Invariant Theory in Computational Complexity and Algebraic Statistics

vorgelegt von
M. Sc.
Philipp Reichenbach
ORCID: 0000-0002-5722-5505

von der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
Dr. rer. nat.

vorgelegte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Boris Springborn
Gutachter:	Prof. Dr. Peter Bürgisser
Gutachter:	Prof. Dr. Jan Draisma

Tag der wissenschaftlichen Aussprache: 08.06.2023

Berlin 2023

"Mit dem Wissen wächst der Zweifel"

Johann Wolfgang von Goethe

Preface

Abstract

Invariant theory is a branch of algebra that is classically intertwined with computation, but it also led to important contributions in mathematics and applications in other sciences. For example, the problem of deciding *null cone membership* (NCM) has, thanks to the general abstract setting of invariant theory, manifold applications in mathematics, physics, computer science and statistics.

In this thesis we study invariant theory in two regards: computational complexity and algebraic statistics. As indicated by its title the thesis consists of three parts.

The first part collects preliminary knowledge on *invariant theory* that is used throughout the thesis. Thus, it only contains known results and concepts.

The second part focuses on the *computational complexity* of current geodesic convex optimization methods for the NCM problem and its "approximate" versions: norm minimization and scaling. First, we prove that complexity parameters, that capture the *required precision* for deciding NCM via optimization methods, are exponentially small for several important group actions. Second, in the high precision regime the *diameter* (*"bit complexity"*) of approximate minimizers may be exponentially large for tensor scaling. The provided bounds exclude, for the respective group actions, polynomial running time of current geodesic methods for the three computational problems. Therefore, our results highly motivate the search for and the advancement of new sophisticated methods for geodesic convex optimization.

The third part builds a bridge between invariant theory and *algebraic statistics*, which establishes novel relations to maximum likelihood estimation (ML estimation). We connect norm minimization under a group action to maximizing the likelihood in a statistical model, which is related to the group action. In particular, norm minimizers yield maximum likelihood estimates. Strikingly, this approach yields a dictionary between stability notions from invariant theory and notions from ML estimation. We obtain fruitful applications on the interplay of invariant theory and statistics. First, we recover known statistical results, and even get some new characterizations, via invariant theory. Second, we obtain algorithmic consequences, e.g., complexity results from invariant theory carry over to statistics. Third, one can translate problems from statistics to invariant theory, and vice versa. This has already been used in the literature with great benefit. Fourth, the invariant theoretic approach fostered the development of new statistical models and the understanding of their ML estimation. Namely, we study the new concepts of *Gaussian group models* and of *RDAG models*.

Zusammenfassung

Invariantentheorie ist ein Bereich der Algebra, der klassischerweise eng verbunden ist mit Berechnungen und Algorithmen, der jedoch auch zu wichtigen Beiträgen in der Mathematik selbst sowie ihren Anwendungsbereichen geführt hat. Zum Beispiel hat das Entscheidungsproblem der *Nullkegel-Zugehörigkeit* (NKZ), dank des allgemeinen, abstrakten Settings der Invariantentheorie, vielfältige Anwendungen in Mathematik, Physik, Informatik und Statistik.

In dieser Dissertation studieren wir Invariantentheorie in zweierlei Hinsicht: bezüglich Komplexitätstheorie und bezüglich Algebraischer Statistik. Wie der Titel bereits andeutet besteht die Arbeit aus drei Teilen.

Der erste Teil umfasst und wiederholt wesentliche Vorkenntnisse aus der *Invariantentheorie*. Deshalb enthält er nur bereits bekannte Resultate und Konzepte.

Der zweite Teil beschäftigt sich mit der *Komplexität* von aktuellen geodätisch-konvexen Optimierungsalgorithmen für das NKZ-Problem sowie seiner „approximativen“ Versionen: Norm-Minimierung und Skalierung. Erstens zeigen wir, dass Komplexitätsparameter, welche die *nötige Präzision* zum Entscheiden des NKZ-Problems mittels Optimierungsalgorithmen erfassen, exponentiell klein sind für mehrere wichtige Gruppenaktionen. Zweitens, im Fall von hoher Präzision kann der *Durchmesser* (*„Bit-Komplexität“*) eines approximativen Minimierers exponentiell groß sein für Tensor-Skalierung. Diese bereitgestellten Schranken schließen, für die jeweiligen Gruppenaktionen, eine polynomiale Laufzeit der aktuellen geodätisch-konvexen Methoden für die drei genannten Probleme aus. Deshalb motivieren die Resultate in hohem Maße die Suche nach und die Weiterentwicklung von ausgeklügelten Methoden für geodätisch-konvexe Optimierung.

Der dritte Teil baut eine Brücke zwischen Invariantentheorie und *Algebraischer Statistik*, welche völlig neuartige Verbindungen zur Maximum-Likelihood-Methode (ML Methode) etabliert. Wir verbinden Norm-Minimierung unter einer Gruppenaktion zum Maximierungsproblem der Plausibilität in einem statistischen Modell, welches in Beziehung zur Gruppenaktion steht. Insbesondere, geben Norm-Minimierer einen zugehörigen Maximum-Likelihood-Schätzer. Bemerkenswerterweise führt dieses Vorgehen zu einem Wörterbuch zwischen Stabilitätsnotationen aus der Invariantentheorie und Notationen bezüglich der ML Methode. Dieses Zusammenspiel von Invariantentheorie und Statistik trägt große Früchte. Erstens erhalten wir bereits bekannte statistische Resultate, und manchmal sogar ganz neue Charakterisierungen, mittels Invariantentheorie. Zweitens gibt es algorithmische Folgerungen, zum Beispiel können komplexitätstheoretische Resultate von Invariantentheorie auf die Statistik übertragen werden. Drittens kann man Probleme der Statistik zu Problemen in Invariantentheorie übersetzen, und vice versa. Dies wurde bereits mit großem Erfolg in anderen wissenschaftlichen Arbeiten verwendet. Viertens hat der Zugang mittels Invariantentheorie die Entwicklung von neuen statistischen Modellen sowie das Verständnis ihrer ML-Methode gefördert. Nämlich studieren wir die neuartigen Konzepte der *Gaußschen Gruppenmodelle* sowie der *RDAG Modelle*.

Acknowledgments

First of all, I would like to express my gratitude towards my advisor Peter Bürgisser. I heartily thank him for the opportunity to join his ERC project, which allowed me to pursue research in one of my favourite mathematical areas. I am especially grateful for his advanced lectures, which were very valuable for my research, and for his constant support and feedback during the intense period of writing the thesis.

My research and PhD position were generously financed by the European Research Council (ERC) under the European Horizon 2020 research and innovation programme (grant agreement 787840). In addition, the available travel money gave me the great chance to attend many conferences and workshops (despite the pandemic in between). Therefore, I am deeply thankful to the ERC.

I am very grateful to my collaborators Carlos Améndola, Gergely Bérczi, Cole Franks, Eloise Hamilton, Kathlén Kohn, Visu Makam, Giorgio Ottaviani and Anna Seigal. During our joint projects I learned a lot from you about mathematics and the academical world, but certainly also beyond that.

The Thematic Einstein Semester (TES) on Algebraic Geometry in the winter term 2019/20 had a huge impact on my research. I thank the organizers Peter Bürgisser, Gavril Farkas and Christian Haase, and all organizers of the special events in the TES. I learned a huge amount of mathematics, in particular, during the fall school and while writing the lecture notes by Giorgio Ottaviani into LaTeX. My research on algebraic statistics was initiated by a talk of Mathias Drton in the TES opening conference. I am very thankful to Bernd Sturmfels and Peter Bürgisser for their encouragement to reach out to Mathias Drton and my future co-authors Anna, Carlos and Kathlén. I am grateful to Mathias Drton for hosting my co-authors and myself in Munich twice, for his hospitality and for the fruitful discussions we had during these days.

I thank Jan Draisma for his agreement and time on reviewing this thesis, and I thank Boris Springborn for his quick response and agreement on serving as the head of my doctoral committee.

For our daily work life, the interesting reading groups and colloquia, and our meetings outside the mathematical context I would like to thank all people from the algebra team(s) at TU Berlin: Matías Bender, Paul Breiding, Dominic Bunnett, Peter Bürgisser, Philipp di Dio, M. Levent Doğan, Alperen Ergür, Gorav Jindal, Mario Kummer, Dirk Kussin, Jonathan Leake, Marco Ramponi, Büşra Sert and Josué Tonelli-Cueto. Special thanks to our secretary Beate Niessen, who always helped with the bureaucracy at TU Berlin.

I thank Harold Nieuwboer and Michael Walter for many interesting discussions, especially during the research visits in Berlin and Bochum. Thank you Anna for the hospitality and the great opportunity to visit you at Harvard.

During the four projects, on which this thesis is based on, my co-authors and/or myself had fruitful discussions with many colleagues: thank you to Jason Altschuler, Dominic Bunnett, Peter Bürgisser, Jan Draisma, Mathias Drton, Robin Evans, Visu Makam, Nikolay Nikolov, Panagiotis Papazoglou, Adam Sawicki, Bernd Sturmfels, Caroline Uhler, Michael Walter and Piotr Zwiernik. I

thank the anonymous referees of these works for useful hints and suggestions.

I thank the Berlin Mathematical School for creating a convenient environment for PhD student in mathematics. I enjoyed the MATH+ Fridays and the BMS student conferences including the social highlight: the BMS Wine and Cheese Evening. I am deeply thankful for the annual BMS Career Events, which are extremely valuable for my career-wise orientation. Thank you to my BMS mentor Jörg Liesen for your constant support and feedback, both on my progress and the decisions I made.

Thank you to my bouldering group and the bouldering gyms in Berlin: this awesome sport and the great atmosphere always made my mind rest. Thanks to the great coffee roasteries in Berlin: you kept my mathematical engine running.

Thank you to all my friends. I am happy about the new friendships I made during bouldering and the board game sessions in Berlin. I am thankful for ongoing friendships from my Bonn times including Julian, Max, Melissa, Richard and Valentin. Certainly, I am very grateful to my close long-term friends Franzi and Sam from my Bachelor studies, and to Eugen, Konsti and Paul from school.

Finally, I am deeply grateful to my whole family for all their love, particularly to my parents and my brother for always having my back. I would like to dedicate this thesis to my whole family. This especially includes my grandparents who attach great importance to this thesis.

Funding

This PhD thesis and its results are part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 787840).

Authorship

I, Philipp Reichenbach, comment in the following on authorship for each chapter. Let me start with some general remarks. The thesis is based on the three publications [AKRS21a; AKRS21b; FR21] and the preprint [MRS21]. In each work, all authors contributed equally. It should be noted that I have rewritten some parts of these works and/or reorganized them, and I added further examples for illustration. This especially applies to Chapters 8, 9 and 10. The thesis connects several mathematical areas and I tried to make it accessible to a wide audience. This results in four preliminary chapters (Chapters 1, 2, 3 and 6), which only contain known results.

The Introduction and its German version ("Deutsche Einleitung") are written by myself. The dictionary (1) is taken from [AKRS21a, Introduction].

Chapter 1 is written by myself. All results are known in the literature and corresponding references are given.

Chapter 2 is partly written by myself, but each section is also based on work with my co-authors and literature as follows. The two proofs in Section 2.1 are taken from [AKRS21b, Appendix A]. The presentation of Section 2.2 mainly follows [BFG+19] and [FR21], but sometimes also [AKRS21a, Section 2.2] and [AKRS21b, Section 3.C]. Section 2.3 is completely based on [AKRS21a, Appendix], while Section 2.4 stems from [MRS21, Appendix B]. All results are known in the literature and corresponding references are given.

Chapter 3 is written by myself, but it is certainly influenced by existing literature on the topic, especially [BFG+19; GO18] and the joint work [FR21] with Cole Franks. All results are known in the literature and corresponding references are given.

Chapters 4 and 5 are almost completely based on [FR21], which is joint work with Cole Franks. Both authors contributed equally.

Chapter 4 presents the bounds from [FR21] on *weight margin* and *gap*. All main proof ideas of these bounds are due to myself. The concept of freeness from Section 4.3 is well-known in the literature and we give corresponding references. I stress that Subsection 4.7.2 presents Theorem 4.7.6, which is due to Cole Franks and Visu Makam and answers [FR21, Problem 4.27] in the affirmative. I thank them for the permission to include their arguments, which are not published and fit nicely into Section 4.7.

Chapter 5 states and discusses the bounds from [FR21] on the *diameter*. All main proof ideas of these bounds are due to my co-author Cole Franks. The exposition is restricted to the main results, their implications and relations to the literature, and a proof outline.

Chapter 6 is partly written by myself. Section 6.2 follows [AKRS21b, Section 2], while Section 6.3 is in parts based on [AKRS21a, Section 2.1] and the extended example on DAG models is based on [MRS21]. All results, with the

possible exception of the formulation of Theorem 6.3.16 ([MRS21, Theorem 4.9]), are known in the literature and references are given.

Chapter 7 is completely based on [AKRS21b], which is joint work with Carlos Améndola, Kathlén Kohn and Anna Seigal. All authors contributed equally.

Chapter 8 is mostly based on [MRS21, Appendix A], which is joint work with Visu Makam and Anna Seigal. All authors contributed equally. Propositions 8.1.2 and 8.1.5 originate from discussions with Visu Makam and Anna Seigal, but do not appear in [MRS21].

Chapter 9 is mainly based on [AKRS21a], which is joint work with Carlos Améndola, Kathlén Kohn and Anna Seigal. All authors contributed equally. It should be noted that this chapter has many parts that have been rewritten, examples were added, and some implicit statements of [AKRS21a] are made explicit. Propositions 9.2.3 and 9.2.4 originate from discussions with my collaborators Gergely Bérczi, Eloise Hamilton, Visu Makam and Anna Seigal. Theorem 9.3.1 has been added and is known in the literature; I thank Harold Nieuwboer for the reference. Section 9.5 is based on [AKRS21a, Section 5], but it also takes knowledge from [MRS21] into account and therefore contains further results. Subsection 9.6.1 is written by myself, while Subsection 9.6.2 is based on [AKRS21b, Section 6].

Chapter 10 presents the preprint [MRS21], which is joint work with Visu Makam and Anna Seigal. All authors contributed equally. The simulations and their implementation in Section 10.5 were completely done by Anna Seigal. It should be noted that [MRS21] usually works over the real numbers, while Chapter 10 works over real and complex numbers in parallel. This contribution by myself explicates [MRS21, Remark 2.11], which states that one may also work over the complex numbers.

Contents

Introduction

"Das Instrument, welches die Vermittlung bewirkt zwischen Theorie und Praxis, zwischen Denken und Beobachten, ist die Mathematik; sie baut die verbindende Brücke und gestaltet sie immer tragfähiger."

David Hilbert[1]

Groups are amongst the most fundamental, organizing objects of mathematics, and appear all over the sciences. From a geometrical perspective, groups provide a framework to encode symmetries and they are often studied themselves via actions on spaces. Invariant theory studies actions of algebraic groups on algebraic varieties, and functions on the variety that remain invariant under this action. It is a branch of algebra that is classically intertwined with computation, but also led to great contributions in mathematics and to applications beyond.

A prime example are Hilbert's landmark papers [Hil90; Hil93] on classical invariant theory. There he proved seminal results of modern algebra and algebraic geometry; most prominently, his Basis Theorem and the Nullstellensatz. Interestingly, the actual objective of Hilbert's papers was to prove a finiteness theorem on the ring of invariants, and to provide an algorithm for computing a generating system. For this, Hilbert introduced in [Hil93] an invariant-theoretic key concept called the *null cone*. It consists of all *unstable* vectors, that is, vectors that cannot be distinguished from zero by invariants. Unstable vectors and further notions of stability play an important role in Geometric Invariant Theory [MFK94] for constructing and studying quotient spaces. Strikingly, there are also many applications beyond algebra itself as we outline below.

In recent years the null cone enjoyed a computational revival. The problem of deciding null cone membership (NCM, see Problem 3.1.2) has been intensively studied, leading to polynomial time algorithms in several important cases. There are algebraic/symbolic methods for deciding NCM, as well as optimization algorithms through "approximate" formulations of NCM: the Norm Minimization Problem 3.1.3 and the Scaling Problem 3.1.4. Thanks to the general abstract setting of invariant theory, these three problems have manifold applications in mathematics, physics, computer science and statistics; thereby connecting seemingly unrelated (computational) problems. This unified framework and its applications serve as a starting point and motivation of this thesis.

The objectives of this thesis are twofold. On one side, we study the *computational complexity* of geodesic convex optimization methods for solving the above three computational problems. On the other hand, we *build a bridge* between invariant theory and *algebraic statistics*, which establishes novel relations to maximum likelihood estimation.

[1] in "Naturerkennen und Logik" (speech from 8th September 1930), see [Hil35, p. 385]

Regarding computational complexity, a prominent example of NCM arises for the so-called operator scaling action[2], where a product of special linear groups acts on (tuples of) matrices. The NCM problem for this action relates to non-rational identity testing, a non-commutative analogue of the famous polynomial identity testing problem. Remarkably, the approach through the NCM problem leads to, both algebraic [DM17; IQS18] and numeric [AGL+18; GGOW16], deterministic polynomial time algorithms for non-rational identity testing![3] However, neither of the current methods is known to run in polynomial time for tensor scaling, the higher dimensional analogue where one acts on (tuples of) tensors. In fact, the main results in Part II prove that this is another example of the "unwritten law" that tensors are (computationally) "more challenging" than matrices.

More precisely, we present the results of [FR21] which give exponentially bad behaved bounds for complexity parameters of current geodesic convex optimization methods [BLNW20; AGL+18; AMS08; Bou23]. First, a parameter capturing the *required precision* for deciding NCM via optimization methods is shown to be exponentially small for tensor scaling. Second, in the high-precision regime the *diameter*, which can be interpreted as the bit-complexity of an approximate minimizer, may be exponentially large for tensor scaling. In contrast, these complexity parameters are known to be only polynomially small (precision) respectively polynomially large (diameter) for operator scaling. Altogether, these bounds exclude polynomial time algorithms for NCM, Norm Minimization and the Scaling Problem via current geodesic methods.

It should be noted that the latter are geodesic analogues of first and second order methods. However, in general, first and second order methods do not even suffice for commutative groups, where the computational problems are convex in the usual sense. Instead, ellipsoid or interior-point algorithms are required to ensure polynomial time [SV14; SV19; BLNW20]. We point out that the very recent works [Hir22; NW23] rigorously study self-concordant functions on manifolds and [NW23] even gives (the main stage of) an interior point method. However, applying this algorithm to the Scaling problem still yields a complexity that depends *linearly* on a diameter bound [NW23, Theorem 1.7]. Hence, the exponential diameter for tensor scaling excludes polynomial running time, making further research necessary [NW23, Outlook]. Altogether, our results highly motivated and keep motivating the search for and the advancement of new sophisticated methods in the regime of geodesic convex optimization.

The part on algebraic statistics focuses on novel relations between invariant theory and maximum likelihood estimation (ML estimation), established in [AKRS21a] and further studies in [AKRS21b; MRS21]. In particular, we add ML estimation to the list of applications of the above computational problems. ML estimation is a common approach to parameter estimation. That is, given a statistical model and some data, one seeks a probability distribution in the model that "best" fits the data. ML estimation chooses a distribution under which it

[2]also called left-right action

[3]In contrast, it remains a major open problem to solve polynomial identity testing in deterministic polynomial time.

is *most likely* to observe the given data. Hereby, "most likely" is encoded by maximizing a likelihood function, and a maximizer of that function is called a maximum likelihood estimate (MLE). Important questions arising in ML estimation are, for example: when does an MLE exist (uniquely)? How often do we have to sample data for almost sure existence of an MLE? How can we compute an MLE?

In this thesis we tackle these questions through invariant theory. This is achieved by providing a dictionary between stability notions under a group action and ML estimation on a corresponding model. For example, certain torus actions relate to *log-linear models*, while the operator and the tensor scaling action correspond to so-called *matrix* and *tensor normal models*, respectively. We always link several notions as in

$$\left\{\begin{array}{c}\text{unstable}\\ \text{semistable}\\ \text{polystable}\\ \text{stable}\end{array}\right\} \quad \longleftrightarrow \quad \left\{\begin{array}{c}\text{likelihood unbounded from above}\\ \text{likelihood bounded from above}\\ \text{MLE exists}\\ \text{MLE exists uniquely}\end{array}\right\} \tag{1}$$

to each other, and for some models we even obtain a full list of equivalences. These connections allow for three main applications.

First, they may be used to recover known results in statistics or even to obtain new characterizations through invariant theory. Second, they yield algorithmic consequences. Namely, we show that norm minimization under a certain group action relates to maximizing the likelihood function over a respective model. Thus, one can use algorithms from invariant theory in ML estimation. Moreover, complexity results, in particular those from the thesis' part on complexity, carry over to statistics. Third, one can translate problems from statistics to invariant theory, and vice versa. This has already been crucially used to compute maximum likelihood (ML) thresholds for matrix normal models [DM21] and for tensor normal models [DMW22]. These thresholds capture how often one should at least sample typically. Highly simplified speaking, the papers [DM21; DMW22] translated the problem on ML thresholds via (1) to a problem in terms of stability notions. Then they solved the latter using invariant-theoretic techniques and translated the result back.

As a summary, invariant theory embraces the thesis' main contributions on computational complexity and algebraic statistics. A prominent link is provided through the three computational problems NCM, Norm Minimization and Scaling. Moreover, important group actions such as torus actions as well as operator and tensor scaling action play a prominent role throughout the thesis.

Organization

As suggested by its title, the thesis consists of three parts. Part I, containing Chapters 1 and 2, collects the required prerequisites from *invariant theory*. In Part II (Chapters 3 – 5) we present the results on *computational complexity*. Fi-

nally, Part III (Chapters 6 – 10) contains the content regarding *algebraic statistics*. In the following we give short descriptions of each chapter.

Chapter 1 presents the necessary background on algebraic groups, matrix Lie groups and the representation theory of these groups. In particular, it defines the concept of (topological) stability notions.

Chapter 2 collects criteria to test stability notions: the Hilbert-Mumford Criterion, King's Criterion for actions on quivers, Popov's Criterion for solvable groups and, of particular importance for this thesis, the Kempf-Ness Theorem.

Chapter 3 gives an introduction to computational invariant theory and its manifold applications. This gives us the opportunity to embed and locate the contributions of this thesis in the research area. We introduce the three computational problems of main interest: Null Cone Membership (NCM) 3.1.2, Norm Minimization 3.1.3 and the Scaling Problem 3.1.4. Furthermore, we discuss known algorithms for these problems and their computational complexity. The latter serves as a preparation of the next two chapters.

Chapter 4 treats the precision parameters *weight margin* and *gap* to solve NCM via optimization methods. We prove (exponentially) small bounds on these parameters for tensor scaling, polynomial scaling and quiver actions, cf. [FR21].

Chapter 5 presents the main result from [FR21] on the *diameter*: it can be exponentially large for tensor scaling. We discuss its implications, related literature, and mention the main proof ideas.[4]

Chapter 6 gives a general introduction to maximum likelihood (ML) estimation, focusing on discrete models and on Gaussian models. It prepares the following four chapters.

Chapter 7 presents results from [AKRS21b]: we link toric invariant theory to ML estimation for *log-linear models*, a huge class of discrete models. In particular, norm minimizers under the action yield the MLE and we obtain a dictionary between some notions in (1).

Chapter 8 sets the stage for the final two chapters. It shows that *any* Gaussian model, that is closed under positive scalars, admits relations to invariant theory which we call the weak correspondence, [MRS21]. The latter provides a dictionary between some notions of (1) and shows that norm minimizers give rise to an MLE, and any MLE is obtained this way. The assumptions notably go *beyond* the setting of groups.

Chapter 9 is based on [AKRS21a] and studies the new concept of *Gaussian group models*. These are Gaussian models induced by a group (action). The group structure allows to extend the results from Chapter 8. In particular, the weak correspondence can be strengthened to an (almost) complete dictionary for two types of models. The first class are Gaussian group models given by a Zariski closed self-adjoint group, and the second consists of Gaussian graphical models on transitive directed acyclic graphs (TDAGs).

[4] All main proof ideas for the diameter bound are due to my co-author Cole Franks. For brevity, we refrain from including all details in this thesis.

Chapter 10 presents the work [MRS21]: it studies symmetries in Gaussian graphical models on directed acyclic graphs (DAGs). The symmetries are given by a graph colouring and the respective models are called *restricted DAG (RDAG) models*. We characterize ML estimation for these models, bound their ML thresholds and compare them to their undirected analogues. The theory was initially inspired by the results of Chapter 9. Indeed, we can extend the weak correspondence from Chapter 8 to a full dictionary, and we discuss connections to Gaussian group models from Chapter 9.

Notation and Conventions

We always work over the real or over the complex numbers. Often we do so in parallel and in that case $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ denotes the ground field. Its group of units is $\mathbb{K}^\times$. For a $\mathbb{K}$-vector space V, the ring of $\mathbb{K}$-linear endomorphisms is denoted $\mathrm{End}(V)$ and its group of units, i.e., the group of $\mathbb{K}$-linear automorphisms is denoted by $\mathrm{GL}(V)$. The projective space of V is denoted by $\mathbb{P}(V)$. Vectors in $\mathbb{K}^m$ are usually viewed as *column* vectors. The space of $m_1 \times m_2$ matrices with entries in $\mathbb{K}$ is denoted by $\mathbb{K}^{m_1 \times m_2}$. Similarly, $\mathbb{K}^{m_1} \otimes_{\mathbb{K}} \cdots \otimes_{\mathbb{K}} \mathbb{K}^{m_d}$ is the space of tensors of order d. Often, we suppress the field over which we are tensoring.

As an important convention, we equip[5] the spaces of (column) vectors, of matrices and of tensors with their standard Euclidean/Hermitian inner product and its induced norm. In particular, $\mathbb{K}^{m_1 \times m_2}$ is equipped with the trace inner product, which induces the Frobenius norm. Furthermore, we follow the convention of [BFG+19] that for $\mathbb{K} = \mathbb{C}$ an inner product is $\mathbb{C}$-linear in the *second*(!) component, and semilinear in the first.

All algebraic groups considered in this thesis are affine, and the same usually applies to varieties. We stress that we work with the $\mathbb{K}$-points of algebraic groups (and varieties). This requires some caution when $\mathbb{K} = \mathbb{R}$ and occasionally we have to consult results from real algebraic geometry. All rational representations of algebraic groups are assumed to be finite-dimensional.

We stress that the *default topology* in this thesis (even in algebraic geometric settings) is the Euclidean topology. We explicitly indicate the Zariski topology, e.g., by writing "Zariski closed". Accordingly, the Euclidean closure is indicated by $\overline{(\cdot)}$, while the Zariski closure is $\overline{(\cdot)}^{\mathrm{Z}}$.

Manifolds and Lie groups are always considered to be smooth.

When working with Gaussian distributions, we stress that we always assume the mean to be zero and known. Furthermore, by convention we work with the *concentration matrix*[6], i.e., the inverse of the covariance matrix.

Let us briefly collect further frequently used notation. A detailed list of symbols is provided at the end of the thesis.

Usually, ε is a positive real number. The imaginary unit is denoted by $\mathfrak{i}$. We denote the set $\{1, 2, \ldots, m\}$ by $[m]$. For $i \in [m]$, the i^{th} canonical unit vector

[5] if not stated otherwise

[6] also called *precision matrix*

in $\mathbb{K}^m$ (with i^{th} entry one, and all other entries zero) is denoted e_i. Similarly, $E_{ij} \in \mathbb{K}^{m_1 \times m_2}$ is the matrix with entry one at position (i, j) and all other entries are zero. The $m \times m$ identity matrix is denoted I_m and the all-ones vector by $\mathbb{1}_m \in \mathbb{K}^m$. Moreover, for $i \in [m]$ we set

$$\epsilon_i := e_i - \frac{1}{m}\mathbb{1}_m \,.$$

The transpose of a matrix is indicated by $(\cdot)^\mathsf{T}$ and the Hermitian transpose by $(\cdot)^\dagger$. For $M \in \mathbb{K}^{m \times m}$, its determinant is $\det(M)$ and $\mathrm{tr}(M)$ is the trace of M.

Finally, we use the following useful notation, which is quite common in statistics. For a tensor $v = (v_{ijk}) \in (\mathbb{C}^m)^{\otimes 3}$ define the "slice sums"

$$v_{i,+,+} := \sum_{k=1}^{m} v_{ijk}, \quad v_{i,j,+} := \sum_{j,k=1}^{m} v_{ijk}, \quad v_{+,+,+} := \sum_{i,j,k=1}^{m} v_{ijk}, \qquad \text{etc.}$$

Similarly, for a vector $x \in \mathbb{K}^m$, x_+ denotes the sum over all entries of x, and for a matrix $M \in \mathbb{K}^{m_1 \times m_2}$, $M_{i,+}$ is the i^{th} row sum, $M_{+,j}$ the j^{th} column sum, and $M_{+,+}$ the sum over all entries of M. Of course, this notation can also be extended to tensors of order $d \geq 4$.

Part I

Invariant Theory

Chapter 1

Algebraic Group Actions

This chapter collects required preliminaries and thereby fixes notation. The presented material covers a wide range, because we need algebraic as well as analytic methods. The aims of the chapter are to allow readers from diverse contexts to follow, and to keep the thesis as self-contained as possible.

Usually, we skip proofs and refer to the literature. References for further reading are provided at the beginning of each section. A reader familiar with the presented material may skip this chapter and only consult it when referenced.

Organization. Section 1.1 recalls linear algebraic groups while Section 1.2 introduces their analytic analogue of (matrix) Lie groups. Afterwards, Section 1.3 reviews aspects of the representation theory of these groups. Finally, Section 1.4 defines the (topological) stability notions and discusses their relation to Geometric Invariant Theory.

1.1 Linear Algebraic Groups

We briefly review the required knowledge on linear algebraic groups. For a detailed treatment the reader is referred to the many textbooks available: e.g., classical books are [Bor91; Hum75; Spr98], a treatment in scheme language is given in [Mil17; Wat79], and a combined treatment of algebraic groups and Lie groups can be found in [Bor06; GW09; OV90; Pro07].

Basic Definitions and $\mathbb{R}$-structures

In this thesis we often study real and complex algebraic settings in parallel. For this, it is convenient to use the concepts of $\mathbb{R}$-structures on complex vector spaces and varieties, compare [Bor91, AG §11 and §12] or [Spr98, Chapter 11]. Given a (not necessarily finite dimensional) complex vector space V, an $\mathbb{R}$-structure on V is an $\mathbb{R}$-vector space $V_{\mathbb{R}} \subseteq V$ such that scalar extension of the inclusion yields $V_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C} = V$. A $\mathbb{C}$-linear map $f\colon V \to W$ of $\mathbb{C}$-vector spaces with $\mathbb{R}$-structures is an $\mathbb{R}$-*morphism* or *defined over* $\mathbb{R}$, if $f(V_{\mathbb{R}}) \subseteq W_{\mathbb{R}}$.

Now, let X be an affine variety over $\mathbb{C}$ with coordinate ring $\mathbb{C}[X]$. An $\mathbb{R}$-structure on X is an $\mathbb{R}$-structure $\mathbb{R}[X]$ on $\mathbb{C}[X]$, which is an $\mathbb{R}$-subalgebra of $\mathbb{C}[X]$. An affine complex variety with $\mathbb{R}$-structure is simply called a $\mathbb{R}$-*variety*. Usually, we identify X with its set $X_{\mathbb{C}}$ of $\mathbb{C}$-rational points, which correspond to $\mathbb{C}$-algebra morphisms $\mathbb{C}[X] \to \mathbb{C}$. If X is an $\mathbb{R}$-variety, then $X_{\mathbb{R}}$ denotes the set of $\mathbb{R}$-rational points, which correspond to $\mathbb{C}$-algebra morphisms $\mathbb{C}[X] \to \mathbb{C}$ that

are defined over $\mathbb{R}$. We note that $X_{\mathbb{R}}$ is a real algebraic variety. Moreover, a morphism $\varphi\colon X \to Y$ of $\mathbb{R}$-varieties is called an $\mathbb{R}$-*morphism* or *defined over* $\mathbb{R}$, if its associated map $\varphi^*\colon \mathbb{C}[Y] \to \mathbb{C}[X]$ on coordinate rings is defined over $\mathbb{R}$.

Note that starting from a real algebraic situation, we naturally obtain by scalar extension a complex algebraic setting with natural $\mathbb{R}$-structures.

Next, we recall linear algebraic groups, which are ubiquitous in this thesis. We remind the reader that in the whole thesis $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

Definition 1.1.1 (Linear algebraic group)**.** A *linear algebraic group* G over $\mathbb{K}$ is an affine algebraic group over $\mathbb{K}$. That is, G is an affine variety over $\mathbb{K}$ endowed with a group structure such that multiplication and inversion are morphisms of varieties over $\mathbb{K}$. ▲

A *morphism of algebraic groups* over $\mathbb{K}$ is a morphism of varieties that is also a group morphism. Such a morphism is an isomorphism of algebraic groups if its inverse is as well a morphism of algebraic groups.

Any Zariski closed subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is a linear algebraic group over $\mathbb{K}$. Actually, the naming originates from the fact that any linear algebraic group over $\mathbb{K}$ is isomorphic to a Zariski closed subgroup of some $\mathrm{GL}_m(\mathbb{K})$, see [Bor91, Proposition 1.10] or [Wat79, Theorem in §3.4]. Since all algebraic groups in this thesis are affine, we often drop the term *"linear"*.

Example 1.1.2. The following are linear algebraic groups over $\mathbb{K}$.

1. The general linear group $\mathrm{GL}_m(\mathbb{K})$ of invertible $m \times m$ matrices over $\mathbb{K}$.
2. The special linear group $\mathrm{SL}_m(\mathbb{K}) := \{g \in \mathrm{GL}_m(\mathbb{K}) \mid \det(g) = 1\}$.
3. The intersection $G \cap H$ of two Zariski closed subgroups $G, H \subseteq \mathrm{GL}_m(\mathbb{K})$.
4. Any torus $(\mathbb{K}^\times)^m$ is linear algebraic. In particular, the groups

$$\begin{aligned} &\mathrm{GT}_m(\mathbb{K}) := \{g \in \mathrm{GL}_m(\mathbb{K}) \mid g \text{ is diagonal}\} \\ \text{and} \qquad &\mathrm{ST}_m(\mathbb{K}) := \mathrm{GT}_m(\mathbb{K}) \cap \mathrm{SL}_m(\mathbb{K}) \cong (\mathbb{K}^\times)^{m-1} \end{aligned}$$

   are linear algebraic groups.
5. The *additive group* $(\mathbb{K}^m, +)$.
6. The group $\mathrm{B}_m(\mathbb{K})$ of invertible upper triangular matrices.
7. The group $\mathfrak{U}_m(\mathbb{K}) := \{g \in \mathrm{B}_m(\mathbb{K}) \mid \forall\, i \in [m]\colon g_{ii} = 1\}$ of unipotent upper triangular matrices.
8. The groups of orthogonal respectively special orthogonal matrices over $\mathbb{K}$:

$$\mathrm{O}_m(\mathbb{K}) := \{g \in \mathrm{GL}_m(\mathbb{K}) \mid g^\mathsf{T} g = \mathrm{I}_m\} \quad \text{and} \quad \mathrm{SO}_m(\mathbb{K}) := \mathrm{O}_m(\mathbb{K}) \cap \mathrm{SL}_m(\mathbb{K}).$$

9. The (semi-)direct product of two linear algebraic groups. ♢

Example 1.1.3. The groups of unitary respectively special unitary matrices

$$\mathrm{U}_m := \left\{ g \in \mathrm{GL}_m(\mathbb{C}) \mid g^\dagger g = \mathrm{I}_m \right\} \qquad \text{and} \qquad \mathrm{SU}_m := \mathrm{U}_m \cap \mathrm{SL}_m(\mathbb{C})$$

are *not* algebraic over $\mathbb{C}$. However, after identifying $\mathbb{C} \cong \mathbb{R}^2$ we see that U_m and SU_m are *real* algebraic subgroups of $\mathrm{GL}_{2m}(\mathbb{R})$. ♢

Examples like $\mathrm{GL}_m(\mathbb{K})$, $\mathrm{GT}_m(\mathbb{K})$ and $\mathrm{O}_m(\mathbb{K})$ indicate that one can often study the real and complex situation in parallel, which is especially useful for Part III on algebraic statistics. In order to do so, we extend $\mathbb{R}$-structures to the setting of algebraic groups.

An $\mathbb{R}$-*group* is a complex algebraic group G that is an $\mathbb{R}$-variety such that multiplication and inversion are defined over $\mathbb{R}$. Thus, given an $\mathbb{R}$-group G, its $\mathbb{K}$-rational points $G_\mathbb{K}$ form an algebraic group over $\mathbb{K}$, i.e., G indeed encodes a real and a complex algebraic group at the same time and $\dim_\mathbb{R} G_\mathbb{R} = \dim_\mathbb{C} G_\mathbb{C}$. Note that all groups given in Example 1.1.2 for $\mathbb{K} = \mathbb{C}$ are naturally $\mathbb{R}$-groups. E.g., $\mathrm{GL}_m(\mathbb{C})$ is an $\mathbb{R}$-group with $\mathbb{R}$-rational points $\mathrm{GL}_m(\mathbb{R})$. An $\mathbb{R}$-*morphism of* $\mathbb{R}$-*groups* G and G' is a morphism $\varphi\colon G \to G'$ of algebraic groups that is defined over $\mathbb{R}$.

Zariski and Euclidean identity component

Given an algebraic group G over $\mathbb{K}$, the *Zariski identity component* $G^{\circ,\mathrm{Z}}$ is the Zariski connected component of G that contains the identity.

Proposition 1.1.4 ([Bor91, Proposition 1.2])**.** *Let G be a complex algebraic group. Then $G^{\circ,\mathrm{Z}}$ is a normal subgroup of finite index in G whose cosets are the Zariski connected as well as irreducible components of G. If G is an $\mathbb{R}$-group, then $G^{\circ,\mathrm{Z}}$ is defined over $\mathbb{R}$ so that $(G_\mathbb{R})^{\circ,\mathrm{Z}} = (G^{\circ,\mathrm{Z}})_\mathbb{R}$.*

Since all points of an algebraic group G over $\mathbb{K}$ are non-singular, G possesses a canonical structure of a Lie group over $\mathbb{K}$, compare [OV90, Sections 3.1.2 and 2.3.4] and Theorem 1.2.4 below. This will become more apparent in Section 1.2. As an upshot, G carries a natural Euclidean topology.

Now, the *Euclidean identity component* G° is the Euclidean connected component of (the Lie group) G that contains the identity. Since the Euclidean topology is finer than the Zariski topology, it holds that $G^\circ \subseteq G^{\circ,\mathrm{Z}}$ and depending on $\mathbb{K}$ we have the following.

1. For $\mathbb{K} = \mathbb{C}$, one always has equality $G^\circ = G^{\circ,\mathrm{Z}}$.
2. For $\mathbb{K} = \mathbb{R}$, the inclusion $G^\circ \subseteq G^{\circ,\mathrm{Z}}$ may be *strict*.

The first item follows from the facts that $G^{\circ,\mathrm{Z}}$ is irreducible, and that any irreducible complex affine variety is connected in the Euclidean topology [Sha13, Theorem 7.1]. The upcoming example provides a strict containment in the real case. Consequently, we need to be careful in the real case whether we mean the Zariski or Euclidean identity component.

Example 1.1.5. The real algebraic group $\mathrm{GL}_m(\mathbb{R})$ is irreducible and therefore Zariski connected. However, it has two Euclidean connected components, namely

$$\begin{aligned} & \mathrm{GL}_m^+(\mathbb{R}) = \{g \in \mathrm{GL}_m(\mathbb{R}) \mid \det(g) > 0\} \\ \text{and} \qquad & \mathrm{GL}_m^-(\mathbb{R}) = \{g \in \mathrm{GL}_m(\mathbb{R}) \mid \det(g) < 0\}. \end{aligned}$$

In particular, $\mathrm{GL}_m(\mathbb{R})^\circ = \mathrm{GL}_m^+(\mathbb{R}) \subsetneq \mathrm{GL}_m(\mathbb{R}) = \mathrm{GL}_m(\mathbb{R})^{\circ,\mathrm{Z}}$. ♢

Nevertheless, also in the real setting the Euclidean identity component G° is a normal subgroup, and its cosets are the finitely many (see next theorem) Euclidean connected components of G.

Theorem 1.1.6 ([Whi57, Theorem 3])**.** *A real algebraic variety $V \subseteq \mathbb{R}^m$ has finitely many Euclidean connected components.*

We note that the preceding theorem holds more generally for semialgebraic subsets of $\mathbb{R}^m$, compare [BCR98, Theorem 2.4.5].

Properties of Morphisms of Algebraic Groups

Proposition 1.1.7. *Let $\varphi\colon G \to G'$ be a morphism of complex algebraic groups.*

(a) $\varphi(G)$ is a Zariski closed subgroup of G'. If φ is an $\mathbb{R}$-morphism of $\mathbb{R}$-groups, then $\varphi(G)$ is defined over $\mathbb{R}$.

(b) $\varphi\big(G^{\circ,\mathrm{Z}}\big) = \varphi(G)^{\circ,\mathrm{Z}}$.

(c) $\ker(\varphi)$ is a Zariski closed normal subgroup of G. If φ is an $\mathbb{R}$-morphism of $\mathbb{R}$-groups, then $\ker(\varphi)$ is defined over $\mathbb{R}$.

(d) $\dim_{\mathbb{C}} G = \dim_{\mathbb{C}} \ker(\varphi) + \dim_{\mathbb{C}} \varphi(G)$. If φ is an $\mathbb{R}$-morphism of $\mathbb{R}$-groups, then $\dim_{\mathbb{R}} G_{\mathbb{R}} = \dim_{\mathbb{R}} \ker(\varphi)_{\mathbb{R}} + \dim_{\mathbb{R}} \varphi(G)_{\mathbb{R}}$ as real algebraic groups.

Proof. Parts (a), (b) and the first part of (d) are [Bor91, Corollary 1.4], while (c) follows from [Spr98, Propositions 2.2.5(i) and 12.1.3]. The second part of (d) follows from $\dim_{\mathbb{C}} H = \dim_{\mathbb{R}} H_{\mathbb{R}}$ for any $\mathbb{R}$-group H. □

Regarding parts (a) and (b) of Proposition 1.1.7 the upcoming example stresses the following. In general, one may have $\varphi(G_{\mathbb{R}}) \subsetneq \varphi(G)_{\mathbb{R}}$ and $\varphi(G_{\mathbb{R}}^{\circ,\mathrm{Z}}) \subsetneq \varphi(G)_{\mathbb{R}}^{\circ,\mathrm{Z}}$, and the image of $\mathbb{R}$-points $\varphi(G_{\mathbb{R}})$ does not need to be Zariski closed. Still, $\varphi(G_{\mathbb{R}})$ is well-behaved as we shall see in Corollary 1.2.6.

Example 1.1.8 (taken from [Bor06, §5.2])**.** Consider the surjective $\mathbb{R}$-morphism

$$\chi\colon\ \mathrm{GL}_m(\mathbb{C}) \mapsto \mathbb{C}^\times, \quad g \mapsto \det(g)^2$$

of Zariski connected $\mathbb{R}$-groups. It is not surjective on the $\mathbb{R}$-rational points, as

$$\chi(\mathrm{GL}_m(\mathbb{R})) = \mathbb{R}_{>0} \subsetneq \mathbb{R}^\times = \chi(\mathrm{GL}_m(\mathbb{C}))_{\mathbb{R}}.$$

We see that $\chi(\mathrm{GL}_m(\mathbb{R}))$ is not real algebraic, but only semialgebraic. ♢

Algebraic Group Actions

Let G be an algebraic group over $\mathbb{K}$ and V an affine variety over $\mathbb{K}$. A *group (left-)action* of G on V is a map

$$\alpha\colon G \times V \to V,\ (g,v) \mapsto \alpha(g,v) =: g \cdot v$$

such that $\mathrm{id} \cdot v = v$ and $(gh) \cdot v = g \cdot (h \cdot v)$ hold for all $v \in V$ and $g, h \in G$. An *algebraic group action* of G on V is a group action α that is also a morphism of varieties over $\mathbb{K}$. As usual, we define the *orbit* of v and the *stabilizer* of v as

$$G \cdot v := \big\{ g \cdot v \mid g \in G \big\} \qquad \text{and} \qquad G_v := \big\{ g \in G \mid g \cdot v = v \big\}, \tag{1.1}$$

respectively. Note that $g \cdot v - v = 0$ gives polynomial equations in the entries of g, since the action is algebraic. Consequently, the stabilizer G_v is a Zariski closed subgroup of G, i.e., is itself an algebraic group over $\mathbb{K}$. In this thesis we focus on the following specific case of algebraic group actions.

Definition 1.1.9 (Rational Representation)**.** Let G be an algebraic group over $\mathbb{K}$ and V a finite dimensional $\mathbb{K}$-vector space. A *rational representation* is a morphism $\pi\colon G \to \mathrm{GL}(V)$ of algebraic groups over $\mathbb{K}$. Equivalently, the induced $\mathbb{K}$-linear action

$$G \times V \to V, (g,v) \mapsto g \cdot v := \pi(g)(v)$$

is algebraic. Note that $\mathbb{K}$-linear algebraic actions of G on V are in one to one correspondence with rational representations $G \to \mathrm{GL}(V)$. ▲

Of course, if G is a complex algebraic $\mathbb{R}$-group and V a complex affine $\mathbb{R}$-variety, an algebraic $\mathbb{R}$-action is an algebraic action α that is an $\mathbb{R}$-morphism. If applicable, this allows to encode algebraic actions over $\mathbb{R}$ and $\mathbb{C}$ at the same time.

The one-dimensional representations of a group are of particular interest.

Definition 1.1.10 (Character)**.** Let G be a complex algebraic group. A *character* of G is an algebraic group morphism $\chi\colon G \to \mathbb{C}^\times = \mathrm{GL}_1(\mathbb{C})$. The set of all characters of G is denoted $\mathfrak{X}(G)$. It becomes an abelian group (written additively) via $(\chi + \chi')(g) := \chi(g)\chi'(g)$ for all $g \in G$. If G is an $\mathbb{R}$-group, then the subgroup of characters defined over $\mathbb{R}$ is denoted $\mathfrak{X}_{\mathbb{R}}(G)$. ▲

Next, we collect some properties of real and complex orbits.

Proposition 1.1.11 ([Bor91, Proposition I.1.8])**.** *Let G be a complex algebraic group acting algebraically on a complex affine variety V. The orbit $G \cdot v$ of $v \in V$ is Zariski-open in its Zariski-closure. Its boundary consists of orbits of strictly lower dimension. In particular, orbits of minimal dimension are Zariski-closed.*

A subset U of a complex affine variety V with U being Zariski open in $\overline{U}^{\mathrm{Z}}$ has Euclidean closure $\overline{U} = \overline{U}^{\mathrm{Z}}$; see [Wal17, Corollary 1.26] or [Kra84, Section AI.7.2]. Thus, an important consequence of Proposition 1.1.11 is the following.

Corollary 1.1.12. *Let G be an algebraic group over $\mathbb{C}$ acting algebraically on a complex affine variety V. For $v \in V$, the Euclidean and the Zariski closure of the orbit coincide: $\overline{G \cdot v} = \overline{G \cdot v}^{\mathrm{Z}}$.*

Remark 1.1.13. We point out that Proposition 1.1.11 and Corollary 1.1.12 fail over $\mathbb{R}$. For this, consider the character given in Example 1.1.8 as an $\mathbb{R}$-algebraic action of $G = \mathrm{GL}_m(\mathbb{C})$ on $V = \mathbb{C}$. For $v = 1 \in V_{\mathbb{R}}$, the orbit $G_{\mathbb{R}} \cdot v = \mathbb{R}_{>0}$ is *not* Zariski open in its Zariski closure $\mathbb{R} = V_{\mathbb{R}}$, and $\overline{G_{\mathbb{R}} \cdot v} = \mathbb{R}_{\geq 0} \subsetneq \mathbb{R} = \overline{G_{\mathbb{R}} \cdot v}^{\mathrm{Z}}$. Moreover, we have the strict containment

$$G_{\mathbb{R}} \cdot v = \mathbb{R}_{>0} \subsetneq \mathbb{R}^{\times} = (G \cdot v) \cap V_{\mathbb{R}},$$

of the real orbit in the $\mathbb{R}$-rational points of the complex orbit. Here, $(G \cdot v) \cap V_{\mathbb{R}}$ is the union of two real orbits, namely $G_{\mathbb{R}} \cdot v$ and $G_{\mathbb{R}} \cdot (-1)$. ▽

Actually, it is a general fact that $(G \cdot v) \cap V_{\mathbb{R}}$ is a finite union of real orbits.

Proposition 1.1.14 ([BH62, Proposition 2.3])**.** *Let G be a connected complex algebraic $\mathbb{R}$-group, $\pi\colon G \to \mathrm{GL}(V)$ a rational representation defined over $\mathbb{R}$ and $v \in V$. Denote the Euclidean identity component of $G_{\mathbb{R}}$ by $(G_{\mathbb{R}})^{\circ}$. If $(G \cdot v) \cap V_{\mathbb{R}}$ is not empty, then it is a finite union of $(G_{\mathbb{R}})^{\circ}$-orbits, which are Euclidean closed if $G \cdot v$ is Euclidean closed.*

Classes of Linear Algebraic Groups

We end this section by presenting different types of linear algebraic groups. Since we usually work with algebraic subgroups $G \subseteq \mathrm{GL}_m(\mathbb{K})$, some definitions are ad-hoc and do not follow usual definitions, but are rather equivalent characterizations that require a proof.

Based on [Bor91, Propositions 8.2 and 8.4] we define the following.

Definition 1.1.15. Let G be an algebraic group over $\mathbb{C}$. We say G is *diagonalizable* if G is isomorphic to a Zariski closed subgroup of $\mathrm{GT}_m(\mathbb{C})$. We say G is a *torus*, if T is isomorphic to some $(\mathbb{C}^{\times})^m \cong \mathrm{GT}_m(\mathbb{C})$. If G is a diagonalizable $\mathbb{R}$-group, we call G *split over* $\mathbb{R}$ if G is $\mathbb{R}$-isomorphic to a Zariski closed subgroup of $\mathrm{GT}_m(\mathbb{C})$. ▲

Example 1.1.16 (Non-split torus, [Bor91, §8.16])**.** We have an isomorphism

$$\mathrm{SO}_2(\mathbb{C}) = \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \middle| \, a, b \in \mathbb{C},\, a^2 + b^2 = 1 \right\} \to \mathbb{C}^{\times}, \quad \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \mapsto a + \mathrm{i}b$$

of complex algebraic groups. Thus, $T := \mathrm{SO}_2(\mathbb{C})$ is a complex torus. It is not split over $\mathbb{R}$: $T_{\mathbb{R}} = \mathrm{SO}_2(\mathbb{R})$ is the compact unit circle, which is not isomorphic to the non-compact set $\mathbb{R}^{\times} = (\mathbb{C}^{\times})_{\mathbb{R}}$. ♢

We note the following. All diagonalizable $\mathbb{R}$-groups in this thesis will be $\mathbb{R}$-split, so we usually drop the term "$\mathbb{R}$-split". Moreover, if we have a real algebraic group G and say it is diagonalizable, then we mean that the complex algebraic group obtained by scalar extension is $\mathbb{R}$-split diagonalizable.

We collect properties of diagonalizable groups and their character groups.

Proposition 1.1.17. *Let G be a complex diagonalizable $\mathbb{R}$-group.*

(a) If H is a Zariski closed subgroup of G, then any character $\chi \in \mathfrak{X}(H)$ extends to a character on G. [Bor91, Proposition 8.2(c)]

(b) G is split over $\mathbb{R}$ if and only if $\mathfrak{X}_{\mathbb{R}}(G) = \mathfrak{X}(G)$. [Bor91, Corollary 8.2]

(c) The character group $M := \mathfrak{X}(G)$ is a finitely generated abelian group ([Spr98, Corollary 3.2.4]) and $\mathfrak{X}(G^{\mathrm{o},Zar}) = M/M^{\mathrm{tor}}$, where M^{tor} denotes the torsion subgroup of M. ([Spr98]: 3.2.5 together with proof of Corollary 3.2.7)

(d) G is a torus if and only if it is Zariski connected. In this case $\mathfrak{X}(G) = \mathbb{Z}^m$, where m is such that $G \cong (\mathbb{C}^\times)^m$. [Bor91, Proposition 8.5]

(e) If G is $\mathbb{R}$-split diagonalizable, then it is isomorphic to the direct product of an $\mathbb{R}$-split torus and a finite group. [Bor91, Proposition 8.7]

Diagonalizable groups are a special case of so-called solvable groups. Thanks to [Bor91, Theorem 15.4] we give the following ad-hoc definition.

Definition 1.1.18 (Solvable Group)**.** An algebraic group G over $\mathbb{K}$ is called *($\mathbb{K}$-split) solvable* if it is isomorphic to a Zariski closed subgroup of $\mathrm{B}_m(\mathbb{K})$. ▲

All solvable groups considered in this thesis are split over $\mathbb{K}$, and we usually drop this term. Another special case of solvable groups are unipotent groups.

Definition 1.1.19 (Unipotent Group)**.** Let G be an algebraic group over $\mathbb{K}$. We say G is *unipotent*, if it is isomorphic to a Zariski-closed subgroup of $\mathfrak{U}_m(\mathbb{K})$. ▲

The preceding ad-hoc definition is justified by [Bor91, Corollary 4.8] (or [Wat79, Theorem in 8.3]).

Proposition 1.1.20 ([Wat79, Corollary in 8.3])**.** *Let U be a unipotent group. Then $\mathfrak{X}(U) = 0$.*

Definition 1.1.21 (Unipotent Radical)**.** Let G be a complex algebraic group. The *unipotent radical* $R_u(G)$ is the maximal Zariski closed, connected, normal unipotent subgroup of G. ▲

Definition 1.1.22 (Reductive Group)**.** Let G be a linear algebraic group over $\mathbb{C}$. We call G a *reductive group* if its unipotent radical is trivial, i.e., $R_u(G) = \{\mathrm{id}\}$. A real linear algebraic group is called *reductive*, if the complex group obtained by scalar extension is reductive. ▲

We stress that we do *not* assume a reductive group to be connected, as is done in some literature.

Example 1.1.23. The following are reductive groups over $\mathbb{K}$.

1. $\mathrm{GL}_m(\mathbb{K})$ and $\mathrm{SL}_m(\mathbb{K})$.
2. $\mathrm{O}_m(\mathbb{K})$ and $\mathrm{SO}_m(\mathbb{K})$.
3. Any diagonalizable group (in particular, any torus) over $\mathbb{K}$ is reductive.
4. The direct product of two reductive groups over $\mathbb{K}$. ♢

Example 1.1.24 (Non-reductive groups)**.** The additive group $\mathbb{K}^m$ is unipotent and hence not reductive. Similarly, $\mathfrak{U}_m(\mathbb{K})$ for $m \geq 2$ is not reductive; note that $\mathfrak{U}_1(\mathbb{K})$ is trivial. The group $\mathrm{B}_m(\mathbb{K})$ of invertible upper triangular matrices is not reductive for $m \geq 2$, as its unipotent radical $\mathfrak{U}_m(\mathbb{K})$ is non-trivial. ♢

Any algebraic group over $\mathbb{K}$ admits the following decomposition, because $\mathbb{K}$ has characteristic zero.

Theorem 1.1.25 (Levi-type decomposition, [Mos56])**.** *Let G be a linear algebraic group over $\mathbb{K}$ with unipotent radical U. Then there is a reductive group R over $\mathbb{K}$ such that G is the semi-direct product of R and U: $G \cong R \ltimes U$. In particular, a solvable group is the semi-direct product of a diagonalizable group and its unipotent radical.*

Example 1.1.26. One has $\mathrm{B}_m(\mathbb{K}) = \mathrm{GT}_m(\mathbb{K}) \ltimes \mathfrak{U}_m(\mathbb{K})$. ♢

1.2 Matrix Lie Groups and Lie Algebras

In this section we collect preliminary knowledge on Lie groups and their Lie algebras. For convenience and brevity, we restrict to so-called matrix Lie groups. This ensures a concrete approach, which is sufficient for this thesis. For further details we refer to textbooks such as [Hal15; Kna96; Lee13], and for a combined treatment of Lie groups and algebraic groups to [Bor06; GW09; OV90; Pro07].

Matrix Lie Groups

Definition 1.2.1 (Matrix Lie group, [Hal15, Definition 1.4])**.**
A *matrix Lie group* is a Euclidean closed subgroup G of $\mathrm{GL}_m(\mathbb{C})$.[1] ▲

Remember that a Lie group in the abstract sense is a smooth manifold with a group structure such that multiplication and inversion are smooth maps. Moreover, a *morphism of Lie groups* is a group morphism that is smooth. Similarly as for algebraic groups, the Euclidean connected component of a (matrix) Lie group containing the identity is denoted G°.

As suggested by the name, any matrix Lie group is a Lie group [Hal15, Corollary 3.45]. This result was first proven by John von Neumann. More generally, one has the following theorem due to Élie Cartan.

Theorem 1.2.2 (Closed Subgroup Theorem, [Lee13, Theorem 2.12])**.**
Let G be a Lie group and $H \subseteq G$ a Euclidean closed subgroup. Then H is an embedded Lie subgroup of G.

Example 1.2.3. Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. The following groups are matrix Lie groups.

1. Any Zariski closed subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is a matrix Lie group, since it is Euclidean closed in $\mathrm{GL}_m(\mathbb{K})$, and hence in $\mathrm{GL}_m(\mathbb{C})$. In particular, all groups in Example 1.1.2 are matrix Lie groups. Moreover, any linear algebraic group over $\mathbb{K}$ is isomorphic to a matrix Lie group.

[1] We stress that we mean the complex general linear group. But, of course, any Euclidean closed subgroup of $\mathrm{GL}_m(\mathbb{R})$ is a Euclidean closed subgroup of $\mathrm{GL}_m(\mathbb{C})$.

2. The groups U_m and SU_m from Example 1.1.3 are Euclidean closed in $\mathrm{GL}_m(\mathbb{C})$ and hence matrix Lie groups.

3. The intersection $G \cap H$ of two matrix Lie groups $G, H \subseteq \mathrm{GL}_m(\mathbb{C})$ is a matrix Lie group.

4. Let $G \subseteq \mathrm{GL}_{m_1}(\mathbb{C})$ and $H \subseteq \mathrm{GL}_{m_2}(\mathbb{C})$ be matrix Lie groups. Under the block-diagonal embedding

$$G \times H \hookrightarrow \mathrm{GL}_{m_1+m_2}(\mathbb{C}), \quad (g,h) \mapsto \begin{pmatrix} g & 0 \\ 0 & h \end{pmatrix}$$

the direct product $G \times H$ is a matrix Lie subgroup of $\mathrm{GL}_{m_1+m_2}(\mathbb{C})$. ♢

Regarding Example 1.2.3 Item 1, one has the following general statement.

Theorem 1.2.4 ([OV90, Theorem 6 in §3.1.2])**.** *Any complex (real) algebraic group is a complex (real) Lie group of the same dimension.*

In general, the image of a Lie group morphism need not be a Lie group. However, in the real algebraic setting this is true and provides an analogue of Proposition 1.1.7(a) in the real situation, also compare Example 1.1.8. Due to the lack of an explicit reference, we provide proofs.

Proposition 1.2.5. *Let $\varphi\colon G \to G'$ be a morphism of real linear algebraic groups. Then $\varphi(G)$ is a closed Lie subgroup of G'.*

Proof. Set $H := \varphi(G)$. We can consider G as a real algebraic subgroup of some $\mathrm{GL}_m(\mathbb{R}) \subseteq \mathbb{R}^{m\times m} \cong \mathbb{R}^{m^2}$, and similarly for G'. In particular, we can view them as matrix Lie groups.[2] By Theorem 1.2.2, the Euclidean closure $\overline{H} \subseteq G'$ is a closed Lie subgroup. Hence, it suffices to show that $H = \overline{H}$. For this, we need several results from Real Algebraic Geometry and refer to [BCR98].

Since φ is a morphism of real algebraic groups, its image H is semialgebraic as a consequence of Tarski-Seidenberg, [BCR98, Proposition 2.2.7]. Thus, $\overline{H} \subseteq G'$ and $\overline{H}\backslash H$ are semialgebraic as well. There is a natural notion of (local) dimension of a semialgebraic set [BCR98, Section 2.8]. We have

$$\dim\left(\overline{H}\backslash H\right) < \dim H = \dim \overline{H}$$

as semi-algebraic sets [BCR98, Propositions 2.8.2 and 2.8.13]. If $\overline{H}\backslash H \neq \emptyset$ then the Lie group $\overline{H}$ has points of different local dimension in the sense of semialgebraic sets. But the local dimension in the semialgebraic sense is equal to the local dimension in the manifold sense (i.e., the dimension of the tangent space at the point); compare proof of [BCR98, Proposition 2.8.14]. This contradicts the fact that $\dim T_h\overline{H} = \dim \mathrm{Lie}\left(\overline{H}\right)$ for all $h \in \overline{H}$. Therefore, $\overline{H}\backslash H$ must be empty. □

Corollary 1.2.6. *Let $\varphi\colon G \to G'$ be an $\mathbb{R}$-morphism of complex linear algebraic $\mathbb{R}$-groups. Then $\varphi(G_\mathbb{R})$ is a closed, semialgebraic Lie subgroup of $G'_\mathbb{R}$, respectively of $\varphi(G)_\mathbb{R}$. It holds that $\dim \varphi(G_\mathbb{R}) = \dim \varphi(G)_\mathbb{R}$ and $(\varphi(G)_\mathbb{R})^\circ \subseteq \varphi(G_\mathbb{R})$.*

[2] They are also Lie groups by Theorem 1.2.4

Proof. On the level of real points we have $\varphi_{\mathbb{R}}\colon G_{\mathbb{R}} \to G'_{\mathbb{R}}$, a morphism of real algebraic groups. By Proposition 1.2.5 and its proof, $\varphi(G_{\mathbb{R}})$ is a closed, semialgebraic Lie subgroup of $G'_{\mathbb{R}}$ and so also of $\varphi(G)_{\mathbb{R}}$. It remains to show $\dim \varphi(G_{\mathbb{R}}) = \dim \varphi(G)_{\mathbb{R}}$. We have

$$\begin{aligned} \dim G_{\mathbb{R}} &= \dim \ker(\varphi)_{\mathbb{R}} + \dim \varphi(G)_{\mathbb{R}} && \text{as real algebraic groups} \\ \dim G_{\mathbb{R}} &= \dim \ker(\varphi_{\mathbb{R}}) + \dim \varphi(G_{\mathbb{R}}) && \text{as Lie groups.} \end{aligned}$$

The first equality is Proposition 1.1.7(d). The second follows since $\varphi_{\mathbb{R}}$ is of constant rank as a morphism of Lie groups [OV90, Theorem 2 in §1.1.6], and its image $\varphi(G_{\mathbb{R}})$ is a Lie group. Clearly, $\ker(\varphi_{\mathbb{R}}) = \ker(\varphi)_{\mathbb{R}}$. We deduce $\dim \varphi(G_{\mathbb{R}}) = \dim \varphi(G)_{\mathbb{R}}$, because real algebraic groups have the same dimension as when viewed as a Lie group. Finally, $\varphi(G_{\mathbb{R}}) \subseteq \varphi(G)_{\mathbb{R}}$ and the equality of dimensions yield $(\varphi(G)_{\mathbb{R}})^\circ \subseteq \varphi(G_{\mathbb{R}})$. □

Lie Algebras

We introduce Lie algebras of matrix Lie groups. For this, we denote the exponential of a matrix $X \in \mathbb{K}^{m\times m}$ by $\exp(X)$ or also by e^X.

Definition 1.2.7 (Lie algebra)**.** Let $G \subseteq \mathrm{GL}_m(\mathbb{C})$ be a matrix Lie group. The *Lie algebra* of G is

$$\mathrm{Lie}(G) := \big\{X \in \mathbb{C}^{m\times m} \mid \forall\, t \in \mathbb{R}\colon \ \exp(tX) \in G\big\}$$

and we equip it with the Lie bracket $[X, Y] := XY - YX$. ▲

We collect some properties of $\mathrm{Lie}(G)$.

Proposition 1.2.8. *Let $G \subseteq \mathrm{GL}_m(\mathbb{C})$ be a matrix Lie group.*

(a) $\mathrm{Lie}(G)$ *is a $\mathbb{R}$-vector space and $[X, Y] \in \mathrm{Lie}(G)$ for all $X, Y \in \mathrm{Lie}(G)$. With the latter bracket* $\mathrm{Lie}(G)$ *becomes a real Lie algebra. Furthermore,* $\mathrm{Lie}(G)$ *is the tangent space at the identity of G (in the sense of smooth manifolds).*

(b) If G is Zariski closed in $\mathrm{GL}_m(\mathbb{C})$*, then* $\mathrm{Lie}(G)$ *is a $\mathbb{C}$-vector space and hence a complex Lie algebra. In this case,* $\mathrm{Lie}(G)$ *is the tangent space at the identity of G (in the sense of algebraic geometry).*

(c) If $H \subseteq \mathrm{GL}_m(\mathbb{C})$ is a matrix Lie group, then $\mathrm{Lie}(G \cap H) = \mathrm{Lie}(G) \cap \mathrm{Lie}(H)$.

(d) For all $X \in \mathrm{Lie}(G)$, e^X lies in the Euclidean identity component G°.

Proof. The first part of (a) is [Hal15, Theorem 3.20] and the second part is [Hal15, Corollary 3.46]. Item (b) is [Wal17, Theorem 2.8]. Part (c) is an immediate consequence of the definition, and part (d) is [Hal15, Proposition 3.19]. □

Example 1.2.9 ([Hal15, Section 3.4])**.** We list some common Lie algebras.

1. $\mathrm{Lie}\big(\mathrm{GL}_m(\mathbb{K})\big) = \mathbb{K}^{m\times m}$
2. $\mathrm{Lie}\big(\mathrm{SL}_m(\mathbb{K})\big) = \{X \in \mathbb{K}^{m\times m} \mid \mathrm{tr}(X) = 0\}$

3. $\mathrm{Lie}\big(\mathrm{GT}_m(\mathbb{K})\big) = \{X \in \mathbb{K}^{m\times m} \mid X \text{ diagonal matrix}\}$
4. $\mathrm{Lie}\big(\mathrm{O}_m(\mathbb{K})\big) = \{X \in \mathbb{K}^{m\times m} \mid X^\mathsf{T} = -X\}$, the space of skew-symmetric matrices. Note that $\mathrm{Lie}\big(\mathrm{O}_m(\mathbb{K})\big) = \mathrm{Lie}\big(\mathrm{SO}_m(\mathbb{K})\big)$ as any skew symmetric matrix has trace zero.
5. $\mathrm{Lie}(\mathrm{U}_m) = \{X \in \mathbb{C}^{m\times m} \mid X^\dagger = -X\} = \mathfrak{i}\,\mathrm{Sym}_m(\mathbb{C})$, the space of skew-Hermitian matrices. Here, $\mathfrak{i} \in \mathbb{C}$ denotes the imaginary unit and $\mathrm{Sym}_m(\mathbb{C})$ the space of $m \times m$ Hermitian matrices.
6. Consider $\mathrm{GT}_m(\mathbb{C}) \cap \mathrm{U}_m$. Using Proposition 1.2.8(c) we obtain that
$$\mathrm{Lie}(\mathrm{GT}_m(\mathbb{C}) \cap \mathrm{U}_m) = \big\{\mathfrak{i}\,\mathrm{diag}(x) \mid x = (x_1, \ldots, x_m) \in \mathbb{R}^m\big\}.$$
Hence, we can identify $\mathfrak{i}\,\mathrm{Lie}(\mathrm{GT}_m(\mathbb{C}) \cap \mathrm{U}_m) = \{\mathrm{diag}(x) \mid x \in \mathbb{R}^m\}$ with $\mathbb{R}^m$. Note that under this identification the Frobenius norm becomes the usual Euclidean norm on $\mathbb{R}^m$.
7. Set $T_K := \mathrm{ST}_m(\mathbb{C}) \cap \mathrm{U}_m$. Similarly to (6) we obtain that
$$\mathrm{Lie}(T_K) = \big\{\mathfrak{i}\,\mathrm{diag}(x) \mid x = (x_1, \ldots, x_m) \in \mathbb{R}^m,\, x_+ = x_1 + \ldots + x_m = 0\big\}.$$
Thus, we can identify $\mathfrak{i}\,\mathrm{Lie}(T_K) = \{\mathrm{diag}(x) \mid x \in \mathbb{R}^m, x_+ = \langle \mathbb{1}_m, x\rangle = 0\}$ with $\mathbb{1}_m^\perp$, the orthogonal complement of the all-ones vector $\mathbb{1}_m$ in $\mathbb{R}^m$. ♢

Given real Lie algebras $\mathfrak{g}$ and $\mathfrak{h}$, a *morphism of Lie algebras* is a $\mathbb{R}$-linear map $\Pi\colon \mathfrak{g} \to \mathfrak{h}$ such that $\Pi([X,Y]) = [\Pi(X), \Pi(Y)]$ holds for all $X, Y \in \mathfrak{g}$. Given a morphism of matrix Lie groups one naturally obtains a morphism of the respective Lie algebras by considering the differential at the identity.

Theorem 1.2.10 ([Hal15, Theorem 3.28])**.** *Let G and H be matrix Lie groups, and $\pi\colon G \to H$ a Lie group morphism. Then there exists a unique $\mathbb{R}$-linear map $\Pi\colon \mathrm{Lie}(G) \to \mathrm{Lie}(H)$ such that $\pi(e^X) = e^{\Pi(X)}$ holds for all $X \in \mathrm{Lie}(G)$. The map Π has the following additional properties:*

1. $\Pi(gXg^{-1}) = \pi(g)\Pi(X)\pi(g)^{-1}$ *for all $X \in \mathrm{Lie}(G)$, $g \in G$.*
2. $\Pi([X,Y]) = [\Pi(X), \Pi(Y)]$ *for all $X, Y \in \mathrm{Lie}(G)$.*
3. $\Pi(X) = \frac{d}{dt}\big|_{t=0} \pi(e^{tX})$ *for all $X \in \mathrm{Lie}(G)$.*

Self-adjoint Groups

We review Zariski closed self-adjoint groups. This is motivated by the fact that reductive subgroups of $\mathrm{GL}_m(\mathbb{K})$ are, up to conjugation, the Zariski closed self-adjoint subgroups; compare Theorem 1.3.10 below. At the end, we present important connections to Riemannian geometry.

Definition 1.2.11 (Self-adjoint Group)**.** A subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is *self-adjoint*, if for all $g \in G$ one has $g^\dagger \in G$. (Note that $g^\dagger = g^\mathsf{T}$ if $\mathbb{K} = \mathbb{R}$.)

More generally, let V be a $\mathbb{K}$-vector space equipped with an inner product $\langle\cdot,\cdot\rangle$ (which is Hermitian if $\mathbb{K} = \mathbb{C}$). A subgroup $G \subseteq \mathrm{GL}(V)$ is called *self-adjoint* if for all $g \in G$ the adjoint g^* with respect to $\langle\cdot,\cdot\rangle$ is contained in G. Thus, $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is self-adjoint if it is self-adjoint with respect to the standard inner product on $\mathbb{K}^m$. ▲

Example 1.2.12. The following groups are Zariski closed and self-adjoint.

1. The groups $\mathrm{GL}_m(\mathbb{K}), \mathrm{SL}_m(\mathbb{K}), \mathrm{GT}_m(\mathbb{K}), \mathrm{ST}_m(\mathbb{K}), \mathrm{O}_m(\mathbb{K})$ and $\mathrm{SO}_m(\mathbb{K})$ are all Zariski closed and self-adjoint subgroups of $\mathrm{GL}_m(\mathbb{K})$.

2. The intersection $G \cap H$ of two Zariski closed self-adjoint subgroups $G, H \subseteq \mathrm{GL}_m(\mathbb{K})$ is Zariski closed and self-adjoint.

3. Let $G \subseteq \mathrm{GL}_{m_1}(\mathbb{K})$, $H \subseteq \mathrm{GL}_{m_2}(\mathbb{K})$ be Zariski closed and self-adjoint. Similar to Example 1.2.3 Item 4, the direct product $G \times H$ is a Zariski closed self-adjoint subgroup of $\mathrm{GL}_{m_1+m_2}(\mathbb{K})$ via block-diagonal embedding. ♢

Remark 1.2.13. For the following compare [Wal17, p. 39]. One can identify $\mathrm{GL}_m(\mathbb{C})$ canonically with $\mathrm{GL}_{2m}(\mathbb{R})$ via

$$\mathrm{GL}_m(\mathbb{C}) \to \mathrm{GL}_{2m}(\mathbb{R}), \quad g = a + \mathfrak{i}\, b \mapsto \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

where $a, b \in \mathbb{R}^{m\times m}$. Note that under this identification the Hermitian transpose becomes the transpose, and that the group of unitary matrices U_m is mapped to the group $\mathrm{O}_{2m}(\mathbb{R})$ of orthogonal matrices. Moreover, under the above identification any (Zariski closed) self-adjoint subgroup $G \subseteq \mathrm{GL}_m(\mathbb{C})$ can be viewed as a (Zariski closed) self-adjoint subgroup of $\mathrm{GL}_{2m}(\mathbb{R})$.

Note that Zariski closed self-adjoint subgroups $G \subseteq \mathrm{GL}_m(\mathbb{K})$ are called *symmetric* in [Wal17]. We refrain from using the latter term to avoid confusion with the usual symmetric groups consisting of permutations. ▽

In the following we deal with some important properties of Zariski closed self-adjoint subgroups. We denote by $\mathrm{Sym}_m(\mathbb{K}) := \{X \in \mathbb{K}^{m\times m} \mid X^\dagger = X\}$ the space of symmetric respectively Hermitian matrices. Recall that $\mathbb{K}^{m\times m}$ is equipped with the trace inner product, if not stated otherwise.

Proposition 1.2.14. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint subgroup. Set $K := \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$ and $\mathfrak{p} := \mathrm{Lie}(G) \cap \mathrm{Sym}_m(\mathbb{K})$. Then*

(a) K is a maximal compact subgroup of G.

(b) If $\mathbb{K} = \mathbb{C}$, then $T := (G \cap \mathrm{GT}_m(\mathbb{K}))^\circ$ is a maximal torus of G, and $T_K := T \cap K$ is a maximal compact torus of K.

(c) $\mathrm{Lie}(G) = \mathrm{Lie}(K) \oplus \mathfrak{p}$ is an orthogonal decomposition with respect to the Euclidean *inner product $(X, Y) \mapsto \mathrm{Re}(\mathrm{tr}(X^\dagger Y))$ on $\mathbb{K}^{m\times m}$.[3] If $\mathbb{K} = \mathbb{C}$ then $\mathfrak{p} = \mathfrak{i}\,\mathrm{Lie}(K)$.*

Proof. Part (a) is a consequence of [Wal17, Theorem 2.29] and part (b) follows from [Wal17, Theorem 2.21]. For (c), note that $\mathfrak{p}$ consists of symmetric (respectively Hermitian) matrices while $\mathrm{Lie}(K)$ consists of skew-symmetric (respectively skew-Hermitian) matrices. If $\mathbb{K} = \mathbb{C}$, then $\mathrm{Lie}(G) = \mathrm{Lie}(K) \oplus \mathfrak{i}\,\mathrm{Lie}(K)$ by [Wal17, Theorem 2.12] and $\mathfrak{i}\,\mathrm{Lie}(K)$ consists of Hermitian matrices. Hence, $\mathfrak{i}\,\mathrm{Lie}(K) = \mathrm{Lie}(G) \cap \mathrm{Sym}_m(\mathbb{K}) = \mathfrak{p}$. □

[3] Here Re denotes the real part. For $\mathbb{K} = \mathbb{R}$, this is the usual inner product on $\mathbb{R}^{m\times m}$. Over $\mathbb{C}$ we need to adjust as $\mathrm{tr}(X^\dagger Y) \in \mathfrak{i}\mathbb{R}$ for X Hermitian and Y skew-Hermitian.

Example 1.2.15. Let $G := \mathrm{GL}_m(\mathbb{K})$. Then $K := \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$ equals $\mathrm{O}_m(\mathbb{R})$ if $\mathbb{K} = \mathbb{R}$, and $K = \mathrm{U}_m$ if $\mathbb{K} = \mathbb{C}$. Moreover, $\mathrm{Lie}(K)$ is the set of skew-symmetric respectively skew-Hermitian matrices, compare Example 1.2.9, while $\mathfrak{p} = \mathrm{Sym}_m(\mathbb{K})$ is the set of symmetric respectively Hermitian matrices. So indeed, if $\mathbb{K} = \mathbb{C}$ then $\mathfrak{p} = \mathrm{Sym}_m(\mathbb{K}) = \mathfrak{i}\,\mathrm{Lie}(K)$. ♢

Next, we recall the polar decomposition [Hal15, Section 2.5]. We denote by $\mathrm{PD}_m(\mathbb{C})$ the cone of Hermitian positive definite matrices, and by $\mathrm{PD}_m(\mathbb{R})$ the cone of symmetric positive definite matrices. The map

$$\mathrm{Sym}_m(\mathbb{K}) = \{X \in \mathbb{K}^{m\times m} \mid X^\dagger = X\} \to \mathrm{PD}_m(\mathbb{K}), \quad X \to e^X$$

is a diffeomorphism. In particular, the *logarithm* $\log(\Psi) \in \mathrm{Sym}_m(\mathbb{K})$ is well-defined for all $\Psi \in \mathrm{PD}_m(\mathbb{K})$. For $G = \mathrm{GL}_m(\mathbb{K})$ set $K := \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$. Then the polar decomposition is given by the diffeomorphism

$$K \times \mathrm{Sym}_m(\mathbb{K}) \to \mathrm{GL}_m(\mathbb{K}), \quad (k, X) \mapsto ke^X.$$

In particular, any $g \in G$ can be uniquely written as $g = kp$, where $k \in K$ and $p \in \mathrm{PD}_m(\mathbb{K})$. The polar decomposition holds more generally for any Zariski closed self-adjoint subgroup.

Theorem 1.2.16 (Polar Decomposition, [Wal17, Theorems 2.12 and 2.16])**.** *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be Zariski closed and self-adjoint, $K = \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$ and $\mathfrak{p} = \mathrm{Lie}(G) \cap \mathrm{Sym}_m(\mathbb{K})$. Then*

$$K \times \mathfrak{p} \to G, \quad (k, X) \mapsto ke^X \tag{1.2}$$

is a diffeomorphism. In particular, any $g \in G$ can be uniquely written as $g = kp$, where $k \in K$ and $p \in P := G \cap \mathrm{PD}_m(\mathbb{K})$. Moreover, G is connected if and only if K is connected.

As an interesting consequence any (not necessarily Zariski closed) subgroup lying in between G° and G is self-adjoint.

Corollary 1.2.17. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be Zariski closed and self-adjoint. Consider a subgroup $H \subseteq G$ with $G^\circ \subseteq H$. Then H is self-adjoint and the polar decomposition can be carried out in H.*

Proof. Define K and $\mathfrak{p}$ as in Theorem 1.2.16 and consider $h \in H \subseteq G$. By Theorem 1.2.16, there exist $k \in K$ and $X \in \mathfrak{p}$ such that $h = k\exp(X)$. We have $\exp(X) \in G^\circ \subseteq H$ by Proposition 1.2.8(d) and hence $k = h\exp(X)^{-1} \in H$. We deduce $h^\dagger = \exp(X^\dagger)k^\dagger = \exp(X)k^{-1} \in H$. □

Now, we briefly recall some Riemannian geometry of $\mathrm{PD}_m(\mathbb{K})$; see [Bha07] or [BH99, Chapter II.10]. We denote by $\Psi^{1/2} \in \mathrm{PD}_m(\mathbb{K})$ (or by $\sqrt{\Psi}$) the *square root* of $\Psi \in \mathrm{PD}_m(\mathbb{K})$; that is the unique matrix in $\mathrm{PD}_m(\mathbb{K})$ whose square equals Ψ. Viewing $\mathrm{PD}_m(\mathbb{K})$ as an open real submanifold of $\mathrm{Sym}_m(\mathbb{K})$ one can define a Riemannian metric on $\mathrm{PD}_m(\mathbb{K})$ via

$$\langle X, Y\rangle_\Psi := \mathrm{tr}\left(\Psi^{-1}X\Psi^{-1}Y\right),$$

where X, Y are in the tangent space $T_\Psi \mathrm{PD}_m(\mathbb{K}) \cong \mathrm{Sym}_m(\mathbb{K})$ at Ψ. Note that

$$\langle X, X\rangle_{\mathrm{I}_m} = \|X\|^2 \quad \text{and} \quad \langle X, Y\rangle_\Psi = \left\langle \Psi^{-1/2} X \Psi^{-1/2}, \Psi^{-1/2} Y \Psi^{-1/2} \right\rangle_{\mathrm{I}_m}.$$

For $\Psi, \Theta \in \mathrm{PD}_m(\mathbb{K})$, the Riemannian manifold $\mathrm{PD}_m(\mathbb{K})$ has a unique *geodesic line* with $\gamma(0) = \Psi$ and $\gamma(1) = \Theta$:

$$\gamma\colon \mathbb{R} \to \mathrm{PD}_m(\mathbb{K}), \quad t \mapsto \Psi^{1/2} e^{tX} \Psi^{1/2} \tag{1.3}$$

where $X := \log(\Psi^{-1/2}\Theta\Psi^{-1/2})$. We call $\gamma([0,1])$ the *geodesic segment* between Ψ and Θ.[4] Consequently, the induced distance function on $\mathrm{PD}_m(\mathbb{K})$ is

$$d(\Psi, \Theta) = \left\| \log\left(\Psi^{-1/2}\Theta\Psi^{-1/2}\right) \right\|.$$

In particular, we have $d(\mathrm{I}_m, \Psi) = \|\log(\Psi)\|$.

A subset $B \subseteq \mathrm{PD}_m(\mathbb{K})$ is called *geodesically convex*, if it contains the geodesic segment between any two point in B. We say an embedded submanifold $M \subseteq \mathrm{PD}_m(\mathbb{K})$ is *totally geodesic*[5], if any geodesic *line* of $\mathrm{PD}_m(\mathbb{K})$ that intersects M in two points is entirely contained in M.

Note that $\mathrm{GL}_m(\mathbb{K})$ acts transitively (from the right) on $\mathrm{PD}_m(\mathbb{K})$ via $(\Psi, g) \mapsto g^\dagger \Psi g$ and the stabilizer of I_m is $K = \{g \in \mathrm{GL}_m(\mathbb{K}) \mid g^\dagger g = \mathrm{I}_m\}$.[6] Furthermore, $\mathrm{PD}_m(\mathbb{K}) = \{g^\dagger g \mid g \in \mathrm{GL}_m(\mathbb{K})\} = \mathrm{GL}_m(\mathbb{K}) \cdot \mathrm{I}_m$. From this one can deduce that the Riemannian manifold G/K is isometric to $\mathrm{PD}_m(\mathbb{K})$. More generally, we have the following.[7]

Theorem 1.2.18 ([BH99, Theorem II.10.58])**.** *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint subgroup. Set $P := G \cap \mathrm{PD}_m(\mathbb{K})$, $K := \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$ and $\mathfrak{p} := \mathrm{Lie}(G) \cap \mathrm{Sym}_m(\mathbb{K})$. Then*

(i) $P = \exp(\mathfrak{p}) = \{g^\dagger g \mid g \in G\}$.

(ii) P *is a totally geodesic submanifold of* $\mathrm{PD}_m(\mathbb{K})$ *and diffeomorphic to* G/K.

(iii) P *is a CAT(0) symmetric space.*[8]

Conversely, if P' is a totally geodesic submanifold of $\mathrm{PD}_m(\mathbb{K})$ with $\mathrm{I}_m \in P'$, then $G := \{g \in \mathrm{GL}_m(\mathbb{K}) \mid g^\dagger P' g = P'\}$ is a Euclidean closed self-adjoint subgroup of $\mathrm{GL}_m(\mathbb{K})$ such that $P' = G \cap \mathrm{PD}_m(\mathbb{K})$.

Remark 1.2.19. For consulting [BH99] we point out the following. In [BH99] a reductive subgroup $G \subseteq \mathrm{GL}_m(\mathbb{R})$ is a Euclidean closed self-adjoint subgroup in our sense, see [BH99, Definition 10.56]. However, the assumptions in [BH99, Theorem II.10.58] are different from ours. First, [BH99, Theorem II.10.58] is only stated over $\mathbb{R}$, but the complex case is actually a special case by Remark 1.2.13, also see [BH99, Example II.10.57 (2)]. Second, the assumptions on G are slightly different, but this is justified by [BH99, Lemma II.10.59]. ▽

[4] One should think of the geodesic segment as a curve representing the shortest path between Ψ and Θ.

[5] also called *geodesically complete*

[6] Of course, one can also consider the left action $g \cdot \Psi = g\Psi g^\dagger$. However, the right action appears naturally in Part III on algebraic statistics.

[7] I thank Harold Nieuwboer for pointing out the reference [BH99, Theorem II.10.58].

[8] We do not give a definition but point out that such spaces have a rigid geometry that is very useful for optimization techniques.

An important application of Theorem 1.2.18 is that norm minimization under G is a geodesically convex optimization problem as follows.

Definition 1.2.20. Let $M \subseteq \mathrm{PD}_m(\mathbb{K})$ be a totally geodesic embedded submanifold, and $f\colon M \to \mathbb{R}$ a smooth map. We say f is *geodesically convex*, if it is convex along all geodesics contained in M. ▲

Example 1.2.21. The following two functions are geodesically convex on $\mathrm{PD}_m(\mathbb{K})$, and hence on all totally geodesic submanifolds of $\mathrm{PD}_m(\mathbb{K})$.

1. For a fixed vector $v \in \mathbb{K}^m$, consider

$$f_v\colon \mathrm{PD}_m(\mathbb{K}) \to \mathbb{R},\ \Psi \mapsto \langle v, \Psi v\rangle = \|\Psi^{1/2}v\|^2.$$

   Then $F_v := \log f_v$ is geodesically convex [BFG+19, Proposition 3.13], and hence also f_v is.[9] Thus, for fixed $v \in \mathbb{K}^m$ and $G \subseteq \mathrm{GL}_m(\mathbb{K})$ a Zariski closed self-adjoint subgroup, the optimization problems

$$\inf_{g\in G} \|gv\|^2 = \inf_{g\in G} \langle v, g^\dagger g v\rangle \quad \text{and} \quad \inf_{g\in G} \log\left(\|gv\|^2\right)$$

   are geodesically convex on $P = \{g^\dagger g \mid g \in G\}$. This observation is important in Section 2.2 and Part II.

2. The two functions $\mathrm{PD}_m(\mathbb{K}) \to \mathbb{R},\ \Psi \mapsto \pm\log\det(\Psi)$ are geodesically convex. Indeed, for a geodesic line γ as in (1.3) consider

$$h(t) := \pm\log\det(\gamma(t)) = \pm\log\det(\Psi) \pm \log\det\left(e^{tX}\right).$$

   Using $\det(\exp(tX)) = \exp(t\,\mathrm{tr}(X))$ one computes that $h'(t) = \pm\,\mathrm{tr}(X)$ and $h''(t) = 0$ for all $t \in \mathbb{R}$. The latter yields that h is convex. ♢

1.3 Representation Theory

We recall required knowledge on representation theory. First, we present examples of representations that are studied in this thesis. Afterwards, we connect reductive groups to Zariski closed self-adjoint subgroups, which justifies our restriction to the latter case. Finally, we review weights and roots. Further material on representation theory is provided, e.g., by [Bor06; FH91; Hal15; OV90; Pro07].

Basic Definitions and Examples

We briefly recall some standard terminology on group representations. Consider a group G (not necessarily endowed with further structure) and let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

A *representation* of G on the $\mathbb{K}$-vector space V is a group morphism $\pi\colon G \to \mathrm{GL}(V)$. Equivalently, G acts $\mathbb{K}$-linearly on V and we write $g \cdot v := \pi(g)(v)$, where $g \in G$ and $v \in V$. A representation π is called *faithful* if it is injective. If G has further structure, then one usually requires additional properties on π: a

[9] A logarithmically convex function is convex.

representation of a matrix Lie group is additionally assumed to be a Lie group morphism. If G is an algebraic group over $\mathbb{K}$, one considers rational representations as in Definition 1.1.9. Note, that if we view an algebraic group over $\mathbb{K}$ as a (matrix) Lie group, then any rational representation is smooth and hence a representation of the (matrix) Lie group.

If $\varrho\colon G \to \mathrm{GL}(W)$ is a representation on the $\mathbb{K}$-vector space W, then the *direct sum* of π and ϱ is defined as

$$\pi \oplus \varrho\colon G \to \mathrm{GL}(V \oplus W), \quad g \mapsto \big((v,w) \mapsto (\pi(v), \varrho(w))\big).$$

The n-fold direct sum of π is denoted $\pi^{\oplus n}$. A *morphism of representations* is a $\mathbb{K}$-linear map $f\colon V \to W$ that is G-*equivariant*, i.e., $f\big(\pi(g)(v)\big) = \varrho(g)\big(f(v)\big)$ holds for all $v \in V$ and all $g \in G$. The representations π and ϱ are *isomorphic*, if there exists a bijective morphism of representations between them.[10]

A *subrepresentation* is a $\mathbb{K}$-vector subspace $W \subseteq V$ that is invariant under G, i.e., $g{\cdot}u \in U$ for all $g \in G$ and all $u \in U$. A representation $\pi\colon G \to \mathrm{GL}(V)$ is called *simple*[11] if its only subrepresentations are $\{0\}$ and V. It is called *semisimple*[12] if it is a direct sum of simple representations.

Now, assume $G \subseteq \mathrm{GL}_m(\mathbb{C})$ is a matrix Lie group and $\pi\colon G \to \mathrm{GL}(V)$ a representation of G. Then we obtain a Lie algebra morphism $\Pi\colon \mathrm{Lie}(G) \to \mathrm{End}(V)$ via the differential, compare Theorem 1.2.10. Such a morphism Π is called a *representation of the Lie algebra* $\mathrm{Lie}(G)$. One can define the above concepts similarly for representations of Lie algebras, but this is not needed here.

Of particular importance in representation theory is the adjoint representation.

Example 1.3.1 (Adjoint Representation)**.** Let G be a matrix Lie group. The *adjoint representation* of G is

$$\mathrm{Ad}\colon G \to \mathrm{GL}(\mathrm{Lie}(G)), \quad g \mapsto (X \mapsto gXg^{-1}).$$

It induces via the differential the *adjoint representation* of $\mathrm{Lie}(G)$

$$\mathrm{ad}\colon \mathrm{Lie}(G) \mapsto \mathrm{End}(\mathrm{Lie}(G)), \quad X \mapsto (Y \mapsto [X,Y]),$$

compare [Hal15, Proposition 3.34]. ♢

Next, we present several important examples of group representations, that are studied in this thesis. We point out that these are all rational representations defined over $\mathbb{K}$ of a reductive group over $\mathbb{K}$. We present these representations in terms of their $\mathbb{K}$-linear algebraic action of G on V. Moreover, we note that one can, of course, replace SL always with GL in these examples. However, the actions of (products of) SL are usually the ones we are interested in this thesis, also compare Example 1.4.3 below.

[10] Note that the inverse of such a morphism is automatically $\mathbb{K}$-linear and G-equivariant.
[11] also called *irreducible*
[12] also called a *completely reducible* representation

Example 1.3.2 (Left Multiplication)**.** The group $G = \mathrm{SL}_m(\mathbb{K})$ acts algebraically on $\mathbb{K}^m$ via left multiplication, i.e., $g \cdot v = gv$ for $g \in G$ and $v \in \mathbb{K}^m$. Note that the n-fold direct sum of this representation is isomorphic (via $(\mathbb{K}^m)^{\oplus n} \cong \mathbb{K}^{m\times n}$) to the left multiplication of G on $\mathbb{K}^{m\times n}$: $g \cdot Y = gY$, where $Y \in \mathbb{K}^{m\times n}$. ♢

Example 1.3.3 (Left-right Action)**.** The *left-right action* of $G = \mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ on $V = (\mathbb{K}^{m_1\times m_2})^n$ is given by

$$g \cdot Y := \left(g_1 Y_1 g_2^{\mathsf{T}}, \ldots, g_1 Y_n g_2^{\mathsf{T}}\right),$$

where $g = (g_1, g_2) \in G$ and $Y = (Y_1, \ldots, Y_n) \in V$. We stress that the transpose g_2^{T} is also considered for $\mathbb{K} = \mathbb{C}$ to ensure an *algebraic* action. Using the Hermitian transpose $g_2^{\dagger}$ would involve complex conjugation, which prevents the action $G \times V \to V$ to be a polynomial function in the coordinates of g and Y. ♢

It is convenient to use the Kronecker product for the upcoming example.

Definition 1.3.4 (Kronecker product of matrices)**.** The Kronecker product $A \otimes B$ of two matrices $A \in \mathbb{K}^{m\times n}$ and $B \in \mathbb{K}^{p\times q}$ is a matrix of size $mp \times nq$. It is defined as the following $m \times n$ block matrix, where each block has size $p \times q$,

$$A \otimes B := \begin{pmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{pmatrix} \in \mathbb{K}^{(mp)\times(nq)}.$$

We index its rows by (i, k) where $i \in [m]$ and $k \in [p]$, and its columns by (j, l), where $j \in [n]$ and $l \in [q]$. Note that by definition the rows are ordered as follows: $(i_1, k_1) < (i_2, k_2)$ if and only if $i_1 < i_2$, or ($i_1 = i_2$ and $k_1 < k_2$). The same applies to the columns. The entry of $A \otimes B$ at index $((i, k), (j, l))$ is $A_{ij}B_{kl}$.

If one views A and B as linear maps, then the Kronecker product $A \otimes B$ is a representing matrix[13] for the tensor product of these linear maps. ▲

We are now able to introduce a natural action on tensors. It contains Examples 1.3.2 and 1.3.3 as special cases.

Example 1.3.5 (Tensor Scaling)**.** The group $G = \mathrm{SL}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{SL}_{m_d}(\mathbb{K})$ acts algebraically on $V = \mathbb{K}^{m_1} \otimes \cdots \otimes \mathbb{K}^{m_d}$ by $\mathbb{K}$-linear extension of

$$(g_1, \ldots, g_d) \cdot (v_1 \otimes \cdots \otimes v_d) = g_1(v_1) \otimes \cdots \otimes g_d(v_d),$$

where $g_i \in \mathrm{SL}_{m_i}(\mathbb{K})$ and $v_i \in \mathbb{K}^{m_i}$. There is a unique way to identify $V \cong \mathbb{K}^{m_1 \cdots m_d}$ such that the tensor scaling action corresponds to the representation

$$\pi_{m_1 \otimes \cdots \otimes m_d} \colon G \to \mathrm{GL}_{m_1 \cdots m_d}(\mathbb{K}), \quad (g_1, \ldots, g_d) \mapsto g_1 \otimes \cdots \otimes g_d,$$

where $g_1 \otimes \cdots \otimes g_d$ denotes the Kronecker product as introduced in Definition 1.3.4. Of course, the n-fold direct sum $\pi^{\oplus n}_{m_1 \otimes \cdots \otimes m_d}$ corresponds to the simultaneous action of G on n many tensors.

[13] With respect to certain ordered bases on $\mathbb{K}^m \otimes \mathbb{K}^p$ and $\mathbb{K}^n \otimes \mathbb{K}^q$.

We note that for $d = 1$ this is just the action by left multiplication. Moreover, if $d = 2$ then $\pi_{m_1 \otimes m_2}^{\oplus n}$ is isomorphic to the left-right action from Example 1.3.3. This will be explained in Example 9.1.6.

We speak of the *tensor scaling action* if $d \geq 3$ and of the *operator scaling action* if $d = 2$. When restricting to the torus $T = \mathrm{ST}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{ST}_{m_d}(\mathbb{K})$, we refer to this action as *array scaling action* if $d \geq 3$ and as *matrix scaling action* if $d = 2$. Finally, if $m = m_1 = \ldots = m_d$ we set $\pi_{m,d} := \pi_{m \otimes \cdots \otimes m}$. ♢

For the last example we first need to introduce quivers and their representations. Detailed information on quiver representations can be found in [DW17].

Definition 1.3.6 ([DW17, Definition 1.1.1])**.** A *quiver* $Q = (Q_0, Q_1, h, t)$ consists of a finite set Q_0 of vertices, a finite set Q_1 of arrows, and two functions $h, t \colon Q_1 \to Q_0$. For $a \in Q_1$, $h(a)$ is the *head* of a and $t(a)$ is the *tail* of a, i.e.,

$$t(a) \xrightarrow{a} h(a)\,.$$

We stress that multiple arrows and multiple loops are allowed. ▲

Definition 1.3.7 (Quiver Representation)**.** Let Q be a quiver with $Q_0 = [d]$. A *representation* of Q is an assignment of a vector space $\mathbb{K}^{m_i}$ to each vertex $i \in [d]$ and a matrix $Y_a \in \mathbb{K}^{m_{h(a)} \times m_{t(a)}}$ to each arrow $a \in Q_1$. The matrix Y_a represents a $\mathbb{K}$-linear map $\mathbb{K}^{m_{t(a)}} \to \mathbb{K}^{m_{h(a)}}$. All information on the vertices is encoded by the *dimension vector* $\alpha = (m_1, \ldots, m_d)$. The vector space

$$\mathscr{R}(Q, \alpha) := \bigoplus_{a \in Q_1} \mathbb{K}^{m_{h(a)} \times m_{t(a)}}$$

is called the *representation space* of α-dimensional representations of Q. ▲

Example 1.3.8 (Action on Representations of a Quiver)**.** Let Q be a quiver with vertex set $Q_0 = [d]$ and fix a dimension vector $\alpha = (m_1, \ldots, m_d)$. Set

$$\mathrm{GL}_\alpha(\mathbb{K}) := \mathrm{GL}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{GL}_{m_d}(\mathbb{K}) \quad \text{and} \quad \mathrm{SL}_\alpha(\mathbb{K}) := \mathrm{SL}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{SL}_{m_d}(\mathbb{K}).$$

$\mathrm{GL}_\alpha(\mathbb{K})$ acts algebraically via base change on the representation space $\mathscr{R}(Q, \alpha)$:

$$g \cdot (Y_a)_{a \in Q_1} := \left(g_{h(a)}\, Y_a\, g_{t(a)}^{-1}\right)_{a \in Q_1},$$

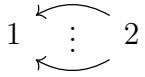
where $g \in \mathrm{GL}_\alpha$ and $(Y_a)_{a \in Q_1} \in \mathscr{R}(Q, \alpha)$. We call this action the GL-*action on the quiver* Q with dimension vector α.[14] If we restrict the action to the subgroup $\mathrm{SL}_\alpha(\mathbb{K})$ then we speak of the SL-*action on the quiver* Q with dimension vector α.

For illustration we consider two examples. First, let Q be the one loop quiver

[14] This may be a non-standard name.

and $\alpha = (m)$. Then $\mathrm{GL}_\alpha(\mathbb{K}) = \mathrm{GL}_m(\mathbb{K})$ and $\mathscr{R}(Q,\alpha) = \mathbb{K}^{m\times m}$. As head and tail of the arrow in Q are the same, we see that the GL-action on the one loop quiver is the conjugation action. If ϱ is the corresponding representation, then $\varrho^{\oplus n}$ is the simultaneous conjugation of $\mathrm{GL}_m(\mathbb{K})$ on n-many matrices. Note that the latter is the GL action on the quiver with one vertex and n loops.

Second, let Q be the *n-Kronecker quiver* with two vertices and n arrows:



and $\alpha = (m_1, m_2)$. Then $\mathrm{GL}_\alpha(\mathbb{K}) = \mathrm{GL}_{m_1} \times \mathrm{GL}_{m_2}(\mathbb{K})$ and $\mathscr{R}(Q,\alpha) = (\mathbb{K}^{m_1\times m_2})^n$. Since vertex 1 is the head and vertex 2 is the tail of all arrows, the GL-action on Q is given by

$$g \cdot Y := \big(g_1 Y_1 g_2^{-1}, \ldots, g_1 Y_n g_2^{-1}\big),$$

where $g = (g_1, g_2) \in \mathrm{GL}_\alpha(\mathbb{K})$ and $Y = (Y_1, \ldots, Y_n) \in (\mathbb{K}^{m_1\times m_2})^n$. One verifies that pre-composition with the automorphism $(g_1, g_2) \mapsto (g_1, g_2^{-\mathsf{T}})$ of $\mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K})$ transforms the GL-action on the n-Kronecker quiver into the GL-left-right action (Example 1.3.3), and vice versa. The same applies to the respective SL-actions, i.e., when restricting to $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$. ♢

Self-Adjoint and reductive groups

We connect the important concepts of self-adjoint groups and reductive groups to each other. Remember the definitions of semisimple representations from the beginning of this Section 1.3.

A linear algebraic group G is called *linearly reductive*, if all its rational representations are semisimple. An important property of reductive groups in characteristic zero is that their rational representations are semisimple (also called completely reducible). In fact, in characteristic zero reductive and linearly reductive are equivalent notions.

Theorem 1.3.9 ([Mil17, Theorem 22.42 and Corollary 22.43])**.**
Let G be a linear algebraic group over $\mathbb{K}$. Then G is reductive if and only if it admits a faithful *semisimple rational representation. Moreover, G is reductive if and only if all finite-dimensional representations of G are semisimple.*

Combining the latter theorem with results from [Mos55] links self-adjoint and reductive groups.

Theorem 1.3.10 ([Mos55, Theorems 7.1 and 7.2])**.**
Let V be a finite dimensional $\mathbb{K}$-vector space and let $G \subseteq \mathrm{GL}(V)$ be an algebraic subgroup over $\mathbb{K}$. Then G is reductive if and only if G is self-adjoint with respect to some inner product on V. Thus, if $V = \mathbb{K}^m$ then $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is reductive if and only if there exists some $h \in \mathrm{GL}_m(\mathbb{K})$ such that hGh^{-1} is self-adjoint (with respect to the standard inner product).

As a consequence of the preceding theorem, the reductive subgroups of $\mathrm{GL}_m(\mathbb{K})$ are, up to conjugation, the Zariski closed self-adjoint subgroups.

Weights and Roots

We present necessary background on weights and roots. These concepts are only needed in the complex case and mainly used in Part II. Thus, we restrict to $\mathbb{K} = \mathbb{C}$ and for an easier comparison we follow the conventions in [BFG+19, Section 2]. For further information we refer to [FH91; GW09; Hal15; Kna96; Pro07] and for a treatment over the reals to [Bor06; OV90].

Thanks to Theorem 1.3.10 we may, for the sake of concreteness, restrict to Zariski closed self-adjoint subgroups when working with reductive groups. Our setting for studying weights and roots is as follows.

Setting 1.3.11. Let $G \subseteq \mathrm{GL}_N(\mathbb{C})$ be a Zariski closed self-adjoint subgroup. Then $K := \{g \in G \mid g^\dagger g = \mathrm{I}_N\}$ is a maximal compact group of G, see Proposition 1.2.14(a). Moreover, $T := (G \cap \mathrm{GT}_N(\mathbb{C}))^\circ$ is a maximal torus of G and $T_K := T \cap K$ is a maximal compact torus in K, Proposition 1.2.14(b). The $\mathbb{R}$-space $\mathfrak{i}\,\mathrm{Lie}(T_K)$ lies in $\mathfrak{i}\,\mathrm{Lie}(\mathrm{GT}_N(\mathbb{C}) \cap \mathrm{U}_N)$ which can be identified with $\mathbb{R}^N$, compare Example 1.2.9 Item 6.

Often, we study the concrete case where $G := \mathrm{SL}_m(\mathbb{C})^d$ is block-diagonally embedded in $\mathrm{GL}_{dm}(\mathbb{C})$ ($N = dm$). In that case $K = (\mathrm{SU}_m)^d$, $T = \mathrm{ST}_m(\mathbb{C})^d$ and $T_K = T \cap K$, which are as well block-diagonally embedded into $\mathrm{GL}_{dm}(\mathbb{C})$. Similarly, their Lie algebras are block-diagonally embedded into $\mathbb{C}^{dm\times dm}$. Considering Example 1.2.9 Item 7, we frequently use the identification

$$\mathfrak{i}\,\mathrm{Lie}(T_K) \cong (\mathbb{1}_m^\perp)^d \subseteq (\mathbb{R}^m)^d,$$

where $\mathbb{1}_m^\perp$ is the orthogonal complement of $\mathbb{1}_m$ in $\mathbb{R}^m$. ▲

Definition 1.3.12 (Weights and Weight Spaces)**.** Consider the Setting 1.3.11. Let $\pi\colon G \to \mathrm{GL}(V)$ be a complex rational representation with corresponding Lie algebra representation by $\Pi\colon \mathrm{Lie}(G) \to \mathrm{End}(V)$, compare Theorem 1.2.10.

We call $\omega \in \mathfrak{i}\,\mathrm{Lie}(T_K)$ a *weight* of π (with respect to the maximal torus T) if there exists a *non-zero* $v_\omega \in V$ such that

$$\forall X \in \mathrm{Lie}(T)\colon \quad \pi\left(e^X\right) v_\omega = e^{\mathrm{tr}(X\omega)} v_\omega$$

or, equivalently (see Theorem 1.2.10),

$$\forall X \in \mathrm{Lie}(T)\colon \quad \Pi\left(X\right) v_\omega = \mathrm{tr}(X\omega) v_\omega\,.$$

We say v_ω is a *weight vector* for weight ω. The *weight space* V_ω contains all weight vectors of ω and the zero vector. We denote by $\Omega(\pi)$ the set of weights of π. ▲

Remark 1.3.13. The set of possible weights forms a lattice which is isomorphic to the character group $\mathfrak{X}(T)$; compare Proposition 2.1.3 and Theorem 3.1.16 from [GW09] with each other. Indeed, [GW09, Proposition 2.1.3] follows the algebraic geometric point of view and defines weights via characters. The Lie group/Lie algebra approach in Definition 1.3.12, which equals the approach in [GW09, Theorem 3.1.16], identifies the characters as points in $\mathfrak{i}\,\mathrm{Lie}(T_K)$.

For example, $\mathfrak{X}(\mathrm{GT}_m(\mathbb{C})) = \mathbb{Z}^m \subseteq \mathbb{R}^m \cong \mathfrak{i}\,\mathrm{Lie}(\mathrm{GT}_m(\mathbb{C}) \cap \mathrm{U}_m)$. In the case $T = \mathrm{ST}_m(\mathbb{C})$ each character in $\mathfrak{X}(T) = \mathbb{Z}^m/\mathbb{Z}\mathbb{1}_m$ is identified via

$$\mathfrak{X}(T_K) \to \mathbb{1}_m^{\perp} \cong \mathfrak{i}\,\mathrm{Lie}(T_K), \quad (\lambda_1, \dots, \lambda_m) \mapsto (\lambda_1, \dots, \lambda_m) - \frac{\lambda_+}{m}\mathbb{1}_m$$

with a rational point in $\mathfrak{i}\,\mathrm{Lie}(T_K)$; also compare Example 1.3.17 below. ▽

We have the following important decomposition of V.

Theorem 1.3.14 (Weight Space Decomposition, [GW09, Theorem 3.1.16][15])**.** *Consider Setting 1.3.11 and let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation. The weight spaces V_ω of V with respect to the torus T decompose V:*

$$V = \bigoplus_{\omega \in \Omega(\pi)} V_\omega\,. \tag{1.4}$$

In particular, the set of weights $\Omega(\pi)$ is finite.

Remark 1.3.15. Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation with weight space decomposition as in (1.4). Then its n-fold direct sum $\pi^{\oplus n}\colon G \to \mathrm{GL}(V^{\oplus n})$ has the weight space decomposition $V^{\oplus n} = \bigoplus_{\omega\in\Omega(\pi)} V_\omega^{\oplus n}$. In particular, we see that $\Omega(\pi) = \Omega(\pi^{\oplus n})$ ▽

Next, we give the set of weights for several rational representations.

Example 1.3.16 (General Action of $\mathrm{GT}_d(\mathbb{C})$)**.** In the following we discuss all *rational* representations of $\mathrm{GT}_d(\mathbb{C})$ up to isomorphism. The notation is adjusted to the one used in Chapter 7.

If $\pi\colon \mathrm{GT}_d(\mathbb{C}) \to \mathrm{GL}(V)$ is a rational representation, then we can identify $V \cong \mathbb{C}^m$ such that the canonical unit vectors e_j, $j \in [m]$ are weight vectors. Let $(a_{1j}, \dots, a_{dj}) \in \mathbb{Z}^d$ be the weight with weight vector e_j. Then $t = \mathrm{diag}(t_1, \dots, t_d) \in \mathrm{GT}_d(\mathbb{C})$ acts on the coordinates $v \in \mathbb{C}^m$ via $v_j \mapsto t_1^{a_{1j}} \cdots t_d^{a_{dj}} v_j$. That is, t acts on v by left-multiplication with the diagonal matrix

$$\begin{pmatrix} t_1^{a_{11}} t_2^{a_{21}} \cdots t_d^{a_{d1}} & & & \\ & t_1^{a_{12}} t_2^{a_{22}} \cdots t_d^{a_{d2}} & & \\ & & \ddots & \\ & & & t_1^{a_{1m}} t_2^{a_{2m}} \cdots t_d^{a_{dm}} \end{pmatrix}. \tag{1.5}$$

We can encode this action uniquely by the *weight matrix* $A = (a_{ij}) \in \mathbb{Z}^{d\times m}$, which contains the weights as columns. Of course, any such matrix A defines an algebraic action via (1.5). Thus, rational representations of $\mathrm{GT}_d(\mathbb{C})$ on $\mathbb{C}^m$ are in one-to-one correspondence with their weight matrix A.

For us, a *linearization* via $b \in \mathbb{Z}^m$ of the above action shifts all weights by the vector $-b$.[16] That is, $t \in \mathrm{GT}_d(\mathbb{C})$ acts on $v \in \mathbb{C}^m$ via

$$v_j \mapsto t_1^{a_{1j}-b_1} \cdots t_d^{a_{dj}-b_d} v_j\,. \tag{1.6}$$

[15] Via rational characters it is [GW09, Proposition 2.1.3]. Further references are [Mil17, Theorem 12.12], [OV90, p. 141], [Spr98, Theorem 3.2.3].

[16] Linearizations are a concept from Geometric Invariant Theory [Dol03, Chapter 7]. In our specific situation the general concept agrees with the definition of linearization presented here, see [AKRS21b, Remark 3.3].

We refer to this action as the *action of* $\mathrm{GT}_d(\mathbb{C})$ *given by matrix* A *with linearization* b. Of course, the action in (1.6) is again encoded by a weight matrix, namely $A - \mathbb{1}_m^\mathsf{T} \otimes b = A - (b, \dots, b) \in \mathbb{Z}^{d \times m}$. However, it is instructive to work with linearizations in Chapter 7. There, the matrix A will encode a *statistical model*[17] and b is a vector that depends on the *observed data* and the matrix A. ♢

Example 1.3.17 (Left Multiplication, [FR21, Example B.2])**.** Consider the rational representation $\pi\colon \mathrm{SL}_m(\mathbb{C}) \to \mathrm{GL}_m(\mathbb{C}), g \mapsto g$, which is the action of $G = \mathrm{SL}_m(\mathbb{C})$ on $\mathbb{C}^m$ by left multiplication. For $i \in [m]$, we set

$$\epsilon_i := e_i - \frac{1}{m}\mathbb{1}_m \in \mathbb{1}_m^\perp \subseteq \mathbb{R}^m \tag{1.7}$$

where $e_i \in \mathbb{R}^m$ is the i^{th} canonical unit vector. Remember that we identify $\mathbb{1}_m^\perp \cong \mathfrak{i}\,\mathrm{Lie}(T_K)$. For all $X = \mathrm{diag}(x_1, \dots, x_m) \in \mathrm{Lie}(T)$ and all $i \in [m]$

$$\pi\left(\exp(X)\right) e_i = \exp(x_i) e_i \overset{(*)}{=} \exp\big(\operatorname{tr}(X\,\mathrm{diag}(\epsilon_i))\big) e_i\,,$$

where we used $x_1 + \ldots + x_m = 0$ in $(*)$. Thus, $\epsilon_i \in \mathbb{1}_m^\perp \cong \mathfrak{i}\,\mathrm{Lie}(T_K)$ is a weight of π with weight vector e_i. Since $\mathbb{C}^m = \bigoplus_i \mathbb{C}e_i$, we deduce $\Omega(\pi) = \{\epsilon_i \mid i \in [m]\}$.

We stress that, although $\pi\left(\exp(X)\right) e_i = \exp\big(\operatorname{tr}(X\,\mathrm{diag}(e_i))\big) e_i$ holds for all $X \in \mathrm{Lie}(T)$, we have $e_i \notin \mathbb{1}_m^\perp \cong \mathfrak{i}\,\mathrm{Lie}(T_K)$ and hence e_i cannot be a weight. ♢

Example 1.3.18 (Tensor Scaling)**.** Consider the tensor scaling action $\pi_{m,d}$, i.e., the natural action of $G = \mathrm{SL}_m(\mathbb{C})^d$ on $V = (\mathbb{C}^m)^{\otimes d}$ from Example 1.3.5. Using the argument from Example 1.3.17 in each tensor factor, one verifies that $(\epsilon_{i_1}, \dots, \epsilon_{i_d})$ is a weight of $\pi_{m,d}$ with weight vector $e_{i_1} \otimes \cdots \otimes e_{i_d}$. Therefore, we deduce

$$\Omega(\pi_{m,d}) = \big\{(\epsilon_{i_1}, \dots, \epsilon_{i_d}) \mid i_1, \dots, i_d \in [m]\big\} \subseteq (\mathbb{R}^m)^d,$$

since the $e_{i_1} \otimes \cdots \otimes e_{i_d}$ span V. ♢

Example 1.3.19 (Actions on Quivers)**.** Recall the SL-action on a quiver Q, i.e., the action of $\mathrm{SL}_\alpha(\mathbb{C})$ on $\mathscr{R}(Q, \alpha) = \bigoplus_{a \in Q_1} \mathbb{C}^{m_{h(a)} \times m_{t(a)}}$ from Example 1.3.8. Since $\mathscr{R}(Q, \alpha)$ is the direct sum of the matrix spaces associated to each arrow $a \in Q$, one can read off the weights for a general quiver by considering the two "building blocks". The latter refers to the two quivers

and

Let π be the action of $G = \mathrm{SL}_m(\mathbb{C})^2$ on the right quiver with dimension vector $\alpha = (m_1, m_2)$, i.e., $(g_1, g_2) \cdot Y = g_1 Y g_2^{-1}$ where $Y \in \mathbb{K}^{m_1 \times m_2}$. For $i \in [m_1]$ and $j \in$

[17] namely, the log-linear model $\mathcal{M}_A^{\ell\ell}$ defined by A

$[m_2]$, denote by $E_{i,j} \in \mathbb{C}^{m_1 \times m_2}$ the matrix with entry one at position (i,j) and all other entries zero. Then for all $i \in [m_1]$, $j \in [m_2]$ and all $X = \operatorname{diag}(x,y) \in \operatorname{Lie}(T)$

$$\begin{aligned}\exp(X) \cdot E_{i,j} &= \exp(x_i - y_j) E_{i,j} \overset{(*)}{=} \exp\big(\langle x, \epsilon_i\rangle - \langle y, \epsilon_j\rangle\big) E_{i,j} \\ &= \exp\big(\operatorname{tr}(X \operatorname{diag}(\epsilon_i, -\epsilon_j))\big) E_{i,j},\end{aligned}$$

where we used in $(*)$ that $x_+ = y_+ = 0$ (i.e., that $X \in \operatorname{Lie}(T)$).[18] Therefore, $(\epsilon_i, -\epsilon_j)$ is a weight with weight vector $E_{i,j}$ and hence

$$\Omega(\pi) = \{(\epsilon_i, -\epsilon_j) \mid i \in [m_1],\, j \in [m_2]\}.$$

Similar computations show that the SL-action on the one loop quiver, i.e., the conjugation action of $\mathrm{SL}_m(\mathbb{C})$ on $\mathbb{C}^{m\times m}$ has the following weights. For $i,j \in [m]$ with $i \neq j$, $(e_i - e_j)$ is a weight with weight vector $E_{i,j}$, and 0 is a weight with weight space $\bigoplus_{i\in[m]} \mathbb{C}E_{i,i}$. ♢

Finally, we define roots and root spaces.

Definition 1.3.20 (Roots and Root Spaces)**.** Let $G \subseteq \mathrm{GL}_m(\mathbb{C})$ be Zariski closed and self-adjoint. Set $T := G \cap \mathrm{GT}_m(\mathbb{K})$ and consider the adjoint representations Ad and ad from Example 1.3.1. The *non-zero* weights $\alpha \in \Omega(\mathrm{Ad})$ are called *roots* of G and the weight spaces $\operatorname{Lie}(G)_\alpha$ are called *root spaces*. Note that $Y \in \operatorname{Lie}(G)$ satisfies $\operatorname{ad}(X)(Y) = [X,Y] = 0$ for all $X \in \operatorname{Lie}(T)$ if and only if $Y \in \operatorname{Lie}(T)$. Hence, $\operatorname{Lie}(T)$ is the weight space of $0 \in \Omega(\mathrm{Ad})$ and with Theorem 1.3.14 we obtain

$$\operatorname{Lie}(G) = \operatorname{Lie}(T) \oplus \bigoplus_{\alpha} \operatorname{Lie}(G)_\alpha,$$

the *root space decomposition* of $\operatorname{Lie}(G)$. ▲

Example 1.3.21 ([FR21, Example B.3])**.** Let $G = \mathrm{SL}_m(\mathbb{C})$ and for $i,j \in [m]$ denote by $E_{i,j} \in \mathbb{C}^{m\times m}$ the matrix with entry one at position (i,j) and all other entries zero. For $i,j \in [m]$ with $i \neq j$ and for all $X = \operatorname{diag}(x_1,\dots,x_m) \in \operatorname{Lie}(T)$ we compute

$$\operatorname{ad}(X)(E_{i,j}) = [X, E_{i,j}] = (x_i - x_j) E_{i,j} = \operatorname{tr}\big(X \operatorname{diag}(e_i - e_j)\big) E_{i,j}.$$

Since $e_i - e_j \in \mathbb{1}_m^\perp \cong \mathfrak{i}\operatorname{Lie}(T_K)$, we deduce $e_i - e_j \in \Omega(\mathrm{Ad})$ with weight vector $E_{i,j}$. Therefore, the set of roots of $G = \mathrm{SL}_m(\mathbb{C})$ is $\{e_i - e_j \mid i,j \in [m], i \neq j\}$, because $\operatorname{Lie}(G) = \operatorname{Lie}(T) \oplus \bigoplus_{i\neq j} \mathbb{C}E_{i,j}$.

More generally, one can deduce that the roots of $G = \mathrm{SL}_m(\mathbb{C})^d$ are the

$$(e_i - e_j, 0_m, \dots, 0_m), (0_m, e_i - e_j, 0_m, \dots, 0_m), \dots, (0_m, \dots, 0_m, e_i - e_j) \in (\mathbb{R}^m)^d$$

for $i,j \in [m]$ with $i \neq j$. ♢

We need the following property of roots, which is proved similarly as [Hal15, Lemma 6.5] and [Kna96, Proposition 5.4(c)].

[18] By abuse of notation, $\epsilon_i = e_i - m_1^{-1}\mathbb{1}_{m_1} \in \mathbb{1}_{m_1}^\perp$ while $\epsilon_j = e_j - m_2^{-1}\mathbb{1}_{m_2} \in \mathbb{1}_{m_2}^\perp$.

Proposition 1.3.22 ([FR21, Proposition B.4]). *Let $G \subseteq \mathrm{GL}_N(\mathbb{C})$ be a Zariski closed self-adjoint subgroup and let α be a root of G with root space $\mathrm{Lie}(G)_\alpha$. Consider a rational representation $\pi\colon G \to \mathrm{GL}(V)$ and its induced differential $\Pi\colon \mathrm{Lie}(G) \to \mathrm{End}(V)$. If V_ω is the weight space of some weight $\omega \in \Omega(\pi)$, then*

$$\Pi\big(\mathrm{Lie}(G)_\alpha\big)(V_\omega) \subseteq V_{\omega+\alpha},$$

where $V_{\omega+\alpha} := \{0\}$, if $\omega + \alpha \notin \Omega(\pi)$.

1.4 Stability Notions

We introduce the (topological) stability notions that play a central role in this thesis. From the perspective of Geometric Invariant Theory (GIT), our definitions in terms of the Euclidean topology may seem unusual. However, this is needed for the Kempf Ness Theorem (Section 2.2) over $\mathbb{R}$. We also comment on connections to GIT and point out that in the complex reductive setting our notions agree with the classical notions from GIT, see Remark 1.4.8. We refer to [Dol03; Hos15; Kra84; Mum77; MFK94; New78; PV94] for further information on GIT.

Let G be a group[19] and V a finite dimensional $\mathbb{K}$-vector space with an inner product. Given a representation $\pi\colon G \to \mathrm{GL}(V)$, define the *capacity* of $v \in V$ as

$$\mathrm{cap}_G(v) := \inf_{g\in G} \|g \cdot v\|^2. \tag{1.8}$$

Note that $\mathrm{cap}_G(v) = \mathrm{cap}_G(g \cdot v)$ holds for all $g \in G$.

Definition 1.4.1 (Toplogical Stability Notions). Let $\pi\colon G \to \mathrm{GL}(V)$ be a representation of a group G, where V a finite-dimensional $\mathbb{K}$-vector space equipped with its Euclidean topology. For $v \in V$, denote its stabilizer by G_v and its orbit by $G \cdot v$. We define the following stability notions under the action of G.

(a) v is *unstable*, if $0 \in \overline{G \cdot v}$. Equivalently, $\mathrm{cap}_G(v) = 0$.

(b) v is *semistable*, if $0 \notin \overline{G \cdot v}$. Equivalently, $\mathrm{cap}_G(v) > 0$.

(c) v is *polystable*, if $v \neq 0$ and $G \cdot v$ is closed.

(d) v is *stable*, if v is polystable and G_v is finite.

Note that polystable implies semistable. The set $\mathcal{N}$ of all unstable points is called *(topological) null cone*. ▲

Usually, we consider the stability notions for a rational representation of an algebraic group over $\mathbb{K}$. In Part III on algebraic statistics we often restrict to the image and work with stability notions under $\pi(G)$.

Remark 1.4.2. We note that (a), (b) and (c) in Definition 1.4.1 only depend on the image $H := \pi(G)$, so these stability notions coincide for the action of G and of H. However, the notion *stable* may change as $H_v = G_v/\ker(\pi)$. Namely, if $\ker(\pi)$ is infinite (and hence $G \supseteq \ker(\pi)$ is), it may be that H_v is finite. Still, if $\ker(\pi)$ is finite, then H_v is finite if and only if G_v is finite. Hence, also the notion of *stable* coincides in this case. ▽

[19] not necessarily endowed with further structure

Example 1.4.3. Let $G = \mathrm{GL}_m(\mathbb{K})$ act on $V = \mathbb{K}^{m\times n}$ via left multiplication. Then any $v \in V$ is unstable: for $\epsilon > 0$ we see that $(\varepsilon \,\mathrm{I}_m) \cdot v = \varepsilon v \to 0$ as $\varepsilon \to 0$. Therefore, this action is in a certain sense "uninteresting" when studying stability notions. This also applies to similar actions of (products of) GL, e.g., left-right action from Example 1.3.3 or the tensor scaling action from 1.3.5. ♢

As a consequence of the preceding example, it is more natural to consider actions of (products of) SL.

Example 1.4.4. Let $G = \mathrm{SL}_m(\mathbb{K})$ act on $V = \mathbb{K}^{m\times n}$ via left multiplication. We argue that $Y \in \mathbb{K}^{m\times n}$ is either unstable or stable, depending on its row rank only.

If Y does not have full row rank m, then by Gaussian elimination one can create a matrix $Y' \in G \cdot Y$ that has a zero row. To ease notation assume the first row of Y' is zero. Then $\mathrm{diag}(\varepsilon^{-m+1}, \varepsilon, \ldots, \varepsilon) \cdot Y' \to 0$ for $\varepsilon \to 0$ and therefore Y is G-unstable. In particular, if $m > n$ then all matrices are unstable.

Now, assume Y has full row rank m, so we must have $m \leq n$ and $Y \neq 0$. We argue that Y is stable under G. If $g \in G_Y$, i.e., $gY = Y$, then g has m linearly independent eigenvectors, which are columns of Y, for eigenvalue one. Hence, we must have $g = \mathrm{I}_m$ and this shows $G_Y = \{\mathrm{I}_m\}$ is finite. To show that the orbit $G \cdot Y$ is Euclidean closed consider first $m = n$. Then

$$G \cdot Y = \{X \in \mathbb{K}^{m\times m} \mid \det(X) = \det(Y)\},$$

where "$\supseteq$" is clear, and conversely given X with $\det(X) = \det(Y)$ just consider $g := XY^{-1} \in G$. We see that $G \cdot Y$ is even Zariski closed. For the general case $m \leq n$, the assumption on Y means that Y has a non-vanishing maximal minor. Without loss of generality assume it is given by the first m columns $Y_1, \ldots, Y_m$. Set $Y_{(1..m)} := (Y_1, \ldots, Y_m) \in \mathbb{K}^{m\times m}$. One verifies that

$$\begin{aligned} G \cdot Y = \big\{X \in \mathbb{K}^{m\times n} \,|\, &\det(X_{(1..m)}) = \det(Y_{(1..m)}), \\ &(X_{(1..m)})(Y_{(1..m)})^{-1}(Y_{m+1}, \ldots, Y_n) = (X_{m+1}, \ldots, X_n)\big\}, \end{aligned}$$

which is again Zariski closed. Altogether, Y is stable if it has full row rank. ♢

In algebraic statistics one is usually interested in the real setting. For this, the next statement is very useful when working with *reductive* groups.

Proposition 1.4.5 ([DM21, Proposition 2.21])**.** *Let G be a connected, complex reductive $\mathbb{R}$-group. Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation of G defined over $\mathbb{R}$ and let $v \in V_{\mathbb{R}}$. Then v is un-/semi-/poly-/stable under $G_{\mathbb{R}}$ if and only if v is un-/semi-/poly-/stable under G.*

In the following, we comment on connections to Geometric Invariant Theory (GIT). In particular, we justify our topological notions of stability by showing that they agree with the "usual" stability notions from GIT, see Remark 1.4.8. First, we need to recall the ring of invariants.

In the following $\pi\colon G \to \mathrm{GL}(V)$ is always a rational representation of a complex reductive group. The representation π induces a natural action of G on the coordinate ring $\mathbb{C}[V]$ of V via

$$(g \cdot f)(v) := f(g^{-1} \cdot v), \quad \text{where } g \in G, f \in \mathbb{C}[V], v \in V.$$

The *ring of invariants* is the set of all fixed points under the latter action:

$$\mathbb{C}[V]^G := \big\{ f \in \mathbb{C}[V] \mid \forall\, g \in G\colon\ g \cdot f = f \big\}.$$

That is, $\mathbb{C}[V]^G$ contains exactly those regular functions on V that are constant on the G-orbits in V. We start with Hilbert's finiteness theorem [Hil90; Hil93]. Modern references are [DK15, Theorem 2.2.10] and [PV94, Theorem 3.5].

Theorem 1.4.6 (Hilbert)**.** *Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation of a complex reductive group. Then $\mathbb{C}[V]^G$ is a finitely generated $\mathbb{C}$-algebra.*

The *invariant-theoretic null cone* is defined as

$$\mathcal{N}^{\mathrm{inv}} := \big\{ v \in V \mid \forall\, f \in \mathbb{C}[V]^G\colon\ f(v) = f(0) \big\}.$$

In words, $\mathcal{N}^{\mathrm{inv}}$ contains all vectors that cannot be distinguished by invariants from the zero vector. A different characterization is obtained with the next theorem.

Theorem 1.4.7. *Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation of a complex reductive group. For $v, w \in V$ it holds that*

$$\overline{G \cdot v}^{\mathrm{Z}} \cap \overline{G \cdot w}^{\mathrm{Z}} = \emptyset \quad \Leftrightarrow \quad \exists\, f \in \mathbb{C}[V]^G\colon\ f(v) \neq f(w).$$

Moreover, any orbit closure contains a unique Zariski closed orbit.

Proof. Note that any $f \in \mathbb{C}[V]^G$ is constant on G-orbits and hence, by continuity, on Zariski closures of G-orbits. Therefore, $\overline{G \cdot v}^{\mathrm{Z}} \cap \overline{G \cdot w}^{\mathrm{Z}} \neq \emptyset$ implies that for all $f \in \mathbb{C}[V]^G$ one has $f(v) = f(w)$. The other direction follows from [Dol03, Lemma 6.1], also see [Wal17, Theorem 3.12].

Let $x \in V$. Since invariants are constant on $\overline{G \cdot x} = \overline{G \cdot x}^{\mathrm{Z}}$, the first part shows that there can be at most one Zariski closed orbit in $\overline{G \cdot x}$. Such an orbit always exists by Proposition 1.1.11. □

In the special case $w = 0$, the above theorem shows that $v \in \mathcal{N}^{\mathrm{inv}}$ if and only if $0 \in \overline{G \cdot v}^{\mathrm{Z}}$. A vector v lying in $\mathcal{N}^{\mathrm{inv}}$ is called unstable (in the GIT sense). More generally, we have the following.

Remark 1.4.8 (Stability Notions in GIT). Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation of a complex reductive group. In Geometric Invariant Theory (GIT), when studying (affine) GIT quotients one usually considers the notions unstable, semistable and stable. They have different equivalent characterizations (as G is reductive), see the excellent Table 1.1 in [Mum77, p. 41]. One characterization is via the ring of invariants $\mathbb{C}[V]^G$, e.g., as for $\mathcal{N}^{\mathrm{inv}}$. Another characterization is topological and exactly as in Definition 1.4.1(a), (b) and (d), but using the Zariski topology instead of the Euclidean; also compare [MFK94, Appendix, p. 194]. Similarly, some modern literature (e.g., [Tho06]) defines polystable as in Definition 1.4.1(c), again using the Zariski topology. Taking into account that Euclidean and Zariski closure of a G-orbit coincide (Corollary 1.1.12), we see that the classical stability notions from GIT agree with the ones in Definition 1.4.1.

We caution the reader to always check the definitions of stability in the literature. Over time the namings have changed: e.g., polystable is called "stable" in the main text of [MFK94], while stable is called there "properly stable". Moreover, polystable is "Kempf-stable" in [Dol03] and "nice semistable" in [Nes84]. ▽

The next example stresses that G being reductive is necessary for the equality of invariant-theoretic and topological null cone.

Example 1.4.9. Let $G = \mathbb{C}$ be the one-dimensional additive group, which is non-reductive (Example 1.1.24). Consider the rational representation

$$\pi\colon G \to \mathrm{GL}_2(\mathbb{C}), \quad g \mapsto \begin{pmatrix} 1 & g \\ 0 & 1 \end{pmatrix}$$

on $V = \mathbb{C}^2$, i.e., g acts on $(x, y) \in \mathbb{C}^2$ via $g \cdot (x, y) = (x + gy, y)$. Denote the coordinate functions on V by $X, Y \in \mathbb{C}[V]$. Then $\mathbb{C}[Y] \subseteq \mathbb{C}[V]^G$ and one verifies that equality holds. Therefore,

$$\mathcal{N}^{\mathrm{inv}} = \big\{(x, 0) \mid x \in \mathbb{C}\big\}.$$

Moreover, any orbit $G \cdot (x, y)$ is either a point (if $y = 0$) or a horizontal affine line (if $y \neq 0$). Thus, all orbits are closed and hence the topological null cone is

$$\mathcal{N} = \big\{(x, y) \in \mathbb{C}^2 \mid 0 \in \overline{G \cdot (x, y)}\big\} = \{0\}.$$

We see that $\mathcal{N} \subsetneq \mathcal{N}^{\mathrm{inv}}$. ♢

Chapter 2

Criteria for Stability Notions

The chapter presents several criteria for testing stability notions from Definition 1.4.1. These criteria are used throughout the thesis. We give corresponding references in each section.

Organization. Section 2.1 contains the Hilbert-Mumford Criterion for tori, and more generally, for reductive groups. In Section 2.2 we introduce moment maps and moment polytopes, and state the Kempf-Ness Theorem, which is of particular importance for this thesis. Afterwards, we deduce from King's Criterion a characterization for being (semi)stable under the left-right action, Section 2.3. While all previous criteria require a reductive group, Popov's Criterion in Section 2.4 can be used to test polystability under a solvable group.

2.1 Hilbert-Mumford Criterion

In the following we formulate the Hilbert-Mumford Criterion for tori and then for general reductive groups. Afterwards, we focus on the torus case and provide two detailed proofs. The latter is mainly based on [AKRS21b, Appendix A].

Let G be a complex algebraic group. An *(algebraic) one-parameter subgroup* (short: 1-psg) of G is a morphism $\lambda\colon \mathbb{C}^\times \to G$ of complex algebraic groups G.

Example 2.1.1. The algebraic one-parameter subgroups of the torus $\mathrm{GT}_d(\mathbb{C})$ are in bijection with $\mathbb{Z}^d$. The 1-psg given by $(\lambda_1, \dots, \lambda_d) \in \mathbb{Z}^d$ is

$$\lambda\colon \mathbb{C}^\times \to \mathrm{GT}_d(\mathbb{C}), \quad t \mapsto \mathrm{diag}\left(t^{\lambda_1}, \dots, t^{\lambda_d}\right). \tag{2.1}$$

By abuse of notation, we denote by λ both the 1-psg and the vector in $\mathbb{Z}^d$. ♢

Theorem 2.1.2 (Hilbert-Mumford for Tori, [Kra84, p. 173])**.**
Let $\pi\colon T \to \mathrm{GL}(V)$ be a rational representation of a complex torus T. Fix $v \in V$ and let $w \in \overline{T \cdot v} \backslash T \cdot v$. Then there exists an algebraic one-parameter subgroup $\lambda\colon \mathbb{C}^\times \to T$ such that

$$\lim_{t \to 0} \lambda(t) \cdot v \in T \cdot w.$$

In particular, if $v \neq 0$ is T-unstable, then choosing $w = 0$ gives $\lim_{t\to 0} \lambda(t) \cdot v = 0$.

We give a proof of the special case of an unstable v and $w = 0$ below in Theorem 2.1.7. Furthermore, Theorem 2.1.2 allows for a characterization of all stability notions under a torus, see Theorem 2.1.9 below. For a general reductive group we have the following statement, also see [Bir71, Theorem 4.2] (proof due to R. Richardson).

Theorem 2.1.3 (Hilbert-Mumford for Reductive Groups, [PV94, Theorem 6.9])**.** *Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation of a complex reductive group G. Fix $v \in V$ and let $G \cdot w$ be the unique closed orbit*[1] *in $\overline{G \cdot v}$. Then there exists an algebraic one-parameter subgroup $\lambda\colon \mathbb{C}^\times \to G$ of G such that*

$$\lim_{t\to 0} \lambda(t) \cdot v \in G \cdot w.$$

In particular, if $v \neq 0$ is G-unstable, then $w = 0$ yields $\lim_{t\to 0} \lambda(t) \cdot v = 0$.

Hence, the Hilbert-Mumford Criterion ensures that being unstable under the action of a reductive group is always witnessed by a one-parameter subgroup.

Remark 2.1.4. Regarding Theorem 2.1.3 we point out the following.

(i) In contrast to case of tori (Theorem 2.1.2), for a reductive group G one can in general *not* choose any G-orbit in $\overline{G \cdot v} \backslash G \cdot v$. Indeed, Example 1 in [PV94, §6.8] shows that the assumption "$G \cdot w$ is the unique closed orbit in $\overline{G \cdot v}$" in Theorem 2.1.3 is necessary.

(ii) If the whole setting in Theorem 2.1.3 is defined over $\mathbb{R}$ and $v \in V_{\mathbb{R}}$, then one can choose a one-parameter subgroup that is defined over $\mathbb{R}$, by a result of Birkes [Bir71, Theorem 5.2]. In fact, it was proven by Kempf that such a rationality result of the Hilbert-Mumford Criterion holds for *any* perfect field, [Kem78, Corollary 4.3].

(iii) The Hilbert-Mumford Criterion is an important proof ingredient for the Kempf-Ness Theorem 2.2.13 , both over the complex and over the real numbers. ▽

We will need the following result, that is often shown as an intermediate step to prove Hilbert-Mumford.

Theorem 2.1.5 ([Wal17, Theorem 3.25])**.** *Let $G \subseteq \mathrm{GL}_N(\mathbb{C})$ be Zariski closed and self-adjoint. Set $K := G \cap \mathrm{U}_N$ and $T := (G \cap \mathrm{GT}_N(\mathbb{C}))^\circ$. Consider a rational representation $\pi\colon G \to \mathrm{GL}(V)$ and fix $v \in V$. Let $G \cdot w$ be the unique closed orbit in $\overline{G \cdot v}$. Then there exists $k \in K$ such that $\overline{T \cdot (k \cdot v)} \cap G \cdot w \neq \emptyset$. In particular, if v is G-unstable, then $w = 0$ and hence $0 \in \overline{T \cdot (k \cdot v)}$.*

Proofs in the Torus Case

We provide a proof of the "classical" Hilbert-Mumford Theorem for a torus, and for characterizations via the so-called weight polytope. The proofs are taken from [AKRS21b, Appendix A] and are intended to be accessible to a wide audience. They illustrate that the Hilbert-Mumford Criterion in the torus case is an instance of linear programming duality and its many variants, compare [Sch86, Chapter 7].

Let $T \subseteq \mathrm{GT}_N(\mathbb{C})$ be a complex sub-torus and set $T_K := T \cap \mathrm{U}_N$. Consider a rational representation $\pi\colon T \to \mathrm{GL}(V)$ with set of weights $\Omega(\pi) \subseteq \mathfrak{i}\,\mathrm{Lie}(T_K) \subseteq \mathbb{R}^N$

[1] compare Theorem 1.4.7

and weight space decomposition $V = \bigoplus_\omega V_\omega$, see Theorem 1.3.14. Given $v \in V$, we write $v = \sum_\omega v_\omega$ with $v_\omega \in V_\omega$. Define the *support* of v with respect to π as

$$\operatorname{supp}(v) := \{\omega \in \Omega(\pi) \mid v_\omega \neq 0\}.$$

Furthermore, the *weight polytope* of v is

$$\Delta_T(v) := \operatorname{conv}\big\{\omega \mid \omega \in \operatorname{supp}(v)\big\} \subseteq \mathrm{i}\operatorname{Lie}(T_K) \subseteq \mathbb{R}^N. \tag{2.2}$$

Using the weight polytope, the Hilbert-Mumford Criterion, Theorem 2.1.2, actually yields a characterization of all stability notions from Definition 1.4.1; compare Theorem 2.1.9 below. Since any torus is isomorphic to $\mathrm{GT}_d(\mathbb{C})$, we restrict for concreteness to this situation.

Let $T = \mathrm{GT}_d(\mathbb{C})$ act on $V = \mathbb{C}^m$ via the matrix $A \in \mathbb{Z}^{d\times m}$, see Example 1.3.16. The weights of this action are the columns A_j of the matrix A with corresponding weight vector $e_j \in \mathbb{C}^m$. Therefore, the weight polytope (2.2) of $v \in \mathbb{C}^m$ is

$$\Delta_A(v) := \Delta_T(v) = \operatorname{conv}\big\{A_j \mid v_j \neq 0\big\}.$$

It is convenient to remember the weight matrix A in the notation.

Now, we head towards proving the special case of Theorem 2.1.2. For this, let λ be a one-parameter subgroup of $T = \mathrm{GT}_d(\mathbb{C})$ as in (2.1). For $v \in \mathbb{C}^m$, the j^{th} entry of $\lambda(t)\cdot v$ is

$$(\lambda(t)\cdot v)_j = t^{\langle\lambda, A_j\rangle} v_j.$$

We consider $\lim_{t\to 0}\lambda(t)\cdot v$. Its j^{th} entry is zero for $j \notin \operatorname{supp}(v)$. For $j \in \operatorname{supp}(v)$, we have three possibilities

$$\Big(\lim_{t\to 0}\ \lambda(t)\cdot v\Big)_j = \begin{cases} 0 & \text{if } \langle\lambda, A_j\rangle > 0 \\ v_j & \text{if } \langle\lambda, A_j\rangle = 0 \\ \infty & \text{if } \langle\lambda, A_j\rangle < 0 \end{cases} \tag{2.3}$$

To prove the Hilbert-Mumford Criterion, we need the following result from the realm of linear programming duality, Farkas' lemma, etc.

Theorem 2.1.6 (Gordan's Transposition Theorem, [Sch86, §7.8 Equation (31)])**.** *Let $\mathbb{F} \in \{\mathbb{Q}, \mathbb{R}\}$ and $B \in \mathbb{F}^{d\times k}$. There is a vector $x \in \mathbb{F}^k$ with $x \geq 0$, $x \neq 0$ and $Bx = 0$ if and only if there is no vector $y \in \mathbb{F}^d$ with $y^\mathsf{T} B > 0$.*

The classical statement of the Hilbert-Mumford Criterion for a torus action is as follows, see e.g., [PV94, Proposition 5.3] and [Bir71, Lemma 3.4].

Theorem 2.1.7. *Consider the action of $\mathrm{GT}_d(\mathbb{C})$ on $\mathbb{C}^m$ via the matrix $A \in \mathbb{Z}^{d\times m}$. Let $v \in \mathbb{C}^m\backslash\{0\}$ with zero in its orbit closure. Then there exists a one-parameter subgroup λ of $\mathrm{GT}_d(\mathbb{C})$ such that $\lim_{t\to 0}\lambda(t)\cdot v = 0$.*

Proof of Theorem 2.1.7. The proof follows [Sur00]. We have $\operatorname{supp}(v) \neq \emptyset$ as $v \neq 0$. After reordering the entries of v, we can assume without loss of generality that $\operatorname{supp}(v) = [k]$ for some $k \leq m$.

We seek a one parameter subgroup $\lambda\colon \mathbb{C}^\times \to \mathrm{GT}_d(\mathbb{C})$ such that $\lim_{t\to 0} \lambda(t)\cdot v$ is zero. From the form of a one parameter subgroup in (2.1) and the limiting behaviour from (2.3), we see that this is equivalent to showing that

$$\exists\, \lambda \in \mathbb{Z}^d\colon \forall\, j \in [k] = \mathrm{supp}(v)\colon \quad \langle \lambda, A_j\rangle > 0. \tag{2.4}$$

Let $B \in \mathbb{Z}^{d\times k}$ be the submatrix consisting of the first k columns of $A = (a_{ij})$. Then (2.4) reformulates as: there exists $\lambda \in \mathbb{Z}^d$ with $\lambda^\mathsf{T} B > 0$. Hence, by Theorem 2.1.6 with $\mathbb{F} = \mathbb{Q}$, (2.4) is equivalent[2] to the following statement:

$$\begin{array}{l}\text{if } x = (x_1,\ldots,x_k) \in \mathbb{Q}^k\backslash\{0\} \text{ is such that } a_{i1}x_1 + \cdots + a_{ik}x_k = 0\\ \text{for all } i \in [d], \quad \text{then at least two entries of } x \text{ are of opposite sign.}\end{array} \tag{2.5}$$

Thus, it remains to prove (2.5). Since $0 \in \overline{\mathrm{GT}_d(\mathbb{C})\cdot v}$, there exists a sequence $t^{(n)} = (t_1^{(n)},\ldots,t_d^{(n)}) \in \mathrm{GT}_d(\mathbb{C})$ with $t^{(n)}\cdot v \to 0$ as $n\to\infty$. In coordinates,

$$\forall\, j \in [k]\colon \quad \left(t_1^{(n)}\right)^{a_{1j}} \cdots \left(t_d^{(n)}\right)^{a_{dj}} \to 0 \quad \text{as} \quad n\to\infty. \tag{2.6}$$

The hypothesis of (2.5) is that we have $x \in \mathbb{Q}^k\backslash\{0\}$ with $x_1a_{i1}+\cdots+x_ka_{ik} = 0$ for all $i \in [d]$. Without loss of generality, we can assume x_1 is non-zero and therefore

$$\forall\, j \in [k]\colon \quad -a_{i1} = \frac{x_2}{x_1}a_{i2} + \cdots + \frac{x_k}{x_1}a_{ik},$$

which implies

$$\prod_{i=1}^d \left(t_i^{(n)}\right)^{-a_{i1}} = \left(\prod_{i=1}^d \left(t_i^{(n)}\right)^{a_{i2}}\right)^{\frac{x_2}{x_1}} \cdots \left(\prod_{i=1}^d \left(t_i^{(n)}\right)^{a_{ik}}\right)^{\frac{x_k}{x_1}}. \tag{2.7}$$

If $x_j/x_1 \geq 0$ for all $j \in \{2,\ldots,k\}$, then the right-hand side of (2.7) either equals one (if all x_j/x_1 are zero) or tends to zero (if there exists some j with $x_j/x_1 > 0$). But the left-hand side of (2.7) tends to infinity as $n\to\infty$, since it is the inverse of (2.6) for $j = 1$. Hence x_j/x_1 must be strictly negative for some j, i.e., x_1 and x_j have opposite signs. □

We note that the generalization in Theorem 2.1.2 can be proven by similar arguments from polyhedral geometry.

Now, let us turn towards Hilbert-Mumford in terms of the weight polytope. We use the following consequence of Gordan's Theorem 2.1.6.

Corollary 2.1.8. *Let $B \in \mathbb{Z}^{d\times k}$ and let $\Delta_B \subseteq \mathbb{R}^d$ be the polytope spanned by the columns of B. Then $0 \notin \Delta_B$ if and only if there exists $\lambda \in \mathbb{Z}^d$ with $\lambda^\mathsf{T} B > 0$.*

Proof. First, note that $0 \in \Delta_B$ is equivalent to the existence of $x \in \mathbb{R}^d\backslash\{0\}$ such that $x \geq 0$ and $Bx = 0$. Thus, if there is $\lambda \in \mathbb{Z}^d$ with $\lambda^\mathsf{T} B > 0$, then $0 \notin \Delta_B$, by Theorem 2.1.6 for $\mathbb{F} = \mathbb{R}$. On the other hand, if $0 \notin \Delta_B$ then there is $y \in \mathbb{R}^d$ with $y^\mathsf{T} B > 0$, again by Theorem 2.1.6. The existence of such a vector y ensures that we can in fact choose $y \in \mathbb{Q}^d$. After multiplying with a common denominator, we obtain some $\lambda \in \mathbb{Z}^d$ with $\lambda^\mathsf{T} B > 0$. □

[2]Note that the existence of a $y \in \mathbb{Q}^d$ with $y^\mathsf{T} B > 0$ is, after multiplying with a common denominator, equivalent to the existence of some $\lambda \in \mathbb{Z}^d$ with $\lambda^\mathsf{T} B > 0$. In [Sur00] the equivalence of (2.4) and (2.5) is stated in Lemma 1.1.

Finally, we prove a full characterization of stability notions via the weight polytope. Its formulation is based on [AKRS21b, Theorem 3.4] and the proof is taken from [AKRS21b, Appendix A]. Given a polytope $P \subseteq \mathbb{R}^d$, we denote its *interior* by $\mathrm{int}(P)$ and its *relative interior* by $\mathrm{relint}(P)$.

Theorem 2.1.9 (Hilbert-Mumford Criterion via the Weight Polytope)**.** *Consider the action of* $\mathrm{GT}_d(\mathbb{C})$ *on* $\mathbb{C}^m$ *given by matrix* $A \in \mathbb{Z}^{d\times m}$*. For* $v \in \mathbb{C}^m$*, we have*

$$\begin{array}{lccc} (a) & v \text{ unstable} & \Leftrightarrow & 0 \notin \Delta_A(v) \\ (b) & v \text{ semistable} & \Leftrightarrow & 0 \in \Delta_A(v) \\ (c) & v \text{ polystable} & \Leftrightarrow & 0 \in \mathrm{relint}(\Delta_A(v)) \\ (d) & v \text{ stable} & \Leftrightarrow & 0 \in \mathrm{int}(\Delta_A(v)) \end{array}$$

If $\mathrm{GT}_d(\mathbb{C})$ *acts on* $\mathbb{C}^m$ *given by matrix* $A \in \mathbb{Z}^{d\times m}$ with linearization $b \in \mathbb{Z}^d$*, then the same statements (a) – (d) apply when replacing zero by* b*.*

Remark 2.1.10. Of course, Theorem 2.1.9 also holds for the setting $T \subseteq \mathrm{GL}_N(\mathbb{C})$ with weight polytope $\Delta_T(v)$ as in (2.2). In that situation, the interior in part (d) has to be taken with respect to the $\mathbb{R}$-vector space $\mathrm{i}\,\mathrm{Lie}(T_K)$. ▽

We give a (hopefully) elementary and accessible proof of Theorem 2.1.9. Other references are [Dol03, Theorem 9.2] and [Szé06, Theorem 1.5.1].

Proof of Theorem 2.1.9. Set $T := \mathrm{GT}_d(\mathbb{C})$. We first prove part (a), and hence (b) as well. If $v = 0$, then the polytope $\Delta_A(v)$ is empty, hence $0 \notin \Delta_A(v)$. Assume $v \neq 0$. Then v is unstable if and only if there exists some $\lambda \in \mathbb{Z}^d$ such that $\langle \lambda, A_j\rangle > 0$ for all $j \in \mathrm{supp}(v)$, by combining Theorem 2.1.7 with (2.3). By Corollary 2.1.8, this is equivalent to $0 \notin \Delta_A(v)$.

For (c), we first prove that if 0 is on the boundary of $\Delta_A(v)$, then v is not polystable. We construct a point in the orbit closure of v, with support strictly smaller than that of v, and hence deduce that the orbit of v is not closed. Since 0 lies on the boundary of $\Delta_A(v)$, it is contained in a minimal face $F \subsetneq \Delta_A(v)$. Since A has integer entries, there is a hyperplane

$$H_\lambda := \{x \in \mathbb{R}^d \mid \langle \lambda, x\rangle = 0\},$$

with $\lambda \in \mathbb{Z}^d$, such that $F = H_\lambda \cap \Delta_A(v)$. We choose the sign of λ so that it has non-negative inner product with all of $\Delta_A(v)$. This ensures that the limit $w := \lim_{t\to 0} \lambda(t)\cdot v$ exists. The limit w has $\mathrm{supp}(w) \subsetneq \mathrm{supp}(v)$, since $\Delta_A(w) \subseteq F$. Hence $w \in \overline{T\cdot v}\backslash T\cdot v$, and $T\cdot v$ is not closed.

For the converse direction of (c), we show that if v is semistable but not polystable, then $0 \notin \mathrm{relint}(\Delta_A(v))$. Let $w' \in \overline{T\cdot v}\backslash T\cdot v$. There exists $\lambda \in \mathbb{Z}^d$ such that $w := \lim_{t\to 0} \lambda(t)\cdot v \in T\cdot w'$, by Theorem 2.1.2. We have $\mathrm{supp}(w) \subseteq \mathrm{supp}(v)$ and, moreover, $\mathrm{supp}(w) \subsetneq \mathrm{supp}(v)$ (otherwise $w = v$ by (2.3), a contradiction). Hence $\langle \lambda, A_j\rangle > 0$ for all $j \in \mathrm{supp}(v)\backslash \mathrm{supp}(w)$, while $\langle \lambda, A_j\rangle = 0$ for all $j \in \mathrm{supp}(w)$, by (2.3). We obtain $\Delta_A(v) \nsubseteq H_\lambda$ and $\Delta_A(w) = H_\lambda \cap \Delta_A(v)$, i.e., $\Delta_A(w)$ is a proper face of $\Delta_A(v)$. We have $T\cdot w = T\cdot w' \subseteq \overline{T\cdot v}$ and so w is semistable as v is semistable. By (b), $0 \in \Delta_A(w)$ and hence 0 is on the boundary of $\Delta_A(v)$.

To prove (d), we can assume v is polystable, i.e., $0 \in \mathrm{relint}(\Delta_A(v))$. We want to show that the dimension of the stabilizer $T_v = \{t \in T \mid t\cdot v = v\}$ is zero

if and only if the interior of $\Delta_A(v)$ equals its relative interior (i.e., if and only if $\Delta_A(v)$ is full-dimensional). Since $0 \in \Delta_A(v)$, the equality of the interior and relative interior holds if and only if $U := \operatorname{span}\{A_j \mid j \in \operatorname{supp}(v)\}$ equals $\mathbb{R}^d$. If T_v is positive dimensional, it must contain a one-parameter subgroup, i.e., some $\lambda \in \mathbb{Z}^d \backslash \{0\}$ with $\lambda(t) \cdot v = v$ for all $t \in \mathbb{C}^\times$. Then $\langle \lambda, A_j \rangle = 0$ for all $j \in \operatorname{supp}(v)$, so the orthogonal complement $U^\perp \subseteq \mathbb{R}^d$ contains a line, and $U \neq \mathbb{R}^d$. Conversely, if $U \neq \mathbb{R}^d$ then there exists non-zero $\lambda \in U^\perp$, which can be chosen to have integer entries, since A has integer entries. Then the image of the non-trivial one parameter subgroup λ lies in T_v, which is therefore positive-dimensional.

Finally, if T acts on $\mathbb{C}^m$ by matrix $A \in \mathbb{Z}^{d \times m}$ with linearization $b \in \mathbb{Z}^d$, then this is the same as the action given by matrix $A' \in \mathbb{Z}^{d \times m}$, where A' has j^{th} column $A_j - b$; see (1.6) in Example 1.3.16. Therefore, we can deduce the last statement by noting that $\Delta_{A'}(v) = \Delta_A(v) - b$. □

2.2 Kempf-Ness Theorem

In this section we present an important analytical tool from invariant theory – the Kempf-Ness Theorem. It plays a crucial role in this thesis and is heavily used both in Part II and Part III. The presentation is based on [BFG+19] and [FR21], sometimes also on [AKRS21a] and [AKRS21b, Appendix B].

First, we introduce the Setting 2.2.2 and define the moment map. Thereby, we follow the conventions used in [BFG+19] for $\mathbb{K} = \mathbb{C}$, which enables a good comparison with that paper in Part II. Afterwards, we compute the moment map in several examples. We continue with Kempf-Ness, Theorem 2.2.13, and deduce several statements from it. Finally, we introduce moment polytopes, which are induced by the moment map and generalize the concept of weight polytopes.

The literature on Kempf-Ness, moment maps and polytopes, and related topics is vast. The following list is certainly incomplete. We refer to [KN79; MFK94; Wal17] for Kempf-Ness over $\mathbb{C}$ and to [RS90; Bil21; BL21; Wal17] for Kempf-Ness over $\mathbb{R}$. Moment polytopes are treated in [Bri87; GS84; Kir84a; OS00; Par20] and related topics can be found e.g., in [HS07; HSS08b; Kir84b; MFK94; Mar01; Nes84; Tho06] and the references therein.

The Moment Map

We need the following fact, see [Kna96, Proposition 4.6] or [Wal17, Theorem 2.9].[3]

Lemma 2.2.1. *Let K be a compact matrix Lie group and let $\pi\colon K \to \mathrm{GL}(V)$ be a continuous group morphism, where V is a finite dimensional $\mathbb{K}$-vector space. Then there exists an inner product $\langle \cdot, \cdot \rangle$ on V such that K acts isometrically, i.e.,*

$$\forall\, k \in K,\, v, w \in V\colon \quad \langle \pi(k)v, \pi(k)w \rangle = \langle v, w \rangle.$$

Equivalently, for all $k \in K$ we have $\pi(k)^* = \pi(k)^{-1} (= \pi(k^\dagger))$, where $\pi(k)^*$ denotes the adjoint of $\pi(k)$ with respect to $\langle \cdot, \cdot \rangle$.

[3] The proof via the Haar measure also works for $\mathbb{K} = \mathbb{R}$.

We are now ready to fix the required data for defining a moment map.

Setting 2.2.2. Let $G \subseteq \mathrm{GL}_N(\mathbb{K})$ be a Zariski closed self-adjoint subgroup. Recall the following from Proposition 1.2.14. The group $K := \{g \in G \mid g^\dagger g = \mathrm{I}_N\}$ is a maximal compact subgroup and setting $\mathfrak{p} := \mathrm{Lie}(G) \cap \mathrm{Sym}_N(\mathbb{K})$ we have an orthogonal decomposition $\mathrm{Lie}(G) = \mathrm{Lie}(K) \oplus \mathfrak{p}$ of *real* vector spaces with respect to the inner product $\mathrm{Re}\big(\mathrm{tr}(X^\dagger Y)\big)$ on $\mathbb{C}^{N\times N}$. Furthermore, $\mathfrak{p} = \mathfrak{i}\,\mathrm{Lie}(K)$ if $\mathbb{K} = \mathbb{C}$.

Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation defined over $\mathbb{K}$ with differential $\Pi\colon \mathrm{Lie}(G) \to \mathrm{End}(V)$. Fix an inner product $\langle\cdot,\cdot\rangle$ on V such that K acts isometrically, and $\Pi(X)$ is self-adjoint for all $X \in \mathfrak{p}$.[4]

A K-invariant inner product always exists by Lemma 2.2.1. If $\mathbb{K} = \mathbb{C}$ then the property on $\Pi(X)$ automatically follows from the K-invariance of $\langle\cdot,\cdot\rangle$, $\mathbb{C}$-linearity of Π and the fact that $\mathfrak{p} = \mathfrak{i}\,\mathrm{Lie}(K)$. If $\mathbb{K} = \mathbb{R}$ the existence of $\langle\cdot,\cdot\rangle$ is ensured by [BH62, Proposition 13.5], also compare [RS90, §2.3]. ▲

We illustrate the general setting in an Example.

Example 2.2.3. Let $G := \mathrm{SL}_m(\mathbb{K})^d$ be block-diagonally embedded in $\mathrm{GL}_{dm}(\mathbb{K})$ ($N = dm$). Depending on $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, we have $K = \mathrm{SO}_m(\mathbb{R})^d$ or $K = (\mathrm{SU}_m)^d$, again block-diagonally embedded in $\mathrm{GL}_{dm}(\mathbb{K})$. Hence, their Lie algebras are block diagonally embedded into $\mathbb{K}^{dm\times dm}$. Consider the tensor scaling action $\pi_{m,d}$ of G on $V = (\mathbb{C}^m)^{\otimes d}$ from Example 1.3.5. For simplicity, let $d = 3$. One verifies that for all $(X, Y, Z) \in \mathrm{Lie}(G)$

$$\Pi(X, Y, Z) = X \otimes \mathrm{I}_m \otimes \mathrm{I}_m + \mathrm{I}_m \otimes Y \otimes \mathrm{I}_m + \mathrm{I}_m \otimes \mathrm{I}_m \otimes Z$$

using the Kronecker product of matrices. As desired, $\Pi(X, Y, Z) \in \mathrm{Sym}_{3m}(\mathbb{K})$ whenever $(X, Y, Z) \in \mathfrak{p}$. One verifies that K acts isometrically on V with respect to the standard inner product. ♢

Given the above setting, remember from Definition 1.4.1 that a vector v is called unstable if its capacity

$$\mathrm{cap}_G(v) := \inf_{g\in G} \|\pi(g)v\|^2 = \inf_{g\in G} \|g \cdot v\|^2$$

equals zero. It is semistable if the capacity is positive. Considering for $v \in V\backslash\{0\}$ the so-called *Kempf-Ness function*

$$F_v\colon G \to \mathbb{R}, \quad v \mapsto \log \|\pi(g)v\| = \frac{1}{2}\log\big(\|\pi(g)v\|^2\big) \tag{2.8}$$

we see that v is semistable if and only if F_v is bounded from below. In particular, if the capacity is positive and attained by some $\hat{g} \in G$, then the differential of F_v should vanish at $\hat{g}$. To make the concept of a differential/gradient more precise, notice that F_v is right-G-equivariant, i.e.,

$$\forall\, g, h \in G\colon \quad F_v(gh) = \log \|\pi(gh)v\| = \log \|\pi(g)\pi(h)v\| = F_{\pi(h)v}(g)\,.$$

[4] In concrete representations this will usually be the standard inner product; except for polynomial scaling in Section 4.6, where one has to take the Bombieri-Weyl inner product.

Furthermore, as K acts isometrically on V the function F_v is left-K-invariant:

$$\forall\, k \in K, g \in G\colon \quad F_v(kg) = \log \|\pi(kg)v\| = \log \|\pi(g)v\| = F_v(g)\,.$$

The G-equivariance ensures that it is enough to consider the differential of F_v at the identity. The latter is the map

$$\mathrm{Lie}(G) = \mathrm{Lie}(K) \oplus \mathfrak{p} \to \mathrm{Lie}(\mathbb{R}) = \mathbb{R}, \quad X \mapsto \frac{d}{dt}\Big|_{t=0} F_v\big(e^{tX}\big)\,,$$

compare Theorem 1.2.10. Now, the K-invariance of F_v implies that the differential is zero on the direct summand $\mathrm{Lie}(K)$, so it suffices to consider the orthogonal complement $\mathfrak{p}$. Altogether, we define the moment map *both* in the real and complex case as the gradient of F_v.[5]

Definition 2.2.4 (Moment Map)**.** Consider the Setting 2.2.2 and define the *moment map* $\mu_G\colon V\backslash\{0\} \to \mathfrak{p}$ as follows. For $v \in V\backslash\{0\}$, $\mu_G(v)$ is the unique element of the real vector space $\mathfrak{p}$, which satisfies for all $X \in \mathfrak{p}$

$$\mathrm{tr}\left(\mu_G(v)X\right) = \frac{d}{dt}\Big|_{t=0} F_v\big(e^{tX}\big) = \frac{\langle v, \Pi(X)v\rangle}{\langle v, v\rangle}.$$

Here we use that the inner product on $\mathfrak{p}$ is $\mathrm{Re}\big(\mathrm{tr}(\mu_G(v)^\dagger X)\big) = \mathrm{tr}(\mu_G(v)X)$, that $\Pi(\cdot)$ is $\mathbb{R}$-linear and that $\langle\cdot,\cdot\rangle$ is linear in the second component.[6] ▲

Remark 2.2.5. In the literature $\mu_G(v)$ is often the differential of F_v rather than the gradient. We follow the conventions in [BFG+19] for an easier comparison in Part II. ▽

Restricting π to some Zariski closed self-adjoint subgroup $H \subseteq G$ we can similarly define the moment map $\mu_H\colon V \setminus \{0\} \to \mathfrak{q}$, where $H_K := H \cap K$ and $\mathfrak{q} := \mathrm{Lie}(H_K) \cap \mathrm{Sym}_N(\mathbb{K}) \subseteq \mathfrak{p}$. The moment maps are related as follows.

Proposition 2.2.6 (based on [FR21, Proposition 4.2])**.** *Let $p\colon \mathfrak{p} \to \mathfrak{q}$ be the orthogonal projection with respect to the inner product* $\mathrm{Re}(\mathrm{tr}(X^\dagger Y)) = \mathrm{tr}(XY)$ *on* $\mathrm{Sym}_N(\mathbb{K})$*. Then $\mu_H = p \circ \mu_G$ and $\|\mu_H(v)\|_F \leq \|\mu_G(v)\|_F$ for all $v \in V \setminus \{0\}$.*

Proof. Since $\mathfrak{q} \subseteq \mathfrak{p}$ the definition of the moment maps gives

$$\mathrm{tr}(\mu_H(v)X) = \frac{\langle v, \Pi(X)v\rangle}{\langle v, v\rangle} = \mathrm{tr}(\mu_G(v)X) = \mathrm{tr}\left(p(\mu_G(v))X\right)$$

for all $X \in \mathfrak{q}$. Therefore, $p(\mu_G(v)) = \mu_H(v)$ and $\|\mu_H(v)\|_F \leq \|\mu_G(v)\|_F$ follows directly from this. □

Another property of the moment map is its K-equivariance.

[5] This definition agrees with [BFG+19, Definition 3.2] and [FR21, Definition 4.1], where only the complex case is considered.

[6] Remember that, by our convention, Hermitian inner products on $\mathbb{C}$-vector spaces are always linear in the second component and semi-linear in the first.

Proposition 2.2.7. *For all $v \in V\backslash\{0\}$ and all $k \in K$, $\mu_G(k \cdot v) = k\mu_G(v)k^\dagger$.*

Proof. Fix $v \in V\backslash\{0\}$ and $k \in K$. Note that $k\mathfrak{p}k^\dagger = \mathfrak{p}$. For all $X \in \mathfrak{p}$,

$$\begin{aligned}\operatorname{tr}\big(\mu_G(k\cdot v)X\big) &= \frac{1}{\|v\|^2}\big\langle \pi(k)v, \Pi(X)\pi(k)v\big\rangle \overset{(*)}{=} \frac{1}{\|v\|^2}\big\langle v, \Pi(k^\dagger Xk)v\big\rangle \\ &= \operatorname{tr}\big(\mu_G(v)k^\dagger Xk\big) = \operatorname{tr}\big(k\mu_G(v)k^\dagger X\big),\end{aligned}$$

where we used in $(*)$ that $\pi(k)^* = \pi(k^\dagger)$ and then Theorem 1.2.10 Item 1. Since $k\mu_G(v)k^\dagger \in \mathfrak{p}$ we must have $\mu_G(k \cdot v) = k\mu_G(v)k^\dagger$. □

Moment Map in Examples

We state the moment maps for several actions and give a detailed computation in some cases. At a first read one may only skim through the results to quickly progress to the Kempf-Ness Theorem.

Example 2.2.8 (Torus Actions)**.** Consider a complex torus $T \subseteq \mathrm{GT}_N(\mathbb{C})$ and set $T_K := \{t \in T \mid t^\dagger t = \mathrm{I}_N\}$. Let $\pi\colon T \to \mathrm{GL}(V)$ be a rational representation. Then π admits a weight space decomposition $V = \bigoplus_{\omega\in\Omega(\pi)} V_\omega$, where $\Omega(\pi) \subseteq \mathfrak{i}\operatorname{Lie}(T_K)$ is the set of weights; compare Theorem 1.3.14. Equip V with an inner product as in Setting 2.2.2. We show that the weight spaces are pairwise orthogonal. Let $\omega, \epsilon \in \Omega(\pi)$ and choose $v_\omega \in V_\omega$, $v_\epsilon \in V_\epsilon$. As $\mathfrak{p} = \operatorname{Lie}(T) \cap \mathrm{Sym}_N(\mathbb{C}) = \mathfrak{i}\operatorname{Lie}(T_K)$ acts via self-adjoint operators and v_ω, v_ϵ are weight vectors (Definition 1.3.12), we compute for all $X \in \mathfrak{p}$

$$\operatorname{tr}(X\omega)\langle v_\omega, v_\epsilon\rangle = \langle \Pi(X)v_\omega, v_\epsilon\rangle = \langle v_\omega, \Pi(X)v_\epsilon\rangle = \operatorname{tr}(X\epsilon)\langle v_\omega, v_\epsilon\rangle.$$

If $\langle v_\omega, v_\epsilon\rangle \neq 0$, then $\operatorname{tr}(X\omega) = \operatorname{tr}(X\epsilon)$ holds for all $X \in \mathfrak{p}$. Since $\omega, \epsilon \in \mathfrak{p}$ we necessarily have $\omega = \epsilon$ by non-degeneracy of the trace inner product on $\mathfrak{p}$. By contraposition, distinct weight spaces are orthogonal. Therefore, writing $v = \sum_\omega v_\omega \in V$ we have for all $X \in \mathfrak{p}$ that

$$\begin{aligned}\operatorname{tr}\big(\mu_T(v)X\big) &= \frac{1}{\|v\|^2}\langle v, \pi(X)\sum_\omega v\rangle = \frac{1}{\|v\|^2}\Big\langle \sum_\epsilon v_\epsilon, \sum_\omega \operatorname{tr}(\omega X)v_\omega\Big\rangle \\ &= \frac{1}{\|v\|^2}\sum_\omega \operatorname{tr}(\omega X)\langle v_\omega, v_\omega\rangle = \operatorname{tr}\left(\sum_\omega \frac{\|v_\omega\|^2}{\|v\|^2}\omega X\right).\end{aligned}$$

Hence, the moment map at v is given by

$$\mu_T(v) = \sum_{\omega\in\Omega(\pi)} \frac{\|v_\omega\|^2}{\|v\|^2}\,\omega\,. \tag{2.9}$$

Let us end by specifying this in two special cases. First, let $T = \mathrm{GT}_d(\mathbb{C})$ act on $V = \mathbb{C}^m$ via the matrix $A \in \mathbb{Z}^{d\times m}$ with linearization $b \in \mathbb{Z}^d$ as in Example 1.3.16. Then $e_j \in \mathbb{C}^m$ is a weight vector for the weight $A_j - b$, where

A_j denotes the j^{th} column of A. For $v = (v_1, \dots, v_m) \in \mathbb{C}^m$, define the vector $v^{[2]} := (|v_1|^2, \dots, |v_m|^2)$. Then (2.9) becomes

$$\mu_T(v) = \sum_{j=1}^{m} \frac{|v_j|^2}{\|v\|^2} (A_j - b) = \frac{1}{\|v\|^2} Av^{[2]} - b = \frac{1}{\|v\|^2} \big(Av^{[2]} - \|v\|^2 b\big) . \tag{2.10}$$

Second, consider the matrix scaling action from Example 1.3.5: $T = \mathrm{ST}_m(\mathbb{C})^2$ acts on $\mathbb{C}^m \times \mathbb{C}^m \cong \mathbb{C}^{m\times m}$ via $\pi_{m,2}$. We know from Example 1.3.18 that $(\epsilon_i, \epsilon_j) \in (\mathbb{1}_m^{\perp})^2$ is a weight with weight vector $e_i \otimes e_j \cong E_{ij}$. Therefore, for $v = (v_ij) \in \mathbb{C}^{m\times m}$ (2.9) becomes

$$\mu_T(v) = \sum_{i,j=1}^{m} \frac{|v_{ij}|^2}{\|v\|^2} (\epsilon_i, \epsilon_j) = \frac{1}{\|v\|^2} \sum_{i,j=1}^{m} |v_{ij}|^2 \big((\epsilon_i, 0) + (0, \epsilon_j)\big)$$

Setting $M_v := (|v_{ij}|^2)_{i,j} \in \mathbb{C}^{m\times m}$, we compute that

$$\sum_{i,j=1}^{m} |v_{ij}|^2 (\epsilon_i, 0) = \sum_{i=1}^{m} (M_v)_{i,+} (e_i - m^{-1}\mathbb{1}_m, 0) = \left(\sum_{i=1}^{m} (M_v)_{i,+} e_i - \frac{M_{+,+}}{m} \mathbb{1}_m, 0 \right) .$$

Note that $M_{+,+} = \|v\|^2$ and that $\sum_i (M_v)_{i,+} e_i$ is the vector of row sums of M_v, which we denote by $r(M_v)$. A similar computation to the above holds for $c(M_v)$, the vector of column sums $(M_v)_{+,j}$. Altogether, we deduce that

$$\mu_T(v) = \frac{1}{\|v\|^2} \left(r(M_v) - \frac{\|v\|^2}{m} \mathbb{1}_m,\, c(M_v) - \frac{\|v\|^2}{m} \mathbb{1}_m \right) . \tag{2.11}$$

is the moment map at v for matrix scaling. ♢

Example 2.2.9 (Left Multiplication)**.** Let π be the action of $G = \mathrm{SL}_m(\mathbb{K})$ on $V = \mathbb{K}^{m\times n}$ via left-multiplication. Then K acts isometrically on V with respect to the Frobenius inner product. Moreover, for $X \in \mathrm{Lie}(G)$ and $Y \in V$ we have $\Pi(X)Y = XY$. In particular, $\Pi(X)$ is self-adjoint for $X \in \mathfrak{p}$. We compute for all $X \in \mathfrak{p}$

$$\begin{aligned} \mathrm{tr}(\mu_G(Y)X) &= \frac{1}{\|Y\|^2} \langle Y, \Pi(X)Y \rangle = \frac{1}{\|Y\|^2} \mathrm{tr}(Y^\dagger XY) \\ &\overset{(*)}{=} \frac{1}{\|Y\|^2} \mathrm{tr}(YY^\dagger X) - \frac{1}{m} \mathrm{tr}(X) = \mathrm{tr}\left(\left(\frac{YY^\dagger}{\|Y\|^2} - \frac{1}{m} \mathrm{I}_m \right) X \right) \end{aligned}$$

where we used $\mathrm{tr}(X) = 0$ in $(*)$. Note that $\mathrm{tr}(YY^\dagger/\|Y\|^2) = 1$ and hence it cannot be $\mu_G(Y) \in \mathfrak{p}$. However, subtracting $m^{-1}\,\mathrm{I}_m$ ensures we get a trace zero matrix in $\mathrm{Sym}_m(\mathbb{K})$, i.e., a matrix in $\mathfrak{p} = \mathrm{Lie}(G) \cap \mathrm{Sym}_m(\mathbb{K})$. Therefore,

$$\mu_G(Y) = \frac{YY^\dagger}{\|Y\|^2} - \frac{1}{m} \mathrm{I}_m = \frac{1}{\|Y\|^2} \left(YY^\dagger - \frac{\|Y\|^2}{m} \mathrm{I}_m \right) \tag{2.12}$$

gives the moment map. ♢

Example 2.2.10 (Action on a Quiver)**.** Consider the quiver Q

$$1 \xrightarrow{B_1} 2 \xleftarrow{B_2} 3 \tag{2.13}$$

with dimension vector $\alpha = (m, m, m)$. The labels in (2.13) indicate how $(B_1, B_2) \in V = \mathscr{R}(Q, \alpha) = (\mathbb{K}^{m\times m})^2$ is associated to the arrows. A group element $g \in G = \mathrm{SL}_\alpha(\mathbb{K}) = \mathrm{SL}_m(\mathbb{K})^3$ acts on V via

$$(g_1, g_2, g_3) \cdot (B_1, B_2) = (g_2 B_1 g_1^{-1}, g_2 B_2 g_3^{-1}),$$

compare Example 1.3.8. Let π be the corresponding representation. Equipping V with the standard inner product[7] the group K acts isometrically on V. Recall that we think of G, K and their Lie algebras as block diagonally embedded into $\mathrm{GL}_{3m}(\mathbb{K})$ respectively $\mathbb{K}^{3m\times 3m}$.[8] For $A \in \mathbb{K}^{m\times m}$ set

$$\Phi_1(A) := -A^\dagger A + \frac{\|A\|_F^2}{m}\,\mathrm{I}_m \qquad \text{and} \qquad \Phi_2(A) := AA^\dagger - \frac{\|A\|_F^2}{m}\,\mathrm{I}_m\,. \tag{2.14}$$

which are in $\mathrm{Sym}_m(\mathbb{K})$ and have trace zero as $\mathrm{tr}(A^\dagger A) = \mathrm{tr}(AA^\dagger) = \|A\|_F^2$. Therefore, $\Phi_1(A), \Phi_2(A) \in \mathfrak{q} := \mathrm{Lie}(\mathrm{SL}_m(\mathbb{K})) \cap \mathrm{Sym}_m(\mathbb{K})$. We will show that the moment map is given by

$$\mu_G(B) = \frac{1}{\|B\|^2}\big(\Phi_1(B_1), \Phi_2(B_1) + \Phi_2(B_2), \Phi_1(B_2)\big)\,. \tag{2.15}$$

First, note that for general $A \in \mathbb{K}^{m\times m}$ and $(X_1, X_2) \in \mathrm{Lie}(\mathrm{SL}_m(\mathbb{K}))^2$ we have

$$\frac{d}{dt}\Big|_{t=0} e^{tX_1} A e^{-tX_2} = \big(X_1 e^{tX_1} A e^{-tX_2} + e^{tX_1} A(-X_2) e^{-tX_2}\big)\big|_{t=0} = X_1 A - A X_2\,.$$

Therefore, $X = (X_1, X_2, X_3) \in \mathrm{Lie}(G)$ acts via

$$\Pi(X_1, X_2, X_3)(B_1, B_2) = (X_2 B_1 - B_1 X_1, X_2 B_2 - B_2 X_3).$$

In particular, $\mathfrak{p}$ acts via self-adjoint operators on V. By Definition 2.2.4, the moment map $\mu_G(B) = \|B\|^{-2}\big(\mu_1(B), \mu_2(B), \mu_3(B)\big) \in \mathfrak{p} \cong \mathfrak{q}^3$, is determined by

$$\begin{aligned}\sum_{i=1}^{3} \mathrm{tr}\,\big(\mu_i(B) X_i\big) &= \big\langle B, \Pi(X)B\big\rangle \\ &= \mathrm{tr}\,\big(B_1^\dagger(X_2 B_1 - B_1 X_1)\big) + \mathrm{tr}\,\big(B_2^\dagger(X_2 B_2 - B_2 X_3)\big)\end{aligned} \tag{2.16}$$

for all $X = (X_1, X_2, X_3) \in \mathfrak{p}$. Thus, using $X = (X_1, 0, 0) \in \mathfrak{p}$ we obtain with $\mathrm{tr}(X_1) = 0$ that

$$\begin{aligned}\mathrm{tr}\,\big(\mu_1(B) X_1\big) &= \mathrm{tr}\,\big(B_1^\dagger(-B_1 X_1)\big) \overset{(*)}{=} \mathrm{tr}\,\big(-B_1^\dagger B_1 X_1\big) + \frac{\|B_1\|_F^2}{m}\,\mathrm{tr}(X_1) \\ &= \mathrm{tr}\,\big(\Phi_1(B_1) X_1\big).\end{aligned}$$

[7] That is, the two copies $\mathbb{K}^{m\times m}$ are orthogonal to each other and each copy is equipped with the Frobenius inner product.

[8] For convenience, this is neglected in the notation; e.g., we write $X = (X_1, X_2, X_3) \in \mathrm{Lie}(G) \cong \mathrm{Lie}(\mathrm{SL}_m(\mathbb{K}))^3$ instead of $X = \mathrm{diag}(X_1, X_2, X_3)$.

Since $\Phi_1(B_1), \mu_1(B) \in \mathfrak{q}$, we deduce $\mu_1(B) = \Phi_1(B_1)$ by non-degeneracy of the trace inner product on $\mathfrak{q}$. Similarly, one shows $\mu_3(B) = \Phi_1(B_2)$. Finally, for $X = (0, X_2, 0) \in \mathfrak{p}$ in Equation (2.16) we obtain

$$\begin{aligned}\operatorname{tr}\left(\mu_1(B)X_2\right) &= \operatorname{tr}\left(B_1^\dagger X_2 B_1\right) + \operatorname{tr}\left(B_2^\dagger X_2 B_2\right) \\ &= \operatorname{tr}\left(B_1 B_1^\dagger X_2\right) + \operatorname{tr}\left(B_2 B_2^\dagger X_2\right) - \frac{\|B_1\|_F^2 + \|B_2\|_F^2}{m} \operatorname{tr}(X_2) \\ &= \operatorname{tr}\left((\Phi_2(B_1) + \Phi_2(B_2))X_2\right).\end{aligned}$$

We deduce $\mu_2(B) = \Phi_2(B_1) + \Phi_2(B_2)$ and hence (2.15) holds.

Analogously, one can consider $\alpha = (m, m, m)$, the quiver Q'

$$1 \xleftarrow{C_1} 2 \xrightarrow{C_2} 3 \tag{2.17}$$

and its associated action of $G = \mathrm{SL}_m(\mathbb{K})^3$ on $V = \mathscr{R}(Q', \alpha) = (\mathbb{K}^{m\times m})^2$. In that case, $g \in G$ acts on $C = (C_1, C_2) \in V$ via $g \cdot C = (g_1 C_1 g_2^{-1}, g_3 C_2 g_2^{-1})$ and

$$\mu_G(C) = \frac{1}{\|C\|^2}\left(\Phi_2(C_1), \Phi_1(C_1) + \Phi_1(C_2), \Phi_2(C_2)\right). \tag{2.18}$$

is the moment map at C. ♢

Example 2.2.11 (Left-Right Action)**.** Consider the left-right action of $G = \mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ on $V = (\mathbb{K}^{m_1\times m_2})^n$ from Example 1.3.3. One computes that

$$\mu_G(Y) = \frac{1}{\|Y\|^2}\left(\sum_{i=1}^n Y_i Y_i^\dagger - \frac{\|Y\|^2}{m_1}\mathrm{I}_{m_1}, \left(\sum_{i=1}^n Y_i^\dagger Y_i\right)^{\mathsf{T}} - \frac{\|Y\|^2}{m_2}\mathrm{I}_{m_2}\right) \tag{2.19}$$

is the moment map at $Y = (Y_1, \ldots, Y_n) \in V$. ♢

Example 2.2.12 (Tensor Scaling)**.** Let $\pi_{m,d}$ be the natural action of $G = \mathrm{SL}_m(\mathbb{K})^d$ on $V = (\mathbb{K}^m)^{\otimes d}$. For a tensor $v = (v_{i_1,\ldots,i_d}) \in V$, consider its flattenings $M_1, \ldots, M_d \in \mathbb{K}^{m\times m^{d-1}}$ into the d many directions, e.g., $(M_1)_{i_1,(i_2,\ldots,i_d)} = v_{i_1,\ldots,i_d}$. One can compute that the moment map of $\pi_{m,d}$ is given by

$$\mu_G(v) = \frac{1}{\|v\|^2}\left(M_1 M_1^\dagger - \frac{\|v\|^2}{m}\mathrm{I}_m, \ldots, M_d M_d^\dagger - \frac{\|v\|^2}{m}\mathrm{I}_m\right). \tag{2.20}$$

The matrices $M_l M_l^\dagger$, $l \in [d]$ are called *(one-body) quantum marginals* of v. Usually, they are considered for $\mathbb{K} = \mathbb{C}$ and they play an important role in quantum information theory, see corresponding references in Section 3.1. Thus, Equation (2.20) links invariant theory via tensor scaling to this research area. ♢

The Theorem of Kempf-Ness

In the following we state the Kempf-Ness Theorem, which gives criteria to detect semi- and polystability. It was first proven by Kempf and Ness in [KN79] over $\mathbb{C}$. The real case is due to Richardson and Slodowy [RS90], and their result allows to deduce the complex case as well [RS90, Remark 4.5(d)].

First, let us give some intuition for the statement. Remember that a vector v is semistable if and only if the Kempf-Ness function F_v, see (2.8), is bounded from below. An important property of F_v is its geodesic convexity on the manifold $P = \{g^\dagger g \mid g \in G\}$ of positive definite matrices in G [BFG+19, Proposition 3.13]; also compare Theorem 1.2.18 and Example 1.2.21. Similarly to convexity in the Euclidean sense, geodesic convex functions on P achieve a global minimum at a point if and only if their gradient vanishes at the point.[9]

There are several statements to which one refers as (part of) Kempf-Ness Theorem. We collect them in Theorem 2.2.13, whose formulation is based on [AKRS21a, Theorem 2.2].

Theorem 2.2.13 (Kempf-Ness Theorem)**.** *Consider the Setting 2.2.2. In particular, $G \subseteq \mathrm{GL}_N(\mathbb{K})$ is Zariski closed and self-adjoint and $K = \{g \in G \mid g^\dagger g = \mathrm{I}_N\}$. Moreover, $\pi\colon G \to \mathrm{GL}(V)$ is a rational representation over $\mathbb{K}$ with moment map μ. For $v \in V\backslash\{0\}$, we have:*

(a) The vector v is of minimal norm in its orbit if and only if $\mu(v) = 0$.

(b) Let v be of minimal norm in its orbit. If $X \in \mathfrak{p}$ satisfies $\|e^X \cdot v\| = \|v\|$, then $X \cdot v = 0$. If $w \in G \cdot v$ is such that $\|v\| = \|w\|$, then $w \in K \cdot v$.

(c) If the orbit $G \cdot v$ is closed, then there exists some $w \in G \cdot v$ with $\mu(w) = 0$.

(d) If $\mu(v) = 0$, then the orbit $G \cdot v$ is closed.

(e) The vector v is polystable if and only if there exists $0 \neq w \in G \cdot v$ with $\mu(w) = 0$.

(f) The vector v is semistable if and only if there exists $0 \neq w \in \overline{G \cdot v}$ with $\mu(w) = 0$.

We can replace G by any Euclidean closed subgroup $H \subseteq G$ with $G^\circ \subseteq H$. In this case, K is replaced by $K' = \{h \in H \mid h^\dagger h = \mathrm{I}_N\}$.

Proof. For $\mathbb{K} = \mathbb{R}$: note that our Setting 2.2.2 fits into the framework of [RS90]. In the latter work, $G \subseteq \mathrm{GL}(E)$ is stable under a Cartan involution, which just means there is an inner product on E to which G is self-adjoint. For us, $E = \mathbb{R}^N$ is equipped with the standard inner product. Our $\mathfrak{p} = \mathrm{Lie}(G) \cap \mathrm{Sym}_N(\mathbb{R})$ is the -1 eigenspace of $\theta\colon \mathrm{Lie}(G) \to \mathrm{Lie}(G), X \mapsto -X^\mathsf{T}$, and hence agrees with the $\mathfrak{p}$ in [RS90]. Moreover, the inner product from Setting 2.2.2 is K-invariant and $\pi(X)$ is self-adjoint for all $X \in \mathfrak{p}$ as required by [RS90, §3].

Now, Part (a) is the equivalence of (i) and (iii) in [RS90, Theorem 4.3]. Item (b) is the last part of [RS90, Theorem 4.3] plus Lemma 4.2, which ensures the statement on $X \in \mathfrak{p}$. [RS90, Theorem 4.4] yields parts (c), (d) and (e). Finally, part (f) follows from the fact that any orbit closure $\overline{G \cdot v}$ contains a unique closed orbit ([Lun75, Theoreme 2.7])[10], which is not the zero orbit if and only if v is semistable.

[9] For this, the facts from Theorem 1.2.18 that P is a totally geodesic manifold and has non-positive curvature are crucial.

[10] also see [RS90, §9.3] or [BL21, Theorem 1.1(iii)]

For $\mathbb{K} = \mathbb{C}$: by [RS90, Remark 4.5(d)] it follows from the real case. Still, let us refer to the original paper [KN79]. Parts (a) respectively (b) are [KN79, Theorem 0.1(a) respectively (b)], while [KN79, Theorem 0.2] yields items (c), (d) and (e).[11] Part (f) again follows from the fact that any orbit closure $\overline{G \cdot v}$ contains a unique closed orbit, Theorem 1.4.7. We note that the assumption in [KN79] of G being connected is unnecessary.[12]

For H being a Euclidean closed subgroup with $G^\circ \subseteq H$, note that H is self-adjoint by Corollary 1.2.17. Thus, for $\mathbb{K} = \mathbb{R}$ it follows from the general setting of [RS90].[13] If $\mathbb{K} = \mathbb{C}$ note that H is Zariski closed, because it consists of several connected components of G that are all Zariski closed as $G^\circ = G^{\circ,\mathrm{Z}}$ over $\mathbb{C}$ (compare Section 1.1). Hence, H is Zariski closed and self-adjoint which puts us again in Setting 2.2.2. □

Remark 2.2.14 (Further Literature). Parts (a)–(d) of Theorem 2.2.13 are the formulations of [Wal17, Theorems 3.26 and 3.28]. However, one needs to be careful: Wallach directly works with a Zariski closed self-adjoint subgroup of $\mathrm{GL}(V)$, but $\pi(G) \subseteq \mathrm{GL}(V)$ may not be Zariski closed for $\mathbb{K} = \mathbb{R}$, compare Example 1.1.8.

Still, if $\mathbb{K} = \mathbb{R}$ we know from Proposition 1.2.5 that $\pi(G) \subseteq \mathrm{GL}(V)$ is a Euclidean closed Lie subgroup. Furthermore, in Setting 2.2.2 the inner product $\langle \cdot, \cdot \rangle$ on V is K-invariant and for all $X \in \mathfrak{p}$ the operator $\Pi(X)$ is self-adjoint. Thus, the polar decomposition on G induces a polar decomposition on $\pi(G)$ and hence $\pi(G)$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle$. Altogether, $\pi(G) \subseteq \mathrm{GL}(V)$ satisfies the assumptions of [Bil21; BL21] and hence one can deduce Kempf-Ness over $\mathbb{R}$ also from the formulations in [BL21, Theorem 1.1] respectively [Bil21, Theorem 1]. ▽

For Computational Invariant Theory an important consequence of Kempf-Ness Theorem 2.2.13(f) is a "duality" between capacity and moment map:

$$\mathrm{cap}_G(v) = 0 \qquad \Leftrightarrow \qquad 0 < \inf_{g \in G} \|\mu_G(g \cdot v)\|_F = \min_{0 \neq w \in \overline{G \cdot v}} \|\mu_G(w)\|_F. \tag{2.21}$$

We revisit this in Part II, where we state a quantitive version in Theorem 3.2.5. Next, let us illustrate Kempf-Ness in an example.

Example 2.2.15. Consider the left multiplication of $G = \mathrm{SL}_m(\mathbb{K})$ on $V = \mathbb{K}^{m \times n}$. We know from Example 1.4.4 that $Y \in V$ is either unstable or stable. The latter case happens if and only if Y has full row rank. Now, assume that Y is stable. To illustrate Kempf-Ness, Theorem 2.2.13, we determine an element of minimal norm in $G \cdot Y$ and, as a sanity check, show that the moment map vanishes.

This problem is classical and we follow the explanations below Equation (2.2) in [BGO+18]. First, note that the AM-GM inequality for the eigenvalues of a positive semi-definite matrix $\Psi \in \mathbb{K}^{m \times m}$ translates to $\mathrm{tr}(\Psi) \geq m(\det(\Psi))^{1/m}$. With this inequality we compute that for all $g \in \mathrm{SL}_m(\mathbb{K})$

$$\|g \cdot Y\|^2 = \mathrm{tr}\left(gYY^\dagger g^\dagger\right) \geq m\big(\det\left(gYY^\dagger g^\dagger\right)\big)^{1/m} = m \det\left(YY^\dagger\right)^{1/m}.$$

[11] Note that "stable" in [KN79] means polystable in our sense.

[12] Indeed, [RS90] does not assume this.

[13] [RS90] also assumes H to be Zariski dense, but this is only needed in [RS90, §6] and not in §3 and §4 which prove Kempf-Ness. Alternatively, one can deduce the statement on H from [BL21], see Remark 2.2.14 below.

Setting $M := YY^\dagger$, we have that $\mathrm{cap}_G(Y) \geq m \det(M)^{1/m}$. In fact, equality holds as follows. As Y has full row rank the matrix M is invertible, so $M \in \mathrm{PD}_m(\mathbb{K})$. Let $M^{1/2} \in \mathrm{PD}_m(\mathbb{K})$ be the square root and set $h := \det(M)^{1/(2m)} M^{-1/2} \in G$. We compute

$$\begin{aligned}(hY)(hY)^\dagger &= \det(M)^{1/m} M^{-1/2} Y Y^\dagger M^{-1/2} \\ &= \det(M)^{1/m} M^{-1/2} M M^{-1/2} = \det(M)^{1/m} \mathrm{I}_m \,.\end{aligned} \tag{2.22}$$

Therefore, $\|h \cdot Y\|^2 = \det(M)^{1/m} \operatorname{tr}(\mathrm{I}_m) = m \det(M)^{1/m}$ and we necessarily have

$$\mathrm{cap}_G(Y) = \|h \cdot Y\|^2 = m \det(M)^{1/m}.$$

We see that $h \cdot Y$ is of minimal norm in $G \cdot Y$ and hence Y is indeed polystable by Kempf-Ness Theorem 2.2.13. Using (2.22) and the value for $\|h \cdot Y\|^2$ we obtain

$$(hY)(hY)^\dagger - \frac{\|hY\|^2}{m} \mathrm{I}_m = \det(M)^{1/m} \mathrm{I}_m - \frac{m \det(M)^{1/m}}{m} \mathrm{I}_m = 0.$$

Hence, $\mu_G(h \cdot Y) = 0$ by Equation (2.12) in Example 2.2.9. ♢

In the following we present three statements which fall into the realm of Kempf-Ness.

Lemma 2.2.16. *Consider the Setting 2.2.2. Let $v \in V$ be of minimal norm in its orbit. Then the stabilizer G_v is Zariski closed and self-adjoint.*

Proof. The same proof as for [Wal17, Corollary 2.25] applies. First, recall from Section 1.1 that G_v is Zariski closed as the action via π is algebraic. To show self-adjointness, use the polar decomposition (Theorem 1.2.16) to write $g = k \exp(X) \in X \in G_V$ with $k \in K$ and $X \in \mathfrak{p}$. Then $X = X^\dagger$ yields $g^\dagger = \exp(X^\dagger) k^{-1} = \exp(X) k^{-1}$. Now, $g \in G_v$ and K acting isometrically imply $\|v\| = \|g \cdot v\| = \|\exp(X) v\|$. Kempf-Ness Theorem 2.2.13(b) yields $\Pi(X) v = 0$ and hence $\pi(\exp(X)) v = \exp(\Pi(X)) v = v$. That is, $\exp(X) \in G_v$ and thus $k = g \exp(X)^{-1} \in G_v$. Altogether, $g^\dagger \in G_v$. □

Proposition 2.2.17. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be Zariski closed and self-adjoint with Euclidean identity component G°. Set $K := \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$.*

(i) Then there exist finitely many $k_1 = \mathrm{I}_m, k_2, \ldots, k_l \in K$ such that the $k_i G^\circ$ are the Euclidean connected components of G.

(ii) If $\pi \colon G \to \mathrm{GL}(V)$ is a rational representation over $\mathbb{K}$ and K acts isometrically on V with respect to some inner product, then the stability notions for G and G° coincide.

Proof. For part (i), remember that G has only finitely many Euclidean connected components, since it is algebraic. Moreover, $\exp(X) \in G^\circ$ for all $X \in \mathrm{Lie}(G)$, compare Proposition 1.2.8. Therefore, the polar decomposition (Theorem 1.2.16) yields part (i).

For part (ii), let $v \in V$. First, we have $\mathrm{cap}_G(v) = \mathrm{cap}_{G^\circ}(v)$ using part (i) and that K acts isometrically on V. Thus, v is G-unstable/semistable if and only is v if G°-unstable/semistable.

For part(iii), note that we can apply Kempf-Ness to G and G°. Combining Theorem 2.2.13(a) and (e) yields that v is polystable if and only if its capacity is positive and attained. Since K acts isometrically, part (i) shows that $\mathrm{cap}_G(v) = \mathrm{cap}_{G^\circ}(v)$ is attained by some $g \in G$ if and only if it is attained by some $g' \in G^\circ$. Hence, v is G-polystable if and only if it is G°-polystable.

Finally, to ensure the same for "stable" it suffices to show that G_v is finite if and only if $(G^\circ)_v$ is finite. If G_Y is finite, then $(G^\circ)_v$ is finite as $(G^\circ)_v \subseteq G_v$. For the converse, note that $(G_v)^\circ \subseteq G^\circ$ and hence $(G_v)^\circ \subseteq (G^\circ)_v$. Moreover, G_v is Zariski closed, so $G_v/(G_v)^\circ$ is finite. Altogether, if $(G^\circ)_v$ is finite, then $(G_v)^\circ$ is finite and so is G_v. □

We end with the fact that, in a complex setting compatible with the real structures, the capacity of a real vector is independent of $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. This has interesting algorithmic implications: when approximating the capacity of a real vector it allows to use to use algorithms over $\mathbb{C}$, e.g., as in [BFG+19].

Let $G_\mathbb{C} \subseteq \mathrm{GL}_N(\mathbb{C})$ be Zariski closed, self-adjoint and defined over $\mathbb{R}$. Then $G_\mathbb{R} := G_\mathbb{C} \cap \mathrm{GL}_N(\mathbb{R})$ is Zariski closed and self-adjoint. Consider a rational representation $\pi\colon G_\mathbb{C} \to \mathrm{GL}(V_\mathbb{C})$ defined over $\mathbb{R}$. Then $\pi_\mathbb{R}\colon G_\mathbb{R} \to \mathrm{GL}(V_\mathbb{R})$ is a rational representation of $G_\mathbb{R}$. Equip $V_\mathbb{C}$ with a Hermitian inner product $\langle \cdot, \cdot \rangle$ on $V_\mathbb{C}$ that is invariant under $K := G \cap \mathrm{U}_N$ and compatible with $V_\mathbb{R}$, i.e., $\langle v, w \rangle \in \mathbb{R}$ for all $v, w \in V_\mathbb{R}$. This puts us into the setting of [RS90, §8].

Proposition 2.2.18 (based on [AKRS21a, Proposition 2.3])**.**
Assume the setting above. Let $\mathrm{cap}_{G_\mathbb{K}}(v)$ *be the capacity of* $v \in V_\mathbb{K}$ *under* $G_\mathbb{K}$ *and let* $\mathcal{N}_\mathbb{K} = \{v \in V_\mathbb{K} \mid \mathrm{cap}_{G_\mathbb{K}}(v) = 0\}$ *be the null cone under the action of* $G_\mathbb{K}$ *on* $V_\mathbb{K}$*.*

(i) For $v \in V_\mathbb{R}$*, we have the equality of capacities* $\mathrm{cap}_{G_\mathbb{R}}(v) = \mathrm{cap}_{G_\mathbb{C}}(v)$*. In particular,* $\mathcal{N}_\mathbb{R} = \mathcal{N}_\mathbb{C} \cap V_\mathbb{R}$*.*

(ii) $\mathcal{N}_\mathbb{R} = V_\mathbb{R}$ *if and only if* $\mathcal{N}_\mathbb{C} = V_\mathbb{C}$*.*

Proof. For part (i), we have $\mathrm{cap}_{G_\mathbb{R}}(v) \geq \mathrm{cap}_{G_\mathbb{C}}(v)$ as $G_\mathbb{R} \subseteq G_\mathbb{C}$. Regarding the converse inequality, the capacity $\mathrm{cap}_{G_\mathbb{K}}(v)$ is attained at all elements of minimal norm in the closed orbit contained in $\overline{G_\mathbb{K} \cdot v}$, by Kempf-Ness Theorem 2.2.13. Hence, we can reduce to studying a closed orbit $G_\mathbb{R} \cdot v$. If w is of minimal norm in $G_\mathbb{R} \cdot v$, then it is of minimal norm in $G_\mathbb{C} \cdot v$ by [RS90, Lemma 8.1]. Thus, $G_\mathbb{C} \cdot w$ is closed by Kempf-Ness and hence $\|w\|^2 = \mathrm{cap}_{G_\mathbb{C}}(v)$. This shows (i).

For part (ii), $\mathcal{N}_\mathbb{C} = V_\mathbb{C}$ directly implies $\mathcal{N}_\mathbb{R} = V_\mathbb{R}$. Conversely, $V_\mathbb{R}$ is Zariski dense in the irreducible complex variety $V_\mathbb{C}$, so $\mathcal{N}_\mathbb{R} = V_\mathbb{R}$ yields that $\mathcal{N}_\mathbb{C}$ contains the Zariski dense subset $\mathcal{N}_\mathbb{R}$. As $\mathcal{N}_\mathbb{C}$ is Zariski closed in $V_\mathbb{C}$ (see Remark 1.4.8), we must have $\mathcal{N}_\mathbb{C} = V_\mathbb{C}$. □

Moment Polytopes

We explain how the moment maps induces so-called moment polytopes. They generalize weight polytopes, which arise in the case of torus actions. These polytopes be used to express the duality in (2.21). Moreover, the combinatorics of

these polytopes captures important complexity measures studied in Chapter 4. In the latter we only work over $\mathbb{C}$. Therefore, we restrict in the following to the complex numbers, and only comment on real moment polytopes in Remark 2.2.21.

As a motivation of moment polytopes, we first describe how weight polytopes arise as images of the moment map. For this, assume the Setting 2.2.2 for $\mathbb{K} = \mathbb{C}$, $G = T$ being a complex torus, and $\pi\colon T \to \mathrm{GL}(V)$ a rational representation with set of weights $\Omega(\pi)$. Remember that V admits a weight space decomposition $V = \bigoplus_{\omega\in\Omega(\pi)} V_\omega$ and hence for $v \in V$ we have $v = \sum_\omega v_\omega$ for some $v_\omega \in V_\omega$. The weight polytope of v is $\Delta_T(v) = \mathrm{conv}\{\omega \mid v_\omega \neq 0\}$. We know from (2.9) in Example 2.2.8 that the moment map at v is

$$\mu_T(v) = \sum_{\omega\in\Omega(\pi)} \frac{\|v_\omega\|^2}{\|v\|^2}\,\omega\,.$$

Moreover, we have seen in Example 2.2.8 that the weight spaces V_ω are pairwise orthogonal. Therefore, $\mu_T(v)$ is a convex combination of the weights and hence $\mu_T(v)$ lies in the *relative* interior of $\Delta_T(v)$, i.e., $\mu_T(v) \in \mathrm{relint}(\Delta_T(v))$. In fact, it was proven independently by Atiyah [Ati82, Theorem 2] and by Guillemin-Sternberg [GS84, Theorem 4] that

$$\mathrm{relint}\,\Delta_T(v) = \mu(\mathrm{GT}_d\cdot v) \qquad \text{and so} \qquad \Delta_T(v) = \overline{\{\mu_T(t\cdot v) \mid t\in T\}}\,. \tag{2.23}$$

The statements in [Ati82; GS84] rather apply to a projectivized setting. We provide a brief translation for readers that are unfamiliar with these topics.

Remark 2.2.19 (based on [AKRS21b, Remark B.1]). Remember that the moment map $\mu_T\colon V\backslash\{0\} \to \mathfrak{p} = \mathfrak{i}\,\mathrm{Lie}(T_K)$ is invariant under non-zero scalars and therefore factors through the projective space $\mathbb{P}(V)$ via a map $\bar\mu\colon \mathbb{P}(V) \to \mathfrak{p}$. For a non-zero $v \in V$, let $[v]$ be the point in $\mathbb{P}(V)$ that represents the line $\mathbb{C}v$. Then T naturally acts on $\mathbb{P}(V)$ and $\bar\mu$ is the moment map for this action. This action fits the setting of [Ati82; GS84], because $\mathbb{P}(V)$ is a compact Kähler manifold.

The results [Ati82, Theorem 2] and [GS84, Theorem 4] give

$$\Delta_T(v) = \bar\mu\left(\overline{T\cdot[v]}\right).$$

For (2.23), we need a statement for the orbit of v rather than the orbit closure of $[v]$. The closure $\overline{T\cdot[v]}$ is the disjoint union of finitely many T orbits. The orbits relate to $\Delta_T(v)$ as follows. For each open face F of $\Delta_T(v)$ the set $\bar\mu^{-1}(F)\cap\overline{T\cdot[v]}$ is a single T-orbit in $\mathbb{P}(V)$, [Ati82, Theorem 2]. In particular, for $F = \mathrm{relint}\,\Delta_T(v)$ we obtain the orbit $T\cdot[v]$. This yields (2.23), since $\bar\mu(T\cdot[v]) = \mu(T\cdot v)$. ▽

We point out how Equation (2.23) connects Hilbert-Mumford and Kempf-Ness for torus actions. By Hilbert-Mumford Theorem 2.1.9(c) polystability is equivalent to $0 \in \mathrm{relint}\,\Delta_T(v)$, which translates with (2.23) to $0 \in \mu(T\cdot v)$. The latter is equivalent to the statement for polystability in Kempf-Ness, Theorem 2.2.13(e).

The fact that the image of the moment map yields a polytope remarkably generalizes to the non-commutative setting, giving so-called *moment polytopes*.

We need the latter only in the case $G = \mathrm{SL}_m(\mathbb{C})^d$. Thus for concreteness, assume the Setting 2.2.2 for $G = \mathrm{SL}_m(\mathbb{C})^d$ and corresponding moment map μ_G. Then for fixed $v \in V \setminus \{0\}$, the set $\{\mu_G(g \cdot v) \mid g \in G\}$ gives rise to a polytope as follows.

Let spec: $\mathrm{Sym}_m(\mathbb{C}) \to \mathbb{R}^m$ be the function sending a Hermitian matrix to its eigenvalues in decreasing order. Recalling that $\mathfrak{i}\,\mathrm{Lie}(K) \subseteq \mathrm{Sym}_m(\mathbb{C})^d$ is block-diagonally embedded in $\mathbb{C}^{dm \times dm}$, we set

$$s\colon \mathfrak{i}\,\mathrm{Lie}(K) \to (\mathbb{R}^m)^d, \quad \mathrm{diag}(X_1, \dots, X_d) \mapsto \big(\mathrm{spec}(X_1), \dots, \mathrm{spec}(X_d)\big).$$

Then for $v \in V \setminus \{0\}$ the set

$$\Delta_G(v) := \overline{\big\{s\big(\mu_G(w)\big) \mid w \in G \cdot v\big\}} \tag{2.24}$$

is a convex polytope with rational vertices, see [Bri87], [GS84], [Kir84a] or [Nes84, Appendix] by Mumford. We call $\Delta_G(v)$ the *moment polytope* of v. Noting that $\|X\|_F = \|\,\mathrm{spec}(X)\|_2$ for any $X \in \mathrm{Sym}_m(\mathbb{C})$ we have $\|\mu_G(v)\|_F = \|s(\mu_G(v))\|_2$ for all $v \in V \setminus \{0\}$. Thus, we can formulate the duality from Equation (2.21) also as follows:

$$\mathrm{cap}_G(v) = 0 \qquad \Leftrightarrow \qquad \mathrm{dist}\big(0, \Delta_G(v)\big) > 0 \qquad \Leftrightarrow \qquad 0 \notin \Delta_G(v). \tag{2.25}$$

This will motivate the definition of two precision parameters in Definition 4.1.1. Moreover, remember that Equation (2.25) for a torus $G = T$ is Theorem 2.1.9(b), which we proved via the Hilbert-Mumford Criterion (Theorem 2.1.2) and a version of linear programming duality (Corollary 2.1.8). Furthermore, we can also obtain it via Kempf-Ness Theorem 2.2.13(f) and Equation (2.23). Therefore, we can regard the duality via Hilbert-Mumford Theorem 2.1.3 and Kempf-Ness Theorem 2.2.13, and the dualities in Equations (2.21) and (2.25), as a generalization of linear programming duality.

Let us briefly comment on how to define $\Delta_G(v)$ for an arbitrary group G.

Remark 2.2.20 (General Definition of $\Delta_G(v)$). For a general group G as in Setting 2.2.2, one can fix a fundamental Weight chamber[14] $C(G) \subseteq \mathfrak{i}\,\mathrm{Lie}(T_K) \subseteq \mathbb{R}^N$, see [Hal15, Definition 8.20] or [GW09, Definition 3.1.11]. For any $X \in \mathfrak{p} = \mathfrak{i}\,\mathrm{Lie}(K)$, this chamber $C(G)$ intersects the $\mathrm{Ad}(K)$-orbit $\{kXk^\dagger \mid k \in K\}$ in a single point, denoted $s(X)$. This yields the moment polytope $\Delta_G(v)$, defined exactly as in (2.24). Note that for any $X \in \mathfrak{p}$ there is some $k \in K$ with $s(X) = kXk^\dagger$, and so $\|s(X)\| = \|kXk^\dagger\| = \|X\|$ by unitary invariance of the Frobenius norm. Thus, Equation (2.25) holds in general.

If $G = \mathrm{SL}_m(\mathbb{C})$, then the positive Weyl chamber is

$$C(G) = \{\mathrm{diag}(x) \mid x \in \mathbb{R}^m,\, x_+ = 0,\, x_1 \geq x_2 \geq \cdots \geq x_m\} \subseteq \mathfrak{i}\,\mathrm{Lie}(T_K) \cong \mathbb{1}_m^\perp\,.$$

For $X \in \mathfrak{i}\,\mathrm{Lie}(\mathrm{SU}_m)$ we indeed have $\{kXk^\dagger \mid k \in \mathrm{SU}_m\} \cap C(G) = \{\mathrm{spec}(X)\}$. $\triangledown$

We end by giving references for moment polytopes in the real case.

[14] It is also called positive Weyl chamber. In our concrete setting $G \subseteq \mathrm{GL}_N(\mathbb{C})$, a natural choice is to take the fundamental Weyl chamber with respect to the group $G \cap \mathrm{B}_N(\mathbb{C})$ of upper triangular matrices in G.

Remark 2.2.21 (Moment Polytopes for $\mathbb{K} = \mathbb{R}$). Interestingly, one can as well consider moment polytopes over the reals, which can then be described as subpolytopes of complex moment polytopes [OS00, Theorem 3.1]. Recent studies on the facets of these real moment polytopes can be found in the preprint [Par20]. We refer to [OS00; Par20] and the literature therein for further information on real moment polytopes. ▽

This remark naturally leads to Question 4.2.3.

2.3 King's Criterion for Quivers

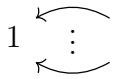
This section is based on [AKRS21a, Appendix A]. Its aim is to characterize stable elements under the left-right action when $\mathbb{K} = \mathbb{C}$. For this, we can use results from [Kin94] on stability of quiver representations. The main result is the following.

Theorem 2.3.1. *Consider the left-right action of $H := \mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ on $V := (\mathbb{C}^{m_1 \times m_2})^n$. Then $Y = (Y_1, \dots, Y_n) \in V$ is stable under H if and only if*

(i) the matrix $(Y_1 | \dots | Y_n) \in \mathbb{C}^{m_1 \times n m_2}$ has rank m_1, and

(ii) for all subspaces $V_1 \subseteq \mathbb{C}^{m_1}$, $\{0\} \subsetneq V_2 \subsetneq \mathbb{C}^{m_2}$ that satisfy $Y_i V_2 \subseteq V_1$ for all $i \in [n]$, one has $m_2 \dim V_1 > m_1 \dim V_2$.

In the following, we explain how to deduce Theorem 2.3.1 from [Kin94]. For concreteness, we directly restrict the general setting in [Kin94] to the quiver of interest. Let Q be the n-Kronecker quiver with two vertices and n arrows:



Recall from Example 1.3.8 that given a dimension vector $\alpha = (m_1, m_2)$ the groups $G := \mathrm{GL}_\alpha(\mathbb{C}) = \mathrm{GL}_{m_1}(\mathbb{C}) \times \mathrm{GL}_{m_2}(\mathbb{C})$ and $H := \mathrm{SL}_\alpha(\mathbb{C}) = \mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ act on $V = \mathscr{R}(Q, \alpha) \cong (\mathbb{C}^{m_1} \times \mathbb{C}^{m_2})^n$ via

$$(g_1, g_2) \cdot (Y_1, \dots, Y_n) = (g_1 Y_1 g_2^{-1}, \dots, g_1 Y_n g_2^{-1}) .$$

After precomposition with the automorphism $(g_1, g_2) \mapsto (g_1, g_2^{-\mathsf{T}})$ this is the left-right action of G (respectively H) on V, compare Example 1.3.3. Thus, $Y \in V$ is semi/poly/stable under the H-Kronecker quiver action if and only if it is semi/poly/stable under the H-left-right action. Hence, we can deduce Theorem 2.3.1 by considering the Kronecker quiver action.

For this, we need another action of $G = \mathrm{GL}_\alpha(\mathbb{C})$ from [Kin94]. Let χ_θ be the character of G given by $\theta := (m_2, -m_1)$, i.e., $\chi_\theta(g_1, g_2) = \det(g_1)^{m_2} \det(g_2)^{-m_1}$. We consider the action of G on $V \times \mathbb{C}$, where G acts on V by the Kronecker quiver action and on $\mathbb{C}$ by the character χ_θ^{-1}, i.e.,

$$g \cdot (X, z) := (g \cdot X, \chi_\theta^{-1}(g) z), \quad \text{where} \quad \chi_\theta^{-1}(g) = \det(g_1)^{-m_2} \det(g_2)^{m_1}. \tag{2.26}$$

Given $Y \in V$, we usually consider this action for $\hat{Y} := (Y, 1)$. Note that $\langle \theta, \alpha \rangle = 0$; an important assumption in [Kin94] which ensures that the central subgroup

$$\Delta := \left\{ (t\,\mathrm{I}_{m_1}, t\,\mathrm{I}_{m_2}) \mid t \in \mathbb{C}^\times \right\} \subseteq G$$

is always contained in the stabilizer $G_{\hat{Y}}$. In [Kin94, Definition 2.1] defines χ_θ-(semi)stability for Y. For us, the following characterizations are important.[15]

Lemma 2.3.2 ([Kin94, Lemma 2.2])**.** *Let $Y \in V = (\mathbb{C}^{m_1 \times m_2})^n = \mathscr{R}(Q, \alpha)$ and set $\hat{Y} := (Y, 1) \in V \times \mathbb{C}$. Then*

(a) Y is χ_θ-semistable if and only if $(V \times \{0\}) \cap \overline{G \cdot \hat{Y}} = \emptyset$.

(b) Y is χ_θ-stable if and only if $G \cdot \hat{Y}$ is closed and $G_{\hat{Y}}/\Delta$ is finite.[16]

To prove Theorem 2.3.1, we will later show that Y is χ_θ-stable if and only if it is H-stable. The items (i) and (ii) from Theorem 2.3.1 stem from the following stability notions.

Definition 2.3.3 ([Kin94, Definition 1.1])**.** Let $Y \in V = (\mathbb{C}^{m_1 \times m_2})^n = \mathscr{R}(Q, \alpha)$. We write $(\mathbb{C}^{m_1}, \mathbb{C}^{m_2}; Y)$ if we want to stress that we view Y as a representation of the Kronecker quiver (Definition 1.3.7). We say Y is *θ-semistable* if for all quiver-subrepresentations of $(\mathbb{C}^{m_1}, \mathbb{C}^{m_2}; Y)$, i.e., all subspaces $V_1 \subseteq \mathbb{C}^{m_1}$, $V_2 \subseteq \mathbb{C}^{m_2}$ with $Y_i V_2 \subseteq V_1$ for all i, we have

$$\langle \theta, (\dim V_1, \dim V_2) \rangle = m_2 \dim V_1 - m_1 \dim V_2 \geq 0. \tag{2.27}$$

Y is *θ-stable* if the inequality in (2.27) is strict for all non-zero proper subrepresentations. Here, non-zero means $V_1 \neq 0$ or $V_2 \neq 0$, while proper means $V_1 \subsetneq \mathbb{C}^{m_1}$ or $V_2 \subsetneq \mathbb{C}^{m_2}$. ▲

The concepts of θ-(semi)stability and χ_θ-(semi)stability agree.

Proposition 2.3.4 ([Kin94, Proposition 3.1])**.** *Let $Y \in V = (\mathbb{C}^{m_1 \times m_2})^n$. Then Y is χ_θ-semistable (respectively χ_θ-stable) if and only if Y is θ-semistable (respectively θ-stable).*

To show that Y is χ_θ-stable if and only if it is H-stable, we provide a lemma.

Lemma 2.3.5 ([AKRS21a, Lemma A.1])**.** *Let $Y \in V = (\mathbb{C}^{m_1 \times m_2})^n$ and $z \in \mathbb{C}^\times$, and set $\hat{Y} := (Y, 1) \in V \times \mathbb{C}$. Fix an $(m_1 m_2)$-root function on $\mathbb{C}$. Then*

(a) $(X, z) \in G \cdot \hat{Y} \quad \Leftrightarrow \quad z^{\frac{1}{m_1 m_2}} X \in H \cdot Y$

(b) $(X, z) \in \overline{G \cdot \hat{Y}} \quad \Leftrightarrow \quad z^{\frac{1}{m_1 m_2}} X \in \overline{H \cdot Y}$

(c) $\left(\exists\, X \in V \colon (X, 0) \in \overline{G \cdot \hat{Y}}\right) \quad \Leftrightarrow \quad 0 \in \overline{H \cdot Y}.$

[15] The reader may regard these characterizations as a definition of χ_θ-(semi)stable.

[16] King works with the Zariski topology, while we apply this result with respect to the Euclidean topology. Thus, we use Corollary 1.1.12 here.

(d) The stabilizer H_Y is finite if and only if $G_{\hat{Y}}/\Delta$ is finite.[17]

Proof. To prove (a), take $g \in G$ with $(X, z) = g \cdot \hat{Y}$. By Equation (2.26), we have $g \cdot Y = X$ and $\det(g_1)^{-m_2} \det(g_2)^{m_1} = z$. The latter shows that there exist some roots[18] $\det(g_1)^{-\frac{1}{m_1}}, \det(g_2)^{-\frac{1}{m_2}} \in \mathbb{C}^\times$ such that

$$\det(g_1)^{-\frac{1}{m_1}} \det(g_2)^{\frac{1}{m_2}} = z^{\frac{1}{m_1 m_2}}, \quad \text{i.e.,} \quad h := \left(\det(g_1)^{-\frac{1}{m_1}} g_1,\, \det(g_2)^{-\frac{1}{m_2}} g_2\right) \in H$$

satisfies $h \cdot Y = z^{\frac{1}{m_1 m_2}} X$. Conversely, given the latter for some $h = (h_1, h_2) \in H$, we define $g := \left(z^{-\frac{1}{m_1 m_2}} h_1,\, h_2\right)$ and compute $g \cdot \hat{Y} = (X, z)$ using (2.26).

Part (b) follows from applying part (a) to a sequence in the respective orbit that tends to a point in the orbit closure.

For part (c), note that if $Y = 0$ then $(0, 0) \in \overline{G \cdot \hat{Y}}$ and $0 \in \overline{H \cdot Y}$. It remains to consider $Y \neq 0$. Take $X \in V$ and let $g^{(k)} \in G$ be a sequence such that $g^{(k)} \cdot \hat{Y}$ tends to $(X, 0)$ as $k \to \infty$. Since $\chi_\theta^{-1}(g^{(k)}) \neq 0$ for all k, we apply (a) to obtain $Y_k := \left[\chi_\theta^{-1}(g^{(k)})\right]^{\frac{1}{m_1 m_2}} g^{(k)} \cdot Y \in H \cdot Y$ for all k. With $g^{(k)} \cdot \hat{Y} \to (X, 0)$ for $k \to \infty$ we conclude that the sequence Y_k tends to $0 \in V$. On the other hand, assume there exist $Y_k \in H \cdot Y$ with $Y_k \to 0$ as $k \to \infty$. Since $Y \neq 0$, we have $Y_k \neq 0$ and hence $c_k := \|Y_k\|^{\frac{m_1 m_2}{2}} \neq 0$ for all k. Thus, setting $X_k := c_k^{-\frac{1}{m_1 m_2}} Y_k$ and applying part (a) to $Y_k = c_k^{\frac{1}{m_1 m_2}} X_k$ gives $(X_k, c_k) \in G \cdot \hat{Y}$. The latter sequence tends to $(0, 0) \in V \times \mathbb{C}$, noting that $\|X_k\| = \|Y_k\|^{\frac{1}{2}}$ by the choice of c_k.

For part (d), first note that any $h = (h_1, h_2) \in H_Y$ stabilizes $\hat{Y}$ under the action (2.26), because h_1 and h_2 have determinant one. Therefore, we have a group morphism

$$\varphi \colon H_Y \to G_{\hat{Y}}/\Delta, \quad (h_1, h_2) \mapsto \overline{(h_1, h_2)}\,.$$

Its kernel is $H_Y \cap \Delta = \{(t\,\mathrm{I}_{m_1}, t\,\mathrm{I}_{m_2}) \mid t \in \mathbb{C}^\times,\, t^{m_1} = t^{m_2} = 1\}$, which is finite. Moreover, φ is surjective by the following. If $g = (g_1, g_2) \in G_{\hat{Y}}$ then $\chi_\theta^{-1}(g) \cdot 1 = 1$ translates to $\det(g_2)^{m_1} = \det(g_1)^{m_2} =: \lambda$. Take an $(m_1 m_2)$-root to obtain

$$t := \lambda^{-\frac{1}{m_1 m_2}} = \det(g_1)^{-\frac{1}{m_1}} = \det(g_2)^{-\frac{1}{m_2}}.$$

Then $h := (tg_1, tg_2) \in H$, but h also stabilizes Y as $g \in G_{\hat{Y}}$, so $h \in H_Y$. By construction, $\varphi(h) = \overline{g} \in G_{\hat{Y}}/\Delta$, hence φ is surjective.

Altogether, $H_Y/\ker(\varphi) \cong G_{\hat{Y}}/\Delta$. Since $\ker(\varphi)$ is finite, we deduce that H_Y is finite if and only if $G_{\hat{Y}}/\Delta$ is finite. □

With the help of Lemma 2.3.5 we finally prove Theorem 2.3.1.

[17] This part is not included in [AKRS21a, Lemma A.1], but appeared later in [AKRS21a, Appendix A]. We note that [AKRS21a] correctly states part (d), but the map $G_{\hat{Y}} \to H_Y$ given in [AKRS21a] is in general not a group morphism, and in general we do not have $G_{\hat{Y}}/\Delta \cong H_Y$. We adjusted the argument using a morphism $\varphi \colon H_Y \to G_{\hat{Y}}$ that shows $H_Y/\ker(\varphi) \cong G_{\hat{Y}}/\Delta$.

[18] Note that in general not all choices of roots will work, but there always exists a certain choice with the desired properties.

Proof of Theorem 2.3.1. By Proposition 2.3.4, the matrix tuple $Y = (Y_1, \dots, Y_n)$ is χ_θ-stable if and only if it is θ-stable. First, we show that the former is equivalent to being H-stable under the Kronecker quiver action. Then we rephrase θ-stability as the (shrunk subspace) conditions (i) and (ii).

Let $G_{\hat{Y}}$ denote the G-stabilizer of $\hat{Y} = (Y, 1)$ under the action (2.26). By Lemma 2.3.2, Y is χ_θ-stable if and only if the orbit $G \cdot \hat{Y}$ is closed and the group $G_{\hat{Y}}/\Delta$ is finite. The group $G_{\hat{Y}}/\Delta$ is finite if and only if H_Y is finite, by Lemma 2.3.5(d). For $Y = 0$, we have $H_Y = H$, which is not finite.

Thus, it remains to show for $Y \neq 0$ that $G \cdot \hat{Y}$ is closed if and only if $H \cdot Y$ is closed. If $G \cdot \hat{Y}$ is closed and $X \in \overline{H \cdot Y}$, then $(X, 1) \in \overline{G \cdot \hat{Y}} = G \cdot \hat{Y}$ using Lemma 2.3.5(b), and hence $X \in H \cdot Y$ by Lemma 2.3.5(a). Conversely, if $H \cdot Y$ is closed with $Y \neq 0$ then $0 \notin \overline{H \cdot Y}$. Thus, Lemma 2.3.5(c) yields $\overline{G \cdot \hat{Y}} \cap \big(V \times \{0\}\big) = \emptyset$. Hence, any $(X, z) \in \overline{G \cdot \hat{Y}}$ must satisfy $z \in \mathbb{C}^\times$, so $z^{\frac{1}{m_1 m_2}} \in \overline{H \cdot Y} = H \cdot Y$ by Lemma 2.3.5(b). We conclude $(X, z) \in G \cdot \hat{Y}$ using Lemma 2.3.5(a).

For Y being θ-stable, recall from Definition 2.3.3 that for all non-zero proper quiver-subrepresentations of $(\mathbb{C}^{m_1}, \mathbb{C}^{m_2}; Y)$ the inequality (2.27) has to be strict:

$$\langle \theta, (\dim V_1, \dim V_2) \rangle = m_2 \dim V_1 - m_1 \dim V_2 > 0.$$

Here, non-zero means $V_1 \neq 0$ or $V_2 \neq 0$, while proper means $V_1 \subsetneq \mathbb{C}^{m_1}$ or $V_2 \subsetneq \mathbb{C}^{m_2}$. Since $V_1 \neq 0$ and $V_2 = 0$ gives strict inequality in (2.27), it is enough to consider $V_2 \neq 0$. On the other hand, strict inequality in (2.27) holds for all proper subrepresentations satisfying $V_1 \subsetneq \mathbb{C}^{m_1}$ and $V_2 = \mathbb{C}^{m_2}$ if and only if there is *no* proper subrepresentation of this form, i.e., if and only if $\operatorname{rank}(Y_1, \dots, Y_n) = m_1$. Hence, by requiring the latter condition we can restrict to the case $V_2 \subsetneq \mathbb{C}^{m_2}$. Altogether, we rephrased the θ-stability of Y as (i) and (ii) in the statement. □

Similarly, we obtain a characterization for being semistable under the left-right action of H. The statement was proven differently in [BD06, Proposition 2.1], and we revisit it in Theorem 9.4.6.

Proposition 2.3.6. *Consider the left-right action of $H := \mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ on $V := (\mathbb{C}^{m_1 \times m_2})^n$. Then $Y = (Y_1, \dots, Y_n) \in V$ is semistable under H if and only if for all subspaces $V_1 \subseteq \mathbb{C}^{m_1}$, $V_2 \subseteq \mathbb{C}^{m_2}$ that satisfy $Y_i V_2 \subseteq V_1$ for all $i \in [n]$, one has $m_2 \dim V_1 \geq m_1 \dim V_2$.*

Proof. This is based on [AKRS21a, Remark A.2]. Remember Y is H-semistable under the left-right action if and only if it is H-semistable under the Kronecker quiver action. By Lemma 2.3.5(c), the latter is equivalent to

$$\big(V \times \{0\}\big) \cap \overline{G \cdot \hat{Y}} \neq \emptyset,$$

which in turn is equivalent to Y being χ_θ-semistable, see Lemma 2.3.2. By Proposition 2.3.4, χ_θ-semistability is equivalent to θ-semistability, and that translates via Definition 2.3.3 to the desired conditions. □

2.4 Popov's Criterion for solvable Groups

In this subsection we present Popov's Criterion for Zariski closed orbits under a connected solvable group. First, we briefly state the criterion in its general form. Afterwards, we specialize it to the very concrete setting in which we will apply it later. Since the criterion requires an algebraically closed field, we end with a lemma that allows to deduce polystability over $\mathbb{R}$, given the complex orbit is closed, and the rational representation is defined over $\mathbb{R}$.

Let G be a connected solvable group over $\mathbb{C}$. Then G is the semi-direct product of its unipotent radical U and a maximal torus T, see Theorem 1.1.25. By Proposition 1.1.20, $\mathfrak{X}(U) = 0$ and hence the character group $\mathfrak{X}(G)$ of G can be identified with $\mathfrak{X}(T)$ via restriction. We use this identification to view $\mathfrak{X}(T)$ as a subset of the coordinate ring $\mathbb{C}[G]$. Assume G acts algebraically on an affine variety Z. For $z \in Z$, consider the orbit map $\nu_{G\cdot z}\colon G \to Z,\ g \mapsto g \cdot z$ and its pullback map $\nu^*_{G\cdot z}\colon \mathbb{C}[Z] \to \mathbb{C}[G]$. Then $R_z := \nu^*_{G\cdot z}(\mathbb{C}[Z])$ is a subalgebra of $\mathbb{C}[G]$. Therefore, $\mathfrak{X}_{G\cdot z} := \{\chi \in \mathfrak{X}(T) \mid \chi \in R_z\}$ is a semigroup, where we identified $\mathfrak{X}(T) \subseteq \mathbb{C}[G]$.

Theorem 2.4.1 (Popov's Criterion, [Pop89, Theorem 4])**.**
Assume G and Z are as above, and let $z \in Z$. The orbit $G \cdot z$ is Zariski closed in Z if and only if the semigroup $\mathfrak{X}_{G\cdot z}$ is a group.

Remark 2.4.2. The Criterion contains the following special cases.

(i) If $G = U$ is unipotent, then $\mathfrak{X}(G)$ is trivial, compare Proposition 1.1.20. Hence, $\mathfrak{X}_{G\cdot z}$ is the trivial group for any $z \in Z$ and therefore all orbits $G \cdot z$ are Zariski closed, compare [Pop89, Corollary 3]. Thus, Popov's Criterion specializes for unipotent groups to the Kostant-Rosenlicht Theorem, see [Ros61, Theorem 2].

(ii) If $G = T$ is a torus, then Popov's Criterion specializes to [Pop89, Corollary 4]. If $Z = V$ is a rational representation of $G = T$, then this gives a reformulation of Theorem 2.1.9(c), the Hilbert Mumford Criterion via the Newton polytope. Further details can be found in [Pop89, p. 386]. ▽

Now, we specialize to the setting in which we will apply Popov's Criterion. The following is similar to [Pop89, Section 9]. Let $G \subseteq \mathrm{GL}_m(\mathbb{C})$ be a subgroup consisting of upper triangular matrices, which acts on $(\mathbb{C}^m)^n \cong \mathbb{C}^{m\times n}$ by left multiplication. Then G is a semi-direct product of T, the group of diagonal matrices in G, and U, the group of unipotent upper triangular matrices in G.

The following notation is unusual from the perspective of algebraic geometry.[19] We adjusted to the notation of Sections 9.5 and 10.7 for an easy comparison. Denote the coordinate functions on G by $x_{i,j} \in \mathbb{C}[G]$, $i, j \in [m]$, and those of $\mathbb{C}^{m\times n}$ by $f_{i,l} \in \mathbb{C}[\mathbb{C}^{m\times n}]$, $i \in [m]$, $l \in [n]$. For matrix $Y \in \mathbb{C}^{m\times n}$, the pullback of the orbit map $\nu_{G\cdot Y}$ is given by

$$\nu^*_{G\cdot Y}(f_{i,l}) = \sum_{j=1}^{m} x_{i,j} Y_{j,l}$$

[19] Usually, small letters denote constants and capital letters denote coordinate functions. Here, it is the other way around.

and therefore

$$R_Y = \nu_{G\cdot Y}^*\big(\mathbb{C}[\mathbb{C}^{m\times n}]\big) = \mathbb{C}\Big[\sum_{j=1}^{m} Y_{j,l}x_{i,j} \mid i \in [m], l \in [n]\Big] \subseteq \mathbb{C}[G]. \tag{2.28}$$

Since $T \subseteq \mathrm{GT}_m(\mathbb{C})$, we have a surjection $\varphi\colon \mathfrak{X}(\mathrm{GT}_m(\mathbb{C})) \cong \mathbb{Z}^m \twoheadrightarrow \mathfrak{X}(T)$ of abelian groups, see Proposition 1.1.17. Therefore, $\mathfrak{X}(T) \cong \mathbb{Z}^m/\ker(\varphi)$ and we may write

$$\mathfrak{X}_{G\cdot Y} = \big\{(d_1,\dots,d_m) \in \mathfrak{X}(T) \mid x_{11}^{d_1}\cdots x_{mm}^{d_m} \in R_Y\big\}$$

using this identification.

We finish the section with an argument how to deduce polystability over the reals if the representation is defined over $\mathbb{R}$. As mentioned in the proofs of [Bir71, Corollary 5.3] and [DM21, Proposition 2.21] the next statement follows from [BH62, Proposition 2.3]. We stress that the group does *not* have to be reductive.

Lemma 2.4.3. *Let G be a connected complex algebraic group and $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation, both defined over $\mathbb{R}$. Let $v \in V_{\mathbb{R}}$ and suppose that $G\cdot v$ is Euclidean closed in V. Then $G_{\mathbb{R}}\cdot v$ is Euclidean closed in $V_{\mathbb{R}}$.*

Proof. Since G is a connected complex algebraic group we can apply [BH62, Proposition 2.3] (Proposition 1.1.14). Hence, we have that $(G\cdot v)\cap V_{\mathbb{R}}$ is[20] a finite union of Euclidean closed $(G_{\mathbb{R}})^\circ$-orbits, where $(G_{\mathbb{R}})^\circ$ denotes the Euclidean identity component. One of these closed orbits must be $(G_{\mathbb{R}})^\circ\cdot v$. As $G_{\mathbb{R}}$ is a real algebraic variety it has finitely many Euclidean-connected components by Theorem 1.1.6. Choose representatives $g_1,\dots,g_k$ of $G_{\mathbb{R}}/(G_{\mathbb{R}})^\circ$. Since $G_{\mathbb{R}}$ is a Lie group, the multiplication with g_i is a homeomorphism and we conclude that

$$G_{\mathbb{R}}\cdot v = \bigcup_{i=1}^{k} g_i\left((G_{\mathbb{R}})^\circ\cdot v\right)$$

is Euclidean closed as a finite union of Euclidean closed sets. □

[20] In general, $(G\cdot v)\cap V_{\mathbb{R}}$ and $G_{\mathbb{R}}\cdot v$ do not have to be equal (see Remark 1.1.13), but the latter is contained in the former

Part II

Computational Complexity

Chapter 3

Computational Invariant Theory

*"Invariant theory has already been pronounced dead several times,
and like the phoenix it has been again and again rising from its ashes."*

Dieudonné and Carrell in [DC70, page 1]

This chapter serves as an introduction to computational invariant theory, and its manifold algorithmic methods and applications. Thereby, we embed and locate the contributions of this thesis in the research area. We stress that an exhaustive discussion of computational invariant theory is not provided and certainly would go beyond this thesis. Instead, we focus and illustrate those aspects especially needed in later chapters. In particular, we provide the necessary background and motivation for Chapters 4 and 5, which present hardness results for geodesic convex methods in invariant theory. Moreover, the presented computational problems and scaling algorithms connect to the algorithmic aspects of maximum likelihood estimation, see Part III on algebraic statistics.

Organization and Assumptions. In Section 3.1 we outline historical developments, state the computational problems studied in this thesis and some of their applications. Afterwards, we discuss scaling algorithms and comment on their complexity to solve these problems in Section 3.2.

We note that the whole chapter uses the assumptions stated below in Setting 3.0.1, which is Setting 2.2.2 over $\mathbb{C}$ and we additionally fix a maximal torus.

Setting 3.0.1 (Assumptions for Part II)**.** We work over $\mathbb{C}$. Let $G \subseteq \mathrm{GL}_N(\mathbb{C})$ be a Zariski closed and self-adjoint subgroup,[1] set $K := G \cap \mathrm{U}_N$ and $\mathfrak{p} := \mathfrak{i}\,\mathrm{Lie}(K) = \mathrm{Lie} \cap \mathrm{Sym}_N(\mathbb{C})$. Moreover, fix a maximal torus $T := (G \cap \mathrm{GT}_N(\mathbb{C}))^\circ$ in G and a maximal compact torus $T_K := T \cap K$ of K, compare Proposition 1.2.14. Consider a rational representation $\pi\colon G \to \mathrm{GL}(V)$ and its differential $\Pi\colon \mathrm{Lie}(G) \to \mathrm{End}(V)$. Equip V with a K-invariant inner product. Finally, let

$$\mu_G\colon V\backslash\{0\} \to \mathfrak{i}\,\mathrm{Lie}(K) \qquad \text{and} \qquad \mu_T\colon V\backslash\{0\} \to \mathfrak{i}\,\mathrm{Lie}(T_K)$$

denote the moment maps for the G-action, respectively T-action, with respect to this inner product. For a concrete instance see Example 2.2.3. ▲

[1] Remember from Theorem 1.3.10 that such groups are reductive, and conversely any reductive group is isomorphic to such a group.

3.1 Computational Problems and Applications

Based on [DC70; Stu08; DK15], we first give a historical overview on some aspects of computational invariant theory. Thereby, we introduce the main computational problems of interest for this thesis. Afterwards we present several applications and cite related literature, that may be consulted for further details. We end with an extended example on matrix scaling, and short comments on its generalizations, to illustrate how the computational problems translate in these cases.

History of Computational Problems

Since its origins in the 19^{th} century invariant theory is inseparably linked to computation. In fact, classical invariant theory from that time was mainly motivated by the following fundamental problems, compare [KP96, Section 1.5] and [Stu08, Section 1.3]. Given a representation $\pi\colon G \to \mathrm{GL}(V)$:

1. Find a finite set $f_1, \dots, f_k$ of generators of the ring of invariants $\mathbb{C}[V]^G$.
2. Determine the algebraic relations, i.e., the syzygies, among $f_1, \dots, f_k$.[2]

Solutions to these problems for concrete actions are usually called the First and Second Fundamental Theorem respectively [KP96]. Many famous mathematicians such as Cayley, Clebsch, Cremona, Gordan and Sylvester contributed to invariant theory in its classical period. The latter culminated in Hilbert's breakthroughs [Hil90; Hil93], in which he proved that $\mathbb{C}[V]^G$ is finitely generated[3] (Theorem 1.4.6) and provided a finite algorithm that computes a system of generators. It is noteworthy, that [Hil90; Hil93] made further outstanding contributions to modern algebra: they contain Hilbert's Nullstellensatz, Hilbert's basis theorem and Hilbert's syzygy theorem. However, the computational methods available were, especially with the lack of modern computers, extremely cumbersome and, if at all, only possible to carry out by hand in tiniest examples.

With some of its main problems being solved and the given computational cost of available algorithms, research in invariant theory (almost) fell asleep for decades. A first revival was initiated by the developments on representations of semisimple groups which realized classical invariant theory as a special case, compare [Wey39]. Latest with Mumford's invention of *Geometric Invariant Theory* (GIT) in 1965 invariant theory was again at the forefront of mathematics [MFK94]. Mumford realized that ideas from Hilbert's paper [Hil93] combined with modern scheme theory enabled him to construct moduli spaces via so-called GIT quotients. Again, this relates to an interesting computational question. Namely, whether two vectors $v, w \in V$ are identified in the affine GIT quotient gives the following decision problem.

Computational Problem 3.1.1 (Orbit Closure Intersection (OCI))**.**
Given $\pi\colon G \to \mathrm{GL}(V)$ *and* $v, w \in V$*, decide whether* $\overline{G \cdot v}^{\mathrm{Z}} \cap \overline{G \cdot w}^{\mathrm{Z}} \neq \emptyset$.

[2] In [Stu08] the following interesting problem is added: give an algorithm that writes any invariant $f \in \mathbb{C}[V]^G$ as a polynomial in the generators $f_1, \dots, f_k$.

[3] where V is a finite dimensional representation of a reductive group G

We note that [Mul17] conjectures that OCI is computable in polynomial time for any rational representation of a reductive group G. An important special case of OCI arises when $w = 0$. This translates to deciding whether v is unstable.

Computational Problem 3.1.2 (Null Cone Membership (NCM)). *Given $\pi\colon G \to \mathrm{GL}(V)$ and $v \in V$, decide whether $0 \in \overline{G \cdot v}^{\mathrm{Z}} = \overline{G \cdot v}$.*

In parallel to Mumford's work, Buchberger's algorithm[4] [Buc70; Buc06] to compute Gröbner bases gave birth to computational commutative algebra as a research field. Soon, Gröbner basis methods fostered many new results in computational invariant theory; the reader is referred to the excellent text books [Stu08; DK15] and the references therein. We remark that Sturmfels' book [Stu08], which marries the ideas of classical invariant theory with Gröbner basis methods, may serve as an introduction to the topic. It is complemented by the monograph [DK15], which treats many modern concepts such as Derksen's algorithm, separating invariants and degree bounds for generating invariants.

We point out that modern methods solve the OCI problem for general reductive groups as follows. One computes a system $f_1, \dots, f_k$ of generators for $\mathbb{C}[V]^G$ using Derksen's algorithm [Der99] and evaluates them at v and w. This decides OCI, as invariants separate orbit closures by Theorem 1.4.7. However, this approach is in general not computationally efficient or often even infeasible. First, Derksen's algorithm crucially involves the computation of a Gröbner basis, which is usually very costly and the basis may be huge. Second, generating invariants can have exponential degree [DM20b], and third, it may be difficult to evaluate them (exactly). Furthermore, an approach via so-called succinct encodings of generating invariants [Mul17] was disproven recently in [GIM+20]. Hence, for general reductive groups it remains open whether the OCI Problem 3.1.1 can be decided in polynomial time.

Complementing the symbolic/algebraic methods, recent years have seen intense study on optimization approaches to computational invariant theory. This already enjoyed several success stories, compare Section 3.2. In the following we present two optimization problems which can be used to decide NCM. For this, recall that $0 \in \overline{G \cdot v}$ if and only if $\mathrm{cap}_G(v) = \inf_{g \in G} \|g \cdot v\|^2 = 0$. Therefore, the NCM problem is naturally linked to approximating the capacity of v.

Computational Problem 3.1.3 (Norm Minimization). *Given $\pi\colon G \to \mathrm{GL}(V)$, $v \in V$ and a precision $\varepsilon > 0$, determine $g \in G$ such that $\|g \cdot v\|^2 \leq \mathrm{cap}_G(v) + \varepsilon$.*

On the other hand, recall that Kempf-Ness gives the duality (Equation (2.21))

$$\mathrm{cap}_G(v) = 0 \quad \Leftrightarrow \quad \inf_{g \in G} \|\mu(g \cdot v)\|^2 > 0.$$

Therefore, norm minimization and deciding (non)-membership in the null cone are related to scaling the moment map to zero.[5]

[4] The algorithm was first published in Buchberger's PhD thesis from 1965. We provide references to the journal version from 1970 and a translation of Buchberger's thesis from 2006.

[5] In fact, a result of [BFG+19] made this quantitive, compare Theorem 3.2.5.

Computational Problem 3.1.4 (Scaling)**.** *Given $\pi\colon G \to \mathrm{GL}(V)$, $v \in V$ with $0 \in \Delta_G(v)$ and a precision $\varepsilon > 0$, determine $g \in G$ such that $\|\mu_G(g \cdot v)\| \leq \varepsilon$.*

The following is of great importance for optimization approaches to computational invariant theory. The Kempf-Ness function F_v, see (2.8), is geodesically convex on the manifold $P = \{g^\dagger g \mid g \in G\}$ of positive definite matrices in G [BFG+19, Proposition 3.13]; also compare Theorem 1.2.18 and Example 1.2.21. Therefore, the Norm Minimization Problem 3.1.3 and the Scaling Problem 3.1.4 fall into the framework of geodesic convex optimization problems! If G is a torus, they are even convex in the usual sense, compare Equation (3.1) below.

Remark 3.1.5. We note that NCM, Norm minimization and Scaling in the above formulations may also be considered over $\mathbb{R}$.[6] In fact, we link these problems over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ to maximum likelihood estimation in the part on algebraic statistics, see e.g., Chapters 7 and 9. Moreover, NCM and Norm minimization still make sense for non-reductive groups[7] and even beyond the group setting; again there are connections to statistics, see Section 9.5 and Chapters 8, 10 respectively. ▽

Finally, we note the following. Another equivalent formulation of the NCM Problem 3.1.2 is to decide whether $0 \notin \Delta_G(v)$. Therefore, NCM is also a special case of the moment polytope membership problem. It asks whether a given rational vector $p \in \mathbb{Q}^N$ is contained in the moment polytope $\Delta_G(v)$ [BFG+19, Problem 1.11]. This problem as well admits a scaling analogue [BFG+19, Problems 1.12], and there are many applications, e.g., to Kronecker polytopes and to Horn's problem. We refer to [BFG+18; BFG+19] for further details.

Applications

We give a brief overview on some applications of the mentioned Computational Problems 3.1.1 – 3.1.4. The interested reader is encouraged to consult for further details the introductions in [BGO+18; BFG+18; BFG+19], [GO18, Section 5] and the references in these papers.

Algebraic Geometry. As already mentioned, the OCI problem plays an important role in the construction of moduli spaces via GIT quotients [MFK94; New78; Hos15]. The NCM problem is of particular interest, since the null cone has to be excluded in the construction of projective GIT quotients.

Convex Optimization. If $G = T$ is a torus, i.e., in the (connected) commutative case, the Norm Minimization Problem 3.1.3 is convex. Indeed, consider the action of $T = \mathrm{GT}_d(\mathbb{C})$ on $\mathbb{C}^m$ via the matrix $A \in \mathbb{Z}^{d\times m}$, see Example 1.3.16.[8] Denote the j^{th} column of A by A_j. Then we have for $v \in \mathbb{C}^m$ that

$$\mathrm{cap}_T(v) = \inf_{t_1,\ldots,t_d \in \mathbb{C}^\times} \sum_{j=1}^{m} |v_j|^2 \prod_{i=1}^{d} |t_i|^{2A_{i,j}} = \inf_{x\in\mathbb{R}^d} \sum_{j=1}^{m} |v_j|^2 \exp\left(\langle x, A_j\rangle\right), \qquad (3.1)$$

[6] Actually, the first two problems admit nice relations between the solutions over $\mathbb{R}$ and those over $\mathbb{C}$, compare Proposition 2.2.18.

[7] However, one needs to be careful. First, if the group is non-reductive then norm minimization and scaling are in general *not* geodesically convex problems. Second, for non-reductive groups the topological null cone and the null cone cut out by invariants do not have to be equal, compare Example 1.4.9.

[8] Recall that, up to identification, any torus action is of this form.

where we used the change of variables $x_i = 2\log(|t_i|)$. Equation (3.1) is a log-convex optimization problem in x known as *unconstrained geometric programming*. This huge class of convex optimization problems itself has manifold applications [DPZ67; Pet76; Eck80; BKVH07]. For example, it covers matrix scaling, matrix balancing and array scaling, which arise in scientific computing and optimal transport [Cut13; PR71]. It also contains commutative polynomial scaling, which recovers Gurvit's polynomial capacity [Gur04b; Gur06].

Physics. Especially the tensor scaling setting has important connections to quantum information theory, see e.g., [Kly06; SOK14; Wal14; BFG+18], and to quantum many-body physics [AMN+22].

Analysis. The Brascamp Lieb inequalities [BL76; Lie90] are a huge family of inequalities which generalize many important inequalities such as Cauchy Schwarz, Hölder and Brunn-Minkowski. Brascamp Lieb inequalities involve an optimal constant known as the BL constant, which is related to invariant theory through certain semi-invariants of the star quiver [GGOW18, Section 4.1]. In this case, the NCM Problem 3.1.2 translates to deciding whether the BL constant is infinite, while the Scaling Problem 3.1.4 means to approximate the BL constant (given it is finite). Via a reduction to operator scaling polynomial time algorithms for both instances are given in [GGOW18].

Computer Science & Complexity Theory. First, we note that geometric complexity theory, an approach to complexity lower bounds, suggests that the OCI Problem 3.1.1 should be in the complexity class P [Mul17]. In fact, [Mul17] gives an algebraic polynomial time algorithm for OCI if the group is *fixed*. Non-rational identity testing, which is a non-commutative analogue of the famous polynomial identity testing (PIT), arises as the NCM problem for operator scaling.[9] This led to several deterministic polynomial time algorithms [GGOW16; DM17; IQS18; AGL+18] for non-rational identity testing.

Statistics. Of course, one important link of the computational problems to statistics is through matrix scaling. We discuss this relation in detail below. In the commutative case the Lagrange dual of the Scaling Problem 3.1.4 covers discrete entropy maximization [SV14; SV19].[10] Moreover, the commutative case connects to maximum likelihood (ML) estimation of log-linear models and iterative proportional scaling[11] [AKRS21b]. The results of [AKRS21b] are presented in Chapter 7. The non-commutative setting is tightly related to ML estimation of so-called Gaussian group models [AKRS21a]. These relations go even beyond the usual setting of reductive groups and are discussed in Chapter 9.[12] Furthermore, connections to operator scaling have been used to obtain results on the sample complexity for Tyler's M estimator [FM20].

[9] The PIT problem is *not* an instance of the NCM problem [MW21].

[10] It is an interesting open problem whether similar connections between the continuous entropy maximization problem (see e.g., [LV20]) and the non-commutative setting hold; private communication with Jonathan Leake.

[11] also known as iterative proportional fitting

[12] Further work was stimulated by these connections [MRS21], which even goes beyond the case of groups. This is discussed in detail in Chapters 8 and 10.

Extended Example: Matrix Scaling

In the following we illustrate how matrix scaling naturally arises when considering the Computational Problems 3.1.2–3.1.4 for the restriction of $\pi_{m,2}$ to $T := \mathrm{ST}_m(\mathbb{C})^2$. Matrix scaling has manifold relations and applications such as optimal transport, bipartite matching and statistics. We refer to the detailed survey [Ide16]. Let us first define what we mean by matrix scaling in the following.[13]

Definition 3.1.6. Let $M \in \mathbb{R}^{m\times m}$ be a matrix with non-negative entries.

1. M is *doubly stochastic* if all row sums $M_{i,+}$ and all column sums $M_{+,j}$ of M are one. The distance of M to doubly stochastic is
$$\mathrm{ds}(M) := \sum_{i=1}^{m}(M_{i,+}-1)^2 + \sum_{j=1}^{m}(M_{+,j}-1)^2. \tag{3.2}$$
2. XMY is called a *scaling* of M if $X, Y \in \mathbb{R}^{m\times m}$ are positive definite diagonal matrices.
3. M is *scalable* (to doubly stochastic), if there is a scaling XMY that is doubly stochastic.
4. M is *approximately scalable* (to doubly stochastic), if for every $\varepsilon > 0$ there exists a scaling XMY such that $\mathrm{ds}(XMY) < \varepsilon$. ▲

Note that we can parametrize X (and similarly Y) as $X = \exp(\mathrm{diag}(x))$, where $x \in \mathbb{R}^m$. Now, matrix scaling arises via $\pi_{m,2}$ restricted to the torus $T = \mathrm{ST}_m(\mathbb{C})^2$.[14] Indeed, remember from Example 1.3.18 that $\pi_{m,2}$ has set of weights
$$\Omega(\pi_{m,2}) = \big\{(\epsilon_i, \epsilon_j) \mid i,j \in [m]\big\} \subseteq (\mathbb{R}^m)^2.$$
Now, for $v \in \mathbb{C}^{m\times m}$ the geometric program
$$\mathrm{cap}_T(v) = \inf_{g,h\in \mathrm{ST}_m(\mathbb{C})} \sum_{i,j=1}^{m} |g_{ii}|^2|v_{ij}|^2|h_{jj}|^2 = \inf_{x,y\in\mathbb{R}^m} \sum_{i,j=1}^{m} |v_{ij}|^2 e^{\langle(\epsilon_i,\epsilon_j),(x,y)\rangle} \tag{3.3}$$
captures matrix scaling for $M_v := (|v_{ij}|^2)_{i,j}$, compare [RS89, Programs I and II]. Perhaps, the connection becomes even more apparent when considering the moment map for this action. Recall from Equation (2.11) in Example 2.2.8 that
$$\mu_T(v) = \frac{1}{\|v\|^2}\left(r(M_v) - \frac{\|v\|^2}{m}\mathbb{1}_m,\, c(M_v) - \frac{\|v\|^2}{m}\mathbb{1}_m\right), \tag{3.4}$$
where $r(M_v), c(M_v) \in \mathbb{R}^m$ are the vectors of row respectively column sums of M_v. Consequently, $\mu_T(v) = 0$ if and only if the matrix $m\|v\|^{-2}M_v$ is doubly stochastic. This allows to link matrix scaling to conditions from Kempf-Ness, Theorem 2.2.13.

[13] Instead of scaling to a doubly stochastic matrix one could, more generally, consider scaling for given vectors r and c of row and column sums.

[14] On first glance, one might wonder why the left-right action of $\mathrm{GT}_m(\mathbb{C})^2$ on $\mathbb{C}^{m\times m}$ is not used. However, this action is not meaningful for NCM as all matrices are unstable.

Proposition 3.1.7. *Let $v \in \mathbb{C}^{m\times m}$ and set $M_v := \big(|v_{ij}|^2\big)_{i,j} \in \mathbb{R}^{m\times m}_{\geq 0}$. Then*

(i) M_v is scalable $\quad\Leftrightarrow\quad \exists\, t \in T\colon\ \|\mu_T(t\cdot v)\| = 0.$

(ii) M_v is approximately scalable $\quad\Leftrightarrow\quad \inf_{t\in T}\|\mu_T(t\cdot v)\| = 0.$

Proof. We prove the first part. Item (ii) follows similarly using continuity of the moment map. First, assume there is some $t \in T = \mathrm{ST}_m(\mathbb{C})^2$ with $\mu_T(v) = 0$. Writing $t = (g, h)$ one computes that

$$\big(M_{t\cdot v}\big)_{ij} = |(t\cdot v)_{ij}|^2 = |g_{ii}|^2|v_{ij}|^2|h_{jj}|^2. \tag{3.5}$$

Therefore, $M_{t\cdot v}$ is a scaling of M_v and so is $m\|t\cdot v\|^{-2}M_{t\cdot v}$. The latter is doubly stochastic as $\mu_T(t\cdot v) = 0$ and we conclude that M is scalable.

Conversely, let XM_vY be a scaling of M_v that is doubly stochastic. Since X, Y are diagonal positive definite matrices we can write $X = \exp(2\,\mathrm{diag}(x))$ and $Y = \exp(2\,\mathrm{diag}(y))$, where $x, y \in \mathbb{R}^m$. We define the determinant one matrices

$$g := \exp\big(-m^{-1}x_+\big)\exp\big(\mathrm{diag}(x)\big) \quad \text{and} \quad h := \exp\big(-m^{-1}y_+\big)\exp\big(\mathrm{diag}(y)\big)$$

to obtain $t := (g, h) \in T$. Via Equation (3.5) we compute $M_{t\cdot v} = \lambda XM_vY$, where $\lambda := \exp\big(-2m^{-1}(x_+ + y_+)\big)$. As XM_vY has row sums equal to one, we get

$$\|t\cdot v\|^2 = \sum_{i\in[m]}\big(M_{t\cdot v}\big)_{i,+} = \lambda\sum_{i\in[m]}\big(XM_vY\big)_{i,+} = \lambda m.$$

Thus, $m\|t\cdot v\|^{-2}M_{t\cdot v} = \lambda^{-1}M_{t\cdot v} = XM_vY$ which is doubly stochastic and hence $\mu_T(t\cdot v) = 0$ as desired. □

As a direct consequence of Kempf-Ness, Theorem 2.2.13 parts (e) and (f), and Hilbert-Mumford Theorem 2.1.9 via weight polytopes,[15] we obtain the following.

Corollary 3.1.8. *Let $v \in \mathbb{C}^{m\times m}$ and set $M_v := \big(|v_{ij}|^2\big)_{i,j} \in \mathbb{R}^{m\times m}_{\geq 0}$. Then*

(i) M_v *is scalable* $\Leftrightarrow$ v *is T-polystable* $\Leftrightarrow$ $0 \in \mathrm{relint}(\Delta_T(v))$

(ii) M_v *is approx. scalable* $\Leftrightarrow$ v *is T-semistable* $\Leftrightarrow$ $0 \in \Delta_T(v)$

Therefore, the NCM Problem 3.1.2 for matrix scaling is deciding whether M_v is not approximately scalable. The Scaling Problem 3.1.4 essentially[16] translates to compute a scaling of M_v that is close to a doubly stochastic matrix.

Moreover, we can relate the Hilbert-Mumford characterization to bipartite matching, also compare [GO18, Corollary 3.5].

Proposition 3.1.9. *For $v \in \mathbb{C}^{m\times m}$, $0 \in \Delta_T(v)$ if and only if the bipartite graph given by the zero pattern of v (equivalently, of M_v) admits a perfect matching.*

[15] also note Remark 2.1.10

[16] up to a rescaling as in the proof of Proposition 3.1.7

Proof. First, the weight polytope (2.2) of v under matrix scaling is given by

$$\Delta_T(v) = \operatorname{conv}\big\{(\epsilon_i, \epsilon_j) \mid v_{ij} \neq 0\big\} \subseteq \mathbb{R}^{2m}.$$

Moreover, v induces the bipartite graph $\mathcal{G}_v = (I = [m], J = [m], E)$ with edges $E = \{(i,j) \in I \times J \mid v_{ij} \neq 0\}$. Now, assume $\mathcal{G}_v$ has a perfect matching, i.e., there is a permutation $\sigma \in \mathfrak{S}_m$ such that $(i, \sigma(i)) \in E$. Using $\sum_i \epsilon_i = 0_m$, we deduce

$$(0_m, 0_m) = \sum_{i \in [m]} \frac{1}{m} (\epsilon_i, \epsilon_{\sigma(i)}) \in \Delta_T(v).$$

Conversely, assume $\mathcal{G}_v$ does not admit a perfect matching. By Hall's marriage theorem [Hal35], there is a set $W \subseteq I$ such that its neighbour set

$$N(W) := \big\{j \in J \mid \exists\, i \in I \colon\ (i,j) \in E\big\} \quad \text{obeys} \quad k := |W| > |N(W)| =: l.$$

Without loss of generality, let $W = [k]$ and $N(W) = [l]$, i.e., v is of the form

$$\begin{pmatrix} A & 0_{k,m-l} \\ B & C \end{pmatrix}, \quad \text{where } A \in \mathbb{C}^{k \times l} \text{ and } C \in \mathbb{C}^{(m-k)\times(m-l)}.$$

Consider $a, b \in \mathbb{Z}^m$ defined by $a_i = -(m-k)$ for $i \in [k]$ and $a_i = k$ for $i > k$; respectively by $b_j = (m-l)$ for $j \in [l]$ and $b_j = -l$ for $j > l$. By construction, $a_+ = \sum_i a_i = \langle a, \mathbb{1}_m \rangle = 0$ and $b_+ = 0$. Therefore, we compute that

$$\langle (a,b), (\epsilon_i, \epsilon_j) \rangle = \langle a, \epsilon_i \rangle + \langle b, \epsilon_j \rangle = \langle a, e_i \rangle + \langle b, e_j \rangle = a_i + b_j\,.$$

Furthermore, we have that $a_i + b_j > 0$ whenever $v_{ij} \neq 0$, since $k - l > 0$ and $k + m - l > 0$. Altogether, (a,b) defines a hyperplane in $\mathbb{R}^{2m}$ which separates 0 from $\Delta_T(v)$. Hence, $0 \notin \Delta_T(v)$ which ends the proof. Let us point out that (a,b) defines a character of $T = \mathrm{ST}_m(\mathbb{C})^2$ that sends v in the limit to zero.[17] □

Combining the characterizations of semistability via Hilbert-Mumford and Kempf-Ness we recover the known link between matrix scaling and bipartite matching, see e.g., [RS89].

Theorem 3.1.10. *A non-negative $M \in \mathbb{R}^{m \times m}$ is approximately scalable if and only if the bipartite graph given by M admits a perfect matching.*

Array, Operator and Tensor Scaling

We briefly outline that the above results on matrix scaling generalize to array, operator and tensor scaling from Example 1.3.5.

Three-dimensional array scaling (i.e., the action of $\mathrm{ST}_m(\mathbb{C})^3$ via $\pi_{m,3}$) translates to scaling the non-negative tensor $p = (|v_{ijk}|^2) \in (\mathbb{R}_{\geq 0})^{\otimes 3}$ to tristochastic. The latter means that all slice sums are one, i.e., $p_{i,+,+} = p_{+,j,+} = p_{+,+,k} = 1$ for all $i, j, k \in [m]$. This generalizes to d-dimensional array scaling. However, we

[17] This construction is also used to characterize instability for operator scaling, compare [BD06, Proof of Theorem 2.1, part one].

note that array scaling does *not* relate to d-partite hypergraph matching. Indeed, the latter is NP-hard, while NCM for array scaling is solvable in polynomial time.

Operator Scaling (i.e., $\pi_{m,2}$) relates to scaling a completely positive map to "doubly stochastic", meaning the two quantum marginals are the identity matrix, [Gur04a], [GGOW16], [GO18, Section 2.2]. It has many applications such as non-rational identity testing [GGOW16] and ML estimation for matrix normal models, Section 9.4. Deciding NCM admits a representation-theoretic translation via so-called shrunk subspaces[18], [Kin94] (see Section 2.3) and [BD06] (Theorem 9.4.6).

Similar to operator scaling, tensor scaling (i.e., $\pi_{m,d}$ for $d \geq 3$) translates to scaling all quantum marginals to the identity, see [BGO+18] and [GO18, Section 2.3]. It has manifold applications such as geometric complexity theory, quantum information theory and ML estimation of tensor normal models, Chapter 9.

3.2 Scaling Algorithms

We discuss several scaling algorithms and their complexity for solving (some of) the Computational Problems 3.1.2-3.1.4 for specific group actions. More precisely, we discuss Sinkhorn scaling and operator scaling as well as convex optimization for the commutative and geodesic convex methods for the non-commutative case. Furthermore, we comment on related algebraic methods.

We highlight that this subsection prepares and connects to other chapters as follows. Sinkhorn scaling and convex optimization methods are related to the study of log-linear models, compare Section 7.3. Similarly, we revisit operator scaling and geodesic convex optimization for ML estimation in Gaussian group models, especially in Section 9.3 and Subsection 9.4.4. Moreover, the detailed discussion of geodesic convex methods and results of [BFG+19] motivates Chapters 4 and 5, which present barriers for geodesic convex methods.

Sinkhorn Scaling

For a non-negative matrix M consider matrix scaling in the approximate sense, i.e., computing a scaling XMY with $\mathrm{ds}(XMY) \leq \varepsilon$, compare Definition 3.1.6. The matrices X and Y, if they exist, can be found by a simple and fast alternating minimization approach. This method was introduced in [Sin64] and is known as *Sinkhorn's algorithm*, see Algorithm 3.1. We note that it admits a natural generalization [SK67] to scale row and column sums to arbitrary marginal vectors.

The work [LSW00] gave the following complexity analysis of Sinkhorn's algorithm, also compare [GO18, Theorem 2.6].

Theorem 3.2.1. *Let $M \in \mathbb{Q}^{m \times m}$ be a non-negative matrix with entries of bit complexity at most b, and let $T = O(m(b + \log(m))\varepsilon^{-1})$. Then Algorithm 3.1 on input (M, T, ε) works correctly.*

[18] We note that the recent preprint [FSG22] gives an alternating minimization procedure to find a shrunk subspace, if existent, in deterministic polynomial time.

Algorithm 3.1: Sinkhorn Scaling

Input : Non-negative matrix $M \in \mathbb{R}^{m\times m}$, a number of iterations N, a precision $\varepsilon > 0$

Output: Either returns "M is not scalable"; or outputs X and Y such that the scaling XMY satisfies $\mathrm{ds}(XMY) \leq \varepsilon$

if *M has a zero row or a zero column* **then**
  return *M is not scalable.*
end
Initialize $X = Y = \mathrm{I}_m$;
for $k = 1$ **to** N **do**
  if $\mathrm{ds}(XMY) \leq \varepsilon$ **then**
    return *X and Y*
  else
    $X \leftarrow \mathrm{diag}\left((XMY)_{1,+}, \ldots, (XMY)_{m,+}\right)^{-1} X$; `/* scale rows */`
    $Y \leftarrow \mathrm{diag}\left((XMY)_{+,1}, \ldots, (XMY)_{+,m}\right)^{-1} Y$; `/* scale col's */`
  end
end
return *M is not scalable.*

We remark that Sinkhorn's algorithm is frequently used in practice, e.g., for quickly approximating the solution to optimal transport problems [Cut13]. Recently, [AGL+21] provided a quantum implementation of Sinkhorn's algorithm.

As discussed in Section 3.1, matrix scaling is captured by the action of $T := \mathrm{ST}_m(\mathbb{C})^2$ via $\pi_{m,2}$. Similarly to Algorithm 3.2 below, the connection via Proposition 3.1.7 allows for a normalized[19] version of Algorithm 3.1 over $\mathbb{C}$, which solves the Scaling Problem 3.1.4 for $\pi_{m,2}|_T$.

Finally, we note that Sinkhorn scaling also generalizes to d-dimensional array scaling. There is a simple and fast alternating minimization algorithm that produces ε-tristochastic scalings in time $O(1/\varepsilon^2)$ [AB22; LHCJ22].

Operator Scaling

The left-right action of $\mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ on $(\mathbb{C}^{m_1\times m_2})^n$ captures operator scaling[20] from [Gur04a]. Algebraic and optimization-based algorithms have, independently and nearly concurrently, resulted in polynomial time algorithms for NCM [GGOW16; IQS18] and even for OCI [AGL+18; DM20a]. The optimization approaches in [GGOW16; AGL+18] also yield polynomial time algorithms for Norm minimization 3.1.3 and Scaling Problem 3.1.4. However, they do not work over fields in arbitrary characteristic like the algebraic methods in [IQS18; DM20a]. We stress that so far neither the algebraic nor the optimization approach solve NCM for 3-tensor scaling in polynomial time.

In [Gur04a] Gurvits' suggested, similar to Sinkhorn's algorithm, an alternating minimization method for operator scaling, also compare [GO18, Section 2.2]. In

[19] to ensure the determinant one condition

[20] Remember that $\pi_{m,2}^{\oplus n}$ is operator scaling for the equidimensional case $m = m_1 = m_2$.

Algorithm 3.2 we present a *normalized* version of Gurvits' algorithm to solve the Scaling Problem 3.1.4 for the left-right action. We compare this algorithm in Subsection 9.4.4 to the flip-flop algorithm from statistics.

Algorithm 3.2: Alternating Minimization for Operator Scaling

Input : $Y \in (\mathbb{C}^{m_1 \times m_2})^n$, a number of iterations N, a precision $\varepsilon > 0$
Output: Either returns "Y is unstable", or outputs
$g \in \mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ with $\|\mu(g \cdot Y)\|_F \leq \varepsilon$

if $\sum_{i=1}^n Y_i Y_i^\dagger$ *or* $\sum_{i=1}^n Y_i^\dagger Y_i$ *is singular* **then**
  return Y *is unstable.*
end
Initialize $g_1 = g_2 = \mathrm{I}_m$;
for $k = 1$ **to** N **do**
  if $\|\mu_G(g \cdot Y)\| \leq \varepsilon$ **then**
    return g
  else
    $\varrho_1 \leftarrow \sum_i (g \cdot Y)_i (g \cdot Y)_i^\dagger$; `/* 1st quantum marginal */`
    $g_1 \leftarrow \det(\varrho_1)^{1/(2m_1)} \varrho_1^{-1/2} g_1$; `/* scale 1st quantum marginal */`
    $\varrho_2 \leftarrow \left(\sum_i (g \cdot Y)_i^\dagger (g \cdot Y)_i\right)^{\mathsf{T}}$; `/* 2nd quantum marginal */`
    $g_2 \leftarrow \det(\varrho_2)^{1/(2m_2)} \varrho_2^{-1/2} g_2$; `/* scale 2nd quantum marginal */`
  end
end
return Y *is unstable.*

Remark 3.2.2. One can verify with Equation (2.19) that, after scaling the first quantum marginal in Algorithm 3.2, the moment map at $g \cdot Y$ is zero in the first component. Similarly, scaling the second quantum marginal results in a zero second component of μ_G at $g \cdot Y$, but this may violate the first component of $\mu_G(g \cdot Y)$ being zero. Therefore, operator scaling, and similarly other alternating minimization methods in computational invariant theory, can be seen as a *"block-coordinate gradient descent method"* [BFG+19, page 12]. ▽

The formulation of Algorithm 3.2 is based on [BGO+18], which generalizes the alternating minimization approaches for matrix and operator scaling to tensor scaling. For fixed $d \geq 3$, this yields a $\mathrm{poly}(m, 1/\varepsilon)$ time algorithm for the Scaling Problem 3.1.4 [BGO+18, Theorem 1] and an $\exp(m \log(m))$ time algorithm for NCM [BGO+18, Theorem 3.8].[21] On the other hand, deciding NCM for operator scaling only requires $\varepsilon = (\mathrm{poly}(m_1, m_2))^{-1}$ precision [Gur04a], so Algorithm 3.2 solves NCM in polynomial time.

Commutative Case

We shortly comment on algorithms in the case that $G = T$ is a torus, also compare [BFG+19, Subsection 1.4.1]. Since a vector v is in the null cone if and only

[21] Theorems 4.2.1 and 4.5.1 certify that deciding NCM for $\pi_{m,d}$, $d \geq 3$, requires *exponential* precision. Therefore, NCM cannot be solved in polynomial time by the methods in [BGO+18].

if $0 \notin \Delta_T(v)$,[22] one can solve NCM in polynomial time via linear programming [Kar84]. Moreover, remember from Equation (3.1) that Norm minimization is unconstrained geometric programming, which admits a convex optimization formulation. Thus, one can use ellipsoid methods, implicitly in [Gur04b; SV14; SV19], and interior point methods [BLNW20] to obtain polynomial time algorithms for the Computational Problems 3.1.2–3.1.4. Actually, the recent paper [BDM+21] provides polynomial time algorithms for the OCI Problem 3.1.1, orbit closure containment and even for orbit equality. These results are obtained by combining linear programming with algebraic methods. Interestingly, efficient optimization approaches to decide OCI seem to be intimately connected to the abc-conjecture [BDM+23].

Geodesic Convex Optimization

Given the success of optimization techniques for the commutative case and the geodesic convex structure in the non-commutative case, it is natural to aim for developing similar geodesic convex methods that solve Problems 3.1.2–3.1.4 for general reductive groups G.

Currently, the only implementable algorithms for Riemannian geodesic convex optimization are analogues of gradient descent (first order) and trust region methods[23] (second order) [AMS08; AGL+18; Bač14; BFG+19; Bou23; ZS16]. In particular, there are no efficiently implementable geodesic convex counterparts to the interior point or cutting plane methods available.[24] Of special interest for computational invariant theory are [AGL+18] and [BFG+19]. The work [AGL+18] provides geodesic second order methods specifically designed for operator scaling. These yield polynomial running time algorithms for OCI, NCM, norm minimization and scaling (Computational Problems 3.1.1–3.1.4).

The second order method of [AGL+18] was generalized to arbitrary reductive groups G in [BFG+19, Algorithm 5.1]. The latter paper also presents a gradient descent method for general reductive G, which can be seen as a generalization of alternating minimization methods, compare Remark 3.2.2. In the following we focus on [BFG+19], since it unifies existing optimization approaches in computational invariant theory, recovers polynomial running time for Computational Problems 3.1.2–3.1.4 in many settings, but also adds several new cases.[25] How-

[22] Recall that the proof of Theorem 2.1.9 was essentially due to Gordan's Theorem - a version of linear programming duality.

[23] also called *box constrained Newton's method*

[24] We remark that, very shortly before the submission of this thesis, the preprint [NW23] appeared. It provides the main stage of an interior point method on Riemannian manifolds. This is achieved by studying self-concordant functions on Riemannian manifolds. These functions are also studied in the related work [Hir22]. Due to time constraints, further details can unfortunately not be provided here. However, we stress that the diameter bound in Theorem 5.1.2 still excludes polynomial running time, compare the Introduction of the thesis as well as the paragraph "Implications of the main Results" in Section 5.1.

[25] In particular, [BFG+19] recovers polynomial running time for matrix scaling, simultaneous conjugation, operator scaling and GL-actions on quiver, while it adds the new cases of SL-actions on quivers with *fixed* number of vertices, and the tensor scaling action of $\mathrm{SL}_m(\mathbb{C}) \times \mathrm{SL}_m(\mathbb{C}) \times \mathrm{SL}_k(\mathbb{C})$ on $(\mathbb{C}^m)^{\otimes 2} \otimes \mathbb{C}^k$ for *fixed* k. Besides, it also recovers certain polynomial

ever, [BFG+19] cannot ensure polynomial time algorithms for tensor scaling.[26]

A very important technical contribution of [BFG+19] is to identify key complexity parameters called weight norm and weight margin. They are used to bound the running time of the first and second order method, to state a quantitative version of Kempf-Ness and to bound the diameter of an approximate minimizer. We outline this in the following.

Definition 3.2.3. Consider $\pi\colon G \to \mathrm{GL}(V)$ with Lie algebra representation Π.

1. [BFG+19, Definition 3.10] The *weight norm* of π is

$$N(\pi) := \max\left\{\|\Pi(X)\|_{\mathrm{op}} \mid X \in \mathfrak{i}\,\mathrm{Lie}(K), \|X\|_F = 1\right\},$$

   where $\|\cdot\|_{\mathrm{op}}$ is the usual operator norm on $\mathrm{End}(V)$.

2. [BFG+19, Definition 3.18] The *weight margin* of π is

$$\gamma_T(\pi) := \min\left\{\,\mathrm{dist}(0, \mathrm{conv}(\Gamma)) \mid \Gamma \subseteq \Omega(\pi),\, 0 \notin \mathrm{conv}(\Gamma)\right\},$$

   where $\mathrm{dist}(0, \mathrm{conv}(\Gamma)) := \min\{\|x\| \mid x \in \mathrm{conv}(\Gamma)\}$ is the Euclidean distance from zero to the polytope $\mathrm{conv}(\Gamma)$. ▲

By [BFG+19, Proposition 3.11], we have

$$N(\pi) = \max\{\|\omega\| \mid \omega \in \Omega(\pi)\}\,, \tag{3.6}$$

which justifies the name *weight norm*. With Eq. (4.5) we see for example that $N(\pi_{m,d}) \leq \sqrt{d}$. In [BFG+19] $N(\pi)$ is used to bound the norm of the moment map, [BFG+19, Lemma 3.12], and to provide a smoothness and a robustness parameter [BFG+19, Propositions 3.13 and 3.15]. These results are then used to control the step size in the algorithms of [BFG+19]. Upper bounds on the weight norm are given in [BFG+19, Lemma 6.1 and Example 6.3]). On the other hand, the weight margin $\gamma_T(\pi)$ is the crucial parameter for running time bounds in [BFG+19] and we report on lower bounds on $\gamma_T(\pi)$ from [BFG+19, Section 6] in Section 4.2.

Remark 3.2.4. Before we state the quantitative version of Kempf-Ness and the diameter bound we note the following.

(i) In [BFG+19] the capacity of v is defined as $\inf_{g\in G}\|g\cdot v\|$, while in this thesis it is the square of the latter: $\mathrm{cap}_G(v) = \inf_{g\in G}\|g\cdot v\|^2$.

(ii) Norm minimization [BFG+19, Problem 1.10] is formulated via *multiplicative* approximation: given $v \in V$ with $\mathrm{cap}_G(v) > 0$ and $\varepsilon > 0$, compute $g \in G$ such that

$$\log\big(\|g\cdot v\|\big) - \frac{1}{2}\log\big(\,\mathrm{cap}_G(v)\big) \leq \varepsilon. \tag{3.7}$$

running times for moment polytope membership, e.g., for Horn's problem.

[26] In fact, the results in Chapters 4 and 5 highly suggest that sophisticated methods, such as geodesic interior point, are necessary to ensure polynomial running time.

For $v \in V$ with $\operatorname{cap}_G(v) > 0$ the solutions between the additive and the multiplicative norm minimization are related as follows. If $g \in G$ solves Computational Problem 3.1.3, then using $\log(1+x) \leq x$ we see that it satisfies

$$\log\left(\frac{\|g \cdot v\|^2}{\operatorname{cap}_G(v)}\right) \leq \log\left(1 + \frac{\varepsilon}{\operatorname{cap}_G(v)}\right) \leq \frac{\varepsilon}{\operatorname{cap}_G(v)}.$$

Hence, g is solves Equation (3.7) for precision $(2\operatorname{cap}_G(v))^{-1}\varepsilon$. On the other hand, if $h \in G$ solves Equation (3.7) for $0 < \varepsilon \leq 1/2$, then

$$\frac{\|h \cdot v\|^2}{\operatorname{cap}_G(v)} \leq \exp(2\varepsilon) \leq 1 + 4\varepsilon,$$

where we used $\exp(x) \leq 1 + 2x$ for $0 \leq x \leq 1$. Thus, h solves Computational Problem 3.1.3 for precision $4\operatorname{cap}_G(v)\varepsilon$. ▽

Theorem 3.2.5 (Quantitative Kempf-Ness, [BFG+19, Theorem 1.17])**.** *Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation and take $v \in V\backslash\{0\}$. Then*

$$1 - \frac{\|\mu_G(v)\|_F}{\gamma_T(\pi)} \leq \frac{\operatorname{cap}_G(v)}{\|v\|^2} \leq 1 - \frac{\|\mu_G(v)\|_F^2}{4N(\pi)^2}. \tag{3.8}$$

Note that Equation (3.8) is indeed a quantitative version of and recovers Kempf-Ness, Theorem 2.2.13(a). An important application of the above theorem is that it connects solutions of norm minimization to those of scaling and vice versa [BFG+19, Corollary 1.18].

Next, we define the diameter. It captures how far a solution for Norm minimization Problem 3.1.3 is away from the identity in the Riemannian manifold $G/K \cong G \cap \mathrm{PD}_N(\mathbb{C})$ (see page 22).

Definition 3.2.6 (Diameter, [FR21, Definition 4.18])**.** Given $\pi\colon G \to \mathrm{GL}(V)$, $v \in V$ and a precision $\varepsilon > 0$. We define the *diameter* as

$$D_v(\varepsilon) := \inf\left\{R > 0 \mid \inf_{g \in B'_R} \|g \cdot v\|^2 \leq \operatorname{cap}_G(v) + \varepsilon\right\},$$

where $B'_R := \{k\exp(X) \mid k \in K, X \in \mathfrak{i}\,\mathrm{Lie}(K), \|X\|_F \leq R\}$.[27] ▲

The following (simplified) diameter bound is obtained from [BFG+19].

Theorem 3.2.7. *As usual, assume Setting 3.0.1. In particular, $\pi\colon G \to \mathrm{GL}(V)$ is a rational representation of a Zariski closed and self-adjoint subgroup $G \subseteq \mathrm{GL}_N(\mathbb{C})$. Let $v \in V$ with $\operatorname{cap}_G(v) > 0$ and assume $0 < \varepsilon \leq 2\operatorname{cap}_G(v)$. Then*

$$D_v(\varepsilon) \leq \frac{\sqrt{N}}{2}\log(2N) + \frac{\sqrt{N}}{2}\gamma_T(\pi)^{-1}\log\left(\frac{2\|v\|^2}{\varepsilon}\right). \tag{3.9}$$

[27] The set $B_R := \{\exp(X) \mid X \in \mathfrak{i}\,\mathrm{Lie}(K), \|X\|_F \leq R\}$ is a geodesic ball of radius R in G/K about the identity. Since K acts isometrically on V, we see that $D_v(\varepsilon)$ indeed captures the distance of an approximate minimizer to the identity.

Proof. Let $\varepsilon' > 0$ and remember Remark 3.2.4(i) for consulting [BFG+19]. By [BFG+19, Proposition 5.6], there exists a group element $g \in G$ that satisfies Equation (3.7), i.e., $\log\big(\|g \cdot v\|\big) - 2^{-1}\log\big(\operatorname{cap}_G(v)\big) \leq \varepsilon'$, and

$$\log \operatorname{reg}(g) \leq \log(2N) + \gamma_T(\pi)^{-1} \log\left(\frac{\|v\|^2}{2\operatorname{cap}_G(v)\varepsilon'}\right), \tag{3.10}$$

where $\operatorname{reg}(g) := \|g\|_F^2 + \|g^{-1}\|_F^2$. We use this to establish a bound on the diameter.

By the polar decomposition (Theorem 1.2.16), $g = k\exp(X)$ for some unitary matrix $k \in G \cap \mathrm{U}_N$ and $X \in \mathfrak{p} \subseteq \mathrm{Sym}_N(\mathbb{C})$. Note that we can assume $g = \exp(X)$, i.e., $k = \mathrm{I}_N$, since $\operatorname{reg}(\cdot)$ is left U_N-invariant and $G \cap \mathrm{U}_N$ acts isometrically on V. Now, if additionally $\varepsilon' \leq 1/2$, then Remark 3.2.4 implies that

$$\|g \cdot v\|^2 \leq \operatorname{cap}_G(v) + \varepsilon' 4 \operatorname{cap}_G(v), \quad \text{hence} \quad D_v\big(\varepsilon' 4 \operatorname{cap}_G(v)\big) \leq \|X\|_F\,. \tag{3.11}$$

To bound $\|X\|_F$, let $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ be the eigenvalues of the Hermitian matrix X such that $|\lambda_1| \leq \cdots \leq |\lambda_N|$. Then $\|X\|_F \leq \sqrt{N}|\lambda_N|$ and the $\exp(\lambda_i)$ are the eigenvalues of $\exp(X)$. Now, we have

$$|\lambda_N| = \begin{cases} \log(\exp(\lambda_N)) \leq \log\big(\|\exp(X)\|_F\big) = \log\big(\|g\|_F\big) & \text{if } \lambda_N \geq 0 \\ \log(\exp(-\lambda_N)) \leq \log\big(\|\exp(-X)\|_F\big) = \log\big(\|g^{-1}\|_F\big) & \text{if } \lambda_N < 0 \end{cases}$$

In any case, $\|X\|_F \leq \sqrt{N}|\lambda_N| \leq \sqrt{N}2^{-1}\log \operatorname{reg}(g)$. Combining the latter with Equations (3.10) and (3.11) yields

$$D_v\big(\varepsilon' 4 \operatorname{cap}_G(v)\big) \leq \|X\|_F \leq \frac{\sqrt{N}}{2}\log(2N) + \frac{\sqrt{N}}{2}\gamma_T(\pi)^{-1}\log\left(\frac{\|v\|^2}{2\operatorname{cap}_G(v)\varepsilon'}\right).$$

Finally, for $0 < \varepsilon \leq 2\operatorname{cap}_G(v)$ we can use $\varepsilon' := \varepsilon(4\operatorname{cap}_G(v))^{-1} \leq 1/2$ to obtain the desired bound (3.9). □

We end with a dichotomy regarding running times for the representation $\pi_{m,d}$. This motivated the work [FR21] which is presented in Chapters 4 and 5. For this, we remark that it is desirable to solve the Norm minimization Problem 3.1.3 for $\pi_{m,d}$ efficiently with *high precision* (HP). That is, solving it in $\operatorname{poly}(m, d, \log(1/\varepsilon))$ time. The state of the art regarding NCM and HP for $\pi_{m,d}$ is given in Table 3.1.

$\pi_{m,d}$	$T = \mathrm{ST}_m(\mathbb{C})^d$: commutative	$G = \mathrm{SL}_m(\mathbb{C})^d$: non-commutative
$d = 2$	**matrix scaling:** HP, NCM (trust region, ellipsoid, IPM)	**operator scaling:** HP, NCM (via trust region)
$d = 3$	**array scaling:** HP, NCM (via IPM and ellipsoid; *not* via trust region)	**tensor scaling:** HP, NCM (no IPM available)

Table 3.1: Dichotomy for $\pi_{m,d}$ between $d = 2$ and $d = 3$. Green indicates polynomial running time, while red means no polynomial time. IPM is a shortcut for interior point method.

This raises the following questions. Can we explain the dichotomy between $d = 2$ and $d = 3$ given in Table 3.1? More specifically:

- Why do gradient descent and trust region methods do not seem to yield polynomial time for HP and NCM when $d = 3$?
- Are known algorithms actually good enough for tensor scaling and only the complexity analysis lacks to show this? Or do we need new algorithmic approaches?

To answer these questions we investigate for NCM bounds on the precision parameters *weight margin* and *gap* in Chapter 4. Regarding HP we provide exponentially large lower bounds on the *diameter* for $\pi_{m,3}$ in Chapter 5. These are the main results of [FR21].[28] They highly suggest that new algorithmic approaches[29] for geodesic convex optimization are necessary to ensure polynomial time for HP and NCM in the case of tensor scaling.

[28] These hardness results align with similar results for the algebraic approach: degree lower bounds for invariant polynomials for the 3-tensor action pose significant challenges [DM20b].

[29] e.g., interior point like methods

Chapter 4

Bounds on Weight Margin and Gap

The material in this chapter is based on [FR21] and contains all upper bounds on weight margin and gap from that paper. We give such upper bounds for tensor scaling, polynomial scaling and SL actions on a certain family of quivers. In the tensor scaling case, these exponentially small bounds explain the dichotomy for null cone membership (NCM) from Table 3.1. Together with the diameter bounds in Chapter 5 they strongly motivate the need of new geodesically convex methods, such as interior-point like algorithms.

All main proof ideas for these upper bounds are due to myself.[1] However, the concept of freeness from Section 4.3 is well-known in the literature and we give corresponding references. Moreover, the lower bound on the gap for a family of quivers in Subsection 4.7.2 was proven by Cole Franks and Visu Makam. I thank them for their permission to include these arguments. Their lower bound showcases an important distinction between weight margin and gap, and answers [FR21, Problem 4.27] in the affirmative.

Organization and Assumptions. In Section 4.1 we introduce the concepts of weight margin and gap from [BFG+19]. A detailed discussion of the main results and related literature is provided in Section 4.2. Afterwards, we present in Section 4.3 the concept of free sets of weights, which is a crucial part of the proof method, Section 4.4. We prove the main results on tensor scaling in several steps, Section 4.5. This in turn allows to deduce similar bounds for polynomial scaling, compare Section 4.6. Finally, Section 4.7 studies the SL-action on a certain family of quivers: we give upper bounds on weight margin and gap, and the lower bound on the gap by Cole Franks and Visu Makam.

The assumptions for this chapter are as in Setting 3.0.1.

4.1 Weight Margin and Gap

In the following we formally define the weight margin and gap, which were first introduced in [BFG+19]. As a motivation, recall the "duality" (2.21), which can be reformulated as (2.25) using the moment polytope $\Delta_G(v)$.

Definition 4.1.1 ([FR21, Definition 4.3])**.** Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational

[1] In contrast, the diameter bounds in Chapter 5 are due to my co-author Cole Franks.

representation. We define the *gap* of π as[2]

$$\begin{aligned}\gamma_G(\pi) &:= \min\left\{\|\mu_G(v)\|_F \mid v \neq 0 \text{ is } G\text{-unstable}\right\}\\ &= \min\left\{\operatorname{dist}\big(0, \Delta_G(v)\big) \mid v \neq 0 \text{ is } G\text{-unstable}\right\},\end{aligned}$$

and the *weight margin* of π as

$$\begin{aligned}\gamma_T(\pi) &:= \min\left\{\|\mu_T(v)\|_F \mid v \neq 0 \text{ is } T\text{-unstable}\right\}\\ &= \min\left\{\operatorname{dist}\big(0, \Delta_T(v)\big) \mid v \neq 0 \text{ is } T\text{-unstable}\right\}.\\ &= \min\left\{\operatorname{dist}\big(0, \operatorname{conv}(\Gamma)\big) \mid \Gamma \subseteq \Omega(\pi),\, 0 \notin \operatorname{conv}(\Gamma)\right\}.\end{aligned}$$

The last equality uses that the weight polytope $\Delta_T(v)$ is $\operatorname{conv}(\Gamma)$ for $\Gamma = \operatorname{supp}(v)$. Hence, the above definition of $\gamma_T(\pi)$ aligns with Definition 3.2.3. ▲

If G is a torus, i.e., $G = T$, then the weight margin is simply the gap. The description of weight margin and gap via weight respectively moment polytopes will enable us to find small upper bounds via extremal combinatorics of the polytopes. Let us state two important remarks on weight margin and gap.

Remark 4.1.2 (Gap and Weight Margin are Precision Parameters).
By definition, the gap $\gamma_G(\pi)$ is the largest constant $C > 0$ with the following property: if $\|\mu_G(v)\|_F < C$ for some vector $v \in V$, then v is G-semistable. The same statement holds for the weight margin $\gamma_T(\pi)$ replacing G by T. Therefore, these notions capture how small $\mu_G(g \cdot v)$ (respectively $\mu_T(t \cdot v)$) must be to certify null-cone non-membership. Hence, $\gamma_G(\pi)$ and $\gamma_T(\pi)$ are the precision parameters if the Scaling Problem 3.1.4 is used to solve the NCM Problem 3.1.2. ▽

The next remark connects the gap to the classical notion of *instability* due to Mumford [MFK94].

Remark 4.1.3 (Gap as Mumford's Instability, [FR21, Remark 4.4]).
Denote the instability of a vector v by $M(v)$, see e.g., [Nes84, Eq. (9)]. It is positive if and only if v is unstable. Now, if v is non-zero and unstable then $\operatorname{dist}(0, \Delta_G(v)) \geq 2M(v)$ by [Nes84, (13)]. Together with [Nes84, Theorem 6.1] this implies

$$\gamma_G(\pi) = \inf\left\{2M(v) \mid v \neq 0, v \text{ is } G\text{-unstable}\right\}.$$

In words, the gap is twice the minimum value of all positive instabilities.

We note that Mumford's instability $M(v)$ is defined as a supremum over one-parameter subgroups (1-psg's) of G, and this supremum is attained. A 1-psg that witnesses the instability $M(v)$ is called *adapted*[3] for v and such 1-psg's play an important role in [Kem78]. As a consequence of the above observation the gap (and weight margin) may be studied via adapted 1-psg's. ▽

Weight margin and gap satisfy the following inequality, also see [BFG+19, Lemma 3.19].

[2]Gap and weight margin are well-defined, i.e., the minimum is attained. Indeed, the moment maps give rise to continuous maps on $\mathbb{P}(V)$ and the non-zero G-unstable (respectively non-zero T-unstable) vectors form a projective subvariety of $\mathbb{P}(V)$; in particular, they form a compact set.

[3]Adapted 1-psg's are also known as Kempf-optimal subgroups.

Proposition 4.1.4 ([FR21, Proposition 4.6])**.** *It holds that* $\gamma_T(\pi) \leq \gamma_G(\pi)$.

Proof. Let $v \neq 0$ be G-unstable. Then there exists $k \in K$ such that $k \cdot v$ is T-unstable; see Theorem 2.1.5. We obtain

$$\|\mu_G(v)\|_F = \|\mu_G(k \cdot v)\|_F \geq \|\mu_T(k \cdot v)\|_F \geq \gamma_T(\pi)\,,$$

where we used $\mu_G(k \cdot v) = k\mu_G(v)k^\dagger$ (Proposition 2.2.7) in the equality, and Proposition 2.2.6 in the first inequality. We deduce $\gamma_G(\pi) \geq \gamma_T(\pi)$ from the displayed equation. □

Further properties of weight margin and gap are listed in Proposition 4.3.10. Let us end this section with an interesting open problem which is already posed in [BFG+19, Remark 3.20].

Problem 4.1.5. *Is the quantitative non-commutative duality from Theorem 3.2.5 still true*[4] *if one replaces the weight margin* $\gamma_T(\pi)$ *by the gap* $\gamma_G(\pi)$*?*

If the question is answered in the affirmative, then the (possibly larger) gap can replace the weight margin in all appearances of running time bounds and the diameter bound in [BFG+19].

4.2 Main Results and related Literature

In this section we present and discuss the main results on weight margin and gap. First, we stress the relevance of these complexity parameters and review known lower bounds. Afterwards, we state the main result on array/tensor scaling, Theorem 4.2.1, and discuss its implications and relation to the literature. We discuss and relate the main results on two other actions, which are studied in Section 4.6 and 4.7 respectively. We end with an open Question 4.2.3 on possible implications for moment polytopes over the real numbers.

Significance of Weight Margin and Gap. We discuss four important features of the complexity parameters weight margin and gap.

First, the weight margin and gap capture the *required precision* needed in the Scaling Problem 3.1.4 in order to decide the NCM Problem 3.1.2, compare Remark 4.1.2. Thus, the smaller the weight margin (respectively gap) is, the higher is the required precision to decide whether the optimization value of the underlying geometric program (respectively geodesic optimization problem) is positive. For an illustration of this fact the reader is referred to the extended example on matrix scaling in Section 3.1.

Second, as a consequence of [BFG+19, Proposition 5.6] the weight margin upper bounds the diameter (Theorem 3.2.7):

$$D_v(\varepsilon) = O\big(\sqrt{N}\log(N) + \gamma_T(\pi)^{-1}\log(\|v\|/\varepsilon)\big).$$

[4] perhaps, in a reasonable adjusted manner

Therefore, the smaller the weight margin is the larger the diameter may be, which can prevent efficient algorithms. We point out that diameter upper bounds play an important role in the literature, compare Section 5.1.

Third, the inverse of the weight margin appears (polynomially) in running time bounds of geodesic methods in [BFG+19]. More precisely, it appears in running time bounds for NCM,[5] e.g., in [BFG+19, Corollary 1.26] and for Norm Minimization, e.g., in [BFG+19, Theorem 1.22]. Therefore, an exponentially small weight margin only ensures exponential running time, while if polynomially small it yields a polynomial time algorithm.

Finally, we recall that the weight margin controls the lower bounds in the quantitative non-commutative duality in Theorem 3.2.5. As a consequence, the weight margin controls when an output for the Scaling Problem 3.1.4 is a valid output for the Norm Minimization Problem 3.1.3 [BFG+19, Corollary 1.18]; and it also characterizes the required precision in Norm Minimization to decide NCM [BFG+19, Corollary 1.19]. Note that the second and third property would also apply to the gap, if the (possibly larger)[6] gap can replace the weight margin in Theorem 3.2.5 (see open Problem 4.1.5).

Known lower Bounds. Before we state the main result for tensor scaling we briefly review known lower bounds for the weight margin $\gamma_T(\pi)$ (and hence the gap by Proposition 4.1.4).

In the case of matrix scaling and operator scaling it is known that

$$\Omega\big(m^{3/2}\big) = \gamma_T(\pi_{m,2}) = \gamma_T\big(\pi_{m,2}^{\oplus n}\big)\,, \tag{4.1}$$

see [LSW00; Gur04a]. This good bound can be attributed to the extraordinary geometry of $\Omega(\pi_{m,2})$: its elements form the columns of a totally unimodular matrix (up to a shift). Similar good bounds on the weight margin are given in [BFG+19, Corollaries 6.11 and 6.18] provided the weight matrix is (up to a shift) totally unimodular.

Moreover, [BFG+19, Theorem 6.24] gives lower bounds for GL-actions and for SL-actions on quivers. The most general lower bounds are provided in [BFG+19, Theorem 6.10]: they hold for any rational representation for a product of GL's, respectively of SL's. The SL-case, i.e., [BFG+19, Theorem 6.10, Item 3], applied to the representation $\pi_{m,d}$ capturing array and tensor scaling yields

$$\Omega\big((m\sqrt{d})^{-md}\big) = \gamma_T(\pi_{m,d}), \tag{4.2}$$

where we used $N(\pi_m, d) \leq \sqrt{d}$ (see Eq. (3.6) and below). Comparing this general bound with Equation (4.1) for the special case $d = 2$ shows a huge discrepancy. This actually relates to the dichotomy presented in Table 3.1 as follows.

Main Result on Tensor Scaling. Given the just mentioned discrepancy, it is natural to ask whether the lower bound for the weight margin (and the gap)

[5] This is tackled by solving the scaling problem with the precision required by the weight margin.

[6] recall Proposition 4.1.4

is too pessimistic. The main result shows that this is not the case: the weight margin *and* the gap become exponentially small in md for $d \geq 3$.

Theorem 4.2.1 (General Tensor Gap, [FR21, Theorems 1.3 and 1.6])**.** *There is a constant $C > 0$, independent of m and d, such that for all $d \geq 3$ and $m \geq 2$, the weight margin and the gap for d-tensor scaling satisfy*

$$\gamma_T(\pi_{m,d}) \leq \gamma_G(\pi_{m,d}) \leq 2^{-Cdm}.$$

A detailed statement on upper bounds for gap and weight margin can be found in Theorem 4.5.1, and we show in Subsection 4.5.4 how to fill in the missing values of m and d to obtain Theorem 4.2.1. We note that the upper bounds in Theorems 4.2.1 and 4.5.1 are provided by constructing free tensors, whose support has $O(md)$ elements.

Remark 4.2.2 (Constant in Theorem 4.2.1)*.* The constant $C = 1/16$ works for all $m \geq 2$, $d \geq 3$. For $m, d \gg 0$ one can choose $C \approx 1/6$, compare Theorem 4.5.1. $\triangledown$

Implications of Main Theorem. Taking the paragraph on the significance of weight margin and gap into account, Theorem 4.2.1 implies the following.

First of all, it shows that exponentially high precision is required to solve NCM for array and tensor scaling. In particular, current first and second order methods do not seem to be able to solve NCM for tensor scaling in poly time. Certainly, current running time bounds are exponential in md. Similarly, the main theorem suggests that ellipsoid and interior point methods are necessary for array scaling to ensure polynomial running time. This explains the dichotomy for NCM that we presented in Table 3.1 (Section 3.2).

Moreover, Theorem 4.2.1 yields that the upper bound on the diameter from Theorem 3.2.7 is exponentially large. In fact, Theorems 5.1.1 and 5.1.2 show that diameter *is* exponentially large in the high precision regime for 3-order array and tensor scaling. Finally, we point out that running time and diameter upper bounds remain exponentially large even if we could replace the weight margin by the gap. Hence, an affirmative answer to Problem 4.1.5 would not help for tensor scaling.

Relation to the Literature. Theorem 4.2.1 aligns with existing results showing that the $d > 2$ array/tensor case is more complex than the matrix case. For example, it is known that the polytope of non-negative arrays with uniform marginals, known as the *d-index axial assignment polytope*, has many more vertices when $d \geq 3$ and that the vertices can have exponentially small entries [Kra07; LL14].[7] In contrast, for $d = 2$ this polytope is the Birkhoff-von Neumann polytope which has integral vertices by the Birkhoff-von Neumann theorem.

Next we discuss the case of local dimension two, i.e., $m = 2$, for which Theorem 4.5.1(a) provides a more concrete bound. For d-dimensional array scaling $\gamma_T(\pi_{2,d})$ is on the order of the weight margin of the d-dimensional hypercube $\{\pm 1\}^d$. Therefore, $\gamma_T(\pi_{2,d}) = d^{-\frac{d}{2}(1+o(1))}$ by [AV97]. This bound is better by a $\log(d)$ factor than the one in Theorem 4.5.1(a). However, an $\exp(-d)$

[7] Actually, we use such a vertex with exponentially small entry from [Kra07] to settle the $d = 3$ case.

for the *gap* $\gamma_G(\pi_{2,d})$ was not known before, also compare Remark 4.5.2. Still, there are interesting results regarding $\gamma_G(\pi_{2,d})$. First, using the algorithm in [MS15] Maciążek and Sawicki numerically found several free[8] tensors of format $(\mathbb{C}^2)^{\otimes d}$ with $\operatorname{dist}(0, \Delta_G(v))$ at most $\exp(-d)$; Theorem 4.2.1 confirms this exponential behaviour for all d (and all m). Second, [MS18, Main result] shows that $\operatorname{dist}(0, \Delta_G(v))^2$, where $0 \notin \Delta_G(v)$, tends for large d to the Gamma distribution $\Gamma(1/2, 2d)$.[9] Therefore, the moment polytopes that witness the exponential behaviour in Theorem 4.5.1(a) are rare. It is an interesting open[10] question whether a similar result holds for other parameter regimes, e.g., tensors of order three.

Finally, note that the exponential rate of decay in Theorem 4.2.1 is tight up to log factors (in the exponent), compare Equation (4.2). One may ask whether the true bound is $2^{-\Theta(md)}$ or $2^{-\Theta(md(\log m + \log d))}$ as in the lower bound. [AV97] shows that the latter is correct in the *commutative* case for $m = 2$.

Weight Margin and Gap results for other group actions

In addition to the tensor scaling action, we also consider two other actions of groups G of interest in computational invariant theory.

Polynomial Scaling. The first is the action of the special linear group on the space of homogeneous d-forms $\mathbb{C}[x_1, \ldots, x_n]_d$, in which $G = \mathrm{SL}_n(\mathbb{C})$ acts by $g \cdot p(x) = p(g^{-1}x)$ for $p \in \mathbb{C}[x_1, \ldots, x_n]_d$, see Section 4.6. This action and its null cone are crucial for constructing a moduli space of hypersurfaces of degree d in $\mathbb{P}^{n-1}(\mathbb{C})$, compare [Hos15, Section 7]. In fact, homogeneous d-forms were among the objects studied earliest in computational invariant theory, and much of the theory was developed to catalogue invariants of the $\mathrm{SL}(n)$ action on forms [Wey39]. Still, deciding null-cone membership for $d = 3$ is challenging. We explain the difficulty by showing that the gap for this action is inverse exponential in n as soon as $d \geq 3$, see Theorem 4.6.2. This shows that the diameter bound from [BFG+19] (Theorem 3.2.7) becomes exponentially large in n.

In the commutative case, i.e., $T = \mathrm{ST}_n(\mathbb{C})$, the capacity $\operatorname{cap}_T(p)$ recovers Gurvit's polynomial capacity [Gur04b; Gur06]. To decide whether the polynomial capacity is positive and for high precision approximations the bounds in Theorem 4.6.2 suggest the following. As soon as $d \geq 3$ sophisticated methods (like ellipsoid and interior point) are required to ensure polynomial running time, while gradient descent and trust region methods do not suffice.

Quiver Action. Second, in Section 4.7 we study the natural SL-action on a family of quivers. We note that quiver representations include the important cases of operator scaling and an action that captures Horn's problem. However, efficient algorithms for SL-actions on quivers are only known for certain cases. In this regard, the family in Section 4.7 is a very interesting example. The quivers in this family have d vertices, each endowed with dimension m, and $d - 1$ arrows. Theorem 4.7.1 gives the bound $O(m^{-d})$ on the weight margin,

[8] see Section 4.3

[9] The moment polytope $\Delta_G(v)$ is distributed as follows: [MS18] chooses d linearly independent vertices uniformly from the 2^d possible vertices; see [MS18, Sections III and IV] for details. We stress that v (called ϕ in [MS18]) is *not* endowed with a distribution.

[10] to the authors knowledge

i.e., it becomes exponentially small as the number of vertices d grows. Hence, the general lower bound [BFG+19, Theorem 6.21, Item 4] cannot be improved in this regard. However, the gap is only polynomially small in m and d, Theorem 4.7.6.[11] Therefore, weight margin and gap differ significantly for this action; and the first order method from [BFG+19] still suffices to decide NCM in polynomial time thanks to the large gap. In contrast, when allowing m copies of each arrow in the constructed quiver, i.e., $m(d-1)$ arrows in total, we can ensure the bound $O(m^{-d})$ for the gap as well, see Theorem 4.7.1. Therefore, current methods do not run in polynomial time for this enlarged quiver.

Outlook

Since all actions studied in this chapter are defined over $\mathbb{R}$ and allow for moment polytopes over $\mathbb{R}$, Remark 2.2.21 naturally leads to the following question.[12]

Question 4.2.3. *Do the upper bounds for the gap via complex moment polytopes from this chapter also hold for a gap defined analogously via real moment polytopes?*

4.3 Free Sets of Weights

We introduce the crucial tool for lifting bounds from the commutative (weight margin and diameter) to the non-commutative case (gap and diameter).

Proposition 4.1.4 states that $\gamma_T(\pi) \leq \gamma_G(\pi)$ and we will see in Section 4.7 that $\gamma_G(\pi)$ can be significantly larger than $\gamma_T(\pi)$. Therefore, an upper bound for the weight margin $\gamma_T(\pi)$ need not necessarily apply to the gap $\gamma_G(\pi)$. Still, many presented bounds in the commutative case transfer to the noncommutative case. For this, we crucially use the notion of a *free* subset of weights, which appears in many references such as [Sja98; Fra02; CVZ23]. In [DK85] freeness is called *strong orthogonality* and in [DM20b] it appears as *uncramped*.

Definition 4.3.1 ([FR21, Definition 4.7])**.** Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation with set of weights $\Omega(\pi)$.

A subset $\Gamma \subseteq \Omega(\pi)$ is called *free* if no two distinct elements of Γ differ by a root[13] of G. In other words, $\Gamma \cap (\Gamma + \alpha) = \emptyset$ holds for all roots α of G.

A vector $v \in V \setminus \{0\}$ is called *free* if its support $\mathrm{supp}(v) \subseteq \Omega(\pi)$ is free. ▲

For concreteness, let us translate the above general definition to the tensor scaling setting given by the representation $\pi_{m,d}$. Recall from Example 1.3.18 that $\Omega(\pi_{m,d}) = \{\epsilon_i \mid i \in [m]\}^d \subseteq (\mathbb{R}^m)^d$.

Definition 4.3.2 (Free sets, [FR21, Definition 4.12])**.** A set $\mathscr{W} \subseteq [m]^d$ is called *free*, if $i = (i_1, \ldots, i_d), j = (j_1, \ldots, j_d) \in \mathscr{W}$ with $i \neq j$ always implies that $|\{i_l \neq j_l \mid l = 1, \ldots, d\}| \geq 2$. ▲

[11] This result is due to Cole Franks and Visu Makam.

[12] The author only recently became aware of the concept of moment polytopes for $\mathbb{K} = \mathbb{R}$.

[13] see Definition 1.3.20

Proposition 4.3.3 ([FR21, Proposition 4.13])**.** *Let $\mathscr{W} \subseteq [m]^d$ and denote the induced subset of weights of $\pi_{m,d}$ by*

$$\Gamma_{\mathscr{W}} := \{(\epsilon_{i_1}, \dots, \epsilon_{i_d}) \mid (i_1, \dots, i_d) \in \mathscr{W}\} \subseteq (\mathbb{R}^m)^d.$$

Then $\mathscr{W}$ is a free set if and only if the set of weights $\Gamma_{\mathscr{W}} \subseteq \Omega(\pi_{m,d})$ is free.

Proof. The set of weights $\Gamma_{\mathscr{W}}$ is free if and only if no two distinct elements of $\Gamma_{\mathscr{W}}$ differ by a root of $G = \mathrm{SL}_m(\mathbb{C})^d$, see Definition 4.3.1. Furthermore, remember from Example 1.3.21 that the roots of G are

$$(e_i - e_j, 0_m, \dots, 0_m), (0_m, e_i - e_j, 0_m, \dots, 0_m), \dots, (0_m, \dots, 0_m, e_i - e_j) \in (\mathbb{R}^m)^d$$

for $i, j \in [m]$ with $i \neq j$. Now, if $\mathscr{W} \subseteq [m]^d$ is not free, then there exist $i = (i_1, \dots, i_d), j = (j_1, \dots, j_d) \in \mathscr{W}$ with $i \neq j$ such that they exactly differ one component. Without loss of generality we assume $i_1 \neq j_1$ and $i_l = j_l$ for $l = 2, \dots, m$. But then

$$(\epsilon_{i_1}, \dots, \epsilon_{i_d}) = (\epsilon_{j_1}, \dots, \epsilon_{j_d}) + (e_{i_1} - e_{j_1}, 0_m, \dots, 0_m),$$

and hence $\Gamma_{\mathscr{W}}$ is not free. The argument can be inverted to show that if $\Gamma_{\mathscr{W}}$ is not free, then $\mathscr{W}$ is not free. □

Remark 4.3.4. We point out that the notion of freeness in [CVZ23] requires less as follows. The authors of [CVZ23] call a tensor free, if there *exist* ordered bases of the tensor factors, such that the support with respect to these bases is free. In contrast, free in this thesis means that the support is free with respect to the ordered standard bases.[14]

Moreover, [CVZ23, Remark 4.17] gives a dimension argument that $(\mathbb{C}^m)^{\otimes 3}$ does contain *non-free* tensors as soon as $m \geq 5$. These non-free tensors from [CVZ23] are by the above also non-free in our sense. ▽

We illustrate consequences of Proposition 4.5.13 in two examples.

Example 4.3.5 (Freeness for Operator Scaling)**.** Let us consider operator scaling, i.e., the representation $\pi_{m,2}^{\oplus n}$. For $n = 1$ and $M \in \mathbb{C}^{m \times m}$, let $s(M) := \{(i, j) \in [m]^2 \mid M_{ij} \neq 0\}$ so that

$$\mathrm{supp}(M) = \Gamma_{s(M)} = \big\{(\epsilon_i, \epsilon_j) \mid M_{ij} \neq 0\big\} \subseteq \Omega(\pi_{m,2}).$$

Now, Proposition 4.3.3 shows that M is free if and only if M has at most one non-zero entry in each row and in each column. In particular, M is free and invertible if and only if $s(M) = s(P)$ for a permutation matrix P.

More generally, for $n \geq 1$ and $M = (M_1, \dots, M_n) \in (\mathbb{C}^{m \times m})^n$ we have

$$\mathrm{supp}(M) = \big\{(\epsilon_i, \epsilon_j) \mid \exists\, k \in [n] \colon\ (M_k)_{ij} \neq 0\big\} \subseteq \Omega\big(\pi_{m,2}^{\oplus n}\big) = \Omega(\pi_{m,2}).$$

Therefore, $M = (M_1, \dots, M_n)$ is free if and only if there is a permutation matrix P such that $s(M_1), \dots, s(M_n) \subseteq s(P)$. ♢

[14] This comes from the fact that we choose the maximal torus T to be in $\mathrm{GT}_N(\mathbb{C}) \subseteq \mathrm{GL}_N(\mathbb{C})$.

Example 4.3.6 (Freeness and Quantum Marginals)**.** Let $d = 3$ and consider a free tensor $v \in (\mathbb{C}^m)^{\otimes 3}$ with respect to tensor scaling $\pi_{m,d}$. Then its quantum marginals (see Example 2.2.12) are diagonal which is exemplified in the following. The first quantum marginal of v is $MM^\dagger$, where $M \in \mathbb{C}^{m \times m^2}$ given by $M_{i,(j,k)} = v_{ijk}$ is a flattening of v. For $s, t \in [m]$ with $s \neq t$ we compute

$$\left(MM^\dagger\right)_{s,t} = \sum_{j,k=1}^{m} M_{s,(j,k)} \, \overline{M_{t,(j,k)}} = \sum_{j,k=1}^{m} v_{s,j,k} \, \overline{v_{t,j,k}} = 0, \tag{4.3}$$

where we used that $v_{s,j,k} \, \overline{v_{t,j,k}} = 0$ holds by freeness of v and Proposition (4.5.13).[15]

This principle generalizes to tensors of any order d. Each off-diagonal entry of a quantum marginal is the inner product between distinct $d-1$-dimensional slices of a tensor, and if the support of the tensor is free then the supports of such slices are entirely disjoint. Hence, the quantum marginals are diagonal. ♢

Recall that for $\pi_{m,d}$ the components of the moment map are, up to addition of a scalar multiple of I_m, given by the quantum marginals, compare Example 2.2.12. Thus, $\mu_G(v)$ is diagonal for a free tensor $v \in (\mathbb{C}^m)^{\otimes d}$ and it follows that $\mu_G(v) = \mu_T(v)$. It is known that this fact generalizes to any rational representation and we use it to transfer bounds for the weight margin to bounds on the gap via Proposition 4.3.7. The latter appears implicitly in, e.g., [Sja98, Lemma 7.1] and [Fra02, Proposition 2.2], but we prove it below for completeness.

Thanks go to Visu Makam for pointing out that the equality $\mu_G(v) = \mu_T(v)$ still holds under a weaker condition on v, when the representation decomposes into orthogonal subrepresentations.[16] This can be used to turn a weight margin upper bound for quivers into a gap upper bound, see Theorem 4.7.1. The weaker condition also appears in [DM20b, Theorem 6.5].

Proposition 4.3.7 ([FR21, Proposition 4.8])**.** *Let $\pi \colon G \to \mathrm{GL}(V)$ be a rational representation over $\mathbb{C}$ and suppose $V = \bigoplus_{i=1}^{k} V_i$ is an orthogonal decomposition into G-subrepresentations with respect to the K-invariant inner product, that is used to define μ_T and μ_G. Let $v = \sum_{i=1}^{k} v_i \in V \setminus \{0\}$, $v_i \in V_i$ be such that all supports $\Gamma_i := \mathrm{supp}(v_i) \subseteq \Omega(\pi)$ are free. Set $\Gamma := \bigcup_i \Gamma_i = \mathrm{supp}(v)$. Then:*

(i) For all $t \in T$ it holds that $\mu_G(t \cdot v) \in \mathfrak{i}\,\mathrm{Lie}(T_K)$ and $\mu_G(t \cdot v) = \mu_T(t \cdot v)$.

(ii) If $0 \notin \Delta_T(v) = \mathrm{conv}(\Gamma)$, then the upper bound $\mathrm{dist}(0, \mathrm{conv}(\Gamma))$ for the weight margin $\gamma_T(\pi)$ also applies to the gap, i.e., $\gamma_G(\pi) \leq \mathrm{dist}(0, \mathrm{conv}(\Gamma))$.

Proof. The action of T preserves the supports Γ_i, and in particular preserves their freeness. Hence, it suffices to show $\mu_G(v) \in \mathfrak{i}\,\mathrm{Lie}(T_K)$, which immediately yields $\mu_G(v) = \mu_T(v)$ by Proposition 2.2.6. Moreover, the orthogonality with respect to the K-invariant inner product shows $\mu_G(v) = H_1 \oplus \cdots \oplus H_k$, where $H_i = \mu_G^{(i)}(v_i)$ is given by the moment map $\mu_G^{(i)}$ of the G-module V_i if $v_i \neq 0$ and otherwise $H_i = 0$. The latter holds similarly for μ_T.

[15] Equation (4.3) suggests why freeness is called *strong orthogonality* in [DK85]. The distinct slices $M_{s,\cdot}$ and $M_{t,\cdot}$ of v are not only orthogonal - actually each summand in (4.3) is zero.

[16] In a preliminary version of [FR21] Proposition 4.3.7 was stated for the case $k = 1$.

Therefore, we may assume $k = 1$, i.e., $v \neq 0$ has free support Γ. We write $v = \sum_{\omega\in\Gamma} v_\omega$ for $v_\omega \in V_\omega$. Recall from Example 1.3.21 the concept of a root α of G and its corresponding root space $\mathrm{Lie}(G)_\alpha$. For any root α of G and all $A \in \mathfrak{i}\,\mathrm{Lie}(K) \cap \mathrm{Lie}(G)_\alpha$ we have $\Pi(A)v_\omega = 0$, by $\Gamma \cap (\Gamma + \alpha) = \emptyset$ (i.e., freeness) and Proposition 1.3.22. Thus, $\Pi(A)v = 0$ and $\mathrm{tr}\big(\mu_G(v)A\big) = 0$ for all roots α and all $A \in \mathfrak{i}\,\mathrm{Lie}(K) \cap \mathrm{Lie}(G)_\alpha$. With the root space decomposition $\mathrm{Lie}(G) = \mathrm{Lie}(T) \oplus \bigoplus_\alpha \mathrm{Lie}(G)_\alpha$ (also see Example 1.3.21) we conclude $\mu_G(v) \in \mathfrak{i}\,\mathrm{Lie}(T_K)$. The first statement is proven.

For the second claim, assume $0 \notin \mathrm{conv}(\Gamma) = \Delta_T(v)$. Then v is T-unstable. In particular, v is G-unstable and thus

$$\gamma_G(\pi) \leq \mathrm{dist}\big(0, \Delta_G(v)\big).$$

On the other hand, we have

$$\mathrm{dist}\big(0, \Delta_G(v)\big) = \inf_{g\in G} \|\mu_G(g \cdot v)\|_F \leq \inf_{t\in T} \|\mu_G(t \cdot v)\|_F \overset{(*)}{=} \mathrm{dist}\big(0, \mathrm{conv}(\Gamma)\big),$$

where we used $\mu_G(t \cdot v) = \mu_T(t \cdot v)$ in $(*)$. We conclude by combining the two inequalities. □

Remark 4.3.8 ([FR21, Remark 4.9]). It is well-known that any rational representation $\pi\colon G \to \mathrm{GL}(V)$ can be decomposed into irreducible subrepresentations that are pairwise orthogonal with respect to the fixed K-invariant inner product. Therefore, to apply Proposition 4.3.7 it suffices to ensure freeness on the irreducible subrepresentations. ▽

A useful consequence of Proposition 4.3.7 is that semi/polystability of a free vector under G may be checked on the torus T.[17] This application of freeness can be found in [DK85, Proposition 1.2] and [DM20b, Theorem 6.5] to construct vectors with closed G-orbit.

Corollary 4.3.9. *Let $v \in V$ be a free vector. If v is T-semistable (respectively T-polystable) then v is G-semistable (respectively G-polystable).*

Proof. Since v is free we have $\mu_T(t\cdot v) = \mu_G(t\cdot v)$ for all $t \in T$, by Proposition 4.3.7. If v is T-polystable, then there exists some $t \in T$ with $0 = \mu_T(t \cdot v) = \mu_G(t \cdot v)$, by Kempf-Ness Theorem 2.2.13(e) for the action of T. But the same part of Kempf Ness for the action of G yields that v is G-polystable as $t \in G$ and $\mu_G(t \cdot v) = 0$.

Similarly, if v is T-semistable we obtain that v is G-semistable using Kempf-Ness, Theorem 2.2.13(f), and continuity of the moment maps μ_T and μ_G. □

We end with an interesting connection between weight margin and gap.

Proposition 4.3.10 ([FR21, Proposition 4.10])**.** *Let $\pi\colon G \to \mathrm{GL}(V)$ be a rational representation over $\mathbb{C}$ and denote its n-fold direct sum by $\pi^{\oplus n}$.*

1. *The weight margin satisfies $\gamma_T(\pi) = \gamma_T(\pi^{\oplus n})$ for all $n \geq 1$.*
2. *The gap satisfies $\gamma_G(\pi^{\oplus n}) \geq \gamma_G\big(\pi^{\oplus(n+1)}\big)$ for all $n \geq 1$.*

[17] I thank M. Levent Doğan for a fruitful discussion, in which we rediscovered this fact.

3. There exists some $n \leq \dim(V)$ such that $\gamma_G(\pi^{\oplus n}) = \gamma_T(\pi^{\oplus n}) = \gamma_T(\pi)$.

Proof. We note that $\pi^{\oplus n}$ is given by the action $g \cdot (v_1, \dots, v_n) = (g \cdot v_1, \dots, g \cdot v_n)$ on V^n. Furthermore, the K-invariant inner product $\langle \cdot, \cdot \rangle$ of V induces naturally a K-invariant product on V^n by

$$\langle (v_1, \dots, v_n), (w_1, \dots, w_n) \rangle_{V^n} := \sum_{i=1}^{n} \langle v_i, w_i \rangle.$$

For the first claim, just remember that $\Omega(\pi^{\oplus n}) = \Omega(\pi)$ by Remark 1.3.15.

For the second claim, let $(v_1, \dots, v_n) \in V^n \setminus \{0\}$ be G-unstable such that $\|\mu_G(v_1, \dots, v_n)\|_F = \gamma_G(\pi^{\oplus n})$. Then $(v_1, \dots, v_n, 0) \in V^{n+1} \setminus \{0\}$ is G-unstable as well, so $\|\mu_G(v_1, \dots, v_n, 0)\|_F \geq \gamma_G\big(\pi^{\oplus (n+1)}\big)$. Moreover, under the inner product $\langle \cdot, \cdot \rangle_{V^{n+1}}$ the first n copies of V are orthogonal to the last copy. Thus, $\mu_G(v_1, \dots, v_n, 0)$ is the 2×2 block-diagonal matrix $\operatorname{diag}(\mu_G(v_1, \dots, v_n), 0)$ and hence $\|\mu_G(v_1, \dots, v_n, 0)\|_F = \|\mu_G(v_1, \dots, v_n)\|_F = \gamma_G(\pi^{\oplus n})$.

Finally, let $\Gamma = \{\omega_1, \dots, \omega_n\} \subseteq \Omega(\pi)$ be a witness of the weight margin, i.e., $0 \notin \operatorname{conv}(\Gamma)$ and $\operatorname{dist}(0, \operatorname{conv}(\Gamma)) = \gamma_T(\pi)$. We have $n \leq |\Omega(\pi)| \leq \dim(V)$ by the weight space decomposition $V = \bigoplus_{\omega \in \Omega(\pi)} V_\omega$, see Theorem 1.3.14. Now, for each $\omega_i \in \Gamma$ fix some weight vector $v_i \in V_{\omega_i} \setminus \{0\}$. Then $v := (v_1, \dots, v_n) \in V^n$ satisfies the assumptions of Proposition 4.3.7, because $\Gamma_i = \{\omega_i\}$ is free and the distinct copies of V are orthogonal under $\langle \cdot, \cdot \rangle_{V^n}$. Thus, we obtain

$$\gamma_G(\pi^{\oplus n}) \leq \operatorname{dist}\big(0, \operatorname{conv}(\Gamma)\big) = \gamma_T(\pi) = \gamma_T(\pi^{\oplus n}),$$

but on the other hand $\gamma_G(\pi^{\oplus n}) \geq \gamma_T(\pi^{\oplus n})$ by Proposition 4.1.4. □

4.4 Proof Method

In this short section we present the main steps how we prove upper bounds on the weight margin $\gamma_T(\pi)$ and the gap $\gamma_G(\pi)$.

1. We exhibit a set of weights $\Gamma \subseteq \Omega(\pi)$ such that $0 \notin \operatorname{conv}(\Gamma)$. Hence, $\gamma_T(\pi) \leq \operatorname{dist}(0, \operatorname{conv}(\Gamma))$ by Definition 4.1.1.

2. We prove an upper bound on $\operatorname{dist}(0, \operatorname{conv}(\Gamma))$ to obtain a bound on $\gamma_T(\pi)$.

3. If Γ satisfies the assumptions of Proposition 4.3.7 (e.g., if Γ is free), then also $\gamma_G(\pi) \leq \operatorname{dist}(0, \operatorname{conv}(\Gamma))$ holds by Proposition 4.3.7(ii). Therefore, the upper bound from the second step also applies to the gap $\gamma_G(\pi)$.

For the first and second step we often use Lemma 4.4.1 below. Recall that an *affine linear combination* of $v_1, \dots, v_k \in \mathbb{R}^m$ is $\lambda_1 v_1 + \dots + \lambda_k v_k$ for $\lambda_i \geq 0, \sum_{i=1}^{k} \lambda_i = 1$. The affine hull $\operatorname{aff}(S)$ of a set $S \subset \mathbb{R}^m$ is the set of all affine linear combinations of finite subsets of S, or equivalently the affine space of lowest dimension containing S. Furthermore, remember that $\epsilon_i = e_i - \frac{1}{m}\mathbb{1}_m$.

Lemma 4.4.1 ([FR21, Lemma 2.2])**.** *In $\mathbb{R}^m$ we have*

$$\sum_{i=1}^{m} \frac{1}{m}\,\epsilon_i = 0_m \tag{4.4}$$

and this is the only affine linear combination of $\epsilon_1, \dots, \epsilon_m$ giving zero.

Proof. One calculates directly that $\sum_i \frac{1}{m}\,\epsilon_i = 0_m$. To show uniqueness of this affine combination, we note that the vectors $e_2, \dots, e_m, \mathbb{1}_m$ are linearly independent. Thus, $\epsilon_2, \dots, \epsilon_m$ are linearly independent. On the other hand, $\epsilon_1, \dots, \epsilon_m$ are linearly dependent. Therefore, $\{(\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m \mid \sum_i \lambda_i\,\epsilon_i = 0_m\}$ is a one-dimensional subspace of $\mathbb{R}^m$, which yields the uniqueness of the affine linear combination. □

Finally, for $\epsilon_i \in \mathbb{R}^m$ we make the simple observation that

$$\|\epsilon_i\|^2 = \left(1 - \frac{1}{m}\right)^2 + (m-1)\frac{1}{m^2} = 1 - \frac{1}{m}, \quad \text{hence } \|(\epsilon_{i_1}, \dots, \epsilon_{i_d})\| \le \sqrt{d}\,. \tag{4.5}$$

4.5 Tensor Scaling

We recall that $\pi_{m,d}$ is the natural representation of $G = \mathrm{SL}_m(\mathbb{C})^d$ on $(\mathbb{C}^m)^{\otimes d}$ (Example 1.3.5), which captures tensor scaling while its restriction to $T = \mathrm{ST}_m(\mathbb{C})^d$ captures array scaling. Moreover, remember from Example 1.3.18 that

$$\Omega(\pi_{m,d}) = \big\{(\epsilon_{i_1}, \dots, \epsilon_{i_d}) \mid i_1, \dots, i_d \in [m]\big\}^d \subseteq (\mathbb{R}^m)^d.$$

The purpose of this section is to prove exponentially small upper bounds on the weight margin $\gamma_T(\pi_{m,d})$ and the gap $\gamma_G(\pi_{m,d})$ for the case $d \ge 3$.

Theorem 4.5.1 (Bounds for Tensor Gap, [FR21, Theorems 2.1 and 4.11])**.** *Let $\pi_{m,d}$ be the natural representation of $G := \mathrm{SL}_m(\mathbb{C})^d$ on $(\mathbb{C}^m)^{\otimes d}$. The weight margin $\gamma_T(\pi_{m,d})$ and the gap $\gamma_G(\pi_{m,d})$ are bounded as follows:*

(a) If $m = 2$ and $d \ge 3$, then $\gamma_T(\pi_{2,d}) \le \gamma_G(\pi_{2,d}) \le 2^{-\frac{d}{2}+1}$.

(b) If $m \ge 3$ and $d = 3$, then $\gamma_T(\pi_{m,3}) \le \gamma_G(\pi_{m,3}) \le 2^{-m+1}$.

(c) If $m \ge 3$ and $d = 6r - 3$ for some integer $r \ge 2$, then

$$\gamma_T(\pi_{m,d}) \le \gamma_G(\pi_{m,d}) \le \frac{\sqrt{6}}{(m-1)\sqrt{r}}\; 2^{-r(m-1)+1} \le 2^{-r(m-1)+1} = 2^{-\frac{(d+3)(m-1)}{6}+1}.$$

We prove parts (a), (b) and (c) of the preceding theorem in Subsections 4.5.1, 4.5.2 and 4.5.3, respectively. To do so, we proceed as described in Section 4.4.

Even though Theorem 4.5.1 only applies to certain $d \ge 3$, we can "pad" tensor factors to obtain similar results for all $d \ge 3$. This padding procedure is described in Subsection 4.5.4 and allows us to conclude Theorem 4.2.1 from the above Theorem 4.5.1.

4.5.1 Local Dimension two: Qubits

In this subsection we prove part (a) of Theorem 4.5.1, which states that $\gamma_T(\pi_{2,d})$ and $\gamma_T(\pi_{2,d})$ are exponentially small in d. We start with a remark on related literature.

Remark 4.5.2. We point out that $\gamma_T(\pi_{2,d}) = 2^{-\Theta(d \log d)}$ follows from [AV97]. This statement is actually stronger than the provided bound from Theorem 4.5.1(a). However, the result in [AV97] is obtained by describing an involved algorithm that constructs ill-conditioned ± 1-matrices. Thus, it is difficult to verify whether their construction produces free sets of weights. The latter is needed to lift the bound to the gap $\gamma_T(\pi_{2,d})$. In contrast, the construction presented here is simpler and proven to be free. ▽

In the following we construct a subset of

$$\Omega(\pi_{2,d}) = \left\{(\epsilon_{i_1}, \dots, \epsilon_{i_d}) \mid i_1, \dots, i_d \in [2]\right\} \subseteq \left(\mathbb{R}^2\right)^d,$$

which witnesses the exponentially small weight margin. For this, we construct a matrix with entries in [2], and each row of the matrix will correspond to an element of $\Omega(\pi_{2,d})$. For example, the row $(1, 2, 2)$ would correspond to $(\epsilon_1, \epsilon_2, \epsilon_2) \in \Omega(\pi_{2,3})$. To do so, we consider the matrices

$$A_2 := \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}, \; B_1 := \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}, \; B_2 := \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}, \; B_3 := \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

and define recursively

$$A_{2r+2} := \begin{pmatrix} & & & B_1 \\ & A_{2r} & & \vdots \\ & & & B_1 \\ B_2 & \cdots & B_2 & B_3 \end{pmatrix} = \begin{pmatrix} A_2 & B_1 & \cdots & B_1 \\ B_2 & B_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & B_1 \\ B_2 & \cdots & B_2 & B_3 \end{pmatrix} \tag{4.6}$$

for $r \geq 1$. Figure 4.1 is supplied as a visualization aid.

We remark that the entry of A_{2r} at position (i, j) is independent of r and denote it by $a(i, j)$. We set for $r \geq 1$

$$\begin{aligned} \Gamma_{2,2r} &:= \left\{\left(\epsilon_{a(i,1)}, \epsilon_{a(i,2)}, \dots, \epsilon_{a(i,2r)}\right) \mid i \in [2r]\right\} \subseteq \Omega(\pi_{2,2r}) \subseteq \left(\mathbb{R}^2\right)^{2r}, \\ \Gamma_{2,2r+1} &:= \left\{\left(\epsilon_{a(i,1)}, \epsilon_{a(i,2)}, \dots, \epsilon_{a(i,2r)}, \epsilon_{\chi(i)}\right) \mid i \in [2r]\right\} \subseteq \Omega(\pi_{2,2r+1}) \subseteq \left(\mathbb{R}^2\right)^{2r+1}, \end{aligned}$$

where $\chi \colon \mathbb{N} \to \{1, 2\}, \; i \mapsto i \mod 2$. That is, $\Gamma_{2,2r}$ is the subset of $\Omega(\pi_{2,2r})$ induced by the rows of A_{2r} and $\Gamma_{2,2r+1}$ is obtained by alternately appending ϵ_1 or ϵ_2 to the $2r$-many elements of $\Gamma_{2,2r}$.

Lemma 4.5.3 ([FR21, Lemma 2.3])**.** *For $r \geq 1$, it holds that $0 \notin \operatorname{aff}(\Gamma_{2,2r})$ and $0 \notin \operatorname{aff}(\Gamma_{2,2r+1})$.*

Proof. By construction, $0 \in \operatorname{aff}(\Gamma_{2,2r+1})$ implies $0 \in \operatorname{aff}(\Gamma_{2,2r})$, since one could choose the same coefficients for the affine linear combination. Hence, it suffices to prove $0 \notin \operatorname{aff}(\Gamma_{2,2r})$. We proceed by induction on $r \geq 1$. For $r = 1$, it is clear

$$A_4 = \left(\begin{array}{cc|cc} * & * & * & * \\ & * & & \\ \hline * & & & * \\ & & * & * \end{array}\right), \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}; \; A_6 = \left(\begin{array}{cc|cc|cc} * & * & * & * & * & * \\ & * & & & & \\ \hline * & & & * & * & * \\ & & * & * & & \\ \hline * & & * & & & * \\ & & & & * & * \end{array}\right), \begin{bmatrix} 1/4 \\ 1/4 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \end{bmatrix}$$

weights for A_8 : $\begin{bmatrix} 1/4 & 1/4 & 1/8 & 1/8 & 1/16 & 1/16 & 1/16 & 1/16 \end{bmatrix}^\mathsf{T}$

Figure 4.1: The positions of the ones in A_4 and A_6 are marked by $*$ and the cells are coloured according to whether they belong to A_2, B_1, B_2 or B_3. In square brackets, we indicated the weights of the convex combination from (4.9).

that $0 \notin \operatorname{aff}(\Gamma_{2,2}) \subseteq \mathbb{R}^2 \times \{\epsilon_1\}$. Now assume that $0 \notin \operatorname{aff}(\Gamma_{2,2r})$. For the sake of contradiction, let

$$\sum_{i=1}^{2r+2} \lambda_i \left(\epsilon_{a(i,1)}, \epsilon_{a(i,2)}, \ldots, \epsilon_{a(i,2r+2)}\right) = 0 \in \left(\mathbb{R}^2\right)^{2r+2} \tag{4.7}$$

be an affine linear combination of $\Gamma_{2,2r+2}$. Then Equation (4.7) gives for any $\mathbb{R}^2$-component, i.e., for any $k \in [2r+2]$, an affine linear combination of ϵ_1, ϵ_2:

$$\sum_{i=1}^{2r+2} \lambda_i \, \epsilon_{a_{i,k}} = 0 \overset{(*)}{=} \frac{1}{2}(\epsilon_1 + \epsilon_2), \tag{4.8}$$

where we used Lemma 4.4.1 in $(*)$. For the following take a look at (4.6) and Figure 4.1. Considering the scalar factor of ϵ_1 in (4.8) for $k \in \{1, 2r+1, 2r+2\}$, we conclude with the construction of A_{2r+2} in (4.6) that

$$\underbrace{\sum_{j=1}^{r+1} \lambda_{2j-1}}_{k=1} = \frac{1}{2} = \underbrace{\lambda_{2r+2} + \sum_{j=1}^{r} \lambda_{2j-1}}_{k=2r+1} = \frac{1}{2} = \underbrace{\lambda_{2r+2} + \sum_{j=1}^{r+1} \lambda_{2j-1}}_{k=2r+2}.$$

Hence, $\lambda_{2r+2} = 0$ combining the cases $k = 1$ and $k = 2r+2$. Furthermore, $k = 2r+1$ and $k = 2r+2$ give $\lambda_{2r+1} = 0$. Therefore, the first $2r$-many components in Equation (4.7) show $0 \in \operatorname{aff}(\Gamma_{2,2r})$, which contradicts our induction hypothesis. □

Lemma 4.5.4 ([FR21, Lemma 2.4])**.** *It holds that* $\operatorname{dist}(0, \operatorname{conv}(\Gamma_{2,2r})) \leq 2^{-r+\frac{1}{2}}$ *and* $\operatorname{dist}(0, \operatorname{conv}(\Gamma_{2,2r+1})) \leq 2^{-r+\frac{1}{2}}$.

Proof. First, we prove the inequality for $\operatorname{conv}(\Gamma_{2,2r})$. For $i \in [2r]$ let $\omega_i := \left(\epsilon_{a(i,1)}, \ldots, \epsilon_{a(i,2r)}\right) \in (\mathbb{R}^2)^{2r}$ be the weight in $\Gamma_{2,2r}$ that corresponds to the i^{th} row of A_{2r}. Consider the convex combination (see Figure 4.1 for an illustration)

$$(x_1, \ldots, x_{2r}) := 2^{-r}(\omega_{2r-1} + \omega_{2r}) + \sum_{l=1}^{r-1} 2^{-l-1}(\omega_{2l-1} + \omega_{2l}) \in \left(\mathbb{R}^2\right)^{2r}. \tag{4.9}$$

Note that $x_i \in \mathbb{R}^2$. We will argue that $(x_1, \ldots, x_{2r}) = 2^{-r+1}(0_2, \ldots, 0_2, \epsilon_1)$. Since x is a convex combination of the elements in $\Gamma_{2,2r}$, the statement then follows from $\|\epsilon_1\| = 2^{-\frac{1}{2}}$, compare (4.5).

We consider A_{2r} like in its construction (4.6) as $r \times r$ block matrix with block entries being 2×2 matrices. For $m \in [r]$ the two weights ω_{2m-1} and ω_{2m} correspond to the m^{th} block row of A_{2r} and have the same scalar factor in (4.9). Hence, whenever for $i \in [2r]$ the i^{th} column of the m^{th} block row of A_{2k} contains exactly one entry equal to one (and so the other entry equals two), then the contributions of ω_{2m-1} and ω_{2m} to x_i cancel due to $\epsilon_1 + \epsilon_2 = 0_2$. In particular, in (4.9) all contributions of block entries equal to B_1 cancel. Therefore the last column of A_{2r} gives

$$x_{2r} = 2^{-r}(\epsilon_1 + \epsilon_1) = 2^{-r+1}\epsilon_1.$$

Furthermore, $x_1 = x_3 = \ldots = x_{2r-1} = 0_2$ using that also the first columns of A_2, of B_2 and of B_3 contain exactly one entry equal to one. For $r = 1$ we are done. If $r \geq 2$, then reading off the second column of A_{2r}, we find

$$x_2 = \underbrace{2^{-2}(\epsilon_1 + \epsilon_1)}_{\text{first block row}} + \underbrace{2^{-r}(\epsilon_2 + \epsilon_2)}_{\text{last block row}} + \sum_{l=2}^{r-1} \underbrace{2^{-l-1}(\epsilon_2 + \epsilon_2)}_{\text{middle rows}} = 2^{-1}(\epsilon_1 + \epsilon_2) = 0_2.$$

Analogously, as B_1 does not contribute we compute for $j = 2, 3, \ldots, r-1$ that

$$x_{2j} = \underbrace{2^{-j-1}(\epsilon_1 + \epsilon_1)}_{j^{th}\text{ block row}} + \underbrace{2^{-r}(\epsilon_2 + \epsilon_2)}_{\text{last block row}} + \sum_{l=j+1}^{r-1} \underbrace{2^{-l-1}(\epsilon_2 + \epsilon_2)}_{\text{in between rows}} = 2^{-j}(\epsilon_1 + \epsilon_2) = 0_2,$$

because the second columns of B_2 and B_3 are, respectively, $(2, 2)^\mathsf{T}$ and $(1, 1)^\mathsf{T}$. This proves the inequality in the case $\Gamma_{2,2r}$.

By construction, for $\Gamma_{2,2r+1}$ the same convex combination works, because the last $\mathbb{R}^2$-component does not contribute as the entries of the weights alternate between ϵ_1 and ϵ_2. □

Noting that for odd $d = 2r + 1$ one has $-r + 1/2 = -(d/2) + 1$, Lemma 4.5.3 and Lemma 4.5.4 together yield the bound from Theorem 4.5.1(a) for the weight margin. It remains to show that the witness sets are free to deduce the same bound for the gap. We use the characterization of freeness from Proposition 4.3.3.

Proposition 4.5.5 ([FR21, Proposition 4.14])**.** *For $r \geq 2$, the rows of A_{2r} form a free subset of $[2]^{2r}$, i.e., $\Gamma_{2,2r}$ is free. Moreover, for $r \geq 1$ the set of weights $\Gamma_{2,2r+1}$ is free.*

Proof. First, note that $\Gamma_{2,3} = \{\epsilon_{1,1,1}, \epsilon_{2,1,2}\}$ is free. Now, let $r \geq 2$. If $\Gamma_{2,2r}$ is free, then $\Gamma_{2,2r+1}$ is also free by construction. Thus, we are left to prove the former.

Consider A_{2r} as defined in Equation (4.6). We must show that distinct rows of A_{2r} differ in at least two entries for all $r \geq 2$. The claim is proven by induction on $r \geq 3$. For $r = 3$, we verify the claim by inspection of A_6. Let a_i be the i^{th} row of A_6; its definition is recalled in the left-hand table below. The right-hand table lists for each pair a_i, a_j with $i < j$ two distinct entries in which a_i and a_j differ, which shows the claim for $r = 3$.

entry	1	2	3	4	5	6
a_1	1	1	1	1	1	1
a_2	2	1	2	2	2	2
a_3	1	2	2	1	1	1
a_4	2	2	1	1	2	2
a_5	1	2	1	2	2	1
a_6	2	2	2	2	1	1

	a_2	a_3	a_4	a_5	a_6
a_1	1,3	2,3	1, 2	2,4	1,2
a_2		1,2	2,3	1,2	5,6
a_3			1,3	3,4	1, 4
a_4				1,4	3,4
a_5					1,3

In fact, the table also proves the claim for $r = 2$, since $a_1, \ldots, a_4$ already pairwise differ in at least two of the first four entries.

Now assume that the claim holds for some fixed $r \geq 3$. Let a_i, a_j be distinct rows of A_{2r+2}; we will show they differ in at least two entries. If $1 \leq i < j \leq 2r$, then by our inductive hypothesis there is nothing to prove because the first $2r$ rows of A_{2r+2} contain A_{2r} as a submatrix.

To complete the proof, it is enough to show that the $4 \times (2r + 2)$ submatrix formed by restricting to the k^{th} block row, $k \in [r]$, and the last block row of A_{2r+2} satisfies the hypothesis, i.e., any two distinct rows of this submatrix differ in at least two entries. This is the case as restricting to its first, k^{th} and last block columns yields a 4×6 submatrix of A_6 if $k \geq 2$, namely

$$\begin{pmatrix} B_2 & B_3 & B_1 \\ B_2 & B_2 & B_3 \end{pmatrix},$$

and a 4×4 submatrix equal to A_4 if $k = 1$. □

4.5.2 Tensors of order three

In this subsection we show part (b) of Theorem 4.5.1, i.e., that $\gamma_T(\pi_{m,3})$ and $\gamma_G(\pi_{m,3})$ are exponentially small in m. To do so, we set

$$\mathscr{W}_{m,3} := \bigcup_{s=2}^{m} \{(s, 1, s), (s, s, 1), (s-1, s, s)\} \subseteq [m] \times [m] \times [m] \tag{4.10}$$

and consider the corresponding subset

$$\Gamma_{m,3} := \Gamma_{\mathscr{W}_{m,3}} = \big\{(\epsilon_i, \epsilon_j, \epsilon_k) \mid (i, j, k) \in \mathscr{W}_{m,3}\big\} \subseteq \Omega(\pi_{m,3}). \tag{4.11}$$

Let us first show that $0 \notin \operatorname{conv}(\Gamma_{m,3})$ by proving the following statement.

Lemma 4.5.6 ([FR21, Lemma 2.8])**.** *It holds that* $0 \notin \operatorname{aff}(\Gamma_{m,3})$.

Proof. For a proof by contradiction we assume $0 \in \operatorname{aff}(\Gamma_{m,3})$. Then there exist $a_s, b_s, c_s \in \mathbb{R}$ for $s = 2, 3, \ldots, m$ such that $\sum_s (a_s + b_s + c_s) = 1$ and

$$\sum_{s=2}^{m} \big(a_s(\epsilon_s, \epsilon_1, \epsilon_s) + b_s(\epsilon_s, \epsilon_s, \epsilon_1) + c_s(\epsilon_{s-1}, \epsilon_s, \epsilon_s) \big) = (0_m, 0_m, 0_m) \in (\mathbb{R}^m)^3.$$

In each of the three $\mathbb{R}^m$-components we obtain 0_m as an affine linear combination of $\epsilon_1, \dots, \epsilon_m$. Applying Lemma 4.4.1 to the coefficient of ϵ_{s-1} in the first component, respectively to the coefficient of ϵ_s in the second and third component yields

$$a_{s-1} + b_{s-1} + c_s = m^{-1} \quad \text{for } s = 2, 3, \dots, m \tag{4.12}$$

$$\text{respectively} \qquad b_s + c_s = a_s + c_s = m^{-1} \quad \text{for } s = 2, 3, \dots, m \tag{4.13}$$

where we necessarily set $a_1 = b_1 := 0$. Equation (4.12) for $s = 2$ is $c_2 = m^{-1}$ and hence $a_2 = b_2 = 0$ by (4.13) for $s = 2$. But now (4.12) for $s = 3$ gives $c_3 = m^{-1}$ and we can proceed inductively to conclude $c_s = m^{-1}$ and $a_s = b_s = 0$ for all $s = 2, 3, \dots, m$. This gives the contradiction $1 = \sum_{s=2}^{m}(a_s + b_s + c_s) = \frac{m-1}{m}$, so we must have $0 \notin \operatorname{aff}(\Gamma_{m,3})$. Another contradiction arises by applying Lemma 4.4.1 to the coefficient of ϵ_m in the first component, which yields $a_m + b_m = m^{-1}$. □

Next, we prove an exponentially small upper bound on $\operatorname{dist}(0, \operatorname{conv}(\Gamma_{m,3}))$. The key combinatorial idea, which is presented in the following lemma, is due to [Kra07, Theorem 1 with $k = 0$].[18] According to [Kra07] the special case $k = 0$ is already contained in [KL05, Theorem 9].

Lemma 4.5.7 ([FR21, Lemma 2.5])**.** *Let $m \geq 3$ and set $\lambda_{i,j,k} := 0$ for all $(i,j,k) \in [m]^3 \setminus \big(\mathscr{W}_{m,3} \cup \{(1,1,1)\}\big)$. Moreover, define*

$$\lambda_{1,1,1} := 2^{-m+1}, \quad \lambda_{1,2,2} := 1 - 2^{-m+1}, \quad \lambda_{m,1,m} = \lambda_{m,m,1} := 2^{-1}$$

and for $s = 2, 3, \dots, m-1$

$$\lambda_{s,1,s} = \lambda_{s,s,1} := 2^{-m+s-1}, \quad \lambda_{s,s+1,s+1} := 1 - 2^{-m+s}\,.$$

Then the following equations hold:

$$\big(\forall i \in [m]\colon \lambda_{i,+,+} = 1\big),\ \big(\forall j \in [m]\colon \lambda_{+,j,+} = 1\big),\ \big(\forall k \in [m]\colon \lambda_{+,+,k} = 1\big)\,. \tag{4.14}$$

In particular, $\lambda_{+,+,+} = \sum_{i,j,k} \lambda_{i,j,k} = m$.

Proof. This is [Kra07, Theorem 1 with $k = 0$]. Alternatively, the statement can be checked by straightforward computation as follows.

For $i = 1$, we have $\lambda_{1,1,1} + \lambda_{1,2,2} = 1$ and for $i = m$, $\lambda_{m,1,m} + \lambda_{m,m,1} = 1$. If $i \in \{2, 3, \dots, m-1\}$, then

$$\lambda_{i,+,+} = \lambda_{i,1,i} + \lambda_{i,i,1} + \lambda_{i,i+1,i+1} = 2 \cdot 2^{-m+i-1} + 1 - 2^{-m+i} = 1\,.$$

[18] In [Kra07] Kravtsov extensively studies so-called complete r-noninteger vertices (r-CNVs) of the three-index axial assignment polytope. For $k \in \{0, 1, \dots, m-2\}$, [Kra07, Theorem 1] states explicitly a $(3m-2-k)$-CNV, among these we use the $(3m-2)$-CNV (i.e., $k = 0$). Moreover, [Kra07, Theorem 2] states that such r-CNVs of the three-index axial assignment polytope actually only occur for $r \in \{2m, 2m+1, \dots, 3m-2\}$, and the later theorems in [Kra07] fully characterize the r-CNVs and study their combinatorial properties.

For the cases $j = 2$, $j \in \{3, 4, \dots, m-1\}$ and $j = m$ we compute, respectively,

$$\begin{aligned}
\lambda_{+,2,+} &= \lambda_{1,2,2} + \lambda_{2,2,1} = 1 - 2^{-m+1} + 2^{-m+2-1} = 1 \\
\lambda_{+,j,+} &= \lambda_{j,j,1} + \lambda_{j-1,j,j} = 2^{-m+j-1} + 1 - 2^{-m+(j-1)} = 1 \\
\lambda_{+,m,+} &= \lambda_{m,m,1} + \lambda_{m-1,m,m} = 2^{-1} + 1 - 2^{-m+(m-1)} = 1\,.
\end{aligned}$$

Finally, for $j = 1$ we get

$$\lambda_{1,1,1} + \left(\sum_{s=2}^{m-1} \lambda_{s,1,s}\right) + \lambda_{m,1,m} = 2^{-m+1} + \left(2^{-m+1} + \ldots + 2^{-2}\right) + 2^{-1} = 1\,.$$

Note that by definition $\lambda_{i,j,k} = \lambda_{i,k,j}$ for all $i, j, k \in [m]$. This ends the proof. □

Example 4.5.8 ([FR21, Example 2.6])**.** To visualize the idea of Lemma 4.5.7 it is helpful to consider the slices Λ_i given by $(\Lambda_i)_{j,k} = \lambda_{i,j,k}$. For $m = 4$ one has

$$\Lambda_1 = \frac{1}{8}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \Lambda_2 = \frac{1}{8}\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\Lambda_3 = \frac{1}{8}\begin{pmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}, \quad \Lambda_4 = \frac{1}{8}\begin{pmatrix} 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

and for $m = 5$ one has

$$\Lambda_1 = \frac{1}{16}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; \Lambda_2 = \frac{1}{16}\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 14 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\Lambda_3 = \frac{1}{16}\begin{pmatrix} 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; \Lambda_4 = \frac{1}{16}\begin{pmatrix} 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{pmatrix}, \; \Lambda_5 = \frac{1}{16}\begin{pmatrix} 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Indeed, we can see that summing over all entries of some Λ_i gives one. Moreover, summing over the entries of the j^{th} row (respectively k^{th} column) of all Λ_i again yields one. ♢

Lemma 4.5.9 ([FR21, Lemma 2.7])**.** *For $m \geq 3$, $\operatorname{dist}\left(0, \operatorname{conv}(\Gamma_{m,3})\right) \leq 2^{-m+1}$.*

Proof. Define $\lambda_{i,j,k} \geq 0$ for all $i, j, k \in [m]$ as in Lemma 4.5.7, which we can apply as $m \geq 3$. Since $\sum_{i=1}^{m} \epsilon_i = 0$ (compare Equation (4.4)), Lemma 4.5.7 yields

$$\begin{aligned}
&\sum_{i,j,k} \lambda_{i,j,k}(\epsilon_i, \epsilon_j, \epsilon_k) = \sum_{i,j,k} \lambda_{i,j,k}\big((\epsilon_i, 0_m, 0_m) + (0_m, \epsilon_j, 0_m) + (0_m, 0_m, \epsilon_k)\big) \\
&= \sum_{i} \lambda_{i,+,+}(\epsilon_i, 0_m, 0_m) + \sum_{j} \lambda_{+,j,+}(0_m, \epsilon_j, 0_m) + \sum_{k} \lambda_{+,+,k}(0_m, 0_m, \epsilon_k) = 0_{3m}.
\end{aligned}$$

Equivalently, we have

$$-2^{-m+1}(\epsilon_1,\epsilon_1,\epsilon_1) = \sum_{(i,j,k)\in\mathscr{W}_{m,3}} \lambda_{i,j,k}(\epsilon_i,\epsilon_j,\epsilon_k).$$

Normalizing the latter equation we obtain

$$x := -c^{-1}\,2^{-m+1}(\epsilon_1,\epsilon_1,\epsilon_1) \in \operatorname{conv}(\Gamma_{m,3}), \quad \text{where } c := \sum_{(i,j,k)\in\mathscr{W}_{m,3}} \lambda_{i,j,k}\,.$$

Finally, $\|(\epsilon_1,\epsilon_1,\epsilon_1)\| \leq \sqrt{3}$, see (4.5), and $c = m - 2^{-m+1} \geq \sqrt{3}$ together imply that $\|x\| \leq 2^{-m+1}$. □

Combining Lemma 4.5.6 and Lemma 4.5.9 shows $\gamma_T(\pi_{m,3}) \leq 2^{-m+1}$. To conclude the same bound for the gap $\gamma_G(\pi_{m,3})$ it remains to show that $\Gamma_{m,3}$ is free. For this, we use the characterization of freeness from Proposition 4.3.3.

Proposition 4.5.10 ([FR21, first part of Proposition 4.15])**.** *For $m \geq 3$ the set $\mathscr{W}_{m,3} \subseteq [m]^3$ is free, i.e., $\Gamma_{m,3} \subseteq \Omega(\pi_{m,3})$ is free.*

Proof. Recall from (4.10) that

$$\mathscr{W}_{m,3} = \big\{(s,1,s),(s,s,1),(s-1,s,s) \mid s = 2,3,\dots,m\big\}.$$

Let $x = (x_1,x_2,x_3), y = (y_1,y_2,y_3) \in \mathscr{W}_{m,3}$ be such that $x \neq y$. We prove by a distinction of cases that x and y differ in at least two entries.

First, we assume $x_1 = y_1$. Then $a := x_1 = y_1 \geq 2$, otherwise $x = (1,2,2) = y$ contradicts $x \neq y$. Thus $x, y \in \{(a,1,a),(a,a,1),(a,a+1,a+1)\}$ and we conclude that x and y differ in the second and third entry as $a \geq 2$.

Second, we assume $x_1 \neq y_1$. There is nothing to show if $x_2 \neq y_2$, so we additionally assume $b := x_2 = y_2$. If $b = 1$, then we are done by $x = (x_1,1,x_1)$ and $y = (y_1,1,y_1)$. On the other hand, $b \geq 2$ yields $x, y \in \{(b,b,1),(b-1,b,b)\}$ and as $x \neq y$, they differ in the first and third entry. □

4.5.3 Tensors of higher order

In this subsection we part (c) of Theorem 4.5.1 by recycling the combinatorial idea of Lemma 4.5.7. Let us give some intuition for our construction. The main idea is to use the construction from the previous subsection for some multiple of m, i.e., considering $\mathscr{W}_{rm,3}$ for $r \geq 2$:

$$\mathscr{W}_{rm,3} := \bigcup_{s=2}^{rm} \{(s,1,s),(s,s,1),(s-1,s,s)\} \subseteq [rm]\times[rm]\times[rm] \tag{4.15}$$

compare (4.10). Thereby, the main challenge is to ensure that the constructed subset of $\Omega(\pi_{m,d})$ does not contain zero in its convex hull. We can try to extend the elements of $\Omega(\pi_{m,3})$ to elements of $\Omega(\pi_{m,d})$. One natural idea is duplicate each component $d/3$ times, i.e., when $d = 6$ the vector $(\epsilon_i,\epsilon_j,\epsilon_k) \in \Omega(\pi_{m,3})$ becomes $(\epsilon_i,\epsilon_i,\epsilon_j,\epsilon_j,\epsilon_k,\epsilon_k) \in \Omega(\pi_{m,6})$. However, we need a subset of $\Omega(\pi_{m,d})$ with rm

many elements to imitate the construction from the previous subsection. We still extend the elements of $\Omega(\pi_{m,3})$ in this way, but will additionally "shift" and "twist" by some functions $\sigma_1, \dots, \sigma_{2r-1} \colon [rm] \to [m]$, so that the elements of our set will look like

$$\Big(\epsilon_{\sigma_1(i)}, \dots, \epsilon_{\sigma_{d/3}(i)}, \epsilon_{\sigma_1(j)}, \dots, \epsilon_{\sigma_{d/3}(j)}, \epsilon_{\sigma_1(k)}, \dots, \epsilon_{\sigma_{d/3}(k)}\Big)$$

for $d/3 = 2r-1$ and $(i,j,k) \in \mathscr{W}_{rm,3}$. We now define the functions σ_k. For this, let $m \geq 3$ and fix a natural number $r \geq 2$. It is convenient to use an *adjusted* modulo m function $\operatorname{mod}' \ m$ that takes values in $[m]$, i.e., instead of zero it outputs m. For $i \in [r]$ we consider

$$\begin{aligned} \sigma_i &\colon [rm] \to [m], \quad j \mapsto \left\lceil \frac{j + (i-1)}{r} \right\rceil \mod' \ m \\ \sigma_{r+i} &:= \sigma_1 \circ (r-i+1 \quad r+1) \colon [rm] \to [m] \end{aligned}$$

where $(r-i+1 \quad r+1)$ denotes the corresponding transposition in the symmetric group of $[rm]$.[19] We only need the first $2r-1$ of these functions and combine them to obtain

$$\sigma \colon [rm] \to [m]^{2r-1}, \quad j \mapsto \big(\sigma_1(j), \sigma_2(j), \dots, \sigma_{2r-1}(j)\big).$$

Example 4.5.11 ([FR21, Example 2.9])**.** For $r = 3$ the functions $\sigma_1, \sigma_2, \dots, \sigma_6$ are sketched by the following table.

j	1	2	3	4	5	6	$\cdots$	$3m-4$	$3m-3$	$3m-2$	$3m-1$	$3m$
σ_1	1	1	1	2	2	2	$\cdots$	$m-1$	$m-1$	m	m	m
σ_2	1	1	2	2	2	3	$\cdots$	$m-1$	m	m	m	1
σ_3	1	2	2	2	3	3	$\cdots$	m	m	m	1	1
σ_4	1	1	2	1	2	2	$\cdots$	$m-1$	$m-1$	m	m	m
σ_5	1	2	1	1	2	2	$\cdots$	$m-1$	$m-1$	m	m	m
σ_6	2	1	1	1	2	2	$\cdots$	$m-1$	$m-1$	m	m	m

For $r = 3$ and $m = 5$ the functions $\sigma_1, \sigma_2, \dots, \sigma_6$ are given by the following table.

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
σ_1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5
σ_2	1	1	2	2	2	3	3	3	4	4	4	5	5	5	1
σ_3	1	2	2	2	3	3	3	4	4	4	5	5	5	1	1
σ_4	1	1	2	1	2	2	3	3	3	4	4	4	5	5	5
σ_5	1	2	1	1	2	2	3	3	3	4	4	4	5	5	5
σ_6	2	1	1	1	2	2	3	3	3	4	4	4	5	5	5

Remark 4.5.12 ([FR21, Remark 2.10]). By construction, each element of $[m]$ is attained exactly r-times by σ_k, $k \in [2r-1]$. Moreover, the definition of $\sigma_1, \dots, \sigma_r$ yields that σ is injective. ▽

[19] We stress that we always take σ_1 (and *not* σ_i) to define σ_{r+i}.

For $i, j, k \in [rm]$ we introduce the short-hand

$$\begin{aligned}\epsilon_{\sigma(i)} &:= \big(\epsilon_{\sigma_1(i)}, \epsilon_{\sigma_2(i)}, \dots, \epsilon_{\sigma_{2r-1}(i)}\big) \in (\mathbb{R}^m)^{2r-1} \\ \epsilon_{\sigma(i),\sigma(j),\sigma(k)} &:= \big(\epsilon_{\sigma(i)}, \epsilon_{\sigma(j)}, \epsilon_{\sigma(k)}\big) \in (\mathbb{R}^m)^{6r-3} \ .\end{aligned}$$

We set

$$\mathscr{J}_r := \big\{(s,1,s), (s,s,1) \mid s = 2,3,\dots,r\big\} \subseteq \mathbb{Z}^3 \,,$$

which we remove from $\mathscr{W}_{rm,3}$, see (4.15), to obtain.

$$\mathscr{W}_{rm,3}\backslash\mathscr{J}_r = \left(\bigcup_{s=2}^{r}\{(s-1,s,s)\}\right) \cup \bigcup_{s=r+1}^{rm}\{(s,1,s),(s,s,1),(s-1,s,s)\} \tag{4.16}$$

In the following we show that the convex hull of the corresponding set of weights

$$\Gamma_{m,6r-3} := \big\{\epsilon_{\sigma(i),\sigma(j),\sigma(k)} \mid (i,j,k) \in \mathscr{W}_{rm,3}\backslash\mathscr{J}_r\big\} \subseteq \Omega(\pi_{m,6r-3}) \subseteq \big(\mathbb{R}^m\big)^{6r-3}$$

does not contain the zero vector, but is very close to it.[20] But first, we prove freeness of $\Gamma_{m,6r-3}$, which is a direct consequence of its construction.

Proposition 4.5.13 ([FR21, second part of Proposition 4.15])**.** *For $m \geq 3$ and $r \geq 2$ the set of weights $\Gamma_{m,6r-3} \subseteq \Omega(\pi_{m,6r-3})$ is free.*

Proof. We use the characterization of freeness from Proposition 4.3.3. The above definition of $\Gamma_{m,6r-3}$ shows that it equals $\Gamma_{\mathscr{W}_{m,6r-3}}$, where

$$\mathscr{W}_{m,6r-3} := \big\{\big(\sigma(i),\sigma(j),\sigma(k)\big) \mid (i,j,k) \in \mathscr{W}_{rm,3}\backslash\mathscr{J}_r\big\} \subseteq [m]^{6r-3} \,.$$

By Proposition 4.5.10, $\mathscr{W}_{rm,3}$ is free and so is its subset $\mathscr{W}_{rm,3}\backslash\mathscr{J}_r$. Now, consider $(i,j,k),(i',j',k') \in \mathscr{W}_{rm,3}\backslash\mathscr{J}_r$ such that $(\sigma(i),\sigma(j),\sigma(k)) \neq (\sigma(i'),\sigma(j'),\sigma(k'))$. We necessarily have $(i,j,k) \neq (i',j',k')$, hence they differ in at least two entries as $\mathscr{W}_{rm,3}\backslash\mathscr{J}_r$ is free. Since σ is injective, also $(\sigma(i),\sigma(j),\sigma(k))$ and $(\sigma(i'),\sigma(j'),\sigma(k'))$ differ in at least two entries. Therefore, $\mathscr{W}_{m,6r-3}$ is free and so is $\Gamma_{m,6r-3}$. □

Thus, $\Gamma_{m,6r-3}$ may also serve as a witness set for upper bounding the gap, by Proposition 4.3.7(ii). However, we need to ensure $0 \notin \operatorname{conv}(\Gamma_{m,6r-3})$, which indeed holds due to the following.

Lemma 4.5.14 ([FR21, Lemma 2.11])**.** *For $m \geq 3$ and $r \geq 2$ it holds that $0 \notin \operatorname{aff}(\Gamma_{m,6r-3})$.*

We defer the proof to the end of this subsection, as it is very technical. Instead, we first give a lower bound on the distance from zero to the convex hull of $\Gamma_{m,6r-3}$.

[20] One could suggest to consider the set $\{\epsilon_{\sigma(i),\sigma(j),\sigma(k)} \mid (i,j,k) \in \mathscr{W}_{rm,3}\}$, but this won't ensure that zero is not in the convex hull. The intuition behind is, that $\Gamma_{m,3}$ from the last subsection is "nearly at the limit", i.e., $0 \notin \operatorname{conv}(\Gamma_{m,3})$ but $0 \in \operatorname{conv}(\Gamma_{m,3} \cup \{(\epsilon_1,\epsilon_1,\epsilon_1)\})$. Now the function σ "introduces $2r-2$ additional linear relations", since $\epsilon_{\sigma(i)} \in (\mathbb{1}_m^\perp)^{2r-1}$ and the orthogonal complement $\mathbb{1}_m^\perp \subseteq \mathbb{R}^m$ has codimension one while $(\mathbb{1}_m^\perp)^{2r-1} \subseteq (\mathbb{R}^m)^{2r-1}$ has codimension $2r-1$. Thus, it is plausible to remove $2r-2$ many elements from $\mathscr{W}_{rm,3}$.

Lemma 4.5.15 ([FR21, Lemma 2.12]). *Let $m \geq 3$ and $r \geq 2$. Then*

$$\operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,6r-3})\big) \leq \frac{\sqrt{6}}{(m-1)\sqrt{r}}\, 2^{-r(m-1)+1} \leq 2^{-r(m-1)+1}.$$

Proof. We set $N := rm$ and for $i,j,k \in [N]$ we define $\lambda_{i,j,k}$ as in Lemma 4.5.7 applied for the dimension N. Then Equation (4.14) of Lemma 4.5.7 yields

$$\begin{aligned}
&\sum_{i,j,k=1}^{N} \lambda_{i,j,k}\, \big(\epsilon_{\sigma(i)}, \epsilon_{\sigma(j)}, \epsilon_{\sigma(k)}\big) \\
&= \sum_{i,j,k=1}^{N} \lambda_{i,j,k}\, \big(\epsilon_{\sigma(i)}, 0, 0\big) + \sum_{i,j,k=1}^{N} \lambda_{i,j,k}\, \big(0, \epsilon_{\sigma(j)}, 0\big) + \sum_{i,j,k=1}^{N} \lambda_{i,j,k}\, \big(0, 0, \epsilon_{\sigma(k)}\big) \\
&= \sum_{i=1}^{N} \big(\epsilon_{\sigma(i)}, 0, 0\big) + \sum_{j=1}^{N} \big(0, \epsilon_{\sigma(j)}, 0\big) + \sum_{k=1}^{N} \big(0, 0, \epsilon_{\sigma(k)}\big) = \sum_{i=1}^{N} \epsilon_{\sigma(i),\sigma(i),\sigma(i)} = 0,
\end{aligned}$$

where we used in the last step Equation (4.4) and Remark 4.5.12, i.e., that each element of $[m]$ is attained exactly r-many times by all $\sigma_k \colon [rm] \to [m]$, $k \in [2r-1]$. Because $\mathscr{W}_{N,3}$ contains the support of λ apart from the element $(1,1,1)$, we have

$$\begin{aligned}
x &:= -\lambda_{1,1,1}\, \epsilon_{\sigma(1),\sigma(1),\sigma(1)} - \sum_{(i,j,k)\in \mathscr{J}_r} \lambda_{i,j,k}\, \epsilon_{\sigma(i),\sigma(j),\sigma(k)} \\
&= \sum_{(i,j,k)\in \mathscr{W}_{N,3}\setminus \mathscr{J}_r} \lambda_{i,j,k}\, \epsilon_{\sigma(i),\sigma(j),\sigma(k)}.
\end{aligned} \tag{4.17}$$

We see that the positive cone of $\Gamma_{m,6r-3} = \{\epsilon_{\sigma(i),\sigma(j),\sigma(k)} \mid (i,j,k) \in \mathscr{W}_{N,3} \setminus \mathscr{J}_r\}$ contains x. Normalizing the latter equation with

$$c := \sum_{(i,j,k)\in \mathscr{W}_{N,3}\setminus \mathscr{J}_r} \lambda_{i,j,k} = \sum_{i,j,k=1}^{N} \lambda_{i,j,k} - \left(\lambda_{1,1,1} + \sum_{(i,j,k)\in \mathscr{J}_r} \lambda_{i,j,k}\right) \geq N - 1$$

shows $c^{-1}x \in \operatorname{conv}(\Gamma_{m,6r-3})$. To bound the norm of $c^{-1}x$ we compute

$$\begin{aligned}
\lambda_{1,1,1} + \sum_{(i,j,k)\in \mathscr{J}_r} \lambda_{i,j,k} &= 2^{-N+1} + \sum_{s=2}^{r} \left(\lambda_{s,1,s} + \lambda_{s,s,1}\right) \\
&= 2^{-N+1} + \sum_{s=2}^{r} \left(2 \cdot 2^{-N+s-1}\right) = \sum_{s=1}^{r} 2^{-N+s} < 2^{-N+r+1}.
\end{aligned}$$

Finally, using $\|\epsilon_{\sigma(i),\sigma(j),\sigma(k)}\| \leq \sqrt{6r-3}$ (see Equation (4.5)) together with the triangle inequality on Equation (4.17) implies

$$\|c^{-1}x\| \leq \frac{\sqrt{6r-3}}{N-1}\, 2^{-N+r+1} \leq \frac{\sqrt{6}}{(m-1)\sqrt{r}}\, 2^{-N+r+1} \leq 2^{-N+r+1} = 2^{-r(m-1)+1},$$

where we used $m \geq 3$ and $r \geq 2$ for $\sqrt{6} \leq (m-1)\sqrt{r}$. □

From Proposition 4.5.13, Lemma 4.5.14 and Lemma 4.5.15 we can deduce Theorem 4.5.1(c). Still, we are left to show Lemma 4.5.14. First, we present a proof for the special case $r = 3$, in which all main ideas of the general proof become apparent and visible. The proof for the general statement is given afterwards and certainly looks technical at a first encounter. Therefore, it is recommended to read the proof for $r = 3$ first. Afterwards, while reading the general proof it may be helpful to compare it in parallel with the proof of the special case.

Proof of Lemma 4.5.14 for $r = 3$. Recall the construction of $\Gamma_{m,15}$ via the set $\mathscr{W}_{3m,3} \setminus \mathscr{J}_3$, see (4.16) and below. Assume $0 \in \operatorname{aff}(\Gamma_{m,15})$ for a proof by contradiction. Then there are coefficients $a_s, b_s, c_s \in \mathbb{R}$, where $2 \leq s \leq 3m$, such that $a_2 = a_3 = b_2 = b_3 = 0$ (due to removing $\mathscr{J}_3$ from $\mathscr{W}_{3m,3}$), $\sum_s (a_s + b_s + c_s) = 1$ and

$$\sum_{s=2}^{3m} \left(a_s \, \epsilon_{\sigma(s),\sigma(1),\sigma(s)} + b_s \, \epsilon_{\sigma(s),\sigma(s),\sigma(1)} + c_s \, \epsilon_{\sigma(s-1),\sigma(s),\sigma(s)} \right) = 0 \in (\mathbb{R}^m)^{15}. \tag{4.18}$$

The bulk of our work will consist of proving the equations

$$b_2 + c_2 = b_3 + c_3 = \ldots = b_{3m} + c_{3m} \tag{4.19}$$

$$a_2 + c_2 = a_3 + c_3 = \ldots = a_{3m} + c_{3m}. \tag{4.20}$$

From here we will derive a contradiction. We now set about proving (4.19) and (4.20). Rewrite the left-hand-side of (4.18) as the collection for $k \in [5]$ of the following affine linear combinations of $\epsilon_1, \ldots, \epsilon_m$ in $\mathbb{R}^m$:

$$\sum_{s=2}^{3m} \left(a_s \, \epsilon_{\sigma_k(s)} + b_s \, \epsilon_{\sigma_k(s)} + c_s \, \epsilon_{\sigma_k(s-1)} \right) = 0 \tag{4.21}$$

$$\sum_{s=2}^{3m} \left(a_s \, \epsilon_{\sigma_k(1)} + b_s \, \epsilon_{\sigma_k(s)} + c_s \, \epsilon_{\sigma_k(s)} \right) = 0 \tag{4.22}$$

$$\sum_{s=2}^{3m} \left(a_s \, \epsilon_{\sigma_k(s)} + b_s \, \epsilon_{\sigma_k(1)} + c_s \, \epsilon_{\sigma_k(s)} \right) = 0. \tag{4.23}$$

If we expand each expression as an affine linear combination of the ϵ_l, then by Lemma 4.4.1 the coefficient of ϵ_l must be m^{-1} for all $l \in [m]$. Translating this for Equation (4.21) with $k = 2$, $l = 2, \ldots, m$ and using Example 4.5.11 we obtain

$$(a_{p-3} + a_{p-2} + a_{p-1}) + (b_{p-3} + b_{p-2} + b_{p-1}) + (c_{p-2} + c_{p-1} + c_p) = \frac{1}{m} \tag{4.24}$$

for $p = 6, 9, 12, \ldots, 3m$ (e.g., $l = 2$ yields (4.24) with $p = 6$). A similar calculation for $k = 1, 3$ and $l = 2, \ldots, m$ shows (4.24) holds for all $5 \leq p \leq 3m + 1$, where we set $c_{3m+1} := 0$.

Similarly for (4.22) with $l = 2, \ldots, m$ and $k = 1, 2, 3$ we obtain for $4 \leq p \leq 3m$ that

$$(b_{p-2} + c_{p-2}) + (b_{p-1} + c_{p-1}) + (b_p + c_p) = \frac{1}{m} \tag{4.25}$$

and the same equations with "b" replaced by "a" when considering (4.23).

In the following we prove (4.19). Subtracting (4.25) from itself with values of p differing by one, we deduce that

$$\begin{aligned} & b_2 + c_2 = b_5 + c_5 = \ldots = b_{3m-1} + c_{3m-1} \\ & b_3 + c_3 = b_6 + c_6 = \ldots = b_{3m} + c_{3m}, \\ \text{and} \qquad & b_4 + c_4 = b_7 + c_7 = \ldots = b_{3m-2} + c_{3m-2}. \end{aligned}$$

Next we deduce (4.19) by showing $b_2 + c_2 = b_3 + c_3 = b_4 + c_4$.

To do so, we apply Lemma 4.4.1 to (4.22) for the coefficient of ϵ_2 using Example 4.5.11, which yields for $k = 4, 5$ the equations

$$(b_3 + c_3) + (b_5 + c_5) + (b_6 + c_6) = \frac{1}{m} \tag{4.26}$$

$$(b_2 + c_2) + (b_5 + c_5) + (b_6 + c_6) = \frac{1}{m} \tag{4.27}$$

respectively. Subtracting the two shows $b_2 + c_2 = b_3 + c_3$, and we have $b_3 + c_3 = b_4 + c_4$ via subtracting (4.26) from (4.25) for $p = 6$. This completes the proof of (4.19); using (4.23) we similarly deduce (4.20).

To get a contradiction we show that $a_s = b_s = c_s = 0$ for all $s = 2, 3, \ldots, 3m$. For this, we set $a := \sum_s a_s$ and $b := \sum_s b_s$, and recall that $a_2 = a_3 = b_2 = b_3 = 0$. This time we use Lemma 4.4.1 applied to the coefficient of ϵ_1 in (4.21), in (4.22) and in (4.23) respectively for $k = 1$ to get

$$c_2 + c_3 + c_4 = \frac{1}{m}, \qquad a + c_2 + c_3 = \frac{1}{m} \qquad \text{and} \qquad b + c_2 + c_3 = \frac{1}{m} \tag{4.28}$$

respectively. We deduce from these three equations that $a = b = c_4$. Furthermore, $b_2 = b_3 = 0$ shows that (4.25) for $p = 4$ is $b_4 + (c_2 + c_3 + c_4) = m^{-1}$. Subtracting from the latter the left-hand equation in (4.28) yields $b_4 = 0$. Similarly, $a_4 = 0$ follows from $a_2 = a_3 = 0$ and the analogous equation of (4.25) with a's replaced by b's.

Now, (4.24) for $p = 5$ simplifies to $c_3 + c_4 + c_5 = m^{-1}$. Thus, $c_2 = c_5$ with (4.28) and therefore $a_5 = b_5 = 0$ by (4.19), (4.20) and $a_2 = b_2 = 0$. This simplifies (4.24) for $p = 6$ to $c_4 + c_5 + c_6 = m^{-1}$. Hence, $c_3 = c_6$ as we also have $c_3 + c_4 + c_5 = m^{-1}$ and we get via (4.19) and (4.20) that $a_6 = b_6 = 0$. The latter in turn shows that (4.24) for $p = 7$ becomes $c_5 + c_6 + c_7 = m^{-1}$, so $c_4 = c_7$ and $a_7 = b_7 = 0$ by, again, (4.19) and (4.20).

It should have become apparent that we can proceed inductively in the same manner with (4.24) for $p = 5, \ldots, 3m + 1$; thereby using (4.19) and (4.20) to deduce $a_s = b_s = 0$ for all $s = 2, 3, \ldots, 3m$. In particular, $a = b = c_4 = 0$. Finally, (4.19) implies $c_4 = c_s$ for all $s = 2, 3, \ldots, 3m$, which gives the desired contradiction. □

Proof of Lemma 4.5.14 for arbitrary r. Recall the construction of $\Gamma_{m,6r-3}$ via the set $\mathscr{W}_{rm,3} \setminus \mathscr{J}_r$, see (4.16) and below. For the sake of contradiction assume that

$0 \in \mathrm{aff}(\Gamma_{m,6r-3})$. Then there are coefficients $a_s, b_s, c_s \in \mathbb{R}$, where $2 \leq s \leq rm$, such that $a_2 = \ldots = a_r = b_2 = \ldots = b_r = 0$, $\sum_s (a_s + b_s + c_s) = 1$ and

$$\sum_{s=2}^{rm} \left(a_s \, \epsilon_{\sigma(s),\sigma(1),\sigma(s)} + b_s \, \epsilon_{\sigma(s),\sigma(s),\sigma(1)} + c_s \, \epsilon_{\sigma(s-1),\sigma(s),\sigma(s)} \right) = 0 \in (\mathbb{R}^m)^{6r-3}. \quad (4.29)$$

The bulk of our work will consist of proving the equations

$$b_2 + c_2 = b_3 + c_3 = \ldots = b_{rm} + c_{rm} \quad (4.30)$$
$$a_2 + c_2 = a_3 + c_3 = \ldots = a_{rm} + c_{rm}. \quad (4.31)$$

From here we will derive a contradiction. We now set about proving (4.31) and (4.30). Rewrite the left-hand-side of (4.29) as the collection for $k \in [2r-1]$ of the following affine linear combinations of $\epsilon_1, \ldots, \epsilon_m$ in $\mathbb{R}^m$:

$$\sum_{s=2}^{rm} \left(a_s \, \epsilon_{\sigma_k(s)} + b_s \, \epsilon_{\sigma_k(s)} + c_s \, \epsilon_{\sigma_k(s-1)} \right) = 0 \quad (4.32)$$
$$\sum_{s=2}^{rm} \left(a_s \, \epsilon_{\sigma_k(1)} + b_s \, \epsilon_{\sigma_k(s)} + c_s \, \epsilon_{\sigma_k(s)} \right) = 0 \quad (4.33)$$
$$\sum_{s=2}^{rm} \left(a_s \, \epsilon_{\sigma_k(s)} + b_s \, \epsilon_{\sigma_k(1)} + c_s \, \epsilon_{\sigma_k(s)} \right) = 0. \quad (4.34)$$

If we expand this expressions as affine linear combinations of the ϵ_l, then by Lemma 4.4.1 the coefficient of ϵ_l must be m^{-1} for all $l \in [m]$. Translating this for Equations (4.32), (4.33) and (4.34) respectively with $2 \leq l \leq m$ and $k \in [r]$, and using for $j \in [r]$ that

$$\sigma_k\big(r(l-1) + j - k + 1\big) = \left\lceil \frac{(r(l-1) + j - k + 1) + (k-1)}{r} \right\rceil = l \quad (4.35)$$

we get for all $k \in [r], l \in \{2, 3, \ldots, m\}$ that

$$\sum_{j=1}^{r} \left(a_{r(l-1)+j-k+1} + b_{r(l-1)+j-k+1} + c_{r(l-1)+j-k+2} \right) = \frac{1}{m} \quad (4.36)$$
$$\sum_{j=1}^{r} \left(b_{r(l-1)+j-k+1} + c_{r(l-1)+j-k+1} \right) = \frac{1}{m} \quad (4.37)$$
$$\sum_{j=1}^{r} \left(a_{r(l-1)+j-k+1} + c_{r(l-1)+j-k+1} \right) = \frac{1}{m} \quad (4.38)$$

respectively, where we set $c_{rm+1} := 0$. Fixing some $l \geq 2$ and subtracting (4.37) with $k = 1$ from (4.37) for $k = 2$, we find a telescoping sum that reduces to

$b_{r(l-1)} + c_{r(l-1)} = b_{rl} + c_{rl}$. Indeed, subtracting the two yields

$$\begin{aligned} 0 &= \sum_{j=1}^{r} \big(b_{r(l-1)+j-1} + c_{r(l-1)+j-1}\big) - \sum_{j=1}^{r} \big(b_{r(l-1)+j} + c_{r(l-1)+j}\big) \\ &= \sum_{j=0}^{r-1} \big(b_{r(l-1)+j} + c_{r(l-1)+j}\big) - \sum_{j=1}^{r} \big(b_{r(l-1)+j} + c_{r(l-1)+j}\big) \\ &= (b_{r(l-1)} + c_{r(l-1)}) - (b_{rl} + c_{rl}). \end{aligned}$$

More generally, for $k \in [r-1]$ combining (4.37) for k and $k \leftarrow k+1$, implies $b_{rl-k+1} + c_{rl-k+1} = b_{r(l-1)-k+1} + c_{r(l-1)-k+1}$ for all $l = 2, \ldots, m$, i.e. for every $k \in [r-1]$ we have

$$c_{r-k+1} = b_{r-k+1} + c_{r-k+1} = b_{2r-k+1} + c_{2r-k+1} = \ldots = b_{rm-k+1} + c_{rm-k+1}. \quad (4.39)$$

We are still missing the value $k = 0$, i.e., the equations

$$b_{r+1} + c_{r+1} = b_{2r+1} + c_{2r+1} = \ldots = b_{r(m-1)+1} + c_{r(m-1)+1}. \quad (4.40)$$

We obtain this by subtracting, for $l = 2, \ldots, m$, (4.37) for $k = 1$ and l from (4.37) with $k = r$ and $l \leftarrow l+1$. Indeed,

$$\begin{aligned} 0 &= \sum_{j=1}^{r} \big(b_{rl+j-r+1} + c_{rl+j-r+1}\big) - \sum_{j=1}^{r} \big(b_{r(l-1)+j} + c_{r(l-1)+j}\big) \\ &= \sum_{j=2}^{r+1} \big(b_{r(l-1)+j} + c_{r(l-1)+j}\big) - \sum_{j=1}^{r} \big(b_{r(l-1)+j} + c_{r(l-1)+j}\big) \\ &= \big(b_{rl+1} + c_{rl+1}\big) - \big(b_{r(l-1)+1} + c_{r(l-1)+1}\big). \end{aligned}$$

Lastly, we are missing the equations $b_2 + c_2 = b_3 + c_3 = \ldots = b_{r+1} + c_{r+1}$ for (4.30). We have not yet used in (4.33) the values $k = r + p$ with $p \in [r-1]$. For this we note that

$$\sigma_{r+p}\big(j\big) = 2 \quad \text{for } \; j \in \{r-p+1\} \cup \{r+2, r+3, \ldots, 2r\}.$$

We use this equation to apply Lemma 4.4.1 to (4.33) for ϵ_2 and $k = r + p$ with $p \in [r-1]$ to obtain

$$b_{r-p+1} + c_{r-p+1} + \sum_{j=2}^{r} \big(b_{r+j} + c_{r+j}\big) = \frac{1}{m}.$$

We need one more equation to eliminate the right-hand term, so we use the following. Lemma 4.4.1 applied to Equation (4.37) for $k = 1$ and $l = 2$ yields

$$\sum_{j=1}^{r} \big(b_{r+j} + c_{r+j}\big) = \frac{1}{m}.$$

Subtracting this equation from the previous one yields, $b_{r-p+1} + c_{r-p+1} = b_{r+1} + c_{r+1}$ for all $p = 1, \dots, r-1$. Together with the Equations (4.39) and (4.40) we conclude (4.30). Analogously, (4.34) and (4.38) can be used to obtain (4.31).

To get a contradiction we show that $a_s = b_s = c_s = 0$ for all $s = 2, 3, \dots, rm$. For this, we set $a := \sum_s a_s$ and $b := \sum_s b_s$. Equation (4.35) still applies for $l = 1, k = 1$, so Lemma 4.4.1 applied to the coefficient of ϵ_1 in (4.32), in (4.33) and in (4.34) respectively for $k = 1$ gives

$$\sum_{j=1}^{r} c_{j+1} = \frac{1}{m}, \qquad a + \sum_{j=1}^{r-1} c_{j+1} = \frac{1}{m} \qquad \text{and} \qquad b + \sum_{j=1}^{r-1} c_{j+1} = \frac{1}{m}$$

respectively. Subtracting the second equation from the first gives $a = c_{r+1}$, and reasoning analogously for the third yields $a = b = c_{r+1}$. Moreover, (4.37) with $k = r$ and $l = 2$ is $\sum_{j=1}^{r}(b_{j+1} + c_{j+1}) = m^{-1}$. Using the latter together with $b_2 = \ldots = b_r = 0$ and $\sum_{j=1}^{r} c_{j+1} = m^{-1}$ yields $b_{r+1} = 0$ and similarly $a_{r+1} = 0$ via (4.38) with $k = r$ and $l = 2$.

Since now also $a_{r+1} = b_{r+1} = 0$, the Equation (4.36) with $k = r$ and $l = 2$ simplifies to $\sum_{j=1}^{r} c_{j+2} = m^{-1}$. In conjunction with $\sum_{j=1}^{r} c_{j+1} = m^{-1}$ we deduce $c_2 = c_{r+2}$ and hence $b_{r+2} = 0 = a_{r+2}$ by (4.30) and (4.31). But now (4.36) with $k = r-1$ and $l = 2$ is $\sum_{j=1}^{r} c_{j+3} = m^{-1}$ and together with $\sum_{j=1}^{r} c_{j+2} = m^{-1}$ we get $c_3 = c_{r+3}$. Continuing inductively we obtain

$$\forall\, j \in [r]\colon \quad c_{j+1} = c_{r+j+1} \quad \text{and} \quad a_{r+j+1} = b_{r+j+1} = 0$$

via (4.36) with $l = 2$, $k \in [r]$ and via (4.30), (4.31). Then (4.36) with $k = r$ and $l = 3$ simplifies to $\sum_{j=1}^{r} c_{r+j+2} = m^{-1}$ and together with $m^{-1} = \sum_{j=1}^{r} c_{j+1} = \sum_{j=1}^{r} c_{r+j+1}$ we have $c_{r+2} = c_{2r+2}$. Hence, $b_{2r+2} = 0 = a_{2r+2}$ via (4.30) respectively (4.31). Continuing inductively in the outlined manner with Equation (4.36) for $k \in [r]$, $l = 3, \dots, m$ and with the Equations (4.30) and (4.31) we conclude $a_s = b_s = 0$ for all $s = 2, 3 \dots, rm$, so $a = b = 0$. Finally, (4.30) implies $c_{r+1} = c_s$ for all $s = 2, \dots, rm$, but $c_{r+1} = b = 0$ giving the desired contradiction. □

4.5.4 Padding of tensor factors

Theorem 4.5.1 only gives bounds on $\gamma_T(\pi_{m,d})$ and $\gamma_G(\pi_{m,d})$ for certain sub-families of $\{(m, d) \mid m \geq 2, d \geq 3\}$. Still, we can deduce Theorem 4.2.1, which gives a bound for all $m \geq 2$ and all $d \geq 3$, via some padding on the number of tensor factors d. That padding is provided in this subsection and used to prove Theorem 4.2.1. Recall that $\Omega(\pi_{m,d}) = \{\epsilon_i \mid i \in [m]\}^d \subseteq (\mathbb{R}^m)^d$.

Proposition 4.5.16 ([FR21, Proposition C.1])**.** *Let $m, d \geq 1$. Consider a set of weights $\Gamma_{m,d} \subseteq \Omega(\pi_{m,d})$ such that $0 \notin \operatorname{conv}(\Gamma_{m,d})$, i.e., $\Gamma_{m,d}$ witnesses the inequality $\gamma_T(\pi_{m,d}) \leq \operatorname{dist}(0, \operatorname{conv}(\Gamma_{m,d}))$.*

(i) Then $\gamma_T(\pi_{m,d+1}) \leq \operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d})\big)$. Thus, $\gamma_T(\pi_{m,d+1}) \leq \gamma_T(\pi_{m,d})$.

(ii) If $\Gamma_{m,d}$ is free, then $\gamma_G(\pi_{m,d+r}) \leq \operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d})\big)$ for all $r \geq 2$.

Proof. To prove the statement we set for $r \geq 1$

$$\Upsilon_r := \big\{(\epsilon_i, \dots, \epsilon_i) \mid i \in [m]\big\} \subseteq (\mathbb{R}^m)^r \quad \text{and} \quad \Gamma_{m,d+r} := \Gamma_{m,d} \times \Upsilon_r \subseteq \Omega(\pi_{m,d+r}).$$

By Equation (4.4) we have $0 \in \operatorname{conv}(\Upsilon_r)$ and therefore

$$\operatorname{conv}(\Gamma_{m,d+r}) = \operatorname{conv}(\Gamma_{m,d}) \times \operatorname{conv}(\Upsilon_r) \supseteq \operatorname{conv}(\Gamma_{m,d}) \times \{0\}.$$

The latter implies

$$\operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d+r})\big) \leq \operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d})\big). \tag{4.41}$$

Since $\operatorname{conv}(\Gamma_{m,d+r}) = \operatorname{conv}(\Gamma_{m,d}) \times \operatorname{conv}(\Upsilon_r)$, the assumption $0 \notin \operatorname{conv}(\Gamma_{m,d})$ yields $0 \notin \operatorname{conv}(\Gamma_{m,d+r})$. The latter shows $\gamma_T(\pi_{m,d+1}) \leq \operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d+1})\big)$ for $r = 1$ and we conclude the desired inequality with (4.41). Taking the minimum over all $\Gamma_{m,d} \subseteq \Omega(\pi_{m,d})$ with $0 \notin \operatorname{conv}(\Gamma_{m,d})$ shows that $\gamma_T(\pi_{m,d+1}) \leq \gamma_T(\pi_{m,d})$.

Assume in addition that $\Gamma_{m,d}$ is free and let $r \geq 2$. Considering Definition 4.3.2 and Proposition 4.3.3 we prove that also $\Gamma_{m,d+r}$ is free. For this, let $\mathscr{W} \subseteq [m]^d$ be such that $\Gamma_{\mathscr{W}} = \Gamma_{m,d}$ and consider $(x, i, \dots, i), (y, j, \dots, j) \in \mathscr{W} \times [m]^r$ with $(x, i, \dots, i) \neq (y, j, \dots, j)$. If $x \neq y$, then x and y differ in at least two components by freeness of $\mathscr{W}$. If $x = y$, then we have $i \neq j$ and so $(x, i, \dots, i)$ and $(y, j, \dots, j)$ differ in at least two components using $r \geq 2$. This shows that $\Gamma_{m,d+r}$ is free for $r \geq 2$. Since also $0 \notin \operatorname{conv}(\Gamma_{m,d+r})$ we obtain with Proposition 4.3.7(ii) that $\gamma_G(\pi_{m,d+r}) \leq \operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d+r})\big)$ holds for all $r \geq 2$. Finally, we deduce the second statement using Equation (4.41). □

The preceding proposition allows to pad the results from Theorem 4.5.1 to almost all tuples (m, d). Since Proposition 4.5.16(ii) requires a step length of at least two, the case $m \geq 3$ and $d = 4$ is missing for the gap $\gamma_G(\pi_{m,d})$.[21]

Proposition 4.5.17 ([FR21, Proposition C.2])**.** *For all $m \geq 3$ it holds that $\gamma_T(\pi_{m,4}) \leq \gamma_G(\pi_{m,4}) \leq 2^{-m+1}$.*

Proof. This result can be obtained by imitating the proof of Theorem 4.5.1(b) in Subsection 4.5.2. Defining

$$\Gamma_{m,4} := \big\{(\epsilon_i, \epsilon_j, \epsilon_k, \epsilon_i) \mid (i, j, k) \in \mathscr{W}_{m,3}\big\} \subseteq \Omega(\pi_{m,4}).$$

we have $0 \notin \operatorname{conv}(\Gamma_{m,4})$ as $0 \notin \operatorname{conv}(\Gamma_{m,3})$ by Lemma 4.5.6. Moreover, one can show with Lemma 4.5.7 (similar to the proof of Lemma 4.5.9) that

$$x := -\frac{1}{c\, 2^{m-1}}(\epsilon_1, \epsilon_1, \epsilon_1, \epsilon_1) \in \operatorname{conv}(\Gamma_{m,4}), \quad \text{where} \quad c = m - 2^{-m+1} \geq 2.$$

Thus, $\|(\epsilon_1, \epsilon_1, \epsilon_1, \epsilon_1)\| \leq \sqrt{4}$ implies $\|x\| \leq c^{-1} 2^{-m+1} \sqrt{4} \leq 2^{-m+1}$. This proves $\gamma_T(\pi_{m,4}) \leq 2^{-m+1}$.

Since $\mathscr{W}_{m,3}$ is free by Proposition 4.5.10, the set $\{(i, j, k, i) \mid (i, j, k) \in \mathscr{W}_{m,3}\}$ is free. Hence, $\gamma_G(\pi_{m,4}) \leq 2^{-m+1}$ by Proposition 4.3.3 and Proposition 4.3.7. □

[21] Given the fact $\gamma_T(\pi_{m,d+1}) \leq \gamma_T(\pi_{m,d})$, it is natural to ask whether the same inequality holds for the gap. This would lead to a more natural argument than the one presented here.

Using Propositions 4.5.16 and 4.5.17 we can deduce Theorem 4.2.1 from Theorem 4.5.1. We provide a proof to justify that the constant $C = 1/16$ always works, compare Remark 4.2.2.

Proof of Theorem 4.2.1. First, note that all upper bounds in Theorem 4.5.1 involve a negative exponent. Even for $m = 2$ and $d = 3$ we have $\gamma_G(\pi_{2,3}) \leq 2^{-1/2}$, see Theorem 4.5.1(a). Moreover, note that thanks to Theorem 4.5.1(c) we need to pad at most seven tensor factors[22] to apply a bound from Theorem 4.5.1. Consequently, Propositions 4.5.16 and 4.5.17 show that a constant $C > 0$ with

$$\forall\, m \geq 2,\, d \geq 3\colon \qquad \gamma_G(\pi_{m,d}) \leq 2^{-Cmd} \tag{4.42}$$

exists. Moreover, as d grows the impact of the padding becomes smaller, and hence for $d, m \gg 0$ we can choose $C \approx 1/6$ by Theorem 4.5.1(c).

By the above arguments, it suffices to show for small d and m (and biggest necessary padding step) that $C := 1/16$ satisfies Eq. (4.42). First, if $m = 2$ then

$$-\frac{d}{2} + 1 \leq -Cmd = -\frac{2d}{16} \qquad \Leftrightarrow \qquad -\frac{3d}{8} \leq -1$$

and the latter holds for all $d \geq 3$. Together with Theorem 4.5.1(a) this settles the case $m = 2$ and $d \geq 3$. The largest padding step when applying the bound from Theorem 4.5.1(b) arises for $d = 10$. In this case

$$-m + 1 \leq -Cmd = -\frac{10m}{16} \qquad \Leftrightarrow \qquad -\frac{3m}{8} \leq -1$$

and the inequality is satisfied for all $m \geq 3$. For $d < 10$ the required lower bound on m gets smaller. Finally, we consider the largest padding step and smallest d when Theorem 4.5.1(c) is applied. This is the case for $d = 16$ and we use the bound with $r = 2$. We have

$$-2m + 3 = -r(m-1) + 1 \leq -Cmd = -\frac{16}{16}m \qquad \Leftrightarrow \qquad -m \leq -3$$

which is equivalent to $m \geq 3$. This ends the proof. □

4.6 Polynomial Scaling

In this subsection we transfer the bounds on weight margin and gap from d-tensors to bounds on polynomial scaling. For this, let $\mathbb{C}[x_1, \ldots, x_n]_d$ denote the $\mathbb{C}$-vector space of homogeneous polynomials of degree d in n variables (including zero). Polynomial Scaling is given by the natural $\mathrm{SL}_n(\mathbb{C})$ action on $\mathbb{C}[x_1, \ldots, x_n]_d$. The corresponding representation[23] is

$$\varrho_{n,d}\colon \mathrm{SL}_n(\mathbb{C}) \to \mathrm{GL}\left(\mathbb{C}[x_1, \ldots, x_n]_d\right),\ g \mapsto \big(p(x) \mapsto p(g^{-1}x)\big).$$

[22] For example, note that we cannot apply the bound from Theorem 4.5.1(c) for $d = 10$ since Proposition 4.5.16 requires a padding step of at least two for the gap. Hence, we have to use the bound for $d = 3$, Theorem 4.5.1(b), and have to pad seven factors.

[23] We note the following, even though we do not explicitly need it here. To ensure $K = \mathrm{SU}_m$ invariance of the inner product under the action of $\varrho_{n,d}$, one has to equip $\mathbb{C}[x_1, \ldots, x_n]_d$ with the Bombieri-Weyl inner product, see e.g., [BC13, Section 16.1]

We remark that applications of polynomial scaling and related literature are discussed in Section 4.2.

Each monomial $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, where $\alpha = (\alpha_1, \ldots, \alpha_n) \in (\mathbb{Z}_{\geq 0})^n$ is a multi-index with $|\alpha| := \sum_i \alpha_i = d$, is a weight vector of $\varrho_{n,d}$ with weight $-\alpha + \frac{d}{n}\mathbb{1}_n$. Therefore,

$$\Omega(\varrho_{n,d}) = \left\{ -\alpha + \frac{d}{n}\,\mathbb{1}_n \;\middle|\; \alpha \in (\mathbb{Z}_{\geq 0})^n \text{ with } |\alpha| = d \right\}$$

as the monomials of degree d span $\mathbb{C}[x_1, \ldots, x_n]_d$.

We transfer the bounds for $\pi_{m,d}$ to $\varrho_{n,d}$ by relating their set of weights as follows. If $n = dm$ for some integer $m \geq 1$ and $i \in [m]$, then $\epsilon_i = e_i - \frac{1}{m}\mathbb{1}_m = e_i - \frac{d}{n}\mathbb{1}_m$. Hence, for any $i_1, \ldots, i_d \in [m]$ we have

$$-\big(\epsilon_{i_1}, \ldots, \epsilon_{i_d}\big) = -\big(e_{i_1}, \ldots e_{i_d}\big) + \frac{d}{n}\big(\mathbb{1}_m, \ldots, \mathbb{1}_m\big) = -\big(e_{i_1}, \ldots, e_{i_d}\big) + \frac{d}{n}\,\mathbb{1}_{dm},$$

which shows $-\Omega(\pi_{m,d}) \subseteq \Omega(\varrho_{n,d})$. Thus, we can transfer bounds on $\gamma_{\mathrm{ST}_m(\mathbb{C})^d}(\pi_{m,d})$ to bounds on $\gamma_{\mathrm{ST}_n(\mathbb{C})}(\varrho_{n,d})$. The next statement ensures the same for the gap.

Proposition 4.6.1 ([FR21, Proposition 4.16])**.** *Let $\Gamma \subseteq \Omega(\pi_{m,d})$ and $n = dm$ for some integer $m \geq 1$. If $\Gamma \subseteq \Omega(\pi_{m,d})$ is free, then $-\Gamma \subseteq \Omega(\varrho_{n,d})$ is free.*

Proof. We prove the statement by contraposition. Assume that $-\Gamma \subseteq \Omega(\varrho_{n,d})$ is not free. Then there exists a root $\alpha = e_i - e_j \in \mathbb{R}^n$ of $\mathrm{SL}_n(\mathbb{C})$, where $i, j \in [n]$ with $i \neq j$, and two distinct weights $\omega, \omega' \in -\Gamma$ such that $\omega = \omega' + e_i - e_j$, equivalently, $-\omega = -\omega' - e_i + e_j$. The latter enforces $-\alpha$ to be of the form

$$(0_m, \ldots, 0_m, e_k - e_l, 0_m, \ldots, 0_m) \in (\mathbb{R}^m)^d \cong \mathbb{R}^n \quad \text{for some } k, l \in [m] \text{ with } k \neq l,$$

because $-\omega, -\omega' \in \Omega(\pi_{m,d}) = \{(\epsilon_{i_1}, \ldots, \epsilon_{i_d}) \mid i_1, \ldots, i_d \in [m]\}$. Thus, $-\alpha$ is a root of $\mathrm{SL}_m(\mathbb{C})^d$ and hence $\Gamma \subseteq \Omega(\pi_{m,d})$ is not free. □

As a consequence we obtain bounds for the gap of polynomial scaling.

Theorem 4.6.2 (Gap for Polynomial Scaling, [FR21, Theorem 4.17])**.**
Let $d \geq 3$ and let $n = dm$ for some integer $m \geq 2$. Set $G := \mathrm{SL}_n(\mathbb{C})$ and $T := \mathrm{ST}_n(\mathbb{C})$. Then there exists a constant $C > 0$, independent of n and d, with

$$\gamma_T(\varrho_{n,d}) \leq \gamma_G(\varrho_{n,d}) \leq 2^{-Cdm} = 2^{-Cn}.$$

More concretely, for $d = 3$ and $m \geq 3$ it holds that

$$\gamma_T(\varrho_{n,3}) \leq \gamma_G(\varrho_{n,3}) \leq 2^{-m+1} = 2^{-\frac{n}{3}+1},$$

and if $m \geq 3$ and $d = 6r - 3$ for some $r \geq 2$, we have

$$\gamma_T(\varrho_{n,6r-3}) \leq \gamma_G(\varrho_{n,6r-3}) \leq 2^{-r(m-1)+1} = 2^{-\frac{(d+3)(m-1)}{6}+1} \approx 2^{-\frac{n}{6}}.$$

Proof. First, remember that $\gamma_T(\varrho_{n,d}) \leq \gamma_G(\varrho_{n,d})$, by Proposition 4.1.4. Furthermore, we recall that Theorem 4.2.1 was proven by padding the results from Theorem 4.5.1. Thus, the bound $\gamma_{\mathrm{SL}_m(\mathbb{C})^d}(\pi_{m,d}) \leq 2^{-Cdm}$ for each $m \geq 2$ and $d \geq 3$ from Theorem 4.2.1 is witnessed by a free set of weights $\Gamma_{m,d} \subseteq \Omega(\pi_{m,d})$, i.e., $0 < \operatorname{dist}(0, \operatorname{conv}(\Gamma_{m,d})) \leq 2^{-Cdm}$. But then we also have $0 \notin \operatorname{conv}(-\Gamma_{m,d})$, and that $-\Gamma_{m,d} \subseteq \Omega(\varrho_{n,d})$ is free by Proposition 4.6.1. Therefore,

$$\gamma_G(\varrho_{n,d}) \leq \operatorname{dist}\big(0, \operatorname{conv}(-\Gamma_{m,d})\big) = \operatorname{dist}\big(0, \operatorname{conv}(\Gamma_{m,d})\big) \leq 2^{-Cdm}.$$

by Proposition 4.3.7. Applying to the latter equation the inequalities from Lemma 4.5.9 respectively Lemma 4.5.15 yields the other two inequalities. □

4.7 Action on a Family of Quivers

In this section we study a certain family of quivers and its corresponding SL-action. For GL-actions on quivers the weight margin (and hence the gap) are large, i.e., inverse polynomial in the number of vertices and the entries of the dimension vector, compare [BFG+19, Theorem 6.21 Item 2]. Therefore, the algorithms in [BFG+19] solve NCM in polynomial time. In the case of SL-actions [BFG+19, Theorem 6.21 Item 4] provides a lower bound on the weight margin, which is exponentially small in the number of vertices. We show that this general lower bound can essentially not be improved: the SL-weight margin for our family of quivers is exponentially small in the number of vertices, see Theorem 4.7.1. Interestingly, its gap is still large as we state in Theorem 4.7.6 – a result due to Cole Franks and Visu Makam.

4.7.1 Upper Bounds on Weight Margin and Gap

For $d \geq 2$ let Q_d be the quiver

$$1 \longleftarrow 2 \longrightarrow 3 \cdots\cdots d-2 \longrightarrow d-1 \longleftarrow d \qquad \text{if } d \text{ even}$$

$$1 \longrightarrow 2 \longleftarrow 3 \cdots\cdots d-2 \longrightarrow d-1 \longleftarrow d \qquad \text{if } d \text{ odd}$$

and let $Q_d^{(k)}$ be the quiver one obtains from Q_d by adding $k-1$ additional copies of each arrow in Q_d. Then $G = \mathrm{SL}_m(\mathbb{C})^d$ (and $T = \mathrm{ST}_m(\mathbb{C})^d$) act on the quiver Q_d with dimension vector $(m, \ldots, m)$ as described in Example 1.3.8. We denote the corresponding representation by

$$\tau_{m,d} \colon \mathrm{SL}_m(\mathbb{C})^d \to \mathrm{GL}\left((\mathbb{C}^{m\times m})^{d-1}\right).$$

Note that the action of G on $Q_d^{(k)}$ with dimension vector $(m, \ldots, m)$ is given by $\tau_{m,d}^{\oplus k}$. In this subsection we prove an upper bound on the weight margin of $\tau_{m,d}$ and on the gap of $\tau_{m,d}^{\oplus m}$. The bound on $\gamma_G(\tau_{m,d}^{\oplus m})$ is thanks to the refinement in Proposition 4.3.7 pointed out by Visu Makam.

Theorem 4.7.1 ([FR21, Theorem 4.25])**.** *Let $m, d \geq 2$. It holds that*

$$\gamma_T(\tau_{m,d}) \leq (m-1)^{-d+1} \qquad \textit{and} \qquad \gamma_G(\tau_{m,d}^{\oplus m}) \leq (m-1)^{-d+1}.$$

Remark 4.7.2 ([FR21, Remark 4.26]). Before proving the theorem, we point out a few consequences.

1. Theorem 4.7.1 shows that $\gamma_T(\tau_{m,d})^{-1}$ and $\gamma_G(\tau_{m,d}^{\oplus m})^{-1}$ are not polynomially bounded in $\dim(\mathbb{C}^{m\times m})^{d-1} = (d-1)m^2$ and $\dim \mathrm{SL}_m(\mathbb{C})^d = d(m^2-1)$. Instead we see for fixed m and $d \to \infty$ an exponential behaviour in the number of vertices d. Thus, our bound shows that the exponential behaviour in d cannot be avoided in general lower bounds for quiver actions like [BFG+19, Theorem 6.21 Item 4]. The latter applied to $\tau_{m,d}$ shows $\gamma_T(\tau_{m,d}) \geq m^{-d^2-(3/2)d}(dm+1)^{-d}$.

2. The proof of Theorem 4.7.1 below shows that for the bound on the gap it is enough to consider the quiver $Q_d^{(m-1)}$ with an additional m^{th} arrow from d to $d-1$.

3. The ideas presented below can be adjusted to prove similar bounds for other dimension vectors. For example, one can show that the gap for the SL-action on $Q_d^{(2)}$ with dimension vector $(1,3,3,\dots,3,2)$ is inverse exponential in d. This aligns with an algebraic barrier for this action; the invariants that cut out the null cone for this action have exponential degree [DM18, Proposition 1.5].

4. In Theorem 4.7.6 we see that the gap $\gamma_G(\tau_{m,d})$ is only polynomially small in m and d. Thus, $\tau_{m,d}$ is an interesting family of representations for which the weight margin and gap differ significantly. ▽

To prove Theorem 4.7.1 we proceed again as described in Section 4.4. Note that the set of weights of $\tau_{m,d}$ viewed as a subset of $(\mathbb{R}^m)^d$ is

$$\big\{\big((-1)^d\epsilon_i, (-1)^{d-1}\epsilon_j, 0,\dots,0\big), \big(0, (-1)^{d-1}\epsilon_i, (-1)^{d-2}\epsilon_j, 0,\dots,0\big),\dots \\ \dots, \big(0,\dots,0,\epsilon_i,-\epsilon_j\big) \mid i,j\in[m]\big\}.$$

We define recursively subsets of weights $\Upsilon_{m,d} \subseteq \Omega(\tau_{m,d})$ via

$$\Upsilon_{m,2} := \{(\epsilon_i,-\epsilon_j) \mid i\in[m-1],\, j\in[m]\}\ ,\ \text{and for } d\geq 3$$
$$\Upsilon_{m,d} := \big\{\big((-1)^d\epsilon_i, (-1)^{d-1}\epsilon_m, 0_m,\dots,0_m\big) \mid i\in[m-1]\big\} \cup \big(\{0_m\}\times\Upsilon_{m,d-1}\big)\,.$$

Remark 4.7.3 ([FR21, Remark 4.28]). We note that for all $d\geq 2$, $\Upsilon_{m,d}$ is *not* free. For instance, we can always write

$$(0_m,\dots,0_m,\epsilon_1,-\epsilon_1) = (0_m,\dots,0_m,\epsilon_1,-\epsilon_2) + (0_m,\dots,0_m,0_m,e_2-e_1),$$

i.e., the weights $(0_m,\dots,0_m,\epsilon_1,-\epsilon_1)$, $(0_m,\dots,0_m,\epsilon_1,-\epsilon_2) \in \Upsilon_{m,d}$ differ by the root $(0_m,\dots,0_m,0_m,e_2-e_1)$ of $\mathrm{SL}_m(\mathbb{C})^d$. Therefore, we *cannot* deduce a bound on the gap $\gamma_G(\tau_{m,d})$ via Proposition 4.3.7. However, the latter allows us to deduce at least a bound on the gap of $\tau_{m,d}^{\oplus m}$. ▽

In the next two lemmas we show that $\Upsilon_{m,d}$ witnesses the bound on $\gamma_T(\tau_{m,d})$ and afterwards we use Proposition 4.3.7 to transfer this bound to $\gamma_G(\tau_{m,d}^{\oplus m})$.

Lemma 4.7.4 ([FR21, Lemma 4.29]). *For all $d \geq 2$ it holds that $0 \notin \operatorname{conv}(\Upsilon_{m,d})$.*

Proof. We prove the statement by induction on $d \geq 2$. For $d = 2$, just note that any element in $\operatorname{conv}(\Upsilon_{m,2}) \subseteq \mathbb{R}^{2m}$ has value $-1/m$ in the m^{th} entry. In particular, $0 \notin \operatorname{conv}(\Upsilon_{m,2})$. For $d \geq 3$ let

$$x = \sum_{\omega \in \Upsilon_{m,d}} \lambda_\omega \, \omega \, , \quad \lambda_\omega \geq 0$$

be a convex combination of the elements in $\Upsilon_{m,d}$. Assume there is an $i \in [m-1]$ such that for

$$\omega_i := \big((-1)^d \epsilon_i, (-1)^{d-1} \epsilon_m, 0_m, \ldots, 0_m\big)$$

one has $\lambda_{\omega_i} > 0$. Then the m^{th} entry of x is non-zero, since ω_i has m^{th} entry $(-1)^{d+1}/m$ and all (other) $\omega \in \Upsilon_{m,d}$ have $(-1)^{d+1}/m$ or zero as m^{th} entry. Hence, $x \neq 0$ in this case. On the other hand, if $\lambda_{\omega_i} = 0$ for all $i \in [m-1]$, then $x \in \{0_m\} \times \operatorname{conv}(\Upsilon_{m,d-1})$ by construction of $\Upsilon_{m,d}$. We conclude $x \neq 0$ by induction hypothesis on $d-1$. □

Lemma 4.7.5 ([FR21, Lemma 4.30]). *For $d \geq 2$ define*

$$x_d := \lambda_d\big((-1)^{d-1} \epsilon_m, 0_m, \ldots, 0_m\big) \in (\mathbb{R}^m)^d, \quad \textit{where } \lambda_d := \left(\sum_{i=1}^{d-1} (m-1)^i\right)^{-1} .$$

Then we have $x_d \in \operatorname{conv}(\Upsilon_{m,d})$ and $\|x_d\|_2 < |\lambda_d| \leq (m-1)^{-d+1}$.

Proof. We proceed by induction on $d \geq 2$. For $d = 2$, we use Equation (4.4) to obtain the convex combination

$$\sum_{i=1}^{m-1} \sum_{j=1}^{m} \frac{1}{(m-1)m} (\epsilon_i, -\epsilon_j) = \frac{1}{m-1} (-\epsilon_m, 0_m) = x_2 \, .$$

Now assume the claim is proven for some $d \geq 2$, hence

$$\lambda_d\big(0_m, (-1)^{d-1} \epsilon_m, 0_m, \ldots, 0_m\big) \in \{0_m\} \times \operatorname{conv}(\Upsilon_{m,d}) \subseteq \operatorname{conv}(\Upsilon_{m,d+1}). \tag{4.43}$$

Setting $\nu := (m-1)\lambda_{d+1}\lambda_d^{-1}$ we have $\nu\lambda_d = (m-1)\lambda_{d+1}$ and $\nu + (m-1)\lambda_{d+1} = 1$. Together with Equations (4.4) and (4.43) we deduce $x_{d+1} \in \operatorname{conv}(\Upsilon_{m,d+1})$ via

$$\begin{aligned}
&\nu\lambda_d\big(0_m, (-1)^{d-1} \epsilon_m, 0_m, \ldots, 0_m\big) + \lambda_{d+1} \sum_{i=1}^{m-1} \big((-1)^{d+1} \epsilon_i, (-1)^d \epsilon_m, 0_m, \ldots, 0_m\big) \\
&= \big(-(-1)^{d+1} \lambda_{d+1} \epsilon_m \, , \, (-1)^{d-1} \big[\nu\lambda_d - (m-1)\lambda_{d+1}\big] \epsilon_m \, , 0_m, \ldots, 0_m\big) \\
&= \big((-1)^d \lambda_{d+1} \epsilon_m, 0_m, 0_m, \ldots, 0_m\big) = x_{d+1} \, .
\end{aligned}$$

This ends the induction. Finally, $\|x_d\|_2 < |\lambda_d|$ as $\|\epsilon_m\|_2 < 1$ by (4.5). □

Proof of Theorem 4.7.1. By Lemma 4.7.4 and Lemma 4.7.5 we have

$$\gamma_T(\tau_{m,d}) \leq (m-1)^{-d+1}.$$

With the fact $\Omega(\tau_{m,d}) = \Omega(\tau_{m,d}^{\oplus m})$ and Proposition 4.3.7 we transfer this bound to the gap of $\tau_{m,d}^{\oplus m}$. To do so, we note that the inner product on $(\mathbb{C}^{m\times m})^{m(d-1)}$, given by the trace inner product on each $\mathbb{C}^{m\times m}$ copy, is invariant under the action of $K = \mathrm{SU}(m)^d$. Distinct $\mathbb{C}^{m\times m}$ copies are orthogonal under this inner product. Thus, to be able to apply Proposition 4.3.7 it is enough to assign to each $\mathbb{C}^{m\times m}$ copy, i.e., to each arrow of $Q_d^{(m)}$, a matrix M_i such that $\mathrm{supp}(M_i)$ is free and $\Upsilon_{m,d} = \bigcup_i \mathrm{supp}(M_i)$. For this, we consider the $m \times m$ matrices

$$M := \begin{pmatrix} \mathrm{I}_{m-1} & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad P := \begin{pmatrix} 0 & \mathrm{I}_{m-1} \\ 1 & 0 \end{pmatrix},$$

and $E_{i,j}$ is the matrix with (i,j)-entry one and all other entries zero. Then $E_{i,i}P = E_{i,\sigma(i)}$, where $\sigma\colon [m] \to [m]$ is the cycle $(1\ 2\ \ldots\ m)$. Therefore, for $k \in [m]$ we have

$$\mathrm{supp}\left(MP^{k-1}\right) = \left\{\left(0_{m(d-2)}, \epsilon_i, -\epsilon_{\sigma^{k-1}(i)}\right) \mid i \in [m-1]\right\};$$
$$\text{and} \quad \{0_{m(d-2)}\} \times \Upsilon_{m,2} = \bigcup_{k\in[m]} \mathrm{supp}\left(MP^{k-1}\right).$$

For fixed $k \in [m]$, $i_1 \neq i_2$ implies $\sigma^{k-1}(i_1) \neq \sigma^{k-1}(i_2)$, so any distinct elements of $\mathrm{supp}(MP^{k-1})$ differ in the last two $\mathbb{R}^m$-components. Hence, each $\mathrm{supp}(MP^{k-1})$ is free and we assign $M, MP, \ldots, MP^{m-1}$ to the m arrows that go from vertex d to vertex $d-1$. For $l \in [d-2]$, we assign to the m arrows between the vertices l and $l+1$ each of the matrices $E_{1,m}, E_{2,m}, \ldots, E_{m-1,m}$ at least once. (Exactly one of the latter matrices is assigned to two of these arrows.) Clearly, the support of $E_{i,m}$, $i \in [m-1]$ is free as it contains just one weight. By construction, this whole assignment gives an element of $(\mathbb{C}^{m\times m})^{m(d-1)})$ such that its support is $\Upsilon_{m,d}$ and so that we can apply Proposition 4.3.7. This shows

$$\gamma_G(\tau_{m,d}^{\oplus m}) \leq (m-1)^{-d+1}.$$

Moreover, the argument shows that $m-1$ arrows between the vertices l and $l+1$, $l \in [d-2]$, would suffice as commented in part two of Remark 4.7.2. □

4.7.2 A large lower Bound on the Gap

We show that the gap $\gamma_G(\tau_{m,d})$ is inverse polynomial in m and d. The presented proof is completely due to Cole Franks and Visu Makam. I heartily thank them for the permission to include their arguments here. The main result is the following.

Theorem 4.7.6. *For all $m, d \geq 2$ it holds that*

$$\gamma_G(\tau_{m,d}) \geq \frac{1}{d^2\, m}\,.$$

As a consequence of the "large" gap, the first order algorithm from [BFG+19] can solve the null-cone membership problem for $\tau_{m,d}$ in $\mathrm{poly}(m,d)$-time. There are also algebraic algorithms for this problem that run in polynomial-time, because Q_d is of finite representation type and has no oriented cycles.[24] This leads to the following interesting question.

Problem 4.7.7. *Let Q be a quiver of finite representation type and consider the* SL*-action on Q with dimension vector $(m_1,\dots,m_d)$. Is the gap of this action inverse polynomial in $m_1,\dots,m_d$ and the number of vertices d?*

To prove Theorem 4.7.6 we explicitly state the moment map of $\tau_{m,d}$. For $A \in \mathbb{C}^{m\times m}$, we recall from (2.14) that

$$\Phi_1(A) = -A^\dagger A + \frac{\|A\|_F^2}{m}\,\mathrm{I}_m \qquad \text{and} \qquad \Phi_2(A) = AA^\dagger - \frac{\|A\|_F^2}{m}\,\mathrm{I}_m\,. \tag{4.44}$$

The Hermitian matrices $\Phi_1(A)$ and $\Phi_2(A)$ are traceless as $\mathrm{tr}(A^\dagger A) = \mathrm{tr}(AA^\dagger) = \|A\|_F^2$. Furthermore, note that each vertex in Q_d is either a *source*, i.e., the vertex only appears as a tail of arrows, or a *sink*, i.e., the vertex only appears as a head. Thus, one can deduce the moment map of $\tau_{m,d}$ from Example 2.2.10. There we computed the moment map (2.15) of the quiver (2.13) with vertex 2 being a sink of two arrows. Moreover, we stated the moment map (2.18) of a similar quiver where vertex 2 is a source. With this knowledge we obtain the following.

Lemma 4.7.8. *Let $B = (B_1, B_2, \dots, B_{d-1}) \in (\mathbb{C}^{m\times m})^{d-1}$. Then the moment map $\mu := \mu_G$ of $\tau_{m,d}$ at B is $\mu(B) = \|B\|^{-2}\big(\mu_1(B),\dots,\mu_d(B)\big)$, where the components $\mu_i(B)$ are given as follows. We have $\mu_d(B) = \Phi_1(B_{d-1})$ and for $i \in [d-1]$*

$$\mu_i(B) = \begin{cases} \Phi_1(B_{i-1}) + \Phi_1(B_i), & \text{if vertex } i \text{ is a source} \\ \Phi_2(B_{i-1}) + \Phi_2(B_i), & \text{if vertex } i \text{ is a sink} \end{cases}$$

where we set $B_0 := 0$, so that $\Phi_1(B_0) = \Phi_2(B_0) = 0$.

Next, we point out that the action of $G = \mathrm{SL}_m(\mathbb{C})^d$ on $(\mathbb{C}^{m\times m})^{d-1}$ via $\tau_{m,d}$ preserves the determinant in each $\mathbb{C}^{m\times m}$ component. In particular, if for $B = (B_1,\dots,B_{d-1}) \in (\mathbb{C}^{m\times m})^{d-1}$ there is $i \in [d-1]$ with $\det(B_i) \neq 0$, then B is G-semistable. Equivalently, if B is G-unstable then $\mathrm{rank}(B_i) < m$ for all $i \in [d-1]$.[25] Thus, the next lemma will allow us to bound $\|\mu(B)\|$ for an unstable B.

Lemma 4.7.9. *Let $A \in \mathbb{C}^{m\times m}$. It holds that $\|\Phi_1(A)\|_F = \|\Phi_2(A)\|_F$, and if $\mathrm{rank}(A) < m$ then $\|\Phi_1(A)\|_F \geq m^{-1}\|A\|_F^2$.*

[24] Personal communication with Visu Makam. There does not seem to be an explicit reference in the literature. It seems plausible that the same is true for the optimization methods from [BFG+19].

[25] Actually, B is unstable if and only if $\mathrm{rank}(B_i) < m$ holds for all $i \in [d-1]$. The "if"-direction may be shown via Schofield invariants, which can be used to prove that the ring of invariants is generated by the $\det(B_i)$, $i \in [d-1]$.

Proof. Let USV be a singular value decomposition of A, i.e., U and V are unitary matrices and $S = \operatorname{diag}(\sigma_1, \dots, \sigma_m)$ with $\sigma_i \in \mathbb{R}_{\geq 0}$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$. Then $A^\dagger A = V^\dagger S^2 V$ and using that the Frobenius norm is invariant under unitary transformations we compute

$$\|\Phi_1(A)\|_F = \left\| V^\dagger \left(-S^2 + \frac{\|A\|_F^2}{m} \mathrm{I}_m \right) V \right\|_F = \left\| \left(S^2 - \frac{\|A\|_F^2}{m} \mathrm{I}_m \right) \right\|_F . \tag{4.45}$$

A similar computation via $AA^\dagger = US^2U^\dagger$ holds for $\|\Phi_2(A)\|_F$, which shows the first claim. If $\operatorname{rank}(A) < m$, then $\sigma_m = 0$ and we obtain with (4.45) that

$$\|\Phi_1(A)\|_F = \Big(\sum_{i=1}^{m} \big(\sigma_i^2 - m^{-1} \|A\|_F^2 \big)^2 \Big)^{1/2} \geq \left| \sigma_m^2 - m^{-1} \|A\|_F^2 \right| = m^{-1} \|A\|_F^2$$

holds as desired. □

Proof of Theorem 4.7.6. Let $B = (B_1, B_2, \dots, B_{d-1}) \in (\mathbb{C}^{m \times m})^{d-1} \setminus \{0\}$ be unstable with respect to $\tau_{m,d}$. To prove the claim it suffices to show $\|\mu(B)\| \geq (d^2\, m)^{-1}$, compare Definition 4.1.1. Since $\mu(\lambda B) = \mu(B)$ holds for all $\lambda \in \mathbb{C}^\times$, we can assume $\|B\| = 1$. Thus, $\mu(B) = \big(\mu_1(B), \dots, \mu_d(B)\big)$, where $\mu_i(B)$ is as in Lemma 4.7.8. We note that $\|\mu(B)\| \geq \|\mu_i(B)\|$ holds for all $i \in [d]$.

First, we prove by induction on $i \in [d-1]$ that

$$i \, \|\mu(B)\| \geq \|\Phi_1(B_i)\| = \|\Phi_2(B_i)\| \tag{4.46}$$

holds. By Lemma 4.7.9, we have $\|\Phi_1(B_i)\| = \|\Phi_2(B_i)\|$, so it suffices to show the inequality for one of them. For $i = 1$, we observe that $\mu_1(B) = \Phi_k(B_1)$ for some $k \in \{1, 2\}$, by Lemma 4.7.8. The claim follows with $\|\mu(B)\| \geq \|\mu_1(B)\|$. Now, assume that Equation (4.46) holds for some $i < d - 1$. Again by Lemma 4.7.8 there exists $k \in \{1, 2\}$ such that $\mu_{i+1}(B) = \Phi_k(B_{i+1}) + \Phi_k(B_i)$ and therefore $\|\Phi_k(B_{i+1}) + \Phi_k(B_i)\| \leq \|\mu(B)\|$. Together with the triangle inequality and the induction hypothesis we conclude

$$\|\Phi_k(B_{i+1})\| \leq \|\Phi_k(B_{i+1}) + \Phi_k(B_i)\| + \| - \Phi_k(B_i)\| \leq \|\mu(B)\| + i \, \|\mu(B)\| .$$

Finally, since $1 = \|B\|^2 = \sum_i \|B_i\|_F^2$, there exists $j \in [d-1]$ such that $\|B_j\|_F^2 \geq (d-1)^{-1} \geq d^{-1}$. We have $\operatorname{rank}(B_j) < m$ as B is unstable. Hence, Lemma 4.7.9 implies $\|\Phi_1(B_j)\| \geq m^{-1} \|B_j\|_F^2 \geq m^{-1} d^{-1}$. As desired, we obtain

$$\|\mu(B)\| \geq \frac{1}{j} \frac{1}{dm} \geq \frac{1}{d^2 m}$$

using Equation (4.46). □

Chapter 5

Bounds on the Diameter

This chapter is based on [FR21] and presents the diameter bounds from this paper. These bounds explain the dichotomy for high precision solutions (HP) from Table 3.1. Hence, they highly motivate, together with the weight margin and gap bounds from Chapter 4, the search for new geodesic convex methods.

Since all main proof ideas for the diameter bounds are due to my co-author Cole Franks, the exposition is restricted to the main results, their implications and relations to the literature, and a proof outline.

Organization and Assumptions. In Section 5.1 we state the main results on diameter bounds, and provide a discussion of their implications and relation to the literature. Afterwards, we give a brief proof outline in Section 5.2.

The whole chapter uses the assumptions stated in Setting 3.0.1; usually applied to the tensor scaling representation $\pi_{m,3}$ from Example 1.3.5.

5.1 Main Results and related Literature

In the following, we discuss the diameter as a complexity parameter and known upper bounds for it. Moreover, we present the main results, i.e., exponential diameter lower bounds for array and tensor scaling, and we discuss their implications and relations to the literature.

We start by recalling Definition 3.2.6. Given a representation $\pi\colon G \to \mathrm{GL}(V)$ of a reductive group G, $v \in V$ and a precision $\varepsilon > 0$, the *diameter* was defined as

$$D_v(\varepsilon) := \inf\left\{R > 0 \mid \inf_{g \in B'_R} \|g \cdot v\|^2 \leq \mathrm{cap}_G(v) + \varepsilon\right\},$$

where $B'_R := \{k \exp(X) \mid k \in K, X \in \mathfrak{i}\,\mathrm{Lie}(K), \|X\|_F \leq R\}$.

Let us illustrate this for the action of $T = \mathrm{ST}_m(\mathbb{C})^3$ via $\pi_{m,3}$, i.e., array scaling. Similarly to matrix scaling (3.3), for the array $p_{ijk} := |v_{ijk}|^2$, $v \in (\mathbb{C}^m)^{\otimes 3}$, the optimization problem

$$\mathrm{cap}_T(v) = \mathrm{cap}(p) := \inf_{x,y,z \in \mathbb{R}^m} f_p(x,y,z) := \inf_{x,y,z \in \mathbb{R}^m} \sum_{i,j,k=1}^{m} p_{ijk}\, e^{(\epsilon_i, \epsilon_j, \epsilon_k)\cdot(x,y,z)}. \quad (5.1)$$

captures scaling p to tristochastic, compare Section 3.1. Note, that we can also restrict to the infimum over $(\mathbb{1}_m^\perp)^3 = \mathfrak{i}\,\mathrm{Lie}(T_K)$. The diameter $D_v(\varepsilon)$ in this case is the infimum over all $R > 0$ such that

$$\inf\left\{f_p(x,y,z) \mid (x,y,z) \in \left(\mathbb{1}_m^\perp\right)^3, \|(x,y,z)\| \leq R\right\} \leq \mathrm{cap}_T(v) + \varepsilon.$$

Remember, a group element (in particular, an approximate minimizer) is recovered by $t(x, y, z) := \exp\big(\operatorname{diag}(x), \operatorname{diag}(y), \operatorname{diag}(z)\big) \in T$.

Significance of the Diameter. The above explanations for array scaling illustrate why one may regard the diameter as a measure for the *bit complexity* of an approximate minimizer.[1] Furthermore, recall that $\|(x, y, z)\|$ measures the distance between $\mathrm{id} = \exp(0)$ and $t(x, y, z)$ in the flat manifold T/T_K. This generalizes to the curved manifold G/K, so $D_v(\varepsilon)$ captures the distance of an approximate minimizer to the identity.[2] This directly regards it as a complexity parameter as follows.

Guarantees for many iterative algorithms in (geodesic) convex optimization require a bound on the distance D from the starting point to an ε-approximate solution.[3] For example, in the commutative setting the diameter bounds in [SV14; SV19] were used to design ellipsoid methods that are tractable even for very large support, and in [BLNW20] they were used to bound the running time of interior point methods. Similarly, diameter bounds were used to bound the running time of geodesic convex optimization methods [AGL+18; BFG+19].

Specifically, gradient descent (first order) and trust region[4] (second order) methods are iterative algorithms that make progress at each step within a usually small distance, say upper bounded by η.[5] This takes at least D/η many steps to produce an ε-approximate solution. Therefore, a polynomially large diameter is a necessary requirement for gradient descent and trust region methods to provide high precision solutions in polynomial time.

Finally, we remark that cutting plane methods typically use diameter bounds to control the volume of a starting region.

Known Diameter upper Bounds. In Table 5.1 we present known diameter upper bounds for matrix, array, operator and tensor scaling. For matrix scaling, we note that w_v is the ratio between the sum of the entries of the matrix v and its least non-zero entry. The upper bound for operator scaling, which also applies to matrix scaling, is obtained by combining Equation (4.1) with the diameter bound from [BFG+19] (see Theorem 3.2.7). Similarly, combining the general weight margin lower bound (4.2) from [BFG+19, Theorem 6.9] with Theorem 3.2.7 yields the bound for tensor scaling, which also applies to array scaling. Another upper bound for array scaling is $\mathrm{poly}(m^{3/2}2^m, \log(1/\varepsilon))$, which follows from the general upper bound of [SV19] on diameter bounds for unconstrained

[1] This is similar to the notion of bit complexity in [SV19].

[2] The set $B_R := \{\exp(X) \mid X \in \mathfrak{i}\,\mathrm{Lie}(K), \|X\|_F \leq R\}$ is a geodesic ball of radius R in G/K about the identity. Since K acts isometrically on V, we see that $D_v(\varepsilon)$ indeed captures the distance of an approximate minimizer to the identity.

[3] Here, this distance is the diameter $D_v(\varepsilon)$. Indeed, the identity is the natural starting point in G (more precisely, G/K) for Norm Minimization 3.1.3 and Scaling 3.1.4. Note that a different starting point $\exp(X) \in G/K$ already involves a "biased" direction $X \in \mathfrak{i}\,\mathrm{Lie}(K)$.

[4] also called *box-constrained Newton's method*

[5] For example, in [BFG+19] the progress of their geodesic first and second order method is controlled by the weight norm $N(\pi)$. Indeed, it bounds the gradient, [BFG+19, Lemma 3.12], and gives a smoothness as well as a robustness parameter [BFG+19, Propositions 3.13 and 3.15].

geometric programming. There is also a diameter bound for array scaling in the multimarginal transport context that is polynomial in the input size, but it assumes that the tensor has *no* non-zero entries [LHCJ22].

$\pi_{m,d}$	$T = \mathrm{ST}_m(\mathbb{C})^d$: commutative	$G = \mathrm{SL}_m(\mathbb{C})^d$: non-commutative
$d = 2$	**matrix scaling:** $O(m \log(w_v/\varepsilon))$ [CMTV17]	**operator scaling:** $O\big(m^2 \log(\|v\|/\varepsilon)\big)$ [BFG+19]
$d = 3$	**array scaling:** $O\big(m^{3/2}2^m, \log(\|v\|/\varepsilon)\big)$ [SV19]	**tensor scaling:** [BFG+19] $O\big(\sqrt{m}(\sqrt{3}m)^{3m} \log(\|v\|/\varepsilon)\big)$

Table 5.1: Diameter upper bounds for $\pi_{m,d}$. In the non-commutative case, we used Theorem 3.2.7 and the weight margin lower bounds (4.1) and (4.2). The non-commutative bounds also apply to the commutative case.

We point out that Table 5.1 captures the dichotomy for solving norm minimization with high precision (HP) as presented in Table 3.1.

Main Results. Given the upper bounds for $d = 3$ in Table 5.1, one is led to ask whether the exponential behaviour in m is too pessimistic or actually required. The following two theorems confirm the latter in the high precision regime, i.e., for ε being exponentially small in some polynomial in m.

For the commutative case, recall the definition of $f_p(x, y, z)$ and $\mathrm{cap}(p)$ from Equation (5.1). We stress that the following theorem is in terms of $p \in (\mathbb{R}^m_{\geq 0})^{\otimes 3}$ (which corresponds to $(|v_{ijk}|^2)_{ijk}$), and not in terms of $v \in (\mathbb{C}^m)^{\otimes 3}$.

Theorem 5.1.1 (Diameter Bound for Array Scaling, [FR21, Theorem 1.1])**.** *There is an absolute constant $C > 0$ and an array $p \in (\mathbb{R}^m_{\geq 0})^{\otimes 3}$ with $O(m)$ non-zero entries, each of bit-complexity $O(m)$, that satisfies the following property. For all $0 < \varepsilon \leq \exp(-Cm^2 \log m)$ and $(x, y, z) \in \mathbb{R}^{3m}$, if*

$$f_p(x, y, z) \leq \mathrm{cap}(p) + \varepsilon$$

then $\|(x, y, z)\| = \Omega\left(2^{m/3} \log(1/\varepsilon)\right)$. Moreover, $\mathrm{cap}(p) = 1/2$.

The final equality emphasizes that the difficulties do not lie in an additive vs multiplicative approximation, see Remark 3.2.4. By a simple duplication trick, the same bound holds for d-dimensional array scaling with $d \geq 3$, see [FR21, Corollary 3.7].

The constructed array p is free, which allows to lift the above theorem to the non-commutative case of tensor scaling. However, due to some required rounding (see Section 5.2) the tensor v *depends* on the precision ε.

Theorem 5.1.2 (Diameter Bound for Tensor Scaling, [FR21, Theorem 1.4])**.** *For the action of $G = \mathrm{SL}_m(\mathbb{C})^3$ via $\pi_{m,3}$, there is a constant $C > 0$ such that the following holds. For all $\varepsilon \leq \exp(-Cm^2 \log m)$, there exists $v = v(\varepsilon) \in (\mathbb{C}^m)^{\otimes 3}$ with $O(m)$ non-zero entries of bit complexity $O(\log m + \log(1/\varepsilon))$ and*

$$D_v(\varepsilon) = \Omega\big(2^{m/3} \log(1/\varepsilon)\big).$$

Moreover, $1/4 \leq \mathrm{cap}_G(v) \leq 1$ and $1/2 \leq \|v\| \leq 1$.

Again, the bounds on $\mathrm{cap}_G(v)$ and $\|v\|$ ensure that the difficulties are not caused by requiring an additive approximation, compare Remark 3.2.4. A duplication trick analogous to [FR21, Corollary 3.7] yields the same diameter bound for $d \geq 3$, but for the action of $G = \mathrm{SL}_m(\mathbb{C})^d$ on tuples of tensors via the representation $\pi_{m,d}^{\oplus m}$, see [FR21, Corollary 4.24].

Implications of the main Results. First, considering the diameter bound from [BFG+19] via the weight margin, compare Theorem 3.2.7, the main results show that $\gamma_T(\pi_{m,3})$ cannot be polynomially small in m. Instead, the weight margin for array and tensor scaling satisfies $\gamma_T(\pi_{m,3}) = \Omega\big(2^{-m/3}\big)$.[6]

Taking the explanations on the significance of the diameter into account, Theorem 5.1.1 shows that gradient descent and trust region methods for 3-dimensional array scaling with constant (or even polynomial) step size cannot provide high precision solutions in $\mathrm{poly}(m, \log(1/\varepsilon))$ time. Therefore, Theorem 5.1.1 explains why ellipsoid and interior point methods are necessary to achieve HP in polynomial time for array scaling.

Analogously, in the non-commutative case Theorem 5.1.2 shows that geodesic gradient descent and trust region methods with constant step size cannot ε-approximate the capacity in $\mathrm{poly}(m, 1/\varepsilon)$ time for 3-tensors. In particular, the first and second order method of [BFG+19] cannot solve norm minimization with high precision for tensor scaling in polynomial time.

Furthermore, Theorem 5.1.2 also indicates that cutting plane methods as suggested in [Rus20] do not suffice for tensor scaling, as follows. Cutting plane methods usually require an exponential bound on the volume of a known region containing an approximate optimizer. This is the case for Rusciano's non-constructive query upper bound for cutting plane methods on manifolds of non-positive curvature [Rus20]. This upper bound is essentially tight due to [HM21][7]. However, the volume of a ball in the manifold $\mathrm{SL}_m(\mathbb{C})^3/(\mathrm{SU}_m)^3$ grows *exponentially* in the radius, see [GN99]. Therefore, the diameter Theorem 5.1.2, which is exponential in m, shows that an approximate minimizer is only contained in a geodesic ball with volume at least *doubly* exponential in m.

The very recent preprints [Hir22; NW23] study self-concordant functions on Riemannian manifolds. Moreover, [NW23] provides (the main stage of) an interior point method on Riemannian manifolds, which can be used to solve Norm Minimization 3.1.3 and Scaling Problem 3.1.4 in the general Setting 3.0.1. However, the complexity still depends *linearly* on a diameter bound [NW23, Theorem 1.7]. Hence, the exponential diameter for tensor scaling from Theorem 5.1.2 excludes polynomial running time, making further research necessary [NW23, Outlook].

Altogether, the provided diameter bounds explain the dichotomy for HP in Table 3.1. Moreover, they highly motivated and keep motivating[8] the search for and the advancement of sophisticated methods in the geodesic convex setting.

[6] We stress that the bound in Theorem 4.5.1(b) is better, it also applies to the gap and has a rather short proof. In contrast, the above diameter bounds and Theorem 3.2.7 have long, technical proofs and in combination they do not yield a bound on the gap.

[7] [HM21] applies to the hyperbolic plane, which is a totally geodesic submanifold of the manifold we consider.

[8] in view of [NW23, Theorem 1.7]

Relation to the Literature. We remark that [BLNW20] bounds the diameter in the commutative case using the inverse of the so-called *facet gap*, [BFG+19, Definition 1.8]. The construction for Theorem 5.1.1 has exponentially small facet gap; see Corollary 5.2.2 below.

Regarding diameter *lower* bounds, it was shown that there is some bounded set $\Gamma \subset \mathbb{Z}^m$ in a poly(m) size ball such that the geometric program with weights given by Γ has *no* ε-approximate minimizers of norm poly$(m, \log 1/\varepsilon)$ [SV19]. We stress that the specific unconstrained geometric program in the latter result is not array scaling; also compare Remark 5.2.1 below. Actually, in [SV19, Section 2.1] the authors ask whether there is some Γ whose elements are Boolean (up to an additive shift) with a superpolynomial diameter lower bound. As subsets of $\Omega(\pi_{m,d})$ are of this form, we answer their open problem in the affirmative.

Comparing with the upper bounds in Table 5.1, we see that the lower bounds from Theorems 5.1.1 and 5.1.2 are tight up to logarithmic factors in the exponent.

It would be interesting to prove a version of Theorem 5.1.2 that holds for ε larger than $2^{-m+1} \geq \gamma_G(\pi_{m,3})$.[9] This would imply that trust region methods cannot solve the null-cone problem for the 3-tensor action in polynomial time.

5.2 Proof Outline

In the following we briefly sketch the proof ideas and methods used to obtain Theorems 5.1.1 and 5.1.2. This is based on [FR21, Subsections 3.1 and 4.5].

First, we sketch how to construct an array $p \in (\mathbb{R}^m_{\geq 0})^{\otimes 3}$ in the commutative case, Theorem 5.1.1. Recall the formulation of array scaling as a geometric program in Equation (5.1). We build both the support $\operatorname{supp}(p) \subseteq \Omega(\pi_{m,3})$ and the entries of p in [FR21, Section 3] in the following way. We choose a set $\Gamma \subseteq \Omega(\pi_{m,3})$, another weight $\hat{\omega} \in \Omega(\pi_{m,3})$, and an array $q \in (\mathbb{R}^m_{\geq 0})^{\otimes 3}$ such that:

1. The set $\Gamma \subseteq \Omega(\pi_{m,3})$ is the support of an array $q \in (\mathbb{R}^m_{\geq 0})^{\otimes 3}$, and mq is tristochastic, i.e., all slice sums of q are equal to m^{-1}. As a consequence, $q_{+++} = 1$ and $\sum_{\omega \in \Gamma} q_\omega \omega = 0$, showing that $0 \in \operatorname{relint}(\operatorname{conv}(\Gamma))$.[10]

2. The affine hull of Γ, should have codimension one[11] in $\mathbb{R}^{3m}$.

3. The vector $\hat{\omega} \in \Omega(\pi_{m,3})$ is at a very small, positive distance η from $\operatorname{aff}(\Gamma)$. Note that this already implies that the *facet gap*[12] of $\Gamma \cup \{\hat{\omega}\}$ is small.

Finally, we define the entries of p by $p_\omega = \frac{1}{2} q_\omega$ for $\omega \in \Gamma$, $p_{\hat{\omega}} = \frac{1}{2}$, and $p_\omega = 0$ elsewhere. Assuming we have found p according to this process, we now give some intuition for the diameter bound.

[9] Note that Theorem 5.1.2 requires a higher precision, namely $\varepsilon \leq \exp(-Cm^2 \log(m))$.

[10] This also follows from Hilbert-Mumford for the array scaling action and the tristochastic array mq; similar to Corollary 3.1.8.

[11] This will not quite apply in our setting, because $\operatorname{aff}(\Omega(\pi_{m,3}))$ is not full-dimensional. Instead, $\operatorname{aff}(\Gamma)$ will be codimension one in $\operatorname{aff}\big(\Omega(\pi_{m,3})\big)$.

[12] This is a concept from [BLNW20, Definition 1.8]: the *facet gap* of $\Omega \subseteq \mathbb{R}^m$ is the largest constant $C > 0$ such that $\operatorname{dist}(\omega, \operatorname{aff}(F)) \geq C$ for any facet F of $\operatorname{conv}(\Omega)$ and $\omega \in \Omega \backslash F$.

Let v be the projection of $\hat{\omega}$ to the orthogonal complement of $\mathrm{aff}(\Gamma)$. Intuitively, the capacity is only approximately attained by vectors very far in the $-v$ direction. Indeed, first note that $\mathrm{cap}(q) = 1$, by the properties from Item 1 together with the weighted AM-GM inequality. We deduce $\mathrm{cap}(p) = 1/2$, because $\mathrm{cap}(p) \geq \frac{1}{2}\mathrm{cap}(q) = \frac{1}{2}$, and $f_p(-tv/\|v\|) = \frac{1}{2} + e^{-\eta t}$ tends to $\frac{1}{2}$ for $t \to \infty$. However, $f_p(-tv/\|v\|)$ tends to $\frac{1}{2}$ slowly if η is small: $f_p(-tv/\|v\|) \leq \frac{1}{2} + \varepsilon$ if and only if $t \geq -\eta^{-1}\log(\varepsilon) = \eta^{-1}\log(1/\varepsilon)$.

To conclude rigorously that the capacity is only approached by vectors very far in the $-v$ direction, we must rule out directions with non-zero components in $\mathrm{aff}(\Gamma)$. For this, we use in [FR21] the assumption that zero is rather deep in the relative interior of $\mathrm{conv}(\Gamma)$. Then any ε-approximate minimizer must have a bounded component in $\mathrm{aff}(\Gamma)$, for otherwise the contribution to f_p from the elements of Γ alone will be larger than $\frac{1}{2} + \varepsilon$.

Remark 5.2.1 (based on [FR21, Subsubsection 1.1.3]). The structure of the argument bears some similarity to that in [SV19], which uses the construction of [AV97]. The main difference is that the set $\Omega(\pi_{m,3})$ in the 3-dimensional array scaling problem consists of weights of very specific structure: up to an additive shift of $-\frac{1}{m}\mathbb{1}_{3m}$, they are Boolean vectors in $\mathbb{R}^{3m}$ with exactly one non-zero entry among indices in the intervals $[1, m], [m+1, 2m]$ and $[2m+1, 3m]$. Thus, our construction of Γ must consist of weights of this special form and not simply bounded integral vectors as in [SV19]. This is the main additional technical contribution of our construction. ▽

We end the commutative case with a consequence on the facet gap from Item 3.

Corollary 5.2.2 (Facet gap of array scaling, [FR21, Corollary 3.6])**.** *There is a subset of $\Omega(\pi_{m,3})$ with facet gap $O(2^{-m/3})$.*

Similarly to lifting bounds from the weight margin to the gap (Chapter 4), we can lift the diameter bound from the commutative to the non-commutative case, if the construction is free.

Theorem 5.2.3 (based on [FR21, Theorem 4.20])**.** *Let $\pi\colon G \to \mathrm{GL}(V)$ be a representation with assumptions as in Setting 3.0.1. Suppose $\mu_T(t \cdot v) = \mu_G(t \cdot v)$ for all $t \in T$ (which holds if $\mathrm{supp}(v) \subseteq \Omega(\pi)$ is free). Then for any $R > 0$*

$$\inf_{g \in B'_R} \|g \cdot v\|^2 = \inf_{t \in T \cap B'_R} \|t \cdot v\|^2, \tag{5.2}$$

where $B'_R := \{k \exp(X) \mid k \in K, X \in \mathfrak{i}\,\mathrm{Lie}(K), \|X\|_F \leq R\}$.

The above theorem is specifically stated for $\pi_{m,3}$ in [FR21], but the arguments of the proof hold in general. Equation (5.2) ensures that for a free vector v one can always choose an approximate minimizer of $\mathrm{cap}_G(v)$ in T. Since the array from Theorem 5.1.1 has free support [FR21, Lemma 4.21], one can deduce Theorem 5.1.2. However, in the latter we need to choose a tensor v such that $p_{ijk} = |v_{ijk}|^2$, which is not solvable over the rationals. Hence, we need some rounding procedure so that the rationals v_{ijk} satisfy $v_{ijk} \approx \sqrt{p_{ijk}}$. Higher precision, i.e., a smaller ε, requires a more precise rounding. Therefore, the tensor v in Theorem 5.1.2 depends on the precision ε. The technical details of the rounding procedure are treated in [FR21, Lemmas 4.22 and 4.23].

For a full proof of the non-commutative case we refer to [FR21, Section 4.5].

Part III

Algebraic Statistics

Chapter 6

Maximum Likelihood Estimation

The task of parameter estimation is ubiquitous in statistics. That is, given a statistical model and observed data, one seeks the parameters of a probability distribution which "best" explains the data and is contained in the model. There are many different concepts of parameter estimation, see e.g., [Jay03; LC98; Ric06]. In this thesis we focus on the approach of maximum likelihood estimation (ML estimation), which was popularized by Ronald Fisher in the early 20th century. ML estimation is built on an intuitive idea and the ML estimator enjoys several asymptotic properties [Cra46; Vaa98]. As a consequence, it is frequently used in practice [Cra86; Mil11; Sev00; WA18].

This chapter provides the necessary background on ML estimation through the lens of algebraic statistics, and thereby it prepares Chapters 7–10. For further information on ML estimation in the context of algebraic statistics the reader is referred to the textbooks [DSS09; PS05; Sul18].

Organization and Assumptions. Section 6.1 provides a brief, general introduction to ML estimation. Afterwards, this is specified for two widely used classes of models: discrete models in Section 6.2 and Gaussian models in Section 6.3. The former prepares Chapter 7 while the latter is needed in Chapters 8, 9 and 10.

We assume some familiarity with probability theory, e.g., the amount of [Sul18, Chapter 2] certainly suffices.

6.1 Parametric Statistical Models

This general introduction on maximum likelihood (ML) estimation closely follows [Sul18, Chapter 5]. Its purpose is to illustrate that Sections 6.2 and 6.3 follow the same concept. Let us start with the definition of a statistical model, which is fundamental for any theory of parameter estimation.

Definition 6.1.1 (Parametric Statistical Model)**.** A collection

$$\mathcal{P}_{\mathcal{M}} := \{P_\Psi \mid \Psi \in \mathcal{M}\}$$

of probability distributions on a fixed sample space $\mathcal{S}$, parametrized by a set $\mathcal{M} \subseteq \mathbb{R}^d$, is called a *parametric statistical model*. We assume that each P_Ψ admits a density function p_Ψ with respect to a fixed measure ν on $\mathcal{S}$. ▲

The notation of the parameter set $\mathcal{M}$ is suggestive: in Sections 6.2 and 6.3 we directly regard the respective parameter sets as statistical models.[1]

Now, given observed data D, the problem of *parameter estimation* is to determine a joint probability distribution from $\mathcal{P}_\mathcal{M}$ explaining the data D. Intuitively, the idea of ML estimation is to search for the probability distribution in $\mathcal{P}_\mathcal{M}$ under which it is *most likely* to observe the data D. Formally, we always assume that $D = (D_1, \ldots, D_n)$ is a tuple of n samples that are independent identically distributed (i.i.d.) according to some unknown $P_\Psi \in \mathcal{P}_\mathcal{M}$. Then the *likelihood function*, given data D, is

$$L_D \colon \mathcal{M} \to \mathbb{R}, \quad L_D(\Psi) = \prod_{i=1}^{n} p_\Psi(D_i) \tag{6.1}$$

and captures how likely it is to witness the data D under the probability distribution P_Ψ. Often, it is convenient to consider the *log-likelihood function*

$$\ell_D(\Psi) := \log\big(L_D(\Psi)\big) = \sum_{i=1}^{n} \log\big(p_\Psi(Y_i)\big). \tag{6.2}$$

The task of ML estimation is to maximize the (log-)likelihood function.

Definition 6.1.2 (Maximum Likelihood Estimator (MLE))**.** Let $\mathcal{P}_\mathcal{M}$ be a parametric statistical model with observed data D. If $\hat{\Psi} \in \mathcal{M}$ satisfies

$$\ell_D(\hat{\Psi}) = \sup_{\Psi \in \mathcal{M}} \ell_D(\Psi)$$

we call $\hat{\Psi}$ a *maximum likelihood estimator* (MLE) given data D. ▲

The next concept captures how observed data interacts with the parameters of a model.

Definition 6.1.3 (Sufficient Statistics)**.** Let $\mathcal{P}_\mathcal{M}$ be a statistical model. We call a function X a *sufficient statistics* for $\mathcal{P}_\mathcal{M}$, if for any $\Psi \in \mathcal{M}$ and i.i.d. samples $D_1, \ldots, D_n \sim P_\Psi$ the joint density of $D = (D_1, \ldots, D_n)$ can be written as

$$\prod_{i=1}^{n} p_\Psi(D_i) = f(D) g\big(X(D), \Psi\big), \tag{6.3}$$

where f and g are non-negative measurable functions.[2] ▲

Note that the left hand side in Equation (6.3) is $L_D(\Psi)$ and hence the log-likelihood is $\ell_D(\Psi) = \log(f(D)) + \log\big(g(X(D), \Psi)\big)$. We see that ML estimation in a model $\mathcal{P}_\mathcal{M}$ only depends on a sufficient statistics.

Following [BBJJ82, Equation (1.2)] and [Bar83, p. 348], we define a concept involving a group action.

[1] This is justified as the models considered in this thesis are identifiable in the sense that the map $\mathcal{M} \to \mathcal{P}_\mathcal{M},\ \Psi \mapsto P_\Psi$ is bijective.

[2] The identity is a trivial sufficient statistics. However, we are interested in sufficient statistics that yield a proper reduction, i.e., that different data tuples may have the same value under X.

Definition 6.1.4 (Transformation Family)**.** Let $\mathcal{P}_\mathcal{M}$ be a statistical model on a sample space $\mathcal{S}$. We call $\mathcal{P}_\mathcal{M}$ a *transformation family*[3] or *transformation model* if there is a group G, consisting of automorphisms of $\mathcal{S}$, that acts transitively on $\mathcal{P}_\mathcal{M}$ via $(g \cdot P)(A) := P(g^{-1}(A))$, where $P \in \mathcal{P}_\mathcal{M}$ and A is a measurable event. ▲

Finally, we mention some interesting, natural questions that arise when studying ML estimation:

1. Is the log-likelihood ℓ_D bounded from above? Does an MLE given data D exist? If an MLE exists, is it unique?
2. How can we compute an MLE?
3. Which sample sizes n guarantee (almost surely) an affirmative answer to the questions from Item 1?

Interestingly, we see in Chapters 7–10 that we can study these questions for several important models through the lens of invariant theory. As a preparation, we focus on discrete models and Gaussian models in the upcoming two sections.

6.2 Discrete Models

In the following we describe ML estimation for models consisting of discrete probability distributions. The presentation is mainly based on [AKRS21b, Section 2].

We consider the sample space $\mathcal{S} = [m] = \{1, 2, \ldots, m\}$ of m states, which we endow with the counting measure. Then a probability distribution on $\mathcal{S} = [m]$ is uniquely determined by its density[4] $p = (p_1, \ldots, p_m)$, where p_j denotes the probability that the j^{th} state occurs. Such a density is a point in the $(m-1)$-dimensional probability simplex:

$$\Delta_{m-1} := \left\{ p \in \mathbb{R}^m_{\geq 0} \mid p_+ = \sum_{j=1}^{m} p_j = 1 \right\}.$$

Using densities as parameters and identifying a model $\mathcal{P}_\mathcal{M}$ of probability distributions on $\mathcal{S}$ with its parameter set leads to the following.

Definition 6.2.1 (Discrete Model)**.** A *discrete model* $\mathcal{M}$ of distributions with m states is a subset $\mathcal{M} \subseteq \Delta_{m-1}$. ▲

Given a tuple $D = (D_1, \ldots, D_n)$ of i.i.d. samples, the likelihood from (6.1) can be written as $L_D(p) = \prod_j p_j^{u_j}$, where $u_j := \{i \in [n] \mid D_i = j\}$ is the number of times that the j^{th} state occurs. We see that the *vector of counts* $u := (u_1, \ldots, u_m) \in \mathbb{Z}^m_{\geq 0}$ is a sufficient statistic for any discrete model, compare

[3] We caution the reader about ambiguities of the term *transformation family* in the statistics literature. For example, there is a distinct well-studied concept called *power transformation families*, see e.g., [CR81; Sak92] and the references therein. But also Definition 6.1.4 is widely studied, see e.g., [Rei95] and the literature therein.

[4] usually called *probability mass function* in the discrete case

Definition 6.1.3.[5] Therefore, we are allowed to regard the vector of counts u as data for discrete models and will do so from now on. Note that the sample size is recovered via $n = u_+ = \sum_{j=1}^m u_j$. Moreover, a vector of counts induces an *empirical distribution* $\bar{u} = \frac{1}{n}u \in \Delta_{m-1}$.

Now, given a discrete model $\mathcal{M} \subseteq \Delta_{m-1}$ and a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$, the (log-)likelihood function[6] becomes

$$L_u(p) = p_1^{u_1} \cdots p_m^{u_m} \quad \text{respectively} \quad \ell_u(p) = \sum_{i=1}^m u_i \log(p_i). \tag{6.4}$$

We use the convention $0^0 = 1$, so that the likelihood is always defined on Δ_{m-1}. This allows MLEs on the relative boundary of Δ_{m-1}, if some entries of u are zero. Furthermore, following [Lau96, Section 4.2.3] we define the concept of extended models and MLEs, which are used in our study of log-linear models, Chapter 7.

Definition 6.2.2 (Extended MLE)**.** Given a discrete model $\mathcal{M} \subseteq \Delta_{m-1}$ and $u \in \mathbb{Z}^m_{\geq 0}$. The extended model of $\mathcal{M}$ is its Euclidean closure $\overline{\mathcal{M}} \subseteq \Delta_{m-1}$ in $\mathbb{R}^m$. By compactness of $\overline{\mathcal{M}}$ and continuity of the likelihood L_u, $\overline{\mathcal{M}}$ admits an MLE $\hat{p}$ given u, which we call an *extended MLE* of $\mathcal{M}$ given u. ▲

Next, we link the log-likelihood to the *Kullback-Leibler (KL) divergence*. The KL divergence from $q \in \mathbb{R}^m_{\geq 0}$ to $p \in \mathbb{R}^m_{\geq 0}$ is

$$\mathrm{KL}(p\|q) = \sum_{j=1}^m p_j \log \frac{p_j}{q_j}.$$

In view of our convention $0^0 = 1$, we also use $0 \log(0/q_j) = 0$ (even, if q_j is zero). Although the KL divergence is not a metric,[7] for $p, q \in \Delta_{m-1}$ it satisfies $\mathrm{KL}(p\|q) \geq 0$, and $\mathrm{KL}(p\|q) = 0$ if and only if $p = q$.

The log-likelihood (6.4) given u can be written, up to additive constant, as

$$\ell_u(p) - \sum_{j=1}^m \log(\bar{u}_j) = -n \sum_{j=1}^m \bar{u}_j \log \frac{\bar{u}_j}{p_j} = -n \ \mathrm{KL}(\bar{u}\|p). \tag{6.5}$$

Therefore, maximizing the log-likelihood is equivalent to minimizing the KL divergence to the empirical distribution $\bar{u}$. In particular, an MLE $\hat{p}$ given u is a point that minimizes, over the model $\mathcal{M}$, the KL divergence to the empirical distribution $\bar{u}$. We use this viewpoint in Section 7.3.

We end this section with two examples of discrete models.

Example 6.2.3 (Saturated discrete model)**.** Consider the model $\mathcal{M} = \Delta_{m-1}$ and a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$. There is a unique MLE $\hat{p}$ given u. By the mentioned properties of the KL-divergence, it is the empirical distribution: $\hat{p} = \bar{u}$. ♢

[5] Here, choose $f \equiv 1$ and $g(p, u) := \prod_j p_j^{u_j}$.

[6] Strictly speaking we would have to multiply the right hand side of (6.4) with the multinomial coefficient $\binom{n}{u}$. However, this does not change the MLE or any other interesting properties of ML estimation.

[7] The KL divergence is *not* symmetric.

Example 6.2.4 (Independence Model)**.** Consider $\Delta_{m_1 m_2 - 1} \subseteq \mathbb{R}^{m_1 \times m_2}$. Then

$$\begin{aligned}\mathcal{M}_{X \perp\!\!\!\perp Y} &= \left\{\alpha^{\mathsf{T}}\beta \mid \alpha \in \Delta_{m_1-1},\, \beta \in \Delta_{m_2-1}\right\} \\ &= \left\{p = (p_{ij}) \in \Delta_{m_1 m_2 - 1} \mid \operatorname{rank}(p) = 1\right\}\end{aligned}$$

is the model of independence of two discrete random variables with m_1 respectively m_2 states. Note that given $p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$, one finds $\alpha = (p_{i,+})_i$ and $\beta = (p_{+,j})_j$ as the marginal distributions.

Let $u \in \mathbb{Z}_{\geq 0}^{m_1 \times m_2}$ be a table of counts obtained from $n = u_{++}$ i.i.d. samples. Then there is a unique MLE $\hat{p}$ given u. It is determined by the table marginals: $\hat{p}_{ij} = u_{i,+}u_{+,j}/n^2$, see [Sul18, Proposition 5.3.8]. We recover this knowledge in Example 7.2.5 using the theory of Chapter 7. ♢

Further important examples are discrete graphical models [Lau96; Sul18] and log-linear models [DSS09; Sul18]. We study the latter class in Chapter 7.

6.3 Gaussian Models

In this section we study ML estimation for Gaussian models. We focus on the necessary prerequisites for Chapters 8–10. In particular, we define maximum likelihood thresholds, consider several examples of Gaussian models and study ML estimation for models given by a directed acyclic graph in detail. The presentation is based on [AKRS21a; MRS21].

We work in parallel over the real and complex numbers: $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. The cone of symmetric respectively Hermitian positive definite matrices is denoted $\mathrm{PD}_m(\mathbb{R})$ respectively $\mathrm{PD}_m(\mathbb{C})$. Recall that $(\cdot)^\dagger$ denotes the Hermitian transpose, which is just the transpose $(\cdot)^{\mathsf{T}}$ if $\mathbb{K} = \mathbb{R}$. We note that complex Gaussian models have been studied in [AHSE95; Goo63] and they are especially interesting for physics applications. Moreover, when relating ML estimation to invariant theory in Chapters 9 and 10 it is natural from the invariant theory perspective to consider complex Gaussian models.

Let us start by recalling the multivariate Gaussian distribution. Consider the sample space $\mathcal{S} = \mathbb{K}^m$ endowed with the Lebesgue measure. We denote by $\mathcal{N}_m(b, \Sigma)$ the m-dimensional multivariate Gaussian distribution[8] with mean $b \in \mathbb{K}^m$ and covariance matrix $\Sigma \in \mathrm{PD}_m(\mathbb{K})$. Its density at $y \in \mathbb{K}^m$ is

$$p_\Sigma(y) = \begin{cases} \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-b)^\dagger \Sigma^{-1}(y-b)\right) & \text{if} \quad \mathbb{K} = \mathbb{R} \\ \det(\pi\Sigma)^{-1} \exp\left(-\frac{1}{2}(y-b)^\dagger \Sigma^{-1}(y-b)\right) & \text{if} \quad \mathbb{K} = \mathbb{C} \end{cases} \tag{6.6}$$

compare [Woo56] or [Goo63, Theorem 3.1] for the complex case. The Gaussian distribution enjoys many nice properties. We shall need the following later on.

Lemma 6.3.1. *If $Y \sim \mathcal{N}_m(b, \Sigma)$ and $g \in \mathrm{GL}_m(\mathbb{K})$, then $gY \sim \mathcal{N}_m(gb, g\Sigma g^\dagger)$. In particular, gY has concentration matrix $(g^\dagger)^{-1}\Sigma^{-1}g^{-1}$.*

Since the exponential and determinant expression in Equation (6.6) involve the inverse of Σ, it is more convenient to work with the *concentration matrix*[9]

[8] Often we drop the index m, if the dimension is clear from the context.

[9] also called *precision matrix*

$\Psi = \Sigma^{-1}$. We follow the latter approach in this thesis. Furthermore, we restrict to mean zero Gaussians, which is justified by Remark 6.3.7 below.

Definition 6.3.2 (Gaussian Model)**.** A *Gaussian model* is a subset $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$, which contains the *concentration matrices* of the respective m-dimensional multivariate Gaussian distributions of *mean zero*. ▲

Example 6.3.3 (Independent univariate Gaussians)**.** The model

$$\mathcal{M} = \big\{ \Psi \in \mathrm{PD}_m(\mathbb{K}) \mid \Psi \text{ is diagonal} \big\} = \big\{ \operatorname{diag}(d_1, \ldots, d_m) \mid d_i \in \mathbb{R}_{>0} \big\}$$

consists of all tuples of m independent univariate Gaussians. ♢

Now, we turn to ML estimation in a Gaussian model $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$. Given a tuple $Y = (Y_1, \ldots, Y_n) \in (\mathbb{K}^m)^n$ of i.i.d. samples, the likelihood function (6.1) at the concentration matrix $\Psi \in \mathcal{M}$ is, up to a scalar factor,

$$L_Y(\Psi) = \prod_{i=1}^{n} p_{\Psi^{-1}}(Y_i) = \big(\det(\Psi)^n\big)^{c(\mathbb{K})} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} Y_i^\dagger \Psi Y_i\right), \tag{6.7}$$

where $c(\mathbb{R}) = 1/2$ and $c(\mathbb{C}) = 1$. Hence, the log-likelihood function can be written, up to additive and positive multiplicative constants, for both $\mathbb{R}$ and $\mathbb{C}$ as

$$\ell_Y(\Psi) = \log\det(\Psi) - \operatorname{tr}(\Psi S_Y), \quad \text{where } S_Y := \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^\dagger \tag{6.8}$$

is the *sample covariance matrix*, an $m \times m$ positive semi-definite matrix. Equations (6.7) and (6.8) both show that the sample covariance matrix gives rise to a sufficient statistics of $\mathcal{M}$, compare Definition 6.1.3. We point out that we view the samples Y_i as column vectors and this canonically identifies Y as a matrix in $\mathbb{K}^{m\times n} \cong (\mathbb{K}^m)^n$. There is no harm in switching between these identifications, and we often do so implicitly.

Remark 6.3.4. One may consider the concept of an extended MLE for Gaussian models, similarly to the discrete case in Definition 6.2.2. However, we note that the Gaussian models considered in this thesis are already Euclidean closed in $\mathrm{PD}_m(\mathbb{K})$. Furthermore, taking the closure in the cone of positive *semi*-definite matrices does not add anything: the supremum of the likelihood cannot be attained at some rank deficient Ψ as then $L_Y(\Psi) = 0$ by Equation (6.7). ▽

Next, we recall maximum likelihood thresholds for Gaussian models.

Definition 6.3.5 (ML Thresholds)**.** Let $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$ be a Gaussian model. We define three maximum likelihood thresholds (ML thresholds).

(i) $\mathrm{mlt}_b(\mathcal{M})$ is the smallest integer n_0, such that for any $n \geq n_0$ the log-likelihood ℓ_Y is bounded from above for almost all[10] $Y \in (\mathbb{K}^m)^n$.

(ii) $\mathrm{mlt}_e(\mathcal{M})$ is the smallest integer n_0, such that for any $n \geq n_0$ an MLE given $Y \in (\mathbb{K}^m)^n$ almost surely exists.

[10] with respect to the Lebesgue measure

(iii) $\mathrm{mlt}_u(\mathcal{M})$ is the smallest integer n_0, such that for any $n \geq n_0$ there exists almost surely a *unique* MLE given $Y \in (\mathbb{K}^m)^n$.

If such an integer n_0 does not exist, we define the respective threshold to be infinity.[11] Note that $\mathrm{mlt}_b(\mathcal{M}) \leq \mathrm{mlt}_e(\mathcal{M}) \leq \mathrm{mlt}_u(\mathcal{M})$. ▲

The above definition matches those in [DKH21; DM21; DMW22]. We see that ML thresholds provide an answer to Question 3 raised on page 125, which concerns sample sizes that (almost surely) guarantee certain properties of the log-likelihood.

Remark 6.3.6. Consider a Gaussian model $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$.

(a) In algebraic settings the desired properties for ML thresholds often hold for *generic* $Y \in (\mathbb{K}^m)^n$ in the sense of algebraic geometry. That is, a generic property holds for all $Y \in (\mathbb{K}^m)^n \backslash Z$ where $Z \subsetneq (\mathbb{K}^m)^n$ is a subvariety of codimension at least one. As a lower dimensional Zariski closed set of $(\mathbb{K}^m)^n$ has Lebesgue measure zero, a generic property also holds almost surely.

(b) If ℓ_Y is bounded from above on $\mathrm{PD}_m(\mathbb{K})$, then it is also bounded on $\mathcal{M}$. A posteriori, we have $\mathrm{mlt}_b(\mathcal{M}) \leq \mathrm{mlt}_b(\mathrm{PD}_m(\mathbb{K})) = m$, see Example 6.3.8. Moreover, for $n \geq m$ the sample covariance matrix S_Y from (6.8) runs through *all* positive semidefinite $m \times m$ matrices, and S_Y is invertible for generic Y. Since S_Y is a sufficient statistics we conclude that either $\mathrm{mlt}_e(\mathcal{M}) \leq m$ or $\mathrm{mlt}_e(\mathcal{M}) = \infty$. This also applies to $\mathrm{mlt}_u(\mathcal{M})$. ▽

Before exploring some examples of Gaussian models we comment on the consequences of our mean zero assumption.

Remark 6.3.7 (Mean Zero Assumption). We stress that we always assume the mean to be *known* and equal to zero. If one allows arbitrary means $b \in \mathbb{K}^m$, then a Gaussian model is a subset of $\mathbb{K}^m \times \mathrm{PD}_m(\mathbb{K})$. The *sample mean* and *sample covariance matrix* for samples $Y_1, \ldots, Y_n$ are

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{and} \qquad \tilde{S}_Y = \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \bar{Y}\big)\big(Y_i - \bar{Y}\big)^\dagger. \tag{6.9}$$

They are a sufficient statistics for any Gaussian model, [And03, Theorem 3.4.1], and give the MLE of the saturated model $\mathbb{K}^m \times \mathrm{PD}_m(\mathbb{K})$, [Sul18, Proposition 5.3.7].

Now, consider $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$. Then the model $\mathbb{K}^m \times \mathcal{M}$ *always* has $\bar{Y}$ as the MLE for the mean parameter [Sul18, Proposition 7.1.9]. Moreover, for all three ML thresholds we have

$$\mathrm{mlt}\left(\mathbb{K}^m \times \mathcal{M}\right) = \mathrm{mlt}_0(\mathcal{M}) + 1, \tag{6.10}$$

where mlt_0 stresses that the mean is known to be zero, compare [DKH21, Remark 1.1].[12] The latter has to be kept in mind whenever consulting results in the literature that deal with *arbitrary* means. ▽

[11] All models in this thesis have finite thresholds. However, it might be that there exists a model $\mathcal{M}$ with an infinite threshold. A posteriori, we have at least Remark 6.3.6(b).

[12] According to [DKH21, Remark 1.1] this follows implicitly from classical results in [And03, Section 3.3].

Proof of Equation (6.10). The proof idea is thanks to Carlos Améndola. For $Y \in \mathbb{K}^{m\times n}$ let $\bar{Y}$ and $\tilde{S}_Y$ be as in (6.9), and S_Y as in (6.8). If we consider ML estimation for the model $\mathbb{K}^m \times \mathcal{M}$, we have to maximize

$$\tilde{\ell}_Y\big(\bar{Y}, \Psi\big) = \log\det(\Psi) - \operatorname{tr}\big(\Psi \tilde{S}_Y\big), \tag{6.11}$$

where we stress once more that $\bar{Y}$ is always the MLE for the mean parameter. Now, the crucial observations are the following. First, Equation (6.11) for $\tilde{\ell}_Y(\bar{Y}, \cdot)$ equals the one for $\ell_Y(\Psi)$ in (6.8), except that S_Y is replaced by $\tilde{S}_Y$. Second, the properties for ML estimation of $\tilde{\ell}_Y(\bar{Y}, \cdot)$ only depend on $\tilde{S}_Y$, similarly for $\ell_Y(\cdot)$ they only depend on S_Y. Third, we have the algebraic subsets of $\overline{\mathrm{PD}_m(\mathbb{K})}$ (cone of positive *semi*definite matrices)

$$\begin{aligned}\big\{\tilde{S}_Y \mid Y \in \mathbb{K}^{m\times n}\big\} &= \big\{S \in \overline{\mathrm{PD}_m(\mathbb{K})} \mid \operatorname{rank}(S) \leq \min\{m,n\} - 1\big\}\\ \big\{S_Y \mid Y \in \mathbb{K}^{m\times n}\big\} &= \big\{S \in \overline{\mathrm{PD}_m(\mathbb{K})} \mid \operatorname{rank}(S) \leq \min\{m,n\}\big\};\end{aligned}$$

and the generic rank[13] of $\tilde{S}_Y$ equals $\min\{m,n\} - 1$, while the generic rank of S_Y is $\min\{m,n\}$. Altogether, we must have Equation (6.10). □

In the following we present several important examples of Gaussian models.

Example 6.3.8 (Saturated Gaussian model)**.** Let $Y \in \mathbb{K}^{m\times n}$ be a sample matrix for the *saturated Gaussian model* $\mathcal{M} = \mathrm{PD}_m(\mathbb{K})$. The following is well-known, see e.g., [Lau96, Theorem 5.1] or [Sul18, Proposition 5.3.7]. The unique maximizer of ℓ_Y over $\mathrm{PD}_m(\mathbb{K})$ is $\hat{\Psi} = S_Y^{-1}$, if the sample covariance matrix S_Y is invertible. If S_Y is not invertible, the likelihood function is unbounded and the MLE does not exist. One verifies that S_Y is invertible if and only if $Y = \mathbb{K}^{m\times n}$ has full row rank. The latter cannot hold if $m > n$, and it holds generically if $m \leq n$.[14] Altogether, we deduce

$$\mathrm{mlt}_b\big(\mathrm{PD}_m(\mathbb{K})\big) = \mathrm{mlt}_e\big(\mathrm{PD}_m(\mathbb{K})\big) = \mathrm{mlt}_u\big(\mathrm{PD}_m(\mathbb{K})\big) = m.$$

We recover these facts in Examples 9.3.8 and 9.5.11 using the theory developed in Chapter 9. ♢

Example 6.3.9 (Matrix and Tensor Normal Models)**.** If one samples matrices $\mathbb{K}^{m_1\times m_2} \cong \mathbb{K}^{m_1 m_2}$, or more generally tensors $\mathbb{K}^{m_1} \otimes \cdots \otimes \mathbb{K}^{m_d} \cong \mathbb{K}^{m_1\cdots m_d}$, then the saturated model $\mathrm{PD}_{m_1\cdots m_d}(\mathbb{K})$ is huge and one needs at least $m_1 \cdots m_d$ many samples for an MLE to exist (almost surely), compare Example 6.3.8. To decrease the ML threshold one can presume structural assumptions on the model. A common approach is to consider the *tensor normal model*

$$\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1,\ldots,m_d) := \{\Psi_1 \otimes \cdots \otimes \Psi_d \mid \Psi_i \in \mathrm{PD}_{m_i}(\mathbb{K})\} \subseteq \mathrm{PD}_{m_1\cdots m_d}(\mathbb{K}), \tag{6.12}$$

where $\otimes$ denotes the Kronecker product of matrices, see Definition 1.3.4. For $d = 2$ the model $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$ is called the *matrix normal model*, which we study in further detail in Section 9.4.

[13] i.e., the rank that is attained for generic $Y \in \mathbb{K}^{m\times n}$

[14] If $m \leq n$ then full row rank holds outside the vanishing locus of the maximal minors of Y.

Recently, there has been a flurry of new results on ML estimation. For matrix normal models, the paper [DKH21] gave new characterizations of ML estimation and new bounds on ML thresholds. By crucially using the relations between invariant theory and ML estimation presented in Section 9.3, [DM21] and [DMW22] completely characterized all ML thresholds for matrix respectively tensor normal models. Furthermore, [FORW21] provide results on almost optimal sample complexity in tensor normal models. ♢

Example 6.3.10 (Undirected Gaussian graphical model)**.** Let $\mathcal{G} = (I, E)$ be an undirected graph with vertex set $I = [m]$. Then

$$\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}} := \left\{ \Psi \in \mathrm{PD}_m(\mathbb{K}) \mid \Psi_{ij} = \Psi_{ji} = 0 \ \text{ whenever } \ (i \text{ — } j) \notin E \right\}$$

is the *undirected Gaussian graphical model*[15] given by $\mathcal{G}$. In words, the undirected edges describe the off-diagonal support pattern of the concentration matrices in the model. Statistically, if $X \sim \mathcal{N}_m(0, \Sigma)$ then for the concentration matrix $\Psi = \Sigma^{-1}$ the condition $\Psi_{ij} = 0$ is equivalent the conditional independence $X_i \perp\!\!\!\perp X_j | X_{[m]\setminus\{i,j\}}$, [Lau96, Proposition 5.2] or [Sul18, Proposition 6.3.2]. This generalizes for distributions in $\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}}$ via so-called Markov properties[16] given by the undirected graph $\mathcal{G}$, see [Lau96], and [Sul18, Chapter 13] for details.

Regarding ML estimation, it is well-known that $\mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}}) = \mathrm{mlt}_u(\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}})$ and a unique MLE exists if the sample covariance matrix S_Y is invertible. In this case, the MLE $\hat{\Psi} \in \mathcal{M}_{\mathcal{G}}^{\mathrm{ud}}$ is given by $\hat{\Psi}_{ij} = (S_Y^{-1})_{ij}$ whenever $i = j$ or $(i \text{ — } j) \in E$ and $\hat{\Psi}_{ij} = 0$ otherwise, see [Lau96, Theorem 5.3]. In particular, $\mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}}) \leq m$. However, in general $\mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}})$ can be strictly smaller. We refer to [BS19; Buh93; GS18; Uhl12] for further results on ML thresholds.

For applications and further details on undirected Gaussian graphical models we refer to [Lau96; Sul18] and for the complex case to [AHSE95]. ♢

Extended Example: DAG models

In the following we introduce Gaussian graphical models given by directed acyclic graphs and study ML estimation for these models. This prepares our studies in Section 9.5 and Chapter 10. In particular, we will generalize Theorem 6.3.16 to the setting of so-called RDAG models, Theorem 10.3.6. The presentation closely follows [MRS21] and [AKRS21a, Section 5].

A *directed graph* is a tuple $\mathcal{G} = (I, E)$, where I is a finite set of vertices and $E \subseteq I \times I$ is a set of directed edges. Here $(j, i) \in E$ means that $\mathcal{G}$ has a directed edge starting at vertex j and pointing towards i. Instead of $(j, i) \in E$ we usually write $j \to i$ and similarly $j \not\to i$ means $(j, i) \notin E$. Note that, if not specified otherwise, the vertex set I of $\mathcal{G}$ is $[m] = \{1, 2, \ldots, m\}$.

A directed graph $\mathcal{G} = (I, E)$ is called *acyclic*, if $\mathcal{G}$ does not contain any cycle, i.e., $\mathcal{G}$ does not contain a directed path $i_0 \longrightarrow i_1 \longrightarrow \cdots \longrightarrow i_k$ with $i_0 = i_k$. In particular, $\mathcal{G}$ does not contain any loop: $i \not\to i$ for all $i \in I$. From now on

[15] also called *covariance selection model*

[16] We remark that pairwise, local and global Markov property are equivalent for multivariate Gaussians [Sul18, Section 13.1].

we abbreviate *directed acyclic graph* to *DAG*. The set of *parents* and the set of *children* of a vertex i are, respectively,

$$\mathrm{pa}(i) := \{j \in I \mid j \to i \text{ in } \mathcal{G}\} \qquad \text{and} \qquad \mathrm{ch}(i) := \{k \in I \mid i \to k \text{ in } \mathcal{G}\}.$$

Definition 6.3.11 (DAG model)**.** A *DAG model* $\mathcal{M}_{\mathcal{G}}^{\to}$ given by a DAG $\mathcal{G}$ is a Gaussian model defined by the linear structural equation

$$y = \Lambda y + \varepsilon, \qquad \text{i.e.,} \qquad y_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ij} y_j + \varepsilon_i, \tag{6.13}$$

where $y \in \mathbb{K}^m$, $\lambda_{ij} = 0$ for $j \not\to i$ in $\mathcal{G}$, and $\varepsilon \sim \mathcal{N}(0, \Omega)$ with $\Omega \in \mathrm{PD}_m(\mathbb{K})$ diagonal. Since $\mathcal{G}$ is acyclic, the matrix $\Lambda \in \mathbb{K}^{m \times m}$ is nilpotent and hence $(\mathrm{I}_m - \Lambda)$ is invertible. Solving Equation (6.13) for y gives $y = (\mathrm{I}_m - \Lambda)^{-1}\varepsilon$. By Lemma 6.3.1, y is multivariate Gaussian with mean zero and concentration matrix

$$\Psi = (\mathrm{I}_m - \Lambda)^{\dagger} \Omega^{-1} (\mathrm{I}_m - \Lambda), \tag{6.14}$$

i.e., $\mathcal{M}_{\mathcal{G}}^{\to} \subseteq \mathrm{PD}_m(\mathbb{K})$ is the set of all concentration matrices of this form. ▲

The coefficient λ_{ij} is a *regression coefficient*, the effect of parent j on child i. Similarly to Example 6.3.10, the model $\mathcal{M}_{\mathcal{G}}^{\to}$ encodes conditional independence: a node is independent of its non-descendants after conditioning on its parents, see [Sul18, Chapter 13] or [VP90].

We note that DAG models are also called *Gaussian Bayesian networks* and they are a special case of linear structural equation models [Drt18], [Sul18, Section 16.2]. DAG models have been applied to cell signalling [SPP+05], gene interactions [FLNP00], causal inference [Pea09], and many other contexts.

Remark 6.3.12 (based on [MRS21, Remark 1.2])*.* Throughout this thesis, we choose an ordering on the vertices of $\mathcal{G}$ so that Λ is strictly upper triangular. That is, if $j \to i$ is an edge in $\mathcal{G}$ then $j > i$. Such an ordering is possible as $\mathcal{G}$ is acyclic. Thinking of a vertex label as its age, the ordering ensures that parents are older than their children. ▽

Next, we relate undirected models from Example 6.3.10 to DAG models. For this, we need the following definition.

Definition 6.3.13 (Unshielded collider)**.** An *unshielded collider* of a directed graph $\mathcal{G}$ is a subgraph $j \to i \leftarrow k$ with *no* edge between j and k. ▲

Given a DAG $\mathcal{G}$, we denote by $\mathcal{G}^u$ the corresponding undirected graph, which is obtained by forgetting the direction of each edge in $\mathcal{G}$. The following theorem is the Gaussian special case of [AMP97, Theorem 3.1] respectively [Fry90, Theorem 5.6]. We give a proof in Section 10.2.

Theorem 6.3.14 ([MRS21, Theorem 3.7])**.** *Let $\mathcal{G}$ be a DAG. The DAG model $\mathcal{M}_{\mathcal{G}}^{\to}$ is equal to the undirected Gaussian graphical model $\mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$ on $\mathcal{G}^u$ if and only if $\mathcal{G}$ has no unshielded colliders.*

Now, we characterize ML estimation for DAG models. To do so, we prove a lemma that will also be used in Chapter 10.

Lemma 6.3.15 ([MRS21, Lemma 4.10]). *Fix $\alpha > 0$ and, for $\gamma \geq 0$, consider the family of functions*

$$f_\gamma \colon \mathbb{R}_{>0} \to \mathbb{R}, \quad x \mapsto \alpha \log(x) + \frac{\gamma}{x}.$$

(i) If $\gamma = 0$, then f_γ is neither bounded from below nor bounded from above.

(ii) If $\gamma > 0$, then f_γ attains a global minimum at $x_0 = \frac{\gamma}{\alpha}$ with function value $f_\gamma(\frac{\gamma}{\alpha}) = \alpha(\log(\gamma) - \log(\alpha) + 1)$.

(iii) Given $\gamma_1 \geq \gamma_2 > 0$, we have $f_{\gamma_1}(\frac{\gamma_1}{\alpha}) \geq f_{\gamma_2}(\frac{\gamma_2}{\alpha})$ at the global minima.

Proof. Part (i) follows from the properties of the logarithm. To prove part (ii), one computes $f'_\gamma(x) = \frac{\alpha}{x} - \frac{\gamma}{x^2}$ for $x > 0$. For $x > 0$ we have

$$f'_\gamma(x) = 0 \quad \Leftrightarrow \quad \frac{\alpha}{x} = \frac{\gamma}{x^2} \quad \Leftrightarrow \quad \alpha x = \gamma \quad \Leftrightarrow \quad x = \frac{\gamma}{\alpha}.$$

Thus $x_0 := \frac{\gamma}{\alpha}$ is the only possible local extremum of f_γ. For $x > 0$,

$$f'_\gamma(x) > 0 \quad \Leftrightarrow \quad \frac{\alpha}{x} > \frac{\gamma}{x^2} \quad \Leftrightarrow \quad \alpha x > \gamma \quad \Leftrightarrow \quad x > \frac{\gamma}{\alpha}.$$

and similarly one has $f'_\gamma(x) < 0$ if and only if $x < \frac{\gamma}{\alpha} = x_0$. Therefore, x_0 is a global minimum of f_γ. One directly verifies the function value for $f_\gamma(x_0)$, and so part (iii) follows from the monotonicity of the logarithm. □

Now, we characterize ML estimation for DAG models via linear independence conditions on the sample matrix. Let $\mathcal{G}$ be a DAG with vertex set $I = [m]$ and let $Y \in \mathbb{K}^{m \times n}$ be a sample matrix, encoding the n samples which are the columns $Y_1, \ldots, Y_n$ of Y. For $i \in [m]$ we denote by $Y^{(i)}$ the i^{th} row of Y, by $Y^{(\mathrm{pa}(i))}$ the sub-matrix of Y with rows indexed by the parents of i in $\mathcal{G}$, and by $Y^{(i \cup \mathrm{pa}(i))}$ the sub-matrix of Y with rows indexed by vertex i and its parents.

Let us compute the log-likelihood ℓ_Y at some $\Psi \in \mathcal{M}_\mathcal{G}^{\rightarrow}$. To do so, write $\Psi = (\mathrm{I}_m - \Lambda)^\dagger \Omega^{-1} (\mathrm{I}_m - \Lambda)$ as in (6.14). We denote the entries of Ω by ω_{ii} and those of Λ by λ_{ij}. First, note that $\det(\mathrm{I}_m - \Lambda) = 1$ and hence $\log(\det(\Psi)) = -\log(\det(\Omega))$. Moreover, since $\Omega^{-1} \in \mathrm{PD}_m(\mathbb{K})$ we can consider its square root $\Omega^{-1/2} \in \mathrm{PD}_m(\mathbb{K})$. Setting $A := \Omega^{-1/2}(\mathrm{I}_m - \Lambda)$, we have $\Psi = A^\dagger A$ and

$$\begin{aligned} \mathrm{tr}(\Psi S_Y) &= \frac{1}{n} \sum_{j=1}^{n} \mathrm{tr}\left(\Psi Y_j Y_j^\dagger\right) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{tr}\left((AY_j)(AY_j)^\dagger\right) = \frac{1}{n} \|AY\|^2 \\ &= \frac{1}{n} \|\Omega^{-1/2}(\mathrm{I}_m - \Lambda) Y\|^2 = \frac{1}{n} \sum_{i=1}^{m} \left\| \omega_{ii}^{-1/2} \Big(Y^{(i)} - \sum_{j \in \mathrm{pa}(i)} \lambda_{ij} Y^{(j)} \Big) \right\|^2. \end{aligned}$$

Altogether, with Equation (6.8) we conclude that for $\Psi \in \mathcal{M}_\mathcal{G}^{\rightarrow}$

$$\ell_Y(\Psi) = -\sum_{i=1}^{m} \left(\log \omega_{ii} + \frac{1}{n\omega_{ii}} \Big\| Y^{(i)} - \sum_{j \in \mathrm{pa}(i)} \lambda_{ij} Y^{(j)} \Big\|^2 \right). \tag{6.15}$$

The next result follows from this equation, which views ML estimation for a DAG model as a collection of several uncoupled regression problems. Although there does not seem to be a classical reference for this result, it is very likely known to experts and contained implicitly in the literature.

Theorem 6.3.16 ([MRS21, Theorem 4.9])**.** *Consider the DAG model on $\mathcal{G}$, with m nodes, and fix a sample matrix $Y \in \mathbb{K}^{m\times n}$. The following possibilities characterize maximum likelihood estimation given Y:*

$$\begin{array}{lccl}
(a) & \ell_Y \text{ unbounded from above} & \Leftrightarrow & \exists\, i \in [m]\colon\ Y^{(i)} \in \operatorname{span}\left\{Y^{(j)} : j \in \operatorname{pa}(i)\right\} \\
(b) & \text{MLE exists} & \Leftrightarrow & \forall\, i \in [m]\colon\ Y^{(i)} \notin \operatorname{span}\left\{Y^{(j)} : j \in \operatorname{pa}(i)\right\} \\
(c) & \text{MLE exists uniquely} & \Leftrightarrow & \forall\, i \in [m]\colon\ Y^{(i\cup\operatorname{pa}(i))} \text{ has full row rank.}
\end{array}$$

Remark 6.3.17 (based on [AKRS21a, Remark 5.4]). We use the convention that the linear hull of the empty set is the zero vector space. So if a vertex i does not have parents in $\mathcal{G}$, then $Y^{(i)} \notin \operatorname{span}\left\{Y^{(j)} : j \in \operatorname{pa}(i)\right\}$ translates to $Y^{(i)} \neq 0$. ▽

Proof of Theorem 6.3.16. We use the notation that was introduced to obtain Equation (6.15) for $\ell_Y(\Psi)$. Note that each of the entries ω_{ii} and λ_{ij} appears in exactly one of the m summands in (6.15). Thus, to maximize the log-likelihood, or equivalently, to minimize the negative log-likelihood, we can minimize each summand

$$\log \omega_{ii} + \frac{1}{n\omega_{ii}}\Big\|Y^{(i)} - \sum_{j\in\operatorname{pa}(i)} \lambda_{ij} Y^{(j)}\Big\|^2 \tag{6.16}$$

for $i \in [m]$ independently. We can first determine $\hat{\lambda}_{ij} \in \mathbb{K}$ with

$$\zeta_i := \Big\|Y^{(i)} - \sum_{j\in\operatorname{pa}(i)} \hat{\lambda}_{ij} Y^{(j)}\Big\|^2 = \inf_{\lambda_{ij}\in\mathbb{K}} \Big\|Y^{(i)} - \sum_{j\in\operatorname{pa}(i)} \lambda_{ij} Y^{(j)}\Big\|^2.$$

Such $\hat{\lambda}_{ij}$ always exist and are determined by

$$P_i = \sum_{j\in\operatorname{pa}(i)} \hat{\lambda}_{ij} Y^{(j)},$$

where P_i is the orthogonal projection of $Y^{(i)}$ onto $\operatorname{span}\{Y^{(j)} \mid j \in \operatorname{pa}(i)\}$. Note that the $\hat{\lambda}_{ij}$, $j \in \operatorname{pa}(i)$ are unique if and only if $Y^{(\operatorname{pa}(i))}$ has full row rank. To finish the proof we apply Lemma 6.3.15 with $\alpha = 1$ and $\gamma = \zeta_i/n$ several times.

Let $Y^{(i)} \in \operatorname{span}\{Y^{(j)} \mid j \in \operatorname{pa}(i)\}$ for some $i \in [m]$, i.e., $\zeta_i = 0$. Then the summand (6.16) is not bounded from below, see Lemma 6.3.15(i). Hence, setting $\omega_{kk} = 1$ and $\lambda_{k,l} = 0$ for all $k \in [m]\backslash\{i\}$ and all $l \in \operatorname{pa}(k)$ we see that $-\ell_Y$ is not bounded from below. This proves "$\Leftarrow$" of (a).

If $Y^{(i)} \notin \operatorname{span}\{Y^{(j)} \mid j \in \operatorname{pa}(i)\}$, i.e., $\zeta_i > 0$, then $\log(\omega_{ii}) + \zeta_i/(n\omega_{ii})$ has a unique minimizer $\hat{\omega}_{ii} = \zeta_i/n$, compare Lemma 6.3.15(ii). Thus, an MLE given by $\hat{\omega}_{ii}$ and $\hat{\lambda}_{ij}$ exists if $Y^{(i)} \notin \operatorname{span}\{Y^{(j)} \mid j \in \operatorname{pa}(i)\}$ for all $i \in [m]$. This shows "$\Leftarrow$" of (b) and hence all of parts (a) and (b) as their right hand sides are opposites and since MLE existence implies ℓ_Y is bounded from above.

Since the $\hat{\omega}_{ii}$ are uniquely determined (if they exist), an MLE is unique if and only if all $\hat{\lambda}_{ij}$ are unique. We have seen that the latter holds if and only if $Y^{(\operatorname{pa}(i))}$ has full row rank for all $i \in [m]$. In combination with part (b) we deduce (c). □

The above theorem will be generalized to so-called RDAG models, see Theorem 10.3.6. Let us shortly illustrate Theorem 6.3.16 and Remark 6.3.17.

Example 6.3.18. Let $\mathcal{G}$ be the DAG $2 \longrightarrow 1 \longleftarrow 3$ and consider a sample matrix $Y \in \mathbb{K}^{3\times n}$. By Theorem 6.3.16(b), there exists an MLE given Y if and only if $Y^{(2)}, Y^{(3)} \neq 0$ and $Y^{(1)} \notin \operatorname{span}\{Y^{(2)}, Y^{(3)}\}$. Otherwise, the log-likelihood ℓ_Y is not bounded from above. Since $Y = Y^{(1\cup \operatorname{pa}(1))}$ we have that there exists a unique MLE given Y if and only if Y has full row rank, compare Theorem 6.3.16(c). $\diamondsuit$

We use Theorem 6.3.16 to determine the ML thresholds of a DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$. The result is known in the graphical models literature, see [Lau96, Section 5.4.1] and [DFKP19, Theorem 1].

Corollary 6.3.19. *For the model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ of a DAG $\mathcal{G}$, we have*

$$\operatorname{mlt}_b\left(\mathcal{M}_{\mathcal{G}}^{\rightarrow}\right) = \operatorname{mlt}_e\left(\mathcal{M}_{\mathcal{G}}^{\rightarrow}\right) = \operatorname{mlt}_u\left(\mathcal{M}_{\mathcal{G}}^{\rightarrow}\right) = 1 + \max_{i\in[m]} |\operatorname{pa}(i)|.$$

Proof. First, assume there is some vertex $i \in [m]$ with $n < 1 + |\operatorname{pa}(i)|$. Then, for a generic $Y \in \mathbb{K}^{m\times n}$ the parent rows $Y^{(j)}$, $j \in \operatorname{pa}(i)$ span $\mathbb{K}^{1\times n}$ as $n \leq |\operatorname{pa}(i)|$. Thus, $Y^{(i)}$ is in the linear span of the $Y^{(j)}$, $j \in \operatorname{pa}(i)$ for generic Y, so ℓ_Y is not bounded from above for generic Y, by Theorem 6.3.16(a). Hence, we have shown

$$\operatorname{mlt}_b(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) \geq 1 + \max_{i\in[m]} |\operatorname{pa}(i)|. \tag{6.17}$$

On the other hand, if $n \geq 1 + \max_{i\in[m]} |\operatorname{pa}(i)|$ then $Y^{(i\cup\operatorname{pa}(i))} \in \mathbb{K}^{(1+|\operatorname{pa}(i)|)\times n}$ does not have full row rank if and only if all its maximal minors vanish. Thus, for generic (and hence almost all) Y we have that for all $i \in [m]$ the matrix $Y^{(i\cup\operatorname{pa}(i))}$ has full row rank. By Theorem 6.3.16(c), this implies

$$\operatorname{mlt}_u(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) \leq 1 + \max_{i\in[m]} |\operatorname{pa}(i)| \tag{6.18}$$

and combining (6.17) and (6.18) yields the claim. $\square$

Chapter 7

Log-linear Models

Log-linear models are widespread in statistics and play a fundamental role in categorical data analysis, with a wide range of applications [BFH07]. They are discrete models, see Section 6.2, and include independence models and discrete graphical models [Lau96]. There is a long history of the study of log-linear models in statistics, with an emphasis on ML estimation [FR12]. Log-linear models play a prominent role in algebraic statistics: the key link to algebra is that the Zariski closure of a log-linear model is a toric variety, defined by a monomial parametrization. Toric varieties have a foundational place among the algebraic varieties studied in algebraic geometry [CLS11].

We study connections between toric invariant theory and maximum likelihood (ML) estimation for log-linear models. Concretely, we use notions of stability under a torus action to characterize existence of the maximum likelihood estimate (MLE), Theorem 7.2.1. Moreover, we show that norm minimization over a torus orbit is equivalent to maximizing the log-likelihood in log-linear models, Theorem 7.2.3. This in turn allows to compare scaling algorithms from statistics and invariant theory. The whole chapter is based on [AKRS21b], which is joint work with with Carlos Améndola, Kathlén Kohn and Anna Seigal.

This is the first instance in this thesis which intimately links invariant theory and ML estimation. In Chapters 8, 9 and 10 we will encounter similar connections between invariant theory and ML estimation for *Gaussian* models. Of special interest to the discrete setting here is the study of Gaussian group models in Chapter 9. The latter is based on [AKRS21a], the companion paper of [AKRS21b]. We find remarkable similarities and differences between the discrete and Gaussian settings, which we discuss in Subsection 9.6.2. The discrete case is presented first, since the study of scaling algorithms for log-linear models motivates and contributes to the algorithmic consequences in Subsection 9.3.1.

Organization and Assumptions. In Section 7.1 we review log-linear models and known results on their ML estimation. Afterwards, we present the main results, Theorems 7.2.1 and 7.2.3, and illustrate them in examples, Section 7.2. Finally, we compare iterative proportional scaling (IPS), a classical method to find the MLE for log-linear models, with approaches to norm minimization and scaling from invariant theory in Section 7.3.

7.1 ML Estimation in log-linear Models

Recall our conventions on discrete models from Section 6.2. First, we define log-linear models following [Sul18, Definition 6.2.1].

Definition 7.1.1. Let $A \in \mathbb{Z}^{d\times m}$. The *log-linear model* given by matrix A is

$$\mathcal{M}_A^{\ell\ell} := \big\{p \in \operatorname{relint}(\Delta_{m-1}) \mid \log p \in \operatorname{rowspan}(A)\big\}. \tag{7.1}$$

where $\log p$ denotes the coordinatewise logarithm, which only applies to p with strictly positive entries. Therefore, $\mathcal{M}_A^{\ell\ell} \subseteq \operatorname{relint}(\Delta_{m-1})$. ▲

The superscript in $\mathcal{M}_A^{\ell\ell}$ stresses that we deal with log-linear models and it distinguishes them from Gaussian models $\mathcal{M}_{\mathcal{A}}^{\mathfrak{g}}$ studied in Chapters 8, 9 and 10. A parametrization of the model $\mathcal{M}_A^{\ell\ell}$ is given by

$$\begin{array}{rccc} \phi^A: & \mathbb{R}^d_{>0} & \longrightarrow & \Delta_{m-1} \\ & \theta & \longmapsto & \left(\frac{1}{Z(\theta)}\prod_{i=1}^d \theta_i^{a_{ij}}\right)_{1\le j\le m} \end{array} \tag{7.2}$$

where $Z(\theta)$ is a normalization factor. Conversely, any discrete model obtained from such a monomial parametrization is a log-linear model. We observe a first connection between the statistical model and a torus action: the map ϕ^A is, up to normalization, the action (1.6) of the real positive torus element θ on the all-ones vector $\mathbb{1}_m$. Furthermore, given a vector of counts $u \in \mathbb{Z}^m_{\ge 0}$, the vector Au is a sufficient statistics for the model $\mathcal{M}_A^{\ell\ell}$ with respect to the parametrization (7.2). This follows, e.g., by considering $\ell_u(\phi^A(\theta))$.

Remark 7.1.2 (Assumption $\mathbb{1}_m^\mathsf{T} \in \operatorname{rowspan}(A)$). For the log-linear model $\mathcal{M}_A^{\ell\ell}$, we assume that the row vector $\mathbb{1}_m^\mathsf{T}$ is in the row span of A; this is a common assumption for statistical, as well as algebraic, reasons. First, such log-linear models are equivalent to discrete exponential families [Sul18, Section 6.2]. Second, the assumption means the uniform distribution $\frac{1}{m}\mathbb{1}_m$ is in the model. Moreover, consider the Zariski closure of $\mathcal{M}_A^{\ell\ell}$ in $\mathbb{C}^m$, defined by the toric ideal

$$I_A = \langle p^x - p^y \mid x, y \in \mathbb{Z}^m_{\ge 0} \text{ such that } Ax = Ay\rangle \tag{7.3}$$

in the polynomial ring $\mathbb{C}[p_1, \dots, p_m]$, where $p^x := \prod_{j=1}^m p_j^{x_j}$ for $x \in \mathbb{Z}^m_{\ge 0}$; compare [Sul18, Proposition 6.2.4]. If $\mathbb{1}_m^\mathsf{T} \in \operatorname{rowspan}(A)$, this becomes a homogeneous ideal: if $r^\mathsf{T}A = \mathbb{1}_m^\mathsf{T}$ for some $r \in \mathbb{R}^d$ then left multiplying $Ax = Ay$ by r^T results in $\mathbb{1}_m^\mathsf{T}x = \mathbb{1}_m^\mathsf{T}y$. ▽

We just mentioned that log-linear models are examples of so-called discrete exponential families, [Sul18, Section 6.2]. Furthermore, log-linear models contain the undirected discrete graphical models as a special case via hierarchical log-linear models, see [Lau96] and [Sul18, Proposition 13.2.5]. In particular, the independence model is a log-linear model, compare Examples 7.1.3 and 7.2.5.

Example 7.1.3 (based on [AKRS21b, Examples 4.1 and 4.9])**.** The independence model of two discrete random variables with m_1 respectively m_2 states is

$$\mathcal{M}_{X\perp\!\!\!\perp Y} = \big\{\alpha^\mathsf{T}\beta \mid \alpha \in \Delta_{m_1-1},\, \beta \in \Delta_{m_2-1}\big\} \subseteq \mathbb{R}^{m_1\times m_2},$$

see Example 6.2.4. For $p = (p_{ij}) \in \mathcal{M}_{X \perp\!\!\!\perp Y}$, we see that the monomial parametrization $p_{ij} = \alpha_i \beta_j$, where $i \in [m_1]$ and $j \in [m_2]$, yields a log-linear model $\mathcal{M}_A^{\ell\ell}$ with $A \in \mathbb{Z}_{\geq 0}^{(m_1+m_2)\times(m_1 m_2)}$, compare (7.2). The matrix A has one row for each of the parameters α_i and β_j, and one column for each state (i, j) of the pair of random variables. For the concrete case $m_1 = 2$ and $m_2 = 3$, we have

$$A = \begin{array}{c} \begin{matrix} 11 & 12 & 13 & 21 & 22 & 23 \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{matrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{matrix} .$$

In general, if we use the same ordering of the parameters and states, we obtain

$$A = \begin{pmatrix} & \mathrm{I}_{m_1} \otimes \mathbb{1}_{m_2}^{\mathsf{T}} & \\ \\ \mathbb{1}_{m_1}^{\mathsf{T}} \otimes \mathrm{I}_{m_2} & & \end{pmatrix} = \begin{pmatrix} \mathbb{1}_{m_2}^{\mathsf{T}} & & \\ & \ddots & \\ & & \mathbb{1}_{m_2}^{\mathsf{T}} \\ \mathrm{I}_{m_2} & \cdots & \mathrm{I}_{m_2} \end{pmatrix} \in \mathbb{Z}^{(m_1+m_2)\times(m_1 m_2)}, \tag{7.4}$$

where we use the Kronecker product as introduced in Definition 1.3.4. Since $\mathcal{M}_A^{\ell\ell}$ lies in the relative interior of $\Delta_{m_1 m_2 - 1}$, it equals the relative interior of $\mathcal{M}_{X \perp\!\!\!\perp Y}$. We recover the independence model as the extended log-linear model $\overline{\mathcal{M}_A^{\ell\ell}} = \mathcal{M}_{X \perp\!\!\!\perp Y}$, compare Definition 6.2.2.

As mentioned after Equation (7.2), $\mathcal{M}_A^{\ell\ell}$ is the intersection of $\mathrm{relint}(\Delta_{m_1 m_2 - 1})$ with the orbit of the all-ones matrix under the action of $\mathrm{GT}_{2m}(\mathbb{R})$ on $\mathbb{R}^{m \times m}$ given by the matrix A in (7.4). Equivalently, $\mathcal{M}_A^{\ell\ell}$ is the orbit of the all-ones matrix under the left-right action of $\mathrm{GT}_m(\mathbb{R}) \times \mathrm{GT}_m(\mathbb{R})$ on $\mathbb{R}^{m \times m}$, again intersected with $\mathrm{relint}(\Delta_{m_1 m_2 - 1})$.

We illustrate this for the special case $m_1 = 2$ and $m_2 = 3$. The action of $\mathrm{GT}_5(\mathbb{R})$ on $\mathbb{R}^{3\times 3}$ given by (1.5), is as follows. The torus element

$$\begin{pmatrix} t_1 & t_2 & t_3 & t_4 & t_5 \end{pmatrix} = \begin{pmatrix} \lambda_1 & \lambda_2 & \nu_1 & \nu_2 & \nu_3 \end{pmatrix}$$

acts on a matrix $x \in \mathbb{R}^{3\times 3}$ by multiplying the entry x_{ij} by $\prod_{k=1}^{5} t_k^{A_{(i,j)}}$ where $A_{(i,j)}$ denotes the column of A with index (i, j). This is the left-right action of $\mathrm{GT}_2(\mathbb{R}) \times \mathrm{GT}_3(\mathbb{R})$ on the space of 2×3 matrices; it sends M_{ij} to $\lambda_i \nu_j M_{ij}$. ♢

Now, we consider ML estimation for log-linear models. Since the model $\mathcal{M}_A^{\ell\ell}$ is not closed, the MLE may not exist. To ensure existence, recall from Definition 6.2.2 the notion of an *extended log-linear model* $\overline{\mathcal{M}_A^{\ell\ell}} \subseteq \Delta_{m-1}$, and the one of an extended MLE $\hat{p} \in \overline{\mathcal{M}_A^{\ell\ell}}$ of $\mathcal{M}_A^{\ell\ell}$, which *always* exists. In fact, for a log-linear model there is a unique extended MLE [Lau96, Proposition 4.7].[1]

[1] It is known that the likelihood function (6.4) is *strictly* concave on $\mathcal{M}_A^{\ell\ell}$, see [Sul18, Corollary 7.3.8].

By [Lau96, Theorem 4.8], the extended MLE given u is the point $\hat{p} \in \overline{\mathcal{M}_A^{\ell\ell}}$ such that $\pi_L(\hat{p}) = \pi_L(\bar{u})$, where $L := \text{rowspan}(A) \subseteq \mathbb{R}^m$, π_L is the orthogonal projection onto L and $\bar{u} = n^{-1}u$ is the empirical distribution. Note that $\ker(\pi_L) = L^\perp = \text{im}(A^\mathsf{T})^\perp = \ker(A)$ and therefore $\pi_L(\hat{p}) = \pi_L(\bar{u})$ holds if and only if $\bar{u} - \hat{p} \in \ker(A)$. Thus, the extended MLE given u is the point $\hat{p} \in \overline{\mathcal{M}_A^{\ell\ell}}$ satisfying

$$A\hat{p} = A\bar{u}. \tag{7.5}$$

We point out that (7.5) is also the sufficient condition for the MLE given u *if it exists*, see [DSS09, Proposition 2.1.5] or [Sul18, Corollary 7.3.9]. In particular, if the MLE given u exists, it is also the extended MLE. Therefore, the MLE given u exists if and only if the extended MLE $\hat{p}$ has positive entries (so that $\hat{p} \in \mathcal{M}_A^{\ell\ell}$).

We give some historical notes on (7.5). Birch [Bir63] was the first to rigorously study MLE existence in the context of multi-way tables, where he observed that u having all entries strictly positive is a sufficient condition for the MLE to exist and derived condition (7.5), sometimes known as Birch's Theorem, see [PS05, Theorem 1.10]. The fact that some entries could still be zero without affecting MLE existence was not fully understood until the work of Haberman, who gave the first characterization of MLE existence in her paper [Hab74].

A modern necessary and sufficient condition is the following, which is stated in [AKRS21b] as Proposition 4.2. For this, the convex hull of the columns $A_j \in \mathbb{Z}^d$ of the matrix A is the polytope

$$\Delta_A := \text{conv}\{A_1, \ldots, A_m\} \subseteq \mathbb{R}^d. \tag{7.6}$$

Proposition 7.1.4 ([Sul18, Theorem 8.2.1])**.** *Let $A \in \mathbb{Z}^{d\times m}$ be such that $\mathbb{1}_m^\mathsf{T} \in \text{rowspan}(A)$ and let $u \in \mathbb{Z}^m_{\geq 0}$ be a vector of counts for the log-linear model $\mathcal{M}_A^{\ell\ell}$. Then the MLE given u exists in $\mathcal{M}_A^{\ell\ell}$ if and only if $A\bar{u} \in \text{relint}(\Delta_A)$.*

In particular, we see that, indeed, if all entries of u are positive then the MLE always exists. The above proposition allows us to link ML estimation in $\mathcal{M}_A^{\ell\ell}$ to stability notions.

7.2 Toric Invariant Theory for ML estimation

We give equivalent characterizations ML estimation in via stability under a torus action, Theorem 7.2.1. Furthermore, we show that a point where the moment map vanishes yields the (extended) MLE in the log-linear model, see Theorem 7.2.3, and we illustrate the results in examples.

Recall from Example 1.3.16 the concept of a $\text{GT}_d(\mathbb{C})$-action on $\mathbb{C}^m$ via a weight matrix $A \in \mathbb{Z}^{d\times m}$ and a linearization $b \in \mathbb{Z}^d$. This allows to obtain the following characterization from Proposition 7.1.4.

Theorem 7.2.1 ([AKRS21b, Theorem 4.3])**.** *Consider a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$ with sample size $u_+ = n$, a matrix $A \in \mathbb{Z}^{d\times m}$ with $\mathbb{1}_m^\mathsf{T} \in \text{rowspan}(A)$, and the vector $b := Au \in \mathbb{Z}^d$. The stability of $\mathbb{1}_m \in \mathbb{C}^m$ under the action of the torus*

$\mathrm{GT}_d(\mathbb{C})$ *given by matrix* nA *with linearization* b *is related to ML estimation in* $\mathcal{M}_A^{\ell\ell}$ *given data* u *as follows.*

(a)	$\mathbb{1}_m$ *unstable*		*does not happen*
(b)	$\mathbb{1}_m$ *semistable*	$\Leftrightarrow$	*extended MLE exists and is unique*
(c)	$\mathbb{1}_m$ *polystable*	$\Leftrightarrow$	*MLE exists and is unique*
(d)	$\mathbb{1}_m$ *stable*		*does not happen*

Remark 7.2.2. We note that the weight matrix nA encodes the *model* $\mathcal{M}_A^{\ell\ell} = \mathcal{M}_{nA}^{\ell\ell}$, while the linearization $b = Au$ depends on the *observed data*. Furthermore, we always consider stability notions for $\mathbb{1}_m$, which neither depends on the model nor the data. We stress that this differs from the Gaussian case. There we always consider the action via left-multiplication, while the stability notions are in terms of the observed data; compare the discussion in Subsection 9.6.2. ▽

Proof of Theorem 7.2.1. Remember that the Hilbert-Mumford Criterion in Theorem 2.1.9 characterizes stability of $\mathbb{1}_m$ under $\mathrm{GT}_d(\mathbb{C})$ in terms of the weight polytope $\Delta_{nA}(\mathbb{1}_m)$ and the linearization b. We have $\Delta_{nA}(\mathbb{1}_m) = \Delta_{nA}$ since $\mathbb{1}_m$ has full support. By Proposition 7.1.4, the MLE given u exists if and only if $A\bar{u} = n^{-1}Au \in \mathrm{relint}(\Delta_A)$. The latter is equivalent to $b = Au \in \mathrm{relint}(\Delta_{nA})$, which is the condition for $\mathbb{1}_m$ being polystable from Theorem 2.1.9. This shows part (c).

Moreover, an extended MLE always exists and it is unique for log-linear models, compare the discussion around Equation (7.5). Thus, it remains to see that the cases of unstable and stable do not occur. First, $b = Au = nA\bar{u}$ lies in the polytope Δ_{nA} and hence $\mathbb{1}_m$ is semistable under $\mathrm{GT}_d(\mathbb{C})$ by Theorem 2.1.9. Second, the stable case cannot arise, which is seen as follows. There exists some $r \in \mathbb{R}^d$ with $r^\mathsf{T}A = \mathbb{1}_m^\mathsf{T}$, by the assumption $\mathbb{1}_m^\mathsf{T} \in \mathrm{rowspan}(A)$. Thus, the columns A_j of A all lie on the affine hyperplane $r_1x_1 + \cdots + r_dx_d = 1$. Therefore, the polytope Δ_A has empty interior in $\mathbb{R}^d$, and so has Δ_{nA}. □

We remark that we could take any other vector of full support in Theorem 7.2.1. The theorem shows that MLE existence can be tested by checking polystability under the group action. Note that we actually need all four stability notions when characterizing ML estimation of certain Gaussian models, compare, e.g., Theorem 9.3.6.

Next, we link the moment map to ML estimation in log-linear models. For this, recall Kempf-Ness Theorem 2.2.13: a vector v is polystable (respectively semistable) if and only if the moment map vanishes at a non-zero vector w contained in the orbit (closure) of v; equivalently, the capacity of v is positive and attained at w. On the other hand, the (extended) MLE maximizes the likelihood function on the (extended) log-linear model.

Therefore, considering the two optimization problems of maximizing the likelihood function in a (extended) log-linear model and of norm minimization in an orbit (closure) under the torus action, Theorem 7.2.1 states that one problem attains its optimum if and only if the other one does. The next theorem describes how these two optima are related via the moment map μ. We remind the reader that for $w \in \mathbb{C}^m$ we write $w^{[2]} := (|w_1|^2, \ldots, |w_m|^2)$, compare Equation (2.10).

Theorem 7.2.3 ([AKRS21b, Theorem 4.7]). *Let $A \in \mathbb{Z}^{d\times m}$ such that $\mathbb{1}_m^{\mathsf{T}} \in \operatorname{rowspan}(A)$ and let $u \in \mathbb{Z}^m_{\geq 0}$ be a vector of counts for $\mathcal{M}_A^{\ell\ell}$ with $u_+ = n$. Consider the torus action of $\mathrm{GT}_d(\mathbb{C})$ given by matrix nA with linearization $b = Au$. If $w \in \overline{\mathrm{GT}_d(\mathbb{C}) \cdot \mathbb{1}_m}\backslash\{0\}$ is such that $\mu(w) = 0$, then the extended MLE given u is*

$$\hat{p} = \frac{1}{\|w\|^2}\, w^{[2]} = \frac{1}{\|w\|^2}\left(|w_1|^2, |w_2|^2, \ldots, |w_m|^2\right) \in \overline{\mathcal{M}_A^{\ell\ell}}. \tag{7.7}$$

If $\mathbb{1}_m$ is polystable, i.e., if $w \in \mathrm{GT}_d(\mathbb{C}) \cdot \mathbb{1}_m$, then $\hat{p}$ is the MLE given u.

Proof. First, recall from Equation (2.10) in Example 2.2.8 that, for the torus action given by matrix nA and linearization b, the moment map at $w \in \mathbb{C}^m$ is given by

$$\mu(w) = \frac{1}{\|w\|^2}\big(nAw^{[2]} - \|w\|^2 b\big). \tag{7.8}$$

Hence, $\mu(w) = 0$ and $b = Au$ yield that

$$nAw^{[2]} = \|w\|^2 Au \quad \text{, equivalently,} \quad A\frac{w^{[2]}}{\|w\|^2} = A\frac{u}{n} = A\bar{u}.$$

Setting $\hat{p} := w^{[2]}/\|w\|^2 \in \Delta_{m-1}$, we see that $\hat{p}$ satisfies the condition (7.5) for the extended MLE given u. It remains to ensure that $\hat{p} \in \overline{\mathcal{M}_A^{\ell\ell}}$.

For this, let $\overline{\mathcal{M}_A^{\ell\ell}}^{\mathrm{Z}}$ be the smallest Zariski closed subset of Δ_{m-1} containing $\mathcal{M}_A^{\ell\ell}$, i.e., the Zariski closure of $\mathcal{M}_A^{\ell\ell}$ in $\mathbb{R}^m$ intersected with the simplex Δ_{m-1}. By [GMS06, Theorem 3.2], we have $\overline{\mathcal{M}_A^{\ell\ell}}^{\mathrm{Z}} = \overline{\mathcal{M}_A^{\ell\ell}}$, so it suffices to show that $\hat{p}$ satisfies the equations in (7.3).

First, we show that w obeys these equations. To do so, recall that $A^{(i)}$ is the i^{th} row of A. For $t \in \mathrm{GT}_d(\mathbb{C})$ and $x \in \mathbb{Z}^m_{\geq 0}$, we compute

$$(t \cdot \mathbb{1}_m)^x = \prod_{j=1}^{m} \left(t^{nA_j - b}\right)^{x_j} = \prod_{j=1}^{m}\prod_{i=1}^{d} t_i^{nA_{ij}x_j - x_j b_i} = \prod_{i=1}^{d} t^{nA^{(i)}x - x_+ b} = t^{nAx - x_+ b},$$

Therefore, $t \cdot \mathbb{1}_m$ satisfies $(t \cdot \mathbb{1}_m)^x = (t \cdot \mathbb{1}_m)^y$ for all $x, y \in \mathbb{Z}^m_{\geq 0}$ with $nAx - x_+ b = nAy - y_+ b$, and the same is true for $w \in \overline{\mathrm{GT}_d(\mathbb{C}) \cdot \mathbb{1}_m}$ by continuity. Now, if $x, y \in \mathbb{Z}^m_{\geq 0}$ are such that $Ax = Ay$, then $x_+ = y_+$ as $\mathbb{1}_m^{\mathsf{T}}$ is in the row span of A, compare Remark 7.1.2. Thus, we have $nAx - x_+ b = nAy - y_+ b$ and we see that w indeed satisfies equations (7.3), i.e., $w^x = w^y$ for all $x, y \in \mathbb{Z}^m_{\geq 0}$ with $Ax = Ay$.

Finally, for each equation $w^x = w^y$, we can take the absolute value squared on both sides to get $(w^{[2]})^x = (w^{[2]})^x$. Multiplying the latter with $\|w\|^{-2x_+}$ and using the equality $x_+ = y_+$ shows that $(\hat{p})^x = (\hat{p})^y$. This proves $\hat{p} \in \mathcal{M}_A^{\ell\ell}$.

In the polystable case, the vector w lies in the orbit of $\mathbb{1}_m$. Hence, all its entries are positive, and so are the entries of $\hat{p}$. Consequently, $\hat{p} \in \mathcal{M}_A^{\ell\ell}$ and therefore it is the MLE given u. □

Theorem 7.2.3 shows that the (extended) MLE can be obtained from norm minimization on an orbit (closure). It suggests to use algorithms from invariant theory for the Norm Minimization Problem 3.1.3 and Scaling Problem 3.1.4 to approximately compute the MLE. We discuss this approach in Section 7.3 and we motivate the study of these algorithms in the next two examples.

Example 7.2.4 ([AKRS21b, Example 4.8])**.** Consider the log-linear model $\mathcal{M}_A^{\ell\ell}$ and vector of counts u with $n = u_+ = 4$, where

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}, \qquad u = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \qquad b = Au = \begin{pmatrix} 5 \\ 3 \end{pmatrix}.$$

This model is the plane conic $x_2^2 = x_1x_3$ in $\mathrm{relint}(\Delta_2) \subseteq \mathbb{R}^3$, compare (7.2). As usual, we consider the action of $\mathrm{GT}_2(\mathbb{C})$ on $\mathbb{C}^3$ via matrix nA and linearization b. Since $A\bar{u} \in \mathrm{relint}(\Delta_A)$, the MLE given u exists (Proposition 7.1.4). Equivalently, the vector $\mathbb{1}_3$ is polystable under $\mathrm{GT}_2(\mathbb{C})$ by Theorem 7.2.1. Thus, there is a vector w of minimal norm in the orbit of $\mathbb{1}_3$ by Kempf-Ness, Theorem 2.2.13. We illustrate how the MLE given u can be obtained from w, by Theorem 7.2.3.

Since w lies in the orbit of $\mathbb{1}_3$, its entries are $w_j = t_1^{na_{1j}-b_1}t_2^{na_{2j}-b_2}$, where $t_i \in \mathbb{C}^\times$. One computes that $w = \begin{pmatrix} \lambda^3 & \lambda^{-1} & \lambda^{-5} \end{pmatrix}^\mathsf{T}$ where $\lambda = t_1/t_2$. Moreover, the moment map vanishes at w, so we have $nAw^{[2]} = \|w\|^2 b$. Combining these, gives $3|\lambda|^6 - |\lambda|^{-2} - 5|\lambda|^{-10} = 0$, or equivalently, the condition $3\nu^2 - \nu - 5 = 0$ for $\nu = |\lambda|^8$. We obtain

$$\hat{p} = \frac{w^{[2]}}{\|w\|^2}\frac{|\lambda|^{10}}{|\lambda|^{10}} = \frac{1}{\nu^2+\nu+1}\begin{pmatrix} \nu^2 \\ \nu \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{31+\sqrt{61}}{4\sqrt{61}+52} \\ \frac{3+3\sqrt{61}}{4\sqrt{61}+52} \\ \frac{9}{2\sqrt{61}+26} \end{pmatrix} \sim \begin{pmatrix} 0.4662 \\ 0.3175 \\ 0.2162 \end{pmatrix}$$

as the MLE given u. ♢

Example 7.2.5 (based on [AKRS21b, Example 4.9])**.** We revisit the independence model $\mathcal{M}_{X\perp\!\!\!\perp Y}$ in terms of log-linear models as in Example 7.1.3. Remember that $\mathcal{M}_{X\perp\!\!\!\perp Y}$ is the extended log-linear model $\overline{\mathcal{M}_A^{\ell\ell}}$, where A is given by (7.4). As a sanity check, we apply Theorem 7.2.3 to $\overline{\mathcal{M}_A^{\ell\ell}}$ and recover the knowledge on the MLE in $\mathcal{M}_{X\perp\!\!\!\perp Y}$ from Example 6.2.4.

Given a data matrix $u \in \mathbb{Z}_{\geq 0}^{m_1\times m_2}$, we consider the orbit of the all-ones matrix $\mathbb{1}_{m_1\times m_2} := \mathbb{1}_{m_1} \otimes \mathbb{1}_{m_2}^\mathsf{T} \in \mathbb{C}^{m_1\times m_2}$, under the action of $\mathrm{GT}_{m_1+m_2}(\mathbb{C})$ given by the matrix nA with linearization $b = Au \in \mathbb{Z}^{m_1+m_2}$, where the sample size is $n = u_{++}$. We seek a matrix $w \in \mathbb{C}^{m_1\times m_2}$ in the orbit closure of $\mathbb{1}_{m_1\times m_2}$ at which the moment map μ for the action vanishes. Identifying $w \in \mathbb{C}^{m_1\times m_2} \cong \mathbb{C}^{m_1m_2}$, the descriptions of μ in (7.8) and of A in (7.4) yield

$$n\begin{pmatrix} w_{1,+}^{[2]} \\ \vdots \\ w_{m_1,+}^{[2]} \\ w_{+,1}^{[2]} \\ \vdots \\ w_{+,m_2}^{[2]} \end{pmatrix} = \|w\|^2 \begin{pmatrix} u_{1,+} \\ \vdots \\ u_{m_1,+} \\ u_{+,1} \\ \vdots \\ u_{+,m_2} \end{pmatrix}, \tag{7.9}$$

where we recall that $w_{ij}^{[2]} = |w_{ij}|^2$. By Theorem 7.2.3, the extended MLE given data u is $\hat{p} = w^{[2]}/\|w\|^2 \in \overline{\mathcal{M}_A^{\ell\ell}}$. Note that $\hat{p}$ is the MLE in $\mathcal{M}_{X\perp\!\!\!\perp Y}$ given

data u, because $\overline{\mathcal{M}_A^{\ell\ell}} = \mathcal{M}_{X \perp\!\!\!\perp Y}$. Now, Equation (7.9) shows that the marginal distributions of $\hat{p}$ are

$$\hat{p}_{i,+} = \frac{w_{i,+}^2}{\|w\|^2} = \frac{u_{i,+}}{n},\ i \in [m_1] \qquad \text{and} \qquad \hat{p}_{+,j} = \frac{w_{+,j}^2}{\|w\|^2} = \frac{u_{+,j}}{n},\ j \in [m_2].$$

Since $\hat{p} \in \mathcal{M}_{X \perp\!\!\!\perp Y}$, its entries are given by the product of the corresponding marginals, compare Example 6.2.4. Hence, we obtain

$$\hat{p}_{i,j} = \hat{p}_{i,+}\hat{p}_{+,j} = \frac{u_{i,+}u_{+,j}}{n^2},$$

which recovers the knowledge on the MLE in $\mathcal{M}_{X \perp\!\!\!\perp Y}$ from Example 6.2.4.

Finally, we note that if $w \in \mathrm{GT}_{m_1+m_2}(\mathbb{C}) \cdot \mathbb{1}_{m_1 \times m_2}$, then all entries of $\hat{p}$ are positive and so $\hat{p}$ is the MLE in $\mathcal{M}_A^{\ell\ell}$ given u; also compare Theorem 7.2.3. ♢

Finally, we point out that [AKRS21b, Propositions 4.4 and 4.5] give characterizations for existence of the MLE via semistability. We do not include these results here for brevity and since they deviate a bit from the main story.

7.3 Scaling Algorithms for log-linear Models

This section presents different possibilities of MLE computation in independence models and, more generally, log-linear models. We focus on known algorithms in the statistics community, and on computational consequences of Theorem 7.2.3. Thereby, we connect ML estimation to scaling algorithms from invariant theory (see Section 3.2). The purpose of this section is "storytelling". In particular, the following discussion contributes to algorithmic consequences in Chapter 9 by comparing Figures 7.1 and 9.1.[2]

We saw in Theorem 7.2.3 that the (extended) MLE in a log-linear model can be obtained from a point of minimal norm in an orbit (closure). This connects two problems:

1. norm minimization in a complex orbit (closure) under a torus action

2. maximum likelihood estimation in a (extended) log-linear model.

Algorithms exist for both problems: the former can be approached with convex optimization methods, and the latter with an algorithm called iterative proportional scaling (IPS). In fact, both families of algorithms can be thought of as generalizations of Sinkhorn scaling. We explain these different generalizations, and how Theorem 7.2.3 completes the circle of algorithms, see Figure 7.1.

[2]It naturally motivates to regard geodesic convex methods from invariant theory as iterative proportional scaling for so-called Gaussian group models (where the group is Zariski closed and self-adjoint).

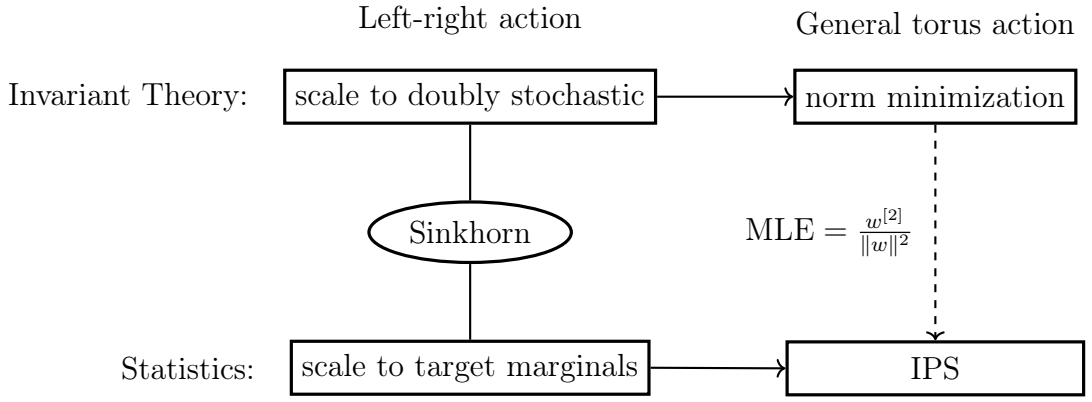


Figure 7.1: [AKRS21b, Figure 4] Overview of different scaling algorithms. The historical progression is from left to right. The dashed line is Theorem 7.2.3.

Sinkhorn Scaling

We recall that the classical scaling algorithm of Sinkhorn in [Sin64], Algorithm 3.1 (approximately) scales a square matrix $M \in \mathbb{R}^{m\times m}$ with non-negative entries to a doubly stochastic matrix. That is, Sinkhorn scaling is a method for matrix scaling as discussed in the extended example of Section 3.1. This is achieved by alternately scaling the row and column marginals to one, see Algorithm 3.1.

A natural extension is to scale the matrix M to other fixed row sums and column sums [SK67]. Both versions of Sinkhorn scaling are depicted on the left of Figure 7.1. These algorithms involve the left-right action of a pair of tori $\mathrm{GT}_{m_1} \times \mathrm{GT}_{m_2}$ on an $m_1 \times m_2$ matrix: the algorithms iterate between updates via the left torus and via the right torus.

Alternately scaling of the rows and columns of a matrix to fixed marginals is an instance of a scaling algorithm which, in the statistics literature, goes back to Deming and Stephan in [DS40]. For the independence model $\mathcal{M}_{X\perp\!\!\!\perp Y}$ on two variables, the algorithm finds the MLE by alternating between scaling the row sums and the column sums to match the marginals of the empirical distribution. Given an observed matrix of counts $u \in \mathbb{Z}_{\geq 0}^{m\times m}$ with sample size $u_{++} = n$, and initialized at the uniform distribution, the algorithm has two steps. The (i, j) entry changes in these two steps as follows:

$$\frac{1}{m^2} \;\mapsto\; \frac{1}{m} \cdot \frac{u_{i+}}{n} \;\mapsto\; \frac{u_{i+}}{n} \cdot \frac{u_{+j}}{n}. \tag{7.10}$$

Its output is the MLE in $\mathcal{M}_{X\perp\!\!\!\perp Y}$ given u, which is extended MLE of the corresponding log-linear model, compare Examples (7.1.3) and 7.2.5. This is the first example of *iterative proportional scaling* (IPS), which we describe next.

Iterative Proportional Scaling (IPS)

We have just seen that alternating scaling of a matrix to fixed row and column sums gives the MLE to the independence model, when initialized at the uniform distribution. This is scaling under a product of tori $\mathrm{GT}_{m_1} \times \mathrm{GT}_{m_2}$. We saw in Examples 7.1.3 and 7.2.5 how the independence model fits into the framework

of log-linear models. In terms of the group action, the left-right action of a pair of tori $\mathrm{GT}_{m_1} \times \mathrm{GT}_{m_2}$ is the action of $\mathrm{GT}_{m_1+m_2}$, acting via (1.5), where A is the matrix in (7.4).

In the following, we explain how Sinkhorn scaling extends to IPS for ML estimation in a general log-linear model, see the bottom arrow of Figure 7.1.

Alternating between matching row and column sums can be extended to hierarchical models, which summarize data by contingency tables [Fie70], by iteratively updating the various marginals. The approach was extended to more general log-linear models by Darroch and Ratcliff in [DR72].

For the log-linear model $\mathcal{M}_A^{\ell\ell}$, the MLE $\hat{p}$ must satisfy the equation $A\hat{p} = A\bar{u}$, (7.5), from Birch's theorem, where $\bar{u} = \frac{u}{n}$ is the empirical distribution. IPS finds the extended MLE in $\mathcal{M}_A^{\ell\ell}$ given an empirical distribution $\bar{u} \in \Delta_{m-1}$. We define IPS for a log-linear model given by a matrix $A \in \mathbb{Z}_{\geq 0}^{d\times m}$ whose column sums are all equal. Starting at the uniform distribution $p^{(0)} = \frac{1}{m}\mathbb{1}_m$, we iterate until the k^{th} update $p^{(k)}$ has sufficient statistics $b^{(k)} = Ap^{(k)}$ close to the target sufficient statistics $b = A\bar{u}$, i.e., until (7.5) holds approximately. The update step is:

$$p_j^{(k+1)} = \prod_{i=1}^{d} \left(\frac{(A\bar{u})_i}{(Ap^{(k)})_i} \right)^{a_{ij}/\alpha} p_j^{(k)}, \tag{7.11}$$

where α is the common column sum of A; see [Sul18, Algorithm 7.3.11].[3] This is the action of a torus element (obtained by componentwise division of $A\bar{u}$ by $Ap^{(k)}$ and then componentwise exponentiation by $1/\alpha$) on the vector $p^{(k)}$. Here the torus action is given by the matrix A with linearization $b = 0$.

The IPS method is a minimization approach: at each step it minimizes the KL divergence to the MLE.

Proposition 7.3.1 ([AKRS21b, Proposition 5.1])**.** *Consider a vector of counts $u \in \mathbb{Z}_{\geq 0}^m$ for the log-linear model $\mathcal{M}_A^{\ell\ell}$, where $A \in \mathbb{Z}^{d\times m}$ has $\mathbb{1}_m^{\mathsf{T}}$ in its row span. Then there exists a matrix $\tilde{A} \in \mathbb{Q}_{\geq 0}^{(d+1)\times m}$, with all column sums equal, such that $\mathcal{M}_A^{\ell\ell} = \mathcal{M}_{\tilde{A}}^{\ell\ell}$, iterative proportional scaling in* (7.11) *with matrix $\tilde{A}$ converges, and at each update step the KL divergence to the MLE decreases.*

Proof. The proof of convergence of IPS is given in [DR72, Theorem 1] in the case where the entries of A are real and non-negative with each column of A summing to one. There, the authors show that each step of IPS decreases the KL divergence $\mathrm{KL}(\hat{p}\|p^{(k)})$ from the k^{th} iterate $p^{(k)}$ to the MLE $\hat{p}$. Since replacing A by $\frac{1}{\alpha}A$ does not change the update step (7.11), the KL divergence also decreases for any matrix with real and non-negative entries and all column sums equal.

We explain how this covers log-linear models defined by integer matrices with $\mathbb{1}_m^{\mathsf{T}}$ in the row span. We modify A without changing its row span, i.e., without changing the model $\mathcal{M}_A^{\ell\ell}$. First, we add a sufficiently large positive integer to every entry of A. For a general choice of integer, this does not change $\mathrm{rowspan}(A)$ since it adds a multiple of the row vector $\mathbb{1}_m^{\mathsf{T}}$, which belongs to $\mathrm{rowspan}(A)$, to every

[3] [Sul18, Algorithm 7.3.11] involves $\phi_i^{A,h}(\theta)$, which is defined in [Sul18, Definition 6.2.2]. Note that $\phi^{A,h}$ in [Sul18] does *not* involve the normalization factor $Z(\theta)$ like in our Equation (7.2).

row. Second, let α be the maximum of the column sums $A_{+,j}$. Add another row to the matrix, with entries $\alpha - A_{+,j}$. The extra row is a linear combination of $\mathbb{1}_m^\mathsf{T}$ and the rows of A, so the augmented matrix has the same row span as A. By construction, the column sums of the augmented matrix $\tilde{A}$ are all equal to α. □

Remark 7.3.2 ([AKRS21b, Remark 5.2]). We saw in Equation (6.5) from Section 6.2 that $\hat{p} = \operatorname{argmin}_{p \in \mathcal{M}_A^{\ell\ell}} \mathrm{KL}(\bar{u} \| p)$. Here, we use KL divergence differently, measuring the KL divergence from iterate $p^{(k)}$ to the MLE: $\mathrm{KL}(\hat{p} \| p^{(k)})$. ▽

Curiously, when IPS for log-linear models in (7.11) is applied to the independence model, we do not recover the classical IPS with Sinkhorn updates, because the column sums of the integer matrix A for the independence model in (7.4) are $\alpha = 2$, hence there is a square root in the update step. If, instead, we did IPS with the same matrix A but $\alpha = 1$ in (7.11) we would recover the two steps in (7.10) in a single step. This leads naturally to the question of which exponents α achieve convergence, and how the choice of α affects the convergence rate. This is the essence of an open problem in algebraic statistics, see [DSS09, Section 7.3].

Norm Minimization

We explain/recall how Sinkhorn scaling generalizes to norm minimization for torus actions in invariant theory; see the top arrow of Figure 7.1.

For this, the extended example on matrix scaling from Section 3.1 is crucial. We recall that given a matrix $v \in \mathbb{C}^{m \times m}$ the left-right action of $T := \mathrm{ST}_m(\mathbb{C})^2$ relates to matrix scaling of $M_v = \left(|v_{ij}|^2\right)$. By Proposition 3.1.7, M_v is (approximately) scalable if and only if the moment map vanishes at some non-zero vector w in the orbit (closure) of v under T. By Kempf-Ness Theorem 2.2.13, the vanishing of the moment map at w is equivalent to the capacity $\mathrm{cap}_T(v) = \inf_{t \in T} \|t \cdot v\|^2$ being positive and attained at w. Therefore, an appropriate normalization[4] of the update steps in Sinkhorn scaling (Algorithm 3.1) solve the Norm minimization Problem 3.1.3 and Scaling Problem 3.1.4 for v under the action of T.

We have seen in Equation (3.1) that norm minimization for any algebraic action of a torus is a convex optimization problem. In the specific situation of log-linear models, the action of $\mathrm{GT}_d(\mathbb{C})$ is given by matrix $A' = nA - \mathbb{1}_m^\mathsf{T} \otimes b \in \mathbb{Z}^{d \times m}$.[5] The vector $\mathbb{1}_m$ is always semistable, see Theorem 7.2.1. By Kempf-Ness, norm minimization converges to a non-zero vector $w \in \overline{\mathrm{GT}_d(\mathbb{C}) \cdot \mathbb{1}_m}$ at which the moment map vanishes. Hence, common algorithms from the vast literature on convex optimization can be used to approximate the capacity and find the (extended) MLE, using Theorem 7.2.3. In particular, one can use the methods mentioned in the paragraph on the commutative case in Section (3.2).

Finally, we recall that the alternating minimization idea from Sinkhorn's algorithm generalizes to operator scaling, Algorithm 3.2. The latter solves norm minimizatio(and scaling) for the left-right action of $\mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ on the space of matrix tuples $(\mathbb{C}^{m_1 \times m_2})^n$. We discuss connections between operator scaling and statistics in Subsection 9.4.4.

[4] similarly to Algorithm (3.2) to ensure the determinant one conditions of $\mathrm{ST}_m(\mathbb{C})^2$

[5] This is the action given by nA with linearization b, compare Definition 1.3.4.

Comparison of Algorithms

We have seen in the previous paragraphs that IPS and norm minimization can be viewed as generalizations of Sinkhorn scaling. Theorem 7.2.3 closes the cycle of algorithms from different communities, by showing how to obtain the (extended) MLE from a complex point of minimal norm in an orbit (closure); see Figure 7.1.

This bridges several differences between IPS and norm minimization. We summarize these differences here. First, when computing the capacity of $\mathbb{1}_m$, the norm is minimized along a *complex* orbit closure (see Theorem 7.2.3), whereas every step in IPS involves *real* numbers. Secondly, the torus action given by matrix nA that is used for computing the capacity is linearized by $b = Au$ (see Theorem 7.2.3), whereas IPS uses the action given by matrix A with trivial linearization $b = 0$. Finally, the objective functions differ: the capacity is defined in terms of the Euclidean norm, which does not appear in IPS; instead IPS minimizes KL divergence (see Proposition 7.3.1). In the following example we see that, while IPS decreases the KL divergence to the MLE, it may increase the Euclidean norm.

Example 7.3.3 ([AKRS21b, Example 5.3])**.** Consider the matrix A and vector of counts u from Example 7.2.4, i.e.,

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}, \qquad u = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

We use IPS to compute the MLE in $\mathcal{M}_A^{\ell\ell}$. We start at the uniform distribution $p^{(0)} = \frac{1}{3}\mathbb{1}_3$ and do update steps as in (7.11) with matrix A. These IPS steps converge by Proposition 7.3.1, since the matrix A has real non-negative entries and all column sums are equal. We obtain $p^{(1)} = \frac{1}{12}\begin{bmatrix} 5 & \sqrt{15} & 3 \end{bmatrix}^{\mathsf{T}}$. Note that the sum of the entries of $p^{(1)}$ is strictly less than one. The KL divergence from the uniform distribution to the MLE is $\mathrm{KL}(\hat{p}\|p^{(0)}) \sim 0.047$, and after the first update it is $\mathrm{KL}(\hat{p}\|p^{(1)}) \sim 0.016$. However, we have $\|p^{(1)}\|^2 = \frac{49}{144}$, which exceeds $\|p^{(0)}\|^2 = \frac{1}{3}$. ♢

Chapter 8

Gaussian Models via Symmetrization

This chapter starts our studies of ML estimation on Gaussian models and sets the stage for Chapters 9 and 10. We define so-called Gaussian models via symmetrization, which in hindsight deserve a treatment on their own. The main result is the weak correspondence, Theorem 8.2.3. It views maximizing the log-likelihood as a norm minimization problem and provides a first dictionary between stability notions and ML estimation in the Gaussian case. The weak correspondence generalizes similar statements of [AKRS21a] and we need this level of generality in Chapter 10.

The chapter is in parts based on discussions with Anna Seigal and Visu Makam and on our joint paper [MRS21, Appendix A].

Organization and Assumptions. In Section 8.1, we define Gaussian models via symmetrization and state simple properties of these. Afterwards, we define stability notions under *sets* and prove the weak correspondence, Section 8.2.

Similarly to Section 6.3 we work in parallel over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Remember that $(\cdot)^\dagger$ is the Hermitian transpose, which equals the transpose $(\cdot)^\mathsf{T}$ if $\mathbb{K} = \mathbb{R}$.

8.1 Examples and first Properties

Definition 8.1.1 (Gaussian model via symmetrization, Gaussian group model)**.** For a subset $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ we define

$$\mathcal{M}^{\mathsf{g}}_{\mathcal{A}} := \left\{a^\dagger a \mid a \in \mathcal{A}\right\} \subseteq \mathrm{PD}_m(\mathbb{K}), \tag{8.1}$$

the *Gaussian model via symmetrization* of $\mathcal{A}$. If $\mathcal{A} = G$ is a subgroup of $\mathrm{GL}_m(\mathbb{K})$ we call $\mathcal{M}^{\mathsf{g}}_G$ a *Gaussian group model*. ▲

The superscript g indicates that $\mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$ is a Gaussian model and distinguishes it from log-linear models $\mathcal{M}^{\ell\ell}_A$, which are studied in Chapter 7. We point out that for $\mathbb{K} = \mathbb{R}$ the Hermitian transpose $a^\dagger$ is just the transpose a^T. Hence, Definition 8.1.1 matches the definition of Gaussian group models over $\mathbb{R}$ respectively $\mathbb{C}$ given in [AKRS21a] and its generalizations to $\mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$ in [MRS21].

A positive definite matrix $\Psi \in \mathrm{PD}_m(\mathbb{K})$ admits a *Cholesky decomposition*: there is a unique upper triangular matrix $\mathrm{chol}(\Psi) \in \mathrm{GL}_m(\mathbb{K})$ with *positive* diagonal entries such that $\Psi = \mathrm{chol}(\Psi)^\dagger \mathrm{chol}(\Psi)$. As a consequence, *any* Gaussian model is of the form $\mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$.

Proposition 8.1.2. *Let $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$ be a Gaussian model. Then there exists a subset $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ with $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$.*

Proof. The subset $\mathcal{A} := \{\mathrm{chol}(\Psi) \mid \Psi \in \mathcal{M}\} \subseteq \mathrm{GL}_m(\mathbb{K})$ satisfies $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$ by construction. Another choice for $\mathcal{A}$ is $\{\Psi^{1/2} \mid \Psi \in \mathcal{M}\}$, where $\Psi^{1/2}$ denotes the square root of Ψ, i.e., the unique positive definite matrix whose square is Ψ. □

Remark 8.1.3. We think of the set $\mathcal{A}$ as a *parametrization* of the model $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$. A Gaussian model may admit many different parametrizations, e.g., whenever we have $\mathcal{A} \subseteq \mathcal{B} \subseteq \{ga \mid a \in \mathcal{A},\, g^\dagger g = \mathrm{I}_m\}$ it holds that $\mathcal{M}^{\mathsf{g}}_{\mathcal{A}} = \mathcal{M}^{\mathsf{g}}_{\mathcal{B}}$. ▽

Example 8.1.4 (Saturated Gaussian Model)**.** The saturated Gaussian model $\mathcal{M} = \mathrm{PD}_m(\mathbb{K})$ can be parametrized by any $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ that contains the group $\mathrm{B}_m(\mathbb{K})$ of invertible upper triangular matrices.[1] In particular, we have $\mathrm{PD}_m(\mathbb{K}) = \mathcal{M}^{\mathsf{g}}_{\mathrm{GL}_m(\mathbb{K})} = \mathcal{M}^{\mathsf{g}}_{\mathrm{B}_m(\mathbb{K})}$. We will see corresponding statistical interpretations of these parametrizations: $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathrm{GL}_m(\mathbb{K})}$ is studied as a Gaussian group model with self-adjoint group in Example 9.3.8; $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathrm{B}_m(\mathbb{K})}$ arises as a directed Gaussian graphical model in Example 9.5.11. ♢

In statistics one often studies Gaussian models which are closed under positive scalars. In this regard, the following proposition[2] justifies the assumption "$\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ is closed under non-zero scalar multiples" of Theorem 8.2.3. The assumption is used there and throughout the thesis to relate ML estimation of the model $\mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$ to norm minimization and to stability notions.

Proposition 8.1.5. *A Gaussian model $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$ is closed under positive scalar multiples if and only if there is some set $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ closed under non-zero scalar multiples such that $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$.*

Proof. To prove the "if"-part, let $\Psi \in \mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$. Then there is some $a \in \mathcal{A}$ with $\Psi = a^\dagger a$. For $\lambda > 0$, we have $\sqrt{\lambda} a \in \mathcal{A}$ by assumption on $\mathcal{A}$ and hence

$$\lambda\Psi = \big(\sqrt{\lambda}a\big)^\dagger \big(\sqrt{\lambda}a\big) \in \mathcal{M}^{\mathsf{g}}_{\mathcal{A}} = \mathcal{M}$$

as claimed.

Conversely, assume that $\mathcal{M}$ is closed under positive scalar multiples. Consider the set

$$\mathcal{A} := \{\tau\, \mathrm{chol}(\Psi) \mid \tau \in \mathbb{K}^\times, \Psi \in \mathcal{M}\},$$

which is closed under non-zero scalar multiples. We have $\mathrm{chol}(\Psi) \in \mathcal{A}$ for all $\Psi \in \mathcal{M}$ and thus $\mathcal{M} \subseteq \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$. On the other hand, for all $\tau \in \mathbb{K}^\times$ and all $\Psi \in \mathcal{M}$,

$$\big(\tau\, \mathrm{chol}(\Psi)\big)^\dagger \big(\tau\, \mathrm{chol}(\Psi)\big) = |\tau|^2\, \Psi \in \mathcal{M},$$

where we used $|\tau|^2 > 0$ and the assumption on $\mathcal{M}$. This shows $\mathcal{M} = \mathcal{M}^{\mathsf{g}}_{\mathcal{A}}$. □

Remark 8.1.6. The proof of Proposition 8.1.5 shows that the statement remains true, if we replace the assumption on $\mathcal{A}$ by "$\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ closed under positive scalar multiples". Indeed, the only necessary adjustment is to consider $\mathcal{A} = \{\lambda\, \mathrm{chol}(\Psi) \mid \lambda > 0, \Psi \in \mathcal{M}\}$ for the "only if"-direction. ▽

[1] These are not all options, e.g., $\{\mathrm{chol}(\Psi) \mid \Psi \in \mathrm{PD}_m(\mathbb{K})\}$ is strictly contained in $\mathrm{B}_m(\mathbb{K})$.

[2] This proposition arose from a discussion with Anna Seigal and Visu Makam.

8.2 The weak Correspondence

In this section we prove the main result of this chapter – the so-called *weak correspondence*[3], Theorem 8.2.3. The weak correspondence casts maximizing the log-likelihood function as a norm minimization problem. This in turn allows to relate ML estimation to stability notions, which we introduce now.

Fix a subset $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ and a tuple of samples $Y = (Y_1, \dots, Y_n) \in (\mathbb{K}^m)^n$. Remember that we view the samples Y_i as column vectors, which identifies Y as a matrix in $\mathbb{K}^{m\times n} \cong (\mathbb{K}^m)^n$. There is no Often we switch implicitly between these identifications. Given some $a \in \mathcal{A}$, we set

$$a \cdot Y := (aY_1, \dots, aY_n) \in (\mathbb{K}^m)^n \cong \mathbb{K}^{m\times n},$$

which is just the multiplication of the matrices a and Y. The dot indicates that we think of the set $\mathcal{A}$ "acting" via left multiplication on $\mathbb{K}^{m\times n} \cong (\mathbb{K}^m)^n$. In analogy to group actions we define the *orbit* of Y and the *stabilizing set* of Y under the *set* $\mathcal{A}$ as

$$\mathcal{A} \cdot Y := \{a \cdot Y \mid a \in \mathcal{A}\} \quad \text{and} \quad \mathcal{A}_Y := \{a \in \mathcal{A} \mid a \cdot Y = Y\}, \tag{8.2}$$

respectively.[4] Analogous to the topological stability notions for group actions, Definition 1.4.1, we make the following definitions.

Definition 8.2.1 (Stability Notions for Sets, [MRS21, Definition A.1])**.**
We say the tuple of samples $Y \in (\mathbb{K}^m)^n \cong \mathbb{K}^{m\times n}$, under the set $\mathcal{A}$, is

(i) *unstable* if $0 \in \overline{\mathcal{A} \cdot Y}$;

(ii) *semistable* if Y is not unstable, i.e., $0 \notin \overline{\mathcal{A} \cdot Y}$;

(iii) *polystable* if $Y \neq 0$ and the set $\mathcal{A} \cdot Y$ is Euclidean closed;

(iv) *stable* if Y is polystable and $\mathcal{A}_Y$ is finite. ▲

If $\mathcal{A}$ is a subgroup of $\mathrm{GL}_m(\mathbb{K})$, then the above stability notions are just the usual topological stability notions under the action on $\mathbb{K}^{m\times n}$ via left-multiplication.

To prove the weak correspondence we need the following lemma.

Lemma 8.2.2. *Fix $m, n > 0$ and, for $\gamma \geq 0$, consider the family of functions*

$$f_\gamma \colon \mathbb{R}_{>0} \to \mathbb{R}, \quad x \mapsto \frac{\gamma}{n}x - m\log(x).$$

(i) If $\gamma = 0$, then $\inf_{x>0} f_\gamma(x) = -\infty$.

(ii) If $\gamma > 0$, then f_γ attains a global minimum at $x_0 = \frac{mn}{\gamma}$ with function value $f_\gamma(\frac{mn}{\gamma}) = m(1 - \log(mn) + \log(\gamma))$.

[3]The name *weak correspondence* was coined by Anna Seigal during discussions with Gergely Bérczi, Eloise Hamilton, Visu Makam and myself.

[4]Since $\mathcal{A}$ is just a *set* one needs to be careful: known results from the theory of group actions do not need to hold. For example, in general the orbits do not form a partition of $\mathbb{K}^{m\times n}$.

(iii) Given $\gamma_1 > \gamma_2 > 0$, we have $f_{\gamma_1}(\frac{\alpha}{\gamma_1}) > f_{\gamma_2}(\frac{\alpha}{\gamma_2})$ at the global minima.

Proof. The first part follows from the properties of the logarithm. To prove part (ii), one computes for $x > 0$ that $f'_\gamma(x) = \frac{\gamma}{n} - \frac{m}{x}$ and $f''_\gamma(x) = \frac{m}{x^2} > 0$. The latter implies that f_γ is strictly convex and hence the former yields that $x_0 = \frac{mn}{\gamma}$ is the unique global minimum. One directly verifies the equation for $f_\gamma(x_0)$, so part (iii) follows from the strict monotonicity of the logarithm. □

Recall Equation (6.8): for the model $\mathcal{M}_A^{\mathrm{g}}$ and tuple of samples $Y \in (\mathbb{K}^m)^n$ the log-likelihood function ℓ_Y at $\Psi = a^\dagger a$, where $a \in \mathcal{A}$, is given by

$$\ell_Y(\Psi) = \ell_Y\big(a^\dagger a\big) = \log\det\big(a^\dagger a\big) - \operatorname{tr}\big(a^\dagger a S_Y\big), \quad \text{where } S_Y = \frac{1}{n}\sum_{i=1}^n Y_i Y_i^\dagger.$$

A key for relating ML estimation to norm minimization is the following observation: for all $a \in \mathcal{A}$ we compute[5]

$$n \operatorname{tr}\big(a^\dagger a S_Y\big) = \sum_{i=1}^n \operatorname{tr}\big(Y_i^\dagger a^\dagger a Y_i\big) = \sum_{i=1}^n (aY_i)^\dagger a Y_i = \|a \cdot Y\|^2 \tag{8.3}$$

where we used in the second equality that $Y_i^\dagger a^\dagger a Y_i = (aY_i)^\dagger a Y_i$ is a scalar. Hence, the log-likelihood ℓ_Y at $\Psi = a^\dagger a$ can be rewritten as

$$\ell_Y\big(a^\dagger a\big) = \log\det\big(a^\dagger a\big) - \frac{1}{n}\|a \cdot Y\|^2. \tag{8.4}$$

We use this equation to prove the weak correspondence. To state it, set

$$\mathcal{A}_{\mathrm{SL}} := \{a \in \mathcal{A} \mid \det(a) = 1\}, \tag{8.5}$$
$$\mathcal{A}_{\mathrm{SL}}^- := \{a \in \mathcal{A} \mid \det(a) = -1\}, \tag{8.6}$$
$$\mathcal{A}_{\mathrm{SL}}^\pm := \{a \in \mathcal{A} \mid \det(a) = \pm 1\}. \tag{8.7}$$

The weak correspondence, Theorem 8.2.3, is based on [MRS21, Proposition A.4] and generalizes [AKRS21a, Proposition 3.4 and Theorem 3.6] to sub*sets* $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$. Its key feature is that it casts maximizing the log-likelihood as a two step optimization problem, compare Equation (8.8). First, one minimizes $\|b \cdot Y\|^2$ over $b \in \mathcal{A}_{\mathrm{SL}}^\pm$, i.e., one computes $\operatorname{cap}_{\mathcal{A}_{\mathrm{SL}}^\pm}(Y)$. Afterwards, one is left with a univariate convex optimization problem. This two step approach in combination with Lemma 8.2.2 allows to connect ML estimation to stability notions.

Theorem 8.2.3 (Weak Correspondence [MRS21, Proposition A.4])**.**
Let $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ be closed under non-zero scalar multiples. The supremum of the log-likelihood ℓ_Y over $\mathcal{M}_{\mathcal{A}}^{\mathrm{g}}$ can be computed as a double infimum:

$$\sup_{a \in \mathcal{A}} \ell_Y\big(a^\dagger a\big) = -\inf_{x \in \mathbb{R}_{>0}} \left(\frac{x}{n} \left(\inf_{b \in \mathcal{A}_{\mathrm{SL}}^\pm} \|b \cdot Y\|^2 \right) - m \log(x) \right). \tag{8.8}$$

[5] Recall that, if not stated otherwise, we consider the norm induced by the standard inner product, so $\mathbb{K}^{m \times n}$ is equipped with the Frobenius norm. Thus, the norm of Y does not change under the identification $\mathbb{K}^{m \times n} \cong (\mathbb{K}^m)^n$.

The MLEs, if they exist, are the matrices $\lambda b^\dagger b$, where $b \in \mathcal{A}_{\mathrm{SL}}^{\pm}$ minimizes the inner infimum and $\lambda \in \mathbb{R}_{>0}$ is the unique *global minimum of the outer infimum. Equation* (8.8) *gives a correspondence between stability under $\mathcal{A}_{\mathrm{SL}}^{\pm}$ and maximum likelihood estimation in the model $\mathcal{M}_{\mathcal{A}}^{\mathtt{g}}$ given sample matrix $Y \in \mathbb{K}^{m\times n}$:*

$$\begin{array}{lccc} (a) & Y \text{ unstable} & \Leftrightarrow & \text{likelihood } \ell_Y \text{ unbounded from above} \\ (b) & Y \text{ semistable} & \Leftrightarrow & \text{likelihood } \ell_Y \text{ bounded from above} \\ (c) & Y \text{ polystable} & \Rightarrow & \text{MLE exists.} \end{array}$$

The whole statement holds for $\mathcal{A}_{\mathrm{SL}}$ replacing $\mathcal{A}_{\mathrm{SL}}^{\pm}$, if

(i) $\mathbb{K} = \mathbb{C}$, *or*

(ii) $\mathbb{K} = \mathbb{R}$ *and for any $a \in \mathcal{A}$ there is an orthogonal matrix $o = o(a)$ such that $o^{\mathsf{T}} a \in \mathcal{A}$ and $\det(o^{\mathsf{T}} a) > 0$.*

Remark 8.2.4. The weak correspondence applies exactly to those Gaussian models $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$ that are closed under positive scalar multiples. Indeed, these models are exactly the ones that admit a set $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$ closed under scalar multiples such that $\mathcal{M} = \mathcal{M}_{\mathcal{A}}^{\mathtt{g}}$, see Proposition 8.1.5. ▽

Proof of Theorem 8.2.3. By Equation (8.4), maximizing ℓ_Y over $\mathcal{M}_{\mathcal{A}}^{\mathtt{g}}$ is equivalent to minimizing the function

$$f\colon \mathcal{A} \to \mathbb{R}, \qquad a \mapsto \frac{1}{n}\|a \cdot Y\|^2 - \log\det(a^\dagger a).$$

Using the assumption on $\mathcal{A}$ we can rewrite an element $a \in \mathcal{A}$ as follows. For $\mathbb{K} = \mathbb{R}$, let $\tau := \sqrt[m]{|\det(a)|} \in \mathbb{R}^\times$, then $b := \tau^{-1}a \in \mathcal{A}_{\mathrm{SL}}^{\pm}$ and $a = \tau b$. If $\mathbb{K} = \mathbb{C}$, let $\tau \in \mathbb{C}^\times$ be some m^{th} root of $\det(a)$, so $b := \tau^{-1}a \in \mathcal{A}_{\mathrm{SL}}^{\pm}$ and $a = \tau b$. (Actually, $b \in \mathcal{A}_{\mathrm{SL}}$ and this leads to the fact that $\mathcal{A}_{\mathrm{SL}}^{\pm}$ may always be replaced by $\mathcal{A}_{\mathrm{SL}}$ given $\mathbb{K} = \mathbb{C}$.) Setting $x := |\tau|^2$, we compute *both* in the real and complex case that

$$f(a) = \frac{|\tau|^2}{n}\|b \cdot Y\|^2 - \log\det\left(|\tau|^2 b^\dagger b\right) = \frac{x}{n}\|b \cdot Y\|^2 - m\log(x).$$

Let $\gamma := \inf_{b\in\mathcal{A}_{\mathrm{SL}}^{\pm}} \|b \cdot Y\|^2$. The above argument and computation yields[6]

$$\inf_{a\in\mathcal{A}} f(a) = \inf_{x>0,\, b\in\mathcal{A}_{\mathrm{SL}}^{\pm}} \frac{x}{n}\|b \cdot Y\|^2 - m\log(x) = \inf_{x>0} \frac{x}{n}\gamma - m\log(x),$$

i.e., we obtain Equation (8.8).

By Lemma 8.2.2, $\inf_{a\in\mathcal{A}} f(a) = -\infty$ if and only if $\gamma = \inf_{b\in\mathcal{A}_{\mathrm{SL}}^{\pm}} \|b \cdot Y\|^2 = 0$, i.e., if and only if Y is unstable under $\mathcal{A}_{\mathrm{SL}}^{\pm}$. This shows parts (a) and (b).

To prove (c), assume that Y is polystable under $\mathcal{A}_{\mathrm{SL}}^{\pm}$. Then $\gamma > 0$ as Y is semistable and hence $x \mapsto \gamma n^{-1}x - m\log(x)$ is minimized by a unique $\lambda > 0$, by Lemma 8.2.2. Since $\mathcal{A}_{\mathrm{SL}}^{\pm} \cdot Y$ is closed in $\mathbb{K}^{m\times n}$, we see that γ is attained by

[6]For a function $F\colon X \times Y \to \mathbb{R}$ one has $\inf_{x\in X, y\in Y} F(x,y) = \inf_{x\in X}\inf_{y\in Y} F(x,y)$. Alternatively, one can use Lemma 8.2.2: the infimum of the function $\mathbb{R}_{>0} \to \mathbb{R}, x \mapsto \gamma n^{-1}x - m\log(x)$ increases as $\gamma \geq 0$ increases.

some b in the compact set $(A_{\mathrm{SL}}^{\pm} \cdot Y) \cap \{Z \in \mathbb{K}^{m\times n} \mid \|Z\|^2 \leq \gamma + 1\}$. Thus, $-f\big(\sqrt{\lambda}b\big) = \sup_{\Psi\in\mathcal{M}_{\mathcal{A}}^{\mathsf{g}}} \ell_Y(\Psi)$ and an MLE given Y, namely $\lambda b^\dagger b$, exists.

Using Equation (8.8) we see that actually any matrix of the form $\lambda b^\dagger b$, where λ and b are as in the statement, is an MLE. Conversely, let $\hat{a} \in \mathcal{A}$ be such that $\hat{\Psi} := \hat{a}^\dagger\hat{a} \in \mathcal{M}_{\mathcal{A}}^{\mathsf{g}}$ is an MLE given Y. Similar to the above, write $\hat{a} = \hat{\tau}\hat{b}$ with $\tau \in \mathbb{K}^\times$ and $\hat{b} \in \mathcal{A}_{\mathrm{SL}}^{\pm}$. Then $\ell_Y(\hat{\Psi}) = -f(\hat{a})$ is the maximum of ℓ_Y, equivalently,

$$\inf_{a\in\mathcal{A}} f(a) = f(\hat{a}) = \frac{|\hat{\tau}|^2}{n}\|\hat{b}\cdot Y\|^2 - m\log\big(|\hat{\tau}|^2\big).$$

Therefore, the inner and outer infima in (8.8) must be attained by $|\hat{\tau}|^2$ and $\hat{b}$, respectively; otherwise we would obtain a contradiction to $\inf_{a\in\mathcal{A}} f(a) = f(\hat{a})$ via Lemma 8.2.2. Altogether, $\hat{\Psi} = |\hat{\tau}|^2\,\hat{b}^\dagger\hat{b}$ has the claimed form.

Finally, we discuss two situations in which $\mathcal{A}_{\mathrm{SL}}^{\pm}$ can be replaced by $\mathcal{A}_{\mathrm{SL}}$. We already mentioned that we can write $a = \tau b$ with $\tau \in \mathbb{C}^\times$ and $b \in \mathcal{A}_{\mathrm{SL}}$, if $\mathbb{K} = \mathbb{C}$. On the other hand, if condition (ii) is satisfied, then any $a \in \mathcal{A}$ can be rewritten as $a = \tau o b$, where $\tau := \sqrt[m]{|\det(a)|}$ and $b := \tau^{-1}o^{\mathsf{T}}a$. We have $b \in \mathcal{A}$, because $o^{\mathsf{T}}a \in \mathcal{A}$ and $\mathcal{A}$ is closed under non-zero scalars. Furthermore, $\det(o^{\mathsf{T}}a) = \det(o)\det(a) > 0$ and $|\det(o)| = 1$ imply $\det(o^{\mathsf{T}}a) = |\det(a)|$, hence $b \in \mathcal{A}_{\mathrm{SL}}$. Noting that $(ob)^\dagger(ob) = b^\dagger b$ and $\|(ob)\cdot Y\|^2 = \|b\cdot Y\|^2$ by orthogonality of o, we obtain *both* under condition (i) and under (ii) Equation (8.8) with $\mathcal{A}_{\mathrm{SL}}^{\pm}$ replaced by $\mathcal{A}_{\mathrm{SL}}$. Moreover, the remaining parts of the proof remain valid under this replacement. □

Remark 8.2.5. Notice that condition (ii) in Theorem 8.2.3 is trivially satisfied if $\mathcal{A}$ only contains matrices with positive determinant (choose $o = \mathrm{I}_m$), or if n is odd. In the latter case, one can choose $o(a) = \mathrm{sgn}(\det(a))\,\mathrm{I}_m$.

From an invariant theory perspective it is more natural to work with $\mathcal{A}_{\mathrm{SL}}$ instead of $\mathcal{A}_{\mathrm{SL}}^{\pm}$. In this regard, condition (ii) seems unpleasant and artificial. However, in this generality it cannot be dropped as we shall see in the next Example 8.2.6 and in Example 9.2.9. Still, apart from these examples all Gaussian models studied in this thesis satisfy condition (ii) and we will work with $\mathcal{A}_{\mathrm{SL}}$ instead of $\mathcal{A}_{\mathrm{SL}}^{\pm}$. ▽

Example 8.2.6. Consider the involutive matrix

$$M := \begin{pmatrix} 1/2 & 3 \\ 1/4 & -1/2 \end{pmatrix}$$

which is *not* orthogonal and has determinant -1. Then

$$G := \bigcup_{\tau\in\mathbb{K}^\times} \{\tau\,\mathrm{I}_2, \tau M\} \tag{8.9}$$

is a subgroup of $\mathrm{GL}_2(\mathbb{K})$, which is closed under non-zero scalars. The Gaussian group model $\mathcal{M}_G^{\mathsf{g}} = \{\lambda\,\mathrm{I}_2, \lambda M^\dagger M \mid \lambda > 0\}$ consists of two rays in $\mathrm{PD}_m(\mathbb{K})$.[7] In

[7] The Gaussian group model is taken from [AKRS21a, Example 3.12], presented in Example 9.3.5, and we use this model several times for illustration.

the following we study the situation $n = 1$ with observed sample $Y = (1, 0)^\mathsf{T}$ and illustrate the differences between the real and complex situation. In particular, we show that violating condition (ii) in the real case can prevent the replacement of $G_{\mathrm{SL}}^{\pm}$ by G_{SL} in Theorem 8.2.3.

Let $\mathbb{K} = \mathbb{R}$. Since scaling a matrix of $\mathrm{GL}_2(\mathbb{R})$ by $\tau \in \mathbb{R}^\times$ scales its determinant with τ^2, we have $G_{\mathrm{SL}} = \{\pm \mathrm{I}_2\}$ and $G_{\mathrm{SL}}^{-} = \{\pm M\}$. Note that there is no orthogonal matrix o with $o^\mathsf{T} M \in G_{\mathrm{SL}}$, i.e., condition (ii) in Theorem 8.2.3 is violated. Indeed, otherwise we would have $o^\mathsf{T} = (o^\mathsf{T} M)M \in G_{\mathrm{SL}}^{-}$, but this contradicts that G_{SL}^{-} does not contain an orthogonal matrix.

We have

$$\|(\pm M) \cdot Y\|^2 = \frac{1}{4} + \frac{1}{16} < 1 = \| \pm Y\|^2$$

and thus $\inf_{g \in G_{\mathrm{SL}}^{\pm}} \|g \cdot Y\|^2 = 5/16$ is attained on $G_{\mathrm{SL}}^{-} \cdot Y$, but not on $G_{\mathrm{SL}} \cdot Y$. This shows that we *cannot* replace $G_{\mathrm{SL}}^{\pm}$ by G_{SL} in Theorem 8.2.3.

By Theorem 8.2.3, an MLE given Y is of the form $\lambda b^\mathsf{T} b$, where $b = \pm M$ and λ is the unique global minimum of $x \mapsto (5/16)x - 2\log(x)$. Lemma 8.2.2 (ii) shows $\lambda = 32/5$. Note that both choices of b give the same MLE, so we conclude that

$$\lambda M^\mathsf{T} M = \frac{32}{5} \begin{pmatrix} 5/16 & 11/8 \\ 11/8 & 37/4 \end{pmatrix} = \begin{pmatrix} 2 & 44/5 \\ 44/5 & 296/5 \end{pmatrix}$$

is the unique MLE given Y.

Next, let $\mathbb{K} = \mathbb{C}$. The main difference to the real case is that we now have $G_{\mathrm{SL}} = \{\pm \mathrm{I}_2, \pm \mathrm{i} M\}$ and $G_{\mathrm{SL}}^{-} = \{\pm M, \pm \mathrm{i}\, \mathrm{I}_2\}$. Therefore, $\inf_{g \in G_{\mathrm{SL}}^{\pm}} \|g \cdot Y\|^2 = 5/16$ is attained *both* on $G_{\mathrm{SL}} \cdot Y$ and on $G_{\mathrm{SL}}^{-} \cdot Y$. By Theorem 8.2.3 for G_{SL}, an MLE given Y is of the form $\lambda b^\dagger b$, where $b = \pm \mathrm{i} M$ and $\lambda = 32/5$ is the unique global minimum of $x \mapsto (5/16)x - 2\log(x)$. Since $|\pm \mathrm{i}|^2 = 1$ and M has only real entries, we see that $\lambda M^\dagger M = \lambda M^\mathsf{T} M$ is again the unique MLE given Y. ♢

Chapter 9

Gaussian Group Models

"Und jedem Anfang wohnt ein Zauber inne"

Hermann Hesse in his poem *Stufen*

Building upon the theory from Chapter 8 we study Gaussian group models and deepen the connections between invariant theory and maximum likelihood (ML) estimation. Recall from Definition 8.1.1 that a Gaussian group model is a Gaussian model via symmetrization $\mathcal{M}_G^{\mathfrak{g}}$, where G is a subgroup of $\mathrm{GL}_m(\mathbb{K})$. The group situation allows to use many further tools, especially since the group G acts on the samples via left multiplication. In particular, we may use different criteria for stability from Chapter 2.

We remark that the starting point of this theory were similarities between operator scaling (Algorithm 3.2) from invariant theory and the flip-flop algorithm for computing MLEs in matrix normal models (Subsection 9.4.4). This algorithmic view stimulated the search for connections between invariant theory and algebraic statistics. Eventually, this lead to a dictionary, like in Equation (1), between stability notions and ML estimation for matrix normal models (Theorem 9.4.1). That in turn fostered research on the existence of such a dictionary at different levels of generality and/or for different assumptions. The current state of this research for Gaussian models is presented in Chapters 8, 9 and 10.

This chapter is mainly based on [AKRS21a], which is joint work with Carlos Améndola, Kathlén Kohn and Anna Seigal. Several results in Section 9.2 were stimulated by discussions with my collaborators Gergely Bérczi, Eloise Hamilton, Visu Makam and Anna Seigal, or are implicitly contained in [AKRS21a]. Moreover, Section 9.5 also takes further knowledge from [MRS21] (see Chapter 10) into account. Finally, we note that [AKRS21a] is the companion paper of [AKRS21b], which studies log-linear models via toric invariant theory and is presented in Chapter 7. We will compare log-linear models and Gaussian group models at the end of this chapter.

Main Results. First, we collect basic properties of Gaussian group models in Propositions 9.2.1, 9.2.3 and 9.2.4. In particular, Gaussian group models are transformation families (Definition 6.1.4), and the stabilizer of a tuple of samples Y naturally acts on the set of MLEs given Y.

Thanks to the group structure, the conditions to work with G_{SL} in the weak correspondence (Theorem 8.2.3) simplify, compare Theorem 9.2.7. Remember

that the weak correspondence casts maximizing the log-likelihood function as a norm minimization problem. The latter means, in the case of a Gaussian group model $\mathcal{M}_G^{\mathfrak{g}}$, to compute the capacity (1.8). Moreover, the weak correspondence yields a first dictionary between stability notions and ML estimation. We extend this for two classes of Gaussian group models to a full list as in Equation (1).

In the first case, the group is assumed to be Zariski closed and self-adjoint, see Theorem 9.3.6. For $\mathbb{K} = \mathbb{C}$ we obtain an exact equivalence between the four notions of stability and the four properties of ML estimation as in the dictionary (1). We call this the *full correspondence*. If $\mathbb{K} = \mathbb{R}$ one implication is missing and we speak of the *strong correspondence* instead.[1] Furthermore, the natural action of the stabilizer on the set of MLEs is transitive for Zariski closed self-adjoint groups, Proposition 9.3.3.

The second case in which we obtain the full correspondence is the situation of Gaussian graphical models on transitive DAGs (TDAGs), see Theorem 9.5.9. We deduce this correspondence by proving equivalences between stability notions and linear independence conditions on the sample matrix, Theorem 9.5.8. Remarkably, for TDAGs the stabilizer of a tuple of samples is even in bijection with the set of MLEs, compare Proposition 9.5.10,

Applications of the Dictionary. We point out three applications of a dictionary between stability notions and ML estimation. In this chapter this is specifically showcased for matrix normal models.

First, such a dictionary may allow to obtain new characterizations and recover known results via an invariant theory perspective. For Example, Theorem 9.4.6 and Corollaries 9.4.7 and 9.4.12 recover known results, while Theorem 9.4.14 is a new characterization for complex matrix normal models; see Subsection 9.4.2.

Second, one can tackle questions on ML thresholds via invariant theory: the problem of computing the three ML thresholds essentially translates to generic semi/poly/stability, respectively. Indeed, we use descriptions of the null cone to give improved bounds on the boundedness threshold mlt_b for matrix normal models, see Theorem 9.4.10 and Corollary 9.4.11. These results were new at their time. In the meantime, the theory from Section 9.4 was successfully used to completely determine the ML thresholds for matrix normal models [DM21]. We state their result in Theorem 9.6.1. In fact, this was generalized to tensor normal models in [DMW22].

Third, the connections lead to algorithmic consequences. We can compare scaling algorithms from invariant theory and ML estimation with each other. In Subsection 9.4.4 we show that operator scaling (Algorithm 3.2) and the Flip-Flop Algorithm 9.1 for matrix normal models are essentially the same. Furthermore, we can regard the geodesically convex methods from [BFG+19] as iterative proportional scaling for Gaussian group models given by Zariski closed self-adjoint groups, see Subsection 9.3.1.

[1] Like the name *weak correspondence*, the names *strong correspondence* respectively *full correspondence* where coined by Anna Seigal during discussions with Gergely Bérczi, Eloise Hamilton, Visu Makam and myself.

Organization and Assumptions. In Section 9.1 we describe how a rational representation of an algebraic group induces a Gaussian group model, and we state several examples of Gaussian group models. Then we collect several basic properties and the weak correspondence in Section 9.2. Afterwards, we study the case of Zariski closed self-adjoint groups in Section 9.3 and illustrate the theory in detail for matrix normal models in Section 9.4. We connect Gaussian graphical models on transitive DAGs to Gaussian group models in Section 9.5. Finally, we discuss related literature and compare Gaussian group models with log-linear models from Chapter 7 in Section 9.6.

As in Section 6.3 we work over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, and $(\cdot)^\dagger$ denotes the Hermitian transpose, which equals the transpose $(\cdot)^\mathsf{T}$ if $\mathbb{K} = \mathbb{R}$.

9.1 Models via Group Actions

Recall from Definition 8.1.1 that for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and a subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$ the Gaussian group model given by G is

$$\mathcal{M}_G^{\mathfrak{g}} = \left\{ g^\dagger g \mid g \in G \right\} \subseteq \mathrm{PD}_m(\mathbb{K}).$$

We have already seen in Example 8.1.4 that the saturated model $\mathrm{PD}_m(\mathbb{K})$ can be seen as a Gaussian group model in several ways, e.g., as $\mathcal{M}^{\mathfrak{g}}_{\mathrm{GL}_m(\mathbb{K})}$ or as $\mathcal{M}^{\mathfrak{g}}_{\mathrm{B}_m(\mathbb{K})}$. Another example of a Gaussian group model is the following.

Example 9.1.1. For the group $G = \mathrm{GT}_m(\mathbb{K})$ of invertible diagonal matrices we obtain $\mathcal{M}_G^{\mathfrak{g}} = \{\Psi \in \mathrm{PD}_m(\mathbb{K}) \mid \Psi \text{ is diagonal}\}$, the model of m independent univariate Gaussians from Example 6.3.3. ♢

The group G naturally acts on the sample space $\mathbb{K}^m$ via left-multiplication. From an invariant theory perspective it is natural to study *general* group actions and associate Gaussian group models to these. This is always possible as follows. Let G be a group acting linearly on an m-dimensional $\mathbb{K}$-vector space V, i.e., we are given a morphism $\pi\colon G \to \mathrm{GL}(V)$ of groups. After choosing an ordered basis of V, or equivalently an isomorphism[2] $V \cong \mathbb{K}^m$, we can view $\pi(G)$ as a subgroup of $\mathrm{GL}_m(\mathbb{K})$ and obtain a corresponding Gaussian group model $\mathcal{M}^{\mathfrak{g}}_{\pi(G)} \subseteq \mathrm{PD}_m(\mathbb{K})$.

Remark 9.1.2. It is important to note that statistics naturally requires a choice how to measure data, e.g., a choice of coordinates as above. We stress that choosing different coordinates affects the statistical meaning. Indeed, a different isomorphism $V \cong \mathbb{K}^m$ gives a different inner product on V by pullback of the standard inner product. In this regard, if V already comes with an inner product then it is natural to choose an ordered *orthonormal* basis, or equivalently an *isometric* isomorphism $V \cong \mathbb{K}^m$.
For an illustration of this remark we refer to Example 9.3.5. ▽

Example 9.1.3. Let $\pi\colon T \to \mathrm{GL}(V)$ be a rational representation of a complex torus. We identify $T \cong \mathrm{GT}_d(\mathbb{C})$, where $d = \dim T$. Remember that V decomposes

[2]Recall that, if not mentioned otherwise, we always equip $\mathbb{K}^m$ with the standard inner product and the standard ordered basis $(e_1, \ldots, e_m)$.

into weight spaces by Theorem 1.3.14 and, as in Example 1.3.16 we can identify $V \cong \mathbb{C}^m$ such that the e_j, $j \in [m]$ are weight vectors. Let $(a_{1j}, \dots, a_{dj}) \in \mathbb{Z}^d$ be the corresponding weights. Then $t = \operatorname{diag}(t_1, \dots, t_d) \in \mathrm{GT}_d(\mathbb{C})$ acts on $V \cong \mathbb{C}^m$ by left multiplication with the diagonal matrix from (1.5), i.e.,

$$\pi(t) = \operatorname{diag}\left(t_1^{a_{11}} \cdots t_d^{a_{d1}},\ t_1^{a_{12}} \cdots t_d^{a_{d2}},\ \dots,\ t_1^{a_{1m}} \cdots t_d^{a_{dm}}\right) \in \mathrm{GL}_m(\mathbb{C}) \cong \mathrm{GL}(V).$$

We get the Gaussian group model $\mathcal{M}^{\mathsf{g}}_{\pi(T)} = \{\pi(t)^\dagger \pi(t) \mid t \in \mathrm{GT}_d(\mathbb{C})\} \subseteq \mathrm{PD}_m(\mathbb{C})$. If the torus T is $\mathbb{R}$-split and π is defined over $\mathbb{R}$, then all identifications can be done in a way that is compatible with the $\mathbb{R}$-structures. We obtain a similar Gaussian group model over the reals: $\{\pi(t)^\mathsf{T} \pi(t) \mid t \in \mathrm{GT}_d(\mathbb{R})\} \subseteq \mathrm{PD}_m(\mathbb{R})$. ♢

In Section 9.3 we study models $\mathcal{M}^{\mathsf{g}}_G$, where $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is a Zariski closed self-adjoint subgroup. Similar to the above construction, the next remark provides a class of group actions that naturally give rise to such Gaussian group models.

Remark 9.1.4 (based on [AKRS21a, Remark 2.4]). Let G be a reductive group over $\mathbb{C}$ and $\pi\colon G \to \mathrm{GL}(V)$ a rational representation on an m-dimensional $\mathbb{C}$-vector space. Then $\pi(G) = \pi(G)_{\mathbb{C}} \subseteq \mathrm{GL}(V)$ is a Zariski closed subgroup by Proposition 1.1.7, and if G and π are defined over $\mathbb{R}$ then $\pi(G)$ is defined over $\mathbb{R}$. Hence, $\pi(G)_{\mathbb{R}} \subseteq \mathrm{GL}(V_{\mathbb{R}})$ is a Zariski closed subgroup as well.

Since G is reductive, π is semisimple by Theorem 1.3.9 and hence $\pi(G)_{\mathbb{K}} \subseteq \mathrm{GL}(V_{\mathbb{K}})$ is a faithful semisimple representation. Therefore, there exists an inner product $\langle \cdot, \cdot \rangle$ on $V_{\mathbb{K}}$ to which $\pi(G)_{\mathbb{K}} \subseteq \mathrm{GL}(V_{\mathbb{K}})$ is self-adjoint, by Theorem 1.3.10. Thus, after fixing an ordered orthonormal basis with respect to $\langle \cdot, \cdot \rangle$ we can view $\pi(G)_{\mathbb{K}}$ as a Zariski closed self-adjoint subgroup of $\mathrm{GL}_m(\mathbb{K})$. We obtain a Gaussian group model $\mathcal{M}^{\mathsf{g}}_{\pi(G)_{\mathbb{K}}}$.

Remember that for $\mathbb{K} = \mathbb{R}$ we may have $\pi(G_{\mathbb{R}}) \subsetneq \pi(G)_{\mathbb{R}}$, see Example 1.1.8. Still, Corollary 1.2.6 yields $\pi(G)^\circ_{\mathbb{R}} \subseteq \pi(G_{\mathbb{R}}) \subseteq \pi(G)_{\mathbb{R}}$. The polar decompositions of $\pi(G)_{\mathbb{R}}$ and its subgroup $\pi(G_{\mathbb{R}})$, Theorem 1.2.16 and Corollary 1.2.17, show that they yield the same Gaussian group model: $\mathcal{M}^{\mathsf{g}}_{\pi(G)_{\mathbb{R}}} = \mathcal{M}^{\mathsf{g}}_{\pi(G_{\mathbb{R}})}$.[3]

We stress once more that the statistical meaning depends on the inner product on V, compare Remark 9.1.2 and see Example 9.3.5 for an illustration. ▽

We showcase the above construction for matrix and tensor normal models.

Example 9.1.5 (Matrix and Tensor Normal Models)**.** Consider the natural group action of the reductive group $\mathrm{GL}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{GL}_{m_d}(\mathbb{K})$ on $\mathbb{K}^{m_1} \otimes \cdots \otimes \mathbb{K}^{m_d}$ via $\mathbb{K}$-linear extension of

$$(g_1, \dots, g_d) \cdot (v_1 \otimes \cdots \otimes v_d) = g_1(v_1) \otimes \cdots \otimes g_d(v_d).$$

Recall that this is the *tensor scaling action* from Example 1.3.5. It induces the Gaussian group model $\mathcal{M}^{\mathsf{g}}_G$ given by the subgroup

$$G = \{g_1 \otimes \cdots \otimes g_d \mid g_i \in \mathrm{GL}_{m_i}(\mathbb{K})\} \subseteq \mathrm{GL}_{m_1 \cdots m_d}(\mathbb{K}),$$

[3] In [AKRS21a, Remark 2.4] it is stated that "$\varrho(G) \subseteq \mathrm{GL}(V)$ is a closed algebraic subgroup" giving a reference to [Mil17, Theorem 5.39] (ϱ is called π in Remark 9.1.4). This is certainly true over $\mathbb{R}$ in the *scheme theoretic sense*. However, we actually would like that the image of the $\mathbb{R}$-rational points of G (i.e., $\varrho(G_{\mathbb{R}})$ respectively $\pi(G_{\mathbb{R}})$) is Zariski closed in the $\mathbb{R}$-rational points of $\mathrm{GL}(V)$. This fails in general as Example 1.1.8 shows. We adjusted the remark correspondingly.

where we used the Kronecker product, see Definition 1.3.4. Note that the use of the Kronecker product implicitly identifies $\mathbb{K}^{m_1} \otimes \cdots \otimes \mathbb{K}^{m_d} \cong \mathbb{K}^{m_1 \cdots m_d}$. Under this identification the group $G \subseteq \mathrm{GL}_{m_1 \cdots m_d}(\mathbb{K})$ is self-adjoint (with respect to the standard inner product on $\mathbb{K}^{m_1 \cdots m_d}$). Moreover, G is Zariski closed in $\mathrm{GL}_{m_1 \cdots m_d}(\mathbb{K})$, even if $\mathbb{K} = \mathbb{R}$.[4] One can deduce Zariski closedness of G for $\mathbb{K} = \mathbb{R}$ by using that Segre embeddings are surjective on $\mathbb{R}$-rational points.[5]

With the properties of the Kronecker product we compute

$$(g_1 \otimes \cdots \otimes g_d)^\dagger (g_1 \otimes \cdots \otimes g_d) = g_1^\dagger g_1 \otimes \cdots \otimes g_d^\dagger g_d$$

and see that $\mathcal{M}_G^{\mathsf{g}}$ is the tensor normal model $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, \ldots, m_d)$ from (6.12) in Example 6.3.9. In the special case $d = 2$ we obtain the matrix normal model. $\diamondsuit$

It is convenient to study matrix normal models via the left-right action. We introduce this viewpoint, which is used in Section 9.4, in the following example.[6]

Example 9.1.6 (Left-right Action and Matrix Normal Models)**.**
The group $G := \mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K})$ acts algebraically on $\mathbb{K}^{m_1 \times m_2}$, which we equip with the Frobenius norm, via

$$(g_1, g_2) \cdot m = g_1 M g_2^{\mathsf{T}}, \tag{9.1}$$

where $g = (g_1, g_2) \in G$ and $M \in \mathbb{K}^{m_1 \times m_2}$. This is the left-right action from Example 1.3.3. We stress that also for $\mathbb{K} = \mathbb{C}$ the transpose g_2^{T} (and *not* the Hermitian transpose $g_2^\dagger$) is used to get an *algebraic* action. Furthermore, this allows for a natural identification via the $\mathbb{K}$-linear isomorphism

$$\mathbb{K}^{m_1} \otimes \mathbb{K}^{m_2} \xrightarrow{\sim} \mathbb{K}^{m_1 \times m_2}, \quad v \otimes w \mapsto v w^{\mathsf{T}} \tag{9.2}$$

which is induced by the $\mathbb{K}$-bilinear map $(v, w) \mapsto v w^{\mathsf{T}}$. Indeed, (9.2) identifies the standard orthonormal bases $e_i \otimes e_j \leftrightarrow E_{i,j}$ and writing $M = \sum_{i=1}^k v_i w_i^{\mathsf{T}} \leftrightarrow \sum_{i=1}^k v_i \otimes w_i$, where $v_i \in \mathbb{K}^{m_1}$ and $w_i \in \mathbb{K}^{m_2}$, we compute

$$(g_1, g_2) \cdot M = \sum_{i=1}^k g_1 v_i w_i^{\mathsf{T}} g_2^{\mathsf{T}} = \sum_{i=1}^k (g_1 v_i)(g_2 w_i)^{\mathsf{T}} \quad \leftrightarrow \quad \sum_{i=1}^k g_1(v_i) \otimes g_2(w_i).$$

The latter shows that the identification from (9.2) is G-equivariant, where G acts on $\mathbb{K}^{m_1} \otimes \mathbb{K}^{m_2}$ as in Example 9.1.5. Therefore, under the identification (9.2) the left-right action induces the matrix normal model

$$\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2) = \left\{ \Psi_1 \otimes \Psi_2 \mid \Psi_j \in \mathrm{PD}_{m_j}(\mathbb{K}) \right\}$$

from Examples 6.3.9 and 9.1.5.

Finally, let us compute the log-likelihood (6.8) in terms of the Kronecker factors Ψ_j. To do so, we need to isometrically identify $\mathbb{K}^{m_1 \times m_2} \cong \mathbb{K}^{m_1} \otimes \mathbb{K}^{m_2}$

[4] This may fail in general, compare Remark 9.1.4.

[5] Note that G is the intersection of $\mathrm{GL}_{m_1 \cdots m_d}(\mathbb{K})$ with the affine cone of a Segre variety.

[6] In [DM21] the left-right action was used to determine the ML thresholds of matrix normal models.

with the space of column vectors $\mathbb{K}^{m_1 m_2}$. By Definition 1.3.4 of the Kronecker product, there is an isomorphism $\mathrm{vec}\colon \mathbb{K}^{m_1 \times m_2} \to \mathbb{K}^{m_1 m_2}$ such that

$$\forall g_j \in \mathrm{GL}_{m_j}(\mathbb{K}),\, M \in \mathbb{K}^{m_1 \times m_2}\colon \qquad \mathrm{vec}\left(g_1 M g_2^{\mathsf{T}}\right) = (g_1 \otimes g_2)\,\mathrm{vec}(M). \tag{9.3}$$

For $A, B \in \mathbb{K}^{m_1 \times m_2}$, one verifies that vec is an isometry:

$$\mathrm{tr}\left(A^\dagger B\right) = \mathrm{vec}(A)^\dagger\,\mathrm{vec}(B) = \mathrm{tr}\left(\mathrm{vec}(A)^\dagger\,\mathrm{vec}(B)\right). \tag{9.4}$$

Given a tuple of samples $Y = (Y_1, \ldots, Y_n) \in (\mathbb{K}^{m_1 \times m_2})^n$, we consider the sample covariance matrix $S_{\mathrm{vec}(Y)}$ for $\mathrm{vec}(Y) := (\mathrm{vec}(Y_1), \ldots, \mathrm{vec}(Y_n))$ and compute

$$\begin{aligned} n\,\mathrm{tr}\left((\Psi_1 \otimes \Psi_2) S_{\mathrm{vec}(Y)}\right) &= \mathrm{tr}\left((\Psi_1 \otimes \Psi_2) \sum_{i=1}^n \mathrm{vec}(Y_i)\,\mathrm{vec}(Y_i)^\dagger\right) \\ &\overset{(9.3)}{=} \sum_{i=1}^n \mathrm{tr}\left(\mathrm{vec}\left(\Psi_1 Y_i \Psi_2^{\mathsf{T}}\right)\mathrm{vec}(Y_i)^\dagger\right) \overset{(9.4)}{=} \sum_{i=1}^n \mathrm{tr}\left(\Psi_1 Y_i \Psi_2^{\mathsf{T}} Y_i^\dagger\right). \end{aligned}$$

As a consequence, the log-likelihood (6.8) becomes

$$\begin{aligned} \ell_Y(\Psi_1 \otimes \Psi_2) &= \log\det(\Psi_1 \otimes \Psi_2) - \mathrm{tr}\left((\Psi_1 \otimes \Psi_2) S_{\mathrm{vec}(Y)}\right) \\ &= m_2\,\log\det(\Psi_1) + m_1\,\log\det(\Psi_2) - \frac{1}{n}\,\mathrm{tr}\left(\Psi_1 \sum_{i=1}^n Y_i \Psi_2^{\mathsf{T}} Y_i^\dagger\right). \end{aligned}$$

We stress the *transpose* Ψ_2^{T} for $\mathbb{K} = \mathbb{C}$, which has to be kept in mind for Algorithm 9.1 over $\mathbb{C}$. (Of course, $\Psi_2^{\mathsf{T}} = \Psi_2$ if $\mathbb{K} = \mathbb{R}$.) ♢

Keeping the constructions of this section in mind, we work with the setting $G \subseteq \mathrm{GL}_m(\mathbb{K})$ in the next two sections.

9.2 MLEs, Stabilizers and weak Correspondence

By convention of this thesis a Gaussian model $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$ is parametrized by its concentration matrices, i.e., the inverses of the covariance matrices. Thus, the set of covariance matrices of a Gaussian group model $\mathcal{M}_G^{\mathbf{g}}$ is

$$\left\{(g^\dagger g)^{-1} = g^{-1}(g^{-1})^\dagger \mid g \in G\right\} = \left\{hh^\dagger \mid h \in G\right\} \tag{9.5}$$

via the reparametrization $h = g^{-1} \in G$. A simple but very useful property of a Gaussian group model $\mathcal{M}_G^{\mathbf{g}}$ is that it admits transitive G-actions on its covariance respectively concentration matrices.

Proposition 9.2.1. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a subgroup. The action of G on $\mathbb{K}^m$ via left multiplication induces a transitive left action on*

(i) the set of covariance matrices from (9.5) via $(g, \Sigma) \mapsto g\Sigma g^\dagger$.

(ii) the set of concentration matrices $\mathcal{M}_G^{\mathbf{g}}$ via $(g, \Psi) := (g^{-1})^\dagger \Psi g^{-1}$.

Proof. Let $Y \sim \mathcal{N}(0, \Sigma)$ be a random vector, where $\Sigma = hh^\dagger$ and $h \in G$. Then for any $g \in G$ we have $g \cdot Y \sim \mathcal{N}(0, g\Sigma g^\dagger)$, by Lemma 6.3.1. Hence, the left action of G on $\mathbb{K}^m$ induces the G-action on the set of covariance matrices given in (i). We compute $g\Sigma g^\dagger = (gh)(gh)^\dagger$ and note that $G \to G, g \mapsto gh$ is surjective. Thus, the action from part (i) is transitive. Similarly, we get an induced transitive action on the level of concentration matrices $\Psi := \Sigma^{-1}$, since $g \cdot Y$ has concentration matrix $(g\Sigma g^\dagger)^{-1} = (g^{-1})^\dagger \Psi g^{-1}$. □

Remark 9.2.2. Regarding Proposition 9.2.1 we remark the following.

(a) Part (i) shows that Gaussian group models are *transformation families* in the sense of [BBJJ82], compare Definition 6.1.4.

(b) Instead of the left action given in part (ii) we often consider the analogous transitive *right* action on $\mathcal{M}_G^{\mathfrak{g}}$ via $(\Psi, g) \mapsto g^\dagger \Psi g$. It has the same orbits as the left action.

(c) The G-actions from Proposition 9.2.1 and part (b) are usually not free. For example, the identity $\mathrm{I}_m \in \mathcal{M}_G^{\mathfrak{g}}$ is fixed by all elements in the compact group $K = \{g \in G \mid g^\dagger = g^{-1}\}$. Furthermore, for $\mathbb{K} = \mathbb{C}$ these G-actions are *not* algebraic due to the Hermitian transpose $(\cdot)^\dagger$. ▽

In the following we study how the transitive group actions of G on $\mathcal{M}_G^{\mathfrak{g}}$ relate to ML estimation for a given tuple of samples $Y \in (\mathbb{K}^m)^n \cong \mathbb{K}^{m \times n}$.[7]

Proposition 9.2.3. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a subgroup and consider the model $\mathcal{M}_G^{\mathfrak{g}}$ with sample matrix $Y \in \mathbb{K}^{m \times n}$. Fix some $h \in G$ with $|\det(h)| = 1$. Then:*

(i) The supremum of ℓ_Y equals the supremum of $\ell_{h \cdot Y}$.

(ii) There exists an MLE given Y if and only if there exists an MLE given $h \cdot Y$. Acting with h on Y changes the set of MLEs according to the left action of h on $\mathcal{M}_G^{\mathfrak{g}}$ from Proposition 9.2.1(ii), i.e.,

$$\{\textit{MLEs given } h \cdot Y\} = (h^{-1})^\dagger \{\textit{MLEs given } Y\} h^{-1}. \tag{9.6}$$

Proof. Recall from Equation (8.4) that for any $g \in G$ we have

$$\ell_Y(g^\dagger g) = \log\left(|\det(g)|^2\right) - \frac{1}{n}\|g \cdot Y\|^2.$$

Since $|\det(h)| = 1$, it holds that $\log\left(|\det(h^{-1})|^2\right) = 0$. We compute

$$\begin{aligned}
\sup_{g \in G} \ell_Y(g^\dagger g) &= \sup_{g \in G} \log\left(|\det(g)|^2\right) + \log\left(|\det(h^{-1})|^2\right) - \frac{1}{n}\|g \cdot Y\|^2 \\
&= \sup_{g \in G} \log\left(\det\left((gh^{-1})^\dagger g h^{-1}\right)\right) - \frac{1}{n}\|(gh^{-1}) \cdot (h \cdot Y)\|^2 \\
&= \sup_{\tilde{g} \in G} \log\left(\det\left(\tilde{g}^\dagger \tilde{g}\right)\right) - \frac{1}{n}\|\tilde{g} \cdot (h \cdot Y)\|^2 = \sup_{\tilde{g} \in G} \ell_{h \cdot Y}(\tilde{g}^\dagger \tilde{g}),
\end{aligned}$$

[7] Note that Equation (9.6) appears in [AKRS21a] in the proofs of Theorems 3.10 and 3.15.

where we used the reparametrization $\tilde{g} = gh^{-1}$, equivalently $g = \tilde{g}h$, in the penultimate equality. This proves (i). Moreover, the above computation shows that $g^\dagger g = h^\dagger(\tilde{g}^\dagger\tilde{g})h$ is an MLE given Y if and only if $\tilde{g}^\dagger\tilde{g} = (h^{-1})^\dagger(g^\dagger g)h^{-1}$ is an MLE given $h \cdot Y$. This proves the second part including Equation (9.6). □

In the setting of Gaussian group models it is a natural question to ask, which role is played by group elements stabilizing Y. Given the above proposition we study the stabilizer H_Y of Y under the group $H := \{g \in G \mid |\det(g)| = 1\}$. Equation (9.6) for $h \in H_Y$ shows that the right action of H_Y on $\mathcal{M}_G^{\mathfrak{g}}$ via $(\Psi, h) \mapsto h^\dagger\Psi h$ restricts to an action on the set of MLEs given Y. Consequently, the set of MLEs given Y is the disjoint union of its H_Y-orbits. The latter is also a consequence of the following statement.

Proposition 9.2.4. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a subgroup and consider the right action $(\Psi, g) \mapsto g^\dagger\Psi g$ on $\mathcal{M}_G^{\mathfrak{g}}$. Set $H := \{g \in G \mid |\det(g)| = 1\}$ and fix some $h \in H_Y$, where $Y \in \mathbb{K}^{m\times n}$ is a sample matrix. Then:*

$$\forall\, \Psi \in \mathcal{M}_G^{\mathfrak{g}}\colon \quad \ell_Y(h^\dagger\Psi h) = \ell_Y(\Psi). \tag{9.7}$$

In particular, ℓ_Y is constant on the H_Y-orbits of $\mathcal{M}_G^{\mathfrak{g}}$ and H_Y acts on the set of MLEs given Y. The statement also holds for the subgroups $(G_{\mathrm{SL}}^{\pm})_Y$ and $(G_{\mathrm{SL}})_Y$ of H_Y.

Proof. As $h \in H_Y$ we have $h \cdot Y = Y$ and $|\det(h)| = 1$. Thus, for all $g \in G$

$$\begin{aligned}\ell_Y(h^\dagger(g^\dagger g)h) = \ell_Y((gh)^\dagger gh) &= \log\left(|\det(gh)|^2\right) - \frac{1}{n}\|(gh)\cdot Y\|^2 \\ &= \log\left(|\det(g)|^2\right) + \log\left(|\det(h)|^2\right) - \frac{1}{n}\|g\cdot(h\cdot Y)\|^2 \\ &= \log\left(|\det(g)|^2\right) - \frac{1}{n}\|g\cdot Y\|^2 = \ell_Y(g^\dagger g).\end{aligned}$$

Since any $\Psi \in \mathcal{M}_G^{\mathfrak{g}}$ is of the form $g^\dagger g$ for some $g \in G$, this shows (9.7). Hence, ℓ_Y is constant on the H_Y-orbits of $\mathcal{M}_G^{\mathfrak{g}}$. Since the MLEs given Y are exactly the $\hat{\Psi} \in \mathcal{M}_G^{\mathfrak{g}}$ with $\ell_Y(\hat{\Psi}) = \sup_{\Psi\in\mathcal{M}_G^{\mathfrak{g}}} \ell_Y(\Psi)$, we see that H_Y acts on the set of MLEs given Y. Alternatively, this can be seen via (9.6) as discussed before this proposition. The arguments are valid for any subgroup of H_Y - in particular for $(G_{\mathrm{SL}}^{\pm})_Y$ and $(G_{\mathrm{SL}})_Y$. □

We note that in general there may exist an MLE $\hat{\Psi}$ given Y and $h \in H_Y$ such that $h^\dagger\hat{\Psi}h = \hat{\Psi}$. In other words, the H_Y-action on the set of MLEs need not to be free, compare Example 9.4.3. Furthermore, the following example shows that the H_Y-action neither needs to be transitive.

Example 9.2.5. Consider the Gaussian group model $\mathcal{M}_G^{\mathfrak{g}}$ from Example 8.2.6:

$$G := \bigcup_{\tau\in\mathbb{K}^\times} \{\tau\,\mathrm{I}_2, \tau M\} \quad \text{and} \quad M := \begin{pmatrix} 1/2 & 3 \\ 1/4 & -1/2 \end{pmatrix}.$$

Assume that the sample size $n = 2$ and that the sample matrix is

$$Y = \begin{pmatrix} 6 & 2 \\ 1 & -1 \end{pmatrix}.$$

For $\mathbb{K} = \mathbb{R}$, we have that $H = G_{\mathrm{SL}}^{\pm} = \{\pm \mathrm{I}_2, \pm M\}$, compare Example 8.2.6. Since $M \cdot Y_1 = Y_1$ and $M \cdot Y_2 = -Y_2$, we have $H_Y = \{\mathrm{I}_2\}$ and $\|\pm Y\|^2 = \|\pm M \cdot Y\|^2 = 42$. Therefore, all elements of the orbit $G_{\mathrm{SL}}^{\pm} \cdot Y$ have the same norm. Hence by Theorem 8.2.3, the MLEs given Y are determined by $\lambda h^{\mathsf{T}} h$, where $h \in G_{\mathrm{SL}}^{\pm}$ and $\lambda = 2/21$ is the unique global minimum of $x \mapsto 21x - 2\log(x)$. As M is not orthogonal there are exactly two MLEs given Y, namely $\lambda \mathrm{I}_2$ and $\lambda M^{\mathsf{T}} M$. Thus, the set of MLEs given Y consists of two H_Y-orbits, because H_Y is trivial. In particular, the H_Y-action is not transitive.

For $\mathbb{K} = \mathbb{C}$, $H = \{g \in G \mid |\det(g)| = 1\} = \{\tau \mathrm{I}_2, \tau M \mid |\tau|^2 = 1\}$ is not finite. Still, the same argument as in the real case can be used to see that $H_Y = \{\mathrm{I}_2\}$. Furthermore, Theorem 8.2.3 for $G_{\mathrm{SL}} = \{\pm \mathrm{I}_2, \pm \mathrm{i} M\}$ yields that, again, $\lambda \mathrm{I}_2$ and $\lambda M^{\dagger} M = \lambda M^{\mathsf{T}} M$ are the MLEs given Y. Consequently, the H_Y-action is not transitive. ♢

The weak correspondence, Theorem 8.2.3, holds in particular for a Gaussian group model $\mathcal{M}_G^{\mathsf{g}}$, if G is closed under non-zero scalar multiples. Thanks to the group structure of G, condition (ii) there admits the following equivalent reformulation.

Lemma 9.2.6. *Let $G \subseteq \mathrm{GL}_m(\mathbb{R})$ be a subgroup that is closed under non-zero scalar multiples. Then the following are equivalent:*

- *Condition (ii) from Theorem 8.2.3;*
- *If G contains a matrix of negative determinant, then it contains an orthogonal matrix of determinant -1.*

Proof. If G only contains matrices of positive determinant, then condition (ii) is trivially satisfied as we can always choose $o = \mathrm{I}_m$, compare Remark 8.2.5. Thus, assume that G contains a matrix $\hat{g}$ with $\det(\hat{g}) < 0$.

If condition (ii) holds, then there is some orthogonal matrix $o = o(\hat{g})$ such that $\det(o^{\mathsf{T}} \hat{g}) > 0$ and $o^{\mathsf{T}} \hat{g} \in G$. The former together with $\det(\hat{g}) < 0$ yields $\det(o) = -1$, while the latter and the group properties imply $o^{\mathsf{T}} = o^{\mathsf{T}} \hat{g} \hat{g}^{-1} \in G$ and hence $o = (o^{\mathsf{T}})^{-1} \in G$. Conversely, if there is some orthogonal $o \in G$ with $\det(o) = -1$, then we have for all $g \in G$ with $\det(g) < 0$ that $\det(o^{\mathsf{T}} g) > 0$ and $o^{\mathsf{T}} g = o^{-1} g \in G$. Therefore, condition (ii) is satisfied. □

As a direct consequence of the preceding lemma and Theorem 8.2.3 we obtain the following statement.

Theorem 9.2.7 (Weak Correspondence for Gaussian group models)**.**
Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a subgroup closed under non-zero scalar multiples. If $\mathbb{K} = \mathbb{R}$ and G contains an element of negative determinant, then additionally assume that there is an orthogonal matrix in G of determinant -1. There is a correspondence

between stability under G_{SL} and maximum likelihood estimation in the model $\mathcal{M}_G^{\mathbf{g}}$ given sample matrix $Y \in \mathbb{K}^{m\times n}$:

$$\begin{array}{lcll} (a) & Y \text{ unstable} & \Leftrightarrow & \text{likelihood } \ell_Y \text{ unbounded from above} \\ (b) & Y \text{ semistable} & \Leftrightarrow & \text{likelihood } \ell_Y \text{ bounded from above} \\ (c) & Y \text{ polystable} & \Rightarrow & \text{MLE exists.} \end{array}$$

The MLEs, if they exist, are the matrices $\lambda h^\dagger h$, where $h \in G_{\mathrm{SL}}$ is such that $\|h \cdot Y\| > 0$ is minimal in $G_{\mathrm{SL}} \cdot Y$ and $\lambda \in \mathbb{R}_{>0}$ is the unique global minimum of

$$\mathbb{R}_{>0} \to \mathbb{R}, \qquad x \mapsto \frac{x}{n}\|h \cdot Y\|^2 - m\log(x).$$

We have already seen in Example 8.2.6 that, in general, we cannot drop the additional assumption of Theorem 9.2.7 if $\mathbb{K} = \mathbb{R}$. In [AKRS21a] a different, perhaps more interesting example is given to show this fact. In the following we present this example [AKRS21a, Example 3.5] in detail. For $M \in \mathbb{R}^{2\times 2}$, define

$$g(M) := \begin{pmatrix} M & & \\ & S_1 M S_1^{-1} & \\ & & S_2 M S_2^{-1} \end{pmatrix}, \text{ where } S_1 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, S_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad (9.8)$$

and set $G := \{\tau g(M) \mid M \in \mathrm{O}_2(\mathbb{R}), \tau \in \mathbb{R}^\times\}$.

Lemma 9.2.8. *G is a subgroup of $\mathrm{GL}_6(\mathbb{R})$, which contains no orthogonal matrix of determinant -1.*

Proof. First, G is a subgroup, because

$$\big(\tau_1 g(M_1)\big)^{-1}\big(\tau_2 g(M_2)\big) = (\tau_1^{-1}\tau_2) g(M_1^{\mathsf{T}} M_2) \in G$$

for all $\tau_1, \tau_2 \in \mathbb{R}^\times$ and all $M_1, M_2 \in \mathrm{O}_2(\mathbb{R})$.

Second, for a proof by contradiction assume that there are $\tau \in \mathbb{R}^\times$ and $M \in \mathrm{O}_2(\mathbb{R})$ such that $\tau g(M) \in \mathrm{O}_6(\mathbb{R})$ and $\det(\tau g(M)) = \tau^6 \det(M)^3 = -1$. Then $|\tau|^6 |\det(M)|^3 = |\tau|^6 = 1$ as $|\det(M)| = 1$. Hence, $\tau \in \{-1, 1\}$ and so $\tau^6 = 1$. Thus, we must have $\det(M) = -1$ and therefore we can write

$$M = \begin{pmatrix} a & b \\ b & -a \end{pmatrix} \text{ for some } a, b \in \mathbb{R} \text{ with } a^2 + b^2 = 1.$$

Since $\tau g(M)$ is orthogonal, we have $\tau^2 g(M)^{\mathsf{T}} g(M) = g(M^{\mathsf{T}}) g(M) = \mathrm{I}_6$. In particular, for $i = 1, 2$ we obtain

$$\mathrm{I}_2 = \big(S_i M S_i^{-1}\big)^{\mathsf{T}}\big(S_i M S_i^{-1}\big) = \big(S_i^{-1} M S_i\big)\big(S_i M S_i^{-1}\big),$$

where we used in the second equality that S_i, S_i^{-1} and M are symmetric. If $i = 2$, then the above equation specializes to

$$\mathrm{I}_2 = \begin{pmatrix} a & 2b \\ 1/2 b & -a \end{pmatrix}\begin{pmatrix} a & 1/2 b \\ 2b & -a \end{pmatrix}.$$

The upper left entry computes as $1 = a^2 + 4b^2$ and we deduce $b = 0$ using $1 = a^2 + b^2$. Now, for $i = 1$ we get

$$\mathrm{I}_2 = \frac{1}{9}\begin{pmatrix} 5a & 4a \\ -4a & -5a \end{pmatrix}\begin{pmatrix} 5a & -4a \\ 4a & -5a \end{pmatrix}$$

and the upper right entry is $0 = -(41/9)a^2$, which contradicts $a^2 = 1$. □

Example 9.2.9 ([AKRS21a, Example 3.5])**.** Consider

$$G := \left\{\tau g(M) \mid M \in \mathrm{O}_2(\mathbb{R}), \tau \in \mathbb{R}^\times\right\},$$

where $g(M)$ is defined as in Equation (9.8). Then G is a subgroup of $\mathrm{GL}_6(\mathbb{R})$ that contains no orthogonal matrix of determinant -1, see Lemma 9.2.8. For the Gaussian group model $\mathcal{M}_G^{\mathsf{g}}$ consider the tuple of four samples given by

$$Y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 2\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 2\sqrt{5} \\ 0 & 0 & \frac{6\sqrt{5}}{5} & \frac{8\sqrt{5}}{5} \end{bmatrix}, \quad \text{with} \quad S_Y = \frac{1}{4}\sum_{i=1}^{4} Y_i Y_i^{\mathsf{T}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & S_2 & 0 \\ 0 & 0 & S_1^2 \end{bmatrix}.$$

By Equation (8.8), the supremum of ℓ_Y can be computed as a double infimum. The inner infimum $\inf_{h \in G_{\mathrm{SL}}^{\pm}} \|h \cdot Y\|^2$ can be rewritten as minimizing the trace $\mathrm{tr}(g^{\mathsf{T}} g S_Y)$ over matrices $g \in G_{\mathrm{SL}}^{\pm}$, by (8.3):

$$\begin{aligned} \inf_{h \in G_{\mathrm{SL}}^{\pm}} \|h \cdot Y\|^2 = 4 \cdot \inf_{M \in \mathrm{O}_2(\mathbb{R})} \big[& \mathrm{tr}\left((S_1 M S_1^{-1})^{\mathsf{T}} (S_1 M S_1^{-1}) S_2\right) \\ & + \mathrm{tr}\left((S_2 M S_2^{-1})^{\mathsf{T}} (S_2 M S_2^{-1}) S_1^2\right) \big]. \end{aligned}$$

We can parametrize the 2×2 special orthogonal matrices by P and the 2×2 orthogonal matrices of determinant -1 by Q where

$$P = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}, \qquad Q = \begin{bmatrix} -a & -b \\ -b & a \end{bmatrix}, \quad \text{with} \quad a, b \in \mathbb{R}, \quad \text{and} \quad a^2 + b^2 = 1.$$

Then the minimization problems over G_{SL} and G_{SL}^{-} can be rewritten as

$$\inf_{h \in G_{\mathrm{SL}}} \frac{1}{4}\|h \cdot Y\|^2 = \min_{a^2+b^2=1} \left(13a^2 - \frac{44}{3}ab + \frac{419}{12}b^2\right),$$

$$\inf_{h \in G_{\mathrm{SL}}^{-}} \frac{1}{4}\|h \cdot Y\|^2 = \min_{a^2+b^2=1} \left(\frac{71}{3}a^2 - \frac{28}{3}ab + \frac{97}{4}b^2\right).$$

We point out that the minimum is justified by compactness of the unit circle. Note that $0 \leq (a-b)^2$ implies $ab \leq (1/2)(a^2+b^2) = 1/2$, equivalently $-ab \geq -1/2$. Thus, substituting $b^2 = 1 - a^2$ in the latter minimum, we see that

$$\begin{aligned} \frac{71}{3}a^2 + \frac{97}{4}(1-a^2) - \frac{28}{3}ab &\geq \frac{97}{4} + \left(\frac{71}{3} - \frac{97}{4}\right)a^2 - \frac{28}{3} \cdot \frac{1}{2} \\ &\geq \frac{97}{4} + \left(\frac{71}{3} - \frac{97}{4}\right) - \frac{28}{3} \cdot \frac{1}{2} = 19, \end{aligned}$$

where we used $71/3 - 97/4 < 0$ with $a^2 \leq 1$ in the second inequality. In contrast, setting $a = 1$ and $b = 0$ in the former minimum gives a value of 13. Hence, $\inf_{h \in G_{\mathrm{SL}}} \|h \cdot Y\|^2 < \inf_{h \in G_{\mathrm{SL}}^-} \|h \cdot Y\|^2$. Multiplying Y by a fixed matrix in G_{SL}^- gives a tuple of samples where the strict inequality is reversed, and the infimum is witnessed *only* at the component G_{SL}^-. Altogether, this example shows that the extra condition for $\mathbb{K} = \mathbb{R}$ in Theorem 9.2.7, equivalently condition (ii) of Theorem 8.2.3, cannot be dropped. ♢

9.3 Self-adjoint Zariski closed groups

In this section we study Gaussian group models $\mathcal{M}_G^{\mathbf{g}}$ with Zariski closed and self-adjoint[8] subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$. Such models arise naturally from rational representations of reductive groups, compare Remark 9.1.4. We have already seen Gaussian group models with Zariski closed self-adjoint subgroup in Examples 9.1.1, 9.1.3, 9.1.5 and 9.1.6.

We stress that the assumptions are properties of the parametrizing subgroup, not the model itself. For example, the saturated model $\mathrm{PD}_m(\mathbb{K})$ is induced by the Zariski closed self-adjoint group $\mathrm{GL}_m(\mathbb{K})$. On the other hand, $\mathrm{PD}_m(\mathbb{K}) = \mathcal{M}_{\mathrm{B}_m(\mathbb{K})}^{\mathbf{g}}$ and the group $\mathrm{B}_m(\mathbb{K})$ of invertible upper triangular matrices is not self-adjoint.

Let us start our study with a simple observation. If G is self-adjoint then

$$\mathcal{M}_G^{\mathbf{g}} = \{g^\dagger g \mid g \in G\} = \{hh^\dagger \mid h \in G\}$$

using the reparametrization $h = g^\dagger \in G$. Statistically this equality means that the set of concentration matrices $\mathcal{M}_G^{\mathbf{g}}$ is equal to the set of covariance matrices of the model $\mathcal{M}_G^{\mathbf{g}}$, compare Equation (9.5).

Next, we reformulate Theorem 1.2.18 to illustrate what the assumption "Zariski closed and self-adjoint" on G means geometrically for the model $\mathcal{M}_G^{\mathbf{g}}$.

Theorem 9.3.1. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint subgroup. Then the Gaussian group model $\mathcal{M}_G^{\mathbf{g}}$ is a totally geodesic submanifold of $\mathrm{PD}_m(\mathbb{K})$. Moreover, $\mathcal{M}_G^{\mathbf{g}}$ is a CAT(0)-symmetric space and equal to $G \cap \mathrm{PD}_m(\mathbb{K})$. In particular, $\mathcal{M}_G^{\mathbf{g}}$ is Euclidean closed in $\mathrm{PD}_m(\mathbb{K})$.*

Conversely, if $\mathcal{M} \subseteq \mathrm{PD}_m(\mathbb{K})$ is a totally geodesic submanifold with $\mathrm{I}_m \in \mathcal{M}$, then $\mathcal{M} = \mathcal{M}_G^{\mathbf{g}}$ for a Euclidean closed self-adjoint subgroup $G \subseteq \mathrm{GL}_m(\mathbb{K})$.

As a consequence of these strong geometric properties and Example 1.2.21 the negative of the log-likelihood is a geodesically convex function on $\mathcal{M}_G^{\mathbf{g}}$. This was observed for matrix normal models in [Wie12]. Geodesic convexity has been used with great benefit for matrix and tensor normal models in [FORW21], see Subsection 9.6.1.

Since $\mathcal{M}_G^{\mathbf{g}}$ is closed in $\mathrm{PD}_m(\mathbb{K})$ if G is Zariski closed and self-adjoint, it does not make sense to speak of *extended* MLEs of $\mathcal{M}_G^{\mathbf{g}}$, compare Remark 6.3.4.

The following lemma ensures that the additional assumption for $\mathbb{K} = \mathbb{R}$ needed in Theorem 9.2.7 is satisfied, if $G \subseteq \mathrm{GL}_m(\mathbb{R})$ is Zariski closed and self-adjoint.

[8] We recall that self-adjoint means that for all $g \in G$ one also has $g^\dagger \in G$.

Lemma 9.3.2 ([AKRS21a, Lemma 3.8])**.** *Let $G \subseteq \mathrm{GL}_m(\mathbb{R})$ be a Zariski closed self-adjoint group, closed under non-zero scalar multiples. If there is an element of G with negative determinant, then G contains an orthogonal matrix of determinant -1. In particular, the weak correspondence, Theorem 8.2.3 respectively Theorem 9.2.7, holds for G_{SL}.*

Proof. Pick $g \in G$ with $\det(g) < 0$. Since G is Zariski closed and self-adjoint, the polar decomposition can be carried out in G, by Theorem 1.2.16. In particular, there is an orthogonal $o \in G$ and a positive definite $p \in G$ such that $g = op$. Then $\det(g) < 0$ implies $\det(o) < 0$, i.e., $\det(o) = -1$. □

In general, the right action of $(G_{\mathrm{SL}})_Y$ from Proposition 9.2.4 on the set of MLEs given Y needs not to be transitive, see Example 9.2.5. However, in the case of self-adjoint groups we have the following sufficient criterion.

Proposition 9.3.3 ([AKRS21a, Propositions 3.9 and 3.14])**.** *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint subgroup which is closed under non-zero scalar multiples. Consider the model $\mathcal{M}_G^{\mathbf{g}}$ with tuple of samples $Y \in (\mathbb{K}^m)^n$. If $\hat{\Psi}$ is an MLE given Y, then*

$$\{\textit{MLEs given } Y\} = \left\{g^\dagger \hat{\Psi} g \mid g \in (G_{\mathrm{SL}})_Y\right\}, \tag{9.9}$$

i.e., the action of $(G_{\mathrm{SL}})_Y$ from Proposition 9.2.4 on the set of MLEs given Y is transitive.

Proof. The weak correspondence, Theorem 9.2.7, holds for G_{SL} by Lemma 9.3.2. Hence, $\hat{\Psi} = \lambda \hat{h}^\dagger \hat{h}$ and any other MLE given Y is of the form $\lambda (h')^\dagger h'$, where $\lambda > 0$ is uniquely determined and $\hat{h}, h' \in G_{\mathrm{SL}}$ satisfy

$$\|h' \cdot Y\|^2 = \inf_{h \in G_{\mathrm{SL}}} \|h \cdot Y\|^2 = \|\hat{h} \cdot Y\|^2. \tag{9.10}$$

$G_{\mathrm{SL}} \subseteq \mathrm{GL}_m(\mathbb{K})$ is Zariski closed and self-adjoint, because G is. Moreover, for $K = \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$ the matrices in $K \cap G_{\mathrm{SL}}$ act isometrically on $\mathbb{K}^{m \times n}$. Therefore, we can apply Kempf-Ness, Theorem 2.2.13(b), to Equation (9.10) and obtain some $k \in K \cap G_{\mathrm{SL}}$ with $k \cdot (\hat{h} \cdot Y) = h' \cdot Y$. Thus, $g := \hat{h}^{-1} k^{-1} h' \in (G_{\mathrm{SL}})_Y$ and using $h' = k\hat{h}g$ we deduce $\lambda (h')^\dagger h' = g^\dagger (\lambda \hat{h}^\dagger \hat{h}) g = g^\dagger \hat{\Psi} g$. □

The following statement is implicitly contained in the proof of [AKRS21a, Theorems 3.10 and 3.15]. Part (i) is explicitly stated and proven in [DMW22, Corollary 2.5].

Proposition 9.3.4. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint subgroup, which is closed under non-zero scalar multiples. Assume that the tuple of samples $Y \in (\mathbb{K}^m)^n$ has an MLE in the model $\mathcal{M}_G^{\mathbf{g}}$. Then:*

(i) If Y has a unique *MLE, then the stabilizer $(G_{\mathrm{SL}})_Y$ is compact.*

(ii) Y has either *a unique MLE* or *infinitely many MLEs.*

Proof. Since Y has an MLE in $\mathcal{M}_G^{\mathsf{g}}$, there is some $h \in G_{\mathrm{SL}}$ such that $h \cdot Y$ is of minimal norm in $G_{\mathrm{SL}} \cdot Y$, by Theorem 9.2.7. Then also $h \cdot Y$ has an MLE in $\mathcal{M}_G^{\mathsf{g}}$, using Proposition 9.2.3(ii), and by Equation (9.6) the set of MLEs given Y has the same cardinality as the set of MLEs given $h \cdot Y$. Moreover, for stabilizers it holds that $(G_{\mathrm{SL}})_{h\cdot Y} = h(G_{\mathrm{SL}})_Y h^{-1}$. As conjugation via h is a homeomorphism we deduce that $(G_{\mathrm{SL}})_Y$ is compact if and only if $(G_{\mathrm{SL}})_{h\cdot Y}$ is compact. Altogether, we argued that, after replacing Y by $h \cdot Y$, we can assume that Y is of minimal norm in its G_{SL}-orbit.

Now, if Y is of minimal norm in $G_{\mathrm{SL}} \cdot Y$, then $\lambda \, \mathrm{I}_m$ is an MLE given Y using Theorem 9.2.7. Proposition 9.3.3 yields that

$$\{\text{MLEs given } Y\} = \{\lambda g^\dagger g \mid g \in (G_{\mathrm{SL}})_Y\}. \tag{9.11}$$

Thus, $\lambda \, \mathrm{I}_m$ is the unique MLE given Y if and only if $(G_{\mathrm{SL}})_Y \subseteq K = \{g \in G \mid g^\dagger g = \mathrm{I}_m\}$. If $(G_{\mathrm{SL}})_Y \subseteq K$, then $(G_{\mathrm{SL}})_Y$ is compact as it is Euclidean (even Zariski) closed in the compact group K. This shows part (i).

On the other hand, assume there is another MLE $\lambda g^\dagger g$, $g \in (G_{\mathrm{SL}})_Y$, with $g^\dagger g \neq \mathrm{I}_m$. The positive definite matrix $g^\dagger g$ admits a decomposition $udu^\dagger$, where $u, d \in \mathrm{GL}_m(\mathbb{K})$ such that $u^{-1} = u^\dagger$ and d is diagonal with *real positive* entries. At least one of the positive diagonal entries of d is not equal to one, as $g^\dagger g \neq \mathrm{I}_m$. This implies that $\{d^N \mid N \in \mathbb{Z}\}$ is an infinite cyclic group. Consequently, the set $\{(g^\dagger g)^{2N} \mid N \in \mathbb{Z}\}$ is infinite. Since G is Zariski closed and self-adjoint, also G_{SL} is, and $K \cap G_{\mathrm{SL}}$ acts isometrically. Hence, Lemma 2.2.16 yields that the stabilizer $(G_{\mathrm{SL}})_Y$ is self-adjoint as Y is of minimal norm in $G_{\mathrm{SL}} \cdot Y$. Thus, for any $g \in (G_{\mathrm{SL}})_Y$ we have $g^\dagger \in (G_{\mathrm{SL}})_Y$ and hence $(g^\dagger g)^N \in (G_{\mathrm{SL}})_Y$ for all $N \in \mathbb{Z}$. Finally, we get infinitely many MLEs

$$\lambda\big((g^\dagger g)^N\big)^\dagger (g^\dagger g)^N = \lambda (g^\dagger g)^N (g^\dagger g)^N = \lambda (g^\dagger g)^{2N}, \quad N \in \mathbb{Z}$$

by (9.11), which ends the proof of part (ii). □

We stress the importance of G being self-adjoint to ensure part (ii) of Proposition 9.3.4. This assumption is needed to conclude that the MLE is unique from the fact that there are finitely many MLEs. Indeed, the following example exhibits a reductive group G, closed under non-zero scalar multiples, and a sample Y with a finite number of MLEs in $\mathcal{M}_G^{\mathsf{g}}$, but not a unique MLE.

Example 9.3.5 ([AKRS21a, Example 3.12])**.** Recall the Gaussian group model $\mathcal{M}_G^{\mathsf{g}}$ from Examples 8.2.6 and 9.2.5:

$$G := \bigcup_{\tau \in \mathbb{K}^\times} \{\tau \, \mathrm{I}_2, \tau M\} \quad \text{and} \quad M := \begin{pmatrix} 1/2 & 3 \\ 1/4 & -1/2 \end{pmatrix}.$$

Assume we have a single sample $Y = (6, 1)^\mathsf{T} \in \mathbb{K}^m$. Remember that we have $G_{\mathrm{SL}}^{\pm} = \{\pm \mathrm{I}_2, \pm M\}$ if $\mathbb{K} = \mathbb{R}$ and $G_{\mathrm{SL}}^{\pm} = \{\pm \mathrm{I}_2, \pm \mathrm{i} \, \mathrm{I}_2, \pm M, \pm \mathrm{i} M\}$ if $\mathbb{K} = \mathbb{C}$, compare Example 8.2.6. The MLEs of Y are $\lambda h^\dagger h$, where $\lambda > 0$ is uniquely determined and $h \in G_{\mathrm{SL}}^{\pm}$ minimizes the norm $\|h \cdot Y\|$ in $G_{\mathrm{SL}}^{\pm} \cdot Y$, by Theorem 8.2.3. Since $M \cdot Y = Y$, we see that all elements in $G_{\mathrm{SL}}^{\pm} \cdot Y$ have the same norm and that there are exactly *two* MLEs given Y, namely $\lambda \, \mathrm{I}_2$ and $\lambda M^\dagger M = \lambda M^\mathsf{T} M$.

Note that the group G is reductive: it is the direct product of $\mathbb{K}^\times$ and the cyclic group $\{\mathrm{I}_2, M\}$. Therefore, there exists an inner product $\langle\cdot,\cdot\rangle$, on $\mathbb{K}^m$ such that G is self-adjoint with respect to $\langle\cdot,\cdot\rangle$, see Theorem 1.3.10. Hence, Proposition 9.3.4 holds for G and $(\mathbb{K}^m, \langle\cdot,\cdot\rangle)$. In particular, Y has a unique MLE. We stress that the *statistical meaning has changed* as the log-likelihood is now computed with respect to a *different* norm: note that $\langle\cdot,\cdot\rangle$ is *not* the standard inner product. This illustrates the general Remarks 9.1.2 and 9.1.4. ♢

The weak correspondence, Theorem 9.2.7, gives a first dictionary between stability notions of G_{SL} and ML estimation for the Gaussian group model $\mathcal{M}_G^{\mathrm{g}}$. We can enlarge this dictionary, if the group $G \subseteq \mathrm{GL}_m(\mathbb{K})$ is additionally Zariski closed and self-adjoint. If $\mathbb{K} = \mathbb{C}$ we obtain a list of *four* equivalences in Theorem 9.3.6(a)–(d), which we call the *full correspondence*. If $\mathbb{K} = \mathbb{R}$ the converse of Theorem 9.3.6(d) does not hold in general, compare Example 9.4.3 from the next section, and we speak instead of the *strong correspondence*.[9]

Theorem 9.3.6 ([AKRS21a, Theorems 3.10 and 3.15])**.**
Let $Y \in (\mathbb{K}^m)^n$ be a tuple of samples, and $G \subseteq \mathrm{GL}_m(\mathbb{K})$ a Zariski closed self-adjoint group that is closed under non-zero scalar multiples. The stability under the action of G_{SL} on $(\mathbb{K}^m)^n$ is related to ML estimation for the Gaussian group model $\mathcal{M}_G^{\mathrm{g}}$ as follows.

(a)	Y *unstable*	$\Leftrightarrow$	ℓ_Y *not bounded from above*
(b)	Y *semistable*	$\Leftrightarrow$	ℓ_Y *bounded from above*
(c)	Y *polystable*	$\Leftrightarrow$	*MLE exists*
(d)	Y *stable*	$\Rightarrow$	*unique MLE exists*

If $\mathbb{K} = \mathbb{C}$, then equivalence *holds in (d).*

Proof. By Lemma 9.3.2 the weak correspondence holds for G_{SL}, see Theorem 9.2.7. Thus, parts (a), (b) and the forward direction of (c) hold. To prove the converse of (c), assume that there is an MLE given Y. By Theorem 9.2.7, this MLE is of the form $\lambda h^\dagger h$ for some $h \in G_{\mathrm{SL}}$ such that $\|h \cdot Y\| > 0$ is minimal in $G_{\mathrm{SL}} \cdot Y$. Hence, $Y \neq 0$ and Kempf-Ness, Theorem 2.2.13, implies that $G_{\mathrm{SL}} \cdot Y$ is Euclidean closed, i.e., Y is polystable. Note that we can apply Kempf-Ness as G_{SL} is Zariski closed and self-adjoint, because G is.

If Y is stable, then there is at least one MLE given Y, by part (c), and $(G_{\mathrm{SL}})_Y$ is finite. The latter and Equation (9.9) imply that there are finitely many MLEs given Y. Hence, Y has a unique MLE, by Proposition 9.3.4(ii). This shows the implication in (d). Finally, assume $\mathbb{K} = \mathbb{C}$ and that there is a unique MLE given Y. By Proposition 9.3.4(i), the stabilizer $(G_{\mathrm{SL}})_Y \subseteq \mathbb{C}^{m\times m}$ is compact, but it is also Zariski-closed in $\mathrm{GL}_m(\mathbb{C})$ (defined by the equations of G and the equations $g \cdot Y = Y$). Hence, $(G_{\mathrm{SL}})_Y$ is Zariski closed and compact in $\mathbb{C}^{m\times m}$, so it must be finite. Furthermore, Y is polystable by part (c). We conclude that Y is stable, which ends the proof. □

[9] Like the name *weak correspondence*, the names *strong correspondence* respectively *full correspondence* where coined by Anna Seigal during discussions with Gergely Bérczi, Eloise Hamilton, Visu Makam and myself.

Remark 9.3.7 (based on [AKRS21a, Remark 3.11]). Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint group that is closed under non-zero scalar multiples. We argue that the results in Theorems 9.2.7 and 9.3.6, and in Propositions 9.3.3 and 9.3.4 are unchanged if we replace G_{SL} by its Euclidean identity component G_{SL}°. First, the stability notions under both groups coincide, by Proposition 2.2.17. Second, by the latter proposition any $h \in G_{\mathrm{SL}}$ is of the form $h = kh'$ for some $k \in K$, $h' \in G_{\mathrm{SL}}^\circ$. Therefore, $\mathrm{cap}_{G_{\mathrm{SL}}}(Y) = \mathrm{cap}_{G_{\mathrm{SL}}^\circ}(Y)$ and as $h^\dagger h = (h')^\dagger k^\dagger k h' = (h')^\dagger h'$ we do not loose any MLEs (if they exist) when replacing G_{SL} by G_{SL}°. Third, we can apply Kempf-Ness also to G°, compare Theorem 2.2.13, and deduce Proposition 9.3.3 similarly. Finally, one can verify that $(G_{\mathrm{SL}})_Y$ is compact if and only if $(G_{\mathrm{SL}}^\circ)_Y$ is. Altogether, this shows the claim.

In fact, we can also replace G_{SL} by any Zariski closed self-adjoint subgroup H of G that satisfies $H^\circ = G_{\mathrm{SL}}^\circ$, because we can repeat the above argument for H and $H^\circ = G_{\mathrm{SL}}^\circ$. We may not have such choices for groups that are not Zariski closed and self-adjoint, see Examples 8.2.6 and 9.2.9. ▽

We illustrate how Theorem 9.3.6 can be used to recover standard knowledge on the saturated Gaussian model $\mathrm{PD}_m(\mathbb{K})$ from Example 6.3.8.

Example 9.3.8. The group $G = \mathrm{GL}_m(\mathbb{K})$ is Zariski closed, self-adjoint and closed under non-zero scalar multiples. Therefore, we can use Theorem 9.3.6 to study ML estimation for the saturated model $\mathcal{M}_G^{\mathbf{g}} = \mathrm{PD}_m(\mathbb{K})$. We have already studied the action of $G_{\mathrm{SL}} = \mathrm{SL}_m(\mathbb{K})$ on $\mathbb{K}^{m\times n}$ via left multiplication in Example 1.4.4. There we have seen that any Y is unstable if $n < m$. Thus, for all Y the log-likelihood ℓ_Y is not bounded from above if $n < m$, by Theorem 9.3.6(a). On the other hand, if $n \geq m$ then Y is stable if and only if Y has full row rank, and it is unstable otherwise. Thus, for $n \geq m$ almost all Y are stable and have a unique MLE by Theorem 9.3.6(d). Altogether, this recovers the results from Example 6.3.8 on ML thresholds:

$$\mathrm{mlt}_b(\mathcal{M}_G^{\mathbf{g}}) = \mathrm{mlt}_e(\mathcal{M}_G^{\mathbf{g}}) = \mathrm{mlt}_u(\mathcal{M}_G^{\mathbf{g}}) = m.$$

Now, let $n \geq m$. The above shows that there exists an MLE given Y (which is then unique) if and only if Y has full row rank. The latter is equivalent to the sample covariance matrix

$$S_Y = \frac{1}{n}\sum_{i=1}^n Y_i Y_i^\dagger = \frac{1}{n} YY^\dagger \in \mathbb{K}^{m\times m}$$

being invertible. Remember from Example 6.3.8 that the MLE given Y, if it exists, is S_Y^{-1}. We deduce this from the weak correspondence, Theorem 9.2.7. For this, fix a sample matrix Y such that S_Y is invertible. Recall from Example 2.2.15 that for $M := YY^\dagger = nS_Y$ and $h := \det(M)^{1/(2m)} M^{-1/2} \in \mathrm{SL}_m(\mathbb{K})$ we have

$$\gamma := \mathrm{cap}_{\mathrm{SL}_m(\mathbb{K})}(Y) = \|h \cdot Y\|^2 = m \det(M)^{1/m}.$$

Therefore, Theorem 9.2.7 yields that $\lambda h^\dagger h$ is the MLE given Y where λ minimizes $x \mapsto \frac{\gamma}{n} x - m\log(x)$. Lemma 8.2.2(ii) shows that $\lambda = mn/\gamma$ and hence

$$\lambda h^\dagger h = \frac{mn}{m\det(M)^{1/m}} \det(M)^{1/m} M^{-1} = n(nS_Y)^{-1} = S_Y^{-1}$$

is the MLE given Y. ♢

9.3.1 Algorithmic Implications

In the following we discuss algorithmic consequences of Theorem 9.3.6. Scaling algorithms are iterative algorithms existing both in statistics and in invariant theory. We already discussed scaling algorithms for computational invariant theory in detail, compare Section 3.2. In statistics one usually refers to scaling algorithms as iterative proportional scaling (IPS)[10].

In Section 7.3, we drew a connection between norm minimization in invariant theory and IPS for log-linear models, also see Figure 7.1. The starting point of this figure is Sinkhorn scaling, an alternating minimization method. On the statistical side, it can be seen as an instance of IPS for the independence model, which generalizes to IPS for any log-linear model. On the invariant theory side it generalizes to norm minimization under a torus action.

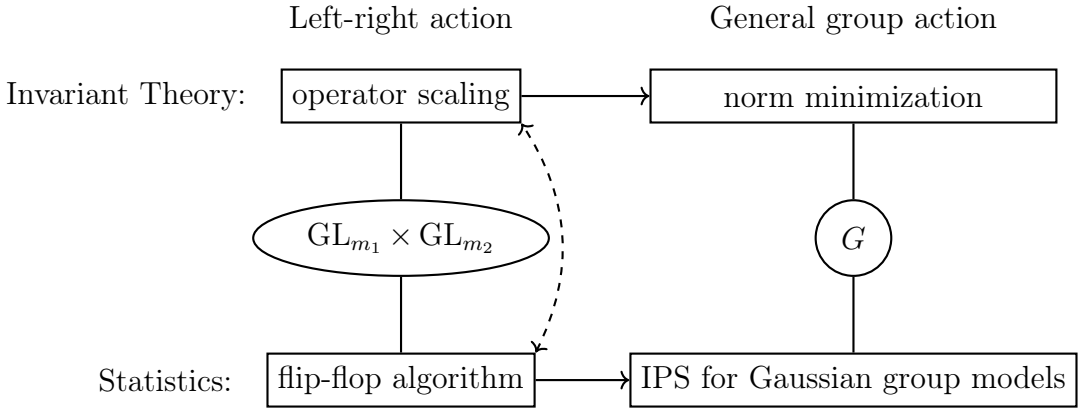


Figure 9.1: [AKRS21a, Figure 1] Overview of different scaling algorithms. For the invariant theory algorithms, we use matrices of determinant one, e.g. $\mathrm{SL}_{m_1} \times \mathrm{SL}_{m_2} \subseteq \mathrm{GL}_{m_1} \times \mathrm{GL}_{m_2}$.

A Gaussian analogue of Figure 7.1 is given in Figure 9.1. Namely, the idea of Sinkhorn scaling generalizes to operator scaling (Algorithm 3.2, [Gur04a; GGOW16]) from invariant theory respectively to the flip-flop algorithm (Algorithm 9.1, [Dut99; LZ05]); see the left of Figure 9.1. In Subsection 9.4.4 below we show that these methods are essentially equivalent. Furthermore, the flip-flop algorithm can be thought of as an instance of IPS [FM81; Cra98].

For complex Gaussian group models $\mathcal{M}_G^{\mathfrak{g}}$ with $G \subseteq \mathrm{GL}_m(\mathbb{C})$ Zariski closed and self-adjoint, we can use the geodesically convex first and second order methods from [BFG+19] to solve Norm Minimization 3.1.3 respectively the Scaling Problem 3.1.4.[11] These algorithms can be thought of as generalizations of operator scaling. Altogether, the above discussion and the comparison of Figures 7.1 and 9.1 motivates to regard these geodesically convex methods as IPS for Gaussian group models (where G is Zariski closed and self-adjoint).

[10] also called *iterative proportional fitting (IPF)*

[11] Note that [BFG+19] requires that G is Zariski closed and self-adjoint. Moreover, without this assumption we do not have a moment map and hence cannot consider the scaling problem.

Remark 9.3.9 (Algorithms for real Gaussian group models). In invariant theory scaling algorithms are usually designed over $\mathbb{C}$ and they minimize the norm over the complex orbit. However, often each update is defined over $\mathbb{R}$ if the input is real, and hence these can also be used for real Gaussian group models with G being Zariski closed and self-adjoint. This crucially uses that in this situation the capacity over $\mathbb{R}$ equals the one over $\mathbb{C}$, compare Proposition 2.2.18.

For example, the alternating minimization method for operator and tensor scaling from [BGO+18] always stays over $\mathbb{R}$ if the input is real. The same applies to the first order method from [BFG+19]. ▽

9.4 Applications to Matrix Normal Models

In this section we illustrate how to apply the theory from Section 9.3 to study matrix normal models. Recall from Example 6.3.9 that a matrix normal model is a sub-model of $\mathrm{PD}_{m_1m_2}(\mathbb{K})$ whose concentration matrices factor as a Kronecker product:

$$\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2) = \left\{ \Psi_1 \otimes \Psi_2 \mid \Psi_j \in \mathrm{PD}_{m_j}(\mathbb{K}) \right\}.$$

Moreover, recall from Example 9.1.6 that the log-likelihood (6.8) computes as

$$\ell_Y(\Psi_1 \otimes \Psi_2) = m_2 \log\det(\Psi_1) + m_1 \log\det(\Psi_2) - \frac{1}{n} \operatorname{tr}\left(\Psi_1 \sum_{i=1}^{n} Y_i \Psi_2^{\mathsf{T}} Y_i^{\dagger} \right). \quad (9.12)$$

An MLE is a concentration matrix $\hat{\Psi}_1 \otimes \hat{\Psi}_2 \in \mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$ that maximizes the log-likelihood function.

9.4.1 Relating norm minimization to ML estimation

In the following we study matrix normal models using the left-right action of $\mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K})$ on $(\mathbb{K}^{m_1 \times m_2})^n$. Remember that the action[12] is given by

$$g \cdot Y := \left(g_1 Y_1 g_2^{\mathsf{T}}, \ldots, g_1 Y_n g_2^{\mathsf{T}} \right), \quad (9.13)$$

where $g = (g_1, g_2) \in \mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K})$ and $Y = (Y_1, \ldots, Y_n) \in (\mathbb{K}^{m_1 \times m_2})^n$. We have seen in Example 9.1.6 that for $n = 1$ this algebraic action, after appropriate identification, induces the rational representation

$$\varrho\colon \mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K}) \to \mathrm{GL}_{m_1 m_2}(\mathbb{K}), \quad (g_1, g_2) \mapsto g_1 \otimes g_2.$$

Hence, for $G := \varrho\big(\mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K}) \big)$ we obtain $\mathcal{M}^{\mathbf{g}}_G = \mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$, the matrix normal model. The subgroup $G \subseteq \mathrm{GL}_{m_1 m_2}(\mathbb{K})$ is Zariski closed,[13] self-adjoint and closed under non-zero scalar multiples. Thus, the results from Section 9.3

[12] We note that the transposes in (9.13) are also used in the complex case to ensure an *algebraic* action, compare Example 9.1.6.

[13] This is even true over $\mathbb{R}$: if $g_j \in \mathrm{GL}_{m_j}(\mathbb{C})$ such that $g_1 \otimes g_2 \in \mathrm{GL}_{m_1 m_2}(\mathbb{R})$, then there exist $h_j \in \mathbb{R}^{m_j \times m_j}$ with $h_1 \otimes h_2 = g_1 \otimes g_2$. The latter uses that Segre embeddings are surjective on $\mathbb{R}$-points. Now, $0 \neq \det(g_1 \otimes g_2) = \det(h_1 \otimes h_2) = (\det h_1)^{m_2} (\det h_2)^{m_1}$ yields $\det(h_j) \neq 0$.

apply to the action of G_{SL}. However, it is possible and more convenient to directly work with the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$. The following theorem makes this precise.

Theorem 9.4.1 (Strong/Full Correspondence, [AKRS21a, Theorem 4.1])**.** *Let $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ be a matrix tuple. The supremum of the log-likelihood ℓ_Y in (9.12) over $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$ is given by the double infimum*

$$-\inf_{x \in \mathbb{R}_{>0}} \left(\frac{x}{n} \left(\inf_{h \in \mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})} \|h \cdot Y\|^2 \right) - m_1 m_2 \log(x) \right). \tag{9.14}$$

The MLEs given Y, if they exist, are the matrices of the form $\lambda h_1^\dagger h_1 \otimes h_2^\dagger h_2$, where $h = (h_1, h_2)$ minimizes $\|h \cdot Y\|$ under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$, and $\lambda \in \mathbb{R}_{>0}$ is the unique value that minimizes the outer infimum.

If $\lambda h_1^\dagger h_1 \otimes h_2^\dagger h_2$ is an MLE, then every (g_1, g_2) in the $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ stabilizer of Y yields an MLE via

$$(g_1 \otimes g_2)^\dagger \big(\lambda h_1^\dagger h_1 \otimes h_2^\dagger h_2\big)(g_1 \otimes g_2) = \lambda \big(g_1^\dagger h_1^\dagger h_1 g_1\big) \otimes \big(g_2^\dagger h_2^\dagger h_2 g_2\big)$$

and, conversely, every MLE given Y is of this form.

The stability under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ is related to ML estimation via:

$$\begin{array}{lccc} (a) & Y \text{ unstable} & \Leftrightarrow & \ell_Y \text{ not bounded from above} \\ (b) & Y \text{ semistable} & \Leftrightarrow & \ell_Y \text{ bounded from above} \\ (c) & Y \text{ polystable} & \Leftrightarrow & \text{MLE exists} \\ (d) & Y \text{ stable} & \Rightarrow & \text{MLE exists uniquely} \end{array}$$

If $\mathbb{K} = \mathbb{C}$, then equivalence holds in (d).

Proof. The proof uses the notation introduced above. Since $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ is Euclidean connected, also $H := \varrho(\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})) \subseteq G$ is Euclidean connected. In fact, it is the Euclidean identity component of G_{SL}: for $\mathbb{K} = \mathbb{C}$ the group H is Zariski closed by Proposition 1.1.7, and one verifies $H = G_{\mathrm{SL}}$. If $\mathbb{K} = \mathbb{R}$, one may have $H \subsetneq G_{\mathrm{SL}}$,[14] but still Corollary 1.2.6 applies and yields $H = G_{\mathrm{SL}}^\circ$.[15] Thus Theorem 9.2.7, Proposition 9.3.3 and Theorem 9.3.6 apply to H as well, by Remark 9.3.7. Furthermore, when restricted to $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ the kernel of ϱ is finite. Hence, the stability notions in Definition 1.4.1(a)–(d) coincide for $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ and H, compare Remark 1.4.2. Thus, we can consider $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ instead of its image H under ϱ. □

Remark 9.4.2. Example 9.1.5 shows that the tensor scaling action (Example 1.3.5) of $\mathrm{GL}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{GL}_{m_d}(\mathbb{K})$ on $\mathbb{K}^{m_1} \otimes \cdots \otimes \mathbb{K}^{m_d}$ gives the tensor normal model $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, \ldots, m_d)$. Analogous arguments as for matrix normal models show that a similar version of Theorem 9.4.1 for $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, \ldots, m_d)$ holds via restricting the tensor scaling action to $\mathrm{SL}_{m_1}(\mathbb{K}) \times \cdots \times \mathrm{SL}_{m_d}(\mathbb{K})$. ▽

[14] This happens, e.g., if $m_1 = m_2 = 2$: one verifies that $\mathrm{diag}(-1, 1) \otimes \mathrm{diag}(-1, 1) \in G_{\mathrm{SL}} \backslash H$.

[15] The corresponding proof in [AKRS21a] states that H is Zariski closed over $\mathbb{R}$. This might not be true, given Example 1.1.8 and the fact that we may have $H \subsetneq G_{\mathrm{SL}}$ for $\mathbb{K} = \mathbb{R}$. Therefore, we adjusted the argument.

Over the complex numbers, the converse of Theorem 9.4.1(d) also holds. However, over the reals there exist matrix tuples Y with a unique MLE but an infinite stabilizer, as the following example shows.

Example 9.4.3 ([AKRS21a, Example 4.2])**.** Set $m_1 = m_2 = n = 2$ and take $Y \in (\mathbb{R}^{2\times 2})^2$, where

$$Y_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

We prove that over the reals the MLE given Y is unique although the stabilizer of Y is infinite. In contrast, Y has infinitely many MLEs for the complex matrix normal model.

First, we show that Y is polystable under the left-right action of $\mathrm{SL}_2(\mathbb{K}) \times \mathrm{SL}_2(\mathbb{K})$, where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Note that any matrix in $\mathrm{SL}_2(\mathbb{K})$ has Frobenius norm at least $\sqrt{2}$. Indeed, if σ_1 and σ_2 are the singular values of $g \in \mathrm{SL}_2(\mathbb{K})$, then $\|g\|^2 = \sigma_1^2 + \sigma_2^2$, where $\sigma_1\sigma_2 = 1$. By the arithmetic mean - geometric mean inequality, we have $\|g\|^2 \geq 2$. Therefore, Y_1 and Y_2 have minimal Frobenius norm in $\mathrm{SL}_2(\mathbb{K})$ and thus Y is of minimal norm in its orbit. By Kempf-Ness, Theorem 2.2.13(d), the matrix tuple Y is polystable and hence an MLE given Y exists.

Next we compute the stabilizer of Y. It consists of matrices $(g_1, g_2) \in \mathrm{SL}_2(\mathbb{K}) \times \mathrm{SL}_2(\mathbb{K})$ with $g_1 Y_i g_2^{\mathsf{T}} = Y_i$. For Y_1, this gives $g_1 g_2^{\mathsf{T}} = \mathrm{I}_2$, i.e., $g_2^{\mathsf{T}} = g_1^{-1}$. From Y_2, we obtain $g_1 Y_2 = Y_2 g_1$ and writing

$$g_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{we get} \quad g_1 Y_2 = \begin{pmatrix} b & -a \\ d & -c \end{pmatrix} = \begin{pmatrix} -c & -d \\ a & b \end{pmatrix} = Y_2 g_1.$$

We deduce $a = d$, $b = -c$ and $\det(g_1) = 1 = a^2 + b^2$. This proves $g_1 \in \mathrm{SO}_2(\mathbb{K})$ and hence $g_2 = g_1^{-\mathsf{T}} = g_1$. Thus, the stabilizer of Y is contained in the infinite set $\{(g, g) \mid g \in \mathrm{SO}_2(\mathbb{K})\}$. In fact, we have equality as $\mathrm{SO}_2(\mathbb{K})$ is commutative and $Y_1, Y_2 \in \mathrm{SO}_2(\mathbb{K})$.

Since Y is of minimal norm in its orbit, $\lambda\, \mathrm{I}_2 \otimes \mathrm{I}_2$ is an MLE by Theorem 9.4.1, where $\lambda > 0$ minimizes the outer infimum in (9.14). For $\mathbb{K} = \mathbb{R}$, transpose and Hermitian transpose agree. Thus, any other MLE is given as $\lambda g^{\mathsf{T}}\, \mathrm{I}_2\, g \otimes g^{\mathsf{T}}\, \mathrm{I}_2\, g$ by some $g \in \mathrm{SO}_2(\mathbb{R})$, where we used the description of the stabilizer of Y. Since $g^{\mathsf{T}} g = \mathrm{I}_2$ we see that Y has unique MLE $\lambda\, \mathrm{I}_2 \otimes \mathrm{I}_2$. Note that the stabilizer $\{(g, g) \mid g \in \mathrm{SO}_2(\mathbb{R})\}$ of Y is indeed compact as predicted by Proposition 9.3.4.

For $\mathbb{K} = \mathbb{C}$, the MLEs involve $g^{\dagger} g$ rather than $g^{\mathsf{T}} g$, hence from the complex stabilizer $\{(g, g) \mid g \in \mathrm{SO}_2(\mathbb{C})\}$ we obtain infinitely many MLEs. ♢

The next example shows that all stability conditions in Theorem 9.4.1(a)–(d) can occur.

Example 9.4.4 ([AKRS21a, Example 4.3])**.** We set $m_1 = m_2 = 2$, and study stability under $\mathrm{SL}_2(\mathbb{K}) \times \mathrm{SL}_2(\mathbb{K})$ on $(\mathbb{K}^{2\times 2})^n$. We use the matrices

$$Y_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad Y_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y_4 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

(a) The matrix Y_4 is unstable and the matrix tuple (Y_4, Y_4) is unstable as well.

(b) The orbit of (Y_1, Y_4) is contained in $\{(g, M) \mid g \in \mathrm{SL}_2(\mathbb{K}),\ M \neq 0\}$. In particular, (Y_1, Y_4) is semistable as $\mathrm{SL}_2(\mathbb{K})$ is Euclidean closed. Moreover, for any $g \in \mathrm{SL}_2(\mathbb{K})$ and $M \in \mathbb{K}^{2\times 2} \setminus \{0\}$ we have

$$\|(g, M)\|^2 = \|g\|^2 + \|M\|^2 \geq 2 + \|M\|^2 > 2,$$

where we used $\|g\|^2 \geq 2$, see Example 9.4.3. On the other hand, we have

$$\left(\begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix}, \begin{pmatrix} \varepsilon^{-1} & 0 \\ 0 & \varepsilon \end{pmatrix}\right) \cdot (Y_1, Y_4) = \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & \varepsilon^2 \\ 0 & 0 \end{pmatrix}\right),$$

which tends to $(Y_1, 0)$ as $\varepsilon \to 0$. Since $\|(Y_1, 0)\|^2 = 2$ the capacity of (Y_1, Y_4) is not attained by an element in the orbit of (Y_1, Y_4), and Y is not polystable.

(c) The matrix $Y_1 = \mathrm{I}_2$ is polystable by Kempf-Ness, Theorem 2.2.13(d), as it is an $\mathrm{SL}_2(\mathbb{K})$ matrix of minimal norm. An MLE is given by $\lambda\, \mathrm{I}_2 \otimes \mathrm{I}_2$, where λ is the minimizer of the outer infimum in (9.14). Furthermore, Y_1 is not stable, because its stabilizer is $\{(g, g^{-\mathsf{T}}) \mid g \in \mathrm{SL}_2(\mathbb{K})\}$. There are infinitely many MLEs given Y of the form $\lambda g^{\mathsf{T}} g \otimes g^{-1} g^{-\mathsf{T}}$ for $g \in \mathrm{SL}_2(\mathbb{K})$.

(d) We show that $Y = (Y_1, Y_2, Y_3)$ is stable. First, any tuple (M_1, M_2, M_3) in the orbit of Y satisfies $M_1, M_2 \in \mathrm{SL}_2(\mathbb{K})$ and $\det(M_3) = -1$. Any 2×2 matrix of determinant ± 1 has Frobenius norm at least $\sqrt{2}$, by the same argument as in Example 9.4.3. Therefore, Y is of minimal norm in its orbit, and hence polystable by Theorem 2.2.13(d). It remains to show that the stabilizer of Y is finite. The discussion from Example 9.4.3 ensures that the stabilizer of Y is contained in $\{(g, g) \mid g \in \mathrm{SO}_2(\mathbb{K})\}$. Given $g \in \mathrm{SO}_2(\mathbb{K})$, the condition $gY_3g^{\mathsf{T}} = Y_3$ is equivalent to $gY_3 = Y_3g$. There exist $a, b \in \mathbb{K}$ with $a^2 + b^2 = 1$ such that

$$g = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \quad \text{and we compute} \quad gY_3 = \begin{pmatrix} b & a \\ a & -b \end{pmatrix} = \begin{pmatrix} -b & a \\ a & b \end{pmatrix} = Y_3 g.$$

Therefore, $b = -b$ which implies $b = 0$ and $a^2 = 1$. We see that $gY_3 = Y_3g$ for $g \in \mathrm{SO}_2(\mathbb{K})$ holds if and only if $g = \pm\, \mathrm{I}_2$. Therefore, the stabilizer of Y is the finite set $\{(\mathrm{I}_2, \mathrm{I}_2), (-\mathrm{I}_2, -\mathrm{I}_2)\}$. Altogether, Y is stable and there is a unique MLE given Y, namely $\lambda'\, \mathrm{I}_2 \otimes \mathrm{I}_2$ where $\lambda' > 0$ is the unique minimizer of the outer infimum in (9.14). ♢

9.4.2 Boundedness of the likelihood via semistability

This subsection[16] illustrates how the dictionary between ML estimation and stability notions can be used to gain new insights and to recover known results on the statistical side. More specifically, we use the equivalence of a bounded

[16] Like the whole Section 9.4 also this subsection closely follows the presentation in [AKRS21a, Section 4]. However, while [AKRS21a, Subsection 4.2] states all results only for $\mathbb{K} = \mathbb{R}$ they are also valid over $\mathbb{C}$. Therefore, the statements and proofs are accordingly adjusted to $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

likelihood with the semistability of a matrix tuple under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$, Theorem 9.4.1(b), to obtain bounds on the ML threshold $\mathrm{mlt}_b(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2))$. The upper bound from Corollary 9.4.11 was new as it appeared, while Corollaries 9.4.7, 9.4.12 and 9.4.13 recover known results from the literature. All these bounds are consequences of Theorems 9.4.6 and 9.4.10, which are proved by using results from [BD06] on the complex null cone under the left-right action.

It is important to point out that the presented results are outdated: Derksen and Makam combined Theorem 9.4.1 with representation theory of quivers to determine all ML thresholds for $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$ [DM21]. We state their main result in Theorem 9.6.1. This was further generalized to determining all ML thresholds of tensor normal models in [DMW22]. Still, this subsection may serve the reader as a first introduction before entering the general concepts in [DM21; DMW22].

Remark 9.4.5. Note that $Y = (Y_1, \ldots, Y_n) \in (\mathbb{K}^{m_1 \times m_2})^n$ is unstable under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ if and only if $Y^\dagger := (Y_1^\dagger, \ldots, Y_n^\dagger) \in (\mathbb{K}^{m_2 \times m_1})^n$ is unstable under the left-right action of $\mathrm{SL}_{m_2}(\mathbb{K}) \times \mathrm{SL}_{m_1}(\mathbb{K})$. Equivalently, ℓ_Y is not bounded from above if and only if $\ell_{Y^\dagger}$ is not bounded from above. Therefore, $\mathrm{mlt}_b(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)) = \mathrm{mlt}_b(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_2, m_1))$ and we may assume that $m_1 \geq m_2$. $\triangledown$

The following theorem gives a characterization of the matrix tuples with unbounded log-likelihood. It has been derived for $\mathbb{K} = \mathbb{R}$ in [DKH21, Theorems 3.1(i) and 3.3(i)] using a different method.

Theorem 9.4.6 ([AKRS21a, Theorem 4.4])**.** *Consider the matrix normal model $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$ with tuple of samples $Y \in (\mathbb{K}^{m_1 \times m_2})^n$. Then ℓ_Y is not bounded from above if and only if there exist subspaces $V_1 \subseteq \mathbb{K}^{m_1}$ and $V_2 \subseteq \mathbb{K}^{m_2}$ with $m_1 \dim_{\mathbb{K}} V_2 > m_2 \dim_{\mathbb{K}} V_1$ such that $Y_i V_2 \subseteq V_1$ for all $i = 1, \ldots, n$.*

Proof. The log-likelihood ℓ_Y is bounded from above if and only if Y is *not* in the null cone $\mathcal{N}_{\mathbb{K}}$ under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$, by Theorem 9.4.1(b). Thus, for $\mathbb{K} = \mathbb{C}$ the statement follows from [BD06, Theorem 2.1] respectively Proposition 2.3.6.

It remains to prove the case $\mathbb{K} = \mathbb{R}$, so let $Y \in (\mathbb{R}^{m_1 \times m_2})^n$. Then $Y \notin \mathcal{N}_{\mathbb{R}}$ if and only if it is not in the complex null cone $\mathcal{N}_{\mathbb{C}}$, by Proposition 2.2.18. The latter is equivalent to the existence of subspaces $W_1 \subseteq \mathbb{C}^{m_1}$ and $W_2 \subseteq \mathbb{C}^{m_2}$ with $m_1 \dim_{\mathbb{C}} W_2 > m_2 \dim_{\mathbb{C}} W_1$ such that $Y_i W_2 \subseteq W_1$ for all $i = 1, \ldots, n$, by [BD06, Theorem 2.1] (respectively Proposition 2.3.6). This is the same condition as in the statement, except with *complex* subspaces. The real condition implies the complex one: if $V_j \subseteq \mathbb{R}^{m_j}$ are real subspaces as in the statement, then $W_j := V_j \oplus \mathrm{i} V_j \subseteq \mathbb{C}^{m_j}$ satisfy the complex conditions. We show the reverse implication following an argument thanks to Jan Draisma.

Given complex subspaces $W_1 \subseteq \mathbb{C}^{m_1}$ and $W_2 \subseteq \mathbb{C}^{m_2}$ as above. Set $V_j := W_j \cap \mathbb{R}^{m_j}$ for $j = 1, 2$. Then also $\mathrm{i} V_j \subseteq W_j$, where i is the imaginary unit. Furthermore, let V_j' be the image of the $\mathbb{R}$-linear map $f_j \colon W_j \to \mathbb{R}^{m_j}$ that sends a complex vector to its real part. Of course, we have $\mathrm{i} V_j \subseteq \ker(f_j)$. Conversely, any $w \in \ker(f_j)$ is of the form $\mathrm{i} v$, where $v \in \mathbb{R}^{m_j}$ but also $-\mathrm{i} w = v \in W_j$. Therefore, $v \in V_j$ and this shows $\ker(f_j) = \mathrm{i} V_j$. The latter implies

$$2 \dim_{\mathbb{C}} W_j = \dim_{\mathbb{R}} W_j = \dim_{\mathbb{R}} V_j + \dim_{\mathbb{R}} V_j'.$$

In particular, we have $m_1 \dim_{\mathbb{R}} V_2 > m_2 \dim_{\mathbb{R}} V_1$ or $m_1 \dim_{\mathbb{R}} V_2' > m_2 \dim_{\mathbb{R}} V_1'$. Since $Y \in (\mathbb{R}^{m_1 \times m_2})^n$ and $Y_i W_2 \subseteq W_1$, both inclusions $Y_i V_2 \subseteq V_1$ and $Y_i V_2' \subseteq V_1'$ hold for all $i = 1, \ldots, n$. Hence, (V_1, V_2) or (V_1', V_2') are real subspaces as in the statement. □

As a consequence we obtain a lower bound on $\mathrm{mlt}_b(\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2))$, which also follows from [DKH21, Lemma 1.2].

Corollary 9.4.7 ([AKRS21a, Corollary 4.5])**.** *If $n < \frac{m_1}{m_2}$, then the log-likelihood function ℓ_Y is unbounded from above for every tuple of samples $Y \in (\mathbb{K}^{m_1 \times m_2})^n$. In particular,*

$$\mathrm{mlt}_b\left(\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)\right) \geq \left\lceil \frac{m_1}{m_2} \right\rceil .$$

Proof. For any one-dimensional subspace $V_2 \subseteq \mathbb{K}^{m_2}$, the dimension of $V_1 := \sum_{i=1}^n Y_i V_2$ is at most n. If $n < \frac{m_1}{m_2}$, Theorem 9.4.6 implies that the log-likelihood ℓ_Y is unbounded. □

To prove further statistical consequences we introduce the cut-and-paste rank from [BD06, Definition 2.2].[17]

Definition 9.4.8. Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. The *cut-and-paste rank* $\mathrm{cp}_{\mathbb{K}}^{(n)}(a, b, c, d)$ over $\mathbb{K}$ of a tuple of positive integers a, b, c, d and n is the maximum rank all $ab \times cd$ matrices of the form $\sum_{i=1}^n X_i \otimes Y_i$, where $X_i \in \mathbb{K}^{c \times a}$ and $Y_i \in \mathbb{K}^{d \times b}$.
By the upcoming remark the cut-and-paste rank does not depend on $\mathbb{K}$ and we therefore drop the index $\mathbb{K}$. ▲

Remark 9.4.9 ([AKRS21a, Remark 4.7]). We have $\mathrm{cp}_{\mathbb{R}}^{(n)}(a, b, c, d) \leq \mathrm{cp}_{\mathbb{C}}^{(n)}(a, b, c, d)$ and equality holds as follows. The condition for the rank of the complex matrix $\sum_{i=1}^n X_i \otimes Y_i$ to drop is given by minors. Thus, $\mathrm{cp}_{\mathbb{C}}^{(n)}(a, b, c, d)$ is witnessed on a Zariski-open subset of $W := (\mathbb{C}^{c \times a})^n \times (\mathbb{C}^{d \times b})^n$ and hence witnessed by some element in $(\mathbb{R}^{c \times a})^n \times (\mathbb{R}^{d \times b})^n$, as the latter is Zariski-dense in W. ▽

We use the cut-and-paste rank to state in Theorem 9.4.10 a necessary and sufficient condition for ℓ_Y to be unbounded from above for every tuple of samples $Y \in (\mathbb{K}^{m_1 \times m_2})^n$; or equivalently, for $\mathcal{N}_{\mathbb{K}} = (\mathbb{K}^{m_1 \times m_2})^n$, where $\mathcal{N}_{\mathbb{K}}$ is the null cone under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$. Note that if $\mathcal{N}_{\mathbb{K}}$ does not fill the irreducible variety $(\mathbb{K}^{m_1 \times m_2})^n$, then $\mathcal{N}_{\mathbb{K}}$ must have positive codimension and hence has Lebesgue measure zero. Consequently, ℓ_Y is *either* not bounded from above for every $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ *or* it is bounded for almost all Y. Therefore, Theorem 9.4.10 solves in principle the problem of determining $\mathrm{mlt}_b\left(\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)\right)$, although in terms of the cut-and-paste rank.[18]

Recall that we can assume that $m_1 \geq m_2$, by Remark 9.4.5. Moreover, since Corollary 9.4.7 shows that the likelihood is unbounded for $m_2 n < m_1$, it suffices to restrict to the range $m_2 \leq m_1 \leq n m_2$.

[17] In [BD06] the cut-and-paste rank is defined over $\mathbb{C}$. For statistical models over the reals it is more natural to define it over $\mathbb{R}$. Actually, both concepts agree, see Remark 9.4.9.

[18] At the time the first preprint of [AKRS21a] appeared this gave a statistical motivation to study the cut-and-paste rank. However, Theorem 9.4.1 quickly led to a full determination of all ML thresholds of $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$ by Derksen and Makam [DM21, Theorem 1.3]; see Theorem 9.6.1. Thus, their result may now in turn be used to understand the cut-and-paste rank.

Theorem 9.4.10 ([AKRS21a, Theorem 4.8]). *Let $0 < m_2 \leq m_1 \leq nm_2$ and consider the matrix normal model $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$. The log-likelihood ℓ_Y is unbounded from above for* every *tuple of samples $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ if and only if there exists $k \in \{1, \ldots, m_2\}$ such that $l = l(k) = \lceil \frac{m_1}{m_2} k \rceil - 1$ satisfies both*

$$\begin{aligned} &m_1 - l \leq n(m_2 - k) \quad \text{and} \\ &\mathrm{cp}^{(n)}(a, b, c, d) = cd, \quad \text{where} \quad (a, b, c, d) = (m_2 - k, k, m_1 - l, nk - l). \end{aligned}$$

Proof. Let $\mathcal{N}_{\mathbb{K}}$ be the null cone under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ on $(\mathbb{K}^{m_1 \times m_2})^n$, where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. By Theorem 9.4.1(a), ℓ_Y is unbounded from above for every tuple of samples $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ if and only if $\mathcal{N}_{\mathbb{K}} = (\mathbb{K}^{m_1 \times m_2})^n$. Moreover, $\mathcal{N}_{\mathbb{C}} = (\mathbb{C}^{m_1 \times m_2})^n$ if and only if $\mathcal{N}_{\mathbb{R}} = (\mathbb{R}^{m_1 \times m_2})^n$. It therefore suffices to characterize when $\mathcal{N}_{\mathbb{C}} = (\mathbb{C}^{m_1 \times m_2})^n$.

For this, define for natural numbers k and l

$$Q_{k,l} := \left\{ (Y_1, \ldots, Y_n) \in (\mathbb{C}^{m_1 \times m_2})^n \mid \exists V \subseteq \mathbb{C}^{m_2} : \dim_{\mathbb{C}} V = k, \dim_{\mathbb{C}} \left(\sum_{i=1}^{n} Y_i V \right) \leq l \right\}.$$

The null cone $\mathcal{N}_{\mathbb{C}}$ is the union of the $Q_{k,l}$ over $1 \leq k \leq m_2$ and $0 \leq l < \frac{m_1}{m_2} k$, by Theorem 9.4.6. We observe that for fixed k the algebraic sets $Q_{k,l}$ become larger as l increases. Hence, it suffices to consider if any of the $Q_{k,l}$ fills $(\mathbb{C}^{m_1 \times m_2})^n$ as k ranges over $1 \leq k \leq m_2$, where the corresponding l is the largest integer strictly smaller than $\frac{m_1}{m_2} k$, i.e., $l = l(k) := \lceil \frac{m_1}{m_2} k \rceil - 1$.

The assumption $m_1 \leq nm_2$ yields $l < nk$. Therefore, [BD06, Proposition 2.4] shows that

$$\dim_{\mathbb{C}} Q_{k,l} = nm_1 m_2 - \left((m_1 - l)(kn - l) - \mathrm{cp}^{(n)}(a, b, \tilde{c}, d) \right),$$

where $a = m_2 - k$, $b = k$, $\tilde{c} = \min\{m_1 - l, n(m_2 - k)\}$ and $d = kn - l$. Thus, $Q_{k,l}$ equals $(\mathbb{C}^{m_1 \times m_2})^n$ if and only if

$$\mathrm{cp}^{(n)}(a, b, \tilde{c}, d) = (m_1 - l)(kn - l).$$

Finally, the latter equation is equivalent to

$$m_1 - l \leq n(m_2 - k) \quad \text{and} \quad \mathrm{cp}^{(n)}(a, b, \tilde{c}, d) = \tilde{c}d,$$

since $\tilde{c} = \min\{m_1 - l, n(m_2 - k)\}$, $d = kn - l \geq 1$ and $\mathrm{cp}^{(n)}(a, b, \tilde{c}, d) \leq \tilde{c}d$. □

We use the above theorem to give an upper bound for $\mathrm{mlt}_b\left(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)\right)$, which was new at its time.

Corollary 9.4.11 ([AKRS21a, Corollary 4.9]). *Let $0 < m_2 \leq m_1$. If*

$$n > \max_{1 \leq k \leq m_2} \left(\frac{l(k)}{k} + \frac{m_2 - k}{m_1 - l(k)} \right), \quad \text{where } l(k) = \left\lceil \frac{m_1}{m_2} k \right\rceil - 1, \tag{9.15}$$

then ℓ_Y is bounded from above for almost all $Y \in (\mathbb{K}^{m_1 \times m_2})^n$. In other words,

$$\mathrm{mlt}_b\left(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)\right) \leq \left\lfloor \max_{1 \leq k \leq m_2} \left(\frac{l(k)}{k} + \frac{m_2 - k}{m_1 - l(k)} \right) \right\rfloor + 1.$$

Proof. First, we observe that (9.15) with $k = m_2$ yields $n > \frac{m_1-1}{m_2}$. The latter is equivalent to $nm_2 \geq m_1$, so we are in the setting of Theorem 9.4.10. Using the notation in that theorem, we see that (9.15) is equivalent to every $k \in \{1, \ldots, m_2\}$ satisfying $cd > ab$. In particular, for every such k we have $\mathrm{cp}^{(n)}(a, b, c, d) \leq ab < cd$. By Theorem 9.4.10, the log-likelihood ℓ_Y cannot be unbounded from above for every tuple Y and hence ℓ_Y is bounded from above for almost all Y. □

We obtain two further upper bounds which are known in the statistics literature, compare [DKH21, Proposition 1.3, Theorem 1.4].

Corollary 9.4.12 ([AKRS21a, Corollary 4.10])**.** *It holds that*

$$\mathrm{mlt}_b\left(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)\right) \leq \left\lceil \frac{m_1}{m_2} + \frac{m_2}{m_1} \right\rceil.$$

Proof. For every $k \in \{1, \ldots, m_2\}$ we have $l(k) < \frac{m_1 k}{m_2}$, which implies that

$$\frac{m_1}{m_2} + \frac{m_2}{m_1} > \frac{l(k)}{k} + \frac{m_2 - k}{m_1 - l(k)}.$$

Thus, the assertion follows from Corollary 9.4.11. □

Corollary 9.4.13 ([AKRS21a, Corollary 4.11])**.** *If m_2 divides m_1, then*

$$\mathrm{mlt}_b\left(\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)\right) = \frac{m_1}{m_2}.$$

Proof. If $n < \frac{m_1}{m_2}$, the log-likelihood is always unbounded from above by Corollary 9.4.7. So we write $m_1 = \gamma m_2$, where $\gamma \in \mathbb{Z}_{\geq 1}$, and assume $n \geq \gamma$. For every $k \in \{1, \ldots, m_2\}$, using the notation from Theorem 9.4.10, we see that $l = l(k) = \gamma k - 1$ and $a < c$. If $n > \gamma$, we also have that $b < d$, so $\mathrm{cp}^{(n)}(a, b, c, d) \leq ab < cd$. If $n = \gamma$, then $m_1 - l(k) > n(m_2 - k)$. In either case, one of the two conditions in Theorem 9.4.10 is not satisfied, so ℓ_Y is bounded from above for almost all Y. □

In Table 9.1 we list the maximum likelihood threshold mlt_b for boundedness of the log-likelihood for small values of m_1, m_2, and compare with the bounds discussed above.[19] We observe that there are cases where our upper bound

$$\alpha = \left\lfloor \max_{1 \leq k \leq m_2} \left(\frac{l(k)}{k} + \frac{m_2 - k}{m_1 - l(k)} \right) \right\rfloor + 1, \qquad \text{where } l(k) = \left\lceil \frac{m_1}{m_2} k \right\rceil - 1,$$

is strictly better than the simple upper bound $U = \lceil \frac{m_1}{m_2} + \frac{m_2}{m_1} \rceil$, e.g., when $(m_1, m_2) = (3, 2)$. In most cases our bound α matches the lower bound $L = \lceil \frac{m_1}{m_2} \rceil$, so that we can determine mlt_b. In addition, when $m_2 | m_1$, one can use Corollary 9.4.13 to determine mlt_b even if the bounds L and α do not coincide, such as in $(m_1, m_2) = (8, 4)$ or in the square cases $m_1 = m_2$. The rest of the values of mlt_b can be filled from [DKH21, Table 1]. We highlight the case $(m_1, m_2) = (8, 3)$: the maximum likelihood threshold $\mathrm{mlt}_b = 3$ was computed in [DKH21] via Gröbner bases, but it is not covered by the general bounds in [DKH21]. Nevertheless, our bound α determines this case.

[19] This comparison represents the status when the first preprint of [AKRS21a] appeared in March 2020. Now, all values of mlt_b in Table 9.1 are determined by [DM21, Theorem 1.3], which we state in Theorem 9.6.1.

m_1	m_2	L	mlt_b	α	U
2	2	1	1	1	2
3	2	2	2	2	3
3	3	1	1	2	2
4	2	2	2	2	3
4	3	2	2	2	3
4	4	1	1	2	2
5	2	3	3	3	3
5	3	2	3	3	3
5	4	2	2	2	3
5	5	1	1	2	2
6	2	3	3	3	4
6	3	2	2	2	3
6	4	2	2	2	3
6	5	2	2	2	3
6	6	1	1	2	2

m_1	m_2	L	mlt_b	α	U
7	2	4	4	4	4
7	3	3	3	3	3
7	4	2	3	3	3
7	5	2	3	3	3
7	6	2	2	2	3
7	7	1	1	2	2
8	2	4	4	4	5
8	3	3	3	3	4
8	4	2	2	3	3
8	5	2	3	3	3
8	6	2	2	2	3
8	7	2	2	2	3
8	8	1	1	2	2
9	2	5	5	5	5
9	3	3	3	3	4

m_1	m_2	L	mlt_b	α	U
9	4	3	3	3	3
9	5	2	3	3	3
9	6	2	2	2	3
9	7	2	3	3	3
9	8	2	2	2	3
9	9	1	1	2	2
10	2	5	5	5	6
10	3	4	4	4	4
10	4	3	3	3	3
10	5	2	2	3	3
10	6	2	3	3	3
10	7	2	3	3	3
10	8	2	2	2	3
10	9	2	2	2	3
10	10	1	1	2	2

Table 9.1: [AKRS21a, Table 1] Bounds for the maximum likelihood threshold mlt_b. $L = \lceil \frac{m_1}{m_2} \rceil$ is the lower-bound from Corollary 9.4.7, $U = \lceil \frac{m_1}{m_2} + \frac{m_2}{m_1} \rceil$ is the upper bound from Corollary 9.4.12, and α is the upper bound from Corollary 9.4.11.

9.4.3 Uniqueness of the MLE via stability

In this short subsection we compare conditions for a stable $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ under left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$ with conditions for existence of a unique MLE given Y in the matrix normal model $\mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$.

Example 9.4.3 shows that for $\mathbb{K} = \mathbb{R}$ existence of a unique MLE given Y in $\mathcal{M}^{\otimes}_{\mathbb{R}}(m_1, m_2)$ is not equivalent to Y being stable. However, such an equivalence holds in the complex setting $\mathbb{K} = \mathbb{C}$, by Theorem 9.4.1. Hence, for the complex model $\mathcal{M}^{\otimes}_{\mathbb{C}}(m_1, m_2)$ we obtain conditions for unique existence of an MLE given $Y \in (\mathbb{C}^{m_1 \times m_2})^n$ from characterizing when Y is stable under the left-right action. Characterizing this stability is a special case of the setting studied in [Kin94], compare Section 2.3. Combining Theorem 2.3.1 and Theorem 9.4.1(d) directly gives the following.

Theorem 9.4.14 ([AKRS21a, Theorem 4.12])**.** *Consider the left-right action of* $\mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ *on* $(\mathbb{C}^{m_1 \times m_2})^n$*, and a tuple* $Y \in (\mathbb{C}^{m_1 \times m_2})^n$ *of* n *samples for the complex matrix normal model* $\mathcal{M}^{\otimes}_{\mathbb{C}}(m_1, m_2)$*. The following are equivalent:*

(a) there exists a unique MLE given Y*;*

(b) the matrix tuple Y *is stable;*

(c) the matrix $(Y_1 | \dots | Y_n) \in \mathbb{C}^{m_1 \times n m_2}$ *has rank* m_1*, and* $m_2 \dim V_1 > m_1 \dim V_2$ *holds for all subspaces* $V_1 \subseteq \mathbb{C}^{m_1}$*,* $\{0\} \subsetneq V_2 \subsetneq \mathbb{C}^{m_2}$ *that satisfy* $Y_i V_2 \subseteq V_1$ *for all* $i = 1, \dots, n$*.*

We note the similarity with the conditions that characterize semistability in Theorem 9.4.6; see also Proposition 2.3.6. However, while Theorem 9.4.6 holds both over $\mathbb{R}$ and $\mathbb{C}$, the same cannot be true for Theorem 9.4.14 by Example 9.4.3. In fact, the real analogue of Theorem 9.4.14(c) characterizes existence of a unique MLE for the real matrix normal model $\mathcal{M}^{\otimes}_{\mathbb{R}}(m_1, m_2)$, see [DKH21, Theorems 3.1(ii) and 3.3(ii)].

9.4.4 Operator Scaling and Flip-Flop Algorithm

In this subsection, we illustrate the algorithmic consequences of the connection between invariant theory and maximum likelihood estimation. We present the flip-flop algorithm for ML estimation that is well-known in statistics, and connect it to Algorithm 3.2 for operator scaling in invariant theory. The connection allows us to give a complexity analysis of the flip-flop algorithm in Theorem 9.4.16.

It is noteworthy to mention that the similarities between operator scaling and the flip-flop algorithm started and stimulated the work on [AKRS21a]. The study first led to Theorem 9.4.1, which was then generalized to the setting of Gaussian group models.

Comparing Operator Scaling and the Flip-Flop Algorithm

Operator scaling, Algorithm 3.2, solves the Scaling Problem 3.1.4 for the left-right action of $\mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$ on $(\mathbb{C}^{m_1 \times m_2})^n$, compare Section 3.2. The method was generalized to tuples of tensors in [BGO+18, Algorithm 1].

The *flip-flop algorithm* [Dut99; LZ05; WJS08], see the bottom left of Figure 9.1, is an alternating maximization procedure to find an MLE in a matrix normal model $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$. It can be thought of as a Gaussian version of IPS for matrix normal models, since one alternately updates the estimates in each marginal. If we consider $\Psi_2 \in \mathrm{PD}_{m_2}(\mathbb{K})$ to be fixed, the log-likelihood in Equation (9.12) becomes, up to constants,

$$m_2 \left[\log\det(\Psi_1) - \operatorname{tr}\left(\Psi_1 \cdot \frac{1}{nm_2} \sum_{i=1}^{n} Y_i \Psi_2^{\mathsf{T}} Y_i^{\dagger} \right) \right].$$

Maximizing the latter with respect to Ψ_1 reduces to the case of a standard multivariate Gaussian model as in (6.8). The unique maximizer over $\mathrm{PD}_{m_1}(\mathbb{K})$, if it exists, is the inverse of the matrix $\frac{1}{nm_2}\sum_{i=1}^{n} Y_i \Psi_2^{\mathsf{T}} Y_i^{\dagger}$, compare Example 6.3.8. In the same way, we can fix Ψ_1 and use $\det(\Psi_2) = \det(\Psi_2^{\mathsf{T}})$ to maximize the log-likelihood with respect to Ψ_2^{T}. Iterating these two steps gives Algorithm 9.1.

We now compare operator scaling, Algorithm 3.2, with the flip-flop algorithm. First, note that operator scaling restricts to matrices of determinant one, in order to stay in the $\mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$-orbit of Y. In comparison, Algorithm 9.1 has constants chosen to minimize the outer infimum in Equation (9.14). In the following we argue that, via the correspondence[20] $g_j^{\dagger} g_j \leftrightarrow \Psi_j$, operator scaling is the same procedure as the flip-flop algorithm, up to scalar factors.[21]

We exemplify this for updating g_2 respectively Ψ_2. Given g and Y, and ignoring the determinant one rescaling, we set $g_2^{\mathrm{new}} := \varrho^{-1/2} g_2$, where

$$\varrho_2 := \left(\sum_{i=1}^{n} \left(g_1 Y_i g_2^{\mathsf{T}}\right)^{\dagger} \left(g_1 Y_i g_2^{\mathsf{T}}\right) \right)^{\mathsf{T}} = g_2 \left(\sum_{i=1}^{n} Y_i^{\dagger} g_1^{\dagger} g_1 Y_i \right)^{\mathsf{T}} g_2^{\dagger}$$

[20] Remember that $g \in G = \mathrm{GL}_{m_1}(\mathbb{K}) \times \mathrm{GL}_{m_2}(\mathbb{K})$ gives $(g_1^{\dagger} g_1) \otimes (g_2^{\dagger} g_2) \in \mathcal{M}_G^{\mathfrak{g}} = \mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$, see Example 9.1.6.

[21] This is similar to classical matrix scaling and its invariant theoretic appearance, compare the extended example in Section 3.1.

Algorithm 9.1: Flip-flop [AKRS21a, Algorithm 4.1]

Input : $Y_1, \dots, Y_n \in \mathbb{K}^{m_1 \times m_2}$, a number of iterations $N \in \mathbb{Z}_{>0}$.
Output: an approximation of an MLE, if it exists.

Initialize $\Psi_2 := \mathrm{I}_{m_2}$;
for $k = 1$ **to** N **do**
the following pair of updates

$$\begin{aligned} \Psi_1 &\leftarrow \left(\frac{1}{nm_2} \sum_{i=1}^{n} Y_i \Psi_2^{\mathsf{T}} Y_i^{\dagger} \right)^{-1} \\ \Psi_2 &\leftarrow \left(\frac{1}{nm_1} \sum_{i=1}^{n} Y_i^{\dagger} \Psi_1 Y_i \right)^{-\mathsf{T}} . \end{aligned} \tag{9.16}$$

end
return $\Psi_1 \otimes \Psi_2$.

is defined as in Algorithm 3.2. We compute

$$(g_2^{\mathrm{new}})^{\dagger} g_2^{\mathrm{new}} = g_2^{\dagger} \varrho^{-1} g_2 = \left(\sum_{i=1}^{n} Y_i^{\dagger} g_1^{\dagger} g_1 Y_i \right)^{-\mathsf{T}} \longleftrightarrow \left(\sum_{i=1}^{n} Y_i^{\dagger} \Psi_1 Y_i \right)^{-\mathsf{T}} ,$$

which indeed corresponds, up to scalar factors, to Ψ_2^{new}, compare Equation (9.16). A similar computation holds for g_1 and Ψ_1.

Conversely, the square-root $\Psi_j^{1/2} =: \hat{g}_j$ satisfies $\hat{g}_j^{\dagger} \hat{g}_j = \Psi_j = g_j^{\dagger} g_j$ and hence $\hat{g}_j$ only differs by a unitary matrix from g_j. Therefore, $\|\hat{g} \cdot Y\| = \|g \cdot Y\|$ and $\|\mu_G(\hat{g} \cdot Y)\| = \|\mu_G(g \cdot Y)\|$. Altogether, Algorithms 3.2 and 9.1 are, up to rescaling, essentially the same.

Although operator scaling is defined over $\mathbb{C}$, when restricting to real inputs it only involves computations over the reals, compare Algorithm 3.2. This allows the computation of MLEs (if they exist) in $\mathcal{M}_{\mathbb{R}}^{\otimes}(m_1, m_2)$ via (9.14), since the capacity of a real matrix tuple is the same under the action of $\mathrm{SL}_{m_1}(\mathbb{R}) \times \mathrm{SL}_{m_2}(\mathbb{R})$ as under the action of $\mathrm{SL}_{m_1}(\mathbb{C}) \times \mathrm{SL}_{m_2}(\mathbb{C})$, see Proposition 2.2.18.

Convergence

Due to the above comparison of the flip-flop algorithm with operator scaling, we can analyse the convergence behaviour of the former. If an update step in Algorithm 9.1 cannot be computed because one of the matrices in (9.16) cannot be inverted, then the matrix tuple $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ is unstable under the action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$. This implies that the log-likelihood ℓ_Y is unbounded, by Theorem 9.4.1(a). Otherwise, the sequence of terms

$$\left\| \left(\det(\Psi_1)^{-1/(2m_1)} \Psi_1^{1/2},\ \det(\Psi_2)^{-1/(2m_2)} \Psi_2^{1/2} \right) \cdot Y \right\|^2$$

converges. If the limit is zero, then the log-likelihood ℓ_Y is unbounded.

Otherwise, the limit is a positive number and Y is semistable. Here, two possibilities can arise. First, if Y is polystable then the minimal norm is attained at an element of the group $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$, and the flip-flop algorithm converges to an MLE, using the fact that the constants in the flip-flop algorithm minimize the outer infimum in (9.14). Second, if Y is semistable but not polystable then the flip-flop algorithm diverges by the following remark.

Remark 9.4.15 ([AKRS21a, Remark 4.14]). If $Y \in (\mathbb{K}^{m_1 \times m_2})^n$ is semistable but not polystable under the left-right action of $\mathrm{SL}_{m_1}(\mathbb{K}) \times \mathrm{SL}_{m_2}(\mathbb{K})$, then the likelihood L_Y (equivalently the log-likelihood ℓ_Y) is bounded from above, but does not attain its supremum. In this case, any sequence $\Psi_N := (\Psi_{1,N} \otimes \Psi_{2,N})$ of concentration matrices with

$$\lim_{N \to \infty} L_Y(\Psi_{1,N} \otimes \Psi_{2,N}) = \sup L_Y > 0$$

diverges by the following. Assume a limit Ψ_∞ exists. If $\Psi_\infty \in \mathrm{PD}_{m_1 m_2}(\mathbb{K})$ then $\Psi_\infty \in \mathcal{M}^{\otimes}_{\mathbb{K}}(m_1, m_2)$ as the latter is Euclidean closed in $\mathrm{PD}_{m_1 m_2}(\mathbb{K})$, by Theorem 9.3.1. This contradicts the supremum of L_Y not being attained. On the other hand, if $\Psi_\infty \notin \mathrm{PD}_{m_1 m_2}(\mathbb{K})$ then it is rank-deficient positive semidefinite, so $\det(\Psi_\infty) = 0$ and (6.7) yield the contradiction $\sup L_Y = L_Y(\Psi_\infty) = 0$. ▽

Complexity

As a direct consequence of the above comparison and convergence analysis, the complexity analysis of operator scaling carries over to the flip-flop algorithm. We adapt [BGO+18, Theorem 1.1] to our notation to derive the following.

Theorem 9.4.16 ([AKRS21a, Theorem 4.15]). *Let $\varepsilon > 0$ and let $Y \in (\mathbb{Z}^{m_1 \times m_2})^n$ with matrix entries of bit size bounded by b. After $\mathrm{poly}(nm_1m_2, b, 1/\varepsilon)$ many steps, the flip-flop algorithm either identifies that the log-likelihood ℓ_Y is unbounded or finds $(\Psi_1, \Psi_2) \in \mathrm{PD}_{m_1}(\mathbb{K}) \times \mathrm{PD}_{m_2}(\mathbb{K})$ such that the matrix tuple $Y' := \left(\det(\Psi_1)^{-1/(2m_1)}\Psi_1^{1/2}, \det(\Psi_2)^{-1/(2m_1)}\Psi_2^{1/2}\right) \cdot Y$ satisfies $\|\mu_G(Y')\| \leq \varepsilon$, where μ_G is as in Equation (2.19).*

If ℓ_Y is bounded from above, taking the limit $\varepsilon \to 0$ in Theorem 9.4.16 gives rise to two possibilities. Either the MLE exists and is the limit of the $\Psi_1 \otimes \Psi_2$ as $\varepsilon \to 0$, or the sequence $\Psi_1 \otimes \Psi_2$ diverges as $\varepsilon \to 0$, by Remark 9.4.15. Thus, in the latter scenario there is no meaningful notion of an approximate MLE.

Outlook

Remember that [BGO+18, Algorithm 1] generalizes operator scaling to scale tensors of format $m_1 \times \cdots \times m_d$ under the action of $\mathrm{SL}_{m_1}(\mathbb{C}) \times \cdots \times \mathrm{SL}_{m_d}(\mathbb{C})$. Thus, it can be used for ML estimation in (real and complex) tensor normal models. Similarly to the above, [BGO+18, Algorithm 1] corresponds to the flip-flop algorithm for tensor normal models, see e.g., [FORW21, Algorithm 2]. The latter algorithm satisfies the following. If Y is a tuple of i.i.d. d-tensor samples from the distribution given by $\Psi \in \mathcal{M}^{\otimes}_{\mathbb{R}}(m_1, \ldots, m_d)$, then the flip-flop algorithm converges linearly with high probability to Ψ, [FORW21, Theorems 2.9 and 2.10].

Finally, we stress again that the geodesic convex methods in [BFG+19] can be used for ML estimation in Gaussian group models $\mathcal{M}_G^{\mathbf{g}}$ where G is Zariski closed and self-adjoint, compare Subsection 9.3.1 and the right hand side of Figure 9.1.

9.5 TDAG models as Gaussian group models

In this section we revisit Gaussian graphical models given by a directed acyclic graph (DAG) $\mathcal{G}$, see Definition 6.3.11. The section is mainly based on [AKRS21a, Section 5], but also presents further results and knowledge from [MRS21]. We focus on the following subclass of DAGs.

Definition 9.5.1. A DAG $\mathcal{G}$ is called *transitive* if whenever $k \to j$ and $j \to i$ in $\mathcal{G}$ then also $k \to i$ in $\mathcal{G}$. We usually abbreviate transitive DAG to TDAG. ▲

First, we connect DAG models to the setting of Gaussian models via symmetrization by defining a natural set $\mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$ such that $\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathbf{g}}$. Afterwards, we characterize when $\mathcal{A}(\mathcal{G})$ is a subgroup of $\mathrm{GL}_m(\mathbb{K})$. It turns out that this is the case if and only if $\mathcal{G}$ is transitive. Therefore, TDAG models are naturally Gaussian group models. However, the group $\mathcal{A}(\mathcal{G})$ is usually not self-adjoint. Still, we can deduce the full correspondence for TDAG models, Theorem 9.5.9, and the G_{SL}-stabilizer of a sample matrix Y is proven to be in bijection with the MLEs given Y, compare Proposition 9.5.10. Finally, we briefly study which undirected Gaussian graphical models from Example 6.3.10 arise as Gaussian group models.

In the following the vertex set I of $\mathcal{G}$ is always $[m] = \{1, 2, \ldots, m\}$. Recall from Definition 6.3.11 that a DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is given by a linear structural equation (6.13). Thus, $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is the set of all concentration matrices of the form

$$(\mathrm{I}_m - \Lambda)^{\dagger} \Omega^{-1} (\mathrm{I}_m - \Lambda),$$

where $\Omega \in \mathrm{PD}_m(\mathbb{K})$ is diagonal and $\lambda_{ij} = 0$ whenever $j \not\to i$ in $\mathcal{G}$, compare Equation (6.14). By acyclicity, we can and will assume that $j > i$ whenever $j \to i$ in $\mathcal{G}$, so Λ is strictly upper triangular, see Remark 6.3.12.

Now, we put DAG models into the context of Gaussian models via symmetrization. Given a DAG $\mathcal{G}$, we define the set of upper triangular matrices

$$\mathcal{A}(\mathcal{G}) = \{a \in \mathrm{GL}_m(\mathbb{K}) \mid a_{ij} = 0 \text{ for } i \neq j \text{ with } j \not\to i \text{ in } \mathcal{G}\}. \tag{9.17}$$

Lemma 9.5.2 ([MRS21, Lemma 2.9])**.** *Let $\mathcal{G}$ be a DAG. The corresponding model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is the Gaussian model given by $\mathcal{A}(\mathcal{G})$: $\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathbf{g}}$.*

Proof. Let $\Psi = (\mathrm{I}_m - \Lambda)^{\dagger} \Omega^{-1} (\mathrm{I}_m - \Lambda) \in \mathcal{M}_{\mathcal{G}}^{\rightarrow}$, where Λ and Ω are as above, and set $a := \Omega^{-1/2}(\mathrm{I}_m - \Lambda)$. By construction, $\Psi = a^{\dagger} a$ and if $i \neq j$ with $j \not\to i$ in $\mathcal{G}$, then $\lambda_{ij} = 0$ and therefore $a_{ij} = -\omega_{ii}^{-1/2} \lambda_{ij} = 0$. This shows $\mathcal{M}_{\mathcal{G}}^{\rightarrow} \subseteq \mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathbf{g}}$.

Conversely, let $\Psi = b^{\dagger} b$ for some $b \in \mathcal{A}(\mathcal{G})$ and set $k_{ii} := \overline{b_{ii}} \, |b_{ii}|^{-1}$ for $i \in [m]$. The latter defines a diagonal matrix k such that $k^{\dagger} k = \mathrm{I}_m$ and $a := kb$ has positive diagonal entries $a_{ii} = |b_{ii}|$. We have $a^{\dagger} a = b^{\dagger} b = \Psi$ and, as multiplication with

k preserves the support, $a \in \mathcal{A}(\mathcal{G})$. Now, consider the positive-definite diagonal matrix $D := \operatorname{diag}(a_{11}^2, \ldots, a_{mm}^2)$ and the unipotent upper triangular matrix $U := \operatorname{diag}(a_{11}^{-1}, \ldots, a_{mm}^{-1})a$. Then $U^\dagger D U = \Psi$, so Ψ is of the form $(\mathrm{I}_m - \Lambda)^\dagger \Omega^{-1}(\mathrm{I}_m - \Lambda)$ for $\Omega = D^{-1}$ and strictly upper triangular $\Lambda = \mathrm{I}_m - U$. It remains to show that $\Lambda_{ij} = 0$ whenever $i \neq j$ such that $j \not\to i$ in $\mathcal{G}$. For such i, j we have $a_{ij} = 0$ since $a \in \mathcal{A}(\mathcal{G})$ and hence $\Lambda_{ij} = a_{ii}^{-1} a_{ij} = 0$. □

Given the previous lemma it is natural to ask when $\mathcal{A}(\mathcal{G})$ is a group, so that $\mathcal{M}^{\mathfrak{g}}_{\mathcal{A}(\mathcal{G})}$ is a Gaussian group model. To prove that transitivity is a necessary and sufficient condition we use the following lemma.

Lemma 9.5.3 ([MRS21, Lemma B.1])**.** *Let $\mathcal{A} = L \cap \mathrm{GL}_m(\mathbb{K})$, where L is a $\mathbb{K}$-linear subspace of $\mathbb{K}^{m\times m}$, and assume $\mathrm{I}_m \in \mathcal{A}$. Then $\mathcal{A}$ is a subgroup of $\mathrm{GL}_m(\mathbb{K})$ if and only if it is closed under multiplication.*

Proof. A group is closed under multiplication. Conversely, if $\mathcal{A}$ is closed under multiplication, we have to show that it is also closed under inverses. For a matrix $a \in \mathcal{A}$ let $f_a(t) = t^m + c_1 t^{m-1} + \cdots + c_m \in \mathbb{K}[t]$ be its characteristic polynomial. We know $c_m \neq 0$ because c_m is, up to sign, the determinant of a. Using the theorem of Cayley-Hamilton we deduce $-c_m^{-1}(a^{m-1} + c_1 a^{m-2} + \cdots + c_{m-1}\,\mathrm{I}_m)a = \mathrm{I}_m$, so

$$a^{-1} = -\frac{1}{c_m}\big(a^{m-1} + c_1 a^{m-2} + \cdots + c_{m-1}\,\mathrm{I}_m\big).$$

By assumption, $\mathrm{I}_m \in L$ and, as $\mathcal{A}$ is closed under multiplication, $a^k \in \mathcal{A} \subseteq L$ for all $k \geq 1$. Since L is a $\mathbb{K}$-vector space, we have $a^{-1} \in L$ and hence $a^{-1} \in \mathcal{A}$. □

Proposition 9.5.4 ([AKRS21a, Proposition 5.1])**.** *Let $\mathcal{G}$ be a DAG. The set of matrices $\mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$ is a group if and only if $\mathcal{G}$ is transitive, i.e., a TDAG.*

Proof. If $\mathcal{G}$ is not transitive, then there exist pairwise distinct indices $i, j, k \in [m]$ such that $j \to i$ and $k \to j$, but $k \not\to i$. Take the matrices $g = \mathrm{I}_m + E_{ij}$ (with ones on the diagonal and at the (i, j) entry, and zero elsewhere) and $h = \mathrm{I}_m + E_{jk}$. We have $g, h \in \mathcal{A}(\mathcal{G})$, but $gh \notin \mathcal{A}(\mathcal{G})$ as $(gh)_{ik} = 1$. Therefore, $\mathcal{A}(\mathcal{G})$ is not a group.

Conversely, assume that $\mathcal{G}$ is transitive. Note that any invertible diagonal matrix, in particular the identity I_m, is contained in $\mathcal{A}(\mathcal{G})$. Thus, it suffices to show that $\mathcal{A}(\mathcal{G})$ is closed under multiplication, by Lemma 9.5.3. Let $g, h \in \mathcal{A}(\mathcal{G})$ and consider $i \neq j$ such that $j \not\to i$. We need to prove that $(gh)_{ij} = 0$ to ensure $gh \in \mathcal{A}(\mathcal{G})$. Using $g_{ij} = h_{ij} = 0$ (as $j \not\to i$) we obtain

$$(gh)_{ij} = \sum_{k\in[m]} g_{ik} h_{kj} = \sum_{k\in[m]\setminus\{i,j\}} g_{ik} h_{kj}.$$

Since $\mathcal{G}$ is transitive we cannot have $g_{ik} \neq 0$ and $h_{kj} \neq 0$ for some $k \in [m]\setminus\{i, j\}$; otherwise $k \to i$ and $j \to k$ would yield $j \to i$, a contradiction. Hence, $(gh)_{ij} = 0$ which ends the proof. □

Example 9.5.5 ([AKRS21a, Example 5.2])**.** Let $\mathcal{G}$ be the TDAG $1 \longleftarrow 3 \longrightarrow 2$. The corresponding group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_3(\mathbb{K})$ consists of invertible matrices g of the form

$$g = \begin{pmatrix} * & 0 & * \\ 0 & * & * \\ 0 & 0 & * \end{pmatrix}.$$

By Proposition 9.5.4, we have that the Gaussian graphical model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is $\mathcal{M}_G^{\mathfrak{g}}$ and one computes that

$$\mathcal{M}_G^{\mathfrak{g}} = \{g^\dagger g \mid g \in G\} = \{\Psi \in \mathrm{PD}_3(\mathbb{K}) \mid \psi_{12} = \psi_{21} = 0\}.$$

is a 5-dimensional linear slice of $\mathrm{PD}_3(\mathbb{K})$. ♢

Given a TDAG $\mathcal{G}$, Proposition 9.5.4 puts us into the setting of Gaussian group models. The group $G := \mathcal{A}(\mathcal{G})$ is Zariski closed but in general *not* self-adjoint as it is upper triangular. Hence, we cannot apply the results from Section 9.3. However, we can prove the full correspondence for TDAG models differently. We start with the following observation.

Remark 9.5.6 (Weak Correspondence for TDAG models)*.*
For a TDAG $\mathcal{G}$ the group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$ is closed under non-zero scalar multiples and contains the orthogonal matrix $\mathrm{diag}(-1, 1, \ldots, 1)$ of determinant -1. Thus, the weak correspondence via the action of G_{SL}, see Theorem 9.2.7 respectively Theorem 8.2.3, holds for the TDAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_G^{\mathfrak{g}}$. ▽

We provide a simple lemma to state equivalences between stability notions under G_{SL} and linear (in)dependence conditions on the rows of $Y \in \mathbb{K}^{m \times n}$.

Lemma 9.5.7. *Let $M \in \mathbb{K}^{(1+\beta)\times n}$ with rows $M^{(0)}, M^{(1)}, \ldots, M^{(\beta)} \in \mathbb{K}^{1\times n}$ such that $M^{(0)} \notin \mathrm{span}\{M^{(1)}, \ldots, M^{(\beta)}\}$. Then there exists $w \in \mathbb{K}^n$ such that for any $x = (x_0, x_1, \ldots, x_n)$ we have*

$$x_0 = \sum_{l=1}^{n} w_l \sum_{j=0}^{\beta} x_j M_{j,l}.$$

Proof. Let $(e_0, e_1, \ldots, e_\beta)$ be the ordered standard basis of $\mathbb{K}^{1+\beta}$. The assumption $M^{(0)} \notin \mathrm{span}\{M^{(1)}, \ldots, M^{(\beta)}\}$ implies that

$$\ker\left(M^\dagger\right) \subseteq \mathrm{span}\{e_1, \ldots, e_\beta\} = \{0\} \times \mathbb{K}^\beta.$$

Therefore, e_0 is in the orthogonal complement of $\ker(M^\dagger)$, i.e., in the image of M. Hence, there is some $w \in \mathbb{K}^n$ with $Mw = e_0$. For the row x, the equation $x_0 = xe_0 = xMw$ yields the claim. □

Next, we give equivalences between stability notions under G_{SL} and the linear (in)dependence conditions on the rows of $Y \in \mathbb{K}^{m\times n}$ encountered in Theorem 6.3.16. For this, recall that $Y^{(i)}$ is the i^{th} row of Y and, by convention, the linear hull of the empty set is the zero vector space. The following statement will be generalized in Theorem 10.6.3.

Theorem 9.5.8. *Let $\mathcal{G}$ be a TDAG with group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$. For $Y \in \mathbb{K}^{m\times n}$, stability under G_{SL} relates to linear independence conditions:*

$$\begin{aligned}
&(a) \quad Y \text{ unstable} \quad \Leftrightarrow \quad \exists\, i \in [m]\colon \ Y^{(i)} \in \operatorname{span}\left\{Y^{(j)} : j \in \mathrm{pa}(i)\right\}\\
&(b) \quad Y \text{ polystable} \quad \Leftrightarrow \quad \forall\, i \in [m]\colon \ Y^{(i)} \notin \operatorname{span}\left\{Y^{(j)} : j \in \mathrm{pa}(i)\right\}\\
&(c) \quad Y \text{ stable} \quad \Leftrightarrow \quad \forall\, i \in [m]\colon \quad Y^{(i\cup \mathrm{pa}(i))} \text{ has full row rank.}
\end{aligned}$$

In particular, Y is semistable if and only if it is polystable.

Proof. First, assume there is some vertex $i \in [m]$ such that the row $Y^{(i)}$ is a $\mathbb{K}$-linear combination of its parent rows:

$$Y^{(i)} = \sum_{j\in \mathrm{pa}(i)} \lambda_j Y^{(j)}$$

Then the i^{th} row of $g \cdot Y$ is zero, where $g \in G_{\mathrm{SL}}$ has diagonal entries equal one and the only non-zero off-diagonal entries are $g_{ij} = -\lambda_j$, $j \in \mathrm{pa}(i)$. For $\varepsilon > 0$, let g_ε be the diagonal matrix with entries $(g_\varepsilon)_{ii} = \varepsilon^{-m+1}$ and $(g_\varepsilon)_{kk} = \varepsilon$, $k \neq i$. By construction, $g_\varepsilon \in G_{\mathrm{SL}}$ and $g_\varepsilon g \cdot Y \to 0$ for $\varepsilon \to 0$, so Y is G_{SL}-unstable.

Conversely, assume that $Y^{(i)} \notin \operatorname{span}\left\{Y^{(j)} : j \in \mathrm{pa}(i)\right\}$ for all vertices $i \in [m]$. Then $Y \neq 0$ and we will show that the G_{SL}-orbit of Y is Euclidean closed, i.e., Y is G_{SL}-polystable. By Lemma 2.4.3, it suffices to prove that $G_{\mathrm{SL}} \cdot Y$ is Zariski closed for $\mathbb{K} = \mathbb{C}$ and we show that via Popov's Criterion from Section 2.4. For this, we use the language of Section 2.4 with respect to the action of G_{SL} on $\mathbb{C}^{m\times n}$. In particular, $T = \mathrm{ST}_m(\mathbb{C})$ and the $x_{i,j} \in \mathbb{C}[G_{\mathrm{SL}}]$ for $i, j \in [m]$ denote the coordinate functions on G_{SL}.

Since $Y^{(i)} \notin \operatorname{span}\left\{Y^{(j)} : j \in \mathrm{pa}(i)\right\}$, we can apply Lemma 9.5.7 to the matrix $Y^{(i\cup \mathrm{pa}(i))}$ and the $x_{i,j}$, $j \in \{i\} \cup \mathrm{pa}(i)$. Hence, there exists $w \in \mathbb{C}^n$ with

$$x_{i,i} = \sum_{l=1}^{n} w_l \sum_{j\in\{i\}\cup \mathrm{pa}(i)} x_{i,j}Y_{j,l} = \sum_{l=1}^{n} w_l \sum_{j=1}^{m} Y_{j,l}x_{i,j}\,,$$

where we used in the final equality that $x_{i,j} = 0$ if $j \notin \{i\} \cup \mathrm{pa}(i)$. By Equation (2.28), this shows that $x_{i,i} \in R_Y$ for all $i \in [m]$ and hence we have

$$\forall\, (d_1, \ldots, d_m) \in \mathbb{Z}^m_{\geq 0}\colon \qquad \prod_{i\in[m]} x_{i,i}^{d_i} \in R_Y.$$

The latter exhaust all characters of $T = \mathrm{ST}_m(\mathbb{C})$ thanks to the fact that $\prod_{i\in[m]} x_{i,i}$ is the trivial character. We conclude $\mathfrak{X}_{G_{\mathrm{SL}}\cdot Y} = \mathfrak{X}(T)$ which is a group. Therefore, the orbit $G_{\mathrm{SL}} \cdot Y$ is Zariski closed by Popov's Criterion (Theorem 2.4.1).

Since the right hand side of (a) and (b) are opposites of each other and polystable implies semistable (the opposite of unstable), we have proven the equivalences in (a) and (b).

To prove part (c) it suffices, by part (b), to show that a polystable Y has finite G_{SL} stabilizer if and only if for all $i \in [m]$ the parent rows $Y^{(j)}$, $j \in \mathrm{pa}(i)$ are

linearly independent. Let Y be polystable. A matrix $g \in G_{\mathrm{SL}}$ is in the stabilizer of Y, i.e., $gY = Y$, if and only if for all $i \in [m]$

$$(gY)^{(i)} = g_{ii}Y^{(i)} + \sum_{j \in \mathrm{pa}(i)} g_{ij}Y^{(j)} = Y^{(i)}. \tag{9.18}$$

Since $Y^{(i)}$ is not in the linear span of its parent rows, Equation (9.18) implies that $g_{ii} = 1$ and $\sum_{j\in\mathrm{pa}(i)} g_{ij}Y^{(j)} = 0$. If $Y^{(j)}$, $j \in \mathrm{pa}(i)$ are linearly independent, then (9.18) has exactly one solution, namely $g_{ii} = 1$ and $g_{ij} = 0$ for all $j \in \mathrm{pa}(i)$. Thus, if for all $i \in [m]$ the $Y^{(j)}$, $j \in \mathrm{pa}(i)$ are linearly independent, then $(G_{\mathrm{SL}})_Y = \{\mathrm{I}_m\}$ is trivial and Y is stable. On the other hand, if there is some $i \in [m]$ such that $Y^{(j)}$, $j \in \mathrm{pa}(i)$ are linearly dependent, then Equation (9.18) has infinitely many solutions. Each solution $g_{ii} = 1$ and g_{ij}, $j \in \mathrm{pa}(i)$ of (9.18) gives rise to a unipotent matrix $g \in (G_{\mathrm{SL}})_Y$ by setting all other off-diagonal entries of g to zero. Therefore, $(G_{\mathrm{SL}})_Y$ is infinite and Y is not stable. □

Parts (a) and (b) of Theorem 9.5.8 constitute [AKRS21a, Theorem 5.3], which is proven in [AKRS21a] more ad-hoc and without using Popov's Criterion.[22] These parts in combination with the weak correspondence, Theorem 9.2.7, prove

$$\mathrm{mlt}_b(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = \mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = 1 + \max_{i\in[m]} |\,\mathrm{pa}(i)|.$$

This is [AKRS21a, Corollary 5.5] and recovers parts of the known Corollary 6.3.19 for *transitive* DAGs *without* using Theorem 6.3.16.

Now, combining Theorem 6.3.16 and Theorem 9.5.8 directly gives the full correspondence for TDAG models, which will be generalized to so-called RDAG models in Theorem 10.6.4.

Theorem 9.5.9 (Full Correspondence for TDAGs)**.** *Let $\mathcal{G}$ be a TDAG with group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$. Consider the TDAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_G^{\mathrm{g}}$ with tuple of samples $Y \in \mathbb{K}^{m\times n}$. Stability under the action of G_{SL} is related to ML estimation as follows.*

$$\begin{array}{lccc}
(a) & Y \text{ unstable} & \Leftrightarrow & \ell_Y \text{ not bounded from above} \\
(b) & Y \text{ semistable} & \Leftrightarrow & \ell_Y \text{ bounded from above} \\
(c) & Y \text{ polystable} & \Leftrightarrow & \text{MLE exists} \\
(d) & Y \text{ stable} & \Leftrightarrow & \text{unique MLE exists}
\end{array}$$

We point out that the equivalence in part (d) also holds for $\mathbb{K} = \mathbb{R}$, which is not the case in the self-adjoint situation, compare Theorem 9.3.6.

Remember that the G_{SL}-stabilizer of Y acts from the right on the set of MLEs given Y, compare Proposition 9.2.4. In the self-adjoint situation this action is transitive (Proposition 9.3.3). This can be further strengthened for TDAG models as follows.

[22] We presented the proof of Theorem 9.5.8 via Popov's Criterion to advertise this algebraic tool for testing polystability. We remark that generalizing this proof led to the concept of augmented sample matrices $M_{Y,s}$, compare Section 10.7 and Lemma 10.7.8. The matrices $M_{Y,s}$ are indispensable for several main results of Chapter 10.

Proposition 9.5.10. *Let $\mathcal{G}$ be a TDAG with group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$. Consider the TDAG model $\mathcal{M}_G^{\mathbf{g}}$ and assume $Y \in \mathbb{K}^{m\times n}$ has an MLE $\hat{\Psi} \in \mathcal{M}_G^{\mathbf{g}}$. Then the group action of $(G_{\mathrm{SL}})_Y$ on the the set of MLEs given Y from Proposition 9.2.4 is free and transitive. In other words, we have a bijection*

$$(G_{\mathrm{SL}})_Y \to \{\textit{MLEs given } Y\}, \quad g \mapsto g^{\dagger}\hat{\Psi}g.$$

A proof is omitted as the statement is a special case of Proposition 10.7.9, which is proven in Section 10.7.

Example 9.5.11 (Saturated model as a TDAG model)**.** Remember that the saturated Gaussian model $\mathcal{M} = \mathrm{PD}_m(\mathbb{K})$ arises as the Gaussian group model $\mathcal{M}^{\mathbf{g}}_{\mathrm{GL}_m(\mathbb{K})}$, studied in Example 9.3.8. However, it is also induced by the group $\mathrm{B}_m(\mathbb{K})$ of upper invertible matrices: $\mathcal{M}^{\mathbf{g}}_{\mathrm{B}_m(\mathbb{K})} = \mathrm{PD}_m(\mathbb{K})$. This is the Gaussian group model given by the "full" TDAG, i.e., the TDAG on vertex set $[m]$ that contains a directed edge $i \leftarrow j$ whenever $i < j$.

An interesting distinction between these two viewpoints arises for the action of the stabilizer $(G_{\mathrm{SL}})_Y$ on the set of MLEs given $Y \in Y \in \mathbb{K}^{m\times n}$, Proposition 9.2.4. For $G = \mathrm{GL}_m(\mathbb{K})$ we have a transitive action by 9.3.3 that is in general not free. In contrast, Proposition 9.5.10 for TDAGs gives a transitive *and free* action for $G = \mathrm{B}_m(\mathbb{K})$. Hence, the restriction to upper triangular matrices excludes possible redundancies, i.e., distinct stabilizer elements giving the same MLE.

The (T)DAG perspective recovers classical knowledge as given in Example 6.3.8. Since vertex 1 has all other $m-1$ vertices as parents, Corollary 6.3.19 yields the known value for the ML thresholds:

$$\mathrm{mlt}_b\big(\mathrm{PD}_m(\mathbb{K})\big) = \mathrm{mlt}_e\big(\mathrm{PD}_m(\mathbb{K})\big) = \mathrm{mlt}_u\big(\mathrm{PD}_m(\mathbb{K})\big).$$

Moreover, we have $Y^{1\cup\mathrm{pa}(1)} = Y$ and hence Theorem 6.3.16(c) shows that there is a unique MLE if and only if Y has full row rank. Otherwise, the log-likelihood ℓ_Y is not bounded from above, by Theorem 6.3.16(a). ♢

Example 9.5.12 (based on[AKRS21a, Example 5.8])**.**
Let $\mathcal{G}$ be the TDAG $2 \longrightarrow 1 \longleftarrow 3$. The corresponding group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_3(\mathbb{K})$ consists of invertible matrices of the form

$$g = \begin{pmatrix} * & * & * \\ 0 & * & 0 \\ 0 & 0 & * \end{pmatrix}.$$

We know from Corollary 6.3.19 that $\mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = 2+1 = 3$ as vertex 1 has two parents. A sample matrix $Y \in \mathbb{K}^{m\times n}$ is G_{SL}-polystable if and only if $Y^{(2)}, Y^{(3)} \neq 0$ and $Y^{(1)}$ is not in the linear span of $Y^{(2)}$ and $Y^{(3)}$, compare Theorem 9.5.8. Otherwise, it is unstable. Furthermore, Y is stable if and only if it has full row rank, since $Y^{(1\cup\mathrm{pa}(1))} = Y$.

Let $n = 2$ and consider the sample matrix

$$Y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

It is polystable and hence there exists an MLE given Y. One can check that Y is of minimal norm in its G_{SL}-orbit. Therefore, $2\,\mathrm{I}_3$ is an MLE given Y using Theorem 9.2.7 and that $\lambda = 2$ minimizes $x \mapsto \frac{3}{2}x - 3\log(x)$, see Lemma 8.2.2(ii). Moreover, the G_{SL}-stabilizer of Y is in bijection with the set of MLEs given Y, by Proposition 9.5.10. We have

$$(G_{\mathrm{SL}})_Y = \left\{ \begin{pmatrix} 1 & t & -t \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} : t \in \mathbb{K} \right\}, \text{ thus } \left\{ 2 \begin{pmatrix} 1 & t & -t \\ \bar{t} & |t|^2+1 & -|t|^2 \\ -\bar{t} & -|t|^2 & |t|^2+1 \end{pmatrix} : t \in \mathbb{K} \right\}$$

is the set of MLEs given Y. Hence, there are infinitely many MLEs given Y. $\diamondsuit$

The next proposition gives a precise criterion when the null cone under the G_{SL} action, i.e., the set of sample matrices $Y \in \mathbb{K}^{m\times n}$ for which ℓ_Y is not bounded from above, is Zariski closed. This extends and clarifies [AKRS21a, Corollary 5.7], which is Proposition 9.5.13(ii).

For this, we use the notion of an unshielded collider from Definition 6.3.13. Furthermore, the *depth* $d(\mathcal{G})$ of a DAG $\mathcal{G}$ is the number of arrows in a maximal directed path in $\mathcal{G}$. Note that $d(\mathcal{G}) \leq m-1$ and if $\mathcal{G}$ is transitive then actually $d(\mathcal{G}) \leq \max_{i\in[m]} |\operatorname{pa}(i)| < \operatorname{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$.

Proposition 9.5.13. *Let $\mathcal{G}$ be a TDAG with group $G := \mathcal{A}(\mathcal{G}) \subseteq \mathrm{GL}_m(\mathbb{K})$ and consider the action of G_{SL} on $\mathbb{K}^{m\times n}$ via left multiplication.*

(i) Let $n < \operatorname{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$. Then the Zariski closure of the null cone is $\mathbb{K}^{m\times n}$. The null cone is Zariski closed, i.e., equal to $\mathbb{K}^{m\times n}$, if and only if $n \leq d(\mathcal{G})$.

(ii) Let $n \geq \operatorname{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$. Then the irreducible components of the Zariski closure of the null cone are determinantal varieties: each component is defined by the maximal minors of the submatrix $Y^{(s\cup\operatorname{pa}(s))}$, where s is a childless vertex. The null cone is Zariski closed if and only if $\mathcal{G}$ has no unshielded colliders.

Proof. First, let $n < \operatorname{mlt}_b(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = \operatorname{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$. By definition of $\operatorname{mlt}_b(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$ and Theorem 9.5.9(a), almost all Y are G_{SL}-unstable, so the null cone is Zariski dense in $\mathbb{K}^{m\times n}$. This shows the first part of (i).

Now, additionally assume $n \leq d := d(\mathcal{G})$. There is some directed path

$$p_0 \longleftarrow p_1 \longleftarrow p_2 \longleftarrow \cdots \longleftarrow p_d$$

in $\mathcal{G}$. The transitivity of $\mathcal{G}$ implies that $p_{j+1}, \ldots, p_d$ are parents of p_j for all $j = 0, 1, \ldots, d-1$. Now, for any $Y \in \mathbb{K}^{m\times n}$ the row vectors $Y^{(p_j)} \in \mathbb{K}^{1\times n}$, $j = 0, 1, \ldots, d$ are linearly dependent as $n < d+1$. Therefore, there is some non-trivial linear combination $\sum_j \lambda_j Y^{(p_j)} = 0$. For the minimal k such that $\lambda_k \neq 0$ we see that $Y^{(p_k)}$ is a linear combination of (some of) its parent rows. Hence, Y is G_{SL}-unstable by Theorem 9.5.8(a). Thus, the null cone equals $\mathbb{K}^{m\times n}$.

Conversely, if $n > d = d(\mathcal{G})$, then we construct a polystable Y as follows. Denote by $d(i)$ the number of arrows in a longest path in $\mathcal{G}$ *starting* at i. Then $0 \leq d(i) \leq d$, $d(i) = 0$ if and only if vertex i is childless, and if $p \to i$ then $d(i) < d(p)$. Fix linear independent row vectors $r_0, r_1, \ldots, r_d \in \mathbb{K}^{1\times n}$ using $n \geq d+1$. Now, define $Y \in \mathbb{K}^{m\times n}$ by setting $Y^{(i)} := r_{d(i)}$ for all $i \in [m]$. By construction,

the parent rows of $Y^{(i)} = r_{d(i)}$ are all contained in $\{r_{d(i)+1}, \ldots, r_{d(\mathcal{G})}\}$. Thus, $Y^{(i)}$ is not in the linear span of its parent rows and hence Y is polystable.

To prove (ii), assume that $n \geq \mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = 1 + \max_{i\in[m]} |\operatorname{pa}(i)|$. The null-cone is the finite union of all

$$\mathcal{L}(i) := \left\{ Y \in \mathbb{K}^{m\times n} \mid Y^{(i)} \in \operatorname{span}\{Y^{(j)} \mid j \in \operatorname{pa}(i)\} \right\}$$

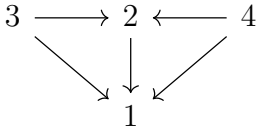
where $i \in [m]$. Taking the Zariski closure commutes with finite unions, hence the Zariski closure of the null cone is the finite union of $\overline{\mathcal{L}(i)}^{Z}$. Since $n \geq 1 + \max_{i\in[m]} |\operatorname{pa}(i)|$, the closure $\overline{\mathcal{L}(i)}^{Z}$ can be described via the the maximal minors of the matrix $Y^{(i\cup\operatorname{pa}(i))}$. Thus, the Zariski closure of the null cone actually contains all matrices that are *not* stable, see Theorem 9.5.8(c). If a vertex i has child c, then by transitivity all parents of i are also parents of c. Hence, $Y^{(i\cup\operatorname{pa}(i))}$ is a submatrix of $Y^{(c\cup\operatorname{pa}(c))}$ and so $\overline{\mathcal{L}(i)}^{Z} \subseteq \overline{\mathcal{L}(c)}^{Z}$. This shows the first part of (ii).

Recall from Remark 6.3.12 that we assume $i < j$ whenever $i \leftarrow j$ in $\mathcal{G}$. Assume $\mathcal{G}$ has no unshielded colliders. Let Y be a matrix in the Zariski closure of the null cone. Then there is some vertex $i = p_0 \in [m]$ such that $Y \in \overline{\mathcal{L}(i)}^{Z}$, i.e., there is a non-trivial linear combination $\sum_{j=0}^{s} \lambda_j Y^{(p_j)} = 0$, where $p_1, \ldots, p_s$ are the parents of $i = p_0$. Let k be the smallest integer with $\lambda_k \neq 0$. Then $Y^{(p_k)}$ is in the linear span of $Y^{(p_{k+1})}, \ldots, Y^{(p_s)}$. If p_t for some $t \in \{k+1, \ldots, s\}$ would not be a parent of p_k, then necessarily $k > 0$ (i.e., $p_k \neq i$) and so $\mathcal{G}$ would have the unshielded collider $p_t \rightarrow i \leftarrow p_k$; a contradiction. Therefore, $Y^{(p_k)}$ is in the linear span of its parent rows and hence Y is unstable. Thus, the null cone is Zariski closed.

On the other hand, assume $\mathcal{G}$ has an unshielded collider $j \rightarrow i \leftarrow k$ where $j < k$. If i has several pairs of parents that give an unshielded collider, then consider a pair $j < k$ where k is maximal. This ensures that any parent p of k is also a parent of j as follows. We have $p > k > j$, so in particular $j \not\rightarrow p$. By transitivity $p \rightarrow k$ and $k \rightarrow i$ show that p is a parent of i. Thus, p must be a parent of j as otherwise $j \rightarrow i \leftarrow p$ would be an unshielded collider with $p > k$, which contradicts the maximality of k. With this we construct a matrix Y which is not in the null cone but in its Zariski closure. Each row of Y, except for the k^{th} row, is chosen such that it is not in the linear span of its parent rows. This is possible as $n \geq \mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$. In particular, the row $Y^{(j)}$ is not in the linear span of its parent rows, which include the parent rows of $Y^{(k)}$ by the above argument. Thus, setting $Y^{(k)} := Y^{(j)}$ ensures that Y is polystable by Theorem 9.5.8(b). Moreover, the parent rows $Y^{(j)}$ and $Y^{(k)}$ of $Y^{(i)}$ are linearly dependent, so Y is contained in $\overline{\mathcal{L}(i)}^{Z}$ and hence in the Zariski closure of the null cone. □

Let us illustrate the previous proposition in an example.

Example 9.5.14. Let $m = 4$ and consider the TDAG $\mathcal{G}$ given by



We have $d(\mathcal{G}) = 2$ and $\mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = 4$, as vertex 1 has three parents. Denote the corresponding group by $G := \mathcal{A}(\mathcal{G})$ and consider the usual G_{SL} action on the sample space $\mathbb{K}^{4\times n}$.

For sample size $n = 1, 2$ the null cone equals $\mathbb{K}^{4\times n}$, by Proposition 9.5.13(i). Alternatively, this can be checked via Theorem 9.5.8(a) by case distinction.

If $n = 3 < \mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow})$, then the null cone is only Zariski dense in $\mathbb{K}^{4\times 3}$ as $3 = n > d(\mathcal{G}) = 2$. For example, the sample matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \tag{9.19}$$

is polystable and was constructed using $d(1) = 0$, $d(2) = 1$, $d(3) = d(4) = 2$ and the recipe from the proof of Proposition 9.5.13.

If $n = \mathrm{mlt}_e(\mathcal{M}_{\mathcal{G}}^{\rightarrow}) = 4$, the Zariski closure of the null cone has one irreducible component given by the sink 1, see Proposition 9.5.13(ii). Since $Y^{(1\cup \mathrm{pa}(1))} = Y$ has exactly one maximal minor, the Zariski closure of the null cone is the set of singular matrices $\{Y \in \mathbb{K}^{4\times 4} \mid \det(Y) = 0\}$. Furthermore, $\mathcal{G}$ has the unshielded collider $3 \to 2 \leftarrow 4$, so the null cone is not Zariski closed. Indeed, if we append a zero column to the matrix from Equation (9.19), then we obtain a polystable Y' that is singular. ♢

Finally, we describe the implications of the above results for undirected Gaussian graphical models from Example 6.3.10, see also [Sul18, Chapter 13]. Remember that a Gaussian graphical model on an *undirected* graph $\mathcal{G}$ is given by all concentration matrices Ψ such that $\Psi_{ij} = 0$ whenever the edge i — j is missing from $\mathcal{G}$. A natural question is to determine which undirected Gaussian graphical models are Gaussian group models, i.e., of the form $\mathcal{M}_G^{\mathbf{g}}$ for some group $G \subseteq \mathrm{GL}_m(\mathbb{K})$. For instance, note that the undirected model corresponding to 1 — 2 — 3 is the same as the directed model from Example 9.5.5. We argue that any undirected model that is a Gaussian group model is covered by TDAGs. The following is very brief and we refer to the mentioned literature.

First, note that the directed model of any TDAG without unshielded colliders equals the undirected model of its underlying undirected graph, see Theorem 6.3.14 or [AMP97, Theorem 3.1]. Conversely, a necessary condition for an undirected graphical model to be a Gaussian group model can be obtained from [LM07, Theorem 2.2]: an undirected Gaussian graphical model is a transformation family[23] if and only if the graph $\mathcal{G}$ has neither 4-cycles nor 4-chains as induced subgraphs. There are two consequences of these conditions. One is that there is a way to *direct* the edges in $\mathcal{G}$ so that there are no unshielded colliders. The other consequence is that this can be done in such a way so that the undirected model coincides with the directed model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$, and the directed graph must be a TDAG, see page 7 of the supplementary material of [DKZ13]. In summary, we have the following equivalence.

[23] Recall that any Gaussian group model is a transformation family, compare Remark 9.2.2(a).

Remark 9.5.15 ([AKRS21a, Remark 5.9]). The undirected graphical models that are Gaussian group models are the TDAG models without unshielded colliders. They are exactly those models whose sets of tuples of n samples with unbounded likelihood are Zariski closed for all $n \geq \mathrm{mlt}_e$, by Proposition 9.5.13. ▽

9.6 Discussion and Outlook

In Subsection 9.6.1 we discuss related literature on Gaussian group models and afterwards, in Subsection 9.6.2, we compare Gaussian group models to log-linear models from Chapter 7.

9.6.1 Related Literature

In the following we comment on literature related to this chapter respectively to [AKRS21a]. We start with works that are contained in this thesis.

The companion paper [AKRS21b], presented in Chapter 7, can be seen as a discrete counterpart of [AKRS21a]. We discuss similarities and differences between the Gaussian setting and the discrete setting of log-linear models in a separate subsection below.

The theory of Gaussian group models and its relation to TDAG models (Section 9.5) stimulated further research on directed Gaussian graphical models [MRS21]. We present this work in detail in Chapter 10.

Now, we focus on papers that are not co-authored by the author of this thesis. Recently, there has been a flurry of new results on ML estimation of matrix and tensor normal models. For matrix normal models, the paper [DKH21] gave new characterizations of ML estimation and new bounds on ML thresholds. In Section 9.4 we compared some of their results to those from [AKRS21a].

All ML thresholds for matrix normal models have been completely characterized in [DM21], by crucially using the relations between invariant theory and ML estimation presented in Section 9.3. Derksen and Makam translate the problem of computing ML thresholds via the dictionary from Theorem 9.4.1 to generic semi/poly/stability and use invariant theory for representations of the n-Kronecker quiver; see Example 1.3.8 for the n-Kronecker quiver.

At first glance, the solution via invariant theory, a completely different mathematical area, is certainly surprising and might seem unnatural. A posteriori, the proof through invariant theory adjusts this first impression. As pointed out in [DM21, Section 1.3], there is a very interesting change of viewpoint thanks to the invariant theory perspective. Consider the matrix normal model $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$. From a statistical point of view, when studying ML thresholds it is natural to fix the dimensions m_1 and m_2, and to let the sample size n vary. Then the behaviour of ML estimation seems to be rather "wild", i.e., it is difficult to spot a pattern; compare [DKH21; DM21]. On the other hand, the representation theory

of the n-Kronecker quiver highly depends on the number n of arrows.[24] Thus, through the lens of invariant theory, when studying generic semi/poly/stability it is natural to fix n and let the dimensions m_1 and m_2 vary. This viewpoint unravels the seemingly wild behaviour and yields a clear picture! For illustration and convenience of the reader, we state the main result here.

Theorem 9.6.1 (ML thresholds for matrix normal model, [DM21, Theorem 1.3])**.** *Consider the matrix normal model $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$, where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. Let d be the greatest common divisor of m_1 and m_2. Set $r := (m_1^2+m_2^2-d^2)/(m_1 m_2)$. Then the ML thresholds for $\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$ satisfy $\mathrm{mlt}_b = \mathrm{mlt}_e$, and existence and uniqueness threshold are given as follows:*

1. *If $m_1 = m_2 = 1$, then $\mathrm{mlt}_e = \mathrm{mlt}_u = 1$.*
2. *If $m_1 = m_2 > 1$, then $\mathrm{mlt}_e = 1$ and $\mathrm{mlt}_u = 3$.*
3. *If $m_1 \neq m_2$ and $r \in \mathbb{Z}$, then $\mathrm{mlt}_e = r$. If $d = 1$, then $\mathrm{mlt}_u = r$, and if $d > 1$, then $\mathrm{mlt}_u = r + 1$.*
4. *If $m_1 \neq m_2$ and $r \notin \mathbb{Z}$, then $\mathrm{mlt}_e = \mathrm{mlt}_u = \left\lceil \frac{m_1^2+m_2^2}{m_1 m_2} \right\rceil$.*

Only shortly afterwards, the strong/full correspondence in Theorem 9.3.6 even led to a full determination of ML thresholds for tensor normal models [DMW22]. There, the authors use that the Castling transform on tensors preserves generic semi/poly/stability [DMW22, Section 3]. The main result [DMW22, Theorem 1.1] contains Theorem 9.6.1 as a special case.

Remember that $\mathcal{M}_G^{\mathrm{g}}$ for a Zariski closed self-adjoint group G is a totally geodesic submanifold of $\mathrm{PD}_m(\mathbb{K})$, Theorem 9.3.1, and that the log-likelihood is a geodesically convex function on $\mathcal{M}_G^{\mathrm{g}}$. This has been observed for matrix normal models in [Wie12]. Geodesic convexity has been applied in [DKH21; FORW21]. Actually, in [FORW21] it is a crucial tool to study (near) optimal sample complexity of matrix and tensor normal models. The main result for tensor normal models is [FORW21, Theorem 2.4], which can be strengthened for matrix normal models [FORW21, Theorem 2.7]. Moreover, the flip-flop algorithm is shown to efficiently compute the MLE with high probability, [FORW21, Theorems 2.9 and 2.10]. Theorem 9.3.1 and the outlined algorithmic consequences in Subsection 9.3.1 raise the following questions on generalizing the studies of [FORW21].

Problem 9.6.2. *Let $G \subseteq \mathrm{GL}_m(\mathbb{K})$ be a Zariski closed self-adjoint group and consider the Gaussian group model $\mathcal{M}_G^{\mathrm{g}}$. Can one, similarly to [FORW21], characterize (near) optimal sample complexity of $\mathcal{M}_G^{\mathrm{g}}$ using geodesic convexity? Moreover, do the first and/or second order method from [BFG+19] yield, with high probability, an efficient computation of the MLE?*

Now, we turn from geodesic convexity to Gaussian group models that are convex in the usual Euclidean sense. Such models are studied in [Ish21]. A complete characterization of Euclidean convex Gaussian group models is provided

[24] Indeed, for $n = 1$ the quiver is *of finite representation type*, for $n = 2$ the quiver is *tame* while for $n = 3$ it is so-called *wild*. We refer to [DW17] for details.

in [Ish21, Proposition 2 and Theorem 2]. Invariant theory is an important proof ingredient; more precisely, Vinberg theory (see [Wal17, Section 3.7]) is applied. Furthermore, the uniqueness threshold mlt_u is computed [Ish21, Theorem 4]. It is also shown that, if there exists a unique MLE, then the MLE is a rational function in the samples [Ish21, Theorem 3].

In Section 9.4 we studied ML estimation for matrix normal models via operator scaling (i.e., the left-right action). We remark that operator scaling was also used in [FM20] to study a different estimator from statistics: Tyler's M estimator for elliptical distributions. The authors prove results on the sample complexity [FM20, Theorems 1.1 and 1.2] of the estimator and they show that Tyler's iterative procedure converges quickly with high probability [FM20, Theorem 1.3].

9.6.2 Comparison with log-linear models

We highlight similarities and differences between the multivariate Gaussian setting from [AKRS21a] studied in this chapter and the discrete setting of log-linear models from [AKRS21b] presented in Chapter 7. This is based on [AKRS21b, Section 6]. We start by comparing the two statistical settings.

In the discrete setting, a model is given as a subset of the $(m-1)$-dimensional probability simplex $\Delta_{m-1} \subseteq \mathbb{R}^m$. In comparison, in the multivariate Gaussian setting, a model is given by a set of concentration matrices in the cone of positive definite matrices $\mathrm{PD}_m(\mathbb{K})$. For a discrete model $\mathcal{M} \subseteq \Delta_{m-1}$ the data/sufficient statistics is a vector of counts $u \in \mathbb{Z}^m_{\geq 0}$ with $u_+ = n$ the total numbers of observations. The log-likelihood given u at $p \in \mathcal{M}$ is $\sum_{j=1}^m u_j \log(p_j)$, see (6.4). In comparison, for a Gaussian model the data is a tuple of samples $Y \in (\mathbb{K}^m)^n$, the sample covariance matrix $S_Y = \frac{1}{n}\sum_{i=1}^n Y_i Y_i^\dagger$ provides a sufficient statistics and the log-likelihood at Y is given by $\log\det(\Psi) - \mathrm{tr}(\Psi S_Y)$, see (6.8).

Stability. In both settings we link notions of stability under a group action to ML estimation in statistical models: for log-linear models in Theorem 7.2.1 and for Gaussian group models in, e.g., Theorems 9.2.7 and 9.3.6. However, a main difference is where the dependence on the data enters. For log-linear models we consider an action of $\mathrm{GT}_d(\mathbb{C})$ on $\mathbb{C}^m$ which *depends on the data*, and we always study stability of the all-ones vector $\mathbb{1}_m$. In contrast, for a Gaussian group model $\mathcal{M}^{\mathbf{g}}_G$, where $G \subseteq \mathrm{GL}_m(\mathbb{K})$, we always use the action of G on the sample space $(\mathbb{K}^m)^n$ via left-multiplication, while we consider stability notions *for the observed data*, i.e., the tuple of samples.

For log-linear models, the log-likelihood is always bounded from above and the all-ones vector cannot be unstable. In contrast, in the Gaussian setting a tuple of samples is unstable if and only if the log-likelihood is not bounded from above. In both cases, semistability is equivalent to the log-likelihood being bounded from above and polystability is equivalent to the existence of an MLE. In the log-linear case, the MLE is unique if it exists, while for Gaussian group models there may be infinitely many. In fact, the existence of a unique MLE for Gaussian group models often relates to stability of a tuple of samples, see Theorems 9.3.6 and 9.5.9. In contrast, for log-linear models the all-ones vector is never stable.

MLE computation. An important similarity between the log-linear and Gaussian settings is that norm minimizers under the respective group actions give an MLE (if it exists), see Theorem 7.2.3 and Theorem 9.2.7. For log-linear models, we compute real MLEs from complex torus orbits. For Gaussian group models, we compute the MLE over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ from orbits over the same field $\mathbb{K}$. If the all-ones vector is semistable but not polystable, Theorem 7.2.3 yields the extended MLE. However, in the Gaussian case, if a tuple of samples Y is semistable but not polystable there is usually no meaningful notion of extended MLE, compare Remark 6.3.4.

Scaling. From the point of view of scaling algorithms, Sinkhorn's algorithm is a common origin to both the log-linear and the Gaussian settings. As we described in Section 7.3, Sinkhorn scaling to target marginals is iterative proportional scaling (IPS) for the independence model and this extends to IPS for a general log-linear model. On the Gaussian side, Sinkhorn scaling generalizes to alternating minimization procedures for computing MLEs of matrix normal models and tensor normal models. This algorithm is used both in invariant theory for norm minimization and in statistics to compute the MLE, compare Subsection 9.4.4.

Since norm minimizers yield an MLE in both settings, one can use scaling algorithms from invariant theory to approximate an MLE; compare right hand side of Figures 7.1 and 9.1. Remember that the above discussion naturally motivates to regard geodesic convex methods for Norm Minimization 3.1.3 and the Scaling Problem 3.1.4 as IPS for Gaussian group models $\mathcal{M}_G^{\mathfrak{g}}$ with Zariski closed self-adjoint group G, compare Subsection 9.3.1.

Exponential Families and Transformation Families. We conclude by pointing out the following with respect to exponential families and transformation families, compare Definition 6.1.4.[25] Remember that log-linear models are discrete regular exponential families [Sul18, Section 6.2]. However, in general they are not transformation families: the group of bijections on the sample space $[m]$ is finite and hence cannot act transitively on an infinite log-linear model.

Gaussian group models are examples of transformation families, compare Remark 9.2.2, and they are *sub*models of the saturated Gaussian model, which is a Gaussian regular exponential family [Sul18, Section 6.3]. In general, a Gaussian group model itself cannot be a regular exponential family. Otherwise an MLE would be unique if it exists [Sul18, Corollary 7.3.8], but this is usually not the case, compare Proposition 9.3.4 or Proposition 9.5.10.

Despite the mentioned differences between the discrete and Gaussian setting, it is interesting and natural to ask the following.

Problem 9.6.3. *Is there a unifying concept that links invariant theory to maximum likelihood estimation, e.g., in the context of (sub)models of exponential families? Or in the context of transformation families?*

[25] [AKRS21b, Section 6] imprecisely states that "log-linear models and the Gaussian group models [...] are examples of exponential transformation families". The paragraph clarifies this, also in view of Definition 6.1.4 used in this thesis. We remind the reader that the term *transformation family* is ambiguous in the literature.

More specifically, is there a unifying theory that covers Chapters 7 and 9 at the same time?[26]

[26] Admittedly, an affirmative answer to this specific question does not seem very likely to the author, given the mentioned differences.

Chapter 10

Restricted DAG Models

Graphical models play a fundamental role in statistics and have manifold applications [Lau96; MDLW19]. Intuitively, the graph encodes the following statistical meaning: each vertex represents a random variable, and the edges between variables reflect their statistical dependence [VP90]. In the Gaussian case we already came across two kinds of graphical models. Example 6.3.10 defined undirected Gaussian graphical models, while Definition 6.3.11 recalled Gaussian graphical models on *directed* acyclic graphs (DAGs). Remember that DAG models are also called Gaussian Bayesian networks and they are linear structural equation models with independent errors, see [Drt18] and [Sul18, Section 16.2]. DAG models have been applied to many different contexts such as cell signalling [SPP+05], gene interactions [FLNP00] and causal inference [Pea09].

In this chapter, we introduce and study Gaussian graphical models on DAGs with coloured vertices and edges. The colours impose symmetries in the model: if two vertices or two edges have the same colour, then their parameters in the model must be the same. We call such models *RDAG models*, where the 'R' stands for restricted, cf. [HL08]. RDAG models contain DAG models as a special case, Remark 10.1.10. In that regard, many results of this chapter generalize statements on (T)DAG models from Sections 6.3 and 9.5.

The whole chapter is based on the preprint [MRS21], which is joint work with Visu Makam and Anna Seigal.

Motivation. We state three main motivations for studying RDAG models.

First, RDAG models are a natural analogue of so-called restricted concentration (RCON) models, which have been introduced in [HL08]; compare Definition 10.2.4 below. RCON models are submodels of undirected Gaussian models (Example 6.3.10) and obey symmetries among the entries of the concentration matrix according to a graph colouring. It is interesting to study possible connections between RDAG and RCON models, similar to the known connection between DAG models and undirected Gaussian graphical models in Theorem 6.3.14. This may allow to study RCON models through the lens of RDAG models.

Second, vertex and edge symmetries appear in various applications, such as in the study of longitudinal data [AFS16; VAAW16], or clustered variables [GM15; HL08]. Therefore, it is desirable to include these symmetries in the model itself. The coloured directed graph gives an intuitive pictorial description of these symmetry conditions.

Third, we aim to decrease the maximum likelihood (ML) thresholds (Definition 6.3.5), since for applications it is desirable to have small ML thresholds.

We comment that innovative ideas have been used to find maximum likelihood thresholds in graphical models [Buh93; DFKP19; GS18; Uhl12] and for estimating MLEs from too few samples [FHT08; WZV+04]. Removing edges from a graph can lower the threshold [Uhl12; Lau96], but there is a trade-off: removing edges imposes more conditional independence among the variables. This is why, instead, we aim to decrease the maximum likelihood threshold by introducing symmetries.

Let us illustrate these motivations in the following running example that we shall use throughout the chapter.

Example 10.0.1. Consider the coloured graph $\textcircled{1} \leftarrow \boxed{3} \rightarrow \textcircled{2}$, with blue (circular) vertices $\{1, 2\}$, black (square) vertex 3 and two red edges. The RDAG model is the linear structural equation model

$$y_1 = \lambda y_3 + \varepsilon_1, \qquad y_2 = \lambda y_3 + \varepsilon_2, \qquad y_3 = \varepsilon_3,$$

where $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, \omega)$ and $\varepsilon_3 \sim \mathcal{N}(0, \omega')$, i.e., ω is the variance of the blue vertices 1 and 2, and ω' is the variance of black vertex 3. The third parameter λ is the regression coefficient given by a red edge.

Regarding our three motivations we note the following. First, Example 10.2.5 will show that the above RDAG model equals its induced RCON model. Hence, one may study the latter through the former. Second, we use this example to model the dependence of two daughters' heights on the height of their mother, and we compute the MLE given some sample data; see Example 10.3.10. Third, in Example 10.3.7 we will see that the ML threshold mlt_u for uniqueness is one. In contrast, if we remove the colours the resulting DAG model has uniqueness threshold two, compare Corollary 6.3.19. ♢

Related Models. To the knowledge of the authors of [MRS21], RDAG models have not been defined before in the literature. We comment on some related models. The assumption of equal variances from [PB14] is the special case of an RDAG model, where all vertex colours are the same. Special colourings encode exchangeability between variables, or invariance under a group of permutations. A graphical model is combined with group symmetries in the directed setting in [Mad00] and in the undirected setting in [AM98; SC12]. RDAG models also relate to the fused graphical lasso [DWW14], which penalises differences between parameters on different edges, whereas in an RDAG model the parameters on edges of the same colour must be equal.

Main Results. As a generalization of Theorem 6.3.16 for DAG models, we characterize the existence and uniqueness of MLEs via linear algebraic properties of the sample data, see Theorem 10.3.6. We give a closed-form formula for MLEs in an RDAG model, as a collection of least squares estimators, see Algorithm 10.1. In Theorem 10.4.9 we provide upper and lower bounds on ML thresholds for RDAG models. Our results show that RDAG thresholds are less or equal to the DAG thresholds, and that high symmetry decreases the thresholds.

Thus, the third motivation about decreasing ML thresholds is achieved. Furthermore, we compare RDAG MLEs to uncoloured DAG MLEs via simulations in Section 10.5. All results hold with an assumption on the graph colouring called *compatibility* (Definition 10.1.6), which allows to view RDAG models in a natural way as Gaussian models via symmetrization. It is an open problem to extend our results to the non-compatible setting, as well as to directed graphs with cycles. It is also an open problem to find the *exact* ML thresholds, see Problem 10.4.10.

Regarding RCON models, the undirected analogue of RDAG models, we note the following. Although a motivation for the graph colouring in RCON models is to lower the maximum likelihood threshold, there are relatively few graphs for which the threshold is known: colourings of the four cycle are studied in [Uhl12, §6], [SU10, §5], while an example with five vertices is [Uhl12, Example 3.2]. In certain cases, RDAG models are equivalent to RCON models. We exactly determine the conditions under which this occurs in Theorem 10.2.8. As a consequence, we obtain an entire class of RCON models where conditions for MLE existence and uniqueness can be found by appealing to our results on RDAGs.

Finally, we draw connections to stability notions and to Gaussian group models, which are studied in Chapter 9. Namely, we extend the dictionary between ML estimation and stability notions to RDAGs in Theorem 10.6.4. This requires the extended concept of stability *under sets* from Definition 8.2.1. Furthermore, we identify RDAGs that are Gaussian group models in Proposition 10.7.3 and generalize a proof via Popov's Criterion from the TDAG setting (Theorem 9.5.9) to RDAGs that are Gaussian group models. We also obtain in the group situation a bijection between the stabilizer and the set of MLEs, Proposition 10.7.9.

While not evident in the final presentation, the "invariant theory perspective" fostered the understanding and created concepts needed to obtain many of the results. For example, trying to link RDAG models in a natural way to Gaussian models via symmetrization lead to the notion of a compatible colouring, while trying to generalize a proof via Popov's Criterion resulted in the concept of augmented sample matrices (Definition 10.3.1).

Organization and Assumptions. Section 10.1 defines RDAG models, compatible colourings and states basic properties. We compare RDAG and RCON models in Section 10.2. Afterwards, we characterize ML estimation for RDAG models in Section 10.3, which enables us in Section 10.4 to bound the ML thresholds. Section 10.5 presents some simulations. We end with connections to stability and to Gaussian group models in Sections 10.6 and 10.7, respectively.

In contrast to the paper [MRS21] we always work in parallel over $\mathbb{R}$ and $\mathbb{C}$.[1] Therefore, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and we remind the reader that $(\cdot)^\dagger$ is the Hermitian transpose, which is just the transpose $(\cdot)^\mathsf{T}$ if $\mathbb{K} = \mathbb{R}$.

[1] [MRS21] usually worked over $\mathbb{R}$, but it was noted that the results extend to the complex case [MRS21, Remark 2.11].

10.1 Introducing RDAG models

In the following we introduce restricted DAG (short: RDAG) models and give several illustrating examples. Moreover, we define the important concept of a *compatible* colouring, which is a common assumption in Chapter 10. In Lemma 10.1.8 we prove important properties of a compatible colouring which we will use throughout. As a main result, we show that an RDAG model admits a natural parametrization as a Gaussian model via symmetrization if and only if the colouring is compatible, see Proposition 10.1.9. This result is analogous to Lemma 9.5.2. We start with the definition of a coloured DAG.

Definition 10.1.1. A *coloured DAG* is a tuple $(\mathcal{G}, c)$, where $\mathcal{G} = (I, E)$ is a DAG on vertices I and directed edges E, and

$$c\colon I \cup E \to \mathscr{C}$$

is a *colouring* of the vertices and edges. Vertex $i \in I$ has colour $c(i) \in \mathscr{C}$, and edge $j \to i$ has colour $c(ij) \in \mathscr{C}$. We sometimes denote the vertex colour $c(i)$ by $c(ii)$, with no ambiguity because a DAG cannot have loops. ▲

In Definition 6.3.11 we introduced DAG models. Similarly, we can define sub-models of these by introducing symmetries among the parameters, which are given by a graph colouring.

Definition 10.1.2 ([MRS21, Definition 2.1])**.** The *restricted DAG (RDAG) model* $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ on the coloured DAG $(\mathcal{G}, c)$ is the set of concentration matrices $\Psi = (\mathrm{I}_m - \Lambda)^{\dagger}\Omega^{-1}(\mathrm{I}_m - \Lambda)$, where $\Lambda \in \mathbb{K}^{m\times m}$ satisfies

1. $\lambda_{ij} = 0$ unless $j \to i$ in $\mathcal{G}$
2. $\lambda_{ij} = \lambda_{kl}$ whenever edges $j \to i$ and $l \to k$ have the same colour

and the diagonal matrix $\Omega \in \mathrm{PD}_m(\mathbb{K})$ has positive entries and satisfies

3. $\omega_{ii} = \omega_{jj}$ if vertices i and j have the same colour.

The model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ is given by the linear structural equation $y = \Lambda y + \varepsilon$, where $y \in \mathbb{K}^m$ and $\varepsilon \sim \mathcal{N}(0, \Omega)$. By construction, $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)} \subseteq \mathcal{M}^{\rightarrow}_{\mathcal{G}}$. ▲

Remember from Remark 6.3.12 that we always assume that $i < j$ whenever $i \leftarrow j$ in $\mathcal{G}$, i.e., "parents are older than children".

Example 10.1.3 ([MRS21, Example 2.2])**.** Let $(\mathcal{G}, c)$ be $\textcircled{1} \leftarrow \boxed{3} \rightarrow \textcircled{2}$, the coloured DAG from Example 10.0.1. The RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)} \subseteq \mathrm{PD}_3(\mathbb{K})$ is parametrized by matrices

$$\Lambda = \begin{pmatrix} 0 & 0 & \lambda \\ 0 & 0 & \lambda \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad \Omega = \begin{pmatrix} \omega & 0 & 0 \\ 0 & \omega & 0 \\ 0 & 0 & \omega' \end{pmatrix}$$

where $\lambda \in \mathbb{K}$ and $\omega, \omega' \in \mathbb{R}_{>0}$. ♢

Remark 10.1.4 (based on [MRS21, Remark 2.3]). Lemma 9.5.2 shows that any DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ admits a natural set $\mathcal{A}(\mathcal{G})$ such that $\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathrm{g}}$. It is desirable to view an RDAG model $\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}$ in a natural, similar way as a Gaussian model via symmetrization. In fact, the obtained parametrization has useful consequences. First, it leads to a condition on the graph colouring, called compatibility, which is indispensable in our results of Sections 10.3 and 10.4. Second, it is helpful when comparing directed and undirected coloured models in Section 10.2. Third, it allows to generalize the connections between TDAG models and stability notions to the setting of RDAG models, see Section 10.6 and 10.7. ▽

Given a coloured DAG $(\mathcal{G}, c)$, we define the set of upper triangular matrices

$$\mathcal{A}(\mathcal{G},c) := \left\{ a \in \mathrm{GL}_m(\mathbb{K}) \;\middle|\; \begin{array}{lll} \text{(I)} & a_{ij} = 0 & \text{for } i \neq j \text{ with } j \not\to i \text{ in } \mathcal{G} \\ \text{(II)} & a_{ii} = a_{kk} & \text{if } c(i) = c(k) \\ \text{(III)} & a_{ij} = a_{kl} & \text{if } c(i \leftarrow j) = c(k \leftarrow l) \end{array} \right\}. \quad (10.1)$$

Note that $\mathcal{A}(\mathcal{G}, c)$ is contained in the set $\mathcal{A}(\mathcal{G})$ from Equation (9.17): their zero patterns agree and $\mathcal{A}(\mathcal{G}, c)$ has further equalities imposed by the colouring c.

Example 10.1.5 ([MRS21, Example 2.4])**.** For the coloured DAG $\textcircled{1} \leftarrow \boxed{3} \rightarrow \textcircled{2}$ we have

$$\mathcal{A}(\mathcal{G},c) = \left\{ \begin{pmatrix} d_1 & 0 & r \\ 0 & d_1 & r \\ 0 & 0 & d_2 \end{pmatrix} : d_1, d_2 \in \mathbb{K}^\times,\ r \in \mathbb{K} \right\}.$$

and hence

$$\mathcal{M}_{\mathcal{A}(\mathcal{G},c)}^{\mathrm{g}} = \left\{ \begin{pmatrix} |d_1|^2 & 0 & r\overline{d_1} \\ 0 & |d_1|^2 & r\overline{d_1} \\ \overline{r}d_1 & \overline{r}d_1 & 2|r|^2 + |d_2|^2 \end{pmatrix} \;\middle|\; d_1, d_2 \in \mathbb{K}^\times,\ r \in \mathbb{K} \right\}. \quad (10.2)$$

is the corresponding Gaussian model via symmetrization. ♢

For DAG models we always have $\mathcal{M}_{\mathcal{G}}^{\rightarrow} = \mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathrm{g}}$, compare Lemma 9.5.2. In contrast, the models $\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}$ and $\mathcal{M}_{\mathcal{A}(\mathcal{G},c)}^{\mathrm{g}}$ do *not* have to be equal. The following assumption on a colouring turns out to be necessary and sufficient for equality.

Definition 10.1.6 ([MRS21, Definition 2.5])**.** A colouring c of a directed graph is *compatible*, if:

(i) vertex colours and edge colours are disjoint; and

(ii) whenever edges $j \to i$ and $l \to k$ have the same colour, then the child vertices i and k have the same colour, i.e., $c(ij) = c(kl)$ implies $c(i) = c(k)$.

Note: compatibility does *not* impose equality of parent colours $c(j)$ and $c(l)$. ▲

Remark 10.1.7 (Statistical meaning of compatibility, [MRS21, Remark 2.6]).
In an RDAG model we do not impose equalities between Ω and Λ. The entry ω_{ii} is a variance, while λ_{kl} is a regression coefficient, so setting them to be equal would be difficult to interpret. Hence the vertex and edge colours can always be thought of as disjoint, as in compatibility condition (i). It ensures that Equation (10.1)

does not impose equalities between a diagonal and an off-diagonal entry. Compatibility condition (ii) has the statistical interpretation that the same regression coefficient appearing in an expression for two variables implies that their error variances agree. This extra assumption is indispensable in many of the upcoming results and proofs. It is a directed analogue to the condition appearing in [HL08, Proposition 1]. ▽

Before we relate $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ and $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$, let us prove important properties of a compatible colouring.[2] These will be frequently, often implicitly, used in the upcoming sections. To state the results, we define for a coloured DAG $(\mathcal{G}, c)$ the set of *parent relationship colours* of vertex colour s as

$$\mathrm{prc}(s) := \{c(ij) \mid \text{there exists } j \to i \text{ in } \mathcal{G} \text{ with } c(i) = s\}. \tag{10.3}$$

In words, the set $\mathrm{prc}(s)$ contains the colours of all edges that point towards some vertex of colour s.

Lemma 10.1.8. *Let $(\mathcal{G}, c)$ be a coloured DAG with compatible colouring c. Then:*

(i) We have a disjoint union $c(E) = \bigsqcup_{s\in c(I)} \mathrm{prc}(s)$. (Note that some of the $\mathrm{prc}(s)$ may be empty.)

(ii) Every matrix $a \in \mathcal{A}(\mathcal{G}, c)$ is uniquely determined by the following data: an entry $a_{s,s} \in \mathbb{K}^\times$ for each vertex colour $s \in c(I)$ and an entry $a_{s,t} \in \mathbb{K}$ for the edge colour encoded by $s \in c(I)$ and $t \in \mathrm{prc}(s)$.
Similarly, matrices Ω and Λ as in Definition 10.1.2 are uniquely determined by entries $\omega_{s,s} \in \mathbb{R}_{>0}$ and $\lambda_{s,t} \in \mathbb{K}$, respectively.

(iii) The set T of diagonal matrices in $\mathcal{A}(\mathcal{G}, c)$ is an algebraic torus and for $t \in T$, $a \in \mathcal{A}(\mathcal{G}, c)$ it holds that $ta \in \mathcal{A}(\mathcal{G}, c)$.

(iv) Let U be the set of unipotent upper triangular matrices in $\mathcal{A}(\mathcal{G}, c)$. For any $a \in \mathcal{A}(\mathcal{G}, c)$ there exist unique $t(a) \in T$ and $u(a) \in U$ with $a = t(a)u(a)$.

(v) The group $T_{\mathrm{SL}} = \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}} \cap T$ is in general just a diagonalizable group, i.e., it does not need to be connected. All other statements in (iii) and (iv) also apply to $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ using T_{SL} and $U_{\mathrm{SL}} = U$.

Proof. To prove (i), let $j \to i$ be an edge in $\mathcal{G}$. Then $c(ij) \in \mathrm{prc}(s_1)$, where $s_1 := c(i)$, and hence $c(E) = \bigcup_{s\in c(I)} \mathrm{prc}(s)$. Moreover, if $c(ij) \in \mathrm{prc}(s_2)$ then there is some $l \to k$ in $\mathcal{G}$ with $c(k) = s_2$. By Definition 10.1.6(ii), compatibility implies $s_1 = s_2$ and therefore the sets $\mathrm{prc}(s)$, $s \in c(I)$ are disjoint.

For (ii), note that by Definition 10.1.6(i) the Equation (10.1) never requires an equality of a diagonal with an off-diagonal entry of $a \in \mathcal{A}(\mathcal{G}, c)$. Therefore, a is uniquely determined by a non-zero diagonal entry for each vertex colour and an entry for each edge colour. The edge colours are in bijection with tuples (s, t) where $s \in c(I)$ and $t \in \mathrm{prc}(s)$, by part (i). This finishes the argument for matrix a and similarly one obtains the claim for Ω and Λ.

[2] These properties occur throughout [MRS21], but were not collected in a separate theorem environment.

For (iii), one directly verifies, using part (ii), that T is a group which is naturally isomorphic to the algebraic torus $(\mathbb{K}^\times)^{|c(I)|}$. Now, let $t \in T$, $a \in \mathcal{A}(\mathcal{G}, c)$ and set $b := ta$. Since $a \in \mathcal{A}(\mathcal{G})$ and multiplication with an invertible diagonal matrix preserves the support, we have $b \in \mathcal{A}(\mathcal{G})$. It remains to check $b_{ij} = b_{kl}$ whenever $c(ij) = c(kl)$. First, note that by condition (i) of compatibility there are no equalities between diagonal and off-diagonal entries of b required. Second, for two vertices i, k with $c(ii) = c(kk)$ we have $t_{ii} = t_{kk}$, $a_{ii} = a_{kk}$ and so $b_{ii} = t_{ii}a_{ii} = t_{kk}a_{kk} = b_{kk}$. Third, for edges $j \to i$ and $l \to k$ of same colour we have $a_{ij} = a_{kl}$ and condition (ii) of compatibility implies $c(ii) = c(kk)$, so $t_{ii} = t_{kk}$. Therefore, $b_{ij} = t_{ii}a_{ij} = t_{kk}a_{kl} = b_{kl}$. This proves (iii).

To show (iv), let $a \in \mathcal{A}(\mathcal{G}, c)$ and, taking part (ii) into account, define $t \in T$ via $t_{ss} := a_{ss}$ for $s \in c(I)$. By part (iii), $t^{-1} \in T$ and $u := t^{-1}a \in \mathcal{A}(\mathcal{G}, c)$. By construction, we have $a = tu$ and $u_{ss} = 1$ for all vertex colours s, so $u \in U$. This shows existence. To prove uniqueness, let $t' \in T$ and $u' \in U$ such that $a = t'u'$. As u' is unipotent we must have $(t')_{ss} = a_{ss}$ for all $s \in c(I)$, so $t = t'$. The latter implies $u' = (t')^{-1}a = t^{-1}a = u$.

For (v), consider the set $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}} = \mathcal{A}(\mathcal{G}, c) \cap \mathrm{SL}_m(\mathbb{K})$. In this situation, T_{SL} is an algebraic group that is naturally isomorphic to the diagonalizable group $\big\{(t_{ss})_s \in (\mathbb{K}^\times)^{|c(I)|} \mid \prod_s t_{ss}^{\alpha_s} = 1\big\}$, where α_s is the number of vertices of colour s. This group does not need to be connected by compare Proposition 1.1.17, as the character group is $\mathfrak{X}(T_{\mathrm{SL}}) \cong \mathbb{Z}^{|c(I)|}/\big(\mathbb{Z} \cdot (\alpha_s)_{s \in c(I)}\big)$ may have torsion elements.[3] Now, for $t \in T_{\mathrm{SL}}$ and $a \in \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$, we have $\det(ta) = 1$ and $ta \in \mathcal{A}(\mathcal{G}, c)$ by part (iii) for $\mathcal{A}(\mathcal{G}, c)$. Thus, $ta \in \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$. Furthermore, any $a \in \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ has a unique decomposition $a = t(a)u(a)$ in $\mathcal{A}(\mathcal{G}, c)$ by part (iv). We have $\det(u(a)) = 1$ as $u(a)$ is unipotent, and thus $\det(a) = 1$ yields $\det(t(a)) = 1$ as well. We deduce that the unique decomposition $a = t(a)u(a)$ lives in $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$. □

A main feature of compatibility is relating the models $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ and $\mathcal{M}^{\mathfrak{g}}_{\mathcal{A}(\mathcal{G},c)}$.[4]

Proposition 10.1.9 ([MRS21, Proposition 2.7])**.** *Fix a coloured DAG $(\mathcal{G}, c)$. The RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ is equal to $\mathcal{M}^{\mathfrak{g}}_{\mathcal{A}(\mathcal{G},c)}$ if and only if colouring c is compatible.*

Remark 10.1.10. A usual DAG model $\mathcal{M}^{\rightarrow}_{\mathcal{G}}$ on $\mathcal{G}$ is an RDAG model with compatible colouring, as follows. Let c be a colouring that assigns to each vertex and to each edge a distinct colour. Then c is compatible, $\mathcal{M}^{\rightarrow}_{\mathcal{G}} = \mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ and $\mathcal{A}(\mathcal{G}) = \mathcal{A}(\mathcal{G}, c)$. In this regard, Proposition 10.1.9 generalizes Lemma 9.5.2. ▽

To prove the proposition, it is instructive to think of $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ and $\mathcal{M}^{\mathfrak{g}}_{\mathcal{A}(\mathcal{G},c)}$ imposing zero patterns and symmetries on certain matrix decompositions.

Recall that the Cholesky decomposition of $\Psi \in \mathrm{PD}_m(\mathbb{K})$ is given by the *unique* upper triangular matrix $a := \mathrm{chol}(\Psi) \in \mathbb{K}^{m \times m}$ with *real-valued, positive* diagonal entries such that $\Psi = a^\dagger a$. The model $\mathcal{M}^{\mathfrak{g}}_{\mathcal{A}(\mathcal{G},c)}$ imposes zeros and symmetries in the Cholesky decomposition, as follows.

[3] $\mathfrak{X}(T_{\mathrm{SL}})$ has torsion if and only if the greatest common divisor of all α_s equals one.

[4] Trying to relate these models was actually how the authors of [MRS21] came up with the concept of a compatible colouring.

Lemma 10.1.11 ([MRS21, Lemma 2.8]). *Fix a coloured DAG $(\mathcal{G}, c)$ with compatible colouring c. Then $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ is the set of positive definite matrices with Cholesky decomposition $a^\dagger a$ for some $a \in \mathcal{A}(\mathcal{G}, c)$.*

Proof. The set $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ consists of all positive definite matrices of the form $a^\dagger a$ for some $a \in \mathcal{A}(\mathcal{G}, c)$, see Equation (8.1), and all matrices in $\mathcal{A}(\mathcal{G}, c)$ are upper triangular by our assumption on the ordering of the vertices.

It remains to show that for any $\Psi = b^\dagger b$, where $b \in \mathcal{A}(\mathcal{G}, c)$, its Cholesky decomposition lies in $\mathcal{A}(\mathcal{G}, c)$. For $i \in [m]$, set $t_{ii} := \overline{b_{ii}}|b_{ii}|^{-1}$. This defines a diagonal matrix t such that $t^\dagger t = \mathrm{I}_m$ and $t \in \mathcal{A}(\mathcal{G}, c)$ as $b \in \mathcal{A}(\mathcal{G}, c)$. Thus, $a := tb \in \mathcal{A}(\mathcal{G}, c)$ using Lemma 10.1.8(iii). By construction, a has positive diagonal entries $a_{ii} = |b_{ii}|$ and hence $a^\dagger a$ is the Cholesky decomposition of Ψ. □

The *LDL decomposition* writes a positive definite matrix $\Psi \in \mathrm{PD}_m(\mathbb{K})$ as $LDL^\dagger$, where D is diagonal with positive entries, and $L \in \mathbb{K}^{m\times m}$ is lower triangular and unipotent (i.e., its diagonal entries are equal to one). With these properties L and D are uniquely determined. The LDL decomposition is closely related to the factorization $\Psi = (\mathrm{I}_m - \Lambda)^\dagger \Omega^{-1}(\mathrm{I}_m - \Lambda)$ from Equation (6.14): the LDL decomposition is $D = \Omega^{-1}$ and $L = (\mathrm{I}_m - \Lambda)^\dagger$. Hence, an RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ imposes zeros and symmetries in the LDL decomposition. The LDL and Cholesky decompositions are are related by:

$$\begin{aligned} &\text{Cholesky from LDL:} && a := \mathrm{chol}(\Psi) = \Omega^{-1/2}(\mathrm{I}_m - \Lambda), \\ &\text{LDL from Cholesky:} && \Omega = \mathrm{diag}(a_{11}^{-2}, \ldots, a_{mm}^{-2}), \quad \Lambda = \mathrm{I}_m - \mathrm{diag}(a_{11}^{-1}, \ldots, a_{mm}^{-1})a \end{aligned}$$

For DAG models, we have shown $\mathcal{M}^{\rightarrow}_{\mathcal{G}} = \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G})}$, Lemma 9.5.2, by comparing the support conditions in the two decompositions. Similarly, we prove Proposition 10.1.9 for RDAG models by comparing zero patterns and symmetries in the LDL and Cholesky decomposition. For this, Lemma 10.1.8(iii) is the crucial property of a compatible colouring.

Proof of Proposition 10.1.9. Let colouring c be compatible. Recall, that condition (i) of compatibility implies that Equation (10.1) does not impose equalities between a diagonal and an off-diagonal entry of $a \in \mathcal{A}(\mathcal{G}, c)$.

First, let $\Psi = (\mathrm{I}_m - \Lambda)^\dagger \Omega^{-1}(\mathrm{I}_m - \Lambda) \in \mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ as in Definition 10.1.2. The colour conditions on $\Omega \in \mathrm{PD}_m(\mathbb{K})$ show that the diagonal matrix $t := \Omega^{-1/2}$ is in $\mathcal{A}(\mathcal{G}, c)$. Moreover, $\mathrm{I}_m - \Lambda \in \mathcal{A}(\mathcal{G}, c)$ as follows. First, it is unipotent upper triangular, as Λ is strictly upper triangular. In particular, the vertex colour conditions are fulfilled as all diagonal entries are equal to one. Second, the support and colour conditions on Λ imply that the off-diagonal entries of $\mathrm{I}_m - \Lambda$ satisfy the corresponding conditions for $\mathcal{A}(\mathcal{G}, c)$. By Lemma 10.1.8(iii), $a := \Omega^{-1/2}(\mathrm{I}_m - \Lambda) \in \mathcal{A}(\mathcal{G}, c)$ and hence $\Psi = a^\dagger a \in \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$.

Conversely, let $\Psi \in \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$. Then the Cholesky decomposition is $\Psi = a^\dagger a$ for $a \in \mathcal{A}(\mathcal{G}, c)$, by Lemma 10.1.11. Since a has positive diagonal entries and $a \in \mathcal{A}(\mathcal{G}, c)$, $\omega_{ii} := a_{ii}^{-2}$ defines a diagonal $\Omega \in \mathrm{PD}_m(\mathbb{K})$ satisfying the colour symmetries. Moreover, $u := \mathrm{diag}(a_{11}^{-1}, \ldots, a_{mm}^{-1})a$ is, by construction, unipotent upper triangular and, by Lemma 10.1.8(iii), $u \in \mathcal{A}(\mathcal{G}, c)$. Therefore, $\Lambda = (\mathrm{I}_m - u)$

is strictly upper triangular and satisfies the support and colour conditions from Definition 10.1.2. This shows $\Psi \in \mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$. Altogether, a compatible colouring implies $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)} = \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$.

Now, assume the colouring is not compatible. We will exhibit some $\Psi \in \mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$, in terms of Ω and Λ, such that $\Psi \notin \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$. For this, let $\Psi = a^{\dagger}a$ be the Cholesky decomposition, i.e.,

$$a_{ij} = \begin{cases} \omega_{ii}^{-1/2} & \text{if } i = j \\ -\omega_{ii}^{-1/2}\lambda_{ij} & \text{if } i \neq j. \end{cases} \tag{10.4}$$

If $\Psi = b^{\dagger}b$ for some $b \in \mathcal{A}(\mathcal{G}, c)$ then, similar to the proof of Lemma 10.1.11, there is some diagonal matrix t with $tb = a$ and $|t_{ii}| = 1$ for all $i \in [m]$.[5] In particular, $|b_{ij}| = |a_{ij}|$ for all $i, j \in [m]$.

First, if Definition 10.1.6(i) does not hold, then there is a vertex $k \in [m]$ and an edge $j \to i$ with $c(kk) = c(ij)$. The RDAG model imposes no relation between ω_{kk} and λ_{ij}, so let Ψ be given by some Ω and Λ with $\omega_{kk} = 1$ and $\lambda_{ij} = 0$. Then $|a_{kk}| = 1$ and $|a_{ij}| = 0$, by (10.4). Hence, $\Psi \notin \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ as otherwise $|b_{kk}| = |a_{kk}| = 1 \neq 0 = |a_{ij}| = |b_{ij}|$ violates the colour conditions for $\mathcal{A}(\mathcal{G}, c)$.

Second, if Definition 10.1.6(ii) does not hold, then there exist edges $j \to i$ and $l \to k$ with $c(ij) = c(kl)$ but $c(i) \neq c(k)$. We choose Ψ given by some Ω and Λ with $\omega_{ii} = 1$, $\omega_{kk} = \frac{1}{4}$ and $\lambda_{ij} = \lambda_{kl} = 1$. Then $|a_{ij}| = 1$ and $|a_{kl}| = 2$, by (10.4). Again, we must have $\Psi \notin \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ as otherwise $|b_{ij}| = |a_{ij}| = 1 \neq 2 = |a_{kl}| = |b_{kl}|$ would violate the colour conditions for $\mathcal{A}(\mathcal{G}, c)$. □

Example 10.1.12 ([MRS21, Example 2.10])**.** We return to the coloured DAG ① ← 3 → ② from Examples 10.1.3 and 10.1.5. The colouring is compatible, because the sets of vertex and edge colours are disjoint, and the children of both red edges have the same colour. Hence, Proposition 10.1.9 shows that $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ from Equation (10.2) is equal to $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$. ♢

10.2 Comparison of RDAG and RCON models

In this section we compare RDAG models to their undirected analogue: restricted concentration (RCON) models which were introduced in [HL08]. Similar to RDAG models, RCON models are sub-models of undirected Gaussian graphical (CON) models, see Example 6.3.10, and impose symmetries on concentration matrices according to a graph colouring. In Theorem 10.2.8 we precisely characterize when an RDAG model equals its induced RCON model. To prove this theorem, we need the similar statement for DAG models and CON models, Theorem 6.3.14. It is well-known in the literature, see [AMP97, Theorem 3.1] or [Fry90, Theorem 5.6]. Still, it is instructive to start with a proof of Theorem 6.3.14, since the presented method generalizes to give a proof of Theorem 10.2.8.

[5] However, we cannot deduce $a \in \mathcal{A}(\mathcal{G}, c)$, because compatibility is needed for Lemma 10.1.8(iii).

Given a DAG $\mathcal{G}$, remember that $\mathcal{G}^u$ denotes the corresponding undirected graph, which is obtained by forgetting the direction of each edge in $\mathcal{G}$. For convenience, we restate Theorem 6.3.14.

Theorem 10.2.1 (Theorem 6.3.14 restated)**.** *Let $\mathcal{G}$ be a DAG. The DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is equal to the undirected Gaussian graphical model $\mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$ on $\mathcal{G}^u$ if and only if $\mathcal{G}$ has no unshielded colliders.*

We prove Theorem 10.2.1 via two propositions. Note that these propositions and their proofs only appear in the *first* arXiv version of [MRS21], e.g., the following is Proposition 3.8 in the first arXiv version.

Proposition 10.2.2. *Let $\mathcal{G}$ be a DAG. Then $\mathcal{M}_{\mathcal{G}}^{\rightarrow} \subseteq \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$ if and only if $\mathcal{G}$ has no unshielded colliders.*

Proof. The DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ equals $\mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathbf{g}}$, by Lemma 9.5.2. Assume $\mathcal{G}$ has an unshielded collider $i \longrightarrow k \longleftarrow j$. In particular, $\mathcal{G}^u$ has no edge between i and j, so $\Psi_{ij} = \Psi_{ji} = 0$ for all $\Psi \in \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$. Let $a \in \mathcal{A}(\mathcal{G})$ be given by $a_{ki} = a_{kj} = 1$, $a_{ll} = 1$ for all $l \in [m]$ and all other entries zero. Then $(a^\dagger a)_{ij} = \overline{a_{ki}} a_{kj} = 1 \neq 0$ and hence $a^\dagger a \notin \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$.

Conversely, if $\mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathbf{g}} = \mathcal{M}_{\mathcal{G}}^{\rightarrow} \not\subseteq \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$ then there is $a \in \mathcal{A}(\mathcal{G})$ with $a^\dagger a \notin \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$. Thus, $a^\dagger a$ violates the off-diagonal zero pattern of $\mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$, i.e., there is a pair of indices $i \neq j$ such that there is no edge between i and j in $\mathcal{G}^u$ but $(a^\dagger a)_{ij} \neq 0$. Since $(a^\dagger a)_{ij} = \sum_{k=1}^{m} \overline{a_{ki}} a_{kj}$, some product $\overline{a_{ki}} a_{kj}$ must be non-zero, i.e., there must exist edges $i \longrightarrow k \longleftarrow j$ in $\mathcal{G}$. This is an unshielded collider, because $\mathcal{G}^u$ (and hence $\mathcal{G}$) has no edge between i and j. □

The following is Proposition 3.9 in the *first* arXiv version of [MRS21].[6]

Proposition 10.2.3. *If a DAG $\mathcal{G}$ has no unshielded colliders, then $\mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}} \subseteq \mathcal{M}_{\mathcal{G}}^{\rightarrow}$.*

Proof. Given a concentration matrix $\Psi \in \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$, we show $\Psi \in \mathcal{M}_{\mathcal{A}(\mathcal{G})}^{\mathbf{g}}$ by proving that its Cholesky decomposition $a := \mathrm{chol}(\Psi)$ satisfies $a \in \mathcal{A}(\mathcal{G})$. The entries of the upper triangular matrix a are determined, for $l \in [m]$ and $i < j \leq m$, by

$$a_{l,l} = \left(\Psi_{l,l} - \sum_{k=1}^{l-1} |a_{k,l}|^2 \right)^{1/2} \quad \text{and} \quad a_{i,j} = \left(\Psi_{i,j} - \sum_{k=1}^{i-1} \overline{a_{k,i}} a_{k,j} \right) a_{i,i}^{-1}, \tag{10.5}$$

see [TB97, Lecture 23].[7] Note that the expression under the square root in (10.5) is indeed a positive real number, compare [TB97, Lecture 23].

We have to ensure the support conditions of $\mathcal{A}(\mathcal{G})$ for the off-diagonal entries of a. For this, let $i, j \in [m]$, $i < j$ such that $j \not\to i$ in $\mathcal{G}$. Then $\mathcal{G}^u$ has no edge between j and i. Therefore, $\Psi_{i,j} = 0$ using that $\Psi \in \mathcal{M}_{\mathcal{G}^u}^{\mathrm{ud}}$. Moreover, we claim that $\overline{a_{k,i}} a_{k,j} = 0$ holds for all $k \in [i-1]$. Indeed, otherwise $i \to k \leftarrow j$ would be an unshielded collider in $\mathcal{G}$, which contradicts the assumption. Altogether, we deduce $a_{i,j} = 0$ using (10.5). This proves $a \in \mathcal{A}(\mathcal{G})$ as desired. □

[6] The proof has been simplified by taking Equation (10.5) for granted.

[7] Equation (10.5) yields an iterative algorithm to compute the Cholesky decomposition a.

Combining Propositions 10.2.2 and 10.2.3 proves Theorem 10.2.1.

Proof of Theorem 10.2.1. An unshielded collider in $\mathcal{G}$ implies $\mathcal{M}^{\rightarrow}_{\mathcal{G}} \not\subseteq \mathcal{M}^{\mathrm{ud}}_{\mathcal{G}^u}$, by Proposition 10.2.2, and hence prevents equality of the models. The absence of unshielded colliders implies $\mathcal{M}^{\rightarrow}_{\mathcal{G}} \subseteq \mathcal{M}^{\mathrm{ud}}_{\mathcal{G}^u}$ (Proposition 10.2.2) and $\mathcal{M}^{\mathrm{ud}}_{\mathcal{G}^u} \subseteq \mathcal{M}^{\rightarrow}_{\mathcal{G}}$ (Proposition 10.2.3). □

Next, we define RCON models as in [HL08]. For this, a coloured undirected graph is a tuple $(\mathcal{G}, c)$, where $\mathcal{G} = (I, E)$ is an undirected graph and the map

$$c : I \cup E \to \mathscr{C}$$

assigns a colour to each vertex and to each edge.

Definition 10.2.4 (see [HL08, §3])**.** The *RCON model* $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G},c)}$ on the coloured undirected graph $(\mathcal{G}, c)$ consists of concentration matrices $\Psi \in \mathrm{PD}_m(\mathbb{K})$ with

(i) $\Psi_{ij} = \Psi_{ji} = 0$ whenever $i \text{ — } j$ is *not* an edge in $\mathcal{G}$

(ii) $\Psi_{ii} = \Psi_{jj}$ whenever $c(i) = c(j)$,

(iii) $\Psi_{ij} = \Psi_{kl}$ whenever $i < j$ and $k < l$ such that $c(i \text{ — } j) = c(k \text{ — } l)$. Note that this implies $\Psi_{ji} = \overline{\Psi_{ij}} = \overline{\Psi_{kl}} = \Psi_{lk}$ since $\Psi^{\dagger} = \Psi$.

By part (i), $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G},c)}$ is a sub-model of the model $\mathcal{M}^{\mathrm{ud}}_{\mathcal{G}}$ from Example 6.3.10. ▲

Let $(\mathcal{G}, c)$ be a coloured DAG. Similarly to the construction of $\mathcal{G}^u$, we obtain a coloured undirected graph $(\mathcal{G}^u, c)$ by forgetting the edge directions in $\mathcal{G}$. All vertex and edge colours are inherited. We call $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$ the RCON model induced by the RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$. Let us compare RDAG models and their induced RCON models in two examples.

Example 10.2.5 (RDAG = RCON, [MRS21, Example 3.1])**.** We revisit our running example $\textcircled{1} \leftarrow \boxed{3} \rightarrow \textcircled{2}$. The corresponding RCON model has coloured undirected graph $\textcircled{1} \text{ — } \boxed{3} \text{ — } \textcircled{2}$, with blue (circular) vertices 1 and 2, black (square) vertex 3, and red edges. By Definition 10.2.4, the RCON model is the set of positive definite matrices of the form

$$\Psi = \begin{pmatrix} \delta_1 & 0 & \varrho \\ 0 & \delta_1 & \varrho \\ \overline{\varrho} & \overline{\varrho} & \delta_2 \end{pmatrix}, \quad \text{where } \varrho \in \mathbb{K} \text{ and } \delta_1, \delta_2 \in \mathbb{R}_{>0}.$$

Since the colouring is compatible, the RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ is equal to $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ from Equation (10.2). Any matrix in $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ satisfies the equalities for the RCON model, so $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)} \subseteq \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$. Conversely, given positive-definite $\Psi \in \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$,

$$\det(\Psi) = \delta_1^2\big(\delta_2 - 2|\varrho|^2\delta_1^{-1}\big) > 0 \quad \text{ and hence } \quad \delta_2 - 2|\varrho|^2\delta_1^{-1} > 0.$$

Setting $d_1 := \sqrt{\delta_1} \in \mathbb{R}_{>0}$, $d_2 := \sqrt{\delta_2 - 2|\varrho|^2\delta_1^{-1}} \in \mathbb{R}_{>0}$ and $r := \varrho/d_1 \in \mathbb{K}$ shows that Ψ is of the form in Equation (10.2), i.e., $\Psi \in \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$. ♢

Example 10.2.6 (RDAG $\neq$ RCON, [MRS21, Example 3.2])**.** Consider the RDAG model on ① $\leftarrow$ ②, the graph with two blue (circular) vertices and a red edge. The colouring is compatible, so by Proposition 10.1.9 the RDAG model is $\mathcal{M}^{\mathbf{g}}_{\mathcal{A}(\mathcal{G},c)}$, where

$$\mathcal{A}(\mathcal{G},c) = \left\{ \begin{pmatrix} d & r \\ 0 & d \end{pmatrix} \;\middle|\; d \in \mathbb{K}^{\times},\, r \in \mathbb{K} \right\}.$$

The induced RCON model is given by ① — ② and consists of all $\Psi \in \mathrm{PD}_2(\mathbb{K})$ with $\Psi_{11} = \Psi_{22}$ and $\Psi_{12} = \Psi_{21}$, by Definition 10.2.4. Neither model is contained in the other: the RCON model contains

$$\Psi' := \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{3} \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & \sqrt{3} \end{pmatrix},$$

but the diagonal entries 2 and $\sqrt{3}$ in the Cholesky decomposition do not satisfy the condition $a_{11} = a_{22}$ for $a \in \mathcal{A}(\mathcal{G},c)$. Therefore, $\Psi' \notin \mathcal{M}^{\mathbf{g}}_{\mathcal{A}(\mathcal{G},c)}$ by Lemma 10.1.11. Conversely, the matrix

$$\Psi'' := \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$$

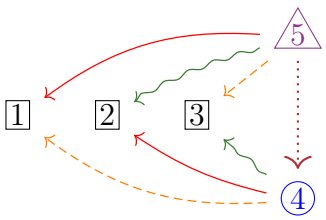
is in the RDAG model, but not the RCON model, since $\Psi''_{11} \neq \Psi''_{22}$. ♢

To characterize when an RDAG model is equal to its corresponding RCON model, we give two constructions of coloured graphs, one that is built from a vertex of a coloured DAG $(\mathcal{G},c)$ and the other from an edge.

Fix a vertex $i \in I$. Recall that $\mathrm{ch}(i)$ is the set of children of i. Consider the subgraph on vertex set $\{i\} \cup \mathrm{ch}(i)$ with edges $i \to k$ for each $k \in \mathrm{ch}(i)$, and colours inherited from $(\mathcal{G},c)$. We denote the coloured subgraph by $\mathcal{G}_i$.

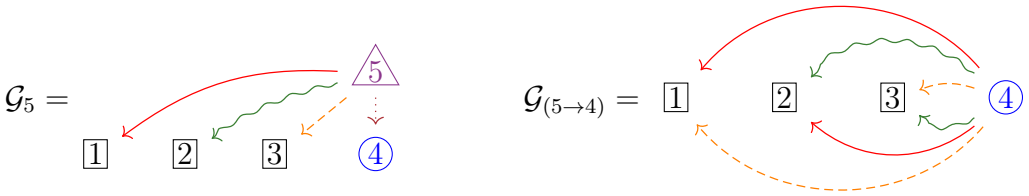
Now, fix an edge $(j \to i)$ in $\mathcal{G}$. Consider the set $\{i\} \cup (\mathrm{ch}(i) \cap \mathrm{ch}(j))$ of vertices with vertex colours inherited from $(\mathcal{G},c)$. For each $k \in \mathrm{ch}(i) \cap \mathrm{ch}(j)$, we introduce two edges $i \to k$, one with colour $c(ki)$ and the other with colour $c(kj)$. We denote this coloured multi-digraph by $\mathcal{G}_{(j \to i)}$. Note that $\mathcal{G}_{(j \to i)}$ only contains the coloured vertex i if vertices i and j do not have common children.

Example 10.2.7 ([MRS21, Example 3.3])**.** Consider the coloured DAG[8]



The vertex construction at vertex 5 and edge construction at edge $5 \to 4$ are:

[8] with three vertex colours (blue/circular, black/square, and purple/triangular) and four edge colours (red/solid, green/squiggly, orange/dashed, and brown/dotted)



♢

Two coloured graphs $(\mathcal{G}, c)$ and $(\mathcal{G}', c')$ are *isomorphic* if the coloured graphs are the same up to relabelling vertices. We denote an isomorphism by $\mathcal{G} \simeq \mathcal{G}'$ when the colouring is clear. Now, we formulate the main theorem of this section.

Theorem 10.2.8 ([MRS21, Theorem 3.4])**.** *Let $(\mathcal{G}, c)$ be a coloured DAG where colouring c is compatible. The RDAG model $\mathcal{M}^{\to}_{(\mathcal{G},c)}$ on $(\mathcal{G}, c)$ is equal to the RCON model $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$ on $(\mathcal{G}^u, c)$ if and only if:*

(a) $\mathcal{G}$ has no unshielded colliders;

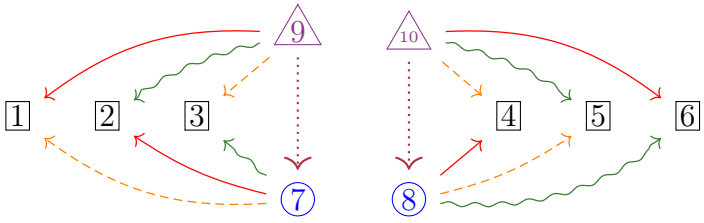
(b) $\mathcal{G}_i \simeq \mathcal{G}_j$ for every pair of vertices i, j of the same colour; and

(c) $\mathcal{G}_{(j\to i)} \simeq \mathcal{G}_{(l\to k)}$ for every pair of edges $j \to i$ and $l \to k$ in $\mathcal{G}$ of same colour.

Before we prove the theorem, we illustrate it in two examples.

Example 10.2.9 ([MRS21, Example 3.5])**.** Our running example 1 ← 3 → 2 satisfies the conditions of Theorem 10.2.8: it has no unshielded colliders and the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ both consist of a single blue vertex. Moreover, $\mathcal{G}_{(3\to 1)}$ and $\mathcal{G}_{(3\to 2)}$ only consist of a blue vertex as 1 and 3 (respectively 2 and 3) do not have common children. The RDAG and RCON models are therefore equal, as we saw in Example 10.2.5. ♢

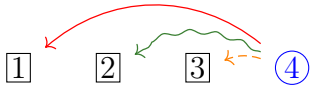
Example 10.2.10 ([MRS21, Example 3.6])**.** The coloured DAG $(\mathcal{G}, c)$ given by



also satisfies the conditions of Theorem 10.2.8:

(a) It has no unshielded colliders.

(b) For the black (square) vertices, the graphs $\mathcal{G}_i$ consist of one black vertex. For the blue (circular) vertices, the $\mathcal{G}_i$ are isomorphic to



The purple (triangular) vertices have $\mathcal{G}_i$ isomorphic to $\mathcal{G}_5$ from Example 10.2.7.

(c) All edges $j \to i$ have $\mathrm{ch}(j) \cap \mathrm{ch}(i) = \emptyset$, except for the two brown edges. For these, $\mathcal{G}_{(10\to 8)}$ and $\mathcal{G}_{(9\to 7)}$ are both isomorphic to $\mathcal{G}_{(5\to 4)}$ from Example 10.2.7.

Hence, the RDAG model $\mathcal{M}^{\to}_{(\mathcal{G},c)}$ equals the induced RCON model $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$. Note that the two connected components of $(\mathcal{G}, c)$ are not isomorphic as coloured directed graphs. We will see why this is not required for the proof of Theorem 10.2.8, i.e., why we can collapse vertices i and j in the definition of $\mathcal{G}_{(j\to i)}$. ♢

Finally, we prove Theorem 10.2.8 in a similar way as Theorem 10.2.1.

Proposition 10.2.11 ([MRS21, Proposition 3.8])**.** *Let $(\mathcal{G}, c)$ be a coloured DAG with compatible colouring c. Then $\mathcal{M}^{\to}_{(\mathcal{G},c)} \subseteq \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$ if and only if conditions (a), (b) and (c) of Theorem 10.2.8 hold.*

Proof. We have $\mathcal{M}^{\to}_{(\mathcal{G},c)} = \mathcal{M}^{\mathbf{g}}_{\mathcal{A}(\mathcal{G},c)}$ since the colouring is compatible, see Proposition 10.1.9. Now, $\mathcal{M}^{\to}_{(\mathcal{G},c)} \subseteq \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$ translates to: for all $a \in \mathcal{A}(\mathcal{G}, c)$ it holds that $a^\dagger a$ lies in the RCON model $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$. Recall that the Definition 10.2.4 of an RCON model involves three conditions: a support condition (i), a vertex colour condition (ii), and an edge colour condition (iii). We show that

$$\text{condition 10.2.8(a) holds} \quad \Leftrightarrow \quad \forall\, a \in \mathcal{A}(\mathcal{G}, c)\colon\ a^\dagger a \ \text{ satisfies 10.2.4(i)}, \tag{10.6}$$

and the analogous equivalences of (10.6) for 10.2.8(b) and 10.2.4(ii), as well as for 10.2.8(c) and 10.2.4(iii). Altogether, we obtain the statement.

First, for the support condition (i), note that Proposition 10.2.2 says that

$$\text{condition 10.2.8(a) holds} \quad \Leftrightarrow \quad \forall\, a \in \mathcal{A}(\mathcal{G})\colon\ a^\dagger a \ \text{ satisfies 10.2.4(i)},$$

where we stress that we have $\mathcal{A}(\mathcal{G})$ (instead of $\mathcal{A}(\mathcal{G}, c)$) on the right-hand side. In particular, condition (a) implies that for all $a \in \mathcal{A}(\mathcal{G}, c)$, $a^\dagger a$ satisfies 10.2.4(i). Conversely, assume condition (a) does not hold, i.e., $\mathcal{G}$ has an unshielded collider $i \longrightarrow k \longleftarrow j$. Thus, $\mathcal{G}^u$ has no edge between i and j, so $\Psi_{ij} = \Psi_{ji} = 0$ for all $\Psi \in \mathcal{M}^{\mathrm{ud}}_{\mathcal{G}^u}$. Let $a \in \mathcal{A}(\mathcal{G}, c)$ be given by: $a_{ll} = 1$ for all $l \in [m]$; $a_{ki} = a_{kj} = 1$; $a_{pq} = 1$ whenever $c(pq) = c(ki)$ or $c(pq) = c(kj)$;[9] and all other entries are zero. Then

$$(a^\dagger a)_{ij} = \sum_{l=1}^{m} \overline{a_{li}} a_{lj} \geq \overline{a_{ki}} a_{kj} = 1 \neq 0,$$

where we used $\overline{a_{li}} a_{lj} \in \{0, 1\}$ in the inequality. The above shows $a^\dagger a$ does *not* satisfy 10.2.4(i). Altogether, we proved (10.6).

Second, the vertex colour condition (ii) on $a^\dagger a$ with $a \in \mathcal{A}(\mathcal{G}, c)$, translates to

$$|a_{ii}|^2 + \sum_{k\in \mathrm{ch}(i)} |a_{ki}|^2 = |a_{jj}|^2 + \sum_{l\in \mathrm{ch}(j)} |a_{lj}|^2 \qquad \text{whenever} \quad c(i) = c(j), \tag{10.7}$$

[9] In comparison with the proof of Proposition 10.2.2, this has to be added to ensure edge colour symmetries, i.e., (10.1)(III).

where we used (10.1)(I) to obtain (10.7). Consider vertices i and j with $c(i) = c(j)$. If condition (b) holds, then $\mathcal{G}_i \simeq \mathcal{G}_j$as coloured directed graphs. The latter implies that (10.7) holds for all $a \in \mathcal{A}(\mathcal{G}, c)$, by definition of $\mathcal{G}_i$ and $\mathcal{G}_j$. Conversely, assume (10.7) holds for all $a \in \mathcal{A}(\mathcal{G}, c)$. The equation over $\mathbb{K} = \mathbb{C}$ implies the equation over $\mathbb{K} = \mathbb{R}$, so it suffices to assume the latter. Over $\mathbb{R}$, (10.7) is a *polynomial identity* in the entries of $a \in \mathcal{A}(\mathcal{G}, c)$. Note that we have $a_{ii} = a_{jj}$ as $a \in \mathcal{A}(\mathcal{G}, c)$. Thus, the sums in (10.7) are equal for all $a \in \mathcal{A}(\mathcal{G}, c)$ only if $|\operatorname{ch}(i)| = |\operatorname{ch}(j)|$ and the edge colours in $\mathcal{G}_i$ and $\mathcal{G}_j$ agree (counted with multiplicity).[10] By compatibility, the corresponding child vertex colours in $\mathcal{G}_i$ and $\mathcal{G}_j$ also agree, hence we have $\mathcal{G}_i \simeq \mathcal{G}_j$. This proves (10.6) for (b) and (ii).

Third, the edge colour condition (iii) on $a^\dagger a$ with $a \in \mathcal{A}(\mathcal{G}, c)$, translates to

$$\overline{a_{ii}}a_{ij} + \sum_{p\neq i,j}^{m} \overline{a_{pi}}a_{pj} = \overline{a_{kk}}a_{kl} + \sum_{q\neq k,l}^{m} \overline{a_{qk}}a_{ql} \qquad \text{whenever} \quad c(ij)=c(kl), \tag{10.8}$$

where we used (10.1)(I) to get (10.8). Let $j \to i$ and $l \to k$ be edges in $\mathcal{G}$ of same colour.[11] Now, if condition (c) holds, then $\mathcal{G}_{(j\to i)} \simeq \mathcal{G}_{(l\to k)}$ as coloured multidigraphs. The latter implies that (10.8) holds for all $a \in \mathcal{A}(\mathcal{G}, c)$, by definition of $\mathcal{G}_{(j\to i)}$ and $\mathcal{G}_{(l\to k)}$. Conversely, assume (10.8) holds for all $a \in \mathcal{A}(\mathcal{G}, c)$. Again, it suffices to assume $\mathbb{K} = \mathbb{R}$. Then (10.8) is a polynomial identity in the entries of a. Note that the compatibility of the colouring gives $a_{ii} = a_{kk}$, hence $a_{ii}a_{ij} = a_{kk}a_{kl}$, and that the other summands in (10.8) vanish unless $p \in \operatorname{ch}(i) \cap \operatorname{ch}(j)$, respectively $q \in \operatorname{ch}(k) \cap \operatorname{ch}(l)$. Hence, the sums are equal for all $a \in \mathcal{A}(\mathcal{G}, c)$ only if $|\operatorname{ch}(i)\cap\operatorname{ch}(j)| = |\operatorname{ch}(k)\cap\operatorname{ch}(l)|$ and the graphs $\mathcal{G}_{(j\to i)}$ and $\mathcal{G}_{(l\to k)}$ are isomorphic on their edge colours. By compatibility, the corresponding child vertex colours must also agree and hence $\mathcal{G}_{(j\to i)} \simeq \mathcal{G}_{(l\to k)}$. This proves (10.6) for (c) and (iii). □

Proposition 10.2.12 ([MRS21, Proposition 3.9])**.** *Let $(\mathcal{G}, c)$ be a coloured DAG with compatible colouring c such that conditions (a), (b) and (c) of Theorem 10.2.8 hold. Then $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)} \subseteq \mathcal{M}^{\to}_{(\mathcal{G},c)}$.*

Proof. We have $\mathcal{M}^{\to}_{(\mathcal{G},c)} = \mathcal{M}^{\mathbf{g}}_{\mathcal{A}(\mathcal{G},c)}$ as colouring c is compatible, see Proposition 10.1.9. Given some $\Psi \in \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$, we show that its unique Cholesky decomposition $a := \operatorname{chol}(\Psi)$ satisfies $a \in \mathcal{A}(\mathcal{G}, c)$. Since $\mathcal{G}$ has no unshielded colliders by condition (a), the proof of Proposition 10.2.3 shows $a \in \mathcal{A}(\mathcal{G})$. Therefore, Equation (10.5) implies that for any vertex l and any edge $i \leftarrow j$ we have

$$a_{l,l} = \Big(\Psi_{l,l} - \sum_{p\in\operatorname{ch}(l)} |a_{p,l}|^2\Big)^{1/2} \tag{10.9}$$

$$a_{i,j} = \Big(\Psi_{i,j} - \sum_{p\in\operatorname{ch}(i)\cap\operatorname{ch}(j)} \overline{a_{p,i}}a_{p,j}\Big)a_{i,i}^{-1}. \tag{10.10}$$

[10] Think of the entries of a as indeterminates.

[11] Note that in (10.8) the terms $a_{ji}a_{jj}$ and $a_{lk}a_{ll}$ do not appear, since the acyclicity ensures $i \not\to j$ and $k \not\to l$ in $\mathcal{G}$. In particular, it does not matter whether $c(j) = c(l)$ holds or not. This explains why the construction of $\mathcal{G}_{(j\to i)}$ does not take vertex j and its colour into account.

We show that a satisfies the symmetries of the colouring. We prove this inductively over the top left $k \times k$ blocks of a. If $k = 1$ there are no symmetries to check. We assume that the top left $k \times k$ submatrix of a satisfies the symmetries. For the induction step, we compare $a_{1,k+1}, a_{2,k+1}, \ldots, a_{k+1,k+1}$ with each other and with $a_{i,j}$, where $i, j \in [k]$.

If there is an edge $(k+1) \to 1$ with same colour as $j \to i$ for $i, j \in [k]$, we need to show that $a_{1,k+1} = a_{i,j}$. First, $a_{11} = a_{ii}$ by compatibility. Second, $\Psi_{i,j} = \Psi_{1,k+1}$ since $i < j$, $1 < k+1$ and $\Psi \in \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$, compare Definition 10.1.2(iii). Third, all $a_{p,q}$ for $p, q \in [k]$ respect the symmetries by induction hypothesis. Therefore, $\mathcal{G}_{(j\to i)} \simeq \mathcal{G}_{(k+1\to 1)}$ as coloured multi-digraphs, condition (c), ensures that the sum over the common children of i and j in (10.10) equals the respective sum over the common children of 1 and $k+1$ in (10.10). Altogether, we deduce that the expressions (10.10) for $a_{i,j}$ and $a_{1,k+1}$ are equal.

Proceeding inductively, we show analogously that all entries $a_{2,k+1}, \ldots, a_{k,k+1}$ respect the symmetries of colouring c. Indeed, for $a_{i',k+1}$ with $i' \in \{2, \ldots, k\}$ the above argument still applies, even if we need to compare to $a_{i,k+1}$ where $i < i'$. This is due to the fact that (10.10) for $a_{i',k+1}$ and for $a_{i,k+1}$ only involves entries of a, which have already been proven to respect the symmetries among each other, namely, $a_{p,q}$ with $p, q \in [k]$ and $a_{1,k+1}, \ldots, a_{i'-1,k+1}$.

Finally, if vertex $k+1$ has same colour as vertex $l \in [k]$, we show $a_{k+1,k+1} = a_{l,l}$. We have $\mathcal{G}_l \simeq \mathcal{G}_{k+1}$ by assumption (b) and $\Psi_{l,l} = \Psi_{k+1,k+1}$, since Ψ is in the RCON model. Furthermore, we have shown that all $a_{p,q}$, where $p \in [k]$ and $q \in [k+1]$, obey colouring c. Altogether, we conclude $a_{l,l} = a_{k+1,k+1}$ using (10.9). □

Proof of Theorem 10.2.8. If any of conditions (a), (b), and (c) do not hold, then $\mathcal{M}^{\to}_{(\mathcal{G},c)} \nsubseteq \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$, by Proposition 10.2.11, and hence the models cannot be equal. If conditions (a), (b) and (c) hold, we have $\mathcal{M}^{\to}_{(\mathcal{G},c)} \subseteq \mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)}$ (by Proposition 10.2.11) and $\mathcal{M}^{\mathrm{ud}}_{(\mathcal{G}^u,c)} \subseteq \mathcal{M}^{\to}_{(\mathcal{G},c)}$ (by Proposition 10.2.12). □

10.3 MLE: existence, uniqueness and an algorithm

In this section we characterize existence and uniqueness of MLEs in an RDAG model via linear dependence conditions on certain augmented sample matrices, see Theorem 10.3.6. This generalizes the characterization of ML estimation in usual DAG models from Theorem 6.3.16. Furthermore, the proof of Theorem 10.3.6 directly gives an algorithm to compute an MLE, if existent, in an RDAG model. Finally we present illustrative examples.

First, we define the augmented sample matrices given a coloured DAG $(\mathcal{G}, c)$ and sample matrix $Y \in \mathbb{K}^{m\times n}$. Let α_s be the number of vertices of colour $s \in c(I)$. Recall the set of *parent relationship colours* of vertex colour s from Equation (10.3):

$$\mathrm{prc}(s) = \{c(ij) \mid \text{there exists } j \to i \text{ in } \mathcal{G} \text{ with } c(i) = s\}, \qquad \beta_s := |\,\mathrm{prc}(s)|.$$

Definition 10.3.1 ([MRS21, Definition 4.1])**.** The *augmented sample matrix* of sample matrix $Y \in \mathbb{K}^{m\times n}$ and vertex colour s, denoted $M_{Y,s}$, has size $(\beta_s+1)\times\alpha_s n$.

We construct it row by row: let $M_{Y,s}^{(i)}$ denote the i^{th} row of $M_{Y,s}$, where we index from 0 to β_s. Each row consists of α_s blocks, each a row vector of length n. Let $i_1 < i_2 < \ldots < i_{\alpha_s}$ be the vertices of colour s. Then the top row of $M_{Y,s}$ is

$$M_{Y,s}^{(0)} := \begin{pmatrix} Y^{(i_1)} & Y^{(i_2)} & \ldots & Y^{(i_{\alpha_s})} \end{pmatrix} \in \mathbb{K}^{1\times(\alpha_s n)},$$

where $Y^{(i)}$ is the i^{th} row of sample matrix Y. The other rows of $M_{Y,s}$ are indexed by the parent relationship colours $t \in \operatorname{prc}(s)$:

$$M_{Y,s}^{(t)} := \begin{pmatrix} \sum\limits_{\substack{i_1 \leftarrow j \\ c(i_1 j)=t}} Y^{(j)} & \sum\limits_{\substack{i_2 \leftarrow j \\ c(i_2 j)=t}} Y^{(j)} & \cdots & \sum\limits_{\substack{i_{\alpha_s} \leftarrow j \\ c(i_{\alpha_s} j)=t}} Y^{(j)} \end{pmatrix}.$$

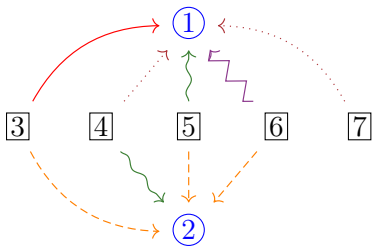
For $k \in [\alpha_s]$, the sum at the k^{th} block of $M_{Y,s}^{(t)}$ is zero if there are no $j \to i_k$ in $\mathcal{G}$ of colour t. Note that we frequently use the following abuse of notation: t is viewed as an edge colour like in $c(i_1 j) = t$, but also as its corresponding number $t \in [\beta_s]$ like in $M_{Y,s}^{(t)}$. ▲

Example 10.3.2 ([MRS21, Example 4.2])**.** For running example $\textcircled{1} \leftarrow \boxed{3} \rightarrow \textcircled{2}$,

$$M_{Y,\circ} = \begin{pmatrix} Y^{(1)} & Y^{(2)} \\ Y^{(3)} & Y^{(3)} \end{pmatrix} \begin{matrix} \circ \\ \rightarrow \end{matrix} \in \mathbb{K}^{2\times 2n} \qquad \text{and} \qquad M_{Y,\square} = \begin{pmatrix} Y^{(3)} \end{pmatrix} \in \mathbb{K}^{1\times n} \tag{10.11}$$

are the two augmented sample matrices, one for each vertex colour. ♢

Example 10.3.3 ([MRS21, Example 4.3])**.** The coloured DAG



has $M_{Y,\circ} = \begin{pmatrix} Y^{(1)} & Y^{(2)} \\ Y^{(3)} & 0 \\ 0 & Y^{(3)}+Y^{(5)}+Y^{(6)} \\ Y^{(5)} & Y^{(4)} \\ Y^{(6)} & 0 \\ Y^{(4)}+Y^{(7)} & 0 \end{pmatrix} \begin{matrix} \circ \\ \rightarrow \\ \dashrightarrow \\ \rightsquigarrow \\ \rightsquigarrow \\ \cdots\rightarrow \end{matrix}$

and $M_{Y,\square} = \begin{pmatrix} Y^{(3)} & Y^{(4)} & Y^{(5)} & Y^{(6)} & Y^{(7)} \end{pmatrix}$ □

as augmented sample matrices for vertex colour blue respectively black. ♢

The following two remarks are implicitly contained in [MRS21].

Remark 10.3.4 ($M_{Y,s}$ recovers $Y^{(i)\cup\operatorname{pa}(i)}$ for usual DAG models). In Remark 10.1.10 we have seen that any DAG model $\mathcal{M}_{\mathcal{G}}^{\rightarrow}$ is an RDAG model $\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}$, where colouring c assigns each vertex and each edge its own distinct colour. Thus, setting $s := c(i)$ for vertex $i \in [m]$, we have $\alpha_s = 1$ and $M_{Y,s}^{(0)} = Y^{(i)}$. Moreover, as each edge has its own colour, any parent of i can be uniquely identified with its parent relationship colour. Therefore, $|\operatorname{pa}(i)| = \beta_s := |\operatorname{prc}(s)|$ and $M_{Y,s}^{(t)} = Y^{(j)}$, where $j \to i$ in $\mathcal{G}$ and $c(ij) = t$. Altogether, $M_{Y,s} = Y^{(i)\cup\operatorname{pa}(i)}$ for a vertex i of $\mathcal{G}$. ▽

Remark 10.3.5. Let $(\mathcal{G}, c)$ be a coloured DAG with compatible colouring. Left-multiplication of $a \in \mathcal{A}(\mathcal{G})$ on $Y \in \mathbb{K}^{m\times n}$ is given by

$$(a \cdot Y)^{(i)} = a_{ii}Y^{(i)} + \sum_{j\in \mathrm{pa}(i)} a_{ij}Y^{(j)}$$

for all vertices $i \in [m]$. The augmented sample matrices are constructed such that the latter generalizes to $\mathcal{A}(\mathcal{G}, c)$. Let $a \in \mathcal{A}(\mathcal{G}, c)$ with vertex colour entries $a_{ss} \in \mathbb{K}^\times$ and edge colour entries $a_{st} \in \mathbb{K}$, where $s \in c(I)$ and $t \in \mathrm{prc}(s)$, compare Lemma 10.1.8(ii). Then $a \cdot Y$ is determined by

$$M^{(0)}_{a\cdot Y,s} = a_{ss}M^{(0)}_{Y,s} + \sum_{t\in \mathrm{prc}(s)} a_{st}M^{(t)}_{Y,s} \tag{10.12}$$

for all vertex colours $s \in c(I)$. ▽

Now, we formulate the main theorem of this section. By Remark 10.3.4, it generalizes Theorem 6.3.16 for DAG models to RDAG models.

Theorem 10.3.6 ([MRS21, Theorem 4.4])**.** *Consider the RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ on $(\mathcal{G}, c)$ where colouring c is compatible, and fix a sample matrix $Y \in \mathbb{K}^{m\times n}$. The following possibilities characterize maximum likelihood estimation given Y:*

(a)	ℓ_Y *unbounded from above*	$\Leftrightarrow$	$\exists\, s \in c(I)\colon\ M^{(0)}_{Y,s} \in \mathrm{span}\left\{M^{(t)}_{Y,s} : t \in [\beta_s]\right\}$
(b)	*MLE exists*	$\Leftrightarrow$	$\forall\, s \in c(I)\colon\ M^{(0)}_{Y,s} \notin \mathrm{span}\left\{M^{(t)}_{Y,s} : t \in [\beta_s]\right\}$
(c)	*MLE exists uniquely*	$\Leftrightarrow$	$\forall\, s \in c(I)\colon$ $M_{Y,s}$ *has full row rank.*

Example 10.3.7 ([MRS21, Example 4.5])**.** For running example ① ← 3 → ②, Theorem 10.3.6 says that the MLE exists uniquely if $Y^{(3)} \neq 0$ and $\begin{pmatrix} Y^{(1)} & Y^{(2)}\end{pmatrix}$ is not parallel to $\begin{pmatrix} Y^{(3)} & Y^{(3)}\end{pmatrix}$. This holds almost surely as soon as we have one sample, i.e., here $\mathrm{mlt}_u = 1$, as we mentioned in Example 10.0.1. ♢

Example 10.3.8 ([MRS21, Example 4.6])**.** Returning to Example 10.3.3, the MLE given Y exists provided $M_{Y,\square} = \begin{pmatrix} Y^{(3)} & \cdots & Y^{(7)}\end{pmatrix} \neq 0$, and $\begin{pmatrix} Y^{(1)} & Y^{(2)}\end{pmatrix}$ is not in the linear hull of the other rows of $M_{Y,\circ}$. The MLE is unique if and only if $M_{Y,\circ}$ is full row rank, since this also implies $M_{Y,\square} \neq 0$. ♢

The proof of Theorem 10.3.6 is analogous to the proof for uncoloured models in Theorem 6.3.16. In particular, we use again Lemma 6.3.15. The following proof also gives Algorithm 10.1 for computing an MLE, and a description of all MLEs, see Corollary 10.3.9.

Proof of Theorem 10.3.6. By Proposition 10.1.9, we have $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)} = \mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ as colouring c is compatible. In particular, for $\Psi = (\mathrm{I}_m - \Lambda)^\dagger\Omega^{-1}(\mathrm{I}_m - \Lambda) \in \mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$, the matrix $a = \Omega^{-1/2}(\mathrm{I}_m - \Lambda)$ giving the Cholesky decomposition $\Psi = a^\dagger a$ is in $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$, compare Lemma 10.1.11. As usual, let $\alpha_s := |c^{-1}(s)|$ and $\beta_s := |\,\mathrm{prc}(s)|$. By Lemma 10.1.8(ii), we can write the entries of the matrices a, Ω and Λ as a_{ss}

and a_{st}, ω_{ss} and λ_{st}, where $s \in c(I)$ and $t \in [\beta_s]$. Using Equation (10.12) with $a_{ss} = \omega_{ss}^{-1/2}$ and $a_{st} = -\omega_{ss}^{-1/2}\lambda_{st}$, and that $\det(\mathrm{I}_m - \Lambda) = 1$, we compute

$$\begin{aligned} -\ell_Y(\Psi) &= -\log\det(\Psi) + \operatorname{tr}(\Psi S_Y) \overset{(8.3)}{=} \log\det(\Omega) + \frac{1}{n}\|a \cdot Y\|^2 \\ &= \log\Big(\prod_{s\in c(I)} \omega_{ss}^{\alpha_s}\Big) + \frac{1}{n}\sum_{s\in c(I)} \Big\|\omega_{ss}^{-1/2}\Big(M_{Y,s}^{(0)} - \sum_{t\in[\beta_s]} \lambda_{s,t} M_{Y,s}^{(t)}\Big)\Big\|^2 \\ &= \sum_{s\in c(I)} \alpha_s \log(\omega_{ss}) + \frac{1}{n\omega_{ss}}\Big\|M_{Y,s}^{(0)} - \sum_{t\in[\beta_s]} \lambda_{s,t} M_{Y,s}^{(t)}\Big\|^2. \end{aligned}$$

An MLE is a minimizer of the above expression. Each parameter occurs in exactly one of the summands over $s \in c(I)$, because the set of edge colours is a disjoint union of the $\mathrm{prc}(s)$, see Lemma 10.1.8(i). We therefore minimize each summand separately, so fix $s \in c(I)$. We can first determine $\hat{\lambda}_{s,t}$, $t \in [\beta_s]$ that minimize

$$\Big\|M_{Y,s}^{(0)} - \sum_{t\in[\beta_s]} \lambda_{s,t} M_{Y,s}^{(t)}\Big\|^2, \tag{10.13}$$

by Lemma 6.3.15(iii). Such $\hat{\lambda}_{s,t}$ always exist: they are coefficients in the orthogonal projection $P_{Y,s}$ of $M_{Y,s}^{(0)}$ onto $\operatorname{span}\{M_{Y,s}^{(t)} : t \in [\beta_s]\}$, i.e.,

$$P_{Y,s} = \sum_{t\in[\beta_s]} \hat{\lambda}_{s,t} M_{Y,s}^{(t)}.$$

Furthermore, $\hat{\lambda}_{s,t}$, $t \in [\beta_s]$ are unique if and only if the vectors $M_{Y,s}^{(t)}$, $t \in [\beta_s]$ are linearly independent. Denote the minimum value of (10.13) by ζ_s. We will apply Lemma 6.3.15 several times with $\gamma_s := \zeta_s/n$.

If $M_{Y,s}^{(0)} \in \operatorname{span}\{M_{Y,s}^{(t)} : t \in [\beta_s]\}$ for some $s \in c(i)$, then $\zeta_s = 0$ and the summand $\alpha_s \log(\omega_{ss}) + \zeta_s/(n\omega_{ss})$ is not bounded from below for $\omega_{ss} > 0$, by Lemma 6.3.15(i). Hence, setting $\omega_{s',s'} = 1$ and $\lambda_{s',t'} = 0$ for all $s' \in c(I) \setminus \{s\}$ and all $t' \in [\beta_{s'}]$ shows that ℓ_Y is not bounded from above. This proves "$\Leftarrow$" of (a).

If $M_{Y,s}^{(0)} \notin \operatorname{span}\{M_{Y,s}^{(t)} : t \in [\beta_s]\}$, equivalently $\zeta_s > 0$, then the summand $\alpha_s \log(\omega_{ss}) + \zeta_s/(n\omega_{ss})$ has unique minimiser $\hat{\omega}_{ss} = \zeta_s/(n\alpha_s)$, by Lemma 6.3.15(ii). Hence, an MLE exists if $\zeta_s > 0$ for all $s \in c(I)$, which proves "$\Leftarrow$" in (b). As the right-hand sides of (a) and (b) are opposites and since MLE existence implies ℓ_Y is bounded from above, we have proved (a) and (b).

Since the $\hat{\omega}_{ss}$ are uniquely determined (if they exist), an MLE is unique if and only if all $\hat{\lambda}_{s,t}$ are unique. The latter is equivalent to: for all $s \in c(I)$ the vectors $M_{Y,s}^{(t)}$, $t \in [\beta_s]$ are linearly independent. In combination with the condition for MLE existence from (b) we deduce (c). □

The above proof of Theorem 10.3.6 gives Algorithm 10.1 and its correctness for finding a MLE in an RDAG model with compatible colouring. The MLE is given in a closed-form formula, as a collection of least squares estimators. It is returned in terms of the matrices Λ and Ω.

The proof of Theorem 10.3.6 also gives a description of the set of MLEs.

Algorithm 10.1: [MRS21, Algorithm 1]
MLE computation for an RDAG model with compatible colouring

Input : A coloured DAG $(\mathcal{G}, c)$ with compatible colouring c, a sample matrix $Y \in \mathbb{K}^{m\times n}$.
Output: An MLE given Y in the RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$, if one exists. Otherwise, returns "MLE does not exist".

for $s \in c(I)$ **do**
  $\alpha_s := |c^{-1}(s)|$;
  $\beta_s := |\operatorname{prc}(s)|$;
  construct matrix $M_{Y,s} \in \mathbb{K}^{(\beta_s+1)\times\alpha_s n}$;
  $P_{Y,s} :=$ orthogonal projection of $M^{(0)}_{Y,s}$ onto span $\{M^{(t)}_{Y,s} : t \in [\beta_s]\}$;
  if $P_{Y,s} = M^{(0)}_{Y,s}$ **then**
    return *MLE does not exist*;
  else
    coefficients $\lambda_{s,t}$ are such that $P_{Y,s} = \sum_{t\in\operatorname{prc}(s)} \lambda_{s,t} M^{(t)}_{Y,s}$;
    $\omega_{s,s} := (\alpha_s n)^{-1} \left\| P_{Y,s} - M^{(0)}_{Y,s} \right\|^2$;
  end
end
return *MLE for* Λ *and* Ω

Corollary 10.3.9 ([MRS21, Corollary 4.8])**.** *Consider the RDAG model on* $(\mathcal{G}, c)$ *where colouring* c *is compatible, with sample matrix* $Y \in \mathbb{K}^{m\times n}$*. If* (Λ, Ω) *and* (Λ', Ω') *are two MLEs, then* $\Omega = \Omega'$ *and*

$$\forall\, s \in c(i): \qquad \sum_{t\in\operatorname{prc}(s)} (\lambda_{s,t} - \lambda'_{s,t}) M^{(t)}_{Y,s} = 0.$$

We end this section with two illustrative examples of RDAG models and the theory presented herein. First, we apply our running example to model the effect of a mother's height on her two daughters' heights.

Example 10.3.10 ([MRS21, Example 4.12])**.** Let $\mathbb{K} = \mathbb{R}$. The RDAG model on the coloured DAG $\textcircled{1} \leftarrow \boxed{3} \rightarrow \textcircled{2}$ is parametrized by $\lambda \in \mathbb{R}$, $\omega, \omega' \in \mathbb{R}_{>0}$ and given by the linear structural equations

$$y_1 = \lambda y_3 + \varepsilon_1, \quad y_2 = \lambda y_3 + \varepsilon_2, \quad y_3 = \varepsilon_3, \quad \text{where } \varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, \omega), \varepsilon_3 \sim \mathcal{N}(0, \omega').$$

Let variable y_3 be the height (in cm) of a woman and let variables y_1 and y_2 be, respectively, the heights of her younger and older daughter. Vertices 1 and 2 both being blue indicates that, conditional on the mother's height, the variance of the daughter's heights is the same. Both edges being red encodes that the dependence of a daughter's height on the mother's height is the same for both daughters.

We saw in Example 10.3.7 that the MLE exists almost surely given one sample. We use Algorithm 10.1 to find the MLE, given one sample where the younger

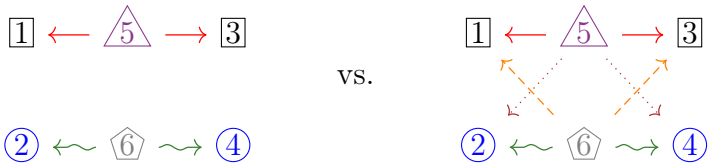
daughter's height is 159.75cm, the older daughter's height is 161.56, the mother's height is 155.32, and the population mean height is 163.83cm.[12] Mean-centring the data gives

$$\begin{pmatrix} Y^{(1)} & Y^{(2)} & Y^{(3)} \end{pmatrix} = \begin{pmatrix} -4.08 & -2.27 & -8.51 \end{pmatrix}.$$

The only black vertex is 3, and it has no parents, hence $\omega' = \|Y^{(3)}\|^2 = 72.42$. The orthogonal projection of $\begin{pmatrix} Y^{(1)} & Y^{(2)} \end{pmatrix}$ onto the line spanned by $\begin{pmatrix} Y^{(3)} & Y^{(3)} \end{pmatrix}$ has coefficient $\lambda = 0.37$ and residual $\omega = \left[(-3.175 + 4.08)^2 + (-3.175 + 2.27)^2\right]/2 = 0.82$. As we would expect, the regression coefficient λ is positive and the variance of the daughters' heights conditional on the mother's height is lower than the variance of the mother's height. ♢

Now, we consider multiple measurements taken in each generation.

Example 10.3.11 ([MRS21, Example 4.13])**.** We consider measurements of the snout length and head length of dogs. These are the first two of the seven morphometric parameters in the study of clinical measurements of dog breeds in [MMB+20]. We compare two RDAG models:



The black/square vertices 1 and 3 are the snout lengths of the two offspring. Blue/circular vertices 2 and 4 are their head lengths. The purple/triangular vertex 5 is the snout length of the parent and grey/pentagonal vertex 6 is the head length of the parent. The edges encode the dependence of the offsprings' traits on those of the parents.

Maximum likelihood estimation in the left hand model is two copies of Example 10.3.10, one on the three odd variables, and one on the three even variables. Thus, given one sample a unique MLE exists almost surely. For the right hand model, Theorem 10.3.6 says that an MLE exists provided $Y^{(5)} \neq 0$, $Y^{(6)} \neq 0$ and neither $\begin{pmatrix} Y^{(1)} & Y^{(3)} \end{pmatrix}$ nor $\begin{pmatrix} Y^{(2)} & Y^{(4)} \end{pmatrix}$ are in $\operatorname{span}\left\{\begin{pmatrix} Y^{(5)} & Y^{(5)} \end{pmatrix}, \begin{pmatrix} Y^{(6)} & Y^{(6)} \end{pmatrix}\right\}$. Hence an MLE exists almost surely with one sample. Moreover, the augmented sample matrices $M_{Y,\circ}$ and $M_{Y,\square}$ have full row rank almost surely provided $n \geq 2$, hence the MLE exists uniquely with two samples, by Theorem 10.3.6. ♢

10.4 Bounds on ML thresholds

In the previous section we gave a characterization of existence and unique existence of an MLE based on linear independence conditions, Theorem 10.3.6. Here

[12] We point out the difference to Remark 6.3.7. The latter discusses that the ML threshold increases by one, if the mean is *unknown* and also part of an MLE. However, here we assume the population mean to be known and use it to mean-venter the data. Hence, we only need to find the concentration matrix.

we use this theorem to give bounds on ML thresholds for RDAG models. These bounds hold whenever the colouring is compatible and there are no edges between vertices of the same colour.

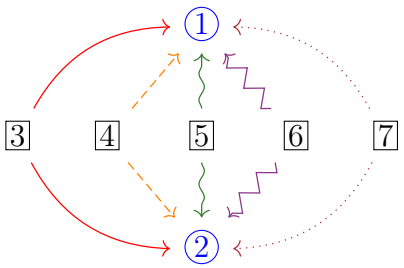
We point out that, similarly to the DAG case, $\mathrm{mlt}_b(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) = \mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)})$ holds by Theorem 10.3.6. However, in contrast to DAG models, we can have $\mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) < \mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)})$ for an RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$. In fact, Example 10.4.12 gives a family of RDAG models for which the gap becomes arbitrarily large.

The section is organized as follows. We start with some definitions. Then we prove two lemmata and two propositions to deduce the main result, Theorem 10.4.9. Afterwards, we discuss examples and end with a randomized method to compute existence and uniqueness threshold.

Definition 10.4.1 ([MRS21, Definition 5.1])**.** Let M_Y be a matrix whose entries are linear combinations of the entries of a matrix $Y \in \mathbb{K}^{m\times n}$. The *generic rank* of M_Y is its rank for generic Y. ▲

We often study the generic rank of M_Y by considering it as a symbolic matrix whose entries are linear forms in the mn indeterminates Y_{ij}.

Example 10.4.2 ([MRS21, Example 5.2])**.** The coloured DAG[13]



has $M_{Y,\circ} = \begin{pmatrix} Y^{(1)} & Y^{(2)} \\ Y^{(3)} & Y^{(3)} \\ Y^{(4)} & Y^{(4)} \\ Y^{(5)} & Y^{(5)} \\ Y^{(6)} & Y^{(6)} \\ Y^{(7)} & Y^{(7)} \end{pmatrix}$ ∘ → ⇢ ⇝ ↝ ⋯→

When $n = 1$, the matrix $M_{Y,\circ}$ has generic rank two. Removing its top row gives a 5×2 matrix of generic rank one. ♢

Let $(\mathcal{G}, c)$ be a coloured DAG. For $s \in c(I)$, let α_s be the number of vertices of colour s and β_s the number of parent relationship colours of s.

Definition 10.4.3. Fix a vertex colour $s \in c(I)$. For sample matrix $Y \in \mathbb{K}^{m\times n}$ let $M_{Y,s} \in \mathbb{K}^{(\beta_s+1)\times\alpha_s n}$ be as in Definition 10.3.1. We define the following.

1. $M'_{Y,s} \in \mathbb{K}^{\beta_s\times\alpha_s n}$ is the submatrix of $M_{Y,s}$ obtained from removing the top row $M^{(0)}_{Y,s}$. In other words, the rows of $M'_{Y,s}$ are $M^{(1)}_{Y,s}, M^{(2)}_{Y,s}, \ldots, M^{(\beta_s)}_{Y,s}$.
2. r_s is the generic rank of $M'_{Y,s}$ when $n = 1$.
3. $\mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right)$ is the smallest n such that $M^{(0)}_{Y,s} \notin \left\{M^{(t)}_{Y,s} \mid t \in \mathrm{prc}(s)\right\}$ holds for almost all $Y \in \mathbb{K}^{m\times n}$.
4. $\mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right)$ is the smallest n such that $M_{Y,s}$ has full row rank $\beta_s + 1$ for almost all $Y \in \mathbb{K}^{m\times n}$. ▲

[13] with two vertex colours (blue/circular and black/square) and five edge colours (red/solid, orange/dashed, green/squiggly, purple/zigzag, and brown/dotted)

In the following we prove bounds on the ML thresholds of an RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$. We proceed as follows. By Theorem 10.3.6, it suffices to give bounds on $\mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right)$ and $\mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right)$ for all $s \in c(I)$. Such bounds are given in Propositions 10.4.8 and 10.4.7 in terms of α_s, β_s and r_s. To obtain these bounds, we show the following two lemmata which study the generic rank of $M'_{Y,s}$ as the sample size n grows.

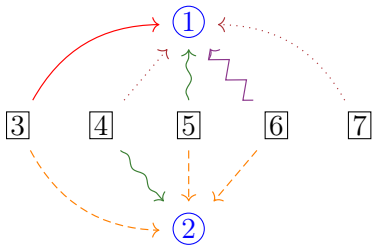
Lemma 10.4.4 ([MRS21, Lemma 5.8])**.** *Consider the RDAG model on $(\mathcal{G}, c)$ where colouring c is compatible, and fix a vertex colour s. For $n \geq \beta_s$ and generic $Y \in \mathbb{K}^{m\times n}$ the row vectors $M^{(1)}_{Y,s}, \ldots, M^{(\beta_s)}_{Y,s}$ are linearly independent.*

Proof. We think of the mn entries of Y as indeterminates and construct an invertible $\beta_s \times \beta_s$ submatrix of $M'_{Y,s} \in \mathbb{K}^{\beta_s \times \alpha_s n}$. We illustrate this construction in Example 10.4.5 below.

Let $i_1 < i_2 < \ldots < i_{\alpha_s}$ be the vertices of colour s. The matrix $M'_{Y,s}$ has α_s many blocks of size $\beta_s \times n$. For each parent relationship colour p_t, $t \in [\beta_s]$ there is some vertex $i_k = i_k(t)$ (where $k \in [\alpha_s]$) such that there is an edge of colour p_t pointing towards vertex $i_k = i_k(t)$. That is, the k^{th} block of $M'_{Y,s}$ has non-zero entries in the t^{th} row. Let $C_t \in \mathbb{K}^{\beta_s \times 1}$ be the t^{th} column of that block, which exists as $n \geq \beta_s$. By construction, the t^{th} entry of C_t is non-zero. We show that the matrix $C = \begin{pmatrix} C_1 & C_2 & \ldots & C_{\beta_s} \end{pmatrix}$, is invertible.

An entry of C is either a sum of variables or it is zero. By construction, column C_t only contains (sums of) elements of the t^{th} column of Y. The same variable $Y_{j,t}$ cannot occur in two different entries of C_t, because there is at most one edge from j to vertex $i_k(t)$. Altogether, the entries of C are (possible empty) sums of variables and each variable occurs in at most one entry of C. Thus, and since the determinant is an alternating sum over products of permutations, it is enough to show that one product is non-zero. By construction, $C_{11}C_{22}\cdots C_{\beta_s\beta_s} \neq 0$. Thus, $M'_{Y,s}$ has generic rank β_s for $n \geq \beta_s$. □

Example 10.4.5. We illustrate the construction of C and its underlying combinatorial idea from the proof of Lemma 10.4.4. Consider the coloured DAG from Example 10.3.3, i.e.,



with $M'_{Y,\circ} = \begin{pmatrix} Y^{(3)} & 0 \\ 0 & Y^{(3)} + Y^{(5)} + Y^{(6)} \\ Y^{(5)} & Y^{(4)} \\ Y^{(6)} & 0 \\ Y^{(4)} + Y^{(7)} & 0 \end{pmatrix}$

Considering vertex colour blue, we have $\alpha_\circ = 2$ and $\beta_\circ = 5$. Let $n \geq \beta_\circ = 5$ and take $Y \in \mathbb{K}^{7\times n}$. The parent relationship colours (prc) are ordered as indicated by $M'_{Y,\circ}$, i.e., red, orange, green, purple and finally brown. Arrows of colour red only point towards vertex 1. Thus, we have to choose the first column (red is first prc) from the first block (i.e., the block for vertex 1) of $M'_{Y,\circ}$. This determines the first column of C. Similarly, we have to choose the second column from the second

block for colour orange, and the fourth and fifth column from the first block for colours purple and brown. Only for the third colour green, we can choose both the first and the second block. We take the first block. Altogether, we obtain

$$C = \begin{pmatrix} Y_{3,1} & 0 & Y_{3,3} & Y_{3,4} & Y_{3,5} \\ 0 & Y_{3,2}+Y_{5,2}+Y_{6,2} & 0 & 0 & 0 \\ Y_{5,1} & Y_{4,2} & Y_{5,3} & Y_{5,4} & Y_{5,5} \\ Y_{6,1} & 0 & Y_{6,3} & Y_{6,4} & Y_{6,5} \\ Y_{4,1}+Y_{7,1} & 0 & Y_{4,3}+Y_{7,3} & Y_{4,4}+Y_{7,4} & Y_{4,5}+Y_{7,5} \end{pmatrix}$$

and $\prod_{i=1}^{5} C_{ii} = Y_{3,1}(Y_{3,2}+Y_{5,2}+Y_{6,2})Y_{5,3}Y_{6,4}(Y_{4,5}+Y_{7,5})$. The matrix C is indeed invertible for generic Y. ♢

The next lemma and its proof are contained (in condensed form) in [MRS21] in the proof of Proposition 5.9.

Lemma 10.4.6. *Let $(\mathcal{G}, c)$ be a coloured DAG with compatible colouring c. For generic $Y \in \mathbb{K}^{m\times n}$ the rank of $M'_{Y,s}$ is at least* $\min\{r_s + n - 1, \beta_s\}$.

Proof. Note that by construction of $M_{Y,s}$ (respectively $M'_{Y,s}$)

$$\mathcal{X}^{\{1,n\}} := \left\{ M'_{Y,s} \mid Y \in \mathbb{K}^{m\times n} \right\}$$

is a linear subspace of $\mathbb{K}^{\beta_s \times \alpha_s n}$. The generic rank of $M'_{Y,s} \in \mathbb{K}^{\beta_s\times\alpha_s n}$, denoted $r_s(n)$, is given by $r_s(n) = \max\{\operatorname{rank}(X) \mid X \in \mathcal{X}^{\{1,n\}}\}$. Note that $r_s = r_s(1)$, by Definition 10.4.3. The space $\mathcal{X}^{\{1,n\}}$ is the so-called $(1, n)$ blow up[14] of $\mathcal{X} := \mathcal{X}^{\{1,1\}}$. In view of the generic matrix $M'_{Y,s} \in \mathbb{K}^{\beta_s\times\alpha_s}$ the $(1, n)$ matrix blow up means that the scalar variables $Y^{(i)}$ are replaced by generic row vectors of length n, to give a $\beta_s \times \alpha_s n$ matrix. As suggested by the notation, this setting fits [DM17, Section 2]. By [DM17, Lemma 2.7 parts (1) and (3)], we have for all n that

$$r_s(n) \le r_s(n+1) \quad \text{and} \quad r_s(n+1) \ge \frac{1}{2}\big(r_s(n) + r_s(n+2)\big), \tag{10.14}$$

i.e., $r_s(n)$ is weakly increasing and weakly concave. Moreover, the maximum rank among the $r_s(n)$ is β_s, which occurs for $n \ge \beta_s$ by Lemma 10.4.4. Now, let n be such that $r_s(n) < \beta_s$ and $r_s(n) = r_s(n+1)$. Then, by the left inequality in (10.14), there exists some integer $2 \le k \le \beta_s - n$ with

$$r_s(n) = r_s(n+1) = \cdots = r_s(n+k-1) < r_s(n+k),$$

but this contradicts $r_s(n+k-1) \ge \frac{1}{2}\big(r_s(n+k-2)+r_s(n+k)\big)$, the right inequality of (10.14). Therefore, $r_s(n) < \beta_s$ implies $r_s(n) + 1 \le r_s(n+1)$. We conclude by induction on n that $r_s(n) \ge \min(r_s + n - 1, \beta_s)$ for all n. □

Equipped with the previous lemmata we prove bounds on $\operatorname{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right)$ and $\operatorname{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right)$.

[14] This is not related to the blow up construction from Algebraic Geometry, e.g., to resolve singularities.

Proposition 10.4.7 ([MRS21, Proposition 5.10]). *Let $(\mathcal{G}, c)$ be a coloured DAG that has no edges between any vertices of colour s, and c is compatible. Then*

$$\left\lfloor \frac{\beta_s}{\alpha_s} \right\rfloor + 1 \leq \mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) \leq \beta_s + 2 - r_s.$$

Moreover, if $r_s \neq \beta_s + 1 - (\beta_s/\alpha_s)$ then $\mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) \leq \beta_s + 1 - r_s$.

Proof. To prove the lower bound, we observe that if $\alpha_s n \leq \beta_s$, then the $\beta_s + 1$ rows of $M_{Y,s}$ will be linearly dependent. Hence, we need at least $n > \beta_s/\alpha_s$ many samples for $M_{Y,s}$ to have generically full row rank.

To prove the upper bound, let $M'_{Y,s}$ and r_s be as in Definition 10.4.3. By Lemma 10.4.6, for n samples we have $\mathrm{rank}(M'_{Y,s}) \geq \min(r_s + n - 1, \beta_s)$ generically. Thus, for $n = \beta_s + 1 - r_s$ the matrix $M'_{Y,s}$ generically has full row rank β_s. It remains to consider the top row of $M_{Y,s}$. We must have $\beta_s \leq \alpha_s n$, otherwise the $\beta_s \times \alpha_s n$ matrix $M'_{Y,s}$ could not have full row rank. We look separately at the possible cases: $\beta_s < n\alpha_s$ and $\beta_s = n\alpha_s$. If $\beta_s < n\alpha_s$, the row vector $M^{(0)}_{Y,s} \in \mathbb{K}^{1\times \alpha_s n}$ is generically not in the span of the β_s rows of $M'_{Y,s}$, because there are no edges between vertices of colour s. Thus, $M_{Y,s}$ generically has full row rank $\beta_s + 1$, and $\mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s) \leq n = \beta_s + 1 - r_s$. If $\beta_s = n\alpha_s$, equivalently if $r_s = \beta_s + 1 - (\beta_s/\alpha_s)$, an additional sample ensures $\mathrm{rank}(M_{Y,s}) = \beta_s + 1$ generically. □

Proposition 10.4.8 ([MRS21, Proposition 5.9]). *Let $(\mathcal{G}, c)$ be a coloured DAG that has no edges between any vertices of colour s, and c is compatible. If $\alpha_s = 1$, then $\mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) = \mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) = \beta_s + 1$, while if $\alpha_s \geq 2$ we have*

$$\left\lfloor \frac{r_s - 1}{\alpha_s - 1} \right\rfloor + 1 \leq \mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) \leq \left\lfloor \frac{\beta_s}{\alpha_s} \right\rfloor + 1.$$

Proof. If $\alpha_s = 1$, then $M^{(0)}_{Y,s} = Y^{(i)}$ where i is the unique vertex of colour s. Moreover, each row $M^{(t)}_{Y,s}$, $t \in [\beta_s]$ is non-zero and a sum of certain $Y^{(j)}$, $j \in \mathrm{pa}(i)$. Note that $Y^{(i)}$ only appears in $M^{(0)}_{Y,s}$ and each parent row of $Y^{(i)}$ appears in exactly one row $M^{(t)}_{Y,s}$, $t \in [\beta_s]$. Similarly to the uncoloured case, the $M^{(t)}_{Y,s}$, $t \in [\beta_s]$ span $\mathbb{K}^{1\times n}$ generically if $n \leq \beta_s$; and $M_{Y,s}$ has generically full row rank if $n \geq \beta_s + 1$. Altogether, $\mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) = \mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) = \beta_s + 1$ if $\alpha_s = 1$.

It remains to consider $\alpha_s \geq 2$. To prove the upper bound, we show that if $n > \beta_s/\alpha_s$, then the top row of $M_{Y,s}$ is generically not in the span of the other rows. Since there are no edges between two vertices of colour s, the $n\alpha_s$ entries of the top row $M^{(0)}_{Y,s}$ are all independent, from each other and from the entries of the other rows. If $\beta_s < \alpha_s n$, the other β_s rows do not span $\mathbb{K}^{n\alpha_s}$, so a generic choice of top row will not lie in their span.

For the lower bound, the generic rank of $M'_{Y,s}$ at least $\min\{r_s + n - 1, \beta_s\}$, by Lemma 10.4.6. Thus, the top row $M^{(0)}_{Y,s}$ is in the span of the other rows whenever $\min(r_s + n - 1, \beta_s) \geq n\alpha_s$. The latter holds in particular, if $\alpha_s n \leq r_s + n - 1 \leq \beta_s$ holds, i.e.,

$$n \leq \min\left(\left\lfloor \frac{r_s - 1}{\alpha_s - 1} \right\rfloor, \beta_s + 1 - r_s\right).$$

Hence, we need at least one more sample to guarantee that $M^{(0)}_{Y,s}$ is not in the row span of $M'_{Y,s}$, i.e.,

$$\operatorname{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) \geq \min\left(\left\lfloor \frac{r_s - 1}{\alpha_s - 1} \right\rfloor + 1, \beta_s + 2 - r_s\right).$$

The minimum must be attained by the former expression, because

$$\operatorname{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) \leq \operatorname{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right) \leq \beta_s + 2 - r_s$$

holds by Proposition 10.4.7. □

As a consequence of Theorem 10.3.6 parts (b) and (c) we have

$$\operatorname{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right) = \max_{s\in c(I)} \operatorname{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right), \quad \operatorname{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right) = \max_{s\in c(I)} \operatorname{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s\right).$$

Taking the maximum of the lower and upper bounds in Propositions 10.4.8 and 10.4.7, over all vertex colours, gives the main theorem of this section.

Theorem 10.4.9 ([MRS21, Theorem 5.3])**.** *Consider the RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ on $(\mathcal{G}, c)$ where colouring c is compatible, and $(\mathcal{G}, c)$ has no edges between vertices of the same colour. For vertex colour s, set $l(s) := (r_s - 1)(\alpha_s - 1)^{-1}$ if $\alpha_s \geq 2$ and $l(s) := \beta_s$ if $\alpha_s = 1$. We have the following bounds on ML thresholds:*

$$\max_{s\in c(I)} \lfloor l(s) \rfloor + 1 \leq \operatorname{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right) \leq \max_{s\in c(I)} \left\lfloor \frac{\beta_s}{\alpha_s} \right\rfloor + 1, \tag{10.15}$$

$$\max_{s\in c(I)} \left\lfloor \frac{\beta_s}{\alpha_s} \right\rfloor + 1 \leq \operatorname{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right) \leq \max_{s\in c(I)} \left(\beta_s + 2 - r_s\right). \tag{10.16}$$

It remains an open problem to turn these bounds into formulae.

Problem 10.4.10 ([MRS21, Problem 5.4])**.** *Determine the maximum likelihood thresholds of an RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$, as formulae involving properties of the DAG $\mathcal{G}$ and its colouring c.*

Remark 10.4.11. We point out the following regarding Theorem 10.4.9.

(i) Recall from Remark 10.1.10 that any DAG model $\mathcal{M}^{\rightarrow}_{\mathcal{G}}$ is an RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ with compatible colouring. In this situation, $\alpha_s = 1$ for all $s \in c(I)$ and $\beta_s = |\operatorname{pa}(i)|$, where i is the unique vertex with $c(i) = s$. Therefore, Theorem 10.4.9 contains Corollary 6.3.19 as a special case.

(ii) The upper bounds for existence threshold and uniqueness threshold are both at most $\max_s \beta_s + 1$.[15] Hence, the RDAG thresholds are always at least as small as the DAG threshold, by part(i).

[15] If $\mathcal{G}$ does not have any edges, i.e., $\beta_s = r_s = 0$ for all $s \in c(I)$, then the uniqueness threshold trivially equals one as then each row vector $M_{Y,s} \in \mathbb{K}^{1\times\alpha_s n}$ has generic rank one.

(iii) Theorem 10.4.9 applies to all RDAG models with compatible colouring that are equal to its induced RCON model, because such models never have edges between vertices of the same colour, as follows. Take the minimal vertex i such that $i \leftarrow j$ in $\mathcal{G}$ and $c(i) = c(j)$. Then no children of i have colour $c(i)$, therefore $\mathcal{G}_i \neq \mathcal{G}_j$, a contradiction to Theorem 10.2.8(b).[16] ▽

We illustrate the threshold bounds in some examples. The first example shows that the existence threshold and uniqueness threshold for an RDAG model can have arbitrarily large distance.

Example 10.4.12 ([MRS21, Example 5.5])**.** We find the existence and uniqueness threshold for the RDAG model on the coloured DAG $(\mathcal{G}, c)$ from Example 10.4.2. Since the black (square) vertices have no parents, the matrix $M_{Y,\square}$ has full rank as soon as $n \geq 1$. Therefore, the thresholds are determined by vertex colour blue. The generic rank of $M'_{Y,\circ}$ is one when $n = 1$, i.e., $r_\circ = 1$. Using $\alpha_\circ = 2$ and $\beta_\circ = 5$, Theorem 10.4.9 and Proposition 10.4.7 give bounds

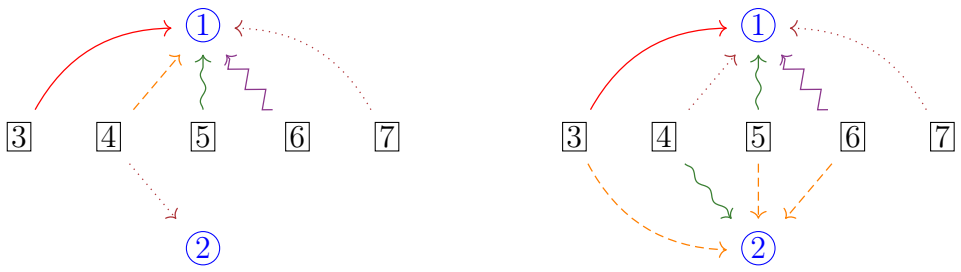
$$\frac{r_\circ - 1}{\alpha_\circ - 1} + 1 = 1 \leq \mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) \quad \text{and} \quad \mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) \leq \beta_\circ + 1 - r_\circ = 5.$$

In fact, both bounds are attained as follows. For all $n \geq 1$, the row $M^{(0)}_{Y,\circ} = (Y^{(1)}, Y^{(2)})$ is almost surely not contained in the span of the other rows of $M_{Y,\circ}$, hence $\mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) = 1$. Moreover, we need $n \geq 5$ samples for generic linear independence of the rows $(Y^{(3)}, Y^{(3)}), \ldots, (Y^{(7)}, Y^{(7)})$. Thus, $\mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) = 5$.

This example extends to an arbitrary number of vertices, i.e., to the coloured DAG with $k + 2$ vertices, 2 blue/circular and k black/square, and $2k$ edges of k colours (arranged as in the $k = 5$ case above). Repeating the above argument gives $\mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) = 1$ and $\mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}) = k$. ♢

We modify the edges from Examples 10.4.2 and 10.4.12 to see how the thresholds change.

Example 10.4.13 ([MRS21, Example 5.7])**.** Consider the following coloured DAGs, both with compatible colouring



Since the black vertices do not have parents, the thresholds are determined by

[16] This is [MRS21, Remark 5.6].

the blue vertices. Given sample matrix $Y \in \mathbb{K}^{7\times n}$, we respectively obtain

$$M_{Y,\circ} = \begin{pmatrix} Y^{(1)} & Y^{(2)} \\ Y^{(3)} & 0 \\ Y^{(4)} & 0 \\ Y^{(5)} & 0 \\ Y^{(6)} & 0 \\ Y^{(7)} & Y^{(4)} \end{pmatrix} \begin{matrix} \circ \\ \longrightarrow \\ \dashrightarrow \\ \rightsquigarrow \\ \wedge\!\!\rightarrow \\ \cdots\!\!\rightarrow \end{matrix} \qquad M_{Y,\circ} = \begin{pmatrix} Y^{(1)} & Y^{(2)} \\ Y^{(3)} & 0 \\ 0 & Y^{(3)}+Y^{(5)}+Y^{(6)} \\ Y^{(5)} & Y^{(4)} \\ Y^{(6)} & 0 \\ Y^{(4)}+Y^{(7)} & 0 \end{pmatrix} \begin{matrix} \circ \\ \longrightarrow \\ \dashrightarrow \\ \rightsquigarrow \\ \wedge\!\!\rightarrow \\ \cdots\!\!\rightarrow \end{matrix}$$

In both cases we have $\alpha_\circ = 2$, $\beta_\circ = 5$, and $r_\circ = 2$. Thus, Theorem 10.4.9 gives in both cases

$$2 = \left\lfloor \frac{r_\circ - 1}{\alpha_\circ - 1} \right\rfloor + 1 \leq \mathrm{mlt}_e \leq \left\lfloor \frac{\beta_\circ}{\alpha_\circ} \right\rfloor + 1 = 3$$
$$3 = \left\lfloor \frac{\beta_\circ}{\alpha_\circ} \right\rfloor + 1 \leq \mathrm{mlt}_u \leq \beta_\circ + 2 - r_\circ = 5.$$

Actually, Proposition 10.4.7 yields $\mathrm{mlt}_u \leq 4$, since $r_\circ \neq \beta_\circ + 1 - (\beta_\circ/\alpha_\circ)$. In the following we determine the precise values of the thresholds.

First, we study the left-hand RDAG. When $n = 2$ the row $Y^{(2)} \in \mathbb{K}^{1\times 2}$ is generically not in the span of $Y^{(4)}$, hence $M^{(0)}_{Y,\circ} = (Y^{(1)}, Y^{(2)})$ is not in the linear span of the other five rows of $M_{Y,\circ}$, so $\mathrm{mlt}_e = 2$. For $n \geq 2$, the submatrix

$$\begin{pmatrix} Y^{(2)} \\ Y^{(4)} \end{pmatrix} \in \mathbb{K}^{2\times n}$$

has generic rank two. Therefore, $M_{Y,\circ} \in \mathbb{K}^{6\times 2n}$ has generic rank at most five if $n = 3$. However, $n = 4$ suffices for $M_{Y,\circ}$ to have full row rank 6 generically. We conclude $\mathrm{mlt}_u = 4$ for the left-hand RDAG.

Next, we study the right-hand RDAG. For $n = 2$, $M^{(0)}_{Y,\circ} = (Y^{(1)}, Y^{(2)})$ is generically contained in the linear span of the other rows of $M_{Y,\circ}$. Together with $\mathrm{mlt}_e \leq 3$ we conclude that $\mathrm{mlt}_e = 3$. For uniqueness, when $n = 3$ the submatrices

$$\begin{pmatrix} Y^{(3)} \\ Y^{(6)} \\ Y^{(4)}+Y^{(7)} \end{pmatrix}, \quad \begin{pmatrix} Y^{(2)} \\ Y^{(3)}+Y^{(5)}+Y^{(6)} \\ Y^{(4)} \end{pmatrix} \in \mathbb{K}^{3\times 3}$$

of $M_{Y,\circ}$ generically have rank three, and the zero pattern ensures that $M_{Y,\circ}$ has full row rank six generically. Combining this with $3 \leq \mathrm{mlt}_u$ gives $\mathrm{mlt}_u = 3$. ♢

We end this section with the following proposition.

Proposition 10.4.14 ([MRS21, Proposition 5.11])**.** *For an RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$, where colouring c is compatible, there is a randomized, polynomial time algorithm for computing the thresholds* $\mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right)$ *and* $\mathrm{mlt}_u\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right)$.

Proof. Remember that $\mathrm{mlt}_e\left(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}\right) = \max_{s\in c(I)} \mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$ and we have the same equality for mlt_u. As $|c(I)| \leq m$ is part of the data, i.e., the coloured DAG $(\mathcal{G}, c)$, it suffices to give a randomized polynomial time algorithm to compute $\mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$ and $\mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$ for a fixed vertex colour s.

The rank of a symbolic matrix can be computed in polynomial time by a randomized algorithm, see e.g., [Lov79; Sch80]. Hence, thinking of the entries of $Y \in \mathbb{K}^{m\times n}$ as indeterminates, we can compute for any $n \geq 1$ the rank of the symbolic $(\beta_s + 1) \times \alpha_s n$ matrix $M_{Y,s}$ as well as the rank of the symbolic $\beta_s \times \alpha_s n$ matrix $M'_{Y,s}$. We obtain $\mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$ as the smallest n such that $\mathrm{rank}(M_{Y,s}) > \mathrm{rank}(M'_{Y,s})$ and $\mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$ as the smallest n such that $\mathrm{rank}(M_{Y,s}) = \beta_s + 1$. The algorithm terminates and has polynomial time by the upper bound of $\beta_s + 1$ for both $\mathrm{mlt}_e(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$ and $\mathrm{mlt}_u(\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}, s)$, i.e., $n \leq \beta_s + 1$ suffices. □

10.5 Simulations

This section is [MRS21, Section 6]. The simulations, their Python implementation and the analysis were completely done by Anna Seigal.

In the previous section, we gave upper and lower bounds for the maximum likelihood thresholds for RDAG models, see Theorem 10.4.9. The bounds quantify how the graph colouring serves to decrease the number of samples needed for existence and uniqueness of the MLE. In this section, we assume that the number of samples is above the maximum likelihood threshold. We explore via simulations the distance of an MLE to the true model parameters. We compare the RDAG model estimate from Algorithm 10.1 to the usual (uncoloured) DAG model MLE.

The details of our simulations are as follows. We used the NetworkX Python package [HSS08a] to build an RDAG model via the following steps. We first build a DAG by generating an undirected graph according to an Erdős–Rényi model that includes each edge with fixed probability, and then directing the edges so that $j \to i$ implies $j > i$. We assign edge colours randomly, after fixing the total number of possible edge colours. We choose the unique vertex colouring with the largest number of vertex colours that satisfies the compatibility assumption from Definition 10.1.6. We sample edge weights λ_{st} from a uniform distribution on $[-1, -0.25] \cup [0.25, 1]$ and we sample noise variances ω_{ss} uniformly from $[0, 1]$. Our code is available at `https://github.com/seigal/rdag`.

The results of the simulations are presented as *violin plots* using the Python package seaborn [Was21]. The following information is taken from `https://seaborn.pydata.org/tutorial/categorical.html#categorical-tutorial`. The violin plot shows the data range and the (smoothed) probability density of the observed data, which gives the characteristic "violin shape", compare Figure 10.1. Moreover, the black rectangle inside a "violin" indicates the quartiles of the distribution: the rectangle itself presents the interquartile range (IQR, middle 50%) while the white point inside the black rectangle presents the median (i.e., the second quartile). The "whiskers" indicate points that lie within 1.5 IQRs of the first and third quartile.

Now, we describe our three simulations and interpret the outcome. First, the RDAG MLE is generally closer to the true model parameters than the DAG MLE, see Figure 10.1. As we would expect, both estimates get closer to the true parameters as the number of samples from the distribution increases. At a high

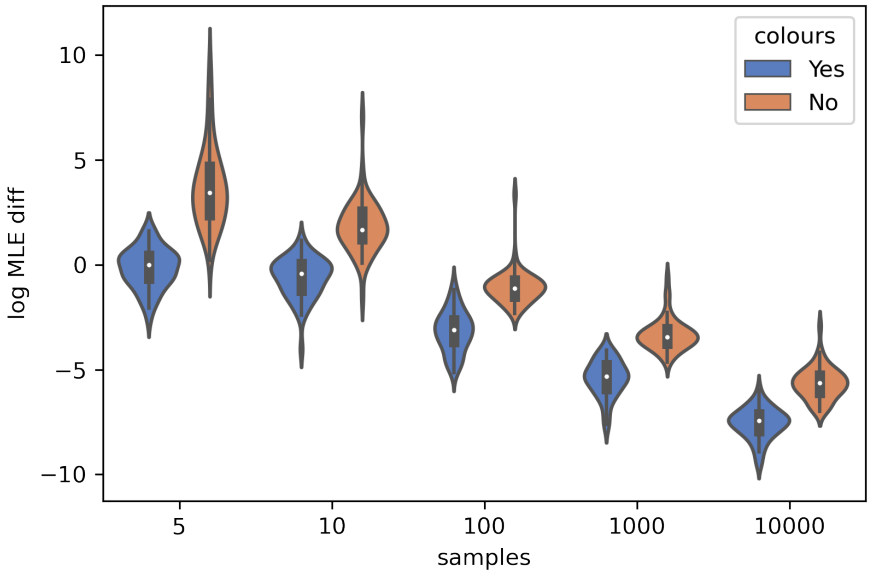


Figure 10.1: [MRS21, Figure 1] We generated RDAGs on 10 vertices, with each edge present with probability 0.5 and 5 edge colours. We sampled from the distribution $n \in \{5, 10, 100, 1000, 10000\}$ times. For each n we generated 50 random graphs and computed the RDAG MLE and the DAG MLE, comparing them to the true parameter values on a log scale. The results are displayed in a violin plot, with blue for the RDAG MLE and orange for the DAG MLE.

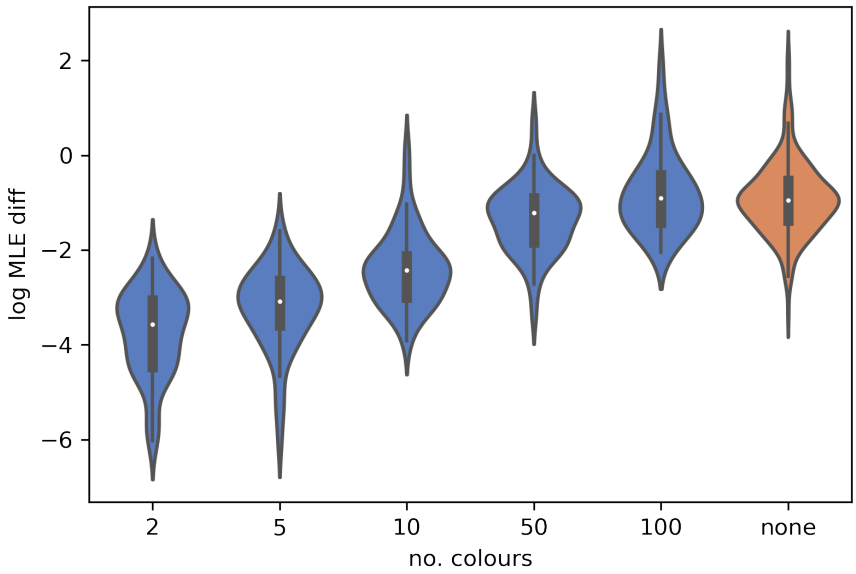


Figure 10.2: [MRS21, Figure 2] We generated RDAGs on 10 vertices, each edge present with probability 0.5 and number of edge colours in $\{2, 5, 10, 50, 100\}$. We sampled from the distribution 100 times and compared the MLE to the true parameter values on a log scale. The DAG MLE is shown in orange for comparison.

number of samples, the difference between the RDAG MLE and the DAG MLE is smaller than at a low number of samples.

Second, we examined how the RDAG MLE was affected by the number of edge colours, see Figure 10.2. The RDAG MLE is closest to the true parameters when the number of edge colours is small; i.e., when there are fewer parameters to learn. As the number of edge colours increases, the difference between the RDAG MLE and the DAG MLE decreases. Note that the DAG model is the setting where each vertex and edge has its own colour.

Third, we looked at how the RDAG MLE and DAG MLE are affected by the edge density of the graph, see Figure 10.3. The RDAG MLEs get closer to the true parameter values as the edge density increases: more edges have the same weight, so more samples contribute to estimating each edge weight. By comparison, the DAG MLEs get further away from the true parameters as the edge density increases, because there are more parameters to learn.

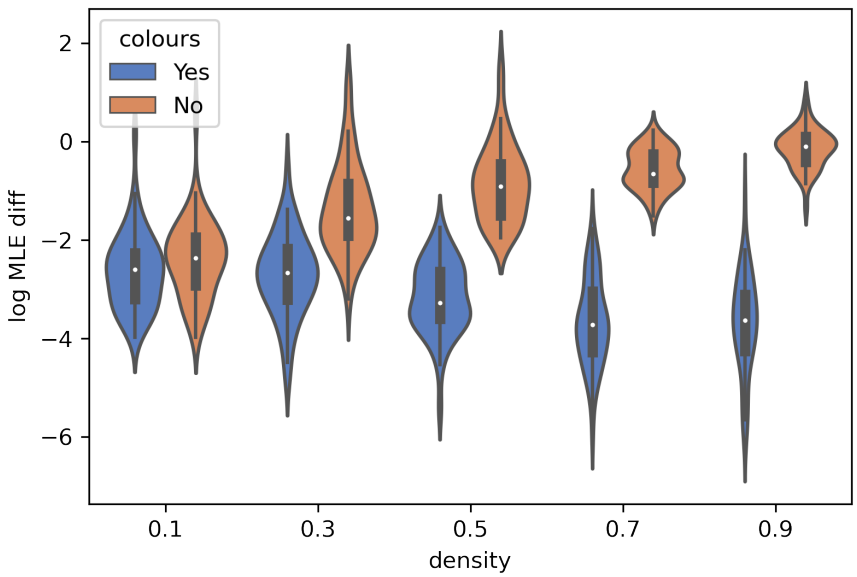


Figure 10.3: [MRS21, Figure 3] We generated RDAGs on 10 vertices, each edge present with probability in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and 5 edge colours. For each edge probability we generated 50 random graphs, sampled from each one 100 times, and compared the RDAG and DAG MLEs. As above, blue is the RDAG MLE and orange is the DAG MLE.

10.6 Connections to Stability Notions

This section characterizes ML estimation for RDAG models via stability notions under *sets* $\mathcal{A} \subseteq \mathrm{GL}_m(\mathbb{K})$, see Definition 8.2.1. We proceed similar to the study of TDAG models in Section 9.5. It is remarkable that the full correspondence extends to RDAG models $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ with compatible colouring c, Theorem 10.6.4.

We prove this by showing that the linear independence conditions from Theorem 10.3.6 are equivalent to stability notions for the sample matrix, see Theorem 10.6.3. Furthermore, we study the set of MLEs in Proposition 10.6.6, which offers an alternative characterization to Corollary 10.3.9. We start with the weak correspondence for RDAG models.

Remark 10.6.1 (Weak Correspondence for RDAG models, [MRS21, Remark A.5]). Let $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ be an RDAG model with compatible colouring c, so $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)} = \mathcal{M}^{\mathbf{g}}_{\mathcal{A}(\mathcal{G},c)}$ by Proposition 10.1.9. The set $\mathcal{A}(\mathcal{G},c) \subseteq \mathrm{GL}_m(\mathbb{K})$ is closed under non-zero scalar multiples. Therefore, the weak correspondence, Theorem 8.2.3, holds for the RDAG model $\mathcal{M}^{\mathbf{g}}_{\mathcal{A}(\mathcal{G},c)}$. Moreover, we can always apply the weak correspondence using $\mathcal{A}(\mathcal{G},c)_{\mathrm{SL}}$ (instead of $\mathcal{A}(\mathcal{G},c)^{\pm}_{\mathrm{SL}}$). Indeed, if $\mathbb{K} = \mathbb{R}$ and α_s is even for all $s \in c(I)$, then $\mathcal{A}(\mathcal{G},c)$ only contains matrices of positive determinant, so $\mathcal{A}(\mathcal{G},c)_{\mathrm{SL}} = \mathcal{A}(\mathcal{G},c)^{\pm}_{\mathrm{SL}}$. On the other hand, if $\mathbb{K} = \mathbb{R}$ and α_s is odd for some vertex colour s, then $\mathcal{A}(\mathcal{G},c)$ contains the diagonal, orthogonal matrix t defined by $t_{s,s} := -1$ and $t_{s',s'} := 1$ for $s' \in c(I)\backslash\{s\}$. We have $ta \in \mathcal{A}(\mathcal{G},c)$ for all $a \in \mathcal{A}(\mathcal{G},c)$, by Lemma 10.1.8(iii). Therefore, condition (ii) in Theorem 8.2.3 is satisfied: choose $o(a) = \mathrm{I}_m$ if $\det(a) > 0$ and otherwise choose $o(a) = t$. ▽

Next, we link the linear independence conditions from Theorem 10.3.6 to stability notions under $\mathcal{A}(\mathcal{G},c)_{\mathrm{SL}}$. First, we prove a condition for polystability.

Lemma 10.6.2 ([MRS21, Lemma A.6]). *Consider the RDAG model on $(\mathcal{G},c)$ where colouring c is compatible, and set $\mathcal{A} := \mathcal{A}(\mathcal{G},c)_{\mathrm{SL}}$. Let $Y \in \mathbb{K}^{m\times n}$ be such that $M^{(0)}_{Y,s} \notin \operatorname{span}\{M^{(1)}_{Y,s},\dots,M^{(\beta_s)}_{Y,s}\}$ for all $s \in c(I)$. Then Y is polystable under $\mathcal{A}$ and $\mathcal{A}\cdot Y$ is Zariski closed.*

Proof. Note that the assumption on Y implies that $Y \neq 0$. To study the orbit $\mathcal{A}\cdot Y$, let T be the set of diagonal matrices in $\mathcal{A}$ and U the set of unipotent matrices in $\mathcal{A}$. We have $\mathcal{A} = T\cdot U$ and actually any $a \in \mathcal{A}$ admits a unique decomposition $a = t(a)u(a)$, where $t(a) \in T$ and $u(a) \in U$, compare Lemma 10.1.8(iv). For $s \in c(I)$, recall the construction of $M_{Y,s} \in \mathbb{K}^{(\beta_s+1)\times\alpha_s n}$ from Definition 10.3.1. Setting $V_s := \mathbb{K}^{1\times\alpha_s n}$ we can identify $\mathbb{K}^{m\times n} \cong \bigoplus_s V_s$ such that the rows of vertex colour s belong to V_s. By Equation (10.12), the set $U\cdot Y$ is $H := \prod_s H_s$ with

$$H_s = \left\{ M^{(0)}_{Y,s} + a_{s,1}M^{(1)}_{Y,s} + \ldots + a_{s,\beta_s}M^{(\beta_s)}_{Y,s} \mid a_{s,t} \in \mathbb{K} \right\}.$$

The affine space H_s equals $M^{(0)}_{Y,s} + X_s$, where $X_s := \operatorname{span}\{M^{(1)}_{Y,s},\dots,M^{(\beta_s)}_{Y,s}\}$. Since by assumption $M^{(0)}_{Y,s} \notin X_s$ for all $s \in c(I)$, we have $K_s := \left(\mathbb{K}M^{(0)}_{Y,s}\right) \oplus X_s \subseteq V_s$ and H_s has at least codimension one in V_s. Since T acts on each V_s with the non-zero scalar for vertex colour s, we have

$$\mathcal{A}\cdot Y = T\cdot(U\cdot Y) = T\cdot H = T\cdot\prod_s H_s \subseteq \bigoplus_s K_s \subseteq \bigoplus_s V_s.$$

It suffices to show that $\mathcal{A}\cdot Y$ is Zariski-closed in $\bigoplus_s K_s$. Each H_s is an affine subspace of K_s with codimension one, by definition of K_s. Therefore, there exists a linear form $p_s \in K_s^*$ such that

$$H_s = \mathbb{V}_{K_s}(p_s - 1),$$

where $\mathbb{V}(\cdot)$ denotes the vanishing locus.

We finish the proof by showing that $\mathcal{A} \cdot Y = \mathbb{V}\big(\prod_s p_s^{\alpha_s} - 1\big)$ in $\bigoplus_s K_s$. First, given $W = (W_s)_s \in \mathcal{A} \cdot Y = T \cdot H$ we can write $W = t \cdot Z$ with $t \in T$ and $Z = (Z_s)_s \in H$. Then

$$\Big(\prod_s p_s^{\alpha_s}\Big)(W) = \prod_s p_s(W_s)^{\alpha_s} = \prod_s \big(t_{ss} p_s(Z_s)\big)^{\alpha_s} = \prod_s (t_{ss})^{\alpha_s} = 1$$

by the choice of $p_s \in K_s^*$ and since $\det(t) = \prod_s t_{ss}^{\alpha_s} = 1$. On the other hand, suppose $W = (W_s)_s \in \mathbb{V}\big(\prod_s p_s^{\alpha_s} - 1\big) \subseteq \bigoplus_s K_s$. Set $t_{ss} := p_s(W_s)$, then we have $\prod_s t_{ss}^{\alpha_s} = 1$, so the t_{ss} define some $t \in T$. Moreover, $t_{ss}^{-1} W_s \in H_s$ by definition of t_{ss}, so $W' := (t_{ss}^{-1} W_s)_s \in H$. Hence, $W = t \cdot W'$ is contained in $T \cdot H = \mathcal{A} \cdot Y$. □

The upcoming theorem links stability notions under $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ to linear independence conditions. It generalizes Theorem 9.5.8 for TDAG models, compare Remark 10.3.4.

Theorem 10.6.3 ([MRS21, Proposition A.7])**.** *Consider an RDAG model on $(\mathcal{G}, c)$ with compatible colouring c and sample matrix $Y \in \mathbb{K}^{m \times n}$. Stability under $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ relates to linear independence conditions on the matrices $M_{Y,s}$:*

$$\begin{array}{lccl}
(a) & Y \text{ unstable} & \Leftrightarrow & \exists\, s \in c(I)\colon \ M_{Y,s}^{(0)} \in \operatorname{span}\big\{M_{Y,s}^{(t)} : t \in [\beta_s]\big\} \\
(b) & Y \text{ polystable} & \Leftrightarrow & \forall\, s \in c(I)\colon \ M_{Y,s}^{(0)} \notin \operatorname{span}\big\{M_{Y,s}^{(t)} : t \in [\beta_s]\big\} \\
(c) & Y \text{ stable} & \Leftrightarrow & \forall\, s \in c(I)\colon \ M_{Y,s} \text{ has full row rank.}
\end{array}$$

In particular, Y is semistable if and only if it is polystable.

Proof. The RDAG model $\mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ equals $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ by compatibility. Recall that the weak correspondence, Theorem 8.2.3, holds for $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$ using $\mathcal{A} := \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$, compare Remark 10.6.1. Therefore, part (a) and the forwards direction of (b) follows in combination with Theorem 10.3.6, while Lemma 10.6.2 gives the backwards direction of (b).

For part (c), it suffices to see that a polystable Y has a trivial stabilizing set $\mathcal{A}_Y$ if and only if for all $s \in c(I)$ the rows $M_{Y,s}^{(1)}, \ldots, M_{Y,s}^{(\beta_s)}$ are linearly independent. So let Y be polystable. By Equation (10.12), a matrix $a \in \mathcal{A}$ satisfies $aY = Y$ if and only if for all $s \in c(I)$

$$a_{s,s} M_{Y,s}^{(0)} + \sum_{t \in [\beta_s]} a_{s,t} M_{Y,s}^{(t)} = M_{Y,s}^{(0)}\,. \tag{10.17}$$

We have $M_{Y,s}^{(0)} \notin \operatorname{span}\big\{M_{Y,s}^{(i)} : i \in [\beta_s]\big\}$ for all $s \in c(I)$, by part (b) and Y being polystable. Therefore, Equation (10.17) implies $a_{s,s} = 1$ and $\sum_{t \in [\beta_s]} a_{s,t} M_{Y,s}^{(t)} = 0$.

If $M_{Y,s}^{(1)}, \ldots, M_{Y,s}^{(\beta_s)}$ are linearly independent, then (10.17) has exactly one solution, namely $a_{s,s} = 1$ and $a_{s,t} = 0$ for all $t \in [\beta_s]$. Thus, if $M_{Y,s}^{(1)}, \ldots, M_{Y,s}^{(\beta_s)}$ are linearly independent for all $s \in c(I)$, then $\mathcal{A}_Y = \{\mathrm{I}_m\}$. On the other hand, if for some $s \in c(I)$ the rows $M_{Y,s}^{(1)}, \ldots, M_{Y,s}^{(\beta_s)}$ are linearly dependent, then $\sum_{t \in [\beta_s]} a_{s,t} M_{Y,s}^{(t)} = 0$ has infinitely many solutions. Distinct solutions give distinct unipotent matrices

$u \in \mathcal{A}$ by setting $u_{s,t} := a_{s,t}$ for $t \in \mathrm{prc}(s)$, and $u_{s',t'} := 0$ for $s' \in c(I)\backslash\{s\}$, $t' \in \mathrm{prc}(s')$. By (10.12), such a unipotent matrix $u \in \mathcal{A}$ satisfies $uY = Y$, since the sets $\mathrm{prc}(s)$ are disjoint, so the $u_{s,t}$ do not affect any rows of Y with a different vertex colour. In conclusion, $\mathcal{A}_Y$ is infinite if $M_{Y,s}^{(1)}, \ldots, M_{Y,s}^{(\beta_s)}$ are linearly dependent for some $s \in c(I)$. □

Combining Theorem 10.6.3 with Theorem 10.3.6 directly yields the following.

Theorem 10.6.4 (Full Correspondence for RDAG models, [MRS21, Theorem A.2])**.** *Consider the RDAG model on $(\mathcal{G}, c)$ with compatible colouring c and sample matrix $Y \in \mathbb{K}^{m\times n}$. Then stability under $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ relates to ML estimation:*

$$\begin{array}{lccc} (a) & Y \text{ unstable} & \Leftrightarrow & \ell_Y \text{ unbounded from above} \\ (b) & Y \text{ semistable} & \Leftrightarrow & \ell_Y \text{ bounded from above} \\ (c) & Y \text{ polystable} & \Leftrightarrow & \text{MLE exists} \\ (d) & Y \text{ stable} & \Leftrightarrow & \text{MLE exists uniquely.} \end{array}$$

Theorem 10.6.4 applies to any DAG model, see Remark 10.1.10. Therefore, Theorem 10.6.4 generalizes Theorem 9.5.9 in two steps. First, it extends from *transitive* DAGs (i.e., $\mathcal{A}(\mathcal{G})$ is a group) to *all* DAGs (i.e., $\mathcal{A}(\mathcal{G})$ a set). Second, it generalizes from uncoloured DAG models to RDAG models.

Now, we link the stabilizing set $\mathcal{A}_Y$ (see Equation (8.2)) to the set of MLEs given Y. Recall that in the case of Gaussian group models given by a self-adjoint group G such a connection is made in Proposition 9.3.3. For its proof Kempf-Ness, Theorem 2.2.13(b), was crucial. To be able to adapt the proof method, the next lemma serves as a substitute of the Kempf-Ness theorem.

Lemma 10.6.5 ([MRS21, Lemma A.8])**.** *Consider the RDAG model on $(\mathcal{G}, c)$ where colouring c is compatible. For $\mathcal{A} := \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ let T and U be the set of diagonal respectively unipotent matrices in $\mathcal{A}$. If $Y \in \mathbb{K}^{m\times n}$ is polystable under $\mathcal{A}$, then the following hold:*

(a) $U \cdot Y$ contains a unique element $\widetilde{Y}$ of minimal norm.

(b) For $t \in T$ and $u \in U$, $\|tu \cdot Y\| \geq \|t \cdot \widetilde{Y}\|$ with equality if and only if $u \cdot Y = \widetilde{Y}$.

(c) Let $a, \widetilde{a} \in \mathcal{A}$ be such that $a \cdot Y$ and $\widetilde{a} \cdot Y$ are of minimal norm in $\mathcal{A} \cdot Y$. Then there is some $t \in T$ such that $t^\dagger t = \mathrm{I}_m$ and $ta \cdot Y = \widetilde{a} \cdot Y$.

Proof. We often use Lemma 10.1.8 in this proof without explicitly referencing it. Since $c(E)$ a disjoint union of the $\mathrm{prc}(s)$, $s \in c(I)$, when minimizing

$$\|uY\|^2 = \sum_{s\in c(I)} \left\| M_{uY,s}^{(0)} \right\|^2 \overset{(10.12)}{=} \sum_{s\in c(I)} \left\| M_{Y,s}^{(0)} + \sum_{t\in[\beta_s]} u_{s,t} M_{Y,s}^{(t)} \right\|^2$$

over $u \in U$ we can minimize each summand separately. For each $s \in c(I)$, the affine space $M_{Y,s}^{(0)} + \mathrm{span}\left\{ M_{Y,s}^{(t)} \colon t \in [\beta_s] \right\}$ has a unique element of minimal norm, call it M_s. Hence, $U \cdot Y$ has a unique element of minimal norm $\widetilde{Y}$, determined by $M_{\widetilde{Y},s}^{(0)} = M_s$ for all $s \in c(I)$.[17] This shows part (a).

[17] Note that there may be several $u \in U$ with $uY = \widetilde{Y}$, i.e., the uniqueness only refers to $\widetilde{Y}$.

To prove part (b), we use (the proof of) part (a) to obtain

$$\left\|M^{(0)}_{tuY,s}\right\|^2 = \left\|t_{ss}\, M^{(0)}_{uY,s}\right\|^2 = |t_{ss}|^2 \left\|M^{(0)}_{uY,s}\right\|^2 \geq |t_{ss}|^2 \left\|M^{(0)}_{\widetilde{Y},s}\right\|^2 = \left\|M^{(0)}_{t\widetilde{Y},s}\right\|^2 \quad (10.18)$$

for all $s \in c(I)$, hence $\|tu \cdot Y\| \geq \|t\widetilde{Y}\|$. The latter inequality is strict if and only if there is strict inequality in (10.18) for at least one s. By $|t_{ss}|^2 > 0$ and uniqueness of $\widetilde{Y}$, this is the case if and only if $uY \neq \widetilde{Y}$.

For (c), write $a = tu$ with $t \in T$ and $u \in U$. Since aY is of minimal norm in $\mathcal{A} \cdot Y$, we deduce $uY = \widetilde{Y}$ using (b). Thus, $aY \in T \cdot \widetilde{Y}$ and similarly $\widetilde{a}Y \in T \cdot \widetilde{Y}$. As $T \cdot \widetilde{Y} \subseteq \mathcal{A} \cdot Y$ the matrices aY and $\widetilde{a}Y$ are also of minimal norm in $T \cdot \widetilde{Y}$. Recall that $T \cong \{(t_{ss})_{s\in c(I)} \in (\mathbb{K}^\times)^{|c(I)|} \mid \prod_s t_{ss}^{\alpha_s} = 1\}$ is a diagonalizable group. In particular, T is reductive. Hence, Kempf-Ness, Theorem 2.2.13(b), for the action of T implies that there is some $t \in T$ with $t^\dagger t = \mathrm{I}_m$ and $taY = \widetilde{a}Y$. □

We finish the section with an alternative description of the set of MLEs via the stabilizing set $\mathcal{A}_Y$.

Proposition 10.6.6 ([MRS21, Proposition A.3])**.** *Fix the RDAG model on $(\mathcal{G}, c)$ with compatible colouring c and set $\mathcal{A} := \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$. Let $\lambda a^\dagger a$ be an MLE given Y, where $a \in \mathcal{A}$ and $\lambda > 0$ as in Theorem 8.2.3. Then we have a bijection*

$$\mathcal{A}_Y \to \{\textit{MLEs given } Y\}, \quad \hat{a} \mapsto \lambda(a + \hat{a} - \mathrm{I}_m)^\dagger (a + \hat{a} - \mathrm{I}_m)\,.$$

Proof. As usual, let T and U be the set of diagonal respectively unipotent matrices in $\mathcal{A}$. If $aY = Y$ for some $a \in \mathcal{A}$, then Equation (10.12) becomes

$$M^{(0)}_{Y,s} = a_{s,s} M^{(0)}_{Y,s} + \sum_{t\in[\beta_s]} a_{s,t} M^{(t)}_{Y,s}.$$

We have $M^{(0)}_{Y,s} \notin \operatorname{span}\{M^{(t)}_{Y,s} \colon t \in [\beta_s]\}$ for all $s \in c(I)$, since Y is polystable. Thus, $a_{s,s} = 1$ for all s, i.e., $a \in U$ and therefore $\mathcal{A}_Y = U_Y$. We set $N_Y := U_Y - \mathrm{I}_m$, which consists of strictly upper triangular matrices. It suffices to show that for fixed MLE $\lambda a^\dagger a$ the map

$$\begin{aligned} \varphi \colon N_Y &\to \{\text{MLEs given } Y\} \\ b &\mapsto \lambda(a+b)^\dagger (a+b) \end{aligned}$$

is well-defined and bijective. For the latter, note that $bY = 0$ for any $b \in N_Y$. Therefore, $(a + b)Y = aY$ is of minimal norm in $\mathcal{A} \cdot Y$ and thus $\varphi(b)$ is also an MLE by the weak correspondence, Theorem 8.2.3.

For surjectivity, let $\lambda\tilde{a}^\dagger\tilde{a}$ be another MLE given Y. Then aY and $\tilde{a}Y$ are of minimal norm in $\mathcal{A} \cdot Y$, hence there is some $t \in T$ with $t^\dagger t = \mathrm{I}_m$ and $aY = t\tilde{a}Y$ by Lemma 10.6.5(c). We set $b := t\tilde{a} - a$ so that $b \cdot Y = 0$ and $(\mathrm{I}_m + b)Y = Y$. By Lemma 10.1.8(iii), we have $t\tilde{a} \in \mathcal{A}$ and thus all entries of $b = t\tilde{a} - a$ obey the colouring c. Thus, we can also use Equation (10.12) for $bY = 0$:

$$0 = M^{(0)}_{bY,s} = b_{s,s} M^{(0)}_{Y,s} + \sum_{t\in[\beta_s]} b_{s,t} M^{(t)}_{Y,s}.$$

The latter implies $b_{s,s} = 0$ for all $s \in c(I)$ by polystability of Y, hence $b \in N_Y$. We compute $\varphi(b) = \lambda(t\tilde{a})^\dagger(t\tilde{a}) = \lambda\tilde{a}^\dagger\tilde{a}$ using $t^\dagger t = \mathrm{I}_m$.

To show injectivity, let $b, b' \in N_Y$ be such that $\varphi(b) = \varphi(b')$. Let $t \in T$ be defined by $t_{s,s} = \overline{a_{s,s}}/|a_{s,s}|$. Then $t^\dagger t = \mathrm{I}_m$ and thus

$$(ta + tb)^\dagger(ta + tb) = (a + b)^\dagger t^\dagger t(a + b) = (a + b)^\dagger(a + b).$$

Similarly, $(ta + tb')^\dagger(ta + tb') = (a + b')^\dagger(a + b')$. Therefore, $\varphi(b) = \varphi(b')$ implies

$$(ta + tb)^\dagger(ta + tb) = (ta + tb')^\dagger(ta + tb'). \tag{10.19}$$

Moreover, tb and tb' are strictly upper triangular and $ta \in \mathcal{A}$ has positive diagonal entries $|a_{s,s}|$, by construction of t. Hence, applying uniqueness of the Cholesky decomposition to (10.19) gives $ta + tb = ta + tb'$, and we deduce $b = b'$. □

10.7 Connections to Gaussian group models

Although many presented results on RDAGs do not need a group structure on $\mathcal{A}(\mathcal{G}, c)$ (see Equation (10.1)) we have more tools available if $\mathcal{A}(\mathcal{G}, c)$ is a group.[18] In this section we illustrate this as follows. We use Popov's Criterion from Section 2.4 to study polystability of a sample matrix Y. Moreover, we give a description of the set of MLEs in an RDAG model via the action of the stabilizer from Proposition 9.2.4. We start with the *butterfly criterion*, which characterizes when $\mathcal{A}(\mathcal{G}, c)$ is a subgroup of $\mathrm{GL}_m(\mathbb{K})$.

The Butterfly Criterion

Recall that for a DAG $\mathcal{G}$ the DAG model is $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G})}$, see Lemma 9.5.2, and in view of Gaussian group models it was natural to ask when $\mathcal{A}(\mathcal{G})$ is a group. By Proposition 9.5.4, the latter is the case if and only if $\mathcal{G}$ is transitive.

Similarly, we have seen that the RDAG model of a coloured DAG $(\mathcal{G}, c)$ with compatible colouring c equals $\mathcal{M}^{\mathrm{g}}_{\mathcal{A}(\mathcal{G},c)}$, compare Proposition 10.1.9. Thus one may ask for an analogous characterization when $\mathcal{A}(\mathcal{G}, c)$ is a group. For this, we define the concept of a butterfly graph.

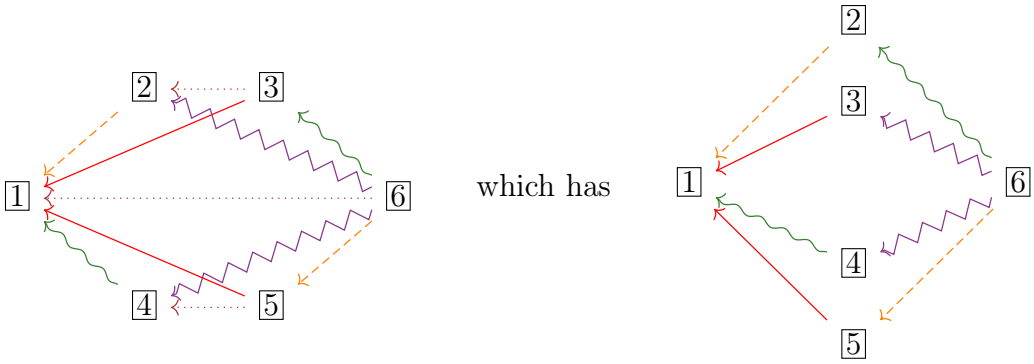
Definition 10.7.1 (Butterfly graph)**.** Let $(\mathcal{G}, c)$ be a coloured DAG. For a pair of vertices $i, j \in [m]$, define the *butterfly body* as

$$b(ij) := \big\{k \in [m] \mid i \leftarrow k, k \leftarrow j \text{ in } \mathcal{G}\big\}.$$

The *butterfly graph* $\mathcal{G}_{b(ij)}$ is defined as the coloured subgraph on $\{i\} \cup \{j\} \cup b(ij)$, with edges $i \leftarrow k, k \leftarrow j$ for each $k \in b(ij)$, and colours inherited from c. ▲

Example 10.7.2. Consider the coloured DAG

[18] In fact, Visu Makam, Anna Seigal and myself first studied RDAG models where $\mathcal{A}(\mathcal{G}, c)$ was assumed to be a group. The results on TDAG models as Gaussian group models served as a guideline and this perspective fostered our understanding to obtain many results of [MRS21].



as butterfly graph $\mathcal{G}_{b(1,6)}$. We point out that the brown (dotted) edges do not appear in the butterfly graph. ♢

We can characterize when $\mathcal{A}(\mathcal{G}, c)$ is a group via the butterfly graphs.

Proposition 10.7.3 (Butterfly Criterion [MRS21, Proposition B.2])**.** *Consider the RDAG model on $(\mathcal{G}, c)$ where colouring c is compatible. The set $\mathcal{A}(\mathcal{G}, c)$ is a group if and only if*

(a) $\mathcal{G}$ is transitive; and

(b) if $c(ij) = c(kl)$ for edges $j \to i$, $l \to k$ in $\mathcal{G}$, then $\mathcal{G}_{b(ij)} \simeq \mathcal{G}_{b(kl)}$.

Remark 10.7.4. Given a DAG $\mathcal{G}$, we know from Remark 10.1.10 that there is a compatible colouring c on $\mathcal{G}$ such that $\mathcal{M}^{\rightarrow}_{\mathcal{G}} = \mathcal{M}^{\rightarrow}_{(\mathcal{G},c)}$ and $\mathcal{A}(\mathcal{G}) = \mathcal{A}(\mathcal{G}, c)$. Since this colouring c assigns to each edge its own *distinct* colour, item (b) of Proposition 10.7.3 is trivially satisfied. Thus, the Butterfly Criterion contains Proposition 9.5.4 as a special case. ▽

Proof of Proposition 10.7.3. By definition in Equation (10.1), $\mathrm{I}_m \in \mathcal{A}(\mathcal{G}, c)$ and there is a $\mathbb{K}$-linear subspace $L \subseteq \mathbb{K}^{m\times m}$ such that $\mathcal{A}(\mathcal{G}, c) = L \cap \mathrm{GL}_m(\mathbb{K})$. Hence, by Lemma 9.5.3 $\mathcal{A}(\mathcal{G}, c)$ is a subgroup of $\mathrm{GL}_m(\mathbb{K})$ if and only if $\mathcal{A}(\mathcal{G}, c)$ is closed under multiplication. We have $gh \in \mathcal{A}(\mathcal{G}, c)$ for $g, h \in \mathcal{A}(\mathcal{G}, c)$ if and only if

(1) $(gh)_{ii} = (gh)_{jj}$ whenever $c(i) = c(j)$;

(2) $(gh)_{ij} = (gh)_{kl}$ whenever $j \to i$, $l \to k$ in $\mathcal{G}$ have $c(ij) = c(kl)$; and

(3) $(gh)_{ij} = 0$ whenever $j \not\to i$ in $\mathcal{G}$.

For (1), observe that $(gh)_{ii} = g_{ii}h_{ii}$. Thus, if $c(i) = c(j)$ then $(gh)_{ii} = (gh)_{jj}$. For (2), take $j \to i$, $l \to k$ in $\mathcal{G}$ with $c(ij) = c(kl)$. Then

$$(gh)_{ij} = g_{ii}h_{ij} + g_{ij}h_{jj} + \sum_{p\in b(ij)} g_{ip}h_{pj} \quad \text{and} \quad (gh)_{kl} = g_{kk}h_{kl} + g_{kl}h_{ll} + \sum_{q\in b(kl)} g_{kq}h_{ql},$$

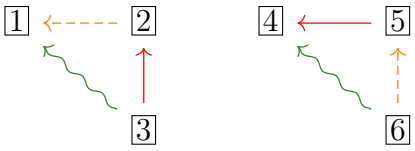
hence $(gh)_{ij} = (gh)_{kl}$ if $\mathcal{G}_{b(ij)} \simeq \mathcal{G}_{b(kl)}$. Conversely, assume $(gh)_{ij} = (gh)_{kl}$ as a polynomial identity in the unknown entries of matrices g and h. As colouring c is compatible and $c(ij) = c(kl)$, we have $c(i) = c(k)$, so $g_{ii}h_{ij} = g_{kk}h_{kl}$. Vertex and edge colours are disjoint and the sums over $b(ij)$ and $b(kl)$ only involve edge

colours. Thus, $(gh)_{ij} = (gh)_{kl}$ implies $g_{ij}h_{jj} = g_{kl}h_{ll}$, so $h_{jj} = h_{ll}$, and the sum over $b(ij)$ must equal the sum over $b(kl)$. This means $c(j) = c(l)$, and the two collections $(c(ip), c(pj)), p \in b(ij)$ and $(c(kq), c(ql)), q \in b(kl)$ of *ordered* pairs[19] counted with multiplicity agree. Compatibility ensures the correct colours on the vertices in $b(ij)$ and $b(kl)$ as well, hence $\mathcal{G}_{b(ij)} \simeq \mathcal{G}_{b(kl)}$.

For (3), we observe that if $j \not\to i$ in $\mathcal{G}$ then $g_{ij} = 0 = h_{ij}$ and therefore $(gh)_{ij} = \sum_{p \in b(ij)} g_{ip}h_{pj}$. The latter is zero for all $g, h \in A(\mathcal{G}, c)$ if and only if $b(ij) = \emptyset$. Thus, condition (3) is equivalent to the following: if $j \not\to i$ in $\mathcal{G}$, then there does not exist $p \in I$ with $j \to p$ and $p \to i$ in $\mathcal{G}$, i.e., $\mathcal{G}$ must be transitive by contraposition. We have shown that (1), (2) and (3) are satisfied if and only if conditions (a) and (b) hold. □

The following example illustrates that the *order* of the colours $c(i \leftarrow k)$ and $c(k \leftarrow j)$ for $k \in b(ij)$ in the butterfly graph $\mathcal{G}_{b(ij)}$ indeed matters.

Example 10.7.5. Consider the coloured TDAG $(\mathcal{G}, c)$ given by



The colouring is compatible as all vertices are black (squared). The butterfly graphs $\mathcal{G}_{b(1,3)}$ and $\mathcal{G}_{b(4,6)}$ for the green (squiggly) edges are

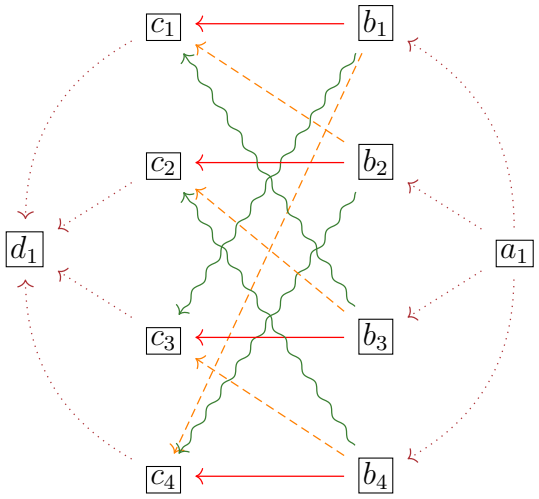
respectively. Due to the different order of red (solid) and orange (dashed) arrows the butterfly graphs $\mathcal{G}_{b(1,3)}$ and $\mathcal{G}_{b(4,6)}$ are not isomorphic. Thus, $\mathcal{A}(\mathcal{G}, c)$ is not a group by the Butterfly Criterion. This can also be checked by hand. Consider the block-diagonal matrices $a := \operatorname{diag}(M_1, M_2)$, $b := \operatorname{diag}(M_2, M_1) \in \mathcal{A}(\mathcal{G}, c)$, where

$$M_1 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad M_2 := \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus, a has entry one for red (solid) and entry zero for orange (dashed) and green (squiggly), while b has entry one for orange (dashed) and entry zero for red (solid) and green (squiggly). We compute $(ab)_{1,3} = (M_1M_2)_{1,3} = 0$ and $(ab)_{4,6} = (M_2M_1)_{1,3} = 1$. Therefore, the matrix ab violates the green (squiggly) colour condition. Hence, $ab \notin \mathcal{A}(\mathcal{G}, c)$ and so $\mathcal{A}(\mathcal{G}, c)$ is not a group. ♢

Example 10.7.6 ([MRS21, Example B.3])**.** Interestingly, two graphs can have all the same butterfly graphs without being isomorphic. We present an example. Consider the following coloured graph with 10 black (square) vertices, and edges that are red (solid), green (squiggly), orange (dashed) or brown (dotted).

[19] The order matters, since the variables in the entries of g are distinct from those in h. Also compare Example 10.7.5 for an illustration.



We add some further edges: four purple edges $a_1 \to c_i$, four blue edges $b_i \to d_1$, and a yellow edge $a_1 \to d_1$. Now, additionally consider the graph obtained by exchanging the green (squiggly) and orange (dashed) edges.

The butterfly graphs for the two graphs are the same, as follows. On the yellow edge, the butterfly graphs both have four paths consisting of a brown edge followed by a blue edge, and four that are a purple edge followed by a brown edge. Similarly, we can check the butterfly graphs at the other edge colours.

However, the two coloured graphs are not isomorphic. Indeed, the only way to get an isomorphism is to permute the b-layer and the c-layer. The red (solid) edges give the identity permutation, the orange (dashed) edges give the cycle $\sigma = (1\ 4\ 3\ 2)$, and the green (squiggly) edges give σ^2. Hence an isomorphism would need to consist of permutations τ_1 and τ_2 of $\{1,2,3,4\}$ with $\tau_1 \mathrm{id} \tau_2 = \mathrm{id}$, $\tau_1 \sigma \tau_2 = \sigma^2$, $\tau_1 \sigma^2 \tau_2 = \sigma$. The first condition implies $\tau_2 = \tau_1^{-1}$, hence σ and σ^2 need to be simultaneously conjugate to σ^2 and σ respectively. This implies $(\sigma^2)^2 = \sigma$, a contradiction because $\sigma^4 = \mathrm{id}$. ♢

Popov's Criterion for RDAGs

If $\mathcal{A}(\mathcal{G}, c)$ is a group we can prove the important Lemma 10.6.2 on polystability differently. Namely, we generalize the proof of Theorem 9.5.8(b) for TDAG models, where we used Popov's Criterion, Theorem 2.4.1. We stress that during the work on [MRS21] this generalization process led to the concept of augmented sample matrices $M_{Y,s}$, which are crucial for several main results on RDAG models. It illustrates once more how the invariant theory perspective can foster the statistical understanding.

Let $(\mathcal{G}, c)$ be a coloured DAG with compatible colouring. Recall that it suffices to work over $\mathbb{C}$ when using Popov's Criterion, compare Lemma 2.4.3. Assume that $\mathcal{A}(\mathcal{G}, c) \subseteq \mathrm{GL}_m(\mathbb{C})$ is a group. Hence, $G := \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ is a group as well and we denote the subgroup of diagonal matrices in G by T. Then T is isomorphic to the diagonalizable group $\{(\lambda_{s,s})_s \in (\mathbb{C}^\times)^{|c(I)|} \mid \prod_s \lambda_{s,s}^{\alpha_s} = 1\}$.

We briefly recall the setting of Section 2.4 for the special case of RDAGs. The group G acts on $\mathbb{C}^{m \times n}$ by left-multiplication and $x_{i,j} \in \mathbb{C}[G]$, $i, j \in [m]$ are the coordinate functions on G. By compatibility and similarly to Lemma 10.1.8, we

can consider the coordinate functions for the colour entries $z_{s,s}$ and $z_{s,t}$, where $s \in c(I)$ and $t \in \mathrm{prc}(s)$. They capture the equalities among the $x_{i,j}$, i.e., $z_{s,s} = x_{i,i}$ whenever $c(i) = s$ and $z_{s,t} = x_{i,j}$ whenever $c(ij) = (s,t)$. For $Y \in \mathbb{C}^{m\times n}$, we recall from Equation (2.28) the $\mathbb{C}$-algebra

$$R_Y = \mathbb{C}\Big[\sum_{j=1}^{m} Y_{j,l}\, x_{i,j} \,\Big|\, i \in [m], l \in [n]\Big] \subseteq \mathbb{C}[G].$$

Using the equalities among the $x_{i,j}$ and that $x_{i,j} = 0$ if $j \notin \{i\} \cup \mathrm{pa}(i)$, we can rewrite the algebra generators of R_Y as follows:

$$\sum_{j=1}^{m} x_{i,j}Y_{j,l} = z_{c(ii)}Y_{i,l} + \sum_{j\in \mathrm{pa}(i)} z_{c(ij)}Y_{j,l} = z_{s,s}Y_{i,l} + \sum_{t=1}^{\beta_s} z_{s,t}\Big(\sum_{\substack{i\leftarrow j\\ c(ij)=t}} Y_{j,l}\Big), \quad (10.20)$$

where $s := c(i)$. The character group of T is $\mathfrak{X}(T) \cong \mathbb{Z}^{|c(I)|}/\big(\mathbb{Z}\cdot(\alpha_s)_{s\in c(I)}\big)$, so that the semigroup $\mathfrak{X}_{G\cdot Y}$ can be written as

$$\mathfrak{X}_{G\cdot Y} = \Big\{(d_s)_{s\in c(I)} \in \mathfrak{X}(T) \,\Big|\, \prod_{s\in c(I)} z_{s,s}^{d_s} \in R_Y\Big\}.$$

Remark 10.7.7 ([MRS21, Remark B.5]). The group $G = \mathcal{A}(\mathcal{G},c)_{\mathrm{SL}} \subseteq \mathrm{GL}_m(\mathbb{C})$ may not be connected as required in Popov's Criterion, Theorem 2.4.1. However, the orbit $G\cdot Y$ is Zariski-closed if $G^\circ \cdot Y$ is Zariski-closed, where G° is the identity component of G.[20] Thus, after restricting to $G^\circ = T^\circ U$ we may assume that G is connected. Restricting to T° amounts to restricting to the torsion-free part of $\mathfrak{X}(T)$, compare Proposition 1.1.17(c). If α is the greatest common divisor of all $\alpha_s, s \in c(I)$, then $T^\circ \cong \big\{(g_s)_{s\in c(I)} \mid \prod_s g_s^{\alpha_s/\alpha} = 1\big\}$ and $\mathfrak{X}(T^\circ) = \mathbb{Z}^{|c(I)|}/\big(\mathbb{Z}\cdot(\alpha_s/\alpha)_{s\in c(I)}\big)$. ▽

We are ready to generalize the proof of Theorem 9.5.8(b) to the RDAG situation. This reproves the part on polystability in Lemma 10.6.2.[21]

Lemma 10.7.8. *Consider the RDAG model on $(\mathcal{G},c)$ where colouring c is compatible. Assume $\mathcal{A}(\mathcal{G},c) \subseteq \mathrm{GL}_m(\mathbb{K})$ is a group and set $G := \mathcal{A}(\mathcal{G},c)_{\mathrm{SL}}$. Let $Y \in \mathbb{K}^{m\times n}$ be such that $M_{Y,s}^{(0)} \notin \mathrm{span}\big\{M_{Y,s}^{(1)},\dots,M_{Y,s}^{(\beta_s)}\big\}$ for all $s \in c(I)$. Then Y is polystable under G.*

Proof. The assumption ensures that $Y \neq 0$, so we need to show that $G\cdot Y$ is Euclidean closed in $\mathbb{K}^{m\times n}$. By Lemma 2.4.3, it is enough to prove that $G\cdot Y$ is Zariski closed for $\mathbb{K} = \mathbb{C}$. We will use Popov's Criterion for this.

Fix a vertex colour $s \in c(I)$. Since $M_{Y,s}^{(0)} \notin \mathrm{span}\big\{M_{Y,s}^{(1)},\dots,M_{Y,s}^{(\beta_s)}\big\}$, we can apply Lemma 9.5.7 to the matrix $M_{Y,s} \in \mathbb{C}^{(1+\beta_s)\times\alpha_s n}$, and for $z_{s,s}$ and the $z_{s,t}$, $t \in [\beta_s]$. Hence, there is $w \in \mathbb{C}^{\alpha_s n}$ with

$$z_{s,s} = \sum_{p=1}^{\alpha_s n} w_p \left(z_{s,s}\big(M_{Y,s}\big)_{0,p} + \sum_{t=1}^{\beta_s} z_{s,t}\big(M_{Y,s}\big)_{t,p}\right). \quad (10.21)$$

[20] Recall that Zariski and Euclidean identity component agree over $\mathbb{C}$, compare Section 1.1

[21] For $\mathbb{K} = \mathbb{R}$ the argument only ensures that the orbit is Euclidean closed.

Fix some $p \in [\alpha_s n]$. By the construction of $M_{Y,s}$ in Definition 10.3.1, there exists a vertex $i = i(p)$ of colour s and some $l = l(p) \in [n]$ such that

$$z_{s,s}\big(M_{Y,s}\big)_{0,p} + \sum_{t=1}^{\beta_s} z_{s,t}\big(M_{Y,s}\big)_{t,p} = z_{s,s}Y_{i,l} + \sum_{t=1}^{\beta_s} z_{s,t}\Big(\sum_{\substack{i\leftarrow j \\ c(ij)=t}} Y_{j,l}\Big) \overset{(10.20)}{=} \sum_{j=1}^{m} x_{i,j}Y_{j,l}\,.$$

Thus, (10.21) shows that $z_{s,s}$ is a $\mathbb{C}$-linear combination of the $\sum_{j=1}^{m} x_{i,j}Y_{j,l}$, where $i \in c^{-1}(s)$ and $l \in [n]$. In particular, $z_{s,s} \in R_Y$ for all $s \in c(I)$ and hence we have

$$\forall (d_s)_s \in \mathbb{Z}_{\geq 0}^{|c(I)|}\colon \qquad \prod_{s\in c(I)} z_{s,s}^{d_s} \in R_Y.$$

Any character of T is of the latter form, since $\prod_s z_{ss}^{\alpha_s}$ is the trivial character.[22] We conclude $\mathfrak{X}_{G\cdot Y} = \mathfrak{X}(T)$ and hence $\mathfrak{X}_{G\cdot Y}$ is a group. Therefore, $G \cdot Y$ is Zariski closed by Popov's Criterion, Theorem 2.4.1. □

Bijection between the Stabilizer and the Set of MLEs

So far we have given two descriptions of the MLEs given Y in an RDAG model. Corollary 10.3.9 gives a linear space of possible Λ, while Proposition 10.6.6 gives an additive bijection between the set of MLEs and the $\mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$-stabilizing set.

Here we give an alternative (multiplicative) bijection. Namely, for a Gaussian group model $\mathcal{M}_G^{\mathrm{g}}$ we have a natural action of the G_{SL}-stabilizer of Y on the set of MLEs given Y, compare Proposition 9.2.4. For Zariski closed self-adjoint groups we have seen in Proposition 9.3.3 that this action is transitive. In the RDAG case the action is even transitive *and* free. The following statement contains Proposition 9.5.10 as a special case, since any TDAG $\mathcal{G}$ arises as a coloured DAG $(\mathcal{G}, c)$ with compatible colouring such that the group $\mathcal{A}(\mathcal{G})$ equals $\mathcal{A}(\mathcal{G}, c)$, see Remark 10.1.10.

Proposition 10.7.9 ([MRS21, Proposition B.6])**.** *Consider the RDAG model on $(\mathcal{G}, c)$ where colouring c is compatible and assume $\mathcal{A}(\mathcal{G}, c)$ is a group. Set $\mathcal{A} := \mathcal{A}(\mathcal{G}, c)_{\mathrm{SL}}$ and let $Y \in \mathbb{K}^{m\times n}$ be polystable under $\mathcal{A}$. Let $\lambda a^\dagger a$ be an MLE given Y, where $a \in \mathcal{A}$ and $\lambda \in \mathbb{R}_{>0}$ are as in Theorem 8.2.3. We have a bijection*

$$\begin{aligned} \varphi\colon \mathcal{A}_Y &\to \{\textit{MLEs given } Y\} \\ g &\mapsto \lambda g^\dagger a^\dagger a g. \end{aligned}$$

In other words, $\mathcal{A}_Y$ acts freely and transitively on the set of MLEs given Y.

Proof. For $g \in \mathcal{A}_Y$ we have $agY = aY$, which is of minimal norm in $\mathcal{A}\cdot Y$ as $\lambda a^\dagger a$ is an MLE. Hence, $\varphi(g) = \lambda(ag)^\dagger(ag)$ is another MLE given Y, by Theorem 8.2.3, and we see that φ is well-defined.

For surjectivity, let $\lambda\tilde{a}^\dagger\tilde{a}$ be another MLE given Y. Then aY and $\tilde{a}Y$ are of minimal norm in $\mathcal{A}\cdot Y$, hence there is some $t \in T$ with $t^\dagger t = \mathrm{I}_m$ such that

[22] In other words, any element of $\mathfrak{X}(T) \cong \mathbb{Z}^{|c(I)|}/\big(\mathbb{Z}\cdot(\alpha_s)_{s\in c(I)}\big)$ admits a representative with non-negative entries.

$taY = \tilde{a}Y$, by Lemma 10.6.5(c). Thus, for $g := a^{-1}t^{-1}\tilde{a}$ we have $gY = Y$ and also $g \in \mathcal{A}$, since $\mathcal{A}(\mathcal{G}, c)$ (and hence $\mathcal{A}$) is a group. Hence, $g \in \mathcal{A}_Y$ and the property $t^\dagger t = \mathrm{I}_m$ gives $\varphi(g) = \lambda g^\dagger a^\dagger a g = \lambda \tilde{a}^\dagger \tilde{a}$.

To prove injectivity, let $g, g' \in \mathcal{A}_Y$ be such that $\varphi(g) = \varphi(g')$. The latter implies $g^\dagger a^\dagger a g = g'^\dagger a^\dagger a g'$, which is equivalent to $h^\dagger h = \mathrm{I}_m$ where $h := ag'g^{-1}a^{-1}$. In the following we show that $h = \mathrm{I}_m$ which implies $g = g'$ as desired.

First, as $\mathcal{A}$ is a group we have $h \in \mathcal{A}$. In particular, h is upper triangular. Together with $h^\dagger h = \mathrm{I}_m$, h is a diagonal matrix by Lemma 10.7.10 below. Moreover, using $g, g' \in \mathcal{A}_Y$ we deduce $haY = aY$, i.e., $h \in \mathcal{A}_{aY}$. Note that Y and aY have the same orbit (closure), where we again use that $\mathcal{A}$ is a group. Thus, aY is polystable as Y is polystable. In particular, for all vertex colours s we must have $M^{(\check{0})}_{aY,s} \neq 0$ by Theorem 10.6.3(b). Finally, combining the latter with Eq. (10.12) for $M^{(0)}_{h\cdot(aY)}$, $h(aY) = aY$ and h being diagonal implies $h = \mathrm{I}_m$. □

We are left to show the following lemma.

Lemma 10.7.10. *Let $h \in \mathrm{GL}_m(\mathbb{K})$ be upper triangular with $h^\dagger h = \mathrm{I}_m$. Then h is a diagonal matrix.*

Proof. We prove the statement by induction on $m \geq 1$. For $m = 1$, there is nothing to show as any 1×1 matrix is diagonal. Now, assume the statement holds for a fixed $m \geq 1$ and let $h \in \mathrm{GL}_{m+1}(\mathbb{K})$ be upper triangular with $h^\dagger h = \mathrm{I}_{m+1}$. Then we have for all $j \in [m+1]$ that

$$\left(h^\dagger h\right)_{1,j} = \sum_{k=1}^{m+1} \overline{h_{k,i}}\, h_{k,j} = \overline{h_{1,1}}\, h_{1,j} = \begin{cases} 1 & \text{, if } j = 1 \\ 0 & \text{, if } j \neq 1 \end{cases}$$

where we used in the middle equality that h is upper triangular. We deduce that $h_{1,1} \neq 0$ and consequently for $j \geq 2$ we must have $h_{1,j} = 0$. Hence, h is a block-diagonal matrix of the form $\mathrm{diag}(1, g)$ with $g \in \mathrm{GL}_m(\mathbb{K})$. The properties of h yield that g must be upper triangular with $g^\dagger g = \mathrm{I}_m$. By induction hypothesis, g is diagonal and therefore h as well. □

Bibliography

[AB22] J. M. Altschuler and E. Boix-Adsera. "Polynomial-time algorithms for multimarginal optimal transport problems with structure". In: *Mathematical Programming* (2022), pp. 1–72. DOI: 10.1007/s10107-022-01868-7.

[AFS16] A. Abbruzzo, V. Fasone, and R. Scuderi. "Operational and financial performance of Italian airport companies: A dynamic graphical model". In: *Transport Policy* 52 (2016), pp. 231–237. DOI: 10.1016/j.tranpol.2016.09.004.

[AGL+18] Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson. "Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 172–181. DOI: 10.1145/3188745.3188942.

[AGL+21] J. van Apeldoorn, S. Gribling, Y. Li, H. Nieuwboer, M. Walter, and R. de Wolf. "Quantum algorithms for matrix scaling and matrix balancing". In: *48th International Colloquium on Automata, Languages, and Programming*. Vol. 198. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2021, Art. No. 110, 17. DOI: 10.4230/LIPIcs.ICALP.2021.110.

[AHSE95] H. H. Andersen, M. Højbjerre, D. Sørensen, and P. S. Eriksen. *Linear and graphical models*. Vol. 101. Lecture Notes in Statistics. For the multivariate complex normal distribution. Springer-Verlag, New York, 1995, pp. x+183. ISBN: 0-387-94521-0. DOI: 10.1007/978-1-4612-4240-6.

[AKRS21a] C. Améndola, K. Kohn, P. Reichenbach, and A. Seigal. "Invariant Theory and Scaling Algorithms for Maximum Likelihood Estimation". In: *SIAM J. Appl. Algebra Geom.* 5.2 (2021), pp. 304–337. DOI: 10.1137/20M1328932.

[AKRS21b] C. Améndola, K. Kohn, P. Reichenbach, and A. Seigal. "Toric invariant theory for maximum likelihood estimation in log-linear models". In: *Algebr. Stat.* 12.2 (2021), pp. 187–211. ISSN: 2693-2997. DOI: 10.2140/astat.2021.12.187.

[AM98] S. Andersson and J. Madsen. "Symmetry and lattice conditional independence in a multivariate normal distribution". In: *Ann. Statist.* 26.2 (1998), pp. 525–572. ISSN: 0090-5364. DOI: 10.1214/aos/1028144848.

[AMN+22] A. Acuaviva et al. *The minimal canonical form of a tensor network*. 2022. DOI: 10.48550/ARXIV.2209.14358.

[AMP97] S. A. Andersson, D. Madigan, and M. D. Perlman. "On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs". In: *Scand. J. Statist.* 24.1 (1997), pp. 81–102. DOI: 10.1111/1467-9469.00050.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. With a foreword by Paul Van Dooren. Princeton University Press, Princeton, NJ, 2008, pp. xvi+224. ISBN: 978-0-691-13298-3. DOI: 10.1515/9781400830244.

[And03] T. W. Anderson. *An introduction to multivariate statistical analysis*. Third. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, 2003, pp. xx+721. ISBN: 0-471-36091-0.

[Ati82] M. F. Atiyah. "Convexity and commuting Hamiltonians". In: *Bull. London Math. Soc.* 14.1 (1982), pp. 1–15. ISSN: 0024-6093. DOI: 10.1112/blms/14.1.1. URL: https://doi.org/10.1112/blms/14.1.1.

[AV97] N. Alon and V. H. Vũ. "Anti-Hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs". In: *J. Combin. Theory Ser. A* 79.1 (1997), pp. 133–160. DOI: 10.1006/jcta.1997.2780.

[Bač14] M. Bačák. *Convex analysis and optimization in Hadamard spaces*. Vol. 22. De Gruyter Series in Nonlinear Analysis and Applications. De Gruyter, Berlin, 2014, pp. viii+185. ISBN: 978-3-11-036103-2. DOI: 10.1515/9783110361629.

[Bar83] O. Barndorff-Nielsen. "On a formula for the distribution of the maximum likelihood estimator". In: *Biometrika* 70.2 (1983), pp. 343–365. ISSN: 0006-3444. DOI: 10.1093/biomet/70.2.343.

[BBJJ82] O. Barndorff-Nielsen, P. Blæsild, J. L. Jensen, and B. Jørgensen. "Exponential transformation models". In: *Proc. Roy. Soc. London Ser. A* 379.1776 (1982), pp. 41–65. DOI: 10.1098/rspa.1982.0004.

[BC13] P. Bürgisser and F. Cucker. *Condition*. Vol. 349. Grundlehren der mathematischen Wissenschaften. Springer, Heidelberg, 2013, pp. xxxii+554. ISBN: 978-3-642-38895-8. DOI: 10.1007/978-3-642-38896-5.

[BCR98] J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*. Vol. 36. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Translated from the 1987 French original, Revised by the authors. Springer-Verlag, Berlin, 1998, pp. x+430. ISBN: 3-540-64663-9. DOI: 10.1007/978-3-662-03718-8.

[BD06] M. Bürgin and J. Draisma. "The Hilbert null-cone on tuples of matrices and bilinear forms". In: *Math. Z.* 254.4 (2006), pp. 785–809. DOI: 10.1007/s00209-006-0008-0.

[BDM+21] P. Bürgisser, M. L. Doğan, V. Makam, M. Walter, and A. Wigderson. "Polynomial Time Algorithms in Invariant Theory for Torus Actions". In: *36th Computational Complexity Conference*. Vol. 200. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2021, Art. No. 32, 30. DOI: 10.4230/LIPIcs.CCC.2021.32.

[BDM+23] P. Bürgisser, M. L. Doğan, V. Makam, M. Walter, and A. Wigderson. *Robust orbit problems for torus actions and the abc-conjecture*. ongoing work; the title may change. 2023.

[BFG+18] P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson. "Efficient algorithms for tensor scaling, quantum marginals, and moment polytopes". In: *59th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2018*. IEEE Computer Soc., Los Alamitos, CA, 2018, pp. 883–897. DOI: 10.1109/FOCS.2018.00088. URL: https://arxiv.org/abs/1804.04739.

[BFG+19] P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter, and A. Wigderson. *Towards a theory of non-commutative optimization: geodesic 1st and 2nd order methods for moment maps and polytopes*. 2019. URL: https://arxiv.org/abs/1910.12375v3.

[BFH07] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice*. With the collaboration of Richard J. Light and Frederick Mosteller, Reprint of the 1975 original. Springer, New York, 2007, pp. viii+557. ISBN: 978-0-387-72805-6.

[BGO+18] P. Bürgisser, A. Garg, R. Oliveira, M. Walter, and A. Wigderson. "Alternating Minimization, Scaling Algorithms, and the Null-Cone Problem from Invariant Theory". In: *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Vol. 94. Leibniz International Proceedings in Informatics (LIPIcs). 2018, 24:1–24:20. ISBN: 978-3-95977-060-6. DOI: 10.4230/LIPIcs.ITCS.2018.24. URL: https://arxiv.org/abs/1711.08039.

[BH62] A. Borel and Harish-Chandra. "Arithmetic subgroups of algebraic groups". In: *Ann. of Math. (2)* 75 (1962), pp. 485–535. DOI: 10.2307/1970210.

[BH99] M. R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*. Vol. 319. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999, pp. xxii+643. ISBN: 3-540-64324-9. DOI: 10.1007/978-3-662-12494-9.

[Bha07] R. Bhatia. *Positive definite matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2007, pp. x+254. ISBN: 978-0-691-12918-1. DOI: 10.1515/9781400827787.

[Bil21] L. Biliotti. "The Kempf-Ness theorem and invariant theory for real reductive representations". In: *São Paulo J. Math. Sci.* 15.1 (2021), pp. 54–74. DOI: 10.1007/s40863-019-00151-6.

[Bir63] M. W. Birch. "Maximum likelihood in three-way contingency tables". In: *J. Roy. Statist. Soc. Ser. B* 25 (1963), pp. 220–233. ISSN: 0035-9246. URL: https://www.jstor.org/stable/2984562.

[Bir71] D. Birkes. "Orbits of linear algebraic groups". In: *Ann. of Math. (2)* 93 (1971), pp. 459–475. ISSN: 0003-486X. DOI: 10.2307/1970884.

[BKVH07] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi. "A tutorial on geometric programming". In: *Optimization and engineering* 8.1 (2007), pp. 67–127. DOI: 10.1007/s11081-007-9001-7.

[BL21] C. Böhm and R. A. Lafuente. "Real geometric invariant theory". In: *Differential geometry in the large*. Vol. 463. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, 2021, pp. 11–49. DOI: 10.1017/9781108884136.003.

[BL76] H. J. Brascamp and E. H. Lieb. "Best constants in Young's inequality, its converse, and its generalization to more than three functions". In: *Advances in Math.* 20.2 (1976), pp. 151–173. ISSN: 0001-8708. DOI: 10.1016/0001-8708(76)90184-5.

[BLNW20] P. Bürgisser, Y. Li, H. Nieuwboer, and M. Walter. *Interior-point methods for unconstrained geometric programming and scaling problems*. 2020. URL: https://arxiv.org/abs/2008.12110v1.

[Bor06] A. Borel. "Lie groups and linear algebraic groups. I. Complex and real groups". In: *Lie groups and automorphic forms*. Vol. 37. AMS/IP Stud. Adv. Math. Amer. Math. Soc., Providence, RI, 2006, pp. 1–49. DOI: 10.1090/amsip/037/01.

[Bor91] A. Borel. *Linear algebraic groups*. Second. Vol. 126. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xii+288. ISBN: 0-387-97370-2. DOI: 10.1007/978-1-4612-0941-6.

[Bou23] N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, Cambridge, 2023, pp. xviii+338. ISBN: 978-1-009-16617-1. DOI: 10.1017/9781009166164.

[Bri87] M. Brion. "Sur l'image de l'application moment". In: *Séminaire d'algèbre Paul Dubreil et Marie-Paule Malliavin (Paris, 1986)*. Vol. 1296. Lecture Notes in Math. Springer, Berlin, 1987, pp. 177–192. DOI: 10.1007/BFb0078526.

[BS19] G. Blekherman and R. Sinn. "Maximum likelihood threshold and generic completion rank of graphs". In: *Discrete Comput. Geom.* 61.2 (2019), pp. 303–324. ISSN: 0179-5376. DOI: 10.1007/s00454-018-9990-3.

[Buc06] B. Buchberger. "An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal". In: *J. Symbolic Comput.* 41.3-4 (2006). Translated from the 1965 German original by Michael P. Abramson, pp. 475–511. ISSN: 0747-7171. DOI: 10.1016/j.jsc.2005.09.007.

[Buc70] B. Buchberger. "Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems". In: *Aequationes Math.* 4 (1970), pp. 374–383. ISSN: 0001-9054. DOI: 10.1007/BF01844169.

[Buh93] S. L. Buhl. "On the existence of maximum likelihood estimators for graphical Gaussian models". In: *Scand. J. Statist.* 20.3 (1993), pp. 263–270. ISSN: 0303-6898. URL: https://www.jstor.org/stable/4616281.

[CLS11] D. A. Cox, J. B. Little, and H. K. Schenck. *Toric varieties*. Vol. 124. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2011, pp. xxiv+841. ISBN: 978-0-8218-4819-7. DOI: 10.1090/gsm/124.

[CMTV17] M. B. Cohen, A. Madry, D. Tsipras, and A. Vladu. "Matrix scaling and balancing via box constrained Newton's method and interior point methods". In: *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*. IEEE Computer Soc., Los Alamitos, CA, 2017, pp. 902–913. DOI: 10.1109/FOCS.2017.88.

[CR81] R. J. Carroll and D. Ruppert. "On prediction and the power transformation family". In: *Biometrika* 68.3 (1981), pp. 609–615. DOI: 10.2307/2335443.

[Cra46] H. Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946, pp. xvi+575. DOI: 10.1515/9781400883868.

[Cra86] J. S. Cramer. *Econometric applications of maximum likelihood methods*. Cambridge University Press, Cambridge, 1986, pp. xiv+208. ISBN: 0-521-25317-9. DOI: 10.1017/CBO9780511572050.

[Cra98] E. Cramer. "Conditional iterative proportional fitting for Gaussian distributions". In: *J. Multivariate Anal.* 65.2 (1998), pp. 261–276. ISSN: 0047-259X. DOI: 10.1006/jmva.1998.1739.

[Cut13] M. Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger. Vol. 26. Curran Associates, Inc., 2013, pp. 2292–2300. URL: https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[CVZ23] M. Christandl, P. Vrana, and J. Zuiddam. "Universal points in the asymptotic spectrum of tensors". In: *J. Amer. Math. Soc.* 36.1 (2023), pp. 31–79. ISSN: 0894-0347. DOI: 10.1090/jams/996.

[DC70] J. A. Dieudonné and J. B. Carrell. "Invariant theory, old and new". In: *Advances in Math.* 4 (1970), 1–80 (1970). ISSN: 0001-8708. DOI: 10.1016/0001-8708(70)90015-0.

[Der99] H. Derksen. "Computation of invariants for reductive groups". In: *Adv. Math.* 141.2 (1999), pp. 366–384. ISSN: 0001-8708. DOI: 10.1006/aima.1998.1787.

[DFKP19] M. Drton, C. Fox, A. Käufl, and G. Pouliot. "The maximum likelihood threshold of a path diagram". In: *Ann. Statist.* 47.3 (2019), pp. 1536–1553. DOI: 10.1214/18-AOS1724.

[DK15] H. Derksen and G. Kemper. *Computational invariant theory*. enlarged. Vol. 130. Encyclopaedia of Mathematical Sciences. With two appendices by Vladimir L. Popov, and an addendum by Norbert A'Campo and Popov, Invariant Theory and Algebraic Transformation Groups, VIII. Springer, Heidelberg, 2015, pp. xxii+366. ISBN: 978-3-662-48420-3. DOI: 10.1007/978-3-662-48422-7.

[DK85] J. Dadok and V. Kac. "Polar representations". In: *J. Algebra* 92.2 (1985), pp. 504–524. DOI: 10.1016/0021-8693(85)90136-X.

[DKH21] M. Drton, S. Kuriki, and P. Hoff. "Existence and uniqueness of the Kronecker covariance MLE". In: *Ann. Statist.* 49.5 (2021), pp. 2721–2754. DOI: 10.1214/21-aos2052.

[DKZ13] J. Draisma, S. Kuhnt, and P. Zwiernik. "Groups acting on Gaussian graphical models". In: *Ann. Statist.* 41.4 (2013), pp. 1944–1969. DOI: 10.1214/13-AOS1130.

[DM17] H. Derksen and V. Makam. "Polynomial degree bounds for matrix semi-invariants". In: *Adv. Math.* 310 (2017), pp. 44–63. DOI: 10.1016/j.aim.2017.01.018.

[DM18] H. Derksen and V. Makam. "Degree bounds for semi-invariant rings of quivers". In: *J. Pure Appl. Algebra* 222.10 (2018), pp. 3282–3292. DOI: 10.1016/j.jpaa.2017.12.007.

[DM20a] H. Derksen and V. Makam. "Algorithms for orbit closure separation for invariants and semi-invariants of matrices". In: *Algebra Number Theory* 14.10 (2020), pp. 2791–2813. ISSN: 1937-0652. DOI: 10.2140/ant.2020.14.2791.

[DM20b] H. Derksen and V. Makam. "An exponential lower bound for the degrees of invariants of cubic forms and tensor actions". In: *Adv. Math.* 368 (2020), pp. 107136, 25. DOI: 10.1016/j.aim.2020.107136.

[DM21] H. Derksen and V. Makam. "Maximum likelihood estimation for matrix normal models via quiver representations". In: *SIAM J. Appl. Algebra Geom.* 5.2 (2021), pp. 338–365. DOI: 10.1137/20M1369348.

[DMW22] H. Derksen, V. Makam, and M. Walter. "Maximum likelihood estimation for tensor normal models via castling transforms". In: *Forum Math. Sigma* 10 (2022), Paper No. e50, 23. DOI: 10.1017/fms.2022.37.

[Dol03] I. Dolgachev. *Lectures on invariant theory*. Vol. 296. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, 2003, pp. xvi+220. ISBN: 0-521-52548-9. DOI: 10.1017/CBO9780511615436.

[DPZ67] R. J. Duffin, E. L. Peterson, and C. Zener. *Geometric programming: Theory and application*. John Wiley & Sons, Inc., New York-London-Sydney, 1967, pp. xi+278.

[DR72] J. N. Darroch and D. Ratcliff. "Generalized iterative scaling for log-linear models". In: *Ann. Math. Statist.* 43 (1972), pp. 1470–1480. ISSN: 0003-4851. DOI: 10.1214/aoms/1177692379.

[Drt18] M. Drton. "Algebraic problems in structural equation modeling". In: *The 50th anniversary of Gröbner bases*. Vol. 77. Adv. Stud. Pure Math. Math. Soc. Japan, Tokyo, 2018, pp. 35–86. DOI: 10.2969/aspm/07710035.

[DS40] W. E. Deming and F. F. Stephan. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known". In: *Ann. Math. Statistics* 11 (1940), pp. 427–444. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731829.

[DSS09] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. Vol. 39. Oberwolfach Seminars. Birkhäuser Verlag, Basel, 2009, pp. viii+171. ISBN: 978-3-7643-8904-8. DOI: 10.1007/978-3-7643-8905-5.

[Dut99] P. Dutilleul. "The MLE algorithm for the matrix normal distribution". In: *J. Stat. Comput. Simul.* 64.2 (1999), pp. 105–123. DOI: 10.1080/00949659908811970.

[DW17] H. Derksen and J. Weyman. *An introduction to quiver representations*. Vol. 184. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2017, pp. x+334. ISBN: 978-1-4704-2556-2. DOI: 10.1090/gsm/184.

[DWW14] P. Danaher, P. Wang, and D. M. Witten. "The joint graphical lasso for inverse covariance estimation across multiple classes". In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 76.2 (2014), pp. 373–397. ISSN: 1369-7412. DOI: 10.1111/rssb.12033.

[Eck80] J. G. Ecker. "Geometric programming: methods, computations and applications". In: *SIAM Rev.* 22.3 (1980), pp. 338–362. ISSN: 0036-1445. DOI: 10.1137/1022058.

[FH91] W. Fulton and J. Harris. *Representation theory*. Vol. 129. Graduate Texts in Mathematics. A first course, Readings in Mathematics. Springer-Verlag, New York, 1991, pp. xvi+551. ISBN: 0-387-97527-6. DOI: 10.1007/978-1-4612-0979-9.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441. DOI: 10.1093/biostatistics/kxm045.

[Fie70] S. E. Fienberg. "An iterative procedure for estimation in contingency tables". In: *Ann. Math. Statist.* 41 (1970), pp. 907–917. ISSN: 0003-4851. DOI: 10.1214/aoms/1177696968.

[FLNP00] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. "Using Bayesian Networks to Analyze Expression Data". In: RECOMB '00. Tokyo, Japan: Association for Computing Machinery, 2000, pp. 127–135. ISBN: 1581131860. DOI: 10.1145/332306.332355.

[FM20] W. C. Franks and A. Moitra. "Rigorous guarantees for Tyler's M-estimator via quantum expansion". In: *Conference on Learning Theory*. PMLR. 2020, pp. 1601–1632. URL: http://proceedings.mlr.press/v125/franks20a.html.

[FM81] S. E. Fienberg and M. M. Meyer. *Iterative Proportional Fitting*. Tech. rep. Carnegie-Mellon University, Pittsburgh PA,Dept. of Statistics, 1981.

[FORW21] C. Franks, R. Oliveira, A. Ramachandran, and M. Walter. *Near optimal sample complexity for matrix and tensor normal models via geodesic convexity*. 2021. URL: https://arxiv.org/abs/2110.07583v2.

[FR12] S. E. Fienberg and A. Rinaldo. "Maximum likelihood estimation in log-linear models". In: *Ann. Statist.* 40.2 (2012), pp. 996–1023. ISSN: 0090-5364. DOI: 10.1214/12-AOS986.

[FR21] W. C. Franks and P. Reichenbach. "Barriers for recent methods in geodesic optimization". In: *36th Computational Complexity Conference*. Vol. 200. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2021, Art. No. 13, 54. DOI: 10.4230/LIPIcs.CCC.2021.13.

[Fra02] M. Franz. "Moment polytopes of projective G-varieties and tensor products of symmetric group representations". In: *J. Lie Theory* 12.2 (2002), pp. 539–549. URL: https://www.emis.de/journals/JLT/vol.12_no.2/16.html.

[Fry90] M. Frydenberg. "The chain graph Markov property". In: *Scand. J. Statist.* 17.4 (1990), pp. 333–353. URL: https://www.jstor.org/stable/4616181.

[FSG22] C. Franks, T. Soma, and M. X. Goemans. *Shrunk subspaces via operator Sinkhorn iteration*. 2022. DOI: 10.48550/ARXIV.2207.08311.

[GGOW16] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson. "A deterministic polynomial time algorithm for non-commutative rational identity testing". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 109–117. DOI: 10.1109/FOCS.2016.95. URL: https://arxiv.org/abs/1511.03730.

[GGOW18] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson. "Algorithmic and optimization aspects of Brascamp-Lieb inequalities, via operator scaling". In: *Geom. Funct. Anal.* 28.1 (2018), pp. 100–145. ISSN: 1016-443X. DOI: 10.1007/s00039-018-0434-2.

[GIM+20] A. Garg, C. Ikenmeyer, V. Makam, R. Oliveira, M. Walter, and A. Wigderson. "Search Problems in Algebraic Complexity, GCT, and Hardness of Generators for Invariant Rings". In: *35th Computational Complexity Conference (CCC 2020)*. Vol. 169. Leibniz International Proceedings in Informatics (LIPIcs). 2020, 12:1–12:17. ISBN: 978-3-95977-156-6. DOI: 10.4230/LIPIcs.CCC.2020.12.

[GM15] X. Gao and H. Massam. "Estimation of symmetry-constrained Gaussian graphical models: application to clustered dense networks". In: *J. Comput. Graph. Statist.* 24.4 (2015), pp. 909–929. ISSN: 1061-8600. DOI: 10.1080/10618600.2014.937811.

[GMS06] D. Geiger, C. Meek, and B. Sturmfels. "On the toric algebra of graphical models". In: *Ann. Statist.* 34.3 (2006), pp. 1463–1492. ISSN: 0090-5364. DOI: 10.1214/009053606000000263.

[GN99] X. Gual-Arnau and A. M. Naveira. "Volume of tubes in noncompact symmetric spaces". In: *Publ. Math. Debrecen* 54.3-4 (1999), pp. 313–320. ISSN: 0033-3883.

[GO18] A. Garg and R. Oliveira. "Recent progress on scaling algorithms and applications". In: *Bull. Eur. Assoc. Theor. Comput. Sci. EATCS* 125 (2018), pp. 14–49. ISSN: 0252-9742. URL: http://eatcs.org/beatcs/index.php/beatcs/article/view/533.

[Goo63] N. R. Goodman. "Statistical analysis based on a certain multivariate complex Gaussian distribution. (An introduction)". In: *Ann. Math. Statist.* 34 (1963), pp. 152–177. ISSN: 0003-4851. DOI: 10.1214/aoms/1177704250.

[GS18] E. Gross and S. Sullivant. "The maximum likelihood threshold of a graph". In: *Bernoulli* 24.1 (2018), pp. 386–407. ISSN: 1350-7265. DOI: 10.3150/16-BEJ881.

[GS84] V. Guillemin and S. Sternberg. "Convexity properties of the moment mapping. II". In: *Invent. Math.* 77.3 (1984), pp. 533–546. DOI: 10.1007/BF01388837.

[Gur04a] L. Gurvits. "Classical complexity and quantum entanglement". In: *J. Comput. System Sci.* 69.3 (2004), pp. 448–484. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2004.06.003.

[Gur04b] L. Gurvits. *Combinatorial and algorithmic aspects of hyperbolic polynomials*. 2004. DOI: 10.48550/ARXIV.MATH/0404474.

[Gur06] L. Gurvits. "Hyperbolic polynomials approach to Van der Waerden/Schrijver-Valiant like conjectures: sharper bounds, simpler proofs and algorithmic applications". In: *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*. 2006, pp. 417–426. DOI: 10.1145/1132516.1132578. URL: https://arxiv.org/abs/math/0510452.

[GW09] R. Goodman and N. R. Wallach. *Symmetry, representations, and invariants*. Vol. 255. Graduate Texts in Mathematics. Springer, Dordrecht, 2009, pp. xx+716. ISBN: 978-0-387-79851-6. DOI: 10.1007/978-0-387-79852-3.

[Hab74] S. J. Haberman. "Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations". In: *Ann. Statist.* 2 (1974), pp. 911–924. ISSN: 0090-5364. URL: https://www.jstor.org/stable/2958058.

[Hal15] B. Hall. *Lie groups, Lie algebras, and representations*. Second. Vol. 222. Graduate Texts in Mathematics. An elementary introduction. Springer, Cham, 2015, pp. xiv+449. ISBN: 978-3-319-13466-6. DOI: 10.1007/978-3-319-13467-3.

[Hal35] P. Hall. "On Representatives of Subsets". In: *J. London Math. Soc.* 10.1 (1935), pp. 26–30. ISSN: 0024-6107. DOI: 10.1112/jlms/s1-10.37.26.

[Hil35] D. Hilbert. *Gesammelte Abhandlungen. Band III: Analysis, Grundlagen der Mathematik, Physik, Verschiedenes, Lebensgeschichte*. Springer-Verlag, Berlin, 1935, pp. vii+435. URL: http://resolver.sub.uni-goettingen.de/purl?PPN237834022.

[Hil90] D. Hilbert. "Ueber die Theorie der algebraischen Formen". In: *Math. Ann.* 36.4 (1890), pp. 473–534. ISSN: 0025-5831. DOI: 10.1007/BF01208503.

[Hil93] D. Hilbert. "Ueber die vollen Invariantensysteme". In: *Math. Ann.* 42.3 (1893), pp. 313–373. ISSN: 0025-5831. DOI: 10.1007/BF01444162.

[Hir22] H. Hirai. *On a manifold formulation of self-concordant functions*. 2022. DOI: 10.48550/ARXIV.2212.10981.

[HL08] S. Højsgaard and S. L. Lauritzen. "Graphical Gaussian models with edge and vertex symmetries". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70.5 (2008), pp. 1005–1027. DOI: 10.1111/j.1467-9868.2008.00666.x.

[HM21] L. Hamilton and A. Moitra. *No-go Theorem for Acceleration in the Hyperbolic Plane*. 2021. DOI: 10.48550/ARXIV.2101.05657.

[Hos15] V. Hoskins. *Moduli Problems and Geometric Invariant Theory*. Lecture Notes. 2015. URL: https://www.math.ru.nl/~vhoskins/M15_Lecture_notes.pdf.

[HS07] P. Heinzner and G. W. Schwarz. "Cartan decomposition of the moment map". In: *Math. Ann.* 337.1 (2007), pp. 197–232. ISSN: 0025-5831. DOI: 10.1007/s00208-006-0032-8.

[HSS08a] A. Hagberg, P. Swart, and D. S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[HSS08b] P. Heinzner, G. W. Schwarz, and H. Stötzel. "Stratifications with respect to actions of real reductive groups". In: *Compos. Math.* 144.1 (2008), pp. 163–185. ISSN: 0010-437X. DOI: 10.1112/S0010437X07003259.

[Hum75] J. E. Humphreys. *Linear algebraic groups*. Graduate Texts in Mathematics, No. 21. Springer-Verlag, New York-Heidelberg, 1975, pp. xiv+247. ISBN: 978-1-4684-9445-7. DOI: 10.1007/978-1-4684-9443-3.

[Ide16] M. Idel. *A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps*. 2016. DOI: 10.48550/ARXIV.1609.06349.

[IQS18] G. Ivanyos, Y. Qiao, and K. V. Subrahmanyam. "Constructive non-commutative rank computation is in deterministic polynomial time". In: *Comput. Complexity* 27.4 (2018), pp. 561–593. ISSN: 1016-3328. DOI: 10.1007/s00037-018-0165-7.

[Ish21] H. Ishi. "On Gaussian group convex models". In: *Geometric science of information*. Vol. 12829. Lecture Notes in Comput. Sci. Springer, Cham, 2021, pp. 256–264. DOI: 10.1007/978-3-030-80209-7_29.

[Jay03] E. T. Jaynes. *Probability theory*. The logic of science, Edited and with a foreword by G. Larry Bretthorst. Cambridge University Press, Cambridge, 2003, pp. xxx+727. ISBN: 0-521-59271-2. DOI: 10.1017/CBO9780511790423.

[Kar84] N. Karmarkar. "A new polynomial-time algorithm for linear programming". In: *Combinatorica* 4.4 (1984), pp. 373–395. ISSN: 0209-9683. DOI: 10.1007/BF02579150.

[Kem78] G. R. Kempf. "Instability in invariant theory". In: *Ann. of Math. (2)* 108.2 (1978), pp. 299–316. ISSN: 0003-486X. DOI: 10.2307/1971168.

[Kin94] A. D. King. "Moduli of representations of finite-dimensional algebras". In: *Quart. J. Math. Oxford Ser. (2)* 45.180 (1994), pp. 515–530. DOI: 10.1093/qmath/45.4.515.

[Kir84a] F. Kirwan. "Convexity properties of the moment mapping. III". In: *Invent. Math.* 77.3 (1984), pp. 547–552. ISSN: 0020-9910. DOI: 10.1007/BF01388838.

[Kir84b] F. C. Kirwan. *Cohomology of quotients in symplectic and algebraic geometry*. Vol. 31. Mathematical Notes. Princeton University Press, Princeton, NJ, 1984, pp. i+211. ISBN: 0-691-08370-3. DOI: 10.2307/j.ctv10vm2m8.

[KL05] M. Kravtsov and E. Lukshin. "On some properties of noninteger vertices of a three-index axial transportation polytope". In: *Tr. Inst. Matematiki NAN Belarusi* 13.2 (2005), pp. 31–36.

[Kly06] A. A. Klyachko. "Quantum marginal problem and N-representability". In: *Journal of Physics: Conference Series*. Vol. 36. 1. IOP Publishing. 2006, p. 014. DOI: 10.1088/1742-6596/36/1/014.

[KN79] G. Kempf and L. Ness. "The length of vectors in representation spaces". In: *Algebraic geometry (Proc. Summer Meeting, Univ. Copenhagen, Copenhagen, 1978)*. Vol. 732. Lecture Notes in Math. Springer, Berlin, 1979, pp. 233–243. DOI: 10.1007/BFb0066647.

[Kna96] A. W. Knapp. *Lie groups beyond an introduction*. Vol. 140. Progress in Mathematics. Birkhäuser Boston, Inc., Boston, MA, 1996, pp. xvi+604. ISBN: 0-8176-3926-8. DOI: 10.1007/978-1-4757-2453-0.

[KP96] H. Kraft and C. Procesi. "Classical invariant theory, a primer". In: *Lecture Notes. Preliminary version* (1996). URL: http://www.math.iitb.ac.in/~shripad/Wilberd/KP-Primer.pdf.

[Kra07] V. M. Kravtsov. "Combinatorial properties of noninteger vertices of a polytope in a three-index axial assignment problem". In: *Kibernet. Sistem. Anal.* 43.1 (2007), pp. 33–44, 189. DOI: 10.1007/s10559-007-0023-0.

[Kra84] H. Kraft. *Geometrische Methoden in der Invariantentheorie*. Aspects of Mathematics, D1. Friedr. Vieweg & Sohn, Braunschweig, 1984, pp. x+308. ISBN: 3-528-08525-8. DOI: 10.1007/978-3-322-83813-1.

[Lau96] S. L. Lauritzen. *Graphical models*. Vol. 17. Oxford Statistical Science Series. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, 1996, pp. x+298.

[LC98] E. L. Lehmann and G. Casella. *Theory of point estimation*. Second. Springer Texts in Statistics. Springer-Verlag, New York, 1998, pp. xxvi+589. ISBN: 0-387-98502-6. DOI: 10.1007/b98854.

[Lee13] J. M. Lee. *Introduction to smooth manifolds*. Second. Vol. 218. Graduate Texts in Mathematics. Springer, New York, 2013, pp. xvi+708. DOI: 10.1007/978-1-4419-9982-5.

[LHCJ22] T. Lin, N. Ho, M. Cuturi, and M. I. Jordan. "On the complexity of approximating multimarginal optimal transport". In: *J. Mach. Learn. Res* 23 (2022), pp. 1–43. URL: http://jmlr.org/papers/v23/19-843.html.

[Lie90] E. H. Lieb. "Gaussian kernels have only Gaussian maximizers". In: *Invent. Math.* 102.1 (1990), pp. 179–208. ISSN: 0020-9910. DOI: 10.1007/BF01233426.

[LL14] N. Linial and Z. Luria. "On the vertices of the d-dimensional Birkhoff polytope". In: *Discrete Comput. Geom.* 51.1 (2014), pp. 161–170. ISSN: 0179-5376. DOI: 10.1007/s00454-013-9554-5.

[LM07] G. Letac and H. Massam. "Wishart distributions for decomposable graphs". In: *Ann. Statist.* 35.3 (2007), pp. 1278–1323. DOI: 10.1214/009053606000001235.

[Lov79] L. Lovász. "On determinants, matchings, and random algorithms". In: *Fundamentals of computation theory (Proc. Conf. Algebraic, Arith. and Categorical Methods in Comput. Theory, Berlin/Wendisch-Rietz, 1979)*. Vol. 2. Math. Res. Akademie-Verlag, Berlin, 1979, pp. 565–574.

[LSW00] N. Linial, A. Samorodnitsky, and A. Wigderson. "A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents". In: *Combinatorica* 20.4 (2000), pp. 545–568. ISSN: 0209-9683. DOI: 10.1007/s004930070007.

[Lun75] D. Luna. "Sur certaines opérations différentiables des groupes de Lie". In: *Amer. J. Math.* 97 (1975), pp. 172–181. ISSN: 0002-9327. DOI: 10.2307/2373666.

[LV20] J. Leake and N. K. Vishnoi. "On the computability of continuous maximum entropy distributions with applications". In: *STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, 2020, pp. 930–943. DOI: 10.1145/3357713.3384302. URL: https://arxiv.org/abs/2004.07403.

[LZ05] N. Lu and D. Zimmerman. "The likelihood ratio test for a separable covariance matrix". In: *Stat. Probab. Lett.* 73.4 (2005), pp. 449–457. DOI: 10.1016/j.spl.2005.04.020.

[Mad00] J. Madsen. "Invariant normal models with recursive graphical Markov structure". In: *Ann. Statist.* 28.4 (2000), pp. 1150–1178. ISSN: 0090-5364. DOI: 10.1214/aos/1015956711.

[Mar01] A. Marian. "On the real moment map". In: *Math. Res. Lett.* 8.5-6 (2001), pp. 779–788. ISSN: 1073-2780. DOI: 10.4310/MRL.2001.v8.n6.a8.

[MDLW19] M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, eds. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019, pp. xviii+536. ISBN: 978-1-4987-8862-5.

[MFK94] D. Mumford, J. Fogarty, and F. Kirwan. *Geometric invariant theory*. Third. Vol. 34. Ergebnisse der Mathematik und ihrer Grenzgebiete (2) [Results in Mathematics and Related Areas (2)]. Springer-Verlag, Berlin, 1994, pp. xiv+292. DOI: 10.1007/978-3-642-57916-5.

[Mil11] R. B. Millar. *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*. John Wiley & Sons, 2011. ISBN: 9780470094822. DOI: 10.1002/9780470094846.

[Mil17] J. Milne. *Algebraic groups*. Vol. 170. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2017, pp. xvi+644. DOI: 10.1017/9781316711736.

[MMB+20] Y. Momozawa et al. "Genome wide association study of 40 clinical measurements in eight dog breeds". In: *Scientific reports* 10.1 (2020), pp. 1–11. DOI: 10.1038/s41598-020-63457-y.

[Mos55] G. Mostow. "Self-adjoint groups". In: *Ann. of Math. (2)* 62 (1955), pp. 44–55. DOI: 10.2307/2007099.

[Mos56] G. D. Mostow. "Fully reducible subgroups of algebraic groups". In: *Amer. J. Math.* 78 (1956), pp. 200–221. ISSN: 0002-9327. DOI: 10.2307/2372490.

[MRS21] V. Makam, P. Reichenbach, and A. Seigal. *Symmetries in Directed Gaussian Graphical Models*. 2021. URL: https://arxiv.org/abs/2108.10058v2.

[MS15] T. Maciążek and A. Sawicki. "Critical points of the linear entropy for pure L-qubit states". In: *Journal of Physics A: Mathematical and Theoretical* 48.4 (2015), p. 045305. DOI: 10.1088/1751-8113/48/4/045305.

[MS18] T. Maciążek and A. Sawicki. "Asymptotic properties of entanglement polytopes for large number of qubits". In: *Journal of Physics A: Mathematical and Theoretical* 51.7 (2018), 07LT01. DOI: 10.1088/1751-8121/aaa4d7.

[Mul17] K. D. Mulmuley. "Geometric complexity theory V: Efficient algorithms for Noether normalization". In: *J. Amer. Math. Soc.* 30.1 (2017), pp. 225–309. ISSN: 0894-0347. DOI: 10.1090/jams/864.

[Mum77] D. Mumford. "Stability of projective varieties". In: *Enseign. Math. (2)* 23.1-2 (1977), pp. 39–110. ISSN: 0013-8584.

[MW21] V. Makam and A. Wigderson. "Singular tuples of matrices is not a null cone (and the symmetries of algebraic varieties)". In: *J. Reine Angew. Math.* 780 (2021), pp. 79–131. ISSN: 0075-4102. DOI: 10.1515/crelle-2021-0044.

[Nes84] L. Ness. "A stratification of the null cone via the moment map". In: *Amer. J. Math.* 106.6 (1984). With an appendix by David Mumford, pp. 1281–1329. DOI: 10.2307/2374395.

[New78] P. E. Newstead. *Introduction to moduli problems and orbit spaces*. Vol. 51. Tata Institute of Fundamental Research Lectures on Mathematics and Physics. Tata Institute of Fundamental Research, Bombay; Narosa Publishing House, New Delhi, 1978, pp. vi+183. ISBN: 0-387-08851-2.

[NW23] H. Nieuwboer and M. Walter. *Interior-point methods on manifolds: theory and applications*. 2023. DOI: 10.48550/ARXIV.2303.04771.

[OS00] L. O'Shea and R. Sjamaar. "Moment maps and Riemannian symmetric pairs". In: *Math. Ann.* 317.3 (2000), pp. 415–457. ISSN: 0025-5831. DOI: 10.1007/PL00004408.

[OV90] A. L. Onishchik and È. B. Vinberg. *Lie groups and algebraic groups*. Springer Series in Soviet Mathematics. Translated from the Russian and with a preface by D. A. Leites. Springer-Verlag, Berlin, 1990, pp. xx+328. ISBN: 3-540-50614-4. DOI: 10.1007/978-3-642-74334-4.

[Par20] P.-E. Paradan. *Moment polytopes in real symplectic geometry I*. 2020. DOI: 10.48550/ARXIV.2012.08837.

[PB14] J. Peters and P. Bühlmann. "Identifiability of Gaussian structural equation models with equal error variances". In: *Biometrika* 101.1 (2014), pp. 219–228. DOI: 10.1093/biomet/ast043.

[Pea09] J. Pearl. *Causality*. Second. Models, reasoning, and inference. Cambridge University Press, Cambridge, 2009, pp. xx+464. ISBN: 978-0-521-89560-6. DOI: 10.1017/CBO9780511803161. URL: https://doi.org/10.1017/CBO9780511803161.

[Pet76] E. L. Peterson. "Geometric programming". In: *SIAM Rev.* 18.1 (1976), pp. 1–51. ISSN: 0036-1445. DOI: 10.1137/1018001.

[Pop89] V. L. Popov. "Closed orbits of Borel subgroups". In: *Mathematics of the USSR-Sbornik* 63.2 (1989), p. 375. DOI: 10.1070/SM1989v063n02ABEH003280.

[PR71] B. N. Parlett and C. Reinsch. "Balancing a matrix for calculation of eigenvalues and eigenvectors". In: *Handbook for Automatic Computation*. Springer, 1971, pp. 315–326. DOI: 10.1007/978-3-642-86940-2_22.

[Pro07] C. Procesi. *Lie groups*. Universitext. An approach through invariants and representations. Springer, New York, 2007, pp. xxiv+596. ISBN: 978-0-387-26040-2. DOI: 10.1007/978-0-387-28929-8.

[PS05] L. Pachter and B. Sturmfels, eds. *Algebraic statistics for computational biology*. Cambridge University Press, New York, 2005, pp. xii+420. ISBN: 978-0-521-85700-0. DOI: 10.1017/CBO9780511610684.

[PV94] V. L. Popov and E. B. Vinberg. "Invariant Theory". In: *Algebraic Geometry IV: Linear Algebraic Groups Invariant Theory*. Ed. by A. N. Parshin and I. R. Shafarevich. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 123–278. ISBN: 978-3-662-03073-8. DOI: 10.1007/978-3-662-03073-8_2.

[Rei95] N. Reid. "The roles of conditioning in inference". In: *Statist. Sci.* 10.2 (1995), pp. 138–157, 173–189, 193–196. ISSN: 0883-4237. URL: https://www.jstor.org/stable/2246182.

[Ric06] J. A. Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[Ros61] M. Rosenlicht. "On quotient varieties and the affine embedding of certain homogeneous spaces". In: *Trans. Amer. Math. Soc.* 101 (1961), pp. 211–223. ISSN: 0002-9947. DOI: 10.2307/1993371.

[RS89] U. G. Rothblum and H. Schneider. "Scalings of matrices which have prespecified row sums and column sums via optimization". In: *Linear Algebra Appl.* 114/115 (1989), pp. 737–764. ISSN: 0024-3795. DOI: 10.1016/0024-3795(89)90491-6.

[RS90] R. W. Richardson and P. J. Slodowy. "Minimum vectors for real reductive algebraic groups". In: *J. London Math. Soc. (2)* 42.3 (1990), pp. 409–429. DOI: 10.1112/jlms/s2-42.3.409.

[Rus20] A. Rusciano. "A Riemannian Corollary of Helly's theorem". In: *J. Convex Anal.* 27.4 (2020), pp. 1261–1275. ISSN: 0944-6532. URL: https://www.heldermann.de/JCA/JCA27/JCA274/jca27067.htm.

[Sak92] R. M. Sakia. "The Box-Cox transformation technique: a review". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 41.2 (1992), pp. 169–178. DOI: 10.2307/2348250.

[SC12] P. Shah and V. Chandrasekaran. "Group symmetry and covariance regularization". In: *Electron. J. Stat.* 6 (2012), pp. 1600–1640. DOI: 10.1214/12-EJS723.

[Sch80] J. T. Schwartz. "Fast probabilistic algorithms for verification of polynomial identities". In: *J. Assoc. Comput. Mach.* 27.4 (1980), pp. 701–717. DOI: 10.1145/322217.322225.

[Sch86] A. Schrijver. *Theory of linear and integer programming*. Wiley-Interscience Series in Discrete Mathematics. A Wiley-Interscience Publication. John Wiley & Sons, Ltd., Chichester, 1986, pp. xii+471. ISBN: 0-471-90854-1.

[Sev00] T. A. Severini. *Likelihood methods in statistics*. Vol. 22. Oxford Statistical Science Series. Oxford University Press, Oxford, 2000, pp. xii+380. ISBN: 0-19-850650-3.

[Sha13] I. R. Shafarevich. *Basic algebraic geometry. 2*. Third. Schemes and complex manifolds, Translated from the 2007 third Russian edition by Miles Reid. Springer, Heidelberg, 2013, pp. xiv+262. ISBN: 978-3-642-38009-9. DOI: 10.1007/978-3-642-38010-5.

[Sin64] R. Sinkhorn. "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *Ann. Math. Statist.* 35 (1964), pp. 876–879. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703591.

[Sja98] R. Sjamaar. "Convexity properties of the moment mapping re-examined". In: *Adv. Math.* 138.1 (1998), pp. 46–91. DOI: 10.1006/aima.1998.1739.

[SK67] R. Sinkhorn and P. Knopp. "Concerning nonnegative matrices and doubly stochastic matrices". In: *Pacific J. Math.* 21 (1967), pp. 343–348. ISSN: 0030-8730. URL: http://projecteuclid.org/euclid.pjm/1102992505.

[SOK14] A. Sawicki, M. Oszmaniec, and M. Kuś. "Convexity of momentum map, Morse index, and quantum entanglement". In: *Rev. Math. Phys.* 26.3 (2014), pp. 1450004, 39. ISSN: 0129-055X. DOI: 10.1142/S0129055X14500044.

[SPP+05] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. "Causal protein-signaling networks derived from multiparameter single-cell data". In: *Science* 308.5721 (2005), pp. 523–529. DOI: 10.1126/science.1105809.

[Spr98] T. A. Springer. *Linear algebraic groups*. Second. Vol. 9. Progress in Mathematics. Birkhäuser Boston, Inc., Boston, MA, 1998, pp. xiv+334. ISBN: 0-8176-4021-5. DOI: 10.1007/978-0-8176-4840-4.

[Stu08] B. Sturmfels. *Algorithms in invariant theory*. Second. Texts and Monographs in Symbolic Computation. SpringerWienNewYork, Vienna, 2008, pp. vi+197. ISBN: 978-3-211-77416-8. DOI: 10.1007/978-3-211-77417-5.

[SU10] B. Sturmfels and C. Uhler. "Multivariate Gaussian, semidefinite matrix completion, and convex algebraic geometry". In: *Ann. Inst. Statist. Math.* 62.4 (2010), pp. 603–638. ISSN: 0020-3157. DOI: 10.1007/s10463-010-0295-4.

[Sul18] S. Sullivant. *Algebraic Statistics*. Vol. 194. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2018, pp. xiii+490. DOI: 10.1090/gsm/194.

[Sur00] B. Sury. "An elementary proof of the Hilbert-Mumford criterion". In: *Electron. J. Linear Algebra* 7 (2000), pp. 174–177. DOI: 10.13001/1081-3810.1053.

[SV14] M. Singh and N. K. Vishnoi. "Entropy, optimization and counting". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 50–59. DOI: 10.1145/2591796.2591803.

[SV19] D. Straszak and N. K. Vishnoi. "Maximum entropy distributions: Bit complexity and stability". In: *Conference on Learning Theory*. PMLR. 2019, pp. 2861–2891. URL: https://proceedings.mlr.press/v99/straszak19a.html.

[Szé06] G. Székelyhidi. "Extremal metrics and K-stability". PhD thesis. Imperial College, University of London, 2006.

[TB97] L. N. Trefethen and D. Bau III. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, pp. xii+361. DOI: 10.1137/1.9780898719574.

[Tho06] R. P. Thomas. "Notes on GIT and symplectic reduction for bundles and varieties". In: *Surveys in differential geometry. Vol. X*. Vol. 10. Surv. Differ. Geom. Int. Press, Somerville, MA, 2006, pp. 221–273. DOI: 10.4310/SDG.2005.v10.n1.a7.

[Uhl12] C. Uhler. "Geometry of maximum likelihood estimation in Gaussian graphical models". In: *Ann. Statist.* 40.1 (2012), pp. 238–261. ISSN: 0090-5364. DOI: 10.1214/11-AOS957.

[Vaa98] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998, pp. xvi+443. ISBN: 0-521-49603-9. DOI: 10.1017/CBO9780511802256.

[VAAW16] V. Vinciotti, L. Augugliaro, A. Abbruzzo, and E. C. Wit. "Model selection for factorial Gaussian graphical models with an application to dynamic regulatory networks". In: *Stat. Appl. Genet. Mol. Biol.* 15.3 (2016), pp. 193–212. ISSN: 2194-6302. DOI: 10.1515/sagmb-2014-0075.

[VP90] T. Verma and J. Pearl. "Causal networks: semantics and expressiveness". In: *Uncertainty in artificial intelligence, 4*. Vol. 9. Mach. Intelligence Pattern Recogn. North-Holland, Amsterdam, 1990, pp. 69–76. DOI: 10.1016/B978-0-444-88650-7.50011-1.

[WA18] M. D. Ward and J. S. Ahlquist. *Maximum Likelihood for Social Science: Strategies for Analysis*. Analytical Methods for Social Research. Cambridge University Press, 2018. DOI: 10.1017/9781316888544.

[Wal14] M. Walter. "Multipartite Quantum States and their Marginals". PhD thesis. 2014. DOI: 10.3929/ETHZ-A-010250985. URL: https://arxiv.org/abs/1410.6820.

[Wal17] N. Wallach. *Geometric invariant theory: Over the real and complex numbers*. Universitext. Springer, 2017, pp. xiv+190. DOI: 10.1007/978-3-319-65907-7.

[Was21] M. L. Waskom. "Seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021.

[Wat79] W. C. Waterhouse. *Introduction to affine group schemes*. Vol. 66. Graduate Texts in Mathematics. Springer-Verlag, New York-Berlin, 1979, pp. xi+164. ISBN: 0-387-90421-2. DOI: 10.1007/978-1-4612-6217-6.

[Wey39] H. Weyl. *The Classical Groups. Their Invariants and Representations*. Princeton University Press, Princeton, N.J., 1939, pp. xii+302. DOI: 10.1515/9781400883905.

[Whi57] H. Whitney. "Elementary structure of real algebraic varieties". In: *Ann. of Math. (2)* 66 (1957), pp. 545–556. DOI: 10.2307/1969908.

[Wie12] A. Wiesel. "Geodesic convexity and covariance estimation". In: *IEEE Trans. Signal Process.* 60.12 (2012), pp. 6182–6189. DOI: 10.1109/TSP.2012.2218241.

[WJS08] K. Werner, M. Jansson, and P. Stoica. "On estimation of covariance matrices with Kronecker product structure". In: *IEEE Trans. Signal Process.* 56.2 (2008), pp. 478–491. ISSN: 1053-587X. DOI: 10.1109/TSP.2007.907834.

[Woo56] R. A. Wooding. "The multivariate distribution of complex normal variables". In: *Biometrika* 43 (1956), pp. 212–215. ISSN: 0006-3444. DOI: 10.1093/biomet/43.1-2.212.

[WZV+04] A. Wille et al. "Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana". In: *Genome biology* 5.11 (2004), pp. 1–13. DOI: 10.1186/gb-2004-5-11-r92.

[ZS16] H. Zhang and S. Sra. "First-order methods for geodesically convex optimization". In: *Conference on Learning Theory*. PMLR. 2016, pp. 1617–1638. URL: https://proceedings.mlr.press/v49/zhang16b.html.

Auxiliary Resources

[1] CTAN. The Comprehensive-TeX-Archive-Network. https://ctan.org

[2] LaTeX Wikibooks. https://en.wikibooks.org/wiki/LaTeX

[3] LEO Dictionary. https://dict.leo.org/englisch-deutsch/

[4] TeX - LaTeX Stack Exchange. https://tex.stackexchange.com

List of Symbols

$\mathbb{1}_m$	the all-ones vector in $\mathbb{R}^m$. 6
$\mathbb{1}_m^\perp$	orthogonal complement of the all-ones vector in $\mathbb{R}^m$. 28
$\overline{(\cdot)}$	closure in the Euclidean topology. 5
$\overline{(\cdot)}^{\mathrm{Z}}$	closure in the Zariski topology. 5
$(\cdot)^\dagger$	the Hermitian transpose of a matrix. 6
$(\cdot)^\mathsf{T}$	the transpose of a matrix. 6
Ad	the adjoint representation of a matrix Lie group G. 24
ad	the adjoint representation of a Lie algebra Lie(G). 24
α_s	the number of vertices of colour s. 216
$\mathcal{A}$	a set of invertible matrices. 149
$\mathcal{A}_{\mathrm{SL}}$	the set of matrices in $\mathcal{A}$ of determinant one. 152
$\mathcal{A}_{\mathrm{SL}}^-$	the set of matrices in $\mathcal{A}$ of determinant -1. 152
$\mathcal{A}_{\mathrm{SL}}^\pm$	the set of matrices in $\mathcal{A}$ of determinant ± 1. 152
$\mathcal{A}(\mathcal{G})$	set of matrices induced by a DAG $\mathcal{G}$. 187
$\mathcal{A}(\mathcal{G}, c)$	set of matrices induced by a coloured DAG $(\mathcal{G}, c)$. 205
$a \cdot Y$	left-multiplication of data Y by the matrix a. 151
$\mathcal{A} \cdot Y$	the "orbit" of Y under the set $\mathcal{A}$. 151
$\mathcal{A}_Y$	the stabilizing set of Y under the set $\mathcal{A}$. 151
β_s	the number of parent relationship colours of vertex colour s. 216
$b(ij)$	the butterfly body for edge $j \to i$. 236
$\mathrm{B}_m(\mathbb{K})$	the group of invertible upper triangular matrices over $\mathbb{K}$. 10
$\mathrm{cap}_G(v)$	the capacity of a vector v under the action of a group G. 32
$\mathrm{cap}(p)$	the capacity of an array p, see (5.1). 115
$\mathbb{C}[V]^G$	the ring of invariants. 34
$\mathrm{ch}(i)$	set of children of a vertex i. 132
$\mathrm{chol}(\Psi)$	the Cholesky decomposition of a positive definite matrix Ψ. 149
$c(i)$	colour of the vertex i. 204
$c(ij)$	colour of the edge $j \to i$ if $i \neq j$, otherwise $c(ii) = c(i)$. 204
$\mathrm{cp}_{\mathbb{K}}^{(n)}(a, b, c, d)$	the cut-and-paste rank. 179
Δ_A	the convex hull of the columns of the matrix A. 140

$\Delta_A(v)$	the weight polytope of v under the action of $\mathrm{GT}_d(\mathbb{C})$ via matrix A. 39
$\Delta_G(v)$	the moment polytope of a vector v. 54
$\Delta_T(v)$	the weight polytope of a vector v. 39
Δ_{m-1}	the $m-1$ dimensional probability simplex in $\mathbb{R}^m$. 125
$\det(M)$	determinant of the matrix M. 6
$d(\mathcal{G})$	the depth of a DAG $\mathcal{G}$. 192
$\mathrm{dist}(0, \mathrm{conv}(\Gamma))$	the Euclidean distance from zero to the polytope $\mathrm{conv}(\Gamma)$. 75
$\mathrm{ds}(M)$	distance to doubly stochastic of a matrix M. 68
$D_v(\varepsilon)$	the diameter for precision ε and vector v. 76
e_i	the i^{th} canonical unit vector. 6
E_{ij}	standard matrix with entry one at position (i,j), all other entries zero. 6
$\mathrm{End}(V)$	ring of $\mathbb{K}$-linear endomorphisms on V. 5
ε	usually a positive real number. 5
ϵ_i	the vector $e_i - m^{-1}\mathbb{1}_m$ in $\mathbb{R}^m$. 6
e^X	the exponential of the matrix X. 18
$\exp(X)$	the exponential of the matrix X. 18
f_p	a function defined in (5.1). 115
F_v	the Kempf-Ness function for a vector v. 43
$\gamma_G(\pi)$	the gap of a representation π. 80
$\gamma_T(\pi)$	the weight margin of a representation π. 75, 80
$\Gamma_{\mathscr{W}}$	a subset of weights of $\pi_{m,d}$ given by $\mathscr{W} \subseteq [m]^d$. 86
$\mathcal{G}$	a graph (further properties, e.g., (un)directed, depend on the context). 131
$(\mathcal{G}, c)$	coloured graph (further properties depend on the context). 204, 211
$\mathcal{G}_i$	a certain coloured subgraph for vertex i. 212
$\mathcal{G}_{(j\to i)}$	a certain coloured subgraph for edge $j \to i$. 212
$\mathcal{G}_{b(ij)}$	the butterfly graph for edge $j \to i$. 236
$\mathcal{G}^u$	the undirected graph induced by a DAG $\mathcal{G}$. 132
$(\mathcal{G}^u, c)$	coloured undirected graph induced by a coloured DAG $(\mathcal{G}, c)$. 211
G°	the Euclidean identity component of an algebraic (or a Lie group) G. 11, 16
$G^{\circ,\mathrm{Z}}$	the Zariski identity component of an algebraic group G. 11
$G_{\mathbb{K}}$	group of $\mathbb{K}$-rational points of an $\mathbb{R}$-group G. 11
$\mathrm{GL}_\alpha(\mathbb{K})$	a product of general linear groups over $\mathbb{K}$ for dimension vector α. 26
$\mathrm{GL}_m(\mathbb{K})$	the general linear group of invertible $m \times m$ matrices over $\mathbb{K}$. 10

$\mathrm{GL}(V)$	group of $\mathbb{K}$-linear automorphisms on V. 5
$\mathrm{GT}_m(\mathbb{K})$	the group of invertible diagonal matrices over $\mathbb{K}$. 10
$g \cdot v$	the group element g acts on vector v. 13
$G \cdot v$	the orbit of v under the group G. 13
G_v	the stabilizer of v under the group G. 13
$\mathfrak{i}$	the imaginary unit. 5
I_A	toric ideal corresponding to the log-linear model given by matrix A. 138
I_m	the $m \times m$ identity matrix. 6
$\mathrm{int}(P)$	the interior of a polytope $P \subseteq \mathbb{R}^d$. 41
$j \not\to i$	indicates that $j \to i$ is not in the corresponding directed graph. 131
$\mathbb{K}$	the field of real or complex numbers. 5
$\mathbb{K}^m$	vector space of m-tuples over $\mathbb{K}$. 5
$\mathbb{K}^{m_1 \times m_2}$	space of $m_1 \times m_2$-matrices with entries in $\mathbb{K}$. 5
$\mathbb{K}^{m_1} \otimes_{\mathbb{K}} \cdots \otimes_{\mathbb{K}} \mathbb{K}^{m_d}$	space of d-order tensors over $\mathbb{K}$. 5
$\mathbb{K}^\times$	the group of units of the field $\mathbb{K}$. 5
$\mathrm{KL}(p\|q)$	Kullback-Leibler (KL) divergence from q to p. 126
L_D	likelihood function given data D for a parametric statistical model. 124
L_u	likelihood function given data u for a discrete model. 126
L_Y	likelihood function given data Y for a Gaussian model. 128
ℓ_D	log-likelihood function given data D for a parametric statistical model. 124
ℓ_u	log-likelihood function given data u for a discrete model. 126
ℓ_Y	log-likelihood function given data Y for a Gaussian model. 128
$\mathrm{Lie}(G)$	the Lie algebra of a matrix Lie group G. 18
$\mathrm{Lie}(G)_\alpha$	root space for the root α of G. 31
$\log(\Psi)$	the logarithm of a positive definite matrix Ψ. 21
$[m]$	the set $\{1, 2, \ldots, m\}$. 5
$\mathcal{M}_{\mathcal{A}}^{\mathrm{g}}$	the Gaussian model via symmetrization of $\mathcal{A}$. 149
$\mathcal{M}_G^{\mathrm{g}}$	the Gaussian group model given by group G. 149, 159
$\mathcal{M}_{\mathcal{G}}^{\rightarrow}$	directed Gaussian graphical model given by a DAG $\mathcal{G}$. 132
$\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}$	restricted DAG (RDAG) model given by the coloured DAG $(\mathcal{G}, c)$. 204
$\mathcal{M}_A^{\ell\ell}$	the log-linear model given by matrix A. 138

$\overline{\mathcal{M}_A^{\ell\ell}}$	the extended log-linear model given by matrix A. 139
$\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, m_2)$	matrix normal model over $\mathbb{K}$. 130
$\mathcal{M}_{\mathbb{K}}^{\otimes}(m_1, \ldots, m_d)$	tensor normal model over $\mathbb{K}$. 130
$\mathcal{M}_{\mathcal{G}}^{\mathrm{ud}}$	undirected Gaussian graphical model for an undirected graph $\mathcal{G}$. 131
$\mathcal{M}_{(\mathcal{G},c)}^{\mathrm{ud}}$	the RCON model given by a coloured undirected graph $(\mathcal{G}, c)$. 211
$\mathcal{M}_{X \perp\!\!\!\perp Y}$	independence model of two discrete random variables. 127
$M_{i,+}$	the i^{th} row sum of the matrix M. 6
$M_{+,j}$	the j^{th} column sum of the matrix M. 6
$M_{+,+}$	sum over all entries of the matrix M. 6
$\mathrm{mlt}_b(\mathcal{M})$	ML boundedness-threshold of a Gaussian model $\mathcal{M}$. 128
$\mathrm{mlt}_e(\mathcal{M})$	ML existence-threshold of a Gaussian model $\mathcal{M}$. 128
$\mathrm{mlt}_u(\mathcal{M})$	ML uniqueness-threshold of a Gaussian model $\mathcal{M}$. 129
$\mathrm{mlt}_e\left(\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}, s\right)$	existence threshold of RDAG model $\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}$ restricted to vertex colour s. 222
$\mathrm{mlt}_u\left(\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}, s\right)$	uniqueness threshold of RDAG model $\mathcal{M}_{(\mathcal{G},c)}^{\rightarrow}$ restricted to vertex colour s. 222
μ_G	moment map for an action of the group G. 44
$M_{Y,s}$	the augmented sample matrix for vertex colour s and data Y. 216
$M'_{Y,s}$	the augmented sample matrix $M_{Y,s}$ without its top row. 222
$M_{Y,s}^{(t)}$	the t^{th} row of the augmented sample matrix $M_{Y,s}$. 217
$\mathcal{N}$	the topological null cone. 32
$\mathcal{N}^{\mathrm{inv}}$	the invariant-theoretic null cone. 34
$\mathcal{N}_m(b, \Sigma)$	m-dimensional multivariate Gaussian distribution with mean b and covariance matrix Σ. 127
$N(\pi)$	the weight norm of a representation π. 75
$\Omega(\pi)$	the set of weights of a representation π. 28
$\Omega(\pi_{m,d})$	the set of weights of the representation $\pi_{m,d}$. 30
$\mathrm{O}_m(\mathbb{K})$	the group of orthogonal matrices over $\mathbb{K}$. 10
$\mathrm{pa}(i)$	set of parents of a vertex i. 132
$\mathrm{PD}_m(\mathbb{R})$	the cone of symmetric positive definite matrices over $\mathbb{R}$. 21
$\mathrm{PD}_m(\mathbb{C})$	the cone of Hermitian positive definite matrices over $\mathbb{C}$. 21
ϕ^A	parametrization of the log-linear model given by matrix A. 138
π	a representation of a group G, usually a rational representation. 13, 23

$\pi^{\oplus n}$	the n-fold direct sum of the representation π. 24
$\pi_{m,d}$	the representation capturing tensor scaling. 26
Π	Lie algebra representation obtained as the differential of π. 19, 24
$\mathcal{P}_{\mathcal{M}}$	a parametric statistical model. 123
$\mathrm{prc}(s)$	the set of parent relationship colours for vertex colour s. 206
$\Psi^{1/2}$	the square root of the positive definite matrix Ψ. 21
$\sqrt{\Psi}$	the square root of the positive definite matrix Ψ. 21
$\mathbb{P}(V)$	the projective space given by the $\mathbb{K}$-vector space V. 5
Q	usually a quiver. 26
$\mathrm{relint}(P)$	the relative interior of a polytope $P \subseteq \mathbb{R}^d$. 41
$\varrho_{n,d}$	the representation capturing polynomial scaling. 107
$\mathscr{R}(Q,\alpha)$	the representation space of α-dimensional representations of the quiver Q. 26
r_s	generic rank of $M'_{Y,s}$ for sample size $n = 1$. 222
$R_u(G)$	the unipotent radical of a complex algebraic group G. 15
R_Y	a $\mathbb{C}$-algebra in the context of Popov's Criterion. 60
$\mathcal{S}$	a sample space. 123
$\mathrm{SL}_\alpha(\mathbb{K})$	a product of special linear groups over $\mathbb{K}$ for dimension vector α. 26
$\mathrm{SL}_m(\mathbb{K})$	the special linear group of determinant one matrices over $\mathbb{K}$. 10
$\mathrm{SO}_m(\mathbb{K})$	the group of special orthogonal matrices over $\mathbb{K}$. 10
$\mathrm{ST}_m(\mathbb{K})$	the group of diagonal matrices over $\mathbb{K}$ with determinant one. 10
SU_m	the group of special unitary matrices over $\mathbb{C}$. 11
$\mathrm{supp}(v)$	the support of a vector v. 39
S_Y	the sample covariance matrix for data Y. 128
$\mathrm{Sym}_m(\mathbb{K})$	the space of symmetric ($\mathbb{K} = \mathbb{R}$), respectively Hermitian ($\mathbb{K} = \mathbb{C}$) matrices. 20, 21
$\tau_{m,d}$	a certain SL-quiver action. 109
$\mathrm{tr}(M)$	trace of the matrix M. 6
$\bar{u}$	the empirical distribution $\frac{1}{n}u$ of a vector of counts u. 126
U_m	the group of unitary matrices over $\mathbb{C}$. 11
$\mathfrak{U}_m(\mathbb{K})$	the group of unipotent upper triangular matrices over $\mathbb{K}$. 10
V	usually a $\mathbb{K}$-vector space. 5

$v^{[2]}$	the vector $(\|v_1\|^2, \ldots, \|v_m\|^2)$ for $v \in \mathbb{C}^m$. 46
$v_{i,+,+}$	the i^{th} slice sum (with respect to the first direction) of a 3-tensor v. 6
$v_{i,j,+}$	the (i,j) "slice" sum (with respect to first and second direction) of a 3-tensor v. 6
$v_{+,+,+}$	sum over all entries of a 3-tensor v. 6
V_ω	the weight space of a representation for weight ω. 28
x_+	sum over all entries of the vector x. 6
$\mathfrak{X}(G)$	the character group of a complex algebraic group G. 13
$\mathfrak{X}_{\mathbb{R}}(G)$	the group of characters defined over $\mathbb{R}$. 13
$\mathfrak{X}_{G \cdot Y}$	a semigroup in the context of Popov's Criterion. 60
$[X, Y]$	Lie bracket of the Lie algebra elements X and Y. 18
$Y^{(i)}$	the i^{th} row of a matrix Y. 133
$Y^{(i \cup \mathrm{pa}(i))}$	the sub-matrix of Y with rows indexed by node i and its parents. 133
$Y^{(\mathrm{pa}(i))}$	the sub-matrix of Y with rows indexed by the parents of i. 133

Index