - e_i]^\top, \qquad i \in \{1, ..., n-1\}$$

where e_i is the i-th column of the $n \times n$ identity matrix. The last block of $\mathbf{H}$ is a full matrix:

$$M_n = \xi \xi^\top$$

The sparsity pattern of $\mathbf{H}$ is visualized for $n = 4$ in Figure 4.2. The matrix $\mathbf{H} = \mathbf{H}_{star}(\xi)$ also has a low-dimensional structure: the $4(n-1)+n^2$ nonzero entries of $\mathbf{H}$ quadratically depend on only $n-1$ parameters $\xi_1, \ldots, \xi_{n-1}$. In Section 4.6, we will demonstrate how linearization of the structure can be applied, while here we provide an analytical proof that deals with the nonlinear low-dimensional structure directly. This proof sheds light on some of the ideas behind Theorem 7.

The 2-RIP constant of the sensing operators that correspond to star topology power networks does not exist due to their sparsity. Therefore, Theorem 6 cannot be applied.

Nevertheless, we will show that the non-convex voltage recovery problem on a system with a star topology possesses no spurious local minima.

Proposition 3. *Consider the problem*

$$\left\{ \underset{x\in\bar{\mathbb{B}}_{1/3}(1)}{\text{minimize}}\ f_{z,\mathbf{A}}(x) \Big| \mathbf{H}=\mathbf{H}_{star}(\xi);\ z\in\bar{\mathbb{B}}_{1/3}(\mathbf{1})\ \textit{and}\ \xi^\top p(z)\neq 0 \right\},$$

or equivalently, (Problem$^{\text{KSP}}$) *under the additional constraint* $\xi^\top p(z)\neq 0$. *No instance of this problem has a spurious second-order critical point.*

Proof. Since $R<\frac{1}{3}$, it holds that $2x_i>x_n$ and $2z_i>z_n$ for all $i\in\{1,\dots,n\}$. We are interested in the landscape of the function

$$\begin{aligned} h(x) &= \text{vec}\,(xx^\top-zz^\top)^\top\mathbf{H}\text{vec}\,(xx^\top-zz^\top) \\ &= \sum_{i=1}^n (x_ix-z_iz)^\top M_i(x_ix-z_iz) \end{aligned}$$

To find the first and second derivatives, consider

$$\begin{aligned} h(x+u) &= \sum_{i=1}^n (x_ix-z_iz+x_iu+u_ix+u_iu)^\top M_i(x_ix-z_iz+x_iu+u_ix+u_iu) \\ &= h(x)+2\sum_{i=1}^n (x_iu+u_ix)^\top M_i(x_ix-z_iz)+ \\ &\quad +\sum_{i=1}^n (x_iu+u_ix)^\top M_i(x_iu+u_ix)+ \\ &\quad +2\sum_{i=1}^n (u_iu)^\top M_i(x_ix-z_iz)+o(|u|^2) \end{aligned}$$

Selecting the term that is linear in u, the gradient takes the form

$$\begin{aligned} \nabla_x h(x) &= 2\sum_{i=1}^n [x_iM_i(x_ix-z_iz)+\text{tr}[x^\top M_i(x_ix-z_iz)]e_i] \\ &= 2\sum_{i=1}^{n-1}\xi_i^2x_i(e_n-e_i)(e_n-e_i)^\top(x_ix-z_iz)+ \\ &\quad +\text{tr}[x^\top(e_n-e_i)(e_n-e_i)^\top(x_ix-z_iz)]e_i+ \\ &\quad +2x_i\xi\xi^\top(x_ix-z_iz)+\text{tr}[x^\top\xi\xi^\top(x_ix-z_iz)]e_i \end{aligned}$$

which can be written in the compact form

$$\nabla_x h(x)=B(x)[p(x)-p(z)]$$

where the (i,j)-th component of the $n\times n$ matrix $B(x)$ is

$$B_{ij}=\begin{cases} \xi_i(2x_i-x_n) & \text{if } i=j, i\neq n \\ \sum_{s=1}^{n-1}\xi_i(2x_n-x_s) & \text{if } i=j=n \\ -\xi_ix_i & \text{if } i\neq j, i=n \text{ or } i\neq j, j=n \\ 0 & \text{otherwise} \end{cases}$$

and $p(x)$ is a vector with its i-th component eqal to $p_i(x)$. If $B(x)$ is non-singular at a point x, then x is a first-order critical point of $h(x)$ in the open ball $\bar{\mathbb{B}}_{1/3}(\mathbf{1})$ if and only if $p(x) - p(z) = \mathcal{A}(xx^\top - zz^\top) = 0$, which implies that it is a global minimum. Therefore, it is essential to identify all points x such that $\det(B) = 0$.

With a slight abuse of notation, we denote $B(x)$ with the shorthand notation B. Let β_n denote the (n,n)-th entry of B, which is equal to $\sum_{s=1}^{n-1} \xi_s(2x_n - x_s)$. Represent the matrix B as a block matrix: $B = \begin{bmatrix} B' & b^\top \\ b & B'' \end{bmatrix}$ with the scalar $B' = \xi_1(2x_1 - x_n)$ and the $(n-1)$-dimensional vector $b^\top = [0, \ldots, 0, -\xi_1 x_1]$. Since $x \in \bar{\mathbb{B}}_{1/3}(\mathbf{1})$, we have $B' \neq 0$. One can write:

$$\det(B) = 0 \quad \Longleftrightarrow \quad \det(B'' - B'^{-1} b b^\top) = 0$$

The new $(n-1) \times (n-1)$ matrix $B'' - B'^{-1} b b^\top$ is equal to B'' in all components but its $(n-1, n-1)$-th entry, which changes to

$$\begin{aligned} \beta_{n-1} &= -\xi_1 \tfrac{x_1 x_n}{2x_1 - x_n} + \beta_n \\ &= -\xi_1 \tfrac{x_1 x_n}{2x_1 - x_n} + \xi_1(2x_n - x_1) + \sum_{s=2}^{n-1} \xi_s(2x_n - x_s) \\ &= -2\xi_1 \tfrac{(x_1 - x_n)^2}{2x_1 - x_n} + \sum_{j=2}^{n-1} \xi_s(2x_n - x_s) \end{aligned}$$

Repeating the above matrix reduction argument $n-1$ times yields that $\det(B) = 0$ if and only if $\beta_1 = 0$, where

$$\beta_1 = -2 \sum_{i=1}^{n-1} \xi_i \frac{(x_i - x_n)^2}{2x_i - x_n}$$

Since $2x_i - x_n > 0$, we conclude that $\beta_1 = 0$ if and only if $x_1 = \cdots = x_n$. Therefore, there are no spurious solutions outside the set $\{x : x_1 = \cdots = x_n\}$. To study this set, we derive the Hessian of $h(x)$ by extracting from $h(x+u)$ the term that is quadratic in u :

$$\begin{aligned} \nabla^2_{xx} h(x) \;= \sum_{i=1}^{n} [\;& x_i^2 M_i + x_i(e_i x^\top M_i + M_i x e_i^\top) + e_i x^\top M_i x e_i^\top + \\ & + M_i(x_i x - z_i z) e_i^\top + e_i(x_i x - z_i z)^\top M_i] \end{aligned}$$

and substitute $x_1 = \cdots = x_n = x'$ or $x = x'1$. At the same time, we substitute $M_i = \xi_i^2 [e_n - e_i][e_n - e_i]^\top$ and $M_n = \xi\xi^\top$, and note that $(e_n - e_i)^\top 1 = 1^\top(e_n - e_i) = 0$ and $\xi^\top 1 = 1^\top \xi = 0$ by construction. After simplification, we obtain that

$$\begin{aligned} \nabla^2_{xx} h(x)\big|_{x=x'1} \;= x'^2 \big[\;& \sum_{i=1}^{n-1} \xi_i^2 (e_n - e_i)(e_n - e_i)^\top + \xi\xi^\top - (\xi e_n^\top + e_n \xi^\top) z_n z^\top \xi - \\ & - \sum_{i=1}^{n-1} \xi_i^2 ((e_n - e_i) e_i^\top + e_i (e_n - e_i)^\top) z_i (z_n - z_i) \big] \end{aligned}$$

Consider the quadratic form $q(s,t) = [s \quad \dots \quad s \quad t]\nabla^2_{xx}h(x)\big|_{x=x'\mathbf{1}}[s \quad \dots \quad s \quad t]^\top$, where $[s \quad \dots \quad s \quad t] \in \mathbb{R}^n$. One can write:

$$\begin{aligned}
q(s,t) &= \sum_{i=1}^{n-1}\xi_i^2(t-s)^2 + \left(\sum_{i=1}^{n-1} s\xi_i + t\xi_n\right)^2 - \\
&\quad -2tz_n\left(\sum_{i=1}^{n-1} s\xi_i + t\xi_n\right)\left(\sum_{i=1}^{n-1} z_i\xi_i + z_n\xi_n\right) - \\
&\quad -2s(t-s)\sum_{i=1}^{n-1}\xi_i^2 z_i(z_n - z_i) = \\
&= (t-s)^2\left[\sum_{i=1}^{n-1}\xi_i^2 + \left(\sum_{i=1}^{n-1}\xi_i\right)^2\right] + \\
&\quad +2(t-s)\left[t\left(\sum_{i=1}^{n-1}\xi_i\right)\left(\sum_{i=1}^{n-1}\xi_i z_n(z_i - z_n)\right) - s\left(\sum_{i=1}^{n-1}\xi_i^2 z_i(z_n - z_i)\right)\right] = \\
&= c_1(t-s)^2 + 2(t-s)(c_2 t - c_3 s) = \\
&= (c_1 + 2c_2)t^2 - 2(c_1 + c_2 + c_3)st + (c_1 + 2c_3)s^2
\end{aligned}$$

where c_1, c_2 and c_3 are some constants introduced to shorten the expression. If $q(s,t)$ takes negative values, then $\nabla^2_{xx}h(x)\big|_{x=x'1}$ has a negative eigenvalue, and therefore $x = x'1$ cannot be a spurious second-order critical point.

Now, consider the polynomials $q(1,t)$ and $q(s,1)$. Suppose that $\sum_{i=1}^{n-1}\xi_i^2 \geq \left(\sum_{i=1}^{n-1}\xi_i\right)^2$ and consider any point $z \in \bar{\mathbb{B}}_{1/3}(1)$. Since $|z_n(z_i - z_n)| \leq \frac{4}{3}\cdot\frac{2}{3} = \frac{8}{9}$, it must hold that

$$c_1 + 2c_2 > \sum_{i=1}^{n-1}\xi_i^2 + \left(\sum_{i=1}^{n-1}\xi_i\right)^2 - 2\frac{8}{9}\left(\sum_{i=1}^{n-1}\xi_i\right)^2 = \sum_{i=1}^{n-1}\xi_i^2 - \frac{6}{9}\left(\sum_{i=1}^{n-1}\xi_i\right)^2 \geq 0$$

Thus, $q(1,t)$ is a polynomial of order 2 with respect to t with a positive leading term. Suppose that $\sum_{i=1}^{n-1}\xi_i^2 < \left(\sum_{i=1}^{n-1}\xi_i\right)^2$. Similarly,

$$c_1 + 2c_3 > \sum_{i=1}^{n-1}\xi_i^2 + \left(\sum_{i=1}^{n-1}\xi_i\right)^2 - 2\frac{8}{9}\sum_{i=1}^{n-1}\xi_i^2 = \left(\sum_{i=1}^{n-1}\xi_i\right)^2 - \frac{6}{9}\sum_{i=1}^{n-1}\xi_i^2 \geq 0$$

and thus $q(s,1)$ is a polynomial of order 2 with respect to s with a positive leading term. At least one of the polynomials $q(1,t)$ or $q(s,1)$ has a positive leading term. Both $q(1,t)$ and $q(s,1)$ have the same determinant and thus the argument to be made below can be made for any of them. Without loss of generality, assume that $q(1,t)$ has a positive leading term and takes negative values if and only if its determinant is strictly positive:

$$4(c_1 + c_2 + c_3)^2 - 4(c_1 + 2c_2)(c_1 + 2c_3) > 0.$$

It is positive if and only if $(c_2 - c_3)^2 > 0$, which is equivalent to $c_2 \neq c_3$. After substituting $c_2 = (\sum_{i=1}^{n-1} \xi_i)(\sum_{i=1}^{n-1} \xi_i z_n(z_i - z_n)) = -(\sum_{i=1}^{n-1} \xi_i) p_n(z)$ and $c_3 = (\sum_{i=1}^{n-1} \xi_i^2 z_i(z_n - z_i)) = -\sum_{i=1}^{n-1} \xi_i p_i(z)$, one can guarantee that $\nabla^2_{xx} h(x)\big|_{x=x'1}$ has a negative eigenvalue unless

$$\xi^\top p(z) = 0.$$

Otherwise, $q(1,t)$ only reaches zero at $t = 1$ and never crosses it. □

The technique of using the properties of the Schur complement to reduce the dimension of a matrix one by one can be applied to any arbitrary network with an acyclic topology. For any such network, the gradient $\nabla_x h(x)$ also takes the form $\nabla_x h(x) = B[p(x) - p(z)]$, but with a different matrix B. Applying elimination of the rows and columns of B that correspond to the leaves first, then to the first layer of parent nodes, then to the second layer of parent nodes and so forth leads to a similar result on the location of first-order critical points. Thus, in a similar way, the conclusion of Proposition 3 can be proven for any arbitrary acyclic network, but for a different value of the radius R that may not be analytically calculable. However, the proof is not generalizable to networks with cycles. The proof of Proposition 3 was based on analyzing the first- and second-order optimality conditions and exploiting the properties of the operator $\mathcal{A}$ that benefits from both sparsity and a low-dimensional structure. The ideas used in the proof help the reader understand Theorem 7. Since Proposition 3 does not apply to networks with cycles, one may instead use Theorem 7 to numerically evaluate the inexistence of spurious solutions for any particular cyclic network. This will be carried out in Section 4.6.

4.5 Combining KSP with RIP

After fixing the hyperparameters $\omega \in \mathbb{R}^{n \times r}$ and $R \in \mathbb{R} \cup \{+\infty\}$ together with the kernel structure of the sensing operators and the RIP constant, we can state the problem studied in this section as follows:

$$\left\{ \underset{x \in \mathbb{B}_R(\omega)}{\text{minimize}} f_{z,\mathcal{A}}(x) \;\middle|\; \mathcal{A} \text{ satisfies } \delta_{2r}\text{-RIP and } \mathcal{T}\text{-KSP},\ z \in \mathbb{B}_R(\omega) \right\}, \qquad (\text{Problem}^{\text{KSP+RIP}})$$

Note that (Problem$^{\text{KSP+RIP}}$) consists in the minimization of a class of functions $f_{z,\mathcal{A}}$ that correspond to some point $z \in \mathbb{B}_R(\omega)$ and some operator $\mathcal{A}$ that satisfies $\mathcal{T}$-KSP and δ_{2r}-RIP simultaneously. 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 8 (KSP+RIP necessary and sufficient conditions)**.** *For all instances of* (Problem$^{\text{KSP+RIP}}$)*, there are no spurious second-order critical points if*

$$\begin{cases} \delta_{2r} < \min\limits_{\substack{x\in\mathbb{B}_R(\omega), z\in\mathbb{B}_R(\omega)\\ xx^\top\neq zz^\top}} \mathbb{O}_P(x,z;\mathcal{T}) & \text{(4.14a)}\\ \delta_{2r} < \min\limits_{\substack{x\in\partial\mathbb{B}_R(\omega), z\in\mathbb{B}_R(\omega)\\ xx^\top\neq zz^\top}} \mathbb{O}_P^{\partial\mathbb{B}}(x,z;\mathcal{T},\omega) & \text{(4.14b)} \end{cases}$$

and only if

$$\begin{cases} \delta_{2r} < \min\limits_{\substack{x\in\mathbb{B}_R(\omega), z\in\mathbb{B}_R(\omega)\\ xx^\top\neq zz^\top}} \mathbb{O}(x,z;\mathcal{T}) & \text{(4.15a)}\\ \delta_{2r} < \min\limits_{\substack{x\in\partial\mathbb{B}_R(\omega), z\in\mathbb{B}_R(\omega)\\ xx^\top\neq zz^\top}} \mathbb{O}^{\partial\mathbb{B}}(x,z;\mathcal{T},\omega) & \text{(4.15b)} \end{cases}$$

Following the results of [120], the necessary and sufficient conditions coincide for the trivial structure operator $\mathcal{T}\equiv 0$ and $R=+\infty$.

Robustness

Consider the scenario where the measurements are corrupted with independent and identically distributed Gaussian noise. More precisely, we assume that the measurement vector b is corrupted by an additive noise that can be written as $\mathcal{A}(V)$ for some random matrix V that is probably full rank (since $\mathcal{A}(\cdot)$ is from the high-dimensional space $\mathbb{S}^n$ to the presumably low-dimensional space $\mathbb{R}^m$, we just need the mild surjectivity assumption). In this section, we show that the resulting recovery error can be bounded with high probability. For simplicity, we consider the case $R=+\infty$, but a similar argument can be used to analyze the case with a finite radius.

Theorem 9. *Consider* (Problem$^{\text{KSP+RIP}}$) *with* $R=+\infty$ *for which the condition* (4.14a) *holds. Let* $V\in\mathbb{S}^n$ *be a random matrix of arbitrary rank. Define the noisy recovery loss*

$$g(x)=\|\mathcal{A}(xx^\top-zz^\top+V)\|.$$

For every $p\in(0,1)$ *and* $\varepsilon>0$*, there exists* $\sigma=\sigma(p,\varepsilon;\mathcal{A},z)>0$ *such that for* $V\sim\mathcal{N}(0,\sigma^2 I)$*, with probability at least* p*, every second-order critical point* $x^\star$ *of* $g(x)$ *satisfies* $\|x^\star x^{\star\top}-zz^\top\|<\varepsilon$

Proof. Expand the recovery loss:

$$\begin{aligned} g(x) &= \langle xx^\top-zz^\top+V,(xx^\top-zz^\top+V)\rangle\\ &= f(x)+\langle V,\mathcal{H}(xx^\top-zz^\top)+xx^\top-zz^\top\rangle+\langle V,\mathcal{H}(V)\rangle \end{aligned}$$

We outline the proof below:

- Split $\mathbb{R}^n$ into four regions according to the behavior of $f(x)$ associated with the noiseless scenario:
  1. ε-neighborhood of the second-order critical points of $f(x)$
  2. some neighborhood of the remaining first-order critical points of $f(x)$
  3. inner compact region where the value of $\|\nabla f\|$ is bounded by some positive constants from below and from above
  4. Outer region, where $\|\nabla f\|$ is large;
- Show the existence of σ such that there are no second-order critical points of $g(x)$ in regions 2, 3 and 4 with high probability;
- Conclude that the only region that contains the second-order critical points of $g(x)$ with high probability is region 1, which coincides with the set $\{x : \|xx^\top - zz^\top\| < \varepsilon\}$.

The illustration of the regions used in the proof can be found in Figure 4.3. To prove formally, first calculate the gradient

$$\nabla_x g(x) = \nabla_x f(x) + \nabla_x \langle V, \mathcal{H}(xx^\top) + xx^\top \rangle$$

For $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, r\}$, one can write:

$$\frac{\partial}{\partial x_{ij}} \langle V, xx^\top \rangle = \langle V, e_i x_j^\top + x_j e_i^\top \rangle = 2 e_i^\top V x_j$$

where e_i is the i-th column of the $n \times n$ identity matrix. Moreover,

$$\frac{\partial}{\partial x_{ij}} \langle V, \mathcal{H}(xx^\top) \rangle = \langle V, \mathcal{H}(e_i x_j^\top + x_j e_i^\top) \rangle$$

The Hessian can also be written as

$$\nabla^2_{xx} g(x) = \nabla^2_{xx} f(x) + \nabla^2_{xx} \langle V, \mathcal{H}(xx^\top) + xx^\top \rangle$$

Similarly, for $i' \in \{1, \dots, n\}$ and $j' \in \{1, \dots, r\}$, we have

$$\frac{\partial^2}{\partial x_{ij} \partial x_{i'j'}} \langle V, xx^\top \rangle = 2\delta_{jj'} \langle V, E_{ii'} \rangle$$

where $\delta_{jj'} \in \mathbb{R}$ is defined as $\delta_{jj'} = \begin{cases} 1 \text{ if } j = j' \\ 0 \text{ otherwise} \end{cases}$, and $E_{ii'} \in \mathbb{R}^{n \times n}$ is a matrix whose (i, i')-th entry is 1 and other entries are 0. Similarly,

$$\frac{\partial^2}{\partial x_{ij} \partial x_{i'j'}} \langle V, \mathcal{H}(xx^\top) \rangle = \delta_{jj'} \langle V, \mathcal{H}(E_{ii'} + E_{i'i}) \rangle$$

By assumption, there exists $\gamma > 0$ such that $\frac{1-\delta_{2r}}{\gamma}\|xx^\top - zz^\top\|_F^2 \leq \|\mathcal{A}(xx^\top - zz^\top)\|_F^2$ for all x. This implies that $f(x)$ is a coercive functions of x for any given $\mathcal{A}$ and z. Moreover, $\|\nabla_x f(x)\|$ is also a coercive function. To show this, using the notation from Section 4.4, consider

$$\begin{aligned}\langle \tfrac{x}{\|x\|_F}, \nabla_x f(x)\rangle &= \tfrac{2}{\|x\|_F}\langle \mathrm{vec}\,(x), \mathbf{X}^\top \mathbf{H}\mathbf{e}\rangle \\ &= \tfrac{2}{\|x\|_F}\langle \mathbf{X}\mathrm{vec}\,(x), \mathbf{H}\mathbf{e}\rangle \\ &= \tfrac{2}{\|x\|_F}\langle xx^\top + xx^\top, \mathcal{H}(xx^\top - zz^\top)\rangle \\ &= \tfrac{4}{\|x\|_F}[f(x) + \langle zz^\top, \mathcal{H}(xx^\top - zz^\top)\rangle]\end{aligned}$$

Knowing that $f(x)$ grows as fast as $\|xx^\top\|_F^2 = \|x\|_F^4$ and $-\langle zz^\top, \mathcal{H}(xx^\top - zz^\top)\rangle$ grows at most as fast as $\|x\|_F^2$, we conclude that $\langle \frac{x}{\|x\|_F}, \nabla_x f(x)\rangle \to \infty$ as $\|x\| \to \infty$, which implies that $\|\nabla_x f(x)\| \to \infty$ as $\|x\| \to \infty$.

For an arbitrary $K' > 0$, define the set $C_{K'} = \{x | f(x) \leq K', \|\nabla_x f(x)\| \leq K'\}$. It is compact due to the coerciveness. The difference $\nabla_x g(x) - \nabla_x f(x)$ is linear in both V and x, while $\nabla_x f(x)$ is cubic in x. Noting that $\|\nabla_x g(x)\| \geq \|\nabla_x f(x)\| - \|\nabla_x f(x) - \nabla_x g(x)\|$, one can conclude that $\|\nabla_x g(x)\|$ is also a coercive function. Therefore, for any $p_K \in (0,1)$ there exist K and $\sigma = \sigma_K$ such that $\|\nabla_x g(x)\| > 0$ over $\mathbb{R}^n \backslash C_K$ with probability p_K. Select $p_K = \sqrt[3]{p}$ and fix the corresponding K and σ_K.

The set O_{fo} of first-order critical points of $f(x)$ is closed due to the closed graph theorem. Moreover, it is bounded due to coercievness of $f(x)$, and thus compact even when $R = +\infty$. Denote the set of second-order critical points of $f(x)$ with $O_{min} \subseteq O_{fo}$. It coincides with the set of global minimizers of $f(x)$ since the condition (4.14a) holds and Theorem 8 can be utilized. Define $U_{min} = \cup_{x \in O_{min}} \bar{\mathbb{B}}_\varepsilon(x)$, and the set $O_{rest} = O_{fo} \backslash U_{min}$ that is compact. Note that the minimum eigenvalue of $\nabla_{xx}^2 f(x)$ is strictly negative for every $x \in O_{rest}$. Since minimum eigenvalue is a continious function, there exists $\bar{\lambda} < 0$ such that $\min_{x \in O_{rest}} \lambda_{min}(\nabla_{xx}^2 f(x)) = \bar{\lambda}$. By continuity of $\nabla_{xx}^2 f(x)$ with respect to x, there exists $\xi > 0$ such that $\lambda_{min}(\nabla_{xx}^2 f(x)) < \frac{\bar{\lambda}}{2}$ for all $x \in U_{rest} = \cup_{x' \in O_{rest}} \bar{\mathbb{B}}_\xi(x')$. The difference of Hessians $\nabla_{xx}^2 g(x) - \nabla_{xx}^2 f(x)$ is linear in V and constant in x. Therefore, for any $\psi > 0$ and $p_\psi \in (0,1)$, there exists σ_ψ such that with probability p_ψ, it holds that $\|\nabla_{xx}^2 g(x) - \nabla_{xx}^2 f(x)\|_F < \psi$ for all $x \in C_K$. Select $\psi = \frac{\bar{\lambda}}{3}$ and $p_\psi = \sqrt[3]{p}$, and fix the corresponding σ_ψ. Notice that under $\sigma \leq \sigma_\psi$, with probability p_ψ there are no second-order critical points of $g(x)$ in U_{rest}.

Denote $U_{fo} = U_{rest} \cup U_{min}$ and notice that $C_K \backslash U_{fo}$ is a compact set that contains no first-order critical points of $f(x)$. Therefore, there exists $\rho > 0$ such that $\|\nabla_x f(x)\| > \rho$ for all $x \in C_K \backslash U_{fo}$. Due to the continuity of $\nabla_x g(x) - \nabla_x f(x)$, for any $\phi > 0$ and $p_\phi \in (0,1)$, there exists $\sigma = \sigma_\phi$ such that with probability p_ϕ it holds that $\|\nabla_x g(x) - \nabla_x f(x)\|_F < \phi$ for all $x \in C_K$. Select $\phi = \rho$ and $p_\phi = \sqrt[3]{p}$, and fix the corresponding σ_ϕ. Notice that under $\sigma \leq \sigma_\phi$, with probability p_ϕ there are no second-order critical points of $g(x)$ in $C_K \backslash U_{fo}$.

To conclude the proof, select $\sigma < \min\{\sigma_K, \sigma_\psi, \sigma_\phi\}$ and observe that with probability at least $p_K \times p_\psi \times p_\phi = p$ there are no second-order critical points of $g(x)$ in the set $\mathbb{R}^n \backslash U_{min} = [\mathbb{R}^n \backslash C_K] \cup U_{rest} \cup [C_K \backslash U_{fo}]$.

□

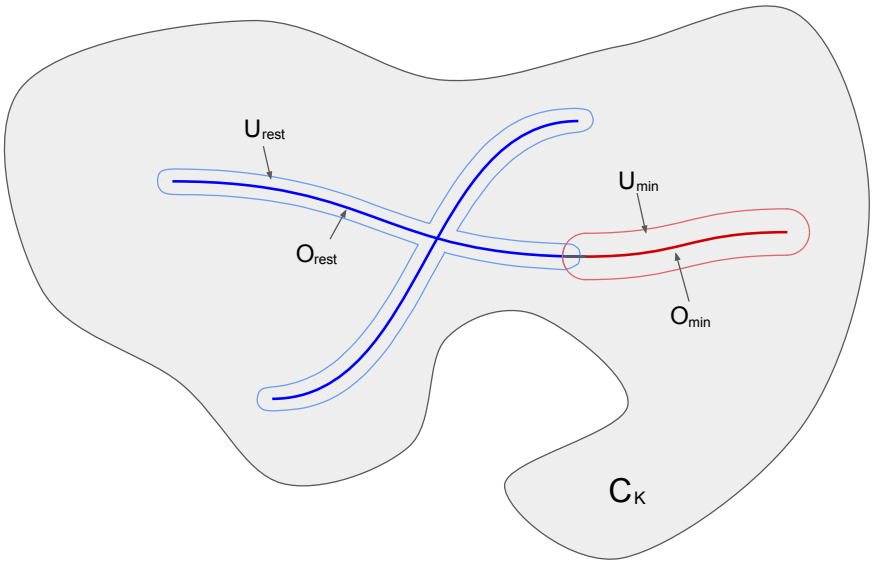


Figure 4.3: Schematic of the domain of the function $f(x)$ with highlighted regions. The grey area denotes the compact region C_K. The bold lines denote the set of first-order critical points named O_{fo} whose subset shown in red corresponds to the set of global minimizers named O_{min}, while the blue part corresponds to O_{rest}. The area countered by the red shaded line is the ε-neighborhood of O_{min}, namely U_{min}, while the area countered by the blue shaded line is the ξ-neighborhood of O_{rest}, namely U_{rest}. The proof finds that with high probability there are no second-order critical points of $g(x)$ outside of C_K (outer region 4), or inside U_{rest} (region 2), or inside $C_K \backslash [U_{rest} \cup U_{min}]$. Therefore, all such points must be located inside U_{min}.

Sparse structure and normalization

Due to Theorem 6 for the rank-1 case, the instances of (Problem$^{\text{KSP+RIP}}$) have no spurious solutions with $\mathcal{T} \equiv 0$ as long as the RIP constant 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 on 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 $\mathcal{H} = \mathcal{A}^\top \mathcal{A} \succ 0$. 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 $xx^\top \neq zz^\top$). Notice that $\mathcal{H} = \mathcal{A}^\top \mathcal{A}$ 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 [119, 120] 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 the RIP constant as discussed below.

Proposition 4. *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 the RIP constant for $\mathbf{H}$ with S' is not more restrictive than the bound for $\mathbf{H}$ with S.*

In other words, a more restrictive 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 greater than 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 $x^\top M_i x \to \frac{x^\top M_i x}{\|M_i\|_F}$. 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.

4.6 Numerical results

It is desirable to numerically study the non-convex matrix recovery in problems with a structured sensing operator. The objective is to show how the general theory developed in Section 4.3 can be applied to a real-world problem, namely the power system state estimation discussed in Section 4.2. In general, optimization problems in (4.14) and (4.15) are non-convex. Thus, we propose to use Bayessian optimization [39] in order to obtain a numerical

estimation of their solutions. 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.

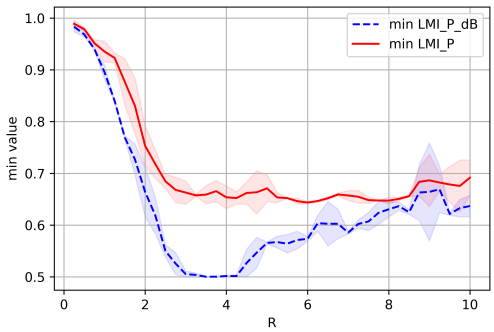
Power systems

In this section, we focus our attention on three networks named `case9`, `case14` and `case30` that are provided in the MATPOWER package [127]. For `case9`, the number of buses is $n = 9$ and there are $m = 63$ possible power measurements that can be collected, while we have $n = 14$ and $m = 98$ for `case14` and have $n = 30$ and $m = 210$ for `case30`. We denote the corresponding sensing operators with $\mathcal{A}^9$, $\mathcal{A}^{14}$ and $\mathcal{A}^{30}$. Both matrices $\mathbf{A}^{30}$ and $\mathbf{H}^{30}$ are visualized in Figure 4.1.

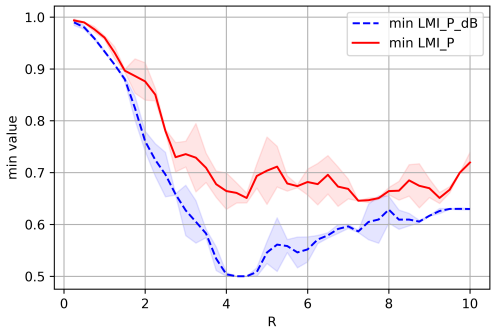
We linearize the low-dimensional structure that was discussed in Section 4.4. Repetition of the nonzero entries of $\mathbf{H}^{\cdot}$ (after some scaling) is considered as a form of low-dimensional structure, instead of the nonlinear dependence on the admittance. For example, if the entries (i, j) and (i', j') of $\mathcal{H}^{30}$ are equal, then $\mathcal{W}^{30}$ is constructed to be such that its kernel consists of matrices, for which the entries (i, j) and (i', j') are equal.

Based on this property, we form the linear operators $\mathcal{T}^9$, $\mathcal{T}^{14}$ and $\mathcal{T}^{30}$. All of the matrices in their kernel subspace are rank deficient. In this case, Theorem 8 can only provide us with the trivial upper bound on the RIP: $\delta_2 < 1$. However, this operator will allow us to use Theorem 8 to find a less conservative bound on RIP to certify the inexistence of spurious solutions for the structured mapping.

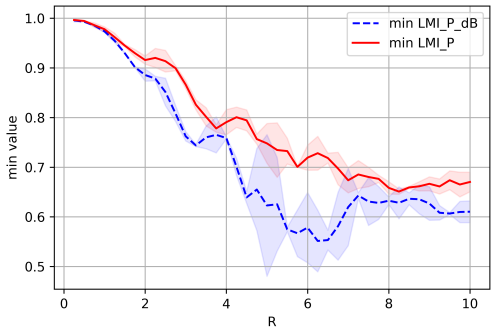
The purpose of the experiment is to study the dependence of δ_{2r} that is sufficient for the absence of spurious solutions in (Problem$^{\text{KSP+RIP}}$) on the radius R of the ball domain. Intuitively, one would expect the dependence to be monotonically decreasing, since the larger the domain is, the more solutions can appear there with some being spurious. However, this is not exactly what can be observed. Figure 4.4 shows the right-hand side of the inequalities in (4.14) from Theorem 8 for a range of values of R for three structure operators: $\mathcal{T}^9$, $\mathcal{T}^{14}$ and $\mathcal{T}^{30}$. In these experiments, the vector ω has the unit entries. The red line provides a guarantee on no spurious solutions in the interior of the domain, while the blue dashed line takes care of the spurious solutions on the boundary. Indeed, the red curve decreases monotonically and converges to a value around 0.64 for all of the experiments, while the blue dashed line decreases to 0.5 and recovers back to the same value afterwards. It turns out that 0.64 is the bound on δ_{2r} for $R = +\infty$ in each of the cases as well. This interesting behavior can be explained qualitatively. Consider a toy example with three cases in Figure 4.5, where the domain grows from Case I to Case III. There are no spurious solutions in case I, whereas one appears in case II and disappears in case III. Notice that the spurious solution can only appear on the boundary, which motivates the steady behavior of the red curve in Figure 4.4. Recall that the threshold 0.5 is valid for the trivial structure operator $\mathcal{T} \equiv 0$ and $R = +\infty$ and the blue curve never goes below it. Therefore, the constructed conditions of the absence of spurious local optimality are strictly superior to the previously known bound.



(a) $\mathcal{T}^9$



(b) $\mathcal{T}^{14}$



(c) $\mathcal{T}^{30}$

Figure 4.4: The outcome of the minimization of $\mathbb{O}_P(x,z)$ and $\mathbb{O}_P^{\partial\mathbb{B}}(x,z)$ with the Bayessian optimization toolbox. The resulting value is the approximation of the right-hand side of the inequalities in (4.14) and can be used in Theorem 8 to estimate the lower bound on the sufficient RIP constant for global optimality. The values of the radius of the domain ball $\mathbb{B}_R(\omega)$ are on the x-axis, and the corresponding approximations of $\min \mathbb{O}_P(x,z)$ and $\min \mathbb{O}_P^{\partial\mathbb{B}}(x,z)$ are on the y-axis. The red line depicts the lowest observed value of the function $\mathbb{O}_P(x,z)$ and the blue dashed line depicts the minimum value of the function $\mathbb{O}_P^{\partial\mathbb{B}}(x,z)$.

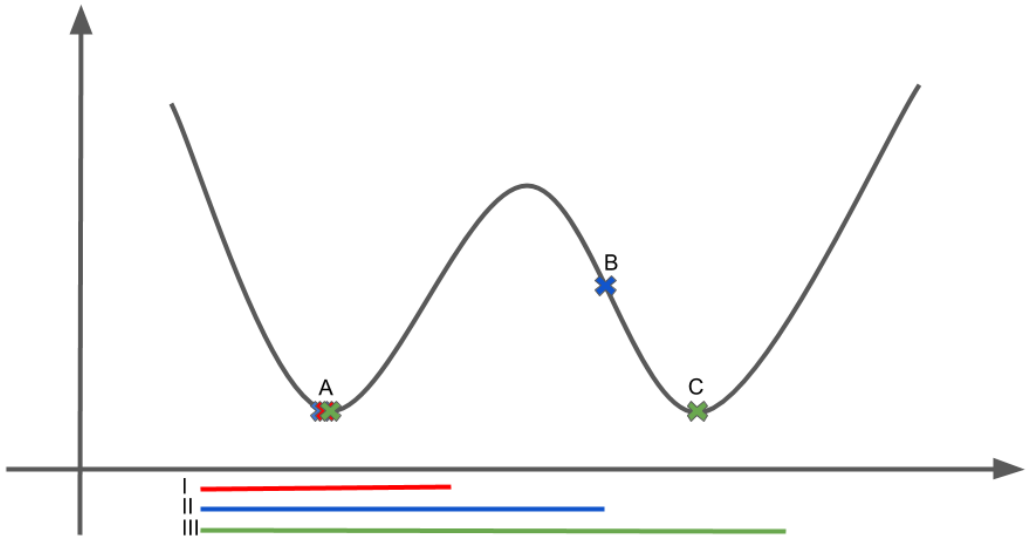


Figure 4.5: Illustration for the local solution on the boundary. Three cases are considered, each marked with a different color. The colored intervals along the x-axis depict the domain in each of the cases, while the colored crosses denote the local solutions.

The above simulations were based on the networks provided in the package MATPOWER 7.0b1 [127]. Keeping the structure of a network, we set the parameters of the lines equal to each other to be able to better visualize the operator $\mathcal{H}$. All of the presented simulations were performed using the MATLAB bayesopt toolbox, and the MATLAB modeling toolbox CVX [45, 46] with SDPT3 [101, 102] as the underlying solver.

Synthetic data

In this subsection, we present numerical studies of the matrix recovery problem for structured sensing operators obtained from random ensembles. For simplicity, we set $R = +\infty$. In this section, the smallest value of δ_r such that $\mathcal{A}$ satisfies the δ_r-RIP property is referred to as *the best RIP constant* of the map $\mathcal{A}$.

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 consider the same form of this operator as in the experiment on power systems data. 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 $\mathbf{H} = h_1 E_1 + \cdots + h_{d_\mathcal{W}} E_{d_\mathcal{W}}$, where 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

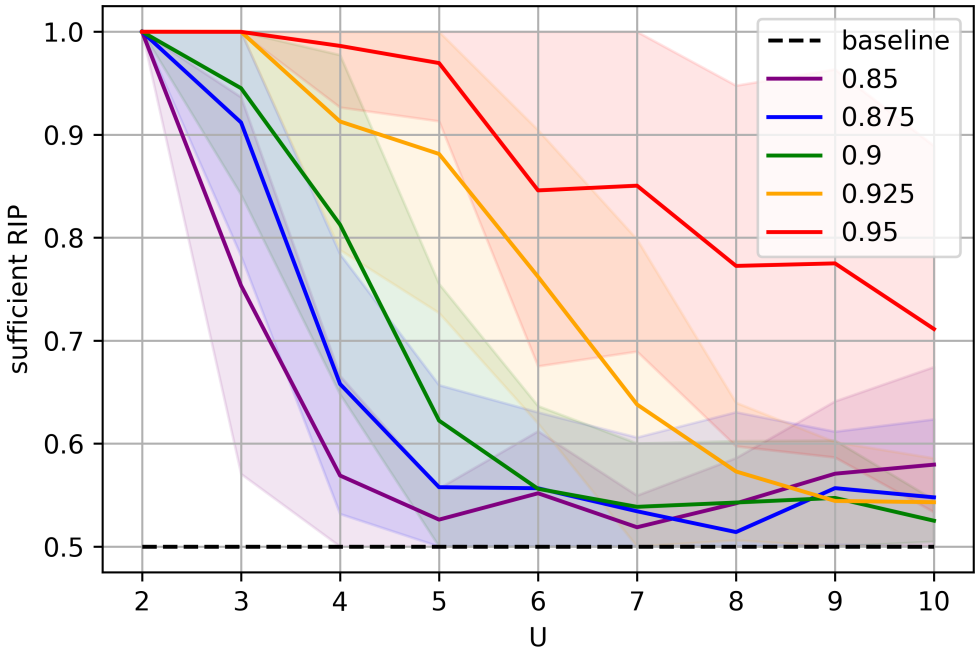


Figure 4.6: The average of sufficient best RIP constant obtained from the developed analytic framework (Theorem 8) for random structures generated from the distribution $RS(p_0, U)$ (each colored line stands for one specific value of p_0), compared with the baseline method from Theorem 6 (shown as black and dashed). Shaded area represents the standard deviation window.

$\{\beta_1 E_1 + \ldots + \beta_{d_{\mathcal{W}}} E_{d_{\mathcal{W}}} | \beta_1, \cdots, \beta_{d_{\mathcal{W}}} \in \mathbb{R}\}$ as its kernel.

We introduce a distribution $RS(p_0, U)$ over the space of structure operators by describing the sampling scheme below. 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 $\mathbf{H}_{\text{st}} = \mathbf{A}_{\text{st}}^\top \mathbf{A}_{\text{st}}$ and construct the sparsity operator $\mathcal{S}$ and the extra structure operator $\mathcal{W}$ as discussed before. 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 a parameter qualifying the amount of sparsity structure in the problem, and U is a parameter qualifying the amount of additional structure.

Figure 4.6 depicts the estimated sufficient RIP to guarantee the existence of no spurious second-order critical points in random problems with different values for the sparsity (p_0) and the unique counter (U). The sufficient RIP is obtained from Theorem 8 by imposing the

KSP. 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 a higher p_0 means more sparsity and a lower U means more extra structure. Although it was observed theoretically that sparsity alone could not guarantee an 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 1, which is a counter-intuitive result. The sufficient RIP seems to decay exponentially as we relax extra structure by increasing U, but with different bases for different p_0. This behavior coincides with the one predicted in Proposition 4. 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 experiment demonstrates that our method can be successfully applied to matrix sensing with randomly generated structure. The key takeaway from this experiment is that our method captures the trade-off between the sparsity and the low-dimensional structural properties of a given mapping. It shows that imposing restrictions on structure significantly affects the sufficient RIP, which leads to certifying the absence of spurious solutions under far less restrictive requirements (by improving the previous RIP bound 0.5 for arbitrary mappings).

Appendix

In this part, we will prove Theorems 7 and 8 via showing the nonexistence of a counterexample. Specifically, given $\mathcal{T}$, for a point x and a parameter value z, we aim to find a value $\delta_{2r}^{x,z}$ for which the following claim holds:

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

$\mathcal{X}$ is equal to $\mathbb{B}_R(\omega)$ in the notation from Section 4.1. The conditions for a point x to be a second-order critical point of a function f over $\mathbb{B}_R(\omega)$ can be expressed in the compact form:

$$\begin{cases} \nabla f(x) = 0, \\ \nabla^2 f(x) \succeq 0 \end{cases} \text{ if } x \notin \partial\mathbb{B}_R(\omega) \text{ or } \begin{cases} \exists\ \mu \leq 0 :\ \nabla f(x) = \mu(x-\omega), \\ P_{x-\omega}\nabla^2 f(x) P_{x-\omega}^\top \succeq 0 \end{cases} \text{ if } x \in \partial\mathbb{B}_R(\omega)$$

where $P_{x-\omega} \in \mathbb{R}^{(nr-1)\times nr}$ is the matrix of orthogonal projection onto the subspace orthogonal to $x-\omega$. With that in mind, we construct the two functions: $\delta(x,z)$ and $\partial\delta(x,z)$ via the following optimization procedures:

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

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

In each of the problems, the first two constraints represent the requirement that x is a second-order critical point of $f_{z,\mathcal{A}}$, the third constraint takes care of the KSP, and the last one is the RIP. It is straightforward to verify that $\min\{\delta, \partial\delta\}$ takes the value of the desired $\delta_{2r}^{x,z}$. Minimization of $\delta_{2r}^{x,z}$ over $\{x \in \mathcal{X}, z \in \mathcal{Z} : xx^\top \neq zz^\top\}$ gives $\delta_{2r}^\star$ such that (Problem$^{\text{KSP+RIP}}$) with δ_{2r} has an instance with a spurious second-order critical point if and only if $\delta_{2r} > \delta_{2r}^\star$. Suppose that we are 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,

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

This inequality shows that $\delta_{2r} \geq \underline{\delta^\star_{2r}}$ is a sufficient, and $\delta_{2r} \leq \overline{\delta^\star_{2r}}$ is a necessary condition for the absence of spurious second-order critical points in the instances of the problem (Problem$^{\text{KSP+RIP}}$). Now, it is desirable to show that $\min\{\mathbb{O}_P^{\partial\mathbb{B}}(x,z;\mathcal{T},\omega), \mathbb{O}_P(x,z;\mathcal{T})\}$ can serve as $\underline{\delta^{x,z}_{2r}}$, and $\min\{\mathbb{O}^{\partial\mathbb{B}}(x,z;\mathcal{T},\omega), \mathbb{O}(x,z;\mathcal{T})\}$ can serve as $\overline{\delta^{x,z}_{2r}}$.

Lemma 23. *The following statements hold all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$:*

$$\mathbb{O}_P(x,z) \leq \delta(x,z) \leq \mathbb{O}(x,z) \tag{4.16a}$$

$$\mathbb{O}_P^{\partial\mathbb{B}}(x,z) \leq \partial\delta(x,z) \leq \mathbb{O}^{\partial\mathbb{B}}(x,z) \tag{4.16b}$$

Proof. Here, we show only inequality (4.16a) since (4.16b) can be shown similarly. Notice that for $P = \text{orth}([x,z])$, the following sequence of inclusions holds:

$$\{PYP^T : Y \in \mathbb{S}^{rank([x\ z])}\} \subseteq \{X \in \mathbb{S}^n : \text{rank}\,(X) \leq 2r\} \subseteq \mathbb{S}^n. \tag{4.17}$$

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

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

The decomposition $\mathcal{H}^* = \mathcal{A}^{*T}\mathcal{A}^*$ exists because $\mathcal{H}^* \succeq 0$. If the inequality holds for all $X \in \mathbb{S}^n$, it must hold when $rank(X) \leq 2r$, as noticed in (4.17). 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 (4.17), the pair $(\mathcal{A}_*^T\mathcal{A}_*, \delta_*)$ is feasible for the problem defining $\mathbb{O}_P(x,z)$. This can be verified after rewriting the last constraint of the problem defining $\mathbb{O}_P$ in the form

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

for all $Y \in \mathbb{S}^{rank(x,z)}$. It is important to notice that the same argument works for an arbitrary choice of $P \in \mathbb{R}^{n\times d}$ with $d \leq 2r$. □

The above lemma completes the proof of Theorem 8. Theorem 7 follows by substituting 1 in the right-hand sides of (4.14) and (4.15). Notice that the linearity of the gradient and the Hessian with respect to the kernel operation matrix is the only property of the objective function that has been extensively used here. It can be exploited for generalization of the developed theory.

Proof of Proposition 4. We write the dual of the problem defining the function $\mathbb{O}$ as:

$$\underset{y,\lambda,U_1\succeq 0,U_2\succeq 0,V\succeq 0}{\text{maximize}} \quad \text{tr}[U_1 - U_2] \tag{4.18a}$$

$$\text{subject to} \quad \text{tr}[U_1 + U_2] = 1, \tag{4.18b}$$

$$\mathcal{L}_{x,z}^\top(y) - \mathcal{M}_{x,z}^\top(V) + \mathcal{T}^\top(\lambda) = U_1 - U_2 \tag{4.18c}$$

This problem is the exact reformulation of

$$\underset{y\in\mathbb{R}^{n\times r},V\succeq 0,\mu\in\mathbb{R}^t}{\text{maximize}} \frac{\sum_{i=1}^{d}(-\lambda_i(\mathcal{L}_{x,z}^{\top}(y)-\mathcal{M}_{x,z}^{\top}(V)+\mathcal{T}^{\top}(\mu)))_+}{\sum_{i=1}^{d}(+\lambda_i(\mathcal{L}_{x,z}^{\top}(y)-\mathcal{M}_{x,z}^{\top}(V)+\mathcal{T}^{\top}(\mu)))_+} \tag{4.19}$$

For details, please refer to Lemma 14 in [120]. Both primal and dual problems are bounded and the dual is strictly feasible. Recall vector e and matrix X from Section 4.4, where for all $u \in \mathbb{R}^{n\times r}$ it holds that

$$e = \text{vec}\,(xx^T - zz^T), \qquad\qquad X\text{vec}\,(u) = \text{vec}\,(xu^T + ux^T)$$

A strictly feasible point of (4.18) can be chosen as $y = 0$, $\lambda = 0$, $V = \varepsilon I$, $U_1 = \eta I - \varepsilon W$ and $U_2 = \eta I + \varepsilon W$, where $2\eta = n^{-2}$, $2W = r[\text{vec}\,(I)e^T + e\text{vec}\,(I)^T] - XX^T$, and ε is sufficiently small to ensure that both U_1 and U_2 are PSD. Consequently, Slater's condition and strong duality hold, and thus the optimal solution of (4.19) coincides with $\mathbb{O}(x, z)$.
If $\mathcal{T} = (\mathcal{S}, \mathcal{W})$, then $\mathcal{T}^{\top}(u, T) = \mathcal{W}^{\top}(u)+\mathcal{S}^{\top}(T)$. If $\mathcal{S}$ represents a sparsity pattern then there is a matrix S such that $\mathcal{S}(T) = \mathcal{S}^{\top}(T) = S \circ T$. Let S and S' be the matrix representations of $\mathcal{S}$ and $\mathcal{S}'$, respectively. $S' \subset S$ means that there exists S^{Δ} such that $S = S' \cup S^{\Delta}$ and therefore $S' = S + S^{\Delta}$. It is straightforward to verify that for every $R \in \mathbb{S}^{n^2}$ there exists $T \in \mathbb{S}^{n^2}$ such that $S \circ T + S^{\Delta} \circ R = S' \circ T$. The opposite is also true: for every $\mathbf{T} \in \mathbb{S}^{n^2}$ there exists $\mathbf{R} \in \mathbb{S}^{n^2}$ such that $S \circ T + S^{\Delta} \circ \mathbf{R} = S' \circ T$.
We introduce the short-hand notation $\mathcal{V}(y, V, u) = \mathcal{L}_{x,z}^{\top}(y) - \mathcal{M}_{x,z}^{\top}(V) + \mathcal{W}^{\top}(u)$. One can verify that the following expression holds:

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

This completes the proof. □

Part II

Learning to Resolve Complexity

Chapter 5

Model-Agnostic Meta Learning as a Path to Tractable Sub-problems

Let us consider a decision process formulated for a system through an optimization problem

$$\{\text{minimize}_{x\in\mathcal{X}(\tau)}\, f(x;\tau) \;:\; \tau\in\mathcal{T}\}$$

where $\mathcal{T}$ is the set of all possible states the system can be found at. Given this optimization problem, a task would be to come up with a tractable algorithm that solves the problem, or at least with a heuristic that solves the problem efficiently in the most important cases. One way to define the importance of cases is to introduce a probability distribution $\mathbb{P}_{\mathcal{T}}$ over $\mathcal{T}$, measuring the probability that the system would be found in the state with the parameters $\tau\in\mathcal{T}$. One idea that may help to automatize the process of designing this heuristic is the idea to teach a machine learning model to take in a dataset of a number of instances $\tau_i\sim\mathbb{P}_{\mathcal{T}}$ and output an optimization algorithm that solves a potentially hard optimization problem very efficiently. Since the search in the space of all of the algorithms is not a tractable problem, we assume that the model would output an algorithm $a:\mathcal{T}\to\mathbb{R}^n$ from a parametric family of efficient algorithms $\mathcal{A}$. For a given algorithm $a\in\mathcal{A}$, the set of parameters that defines the largest problem which can be solved with a is

$$\mathcal{T}^\star(a)=\text{Arg}\max_{S\subseteq\mathcal{T}}\left\{\mathbb{P}_{\mathcal{T}}[S] \;:\; a(\tau)\in\text{Arg}\min_{x\in\mathcal{X}(\tau)} f(x;\tau)\;\forall\tau\in S\right\}$$

The Bayesian optimal rule for this ML formulation would take in the pair $(\mathcal{T},\mathbb{P}_{\mathcal{T}})$ and return

$$a^\star_{glob}\in\text{Arg}\max_{a\in\mathcal{A}}\mathbb{P}_{\mathcal{T}}[\mathcal{T}^\star(a)]$$

which implies that it would need to recognize the largest subproblem of $\mathcal{T}$ that can be solved with an algorithm from $\mathcal{A}$.

The global optimization property over a sub-problem is a desirable feature of an algorithm a, especially for safety-critical systems. However, there are high-scale applications of data

analysis, where an efficient heuristic can be of a greater value than an exact algorithm. For designing a heuristic, an alternative Bayes optimal rule should be formulated:

$$a^\star \in \operatorname{Arg\,max}_{a \in \mathcal{A}} \int_{\tau \in \mathcal{T}} f(a(\tau), \tau) d\mathbb{P}_{\mathcal{T}}(\tau)$$

The corresponding optimization problem characterized with all couples $(\mathcal{A}, \mathcal{T})$ of interest is the problem we are considering in this chapter. A practical machine learning model, at the exploitation stage, would take in a dataset $\{\tau_i\}_{i \in \mathcal{D}}$ and return an estimate $\hat{a}$ of $a^\star$. One example of a successful application of this approach lay inside the field of Machine Learning itself and called Meta-Learning, with the Model-Agnostic Machine Learning algorithm being a prominent representative. Therefore, in this chapter, we study the non-convex optimization problem solved within the MAML framework to understand its tractability with respect to local search methods. We consider $\theta = (\mathcal{A}, \mathcal{T})$ to be a single instance of the MAML problem, where a task set $\mathcal{T}$ is a subset of instances of the Linear Quadratic Regulator problem due to MAML's typical area of application lying in Reinforcement Learning and Control. To align well with the previous research, we consider $\mathcal{A}$ that consists of single-step gradient descent routines with a fixed step size but different initialization points. Nevertheless, the findings reported here can be easily generalized to $\mathcal{T}$ consisting of instances with benign optimization landscapes and $\mathcal{A}$ made of multi-step local search methods with small enough step sizes.

This chapter shows that the MAML objective inherits the benign optimization landscape from the underlying tasks if their objectives are close pointwise. This desirable property fails to hold for the objectives that coincide up to a multiplicative constant. We proposed a modification of MAML that does not possess this drawback. We conclude that for the choice of $\mathcal{A}$ that is common in the machine learning practice and even for elementary practically important tasks $\mathcal{T}$, MAML is unlikely to obtain the Bayes optimal rule $a^\star$ presented above.

5.1 Introduction

Meta-learning, along with transfer learning, is a rapidly developing research area in machine learning which aims to design algorithms that gain computational advantages out of inherent similarities between optimization problem instances of learning, otherwise referred to as tasks. In this work, we study one of the most popular meta-learning algorithms, named *Model-Agnostic Meta-Learning (MAML)*, which has been developed by Finn, Abbeel, and Levine [36]. In the reinforcement learning domain, the algorithm is expected to rapidly adapt a pre-learned policy to a new task. However, measuring the quality of adaptation is vague and therefore the domain of application of meta-learning remains uncharted. In the core of meta-learning, there is an optimization problem that is concerned with the expected generalization performance averaged among considered tasks. In this chapter, we propose to measure the performance of MAML by the optimality gap of the corresponding optimization problem. We consider a space of tasks to be suitable for meta-learning if the algorithm converges to a global optimizer of the meta-learning objective or to a point with a similar value. This approach

enables distinguishing those meta-learning problems that are solvable by MAML from the problems that are not. Intuitively, a particular algorithm should perform satisfactorily on a meta-learning problems that consists of tasks united by a particular type of similarity. For the purpose of demonstration, we consider linear quadratic control problems, although our theory applies to a broad class of RL tasks with benign optimization landscape. The term benign landscape in this chapter is used to describe optimization problems which can be (approximately) solved with common tractable optimization techniques. It is formalized in Section 5.2. We aim to theoretically study the global convergence properties of the original MAML algorithm on sequential decision-making tasks. In short, our findings can be summarised as follows:

- Meta-Learning objective inherits a benign landscape from the objectives of the individual tasks if they are similar pointwise. As a result, the original MAML and other meta-learning algorithms that rely on local search are guaranteed to perform well on the corresponding problems.
- As a strongly negative result, those problems consisting of linear quadratic tasks that coincide up to a scaling of the reward function are not solvable by MAML. We propose an alternative scheme that addresses the issue with this type of similarity.

For the clarity of explanation, we investigate discrete stationary infinite-horizon decision problems and note that the generalization of the results to non-stationary and finite-horizon cases is straightforward. A stationary discrete dynamical system is described as

$$s_{t+1} \sim T(s_t, a_t)\,,$$

where T is a probability distribution over the next state $s_{t+1} \in \mathbb{R}^d$ given the current state $s_t \in \mathbb{R}^d$ and action $a_t \in \mathbb{R}^r$. The initial state s_0 is assumed to follow the distribution T_0. The objective of the infinite-horizon problem is to find a control input a_t minimizing the total discounted cost (the negation of the reward)

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{s_0 \sim T_0}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \\ \text{subject to} \quad & s_{t+1} \sim T(s_t, a_t) \qquad \forall\, t \in \{0, 1, \ldots\} \end{aligned}$$

where $\gamma \in (0, 1]$ is the discount factor that is assumed to be 1 for the LQR problem and $\mathbb{E}[\cdot]$ denotes the expectation operator.

Linear-Quadratic Regulator

One of the important examples of decision-making problems is related to the control of linear dynamical systems with a quadratic objective, referred to as *linear-quadratic regulator*

(LQR). LQR and Iterative LQR (iLQR) [100] are fundamental tools for model-based reinforcement learning. Moreover, linear-quadratic problems enjoy a rich theoretical foundation with a large number of provable guarantees, which make them a compelling benchmark for the mathematical analysis of novel learning algorithms.

We consider the following infinite-horizon exact LQR problem:

$$\begin{aligned} &\text{minimize} \quad \mathbb{E}_{x_0}\left[\sum_{t=0}^{\infty}(s_t^\top Q s_t + a_t^\top R a_t)\right] \\ &\text{subject to} \quad s_{t+1} = As_t + Ba_t, \quad \forall\, t \in \{0,1,\ldots\} \end{aligned} \tag{5.1}$$

where Q and R are positive semidefinite matrices. Assuming that the matrices A and B are such that the optimal cost is finite, it is a classic result that the optimal control policy is deterministic and linear in the state, i.e.,

$$a_t = -W^\star s_t$$

where $W^\star \in \mathbb{R}^{r\times d}$ [9]. Moreover, the matrix $W^\star$ can be found from the model parameters by solving the Algebraic Riccati Equation (ARE)

$$P = A^\top P A + Q - A^\top P B(B^\top P B + R)^{-1} B^\top P A\,,$$

and substituting the positive-definite root P into

$$W^* = (B^\top P B + R)^{-1} B^\top P A.$$

This implies that in order to find the solution of LQR, it suffices to only search over deterministic policies of the form $a = -Ws$ parameterized with a matrix $W \in \mathbb{R}^{r\times d}$.

In an effort to build a bridge between practical RL algorithms and the optimal control theory, [35] shows that W^* can be found by applying the policy gradient algorithm to a reformulated cost function. The LQR cost of a linear deterministic policy with respect to W can be defined as

$$C(W) := \mathbb{E}_{s_0\sim T_0}\left[\sum_{t=0}^{\infty}(s_t^\top Q s_t + a_t^\top R a_t)\right]$$

where $a_t = -Ws_t$ and $s_{t+1} = (A - BW)s_t$. This can be reformulated as

$$C(W) = \mathbb{E}_{s_0\sim T_0}\, s_0^\top P_W s_0$$

where P_W is the solution of $P_W = Q + W^\top R W + (A - BW)^\top P_W (A - BW)$. We only consider the cost of stable policies, and assume the cost to be infinite for unstable ones.

Model-Agnostic Meta-Learning

Given a set of tasks $\mathcal{T}$, each represented by an objective function $\mathcal{L}_\tau$, and a probability distribution $\mathbb{P}_\mathcal{T}$ over the tasks, [36] proposes an algorithm for finding an initialization of the policy gradient method that allows a fast adaptation to a task through just several gradient updates. In case the task consists in regression, classification, or clusterization, $\mathcal{L}_\tau$ is a risk or an empirical risk. In case of a reinforcement learning task, $\mathcal{L}_\tau$ is the cost (the negated return) of a policy.

If the space of considered policies is parameterized via $w \in \mathcal{W}$, then a single-shot MAML (which uses just one gradient update) aims to minimize the objective

$$L(w) = \mathbb{E}_{\tau \in \mathcal{T}} \left[f_\tau (w - \eta \nabla g_\tau (w)) \right] \tag{5.2}$$

where f_τ and g_τ can be two different approximations of the objective function of the task τ and ∇ is the gradient operator. Similarly, a multi-shot version applies multiple gradient updates within the adaptation procedure. For example, if $\mathcal{L}_\tau$ is the risk of a learning problem, then f_τ can be the empirical risk conditioned on a large dataset, while g_τ is the empirical risk conditioned on a smaller dataset. The generality of this formulation will be used in Section 5.2, but for the study of LQR we will assume that the functions $\mathcal{L}_\tau$, f_τ and g_τ all coincide and are equal to $C(W)$. When it is clear from the context which task is being discussed, we will omit the subscript.

Being based upon gradient descent with a constant step size, Algorithm 3 is a basic version of a few-shot MAML although other versions have been developed in the literature, e.g. FO-MAML, HF-MAML [32], iMAML [87], Reptile [80] and FTML [37]. We do not directly address the other formulations in the chapter, but the conclusions of this work are applicable to them as well since they are created to minimize essentially the same objective function (5.2) and the convergence to at least a first-order stationary point has been proven for the majority of these variants.

An interesting modification of MAML is based on exact proximal optimization as an alternative to a gradient update. [125] proposes the proximal update MAML algorithm with the adaptation procedure that finds the best set of parameters in the neighborhood of initialization, while [107] proves that a similar idea can be shaped into a meta-RL algorithm that is proven to converge globally under some overparametrization assumption. However, the exact optimization in the adaptation procedure may be problematic in meta-learning setup, since adaptation is supposed to be fast, meaning that there are sharp constraints on sample complexity and the amount of computation allowed for it.

As meta-learning seeks to improve learning performance by exploiting similarities between tasks, it is important to understand what type of similarities a particular meta-learning algorithm can take advantage of. A highly desirable feature of a meta-learning algorithm is an acceptable meta-test performance at least on the tasks it has been meta-trained on. In other words, if the set of tasks $\mathcal{T}$ is finite and the meta-training procedure has access to all of them, then it should succeed at global optimization on the meta-testing stage. For MAML, this translates into a requirement of successful minimization of the objective (5.2).

Algorithm 3 Model-Agnostic Meta-Learning (MAML)

Require: $p(\mathcal{T})$: Probabilistic task generator
Require: η, β: Step size hyperparameters
1: Randomly initialize w
2: **while** not done **do**
3: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
4: **for all** $\mathcal{T}_i$ **do**
5: Evaluate $\nabla g_{\mathcal{T}_i}(w)$
6: Compute adapted parameters with gradient descent: $w_i' = w - \eta \nabla_w g_{\mathcal{T}_i}(w)$
7: **end for**
8: Update $w \leftarrow w - \beta \nabla_w \sum_{\mathcal{T}_i \sim p(\mathcal{T})} f_{\mathcal{T}_i}(w_i')$
9: **end while**

[37] shows that the global minimum is achieved in case the functions $f_\tau = g_\tau$ are all smooth and strongly convex. However, the objective of many practical decision-making tasks are not convex, although may possess benign landscape, like the LQR objective (5.1). Tasks with non-convex objectives are substantially harder to analyze, and thus [32] only shows convergence of MAML to a first-order stationary point of (5.2) if f_τ are non-convex. We aim to study the global convergence properties of MAML applied to tasks with non-convex objectives.

The landscape of (5.1) has been studied in [35], which has observed that there exist instances of LQR that are not convex, quasi-convex, or star-convex, which means that none of the existing results on global convergence of MAML can be applied even to such a basic decision problem as LQR. However, (5.1) possess benign landscape, and our study shows that this property can be transferred to (5.2).

The purpose of this chapter is not to design a new algorithm or conduct a study of its application on a specific real-world case, but rather to prove the basic properties of a popular existing algorithm for a vast variety of cases. Meta-Learning algorithms are supposed to capture similarities between tasks, although there is no clear way to determine whether or not a similarity has been captured. For MAML, we propose a criterion that is based on the properties of the landscape of the optimization associated with MAML. Specifically, we declare that a version of MAML captures the similarities between the tasks in $\mathcal{T}$ if the objective of the algorithm on $\mathcal{T}$ has benign landscape. Otherwise, the tasks in $\mathcal{T}$ are recognized to be too distinct for this particular version of MAML.

The advantage of the above criterion is that it correlates with the computational complexity of the problem that MAML aims to solve. If the objective function has benign landscape, then the optimization problem has a low computational complexity and one can solve it by a local search method, which is implemented within MAML. If the objective does not have benign landscape, then MAML can become stuck in a spurious local minimum, which can potentially be arbitrarily worse than the optimal solution.

A drawback of this approach is that it does not allow to compare Meta-Learning algo-

rithms against each other, since it does not take into account the computational complexity of the adaptation algorithm. In this chapter, we do not consider this aspect because of our focus specifically on the few-shot MAML. We consider MAML primarily with applications to linear-quadratic systems because they are realistic and yet easy to analyze although our conclusions go far beyond this application.

5.2 Main Results

In this section, we study the MAML algorithm under four different scenarios. We consider MAML applied to a single task and to several identical tasks. Afterwards, we introduce a metric between tasks and extend the study to a number of close tasks, and, finally, we study the convergence of MAML on a large number of distant LQR tasks.

We provide theoretical results for general multi-dimensional systems, while all of the presented examples and counter-examples are on one-dimensional LQR tasks since they are easy to visualize. Note that these examples are extendable to multi-dimensional systems as well. The details on the exact tasks used for the computations are moved to the Appendix.

Single task

We begin by analyzing MAML applied to a singleton task set $\mathcal{T}$. If MAML fails under this scenario, then its global convergence for the multiple-task scenario becomes questionable. For the single task, we rewrite the MAML objective (5.2) as

$$h(w) = f(w - \eta \nabla g(w))$$

where f is assumed to be a continuously differentiable function and g is assumed to be twice continuously differentiable. As noticed in Section 5.1, the existing results on global convergence of MAML are not applicable to LQR. Figure 5.1 demonstrates an example of the MAML objective (5.2) applied to a single LQR task. It is non-convex and has three distinct strict local minimizers. Nevertheless, all of these three points are also global minimizers, which implies that Algorithm 3 would converge to its global minimizer from almost any initial point. The minimizer in the middle corresponds to $W^\star$ of the task, while the rightmost and leftmost minimizers are some points W such that $W - \eta \nabla C(W) = W^\star$ and $\nabla C(W) \neq 0$. The minimizers on both sides rely on the rapid adaptation during the meta-testing stage, which has been assumed in the design of the algorithm.

This example gives rise to a hypothesis that the MAML objective for a single LQR possesses some sort of benign landscape and, more generally, that benign landscape of the cost of the underlying task results in benign landscape of the resulting MAML objective. Following [53], we formalize the notion of benign landscape by defining global functions

Definition 13. *A continuous function $\ell : \mathcal{Z} \to \mathbb{R}$ is called global if every local minimizer of ℓ is a global minimizer.*

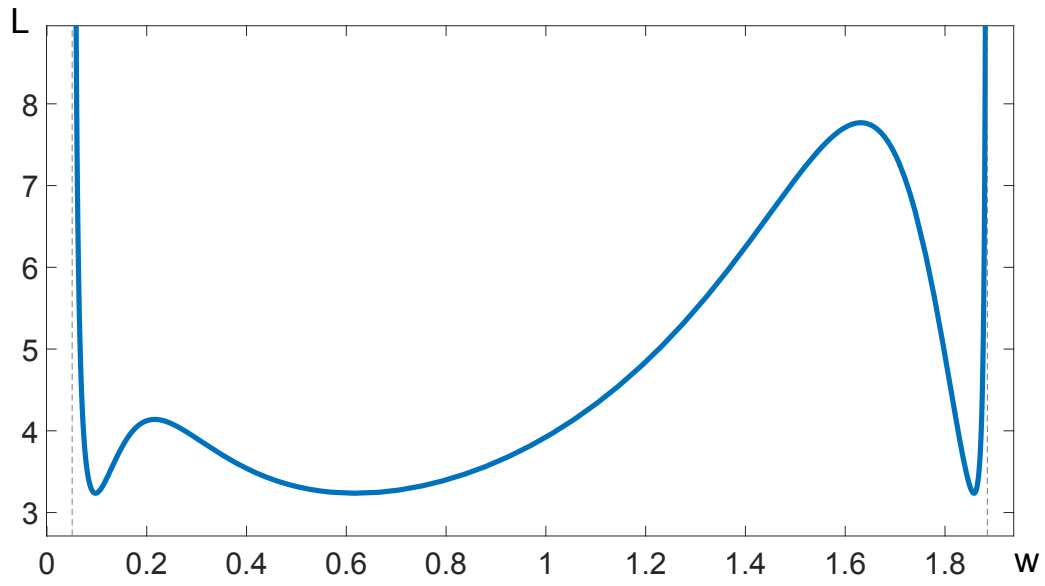


Figure 5.1: MAML objective (5.2) on a single LQR task.

This property is also referred as having no spurious local minima and generalizes the notions of convexity, quasiconvexity, and star-convexity. As a relaxation of this property, we define ε-global function:

Definition 14. *A continuous function $\ell : \mathcal{Z} \to \mathbb{R}$ is called ε-global if for every local minimizer $\bar{z}$ of ℓ it holds that*

$$\ell(\bar{z}) - min_{z \in \mathcal{Z}} \ell(z) \leq \varepsilon$$

Being global is equivalent to being 0-global. This property is more likely to be satisfied than a perfect no-spurious property for cost functions coming from real-world applications, providing with a broader way to formalize the intuitive meaning of a benign landscape. These two notions of benign landscape characterize when a coercive function is easy to optimize using local optimization methods based on gradient descent. The following theorem shows that the benign landscape of the cost of the underlying task indeed leads to a benign landscape for the resulting MAML objective.

Theorem 10. *Let $f : \mathcal{W} \to \mathbb{R}$ be global and $g : \mathcal{W} \to \mathbb{R}$ be twice continuously differentiable with $\|\nabla^2 g(w)\| \leq M < \infty$ for all $w \in \mathcal{W}$. If Algorithm 3 with the parameter η chosen to be smaller than $\frac{1}{M}$ converges to a local miminizer $w^\star \in \mathcal{W}$ of h, then $w^\star$ is a global minimizer of h.*

Theorem 10 is proven in the appendix, and its proof relies significantly on the following technical Lemma:

Lemma 24. *Let $\ell : \mathcal{Z} \to \mathbb{R}$ be an ε-global function, and consider a continuous map $\mathcal{F} : \mathcal{W} \to \mathcal{Z}$ with $\mathcal{Z} = range(\mathcal{F})$ that is locally open at a local minimizer $\bar{w}$ of $\ell \circ \mathcal{F}$. Then, it holds that*

$$\ell(\mathcal{F}(\bar{w})) - \min_{w \in \mathcal{W}} \ell(\mathcal{F}(w)) \leq \varepsilon$$

In the context of LQR, Theorem 10 results in the statement below.

Theorem 11. *Let $w^\star \in \mathbb{R}^{r \times d}$ be the limit point of the sequence generated by Algorithm 3 (MAML) with $\eta < \frac{1}{\|\nabla^2 C(w^\star)\|_2}$ applied to a single LQR task, meaning that $h(W) = C(W - \eta \nabla C(W))$. Then, $w^\star$ is the global minimizer of h, which implies that $C(W - \eta \nabla C(W))$ is a global function.*

Proof. It is shown in [32] that MAML converges to a first-order stationary point of a smooth nonconvex loss. Therefore, if the limit point $w^\star$ exists, it must be a stationary point of $C(W - \eta \nabla C(W))$. Consider the first-order stationarity condition for $w^\star$:

$$\begin{aligned} 0 =& \nabla C(w^\star - \eta \nabla C(w^\star)) = \\ =& [\mathcal{I} - \eta \nabla^2 C(w^\star)]^\top \nabla C(W)\big|_{W = w^\star - \eta \nabla C(w^\star)} \end{aligned}$$

Similar to the proof of Theorem 10, $\mathcal{I} - \eta \nabla^2 C(w^\star)$ is a full-rank marix, meaning that $w^\star$ is a first-order stationary point for the MAML objective if and only if $\nabla C(W)\big|_{W = w^\star - \eta \nabla C(w^\star)} = 0$.

Theorem 7 of [35] states that the Gradient descent algorithm finds an ε-approximation of the global optimum of $C(W)$ in polynomial time for any initial point with a finite value. This directly implies that all first-order stationary points of $C(W)$ are the global minimizers of $C(W)$ because otherwise we could initialize the Gradient descent algorithm at a stationary point and since it converges to the point of initialization, it will lead to a contradiction. Being a first-order stationary point is a sufficient condition of local minimality, and hence guaranties that $C(W)$ is a global function. By Theorem 10, the MAML objective is global for a sufficiently small η.

Since $w^\star - \eta \nabla C(w^\star)$ is a first-order stationary point of C, it is a global minimizer of C and since $\min_W C(W) \leq \min_W C(W - \eta \nabla C(W))$, the point $w^\star$ is the global minimizer of the MAML objective. □

Theorem 10 has implications far beyond the study of LQR. For example, Theorem 4.3 in [118] states that, under certain conditions, the objective of $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee is a global function. Applying Theorem 24 to this objective yields that MAML on the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ state-feedback control design will also have no spurious local minima under the corresponding conditions. It is also applicable in case f is not a reinforcement learning objective but an objective of a regression or a classification task. It also has a straightforward extension to the multi-shot MAML and other variants of the MAML algorithm. Theorem 10 can also be viewed as a practical guideline for proving global convergence post-factum. If the function f is global and one runs Algorithm 3 with a parameter η that converges to a point $w^\star$ such that $\nabla h(w^\star) = 0$, then one can check whether

$\nabla^2 h(w^\star) \succ 0$ and $\eta < \frac{1}{\|\nabla^2 g(w^\star)\|_2}$ and if so, $w^\star$ is guaranteed to be a global minimizer of h. In case f is not global but its landscape has benign properties, then the following generalization takes place:

Proposition 5. *If f is ε-global for some $\varepsilon > 0$ and $w^\star \in \mathcal{W}$ is a local minimum and $\eta < \frac{1}{\|\nabla^2 g(w^\star)\|_2}$, then*

$$h(\bar{w}) - \min_{w \in \mathcal{W}} h(w) \leq \varepsilon$$

Proof. The mapping $\mathcal{F}(w) = w - \eta \nabla g(w)$ is continuously differentiable over $\mathcal{W}$. The Jacobian of this mapping is $\nabla \mathcal{F}(w) = \mathcal{I} - \eta \nabla^2 g(w)$. By assumption, $\eta < \frac{1}{\|\nabla^2 g(w^\star)\|_2}$ and consequently there exists δ such that $\eta < \frac{1}{\|\nabla^2 g(w)\|_2}$ for all $w \in \bar{\mathbb{B}}_\delta(w^\star)$. Similar to the proof of Theorem 10, $\nabla \mathcal{F}(w)$ is positive definite for all $w \in \bar{\mathbb{B}}_\delta(w^\star)$, and therefore by Theorem 9.25 in [92], $\mathcal{F}\big|_{\bar{\mathbb{B}}_\delta(w^\star)}$ is an open mapping and hence locally open at $w^\star$. By Lemma 24, it means that $h(\bar{w}) - \min_{w \in \mathcal{W}} h(w) \leq \varepsilon$. □

So far, we have shown that the benign landscape properties of MAML applied to a singleton $\mathcal{T}$ are inherited from the benign landscape of the objective of the task to which MAML has been applied. In particular, this holds true for the LQR tasks.

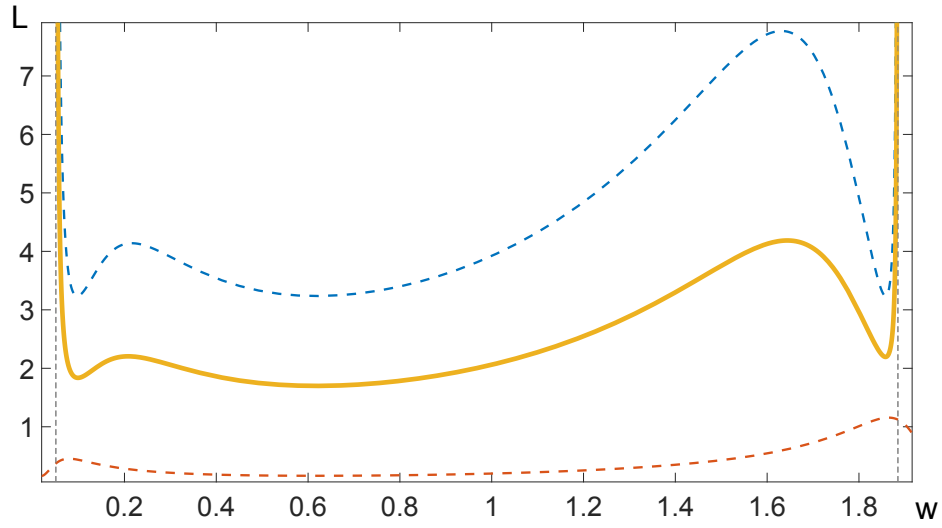
Several identical tasks

Results obtained for the single-task scenario give rise to a hypothesis that the benign landscape of every individual task would help with the convergence of MAML in a multi-task setting as well, provided that all of the tasks have a similar structure. Thus, in this part we study multiple-task learning for which the MAML has originally been designed for. Starting with some tasks that are the most similar to each other, we consider LQR tasks that coincide up to multiplication of the cost by a positive scalar. This is the highest degree of similarity one may hope to have between sequential decision-making problems. More precisely, the landscape features of the cost function (local and global minimizers, maximizers and saddle points) are preserved under this transformation, and therefore we refer to tasks of these types as identical henceforth.

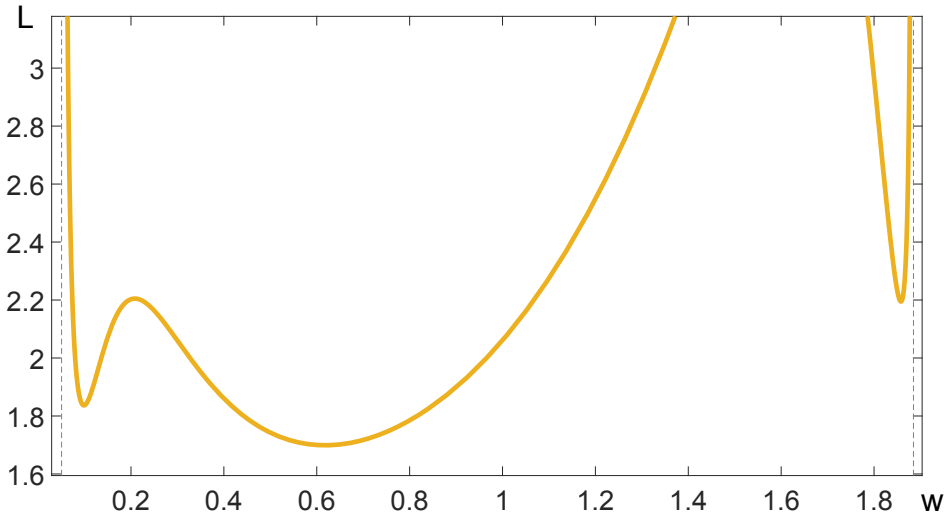
We consider a finite set of LQR tasks with a uniform distribution among them, which allows us to reduce the MAML objective (5.2) to the form

$$\frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} C_\tau(W - \eta \nabla C_\tau(W))$$

Figure 5.2 demonstrates an example of the MAML objective applied to two identical LQR tasks. Although the MAML objective of each individual task is global, only one global minimizer coincides among all of them. This is the minimizer that corresponds to $W^\star$. The two minimizers on the sides are shifted and after interference they produce spurious local minima for the total objective function. Hence, we can conclude that MAML fails to capture



(a)



(b)

Figure 5.2: Two MAML objective functions (5.2) for identical LQR tasks (dashed lines) and the MAML objective for the uniform distribution among them (solid line). 5.2b demonstrates spurious local minima of the solid line.

this type of similarity between the tasks. One practical lesson to learn from this figure is that keeping the cost (or the reward) function of the considered tasks normalized may improve the quality of the solution provided by MAML. Another lesson is that the design of meta-learning algorithms may benefit from considering the analysis of identical tasks as the simplest form of common structure in the tasks.

As an alternative, we propose a modification of MAML with normalized adaptation step. It turns out that this modification manages to capture the similarity between tasks with scaled rewards. The objective function for this version of MAML under single-task scenario is h' :

$$h'(w) = f\left(w - \eta \frac{\nabla g(w)}{\|\nabla g(w)\|}\right)$$

For which a statement similar to Proposition 5 holds.

Proposition 6. *If f is ε-global for some $\varepsilon > 0$ and $w^\star \in \mathcal{W}$ is a local minimizer of $h(w)$ with $\eta < \frac{\|\nabla g(w^\star)\|}{\|\nabla^2 g(w^\star)\|}$, then*

$$h'(\bar{w}) - \min_{w \in \mathcal{W}} h'(w) \leq \varepsilon$$

Proof. The mapping $\mathcal{F}(w) = w - \eta \frac{\nabla g(w)}{\|\nabla g(w)\|}$ is continuously differentiable over $\mathcal{W}$. The Jacobian of this mapping is

$$\nabla \mathcal{F}(w) = \mathcal{I} - \eta \left[\frac{\nabla^2 g}{\|\nabla g\|^3}\left(\|\nabla g\|^2 I - \nabla g \nabla g^\top\right)\right]$$

Observe that

$$\left\|\frac{\nabla^2 g}{\|\nabla g\|^3}\left(\|\nabla g\|^2 \mathcal{I} - \nabla g \nabla g^\top\right)\right\| \leq \frac{\|\nabla^2 g\|}{\|\nabla g\|^3}\left\|\|\nabla g\|^2 \mathcal{I} - \nabla g \nabla g^\top\right\|$$

Since $\nabla g \nabla g^\top$ is a rank-one matrix, the largest eigenvalue of $\|\nabla g\|^2 I - \nabla g \nabla g^\top$ must be equal to $\|\nabla g\|^2$ and therefore

$$\left\|\|\nabla g\|^2 \mathcal{I} - \nabla g \nabla g^\top\right\| = \|\nabla g\|^2$$

Hence, $\eta < \frac{\|\nabla g(w^\star)\|}{\|\nabla^2 g(w^\star)\|}$ is a sufficient to conclude that there exists δ such that $\nabla \mathcal{F}(w)$ is positive definite for all $w \in \bar{\mathbb{B}}_\delta(w^\star)$, and therefore by Theorem 9.25 in [92], $\mathcal{F}\big|_{\bar{\mathbb{B}}_\delta(w^\star)}$ is an open mapping and thus locally open at $w^\star$. By Lemma 24, it follows that $h'(w^\star) - \min_{w \in \mathcal{W}} h'(w) \leq \varepsilon$. □

In general, the objective of this modification of MAML can be written as

$$\mathbb{E}_\tau f\left(w - \eta \frac{\nabla g_\tau(w)}{\|\nabla g_\tau(w)\|}\right)$$

and its first-order stationary point can be found by utilizing the gradient descent with Armijo rule as noticed by [8] in Proposition 1.2.1. The claim that MAML with normalized gradient step inherits the benign landscape from the identical tasks can be formulated for LQR tasks as follows:

Theorem 12. *Let the normalized-gradient version of MAML be applied to k LQR tasks $\{(A_i, B_i, Q_i, R_i)\}_{i=1}^k$ with scaled dynamics and rewards, meaning that there exist $\alpha_1, \dots, \alpha_k > 0$ and $\beta_1, \dots, \beta_k > 0$ such that*

$$A_1 = \cdots = A_k;$$
$$B_1 = \cdots = B_k;$$
$$\alpha_1 Q_1 = \cdots = \alpha_k Q_k;$$
$$\alpha_1 R_1 = \cdots = \alpha_k R_k,$$

and $h(W) = \sum_{i=1}^k \omega_i C_i \left(W - \eta \frac{\nabla C_i(W)}{\|\nabla C_i(W)\|}\right)$. Let $w^\star \in \mathbb{R}^{r \times d}$ be a local minimal point of $h(W)$ with $\eta < \frac{\|\nabla C_i(w^\star)\|}{\|\nabla^2 C_i(w^\star)\|}$ for some $i \in \{1 \dots k\}$. Then, $w^\star$ is the global minimizer of h.

Proof. From the construction of the function C_i, it follows that for all $W \in \mathbb{R}^{r \times d}$ there exists $C(W)$ such that

$$C(W) = \frac{C_1(W)}{\alpha_1} = \cdots = \frac{C_k(W)}{\alpha_k}$$

Consequently, $\nabla C(W) = \frac{\nabla C_1(W)}{\alpha_1} = \cdots = \frac{\nabla C_k(W)}{\alpha_k}$ and $\nabla^2 C(W) = \frac{\nabla^2 C_1(W)}{\alpha_1} = \cdots = \frac{\nabla^2 C_k(W)}{\alpha_k}$ for all $i \in \{1 \dots k\}$. Hence,

$$h(W) = \left[\sum_{i=1}^k w_i \alpha_i\right] C\left(W - \eta \frac{\nabla C(W)}{\|\nabla C(W)\|}\right)$$

Since $w^\star$ is a local minimum, by Proposition 6, we conclude that $w^\star$ is also a global minimum. □

Several similar tasks

Moving forward, it is desirable to consider a different type of similarity between different tasks in the MAML setting. Therefore, we study the landscape of MAML on those tasks that are similar to each other in terms of the norm of the difference between parameters. For LQR, the parameters are the matrices A, B, Q and R. Our result states that the benign landscape of the underlying tasks carries over to the MAML objective if the tasks are sufficiently close to each other. For LQR, it will be informally stated below.

Proposition 7. *For every $\varepsilon > 0$ and $k \in \mathbb{N}$, there exists $\delta > 0$ such that the MAML objective (5.2) is ε-global for almost any set $\mathcal{T}$ of k LQR tasks defined through the parameters $\{(A_i, B_i, Q_i, R_i)\}_{i=1}^k$ such that for all $i, j \in \{1 \dots k\}$*

$$\|A_i - A_j\| + \|B_i - B_j\| + \|Q_i - Q_j\| + \|R_i - R_j\| \leq \delta$$

The formal statement along with the proof of the result and additional discussions are provided in the Appendix. Intuitively, as long as the dependence of the cost of a single task on the parameters is continuous, for a set of tasks that are close to each other, the multi-task landscape of MAML remains close to the landscape of MAML for just one of them. However, for this reasoning to hold true, we need to determine the continuity properties of a manifold of parametric ε-global functions. In order to do that, we make an assumption below.

Assumption 2. *Let $\ell : \mathcal{X} \times \mathbb{R}^m \to \mathbb{R}$ with a compact set $\mathcal{X} \subset \mathbb{R}^n$ be a twice continuously differentiable function with respect to $x \in \mathcal{X}$ with $\ell, \nabla_x \ell$ and $\nabla^2_{xx}\ell$ being continuous with respect to $t \in \mathbb{R}^m$. Assume that $\ell(\cdot, t)$ has a finite number of first-order stationary points for all $t \in \mathbb{R}^m$ and that the Hessian is non-singular ($det[\nabla^2_{xx}\ell(x,t)] \neq 0$) for all $t \in \mathbb{R}^m$ and all $x \in \mathcal{X}$ such that $\nabla_t \ell(x,t) = 0$.*

The following Theorem determines the continuity property of ε-globality over the manifold of parametric functions satisfying Assumption 2.

Theorem 13. *If for some $\bar{t}$ and $\varepsilon > 0$ the function $\ell(\cdot, \bar{t})$ satisfying Assumption 2 is ε-global, then for any $\varepsilon' > 0$ such that $\varepsilon' > \varepsilon$ there exists $\delta > 0$ for which the function $\ell(\cdot, t)$ is ε'-global for all $t \in \bar{\mathbb{B}}_\delta(\bar{t})$.*

Proof. We prove the theorem in three steps. Step 1: $\ell(\cdot, t)$ has stationary points in the neighoborhoods of the stationary points of $\ell(\cdot, \bar{t})$. Step 2: the types of the stationary points coincide. Step 3: There are no stationary points outside of the considered neighborhoods.

A finite number of stationary points means that all of them are isolated. Consider a stationary point $\bar{x}$ of $\ell(\cdot, \bar{t})$. By Theorem 9.28 of [92] (Generalized implicit function theorem), there exist $\psi > 0$ and $\phi > 0$ such that for all $t \in \bar{\mathbb{B}}_\psi(\bar{t})$ there exists a unique $x(t) \in \bar{\mathbb{B}}_\phi(\bar{x})$ with the property that $\nabla_x \ell(x(t), t) = 0$. This implies that for every function $\ell(\cdot, t)$ with $t \in \bar{\mathbb{B}}_\psi(\bar{t})$ there is a unique stationary point over $\bar{\mathbb{B}}_\phi(\bar{x})$.

Note that for any $v \in \mathbb{R}^n$ the function $h(x, t, v) = v^\top [\nabla^2_{xx}\ell(x,t)]v$ is continuous in both x and t. By the assumption of the theorem, $\det[\nabla^2_{xx}\ell(\bar{x}, \bar{t})] \neq 0$, and therefore $\nabla^2_{xx}\ell(\bar{x}, \bar{t}) \succ 0$ if $\bar{x}$ is a local minimum, $\nabla^2_{xx}\ell(\bar{x}, \bar{t}) \prec 0$ if it is a local maximum, and $\nabla^2_{xx}\ell(\bar{x}, \bar{t})$ indefinite if it is a saddle point. In each of these cases, we describe how to find values δ' and ϕ' such that ℓ has a bounded value of local minima over $\bar{\mathbb{B}}_{\phi'}(\bar{x}) \times \bar{\mathbb{B}}_{\delta'}(\bar{t})$.

<u>*Case 1:*</u> $\bar{x}$ is a local minimum. The value of $h(\bar{x}, \bar{t}, v)$ is positive for all $v \in \mathbb{R}^n \backslash \{0\}$. By continuity, there exist $\psi' > 0$ and $\phi' > 0$ such that $\psi' < \psi$ and $\phi' < \phi$ and $h(x, t, v) > 0$ for all $x \in \bar{\mathbb{B}}_{\phi'}(\bar{x})$, $t \in \bar{\mathbb{B}}_{\psi'}(\bar{t})$ and $v \in \mathbb{R}^n \backslash \{0\}$. This way, $\nabla^2_{xx}\ell(x(t), t) \succ 0$ and therefore $x(t)$ is a local minimum of $\ell(\cdot, t)$. By continuity of $\ell(x, t)$ with respect to x and t, there exists $\delta' > 0$ such that $\delta' < \psi'$ and $|\ell(x(t), t) - \ell(\bar{x}, \bar{t})| < \frac{\varepsilon' - \varepsilon}{2}$ for all $t \in \bar{\mathbb{B}}_{\delta'}(\bar{t})$.

<u>*Case 2:*</u> $\bar{x}$ is a local maximum. The value of $h(\bar{x}, \bar{t}, v)$ is negative for all $v \in \mathbb{R}^n \backslash \{0\}$. By continuity, there exist $\psi' > 0$ and $\phi' > 0$ such that $\psi' < \psi$ and $\phi' < \phi$ and $h(x, t, v) < 0$ for all $x \in \bar{\mathbb{B}}_{\phi'}(\bar{x})$; $t \in \bar{\mathbb{B}}_{\psi'}(\bar{t})$ and $v \in \mathbb{R}^n \backslash \{0\}$. This way, $\nabla^2_{xx}\ell(x(t), t) \prec 0$ and therefore $x(t)$ is a local maximum of $\ell(\cdot, t)$. In this case, we assign to the point $\bar{x}$ the value $\delta' = \psi'$.

Case 3: $\bar{x}$ is a saddle point. There exist $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^n$ such that $h(\bar{x}, \bar{t}, v) > 0$ and $h(\bar{x}, \bar{t}, u) < 0$. By continuity, there exist $\psi' > 0$ and $\phi' > 0$ such that $\psi' < \psi$ and $\phi' < \phi$ yet $h(x, t, v) > 0$ and $h(x, t, u) < 0$ for all $x \in \bar{\mathbb{B}}_{\phi'}(\bar{x})$, $t \in \bar{\mathbb{B}}_{\psi'}(\bar{t})$ and $v \in \mathbb{R}^n \backslash \{0\}$. This way, $\nabla^2_{xx}\ell(x(t), t)$ is indefinite and therefore $x(t)$ is a saddle point of $\ell(\cdot, t)$. In this case, we assign to the point $\bar{x}$ the value $\delta' = \psi'$.

As a result, having selected a single stationary point $\bar{x}$ of $\ell(\cdot, \bar{t})$, we can find δ' and ϕ' such that all the stationary points of $\ell(\cdot, t)$ for $t \in \bar{\mathbb{B}}_{\delta'}(\bar{t})$ are of the same type as $\bar{x}$, and in case they are local minimizers, they have a value that is not too different from $\ell(\bar{x}, \bar{t})$. One can repeat this argument for all the stationary points and therefore form sets of numbers $\{\delta'_i\}_{i=1}^N$ and $\{\phi'_i\}_{i=1}^N$, where i corresponds to the index of each of the N stationary points $\bar{x}_i$ of $\ell(\cdot, \bar{t})$.

Consider the set $\mathcal{Y} = \mathcal{X} \backslash \cup_{i \in \{1 \dots N\}} \bar{\mathbb{B}}_{\phi'_i}(\bar{x}_i)$. It is a compact set as a compact set minus an open set and $\|\nabla_x \ell(x, \bar{t})\| > 0$ for all $x \in \mathcal{Y}$. Since $\|\nabla_x \ell(x, \bar{t})\|$ is continuous in x over a compact set, it is uniformly continuous and it reaches its lower bound, meaning that there exists $\xi > 0$ such that $\|\nabla_x \ell(x, \bar{t})\| \geq \xi$ for all $x \in \mathcal{Y}$. By continuity of $\|\nabla_x \ell(x, t)\|$ with respect to t, there exists δ'' such that $\|\nabla_x \ell(x, t)\| > \frac{\xi}{2}$ over $\mathcal{Y} \times \bar{\mathbb{B}}_{\delta''}(\bar{t})$ and therefore there are no stationary points of ℓ over $\mathcal{Y} \times \bar{\mathbb{B}}_{\delta''}(\bar{t})$.

We select $\delta = \min[\{\delta'_i | i \in \{1 \dots N\}\} \cup \{\delta''\}]$ and observe that for all $t \in \bar{\mathbb{B}}_{\delta}(\bar{t})$ the only local minimizers of $\ell(\cdot, t)$ are those close to the local minimizers of $\ell(\cdot, \bar{t})$. In *Case 1*, the corresponding δ' was selected such that, given a local minimizer $x(t)$ of $\ell(\cdot, t)$ that is neighboring a local minimizer $\bar{x}$ of $\ell(\cdot, \bar{t})$ and a global minimizer $x'(t)$ of $\ell(\cdot, t)$ that is neighboring a local minimizer $\bar{x}'$ of $\ell(\cdot, \bar{t})$, it holds that

$$\begin{aligned}
\ell(x(t), t) - \min_x \ell(x, t) = \ell(x(t), t) - \ell(x'(t), t) = \\
\ell(x(t), t) - \ell(\bar{x}, \bar{t}) + \ell(\bar{x}, \bar{t}) - \ell(\bar{x}', \bar{t}) + \\
+ \ell(\bar{x}', \bar{t}) - \ell(x'(t), t) \leq \\
|\ell(x(t), t) - \ell(\bar{x}, \bar{t})| + |\ell(\bar{x}, \bar{t}) - \ell(\bar{x}', \bar{t})| + \\
+ |\ell(\bar{x}', \bar{t}) - \ell(x'(t), t)| < \\
\frac{\varepsilon' - \varepsilon}{2} + \varepsilon + \frac{\varepsilon' - \varepsilon}{2} = \varepsilon'
\end{aligned}$$

□

Proposition 7 goes beyond the uniform distribution and holds for any distribution on a finite number of tasks. An example of the MAML objective for five similar LQR tasks is demonstrated in Figure 5.3a. Similarly to the results for single-task scenario, the statements of this section are generalizable to other tasks with benign optimization landscape. In the end, we conclude that MAML is able to capture the common structure among different tasks given through similar values of the parameters.

A number of tasks with common dynamics

Figure 5.3 demonstrates the MAML objective for several LQR tasks that share the same dynamics but have different cost functions. Increasing the number of considered tasks improves the features of the landscape of the total MAML objective. In the eleven-task scenario, MAML learns the mean of the optimal policies $W^\star$ for the tasks, which implies that instead of learning to rapidly adapt, MAML learns the average policy by effectively finding the minimum of $\frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} C_\tau(W)$. However, for the two-task scenario, the global minimizer of the MAML objective appears to correspond to a policy W that has $\nabla C_\tau(W)$ far from zero for every considered τ, and thus the algorithm that converged to that point would learn to adapt to a task during the meta-testing phase.

We observe that the weighted sum of a large number of functions that have minimizers in a close proximity ends up being almost global with the minimizer in the same region. Therefore, in practice, increasing the number of similar tasks on meta-training stage may improve the properties of the landscape of the MAML objective and assure convergence to the global minimum.

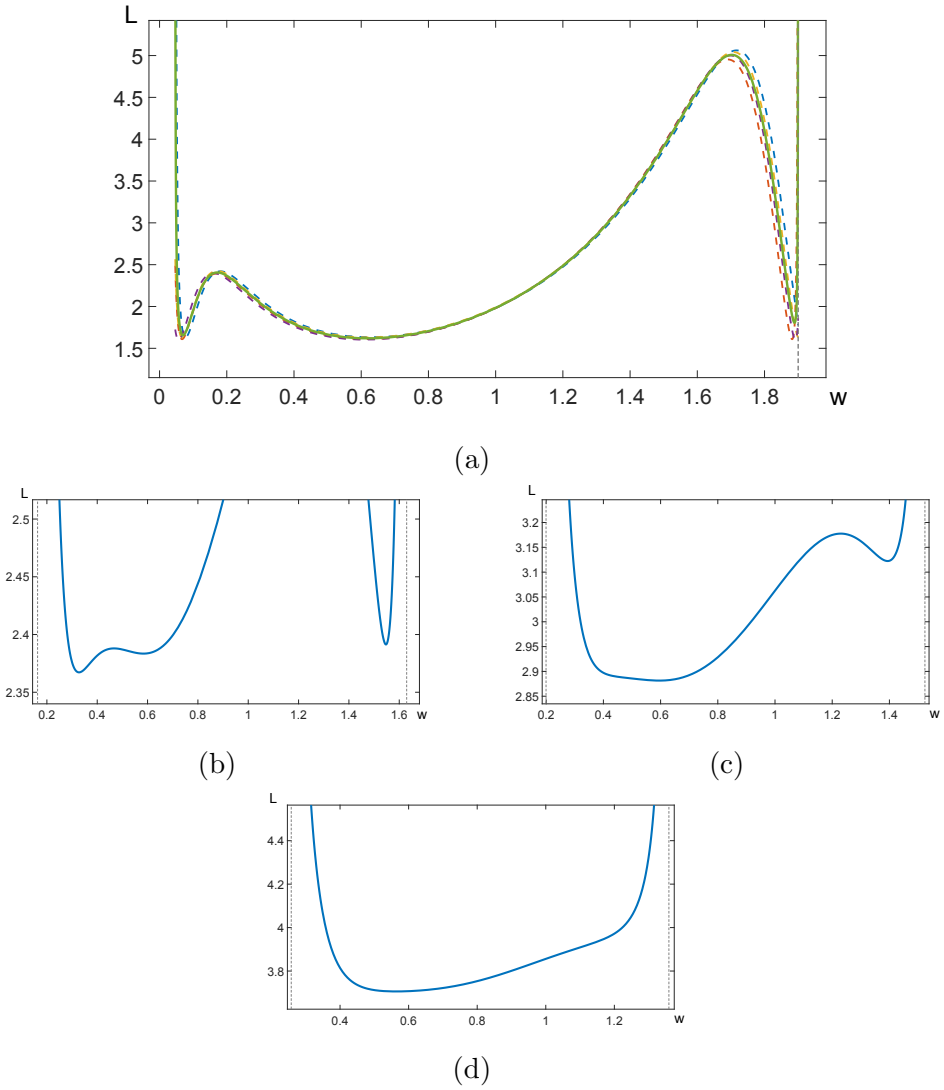


Figure 5.3: Five MAML objective functions (5.2) for LQR tasks with similar values of A, B, Q and R (dashed lines) and the MAML objective for the uniform distribution among them (solid line) **(plot 5.3a)**. MAML objective (5.2) for the uniform distribution among two **(plot 5.3b)**, five **(plot 5.3c)** and eleven **(plot 5.3d)** different LQR tasks that share the same dynamics A and B but have different cost matrices Q and R.

Appendix

Lemma 25. *Given $\lambda_i > 0$ and $\sum_{i=1}^{k} \lambda_i = 1$, if for some $\bar{t}$ and $\varepsilon > 0$ the function $\ell(\cdot, \bar{t})$ satisfying Assumption 2 is ε-global, then for any $\varepsilon' > 0$ such that $\varepsilon' > \varepsilon$ there exists $\delta > 0$ for which the convex combination $\lambda_1 \ell(\cdot, t_1) + \cdots + \lambda_k \ell(\cdot, t_k)$ is ε'-global for all $t_1, \ldots, t_k \in \bar{\mathbb{B}}_\delta(\bar{t})$.*

Proof. We provide the proof for $k = 2$, but the argument holds true for other finite values of k. By Theorem 13, $\ell(\cdot, t_1)$ can be assumed to be ε''-global for $\varepsilon' > \varepsilon'' > \varepsilon$. Therefore, without loss of generality, we assume that $t_1 = \bar{t}$, and then aim to prove that for a given $\lambda \in [0, 1]$ there exists $\delta > 0$ such that $\lambda \ell(x, \bar{t}) + (1 - \lambda)\ell(x, t)$ is ε'-global. We introduce $\mathfrak{r}(x, t) = \lambda \ell(x, \bar{t}) + (1 - \lambda)\ell(x, t)$ and note that $\mathfrak{r}(x, \bar{t}) = \ell(x, \bar{t})$, which means that $\mathfrak{r}(\cdot, \bar{t})$ is ε-global and satisfies Assumption 2. Theorem 13 applied to $\mathfrak{r}$ yields that there exists $\delta > 0$ such that $\lambda \ell(x, \bar{t}) + (1 - \lambda)\ell(x, t)$ is ε'-global. □

Proposition 8. *Given $\lambda_i > 0$ and $\sum_{i=1}^{k} \lambda_i = 1$, if for all the values of t in a compact set $\mathcal{C} \subset \mathbb{R}^m$ and some $\varepsilon > 0$ the function $\ell(\cdot, t)$ satisfying Assumption 2 (restricted from $\mathbb{R}^m$ to $\mathcal{C}$) is ε-global, then for any $\varepsilon' > 0$ such that $\varepsilon' > \varepsilon$ there exists $\delta > 0$ for which any convex combination $\lambda_1 \ell(\cdot, t_1) + \cdots + \lambda_m \ell(\cdot, t_m)$ is ε'-global for all $t_1, \ldots, t_m$ such that $\|t_i - t_j\| < \delta$.*

Proof. To prove by contradiction, suppose that there exists $\varepsilon' > 0$ such that $\varepsilon' > \varepsilon$ and for all $\delta > 0$ there are $t_1(\delta), \ldots, t_k(\delta)$ such that $\|t_i - t_j\| < \delta$ and $\lambda_1 \ell(\cdot, t_1) + \ldots + \lambda_m \ell(\cdot, t_m)$ is not ε'-global. From the sequence $t_1(\frac{1}{l})$, one can extract a converging sub-sequence $t_1(\delta_l)$ since $\mathcal{C}$ is compact. By Lemma 25, for the point $\bar{t} = \lim_{l \to \infty} t_1(\delta_l)$ there exists $\delta' > 0$ such that for all $t_1, \ldots, t_k \in \bar{\mathbb{B}}_{\delta'}(\bar{t})$ the convex combination $\lambda_1 \ell(\cdot, t_1) + \ldots + \lambda_k \ell(\cdot, t_k)$ is ε'-global. Therefore, for the $t_1(\delta_l), \ldots, t_k(\delta_l)$ such that $\delta_l < \frac{\delta'}{2k}$ and $\|t_1(\delta_l) - \bar{t}\| < \frac{\delta'}{2k}$, this convex combination is ε'-global, which is a contradiction. □

Proposition 9 (Formal statement of Proposition 7)**.** *Consider the instances of LQR such that their parameters A, B, Q and R belong to a compact set. Assume that none of them produces a cost function $C(W)$ with a singular Hessian at a stationary point. For any $\varepsilon > 0$ and $k \in \mathbb{N}$, there exists $\delta > 0$ such that for any set of k considered instances of LQR defined through the parameters $\{(A_i, B_i, Q_i, R_i)\}_{i=1}^{k}$ with*

$$\|A_i - A_j\| + \|B_i - B_j\| + \|Q_i - Q_j\| + \|R_i - R_j\| \leq \delta,$$

for all $i, j \in \{1 \ldots m\}$, the MAML objective (5.2) is ε-global.

Proof. Take $t = (A, B, Q, R)$. By Theorem 10, every function $C(W, t)$ satisfies Assumption 2. Therefore, $C(W, t)$ satisfies all of the assumptions of Proposition 8 and the proof follows immediately. □

Given a matrix $\bar{W}$, the system of equations

$$\nabla C(\bar{W}) = 0; \quad \det\left(\nabla^2 C(\bar{W})\right) = 0 \tag{5.3}$$

Figure	LQR $(A, B, Q, R, s_0 s_0^\top)$
Figure 5.1	$(1, 1, 2, 2, 1)$, $\eta = 0.1$
Figure 5.2	$(1, 1, 2, 2, 1), (1, 1, 0.1, 0.1, 1)$, $\eta = 0.1$
Figure 5.3a	$(1.01, 1, 1, 1, 1)$, $(1, 1.01, 1, 1, 1)$, $(1, 1, 1.01, 1, 1)$, $(0.99, 1, 1, 1, 1)$, $\eta = 0.1$
Figure 5.3b	$(1, 1, 1, 2, 1)$, $(1, 1, 2, 1, 1)$, $\eta = 0.1$
Figure 5.3c	$(1, 1, 1, 1, 1)$, $(1, 1, 1, 2, 1)$, $(1, 1, 2, 1, 1)$, $(1, 1, 2, 3, 1)$, $(1, 1, 3, 2, 1)$, $\eta = 0.1$
Figure 5.3d	$(1, 1, 1, 1, 1)$, $(1, 1, 1, 2, 1)$, $(1, 1, 2, 1, 1)$, $(1, 1, 2, 3, 1)$, $(1, 1, 3, 2, 1)$, $(1, 1, 3, 1, 1)$, $(1, 1, 1, 3, 1)$, $(1, 1, 4, 1, 1)$, $(1, 1, 1, 4, 1)$, $(1, 1, 5, 3, 1)$, $(1, 1, 3, 5, 1)$, $\eta = 0.1$

Table 5.1: To simplify the visualization, all of the examples and counterexamples in the chapter were given for one-dimensional LQR systems. As a result, the parameters A, B, Q and R were scalar values, and so were the state and the action. For reproducibility, this table collects the parameters used to construct each of the examples.

with respect to the parameters A, B, Q and R defines a low-dimensional manifold in the space of LQR tasks denoted by $HS(\bar{W})$. All the LQR systems that have at least one stationary point with a singular Hessian are contained in $\cup_W HS(W)$. Now, notice that the union occurs over an $n \times n$ dimensional space, while $HS(W)$ is defined with a system of $(n \times n)+1$ equations. This implies that $\cup_W HS(W)$ is a low-dimensional manifold and thus almost no LQR has a stationary point with a singular Hessian. Thus, the compact domain that is mentioned in the assumption of Proposition 9 can be a large closed set with $\cup_W HS(W)$ excluded together with its small neighborhood. Thus, Proposition 7 is an informal restatement of Proposition 9.

Details on the experiments

For the numerical experiments, we computed the MAML objective explicitly using the formulas provided in the Section on LQR. The initial state s_0 was chosen to be deterministic. Further details are provided in Table 5.1.

Bibliography

[1] Ali Abur and Antonio Gomez Exposito. *Power system state estimation: theory and implementation*. CRC press, 2004.

[2] Alekh Agarwal et al. "Learning sparsely used overcomplete dictionaries via alternating minimization". In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2775–2799.

[3] Albert Akhriev, Jakub Marecek, and Andrea Simonetto. "Pursuit of low-rank models of time-varying matrices robust to sparse and measurement noise". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 3171–3178.

[4] Farid Alizadeh and Donald Goldfarb. "Second-order cone programming". In: *Mathematical programming* 95.1 (2003), pp. 3–51.

[5] Martin S Andersen, Anders Hansson, and Lieven Vandenberghe. "Reduced-complexity semidefinite relaxations of optimal power flow problems". In: *IEEE Transactions on Power Systems* 29.4 (2014), pp. 1855–1863.

[6] Morteza Ashraphijuo, Ramtin Madani, and Javad Lavaei. "Characterization of rank-constrained feasibility problems via a finite number of convex programs". In: *IEEE 55th Conference on Decision and Control (CDC)*. IEEE. 2016, pp. 6544–6550.

[7] Morteza Ashraphijuo, Ramtin Madani, and Javad Lavaei. "Inverse function theorem for polynomial equations using semidefinite programming". In: *IEEE 54th Annual Conference on Decision and Control (CDC)*. IEEE. 2015, pp. 6589–6596.

[8] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, MA, 1999.

[9] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2017.

[10] Kush Bhatia, Prateek Jain, and Purushottam Kar. "Robust regression via hard thresholding". In: *Advances in Neural Information Processing Systems*. 2015, pp. 721–729.

[11] Kush Bhatia et al. "Consistent Robust Regression". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2107–2116.

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

[13] Léon Bottou and Olivier Bousquet. "The tradeoffs of large scale learning". In: *Advances in neural information processing systems*. 2008, pp. 161–168.

[14] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[15] T Tony Cai and Anru Zhang. "Compressed sensing and affine rank minimization under restricted isometry". In: *IEEE Transactions on Signal Processing* 61.13 (2013), pp. 3279–3290.

[16] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. "Phase retrieval via Wirtinger flow: Theory and algorithms". In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1985–2007.

[17] Emmanuel J Candes and Terence Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215.

[18] Emmanuel J Candes et al. "Phase retrieval via matrix completion". In: *SIAM review* 57.2 (2015), pp. 225–251.

[19] Emmanuel J Candès and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Foundations of Computational mathematics* 9.6 (2009), p. 717.

[20] Emmanuel J Candès et al. "Robust principal component analysis?" In: *Journal of the ACM (JACM)* 58.3 (2011), p. 11.

[21] Jinghui Chen et al. "Robust Wirtinger flow for phase retrieval with arbitrary corruption". In: *arXiv preprint arXiv:1704.06256* (2017).

[22] Yudong Chen, Constantine Caramanis, and Shie Mannor. "Robust sparse regression under adversarial corruption". In: *International Conference on Machine Learning*. 2013, pp. 774–782.

[23] Yudong Chen et al. "Low-rank matrix recovery from errors and erasures". In: *IEEE Transactions on Information Theory* 59.7 (2013), pp. 4324–4337.

[24] Yuxin Chen et al. "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval". In: *Mathematical Programming* (2018), pp. 1–33.

[25] Yuxin Chen et al. "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization". In: *SIAM journal on optimization* 30.4 (2020), pp. 3098–3121.

[26] Yuejie Chi, Yue M Lu, and Yuxin Chen. "Nonconvex optimization meets low-rank matrix factorization: An overview". In: *IEEE Transactions on Signal Processing* 67.20 (2019), pp. 5239–5269.

[27] Eldan Cohen and J Christopher Beck. "Problem Difficulty and the Phase Transition in Heuristic Search." In: *AAAI*. 2017, pp. 780–786.

[28] Angang Cui, Jigen Peng, and Haiyang Li. "Exact recovery low-rank matrix via transformed affine matrix rank minimization". In: *Neurocomputing* 319 (2018), pp. 1–12.

[29] Arnak Dalalyan and Yin Chen. "Fused sparsity and robust estimation for linear models with unknown variance". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1259–1267.

[30] Hadi Daneshmand et al. "Escaping saddles with stochastic gradients". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1155–1164.

[31] Deepjyoti Deka, Ross Baldick, and Sriram Vishwanath. "Optimal data attacks on power grids: Leveraging detection & measurement jamming". In: *International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2015, pp. 392–397.

[32] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. "On the convergence theory of gradient-based model-agnostic meta-learning algorithms". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1082–1092.

[33] Salar Fattahi and Somayeh Sojoudi. "Data-Driven Sparse System Identification". In: *56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2018.

[34] Maryam Fazel. "Matrix rank minimization with applications". PhD thesis. Stanford University, 2002.

[35] Maryam Fazel et al. "Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

[36] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1126–1135.

[37] Chelsea Finn et al. "Online meta-learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1920–1930.

[38] Stéphane Fliscounakis et al. "Contingency ranking with respect to overloads in very large power systems taking into account uncertainty, preventive, and corrective actions". In: *IEEE Transactions on Power Systems* 28.4 (2013), pp. 4909–4917.

[39] Peter I Frazier. "A tutorial on bayesian optimization". In: *arXiv preprint arXiv:1807.02811* (2018).

[40] Mituhiro Fukuda et al. "Exploiting sparsity in semidefinite programming via matrix completion I: General framework". In: *SIAM Journal on Optimization* 11.3 (2001), pp. 647–674.

[41] Rong Ge, Chi Jin, and Yi 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.

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

[43] Rong Ge et al. "Escaping from saddle points—online stochastic gradient for tensor decomposition". In: *Conference on learning theory*. PMLR. 2015, pp. 797–842.

[44] Arpita Ghosh, Stephen Boyd, and Amin Saberi. "Minimizing effective resistance of a graph". In: *SIAM review* 50.1 (2008), pp. 37–66.

[45] Michael Grant and Stephen Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. Mar. 2014.

[46] Michael Grant and Stephen Boyd. "Graph implementations for nonsmooth convex programs". In: *Recent Advances in Learning and Control*. Ed. by V. Blondel, S. Boyd, and H. Kimura. Lecture Notes in Control and Information Sciences. Springer-Verlag Limited, 2008, pp. 95–110.

[47] Nima Hamidi and Mohsen Bayati. "On low-rank trace regression under general sampling distribution". In: *arXiv preprint arXiv:1904.08576* (2019).

[48] Paul Hand and Vladislav Voroninski. "Corruption robust phase retrieval via linear programming". In: *arXiv preprint arXiv:1612.03547* (2016).

[49] Roger Horn. *Matrix analysis*. New York: Cambridge University Press, 2013.

[50] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. "Guaranteed rank minimization via singular value projection". In: *Advances in Neural Information Processing Systems*. 2010, pp. 937–945.

[51] Prateek Jain, Praneeth Netrapalli, and Sujay 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.

[52] Ming Jin, Javad Lavaei, and Karl Johansson. "A semidefinite programming relaxation under false data injection attacks against power grid AC state estimation". In: *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2017, pp. 236–243.

[53] Cedric Josz et al. "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization". In: *Advances in neural information processing systems*. 2018, pp. 2441–2449.

[54] Cédric Josz et al. "AC power flow data in MATPOWER and QCQP format: iTesla, RTE snapshots, and PEGASE". In: *arXiv preprint arXiv:1603.01533* (2016).

[55] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. "Matrix completion from a few entries". In: *IEEE transactions on information theory* 56.6 (2010), pp. 2980–2998.

[56] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. "Matrix completion from noisy entries". In: *Journal of Machine Learning Research* 11.Jul (2010), pp. 2057–2078.

[57] Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. "Robust matrix completion". In: *Probability Theory and Related Fields* 169.1-2 (2017), pp. 523–564.

[58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[59] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. "Low rank matrix recovery from rank one measurements". In: *Applied and Computational Harmonic Analysis* 42.1 (2017), pp. 88–116.

[60] Jason D Lee et al. "Gradient descent only converges to minimizers". In: *Conference on learning theory*. 2016, pp. 1246–1257.

[61] Dawei Li, Tian Ding, and Ruoyu Sun. "Over-Parameterized Deep Neural Networks Have No Strict Local Minima For Any Continuous Activations". In: *arXiv preprint arXiv:1812.11039* (2018).

[62] Qiuwei Li, Zhihui Zhu, and Gongguo Tang. "Alternating minimizations converge to second-order optimal solutions". In: *International Conference on Machine Learning*. 2019, pp. 3935–3943.

[63] Xingguo Li et al. "Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization". In: *IEEE Transactions on Information Theory* (2019).

[64] Ramtin Madani, Morteza Ashraphijuo, and Javad Lavaei. "Promises of conic relaxation for contingency-constrained optimal power flow problem". In: *IEEE Transactions on Power Systems* 31.2 (2016), pp. 1297–1307.

[65] Ramtin Madani, Abdulrahman Kalbat, and Javad Lavaei. "A Low-Complexity Parallelizable Numerical Algorithm for Sparse Semidefinite Programming". In: *IEEE Transactions on Control of Network Systems* (2017).

[66] Ramtin Madani, Javad Lavaei, and Ross Baldick. "Convexification of power flow equations in the presence of noisy measurements". In: *IEEE Transactions on Automatic Control* 64.8 (2019), pp. 3101–3116.

[67] Ramtin Madani, Somayeh Sojoudi, and Javad Lavaei. "Convex relaxation for optimal power flow problem: Mesh networks". In: *IEEE Transactions on Power Systems* 30.1 (2014), pp. 199–211.

[68] Ramtin Madani et al. "Finding low-rank solutions of sparse linear matrix inequalities using convex optimization". In: *SIAM Journal on Optimization* 27.2 (2017), pp. 725–758.

[69] Ramtin Madani et al. "Power system state estimation and bad data detection by means of conic relaxation". In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2017.

[70] E Marianna, Monique Laurent, Antonios Varvitsiotis, et al. "Complexity of the positive semidefinite matrix completion problem with a rank constraint". In: *Discrete Geometry and Optimization*. Springer, 2013, pp. 105–120.

[71] Brian McWilliams et al. "Fast and robust least squares estimation in corrupted linear models". In: *Advances in Neural Information Processing Systems*. 2014, pp. 415–423.

[72] Song Mei, Yu Bai, and Andrea Montanari. "The landscape of empirical risk for nonconvex losses". In: *The Annals of Statistics* 46.6A (2018), pp. 2747–2774.

[73] Hyde M Merrill and Fred C Schweppe. "Bad data suppression in power system static state estimation". In: *IEEE Transactions on Power Apparatus and Systems* 6 (1971), pp. 2718–2725.

[74] Karthik Mohan and Maryam Fazel. "Iterative reweighted algorithms for matrix rank minimization". In: *Journal of Machine Learning Research* 13.Nov (2012), pp. 3441–3473.

[75] Matthew W Morency and Sergiy A Vorobyov. "An Algebraic Approach to a Class of Rank-Constrained Semi-Definite Programs With Applications". In: *arXiv preprint arXiv:1610.02181* (2016).

[76] Kazuhide Nakata et al. "Exploiting sparsity in semidefinite programming via matrix completion II: Implementation and numerical results". In: *Mathematical Programming* 95.2 (2003), pp. 303–327.

[77] Nasser M Nasrabadi, Trac D Tran, and Nam Nguyen. "Robust lasso with missing and grossly corrupted observations". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1881–1889.

[78] Balas Kausik Natarajan. "Sparse approximate solutions to linear systems". In: *SIAM journal on computing* 24.2 (1995), pp. 227–234.

[79] Nam H Nguyen and Trac D Tran. "Exact Recoverability From Dense Corrupted Observations via L_1-Minimization". In: *IEEE transactions on information theory* 59.4 (2013), pp. 2017–2035.

[80] Alex Nichol, Joshua Achiam, and John Schulman. "On first-order meta-learning algorithms". In: *arXiv preprint arXiv:1803.02999* (2018).

[81] Maher Nouiehed and Meisam Razaviyayn. "Learning deep models: Critical points and local openness". In: *INFORMS Journal on Optimization* (2021).

[82] Alain Pajor. "Metric entropy of the Grassmann manifold". In: *Convex Geometric Analysis* 34 (1998), pp. 181–188.

[83] Panos M. Pardalos and Stephen A. Vavasis. "Quadratic programming with one negative eigenvalue is NP-hard". In: *Journal of Global Optimization* 1.1 (Mar. 1991), pp. 15–22.

[84] Dohyung Park et al. "Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 65–74.

[85] Santiago Paternain, Aryan Mokhtari, and Alejandro Ribeiro. "A newton-based method for nonconvex optimization with fast evasion of saddle points". In: *SIAM Journal on Optimization* 29.1 (2019), pp. 343–368.

[86] Frank Permenter and Pablo Parrilo. "Partial facial reduction: simplified, equivalent SDPs via approximations of the PSD cone". In: *Mathematical Programming* (2014), pp. 1–54.

[87] Aravind Rajeswaran et al. "Meta-learning with implicit gradients". In: *Advances in Neural Information Processing Systems*. 2019, pp. 113–124.

[88] Pradeep Ravikumar et al. "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence". In: *Electronic Journal of Statistics* 5 (2011), pp. 935–980.

[89] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[90] Benjamin Recht, Weiyu Xu, and Babak Hassibi. "Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization". In: *IEEE 47th Conference on Decision and Control (CDC)*. 2008, pp. 3065–3070.

[91] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. Vol. 589. John wiley & sons, 2005.

[92] Walter Rudin. *Principles of mathematical analysis (third edition)*. McGraw-Hill Inc., 1976.

[93] Somayeh Sojoudi et al. "Graph-theoretic algorithms for polynomial optimization problems". In: *IEEE 53rd Conference on Decision and Control*. IEEE. 2014, pp. 2257–2271.

[94] Mahdi Soltanolkotabi. "Learning relus via gradient descent". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2007–2017.

[95] Christoph Studer et al. "Recovery of sparsely corrupted signals". In: *IEEE Transactions on Information Theory* 58.5 (2012), pp. 3115–3130.

[96] Ju Sun, Qing Qu, and John Wright. "A geometric analysis of phase retrieval". In: *Foundations of Computational Mathematics* 18.5 (2018), pp. 1131–1198.

[97] Ju Sun, Qing Qu, and John Wright. "Complete dictionary recovery using nonconvex optimization". In: *International Conference on Machine Learning*. 2015, pp. 2351–2360.

[98] Ruoyu Sun and Zhi-Quan Luo. "Guaranteed matrix completion via non-convex factorization". In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 6535–6579.

[99] Stanislaw J Szarek. "Nets of Grassmann manifold and orthogonal group". In: *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*. Vol. 169. 1982, p. 185.

[100] Emanuel Todorov and Weiwei Li. "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems". In: *American Control Conference*. IEEE. 2005, pp. 300–306.

[101] Kim-Chuan Toh, Michael J Todd, and Reha H Tütüncü. "SDPT3—a MATLAB software package for semidefinite programming, version 1.3". In: *Optimization methods and software* 11.1-4 (1999), pp. 545–581.

[102] Reha H Tütüncü, Kim-Chuan Toh, and Michael J Todd. "Solving semidefinite-quadratic-linear programs using SDPT3". In: *Mathematical programming* 95.2 (2003), pp. 189–217.

[103] Namrata Vaswani, Seyedehsara Nayer, and Yonina C Eldar. "Low-rank phase retrieval". In: *IEEE Transactions on Signal Processing* 65.15 (2017), pp. 4059–4074.

[104] Jan Ámos Višek. "The least trimmed squares. Part I: Consistency". In: *Kybernetika* 42.1 (2006), pp. 1–36.

[105] Martin J Wainwright. "Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE transactions on information theory* 55.5 (2009), pp. 2183–2202.

[106] Jun-Kun Wang and Shou-De Lin. "Robust Inverse Covariance Estimation under Noisy Measurements". In: *International Conference on Machine Learning*. 2014, pp. 928–936.

[107] Lingxiao Wang et al. "On the global optimality of model-agnostic meta-learning". In: *International conference on machine learning*. PMLR. 2020, pp. 9837–9846.

[108] Ke Wei et al. "Guarantees of Riemannian optimization for low rank matrix recovery". In: *SIAM Journal on Matrix Analysis and Applications* 37.3 (2016), pp. 1198–1222.

[109] Yang Weng et al. "Convexification of bad data and topology error detection and identification problems in AC electric power systems". In: *IET Generation, Transmission & Distribution* 9.16 (2015), pp. 2760–2767.

[110] John Wright and Yi Ma. "Dense Error Correction Via L_1-Minimization". In: *IEEE Transactions on Information Theory* 56.7 (2010), pp. 3540–3560.

[111] Bo Xin and David Wipf. "Pushing the limits of affine rank minimization by adapting probabilistic PCA". In: *International Conference on Machine Learning*. 2015, pp. 419–427.

[112] Chen Xu, Zhouchen Lin, and Hongbin Zha. "A Unified Convex Surrogate for the Schatten-p Norm." In: *AAAI*. 2017, pp. 926–932.

[113] Huan Xu, Constantine Caramanis, and Shie Mannor. "Robust regression and lasso". In: *Advances in Neural Information Processing Systems*. 2009, pp. 1801–1808.

[114] Wenzhuo Yang and Huan Xu. "A unified robust regression model for Lasso-like algorithms". In: *International Conference on Machine Learning*. 2013, pp. 585–593.

[115] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. "Global Optimality Conditions for Deep Neural Networks". In: *International Conference on Learning Representations*. 2018.

[116] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. "Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow". In: *International conference on machine learning*. 2016, pp. 1022–1031.

[117] Jie Zhang and Lijun Zhang. "Efficient Stochastic Optimization for Low-Rank Distance Metric Learning." In: *AAAI*. 2017, pp. 933–940.

[118] Kaiqing Zhang, Bin Hu, and Tamer Basar. "Policy optimization for H_2 linear control with H_∞ robustness guarantee: Implicit regularization and global convergence". In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 179–190.

[119] Richard Zhang et al. "How much restricted isometry is needed in nonconvex matrix recovery?" In: *Advances in neural information processing systems*. 2018, pp. 5591–5602.

[120] Richard Y. Zhang, Somayeh Sojoudi, and Javad Lavaei. "Sharp Restricted Isometry Bounds for the Inexistence of Spurious Local Minima in Nonconvex Matrix Recovery". In: *Journal of Machine Learning Research* 20.114 (2019), pp. 1–34.

[121] Yu Zhang, Ramtin Madani, and Javad Lavaei. "Conic relaxations for power system state estimation with line measurements". In: *IEEE Transactions on Control of Network Systems* 5.3 (2018), pp. 1193–1205.

[122] Yu Zhang, Ramtin Madani, and Javad Lavaei. "Conic relaxations for power system state estimation with line measurements". In: *IEEE Transactions on Control of Network Systems* 5.3 (2018), pp. 1193–1205.

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

[124] Qinqing Zheng and John 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.

[125] Pan Zhou et al. "Efficient meta learning via minibatch proximal update". In: *Advances in Neural Information Processing Systems*. 2019, pp. 1534–1544.

[126] Zhihui Zhu et al. "Global optimality in low-rank matrix optimization". In: *IEEE Transactions on Signal Processing* 66.13 (2018), pp. 3614–3628.

[127] Ray D. Zimmerman, Carlos E. Murillo-Sánchez, and Robert J. Thomas. "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education". In: *IEEE Transactions on power systems* 26.1 (2011), pp. 12–19.