Normal Forms and Squeezing in Continuous Variable Quantum Information Theory

Martin Peter Idel

Zentrum Mathematik M5

Technische Universität München

A thesis submitted for the degree of

Doctor rerum naturalium

February 2017

TECHNISCHE UNIVERSITÄT MÜNCHEN
ZENTRUM MATHEMATIK

Lehrstuhl für Mathematische Physik

Normal Forms and Squeezing in Continuous Variable Quantum Information Theory

Martin Peter Idel

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzende: Prof. Dr. Caroline Lasser

Prüfer der Dissertation:

1. Prof. Dr. Michael M. Wolf
2. Prof. Dr. Jens Eisert

Die Dissertation wurde am 29.09.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 17.01.2017 angenommen.

Acknowledgements

Firstly, I would like to thank Prof. Michael Wolf for giving me the opportunity to write this thesis in his research group.

Secondly, I would like to thank my other coauthors and collaborators on the papers which constitute the main part of this thesis. I especially want to thank Robert König for suggesting an interesting topic and for the very enjoyable collaboration. I also thank him for his thorough reading and valuable comments on the introduction of this thesis.

Furthermore, I would like to thank my (current as well as former) colleagues at M5 for the very agreeable work environment, the many after-lunch debates and the great group retreats as well as their enduring patience when I mentioned "Sinkhorn normal forms". I especially thank Javier Cuesta for reading the introduction of this thesis.

Last but not least, I would like to thank my family and my partner Bettina Schleiermacher for their support even when I was very unhappy about the progress of this work or other matters related to being a PhD student.

List of contributed articles

I) M. Idel, D. Lercher, M. M. Wolf,
An operational measure for squeezing
Journal of Physics A: Mathematical and Theoretical, vol. 49, no. 44, Oct. 2016

II) M. Idel, S. Soto Gaona, M. M. Wolf,
Perturbation Bounds for Williamson's Symplectic Normal Form,
submitted to Linear Algebra and Its Application, arXiv:1609.01338 [math.SP]

III) M. Idel, R. König,
On additive Gaussian quantum channels,
submitted to Quantum Information and Computation, arXiv:1608.04305 [math-ph]

IV) M. Idel, M. M. Wolf,
Sinkhorn normal form for unitary matrices,
Linear Algebra and its Applications, vol. 471, pp. 76-84, Apr. 2015

V) M. Idel
A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps,
preprint, arXiv:1609.06349 [math.RA]

For all articles I am principal author.

Contents

1

Introduction

Normal forms are invariants under (group) operations, which help to simplify classification of the objects they refer to. Classifications can reduce the number of cases to be considered in a given problem and organise approaches, which allows to tackle hard and interesting problems. Therefore they are ubiquitous in mathematics and mathematical physics and occur at any level, from matrix diagonalisation to Dynkin diagram classification to graph canonisation.

This thesis is concerned with the introduction, extension, and application of normal forms mostly in the context of squeezing in continuous variable (CV) quantum information theory. Quantum information theory is concerned with the transmission and storage of information (for instance classical information encoded into bitstrings) in quantum mechanical systems. Such tasks arise naturally in real life, when two people who are spatially separated want to communicate with each other. In that case, they need to use a medium in order to communicate such as telephones connected by cables. Since communication will be disturbed by problems such as imperfections in the production process of the telephones or animals chewing on the cables, it is important to consider robust encondings to transmit information nonetheless. In information theory, such communication devices are modeled as "channels", maps that transform an input into a possibly randomised output based on some noise model. Trying to send information despite the fact that these channels may transform the message seemingly beyond recognition is called "coding". In recent decades, it has transpired that quantum systems such as polarised photons can also be used to transmit information. The study of "quantum channels" and coding is interesting for several reasons, for instance: Quantum channels offer the advantage that the information can be physically secured from eavesdroppers, whereas secure information transmission over classical channels relies on unproven "hardness" results for number theoretic

problems (e.g. finding prime factorisations for very large numbers is next to impossible in short time). Furthermore, smaller and smaller systems may require the study of quantum mechanics just by the fact that they are too small to be sufficiently well-described by classical mechanics. Since light is already one of the most useful information carriers in current technology it seems natural to use quantum mechanical descriptions of light and try to encode information with quantum optical devices. This is one of the applications of the subfield of continuous variable quantum information that is most promising for applications - in fact, commercial applications already exist (see the Swiss company ID Quantique).

Assume for a moment we wanted to build a quantum telephone. A very rough path towards good quantum communication includes the following steps:

1. Identify physical systems and operations that can be used for coding and transmission and model the systems including noise models.

2. Calculate information capacities (i.e achievable error-free transmission rates) of realistic channels.

3. Create concrete coding schemes to implement communication protocols including potential error correction.

4. Prove the robustness of schemes. This is necessary to ensure that small differences in the model will not completely alter information-transmission behaviour.

5. Implement them in a lab.

While the first and last step are purely physical, the other steps rely only on the underlying mathematical model. The study of normal forms enters the picture at various points:

1. Calculating channel capacities turns out to be a hard task in quantum information theory. No single letter formula is known to date for most channels and the lack of additivity (cf. [SY08]) renders a computable formula which holds for all channels impossible. Recent advances in calculating channel capacities in CV systems (see [GGPCH14]) relied on a normal form for a large class of quantum channels showing that such channels are always a concatenation of channels of a certain type.

2. Channels can be seen as states using the Choi-Jamiolkowski isomorphism; therefore it is also beneficial to study normal forms for states. Instead of studying the channel directly, normal forms of states can help to see how a channel acts. For instance, the minimum

output entropy of a channel, which is connected to its capacity, relies only on eigenvalues of the output.

3. Implementing coding schemes and error correcting codes requires the implementation of large classes of (unitary) operations. Usually, these are implemented by using just a small number of "gates" which can then be perfected in the lab and reused. Finding such gate systems or decomposing unitaries into gates can benefit from the invention of normal forms.

4. Robustness is usually implied by continuity of the normal forms.

1.1 Summary

In this dissertation I consider normal forms from several perspectives including proving new normal forms, proving the stability of known normal forms, and applying normal forms to study information theoretic problems. The main focus is twofold: The application of normal forms to questions related to squeezing, and the development of normal forms for quantum channels (mostly connected to squeezing).

Squeezing is a quantum optical operation which requires the use of nonlinear media and is considered a very hard task, also because squeezed states decohere quickly. By studying squeezing in various contexts I provide tools to answer questions about how much squeezing is necessary to perform a given task such as the creation of a quantum state. Squeezing is also closely related to entanglement, arguably the most important resource in quantum information theory. In addition, squeezing can itself be seen as a resource with connections to general resource theories.

In this section I will briefly sketch the results of each paper and how it relates to the overarching theme of the thesis and the other articles written during my graduate studies.

An operational measure of squeezing The article studies squeezing in itself and as a resource. The main goal is to answer the following question: If we only allow single-mode squeezing or single-mode squeezed resource states, how much squeezing is needed to prepare an arbitrary (Gaussian) state? As a preliminary result, we also cover the question how much squeezing is necessary to implement an arbitrary Gaussian unitary.

The question of implementing arbitrary Gaussian unitaries is translated to a question of decomposing symplectic matrices and we prove that the optimum is achieved by the Euler

decomposition, a well-known normal form of symplectic matrices. This was to be expected but has never been formally proven. We prove the result also for arbitrary paths on the symplectic group using results from ordinary differential equations in [Son98]. We then propose multi-mode squeezing measures, which turn out to be convex and superadditive. The main technical tool in this part is a matrix version of the Cayley transform from complex analysis.

Finally, we prove that the measure indeed answers the problem of minimising squeezing. Unfortunately, we have not found a simple and analytically computable formula for the minimisation but instead provide a MATLAB program which can compute the optimum approximately. The approximation was shown to be very precise in various numerical tests. To assess the performance of this algorithm, we also provide lower and upper bounds for the measure. These bounds are in part an application of spectral theory and Williamson's normal form, a symplectic diagonalisation of positive matrices.

To my knowledge, this is the first work which considers multi-mode squeezing measures and not only the smallest eigenvalue of the covariance matrix as in [KGLC03]. It is not clear how relevant the measures in the paper are for experimental purposes since single-mode squeezers might not capture what is actually done in the lab. We give a short discussion of this question and provide results about which modifications are possible without the need of new proofs. We are confident that the tools of this paper can be used in subsequent developments of squeezing measures and the resource theory of squeezing.

Perturbation Bounds for Williamson's Symplectic Normal Form This article grew out of the Bachelor thesis of Sebastián Soto Gaona, who proved that the symplectic spectrum, the diagonal entries of Williamson's normal form, are stable with respect to small perturbations. A similar bound also appeared in [BJ15]. We extend these results to give a full discussion of the perturbation theory of Williamson's normal form.

We give bounds on the stability of the symplectic spectrum and prove that the scaling of the constants, which depend on the condition number of the matrices involved, cannot be improved very much. We prove that the diagonalising matrix is stable if the spectrum is nondegenerate and give a counterexample otherwise. Finally, we prove that if S is a diagonlising matrix, then $S^{-T}S^{-1}$ is stable. We mainly employ techniques from usual perturbation theory (cf. [Bha96]) to prove our stability results.

This result is actually needed to prove termination of the algorithm in the previous article: When giving the MATLAB program, the starting point for the convex optimisation algorithm

is exactly $S^{-T}S^{-1}$ for some matrices. If this combination were not stable, the starting point could turn out to be an infeasible point in the constraint set of the algorithm and it is not clear that the algorithm converges.

Furthermore, the normal form is very important in continuous variable (CV) quantum information because the entropy can be expressed solely in terms of the symplectic spectrum. Therefore, the stability of Williamson's normal form will be interesting for the robustness of certain protocols in CV quantum information.

On additive Gaussian quantum channels This paper studies so called *additive Gaussian channels*, a class of channels consisting primarily of environment modes and beam splitters which couple the environment to the system. We approach the subject from the angle of the resource theory of squeezing: Which channels can be implemented using only passive (i.e. non-squeezing) unitaries but squeezed environments?

We provide a classification of all the channels with squeezed and unsqueezed environment and passive unitaries and in addition provide a normal form which identifies them as concatenations of beam splitters and Gaussian unitaries. Given our measures for squeezing from the first article, we could now ask the question of which states can be prepared using how much squeezed resources.

In addition, given the recent advances on computing channel capacities for covariant channels (cf. [GGPCH14] and similar papers), additive Gaussian channels might be a natural class to extend the results. They retain part of the properties of covariant channels which made them attractive, while essentially only adding the beam splitter with a squeezed environment mode (Efforts to extend the results of [GGPCH14] have however been futile so far).

Sinkhorn normal form for unitary matrices This article studies several new normal forms for unitary matrices, proving a conjecture by [DVDB14a]. The main tool for the proof is a deep result in symplectic geometry stating that a displacement of the so called Clifford torus from itself using Hamiltonian symplectomorphisms is impossible. It transpired later that the same proof had been used in a different context in [KJR14].

We use the normal form to derive a number of other decompositions of unitary matrices. Such decompositions can be interesting for coding schemes, where one constructs unitaries out of a limited number of gates. In our case, the normal form implies that it is enough to be able to implement phase shifters and Fourier transforms on any number of modes in order to

generate all (Gaussian) unitaries. In addition, the normal forms reveal structural properties of the unitary group which might help to solve open questions concerning unbiased bases or (un)certainty relations.

A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps This review is a condensed version of everything I have come to know about Sinkhorn's normal form. Since it is a review, it contains hardly any new material. It closes a gap in the literature because many results have been rediscovered several times and the general literature covers at least 200 papers over almost 80 years. A full overview about different approaches has never been achieved since the 70s. Despite the length and detail, it is focused on the mathematical questions surrounding the problem and only skims applications to give pointers to relevant literature. The article originated in the original research done in my Master's thesis and the paper on the Sinkhorn normal form for unitary matrices.

The topic of Sinkhorn normal forms touches upon quantum mechanics in several ways: We already discussed the normal forms for unitary matrices that were developed in Article IV and its connection to CV quantum information. The original Sinkhorn theorem for positive matrices was used by Aaronson in [Aar05] to study certain hidden-variable models. Most importantly, the extension of Sinkhorn's theorem to positive maps provides new normal forms for arbitrary quantum channels for finite dimensional systems. In this sense, the largest part of this paper is rather different from the rest of the thesis, as it concerns mostly finite dimensional and not continuous variable quantum information. However, it fits neatly into the discussion about normal forms and their impact on the quest to quantum information processing. I was interested in the problem in the hope to apply it to certain classes of quantum channels.

1.2 Outline

The next section introduces the mathematical framework underlying the papers. After introducing basic notation, I will review normal forms in symplectic geometry which are applied in the articles. A second section gives an introduction to continuous variable quantum information, in particular Gaussian states. I introduce resources and resource theories in a third section, sketching the best known example of entanglement theory and its connection to squeezing.

This chapter is followed by the articles ordered according to the list preceding this introduction. The list is ordered contextually to highlight connections between papers. Each article

is introduced by a short summary highlighting the technical advances of the paper as well as a short description of my own contribution to the results.

2

Mathematical foundations

This chapter explores the mathematical foundations underlying my work and the connections of said work to the literature. As the title of the dissertation suggests, there are three main topics to be explored: Normal forms, continuous variable (CV) quantum information, and squeezing as a particular subproblem of CV quantum information. As normal forms are ubiquitous in mathematics, we will not dedicate a complete chapter to them but rather treat them whenever they appear.

Since all topics explored in my work use aspects of symplectic geometry, we will fix notation and then start with a quick recapitulation of symplectic geometry. After that, we introduce quantum mechanics and CV quantum information before discussing resources and resource theories. Squeezing can be considered as a resource theory, which we do peripherally in Paper I. Since the topic of resource theories has drawn a lot of interest in recent years, it is beneficial to study squeezing from this perspective and connect it to the existing body of work in resource theories.

Notation and remarks Throughout this introduction, we use the notation $A \geq B$ if $A - B$ is positive semidefinite. Furthermore, $\mathcal{H}$ will always be a separable Hilbert space and $\mathcal{B}(\mathcal{H})$ will denote the bounded operators on that space, while $\mathcal{S}_1(\mathcal{H})$ denotes the Schatten-1 or trace-class operators. Furthermore, the Hermitian conjugate is denoted by †, while the adjoint map for maps $\mathcal{T} : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$ is marked with a *.

In addition, the identity operator will always be denoted by $\mathbb{1}$ or $\mathbb{1}_n$, while the standard symplectic form $\sigma_{2n} \in \mathbb{R}^{2n\times 2n}$ is given by

$$\sigma_{2n} := \begin{pmatrix} 0_n & \mathbb{1}_n \\ -\mathbb{1}_n & 0_n \end{pmatrix}. \tag{2.1}$$

None of the material in this section is original and most of it is fairly basic. Nevertheless, I tried to give references for most results which are not covered in standard undergraduate courses in physics or mathematics.

2.1 Symplectic geometry with a focus on linear theory

Symplectic geometry arose from the study of Hamiltonian mechanics: Phase space with its canonical variables (q, p) is a natural symplectic manifold equipped with the Poisson bracket. Symplectic geometry therefore allows to study integrable systems of classical mechanics geometrically.

2.1.1 Symplectic manifolds and applications to quantum mechanics

Given a differentiable manifold M, a *symplectic form* ω is a closed non-degenerate differential 2-form and the tuple (M, ω) is known as a *symplectic manifold*. If $M = \mathbb{R}^{2n\times 2n}$, then the matrix σ_{2n} from Equation (2.1) defines such a symplectic form (also called the *standard form*): It is a nonsingular skew-symmetric matrix (i.e. a nondegenerate 2-form), which is closed as all forms on $\mathbb{R}^{2n\times 2n}$ are closed. The tuple $(\mathbb{R}^{2n\times 2n}, \sigma_{2n})$ defines the linear symplectic space which can be studied via *linear symplectic geometry* (for further information see [MS98, dG06]).

Given a symplectic manifold (M, ω), one can ask for diffeomorphisms $\phi : M \to M$ whose pullback (i.e. the natural map on differential forms derived from ϕ) leaves ω invariant. Such maps are called *symplectomorphisms*. A symplectomorphism is *Hamiltonian* if it is derived as the flow of an exterior derivative of a smooth (Hamiltonian) function. There are at least two remarkable facts about symplectomorphisms:

1. Every symplectic form is locally diffeomorphic to the standard form (Darboux' theorem [Dar82]).

2. If $B(r)$ is the Euclidean ball of radius r and $Z(R) := \{z = (x_1, y_1, x_2, y_2, \ldots) \in \mathbb{R}^{2n} | x_1^2 + y_1^2 \leq R\}$, then any symplectic embedding $B(r) \hookrightarrow Z(R)$ implies $r \leq R$ (Nonsqueezing theorem [Gro85]).

Aside from these two theorems, symplectic geometry was largely inspired by two broad conjectures known as *Arnold conjectures* (cf. [MS98], Chapter 11):

Conjecture 2.1.1. *Let (M, ω) be a symplectic manifold. Then every non-degenerate Hamiltonian symplectomorphism of M has at least as many fixed points as the number of critical poitns of a Morse function.*

This conjecture is interesting from the physical perspective because fixed points of Hamiltonian symplectomorphisms are stationary orbits of the associated Hamiltonian. It is interesting from a mathematical perspective because it connects symplectic topology to Morse theory. The second conjecture concerns so-called *Lagrangian submanifolds*. A submanifold L of a symplectic manifold (M, ω) is called *Lagrangian*, if $\omega|_L = 0$ (the symplectic form vanishes on the submanifold), and $\dim(L) = \dim(M)/2$. These manifolds also appear as flow manifolds of Hamiltonian functions in classical mechanics (cf. [hf11]).

Conjecture 2.1.2 ([MS98], Conjecture 11.17)**.** *Let $L_0, L_1 \subset M$ be compact Lagrangian manifolds of a symplectic manifold (M, ω). Suppose that ψ_t is an isotopy derived from a compactly supported Hamiltonian function (precise definition in the reference) such that $\psi_0 = \mathrm{id}$ and $\psi_1(L_0) = L_1$. Then L_0 and L_1 must have at least as many intersection points as a Morse function on L_0 has critical points.*

This conjecture is also known as *Arnold-Givental conjecture*. Both conjectures have generated the invention of powerful tools such as Floer homology (see [Sal99]). While a weaker form of the first conjecture has been proven (cf. [LT98, FO99]), the second conjecture remains open in most interesting cases.

As a special case that [KJR14] and I independently used in applications, let $\mathbb{C}P^n$ be the complex projective space of n complex dimensions and $T^n := \{[x_0, \ldots, x_{n+1}] | |x_0| = \ldots = |x_n|\}$ the Clifford torus. Then:

Theorem 2.1.3 ([BEP04], Theorem 1.3)**.** *The Clifford torus $T^n \subset \mathbb{C}P^n$ cannot be displaced from itself by a Hamiltonian isotopy (i.e. ψT^n must intersect T^n for any Hamiltonian isotopy ψ).*

This is a special case of the Arnold conjecture because it states that two Clifford tori which are related by a Hamiltonian isotopy must always intersect since the number of critical points of the torus is always larger than zero (another proof is given in [Cho04]).

The theorem can be applied to the problem of finding mutually unbiased bases (cf. [AB15, FR15]) and (un)certainty relations (cf. [KJR14, PRC$^+$15]), as well as to find normal forms of unitary matrices (cf. [DVDB14a, DVDB14b, DVDB16]). However, all these problems have an intrinsically linear structure. The introduction of nonlinear symplectic geometry is only necessary because a linear algebra proof of Theorem 2.1.3 for linear symplectomorphisms remains elusive. For most other applications in this thesis, it suffices to consider linear symplectic geometry, which we now review in slightly more detail.

2.1.2 The linear symplectic group

For the linear symplectic space, the linear symplectomorphisms form a matrix group called the *(real) symplectic group* denoted by $Sp(2n,\mathbb{R}) \in \mathbb{R}^{2n\times 2n}$. Its matrices are given by the property $S^T\sigma_{2n}S = \sigma_{2n}$. The symplectic group has been studied extensively as one of the fundamental Lie groups. A good overview for quantum physicists is given in [ADMS95a]. Since we are particularly interested in normal forms, we want to study some normal forms involving the symplectic group.

First, let $O(2n,\mathbb{R})$ be the real orthogonal group, then we define the following subsets of $Sp(2n)$:

$$\begin{aligned} K(n) &:= Sp(2n,\mathbb{R}) \cap O(2n,\mathbb{R}) \\ Z(n) &:= \left\{\mathrm{diag}(\mathbb{1}_{i-1}, s, \mathbb{1}_{n-i-1}, \mathbb{1}_{i-1}, s^{-1}, \mathbb{1}_{n-i-1}) | s \geq 0, i = 1,\ldots,n\right\} \\ \Pi(n) &:= \{S \in Sp(2n,\mathbb{R}) | S \geq 0\} \end{aligned}$$

The first subset is an orthogonal-symplectic subgroup of $Sp(2n)$, which is also the maximal compact subgroup. The second set of matrices contains the single-mode squeezers, where the name will be explained later. It generates a maximal abelian subgroup of $Sp(2n)$. Finally, the last set is the set of positive semidefinite symplectic matrices.

Since $Sp(2n)$ is a Lie group, it has a Lie algebra $\mathfrak{sp}(2n)$:

Proposition 2.1.4. *The Lie algebras of $Sp(2n)$, $K(n)$ and the subset $\pi(n) \subset \mathfrak{sp}(2n)$ corresponding to the set $\Pi(n)$ under the exponential maps are given by*

1. $\mathfrak{sp}(2n) := \{T \in \mathbb{R}^{2n\times 2n} | \sigma_{2n}T + T\sigma_{2n} = 0\}$*, Lie algebra of $Sp(2n)$,*
2. $\mathfrak{k}(n) := \{A \in \mathbb{R}^{2n\times 2n} | A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}, a = -a^T, b = b^T\}$*, the Lie algebra of $K(n)$,*
3. $\pi(n) := \{A \in \mathbb{R}^{2n\times 2n} | A = \begin{pmatrix} a & b \\ b & -a \end{pmatrix}, a = a^T, b = b^T\}$*, the subspace of the Lie algebra $\mathfrak{sp}(2n)$ corresponding to $\Pi(n)$,*

where the Lie bracket is always given by the commutator.

Using standard arguments from Lie algebras, one can now derive a number of decompositions of the Lie algebra, which in turn define decompositions of the Lie group (cf. [Hel01]). We only list two, which are most important for us:

Proposition 2.1.5 (Polar decomposition [ADMS95a])**.** *For every symplectic matrix $S \in Sp(2n)$ there exists a unique $U \in K(n)$ and a unique $P \in \Pi(n)$ such that $S = UP$.*

Proof. Since $\det(S^T\sigma_{2n}S) = \det(\sigma_{2n})$, we have $\det(S) = \pm 1$ (in fact, $\det(S) = 1$ as seen in the proof of the Euler decomposition). In particular, S is invertible and we can write $S = S(S^TS)^{-1/2}(S^TS)^{1/2}$. By construction, setting $P := S^TS^{1/2}$, $P \in \Pi(n)$ and $U := S(S^TS)^{-1/2} \in Sp(2n)$. However, we also have $U \in O(2n)$ since $UU^T = S(S^TS)^{-1}S^T = \mathbb{1}$. The uniqueness of this decomposition follows from the uniqueness of the (positive) square root of a positive definite matrix. □

Proposition 2.1.6 (Euler decomposition, Bloch-Messiah decomposition [ADMS95a])**.** *Let $S \in Sp(2n)$, then there exist $K, K' \in K(n)$ and $A = A_1 \cdots A_n$ with $A_i \in Z(n)$ for each i such that $S = KAK'$.*

Proof. Using the polar decomposition, we can write $S = UP$ with $U \in K(n)$ and $P \in \Pi(n)$. Since P is positive semidefinite, we can diagonalise it. The eigenvalues come in pairs $\pm\lambda_i$: For every eigenvector v of P to eigenvalue λ, $\sigma_{2n}v$ is also an eigenvector of P to the eigenvalue λ^{-1} as $\lambda P^T\sigma_{2n}v = \sigma_{2n}v$ since P is symplectic. If $\{v_i\}$ is an orthonormal set of eigenvectors for the positive eigenvalues of P, then $K = \begin{pmatrix} v_1 & \dots & v_n & \sigma_{2n}v_1 & \dots & \sigma_{2n}v_n \end{pmatrix}$ satisfies $K^T\sigma_{2n}K = \sigma_{2n}$, hence $K \in K(n)$. Furthermore, K diagonalises P where the resulting matrix $D = K^TPK = \operatorname{diag}(\lambda_1, \dots, \lambda_n, \lambda_1^{-1}, \dots, \lambda_n^{-1})$ is a product of matrices in $Z(n)$. In total, we have $S = UKDK^T$ and thus the Euler decomposition. □

There exist many more normal forms, such as the Cartan decomposition or the Iwasawa decomposition which we leave out here and refer the reader to [ADMS95a, Hel01].

2.2 Continuous variable quantum information

Classical information theory studies how to manipulate information, for instance in order to store or transmit it. It is based on the concept of message (usually a bitstring) and channel (usually a stochastic matrix transforming a bitstring into another bitstring probabilistically).

Following the seminal paper of Shannon [Sha48], the (classical) information content of a message can be quantified using the *Shannon entropy*. The operational interpretation of this measure is known as the *Shannon Coding Theorem*: The Shannon entropy of a message is the least amount of space (mostly measured in bits) necessary to store a message with lossless compression. For information transmission, the *Noisy Channel Theorem* of Shannon provides a measure computing the maximal asymptotic capacity of a classical channel. Here, *capacity* means information content per bit such that the message is transmitted without error in the limit of many channel uses. These two theorems form the cornerstone of information theory as they provide bounds for its most important tasks: Compression for storage and coding for sending information (for an overview, see [CT06]).

Quantum information extends this theory from classical to quantum systems, which are modeled by so called *quantum states* (quantum messages) and *CPTP-maps* (quantum channels). This is interesting for example as it allows physically secure information transmission (see [NC00] for an introduction and [BB84, Eke91] for specifics on quantum cryptography). For information content, the Shannon entropy is replaced by the *von Neumann entropy* and the coding theorem becomes the *Schumacher compression theorem* (cf. [Sch95]), while the noisy channel theorem is replaced by the *HSW theorem* (cf. [SW97, Hol98]) for classical and the *LSD theorem* (cf. [Dev05]) for quantum information. However, many questions remain, such as how to efficiently compute *quantum capacities* of channels or which coding schemes are simple but achieve approximately optimal results.

Quantum information is an abstract framework which can be implemented in many ways. For instance, atomic spins might be interesting for storage while laser light is already used for classical communication although it also provides a perfectly quantum system. While spins can be modeled in finite dimensional systems, light needs infinite dimensions. Continuous variable quantum information is the subfield of quantum information which studies infinite dimensional systems with finitely many degrees of freedoms such as modes of light in a cavity. The name derives from the fact that the observables studied have a continuous spectrum.

There exist many good overviews of continuous variable quantum information such as [BvL05, WPGP$^+$12, ARL14]. Since the topic is huge, we will only review a selection of facts and leave out many important results. The reader may wish to consult the cited overviews.

2.2.1 Quantum mechanics in infinite dimensions

As quantum mechanics is a physical theory, it must be founded in experiments. An experiment can roughly be modelled in two steps: A system is set up and controlled (preparation) and then some part of it or the complete system is measured (measurement), i.e. some numbers are produced. Sometimes, it is beneficial to include a third step in between preparation and measurement, where the system evolves (un)controlled (time evolution). The goal of a physical theory is to predict the outcome given the preparation procedure.

In quantum information, the preparation procedure can roughly be seen as an encoding of states, the time evolution could be a storage time or a signal transmission, and the measurement is equal to the decoding scheme (maybe excluding postprocessing).

2.2.1.1 The foundations of quantum mechanics

We specify the mathematical model underlying quantum mechanics by specifying preparation, time evolution, and measurement:

The preparation procedure is subsumed in the concept of a *quantum state*. Two quantum states are the same if the predictions for any measurement one can perform on the two systems are the same. Hence we can see quantum states as equivalence classes of preparation procedures which capture the result of a real system preparation.

Postulate 2.2.1. A *quantum state* or *density matrix* ρ is an operator $\rho \in \mathcal{S}_1(\mathcal{H})$ over a separable Hilbert space $\mathcal{H}$ such that $\rho \geq 0$ and $\mathrm{tr}(\rho) = 1$. The Hilbert space $\mathcal{H}$ contains the description of the system. If two systems $\mathcal{H}_1$, $\mathcal{H}_2$ are considered, the combined system is given by $\mathcal{H}_1 \otimes \mathcal{H}_2$. Conversely, the restriction of a state $\rho \in \mathcal{S}_1(\mathcal{H}_1 \otimes \mathcal{H}_2)$ to one of its subsystems is given by the partial trace.

Furthermore, a *pure quantum state* is a normalised element $v \in \mathcal{H}$ or equivalently (modulo phase) the rank one density matrix $vv^\dagger$.

Since any positive definite matrix can be decomposed as a weighted sum of rank one matrices, the set of states is a convex set with pure states as extremal points.

Measurement procedures in quantum mechanics are intrinsically probabilistic. A measurement of a state ρ is specified by a set of possible outcomes and a probability distribution over the outcomes (depending on ρ). For quantum mechanics, such measurements turn out to be positive operator valued measures (see for instance [Hol11], Prop. 1.6.1).

Postulate 2.2.2. A *positive operator valued measure* (POVM) is an index set I with a map $P : \mathcal{P}(I) \to \mathcal{B}(\mathcal{H})$ from the Borel set of I to the bounded operators on $\mathcal{H}$ such that $P(J) \geq 0$ for all $J \subseteq I$ and $P(I) = \mathbb{1}$. The probability of measuring a value in $J \subseteq I$ of state ρ is then given by $p(J) = \mathrm{tr}(P(J)\rho)$.

A very easy way to obtain POVMs is to take a self-adjoint (not necessarily bounded) operator A on $\mathcal{H}$ with the outcomes defined by the spectrum of A and the measure defined by the spectral projections. Such measures are also called *projection valued measures* (PVM).

Finally, the time evolution of a system is governed by the most famous equation of quantum mechanics, the *Schrödinger equation* (or *von Neumann equation* for density matrices):

Postulate 2.2.3. Every closed quantum mechanical system can be specified by a self-adjoint linear operator H on $\mathcal{H}$ such that the time evolution of a quantum state $\rho \in \mathcal{S}_1(\mathcal{H})$ is given by

$$i\partial_t \rho = [H, \rho] \tag{2.2}$$

where we use natural units with $\hbar = 1$.

Very often, the Hamiltonian of a system can be guessed from the classical equations of the same system (that is, if a classical description exists). The formal ideas behind this are known as *canonical quantisation*. The basic idea is to start with the so called Hamiltonian function of the Hamiltonian formalism of classical mechanics and upgrade variables (except the time variable) to operators. To illustrate this procedure, the variables of position and momentum (q and p) are upgraded to operators of position and momentum (Q and P). Instead of the Poisson bracket, position and momentum must now fulfil the commutator equation

$$[P, Q] = \frac{i}{2}\mathbb{1}. \tag{2.3}$$

This commutator equation is at the heart of what is known as *Heisenberg's uncertainty principle* and is necessary from the fact that P is the Fourier transform of Q. It turns out (cf. [BEH08], Section 8.2) that this equation can only be fulfiled if both P and Q are unbounded operators, thus $\mathcal{H}$ has to be an infinite dimensional Hilbert space.

2.2.1.2 Quantum mechanics in phase space

One of the most important systems for information carriage in quantum information is light. To have a fully quantum theory of light, the electromagnetic field must be quantised. This can be done using a collection of non-interacting quantum harmonic oscillators with different frequencies. This is a standard *second quantisation approach* in quantum field theory (cf. [SZ97, Lou00]). For our purposes, it is sufficient to consider a finite number of these frequencies (also called "modes") to avoid further mathematical complications. In reality, such a system can arise as an electromagnetic field in a cavity.

Each oscillator/mode then defines annihilation and creation operators a_k and $a_k^\dagger$ such that the Hamiltonian of the full electromagnetic field is given by

$$H = \sum_{k=1}^{n} H_k, \qquad H_k = \omega_k \left(a_k^\dagger a_k + \frac{1}{2} \right) \tag{2.4}$$

where ω_k are the base frequencies of the oscillators (see for instance [ARL14]). As the force carriers of the electromagnetic field - the photons - are bosons, the annihilation and creation operators fulfil the CCR $\left[a_k, a_l^\dagger\right] = \delta_{kl}\mathbb{1}$ and $[a_k, a_l] = \left[a_k^\dagger, a_l^\dagger\right] = 0$ (this follows directly from the definition of $a_k, a_k^\dagger$ via a standard computation).

In order to connect with the previous section, we can set

$$Q_k := \frac{(a_k + a_k^\dagger)}{\sqrt{2}}, \quad P_k := \frac{(a_k - a_k^\dagger)}{i\sqrt{2}} \tag{2.5}$$

to obtain the so called *quadratures*, which behave just like position and momentum operators (fulfiling for instance (2.3)). If we write $(R_{2k-1}, R_{2k}) = (Q_k, P_k)$ and $R = (R_1, \ldots, R_{2n})$, then we have

$$[R_k, R_l] = i(\sigma_{2n})_{kl} \frac{\mathbb{1}}{2}. \tag{2.6}$$

Put together, (R, σ_{2n}) define a symplectic structure similar to the standard symplectic structure of classical Hamiltonian mechanics (cf. [Arn89]).

The standard quantum mechanical representation of the CCR is known as the *Schrödinger representation*, where each mode consists of a Hilbert space $\mathcal{H} = L^2(\mathbb{R})$, Q_k is the usual position multiplication operator and $P_k = i\partial/(\partial Q_k)$.

However, this representation has at least two problems:

- Since P, Q fulfil Equation (2.3), they must be unbounded operators on an infinite dimensional Hilbert space and are therefore difficult to work with.
- The representation is not unique.

It is therefore useful to consider bounded representations of the operators. Such a representation is the *Weyl system*

$$W_\xi := \exp(i\xi \cdot \sigma_{2n} R), \quad \xi \in \mathbb{R}^{2n} \tag{2.7}$$

which fulfils the Weyl relations

$$W_\xi W_\eta = \exp\left(-\frac{i}{2}\xi \cdot \sigma_{2n}\eta\right) W_{\xi+\eta}. \tag{2.8}$$

The operators are also called *displacement operators* or *Glauber operators*. By the Stone-von Neumann theorem (cf. [BEH08], Theorem 8.2.4), this definition is the unique strongly continuous and irreducible representation of the CCR algebra defined through R up to unitary equivalence (note that this statement only holds for finitely many modes).

Using the commutation relations, we can easily see that W corresponds to a translation in phase space:

$$W_\xi R_k W_\xi^* = R_k + \xi_k \mathbb{1} \quad \forall \xi \in \mathbb{R}^{2n}. \tag{2.9}$$

The Weyl system looks similar to the usual Fourier plane waves $\exp(ikx)$ and indeed, similar to the (commutative) Fourier transform, we can define the (noncommutative) *Weyl transform* to

obtain phase space functions for quantum states. Given a quantum state $\rho \in \mathcal{S}_1(\mathcal{H})$, we define the characteristic function

$$\chi_\rho(\xi) := \mathrm{tr}(W_\xi \rho), \quad \xi \in \mathbb{R}^{2n}. \tag{2.10}$$

Many results for Fourier transforms have similar versions for the Weyl transform. For instance, we have the following characterisation of quantum states:

Theorem 2.2.4 (Quantum Bochner-Khinchin, [Hol11] Theorem 5.4.1.)**.** *Let $\chi \in L^2(\mathbb{R}^{2n})$ be a function, then it is the characteristic function of a quantum state if and only if*

1. *$\chi(0) = 1$ and $\chi(\xi)$ is continuous at $\xi = 0$,*
2. *For every $m \in \mathbb{N}$, $\xi_1, \ldots, \xi_m \in \mathbb{R}^{2n}$ and $c_1, \ldots, c_m \in \mathbb{C}$ we have*

$$\sum_{k,l=1}^{m} c_k \bar{c}_l \chi_\rho(\xi_k - \xi_l) \exp\left(\frac{i}{2}\xi_k \cdot \sigma_{2n} \xi_l\right) \geq 0$$

The name derives from the classical Bochner-Khinchin theorem which characterises the Fourier transform of positive finite Borel measures on $\mathbb{R}$ which can be seen as classical states.

Finally, using the characteristic function one can define the well-known *Wigner function* (cf. [Wig32]) of a state ρ:

$$\mathcal{W}_\rho(\xi) := (2\pi)^{-2n} \int \exp(i\xi \cdot \sigma_{2n}\eta)\chi_\rho(\eta) d^{2n}\eta.$$

Necessary and sufficient conditions for a function to be the Wigner function of a state can be computed from the Quantum Bochner-Khinchin theorem, but they are much more difficult to state, so we omit them here and refer the readers to [NO86, CFZ14].

2.2.2 Gaussian states and covariance matrices

A very natural class of states are those where characteristic or Wigner function are Gaussian. Those states are therefore called *Gaussian states*. They are especially important in quantum optics as the coherent states of lasers are in particular Gaussian states. They are also important in mathematical physics as the coherent states, a subset of Gaussian states, form an overcomplete set in Hilbert space. Furthermore, thermal states of quadratic Hamiltonians are mixtures of coherent states and therefore Gaussian (cf. [Oli12]).

From classical probability theory, it is well-known that a Gaussian is characterised by its first and second moments. The same is true for Gaussian states, where the first and second

moments are defined as

$$d_k := \mathrm{tr}(\rho R_k) \tag{2.11}$$

$$\gamma_{kl} := \mathrm{tr}(\rho\{R_k - d_k\mathbb{1}, R_l - d_l\mathbb{1}\}_+) \tag{2.12}$$

with the anticommutator $\{\cdot,\cdot\}_+$. Similarly to classical characteristic functions, they can also be defined as the first and second partial derivatives of the (quantum) characteristic function. The first moments are sometimes called *displacement vector*, the matrix of second moments is called the *covariance matrix*. While a Gaussian state can have any displacement vector $d \in \mathbb{R}^{2n}$, the covariance matrix is more limited by Heisenberg's principle:

Theorem 2.2.5. *For any $\gamma \in \mathbb{R}^{2n\times 2n}$ there exists a quantum state ρ with covariance matrix γ if and only if $\gamma \geq \frac{i}{2}\sigma_{2n}$.*

Let us now study a number of normal forms of covariance matrices and Gaussian states. Given a covariance matrix γ of a Gaussian state, since $\gamma \geq i\sigma_{2n}/2$ implies in particular that γ is Hermitian we can of course study eigenvalues and eigenvalue decompositions. It turns out that this leads to the notion of squeezed states: Assume $\gamma = \mathrm{diag}(\lambda_1, \ldots, \lambda_{2n})$ for some values $\lambda_i \in \mathbb{R}$. The inequality $\gamma \geq i\sigma_{2n}/2$ then leads to $\lambda_i\lambda_{n+i} \geq 1/4$ for every $i = 1, \ldots, n$. This is just Heisenberg's uncertainty (2.3) in disguise. If $\lambda_i < 1/2$ or $\lambda_{n+i} < 1/2$, this implies that the uncertainty in this mode (P_i or Q_i) is lower than possible if the variance of P_i and Q_i were the same. The corresponding state is said to be *squeezed*.

Note that this does not contradict the non-squeezing theorem: In fact, from a mathematical perspective Heisenberg's principle is nothing else than the assertion that a quantum state must have a *symplectic capacity*, a notion made possible by the non-squeezing theorem (see [dGL09]). Hence we define:

Definition 2.2.6. Let γ be the covariance matrix of a Gaussian state. Then γ is called *squeezed* if it has an eigenvalue $\lambda < 1/2$.

Other definitions appear in the literature but are known to be equivalent (cf. [ADMS95b]).

An important normal form for covariance matrices is given by

Theorem 2.2.7 (Williamson's theorem [Wil36])**.** *Let $M \in \mathbb{R}^{2n\times 2n}$ be a positive definite matrix. Then there exists a nonnegative diagonal matrix $D \in \mathbb{R}^{n\times n}$ and a symplectic matrix $S \in Sp(2n)$ such that*

$$S^T M S = \mathrm{diag}(D, D). \tag{2.13}$$

The entries of D are often called symplectic eigenvalues *of M and they are the positive eigenvalues of $i\sigma_{2n}M$.*

Proof. We sketch the proof of [SCS99] similar to the sketch in my own papers:

Let $\mathrm{diag}(D, D) =: \tilde{D}$. Since D and M are positive definite, consider $S = M^{-1/2} K \tilde{D}^{1/2}$, where $K \in O(2n)$. By construction $S^T M S = \tilde{D}$. We need to choose K such that S is symplectic: This is equivalent to

$$K^T (M^{-1/2} \sigma_{2n} M^{-1/2}) K = \begin{pmatrix} 0 & D^{-1} \\ -D^{-1} & 0 \end{pmatrix}.$$

Using that $(M^{-1/2}\sigma_{2n}M^{-1/2})^T = -M^{-1/2}\sigma_{2n}M^{-1/2}$, we know that we can indeed find an orthogonal K achieving this construction since $iM^{1/2}\sigma_{2n}M^{1/2}$ is a Hermitian matrix and therefore diagonalisable by a unitary $U \in U(2n)$. Eigenvalues come in pairs $\pm\lambda_j$ with eigenvectors $x_j \pm i v_j$ for $j = 1, \ldots, n$ and $x_j, y_j \in \mathbb{R}^{2n}$ and hence $K = (x_1, \ldots, x_n, y_1, \ldots, y_n)$ using $\sigma_{2n} x_j = y_j$ and $\sigma_{2n} y_j = -x_j$. □

Symplectic matrices, by definition, leave the CCR invariant and are therefore an interesting set of physical operations (see also the discussion in Section 2.2.3). The symplectic spectrum is invariant under symplectic conjugation (by construction) and is therefore more important than the eigenvalue spectrum.

The Williamson normal form leads to a decomposition of Gaussian states known as *normal mode decomposition*

Theorem 2.2.8 (see for instance [Oli12])**.** *Let ρ be a density matrix of an n-mode Gaussian state. Then there exists a canonical basis in phase space with corresponding annihilation and creation operators a, $a^\dagger$ such that ρ takes on the form*

$$\rho = \bigoplus_{k=1}^{n} \frac{\exp(-\beta\omega_k(a^\dagger a + 1/2))}{\mathrm{tr}(\exp(-\beta\omega_k(a^\dagger a + 1/2)))}. \tag{2.14}$$

In other words, ρ is a tensor product of thermal states of different temperature in this basis, where $\beta\omega_k$ are given by $1 + 2(\exp(-\beta\omega_k) - 1)^{-1} = 2s_k$ with the symplectic eigenvalues s_k of the covariance matrix of ρ.

Proof. We first consider the thermal states of a one-mode quantum harmonic oscillator: For an inverse temperature β, those are given by the Gibbs state

$$\rho_{\mathrm{Gibbs}} = \frac{\exp(-\beta H)}{\mathrm{tr}(\exp(-\beta H))} = \frac{\exp(-\beta\omega(a^\dagger a + 1/2))}{\mathrm{tr}(\exp(-\beta\omega(a^\dagger a + 1/2)))}.$$

We can now use

1. $W_\xi = \exp(i/\sqrt{2}((\xi_1 + i\xi_2)a^\dagger - (\xi_1 - i\xi_2)a))$,
2. the Fock states $|n\rangle = (a^\dagger)^n/\sqrt{n!}|0\rangle$ form a basis,
3. $\langle n|W_\xi|m\rangle = \delta_{nm}\exp(-|\xi|^2/4)L_n(|\xi|^2/2)$ with the Laguerre polynomials L_n (cf. [Oli12]),

to obtain

$$\chi(\xi) = \mathrm{tr}(W_\xi \rho_{\mathrm{Gibbs}}) = c \exp(-(\xi_1^2 + \xi_2^2)(1 + 2(\exp(-\beta\omega) - 1)^{-1})/4)$$

with some constant c which does not depend on ξ and fixes the normalisation. Hence a thermal one-mode state is a Gaussian state with covariance matrix $\gamma = (1 + 2(\exp(-\beta\omega) - 1)^{-1})\mathbb{1}_1/2$.

Given any ρ, Williamson's normal form implies that there exists a basis where $\gamma_\rho = \bigoplus_{k=1}^n s_k \mathbb{1}_2$ with symplectic eigenvalues s_k and without any displacement. But this implies that ρ is of the form of Equation (2.14) with $1 + 2(\exp(-\beta_k\omega) - 1)^{-1} = 2s_k$. □

Williamson's normal form gives us an opportunity to describe all Gaussian pure states:

Corollary 2.2.9. *A covariance matrix $\gamma \in \mathbb{R}^{2n\times 2n}$ is a covariance matrix of a Gaussian pure state if and only if* $\det(2\gamma) = 1$ *or equivalently,* $2\gamma \in Sp(2n)$.

Proof. Note that a state ρ is pure if and only if $\rho^2 = \rho$. Using the basis where γ is diagonal and $d = 0$, we have that ρ corresponding to γ is a tensor product of thermal states. A simple calculation shows that a thermal state is pure if and only if $\beta_k \to \infty$ (at zero temperature) or else if and only if $d_k = 1/2$ for all $k = 1, \ldots, n$. □

Note that the set of Gaussian pure states is also the set of extremal Gaussian states.

Williamson's normal form is also useful to derive a simple function for entropy. Recall that the von Neumann entropy of a quantum state is given by the Shannon entropy of its spectrum (which immediately tells us that pure states have zero entropy). Similarly, on the level of covariance matrices, we obtain:

Theorem 2.2.10 ([HSH99])**.** *Let $\rho \in \mathcal{B}(\mathcal{H})$ be a Gaussian quantum state. Then its entropy is described by*

$$S(\gamma) = \sum_{k=1}^{n} \left(g\left(d_k + \frac{1}{2}\right) - g\left(d_k - \frac{1}{2}\right)\right) \tag{2.15}$$

where $g(x) = x\log(x)$ and the d_k are the symplectic eigenvalues of the covariance matrix γ_ρ of ρ.

Many other decompositions and normal forms can be described that might help in different circumstances. For instance, the well-known singular value decomposition has a symplectic analogue described in [Wol08].

2.2.3 Quantum Channels

We already stated that time evolution is governed by the Schrödinger equation. However, this requires knowledge of the total Hamiltonian and the state of the system, something which is

infeasible in most experiments. To remedy this we introduce the idea of *open quantum systems* and *quantum channels*. The idea is to divide the whole closed system into one part (called the "system") that is interesting in an experiment and can be manipulated and the rest (dubbed the "environment"). Consequently, a *quantum channel* is a function which sends a state of the system to a different state taking the environment into account.

Given a state of the system ρ and a state of the environment ρ_E and assuming that we know the time evolution $U = \exp(iHt)$ of the whole system such a channel would be

$$\mathcal{T}(\rho) = \mathrm{tr}_E(U(\rho \otimes \rho_E)U^\dagger). \tag{2.16}$$

Here, tr_E denotes the partial trace over the environment and ρ_E is the state of the environment. Usually, the assumption is made that the system starts out in a product state, which can be achieved by a careful state preparation, but we do not go into details here. It turns out that these maps $\mathcal{T} : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$ are *completely positive*, meaning that $\mathrm{id}_n \otimes \mathcal{T} : \mathbb{C}^{n\times n} \otimes \mathcal{B}(\mathcal{H}) \to \mathbb{C}^{n\times n} \otimes \mathcal{B}(\mathcal{H})$ is a positive map for any $n \in \mathbb{N}$. In order to maintain normalisation of the state, $\mathcal{T}$ must be trace-preserving, too.

Definition 2.2.11. A *quantum channel* $\mathcal{T} : \mathcal{S}_1(\mathcal{H}) \to \mathcal{S}_1(\mathcal{H}')$ is a completely positive trace preserving linear map. A *Gaussian channel* is a quantum channel that sends Gaussian states to Gaussian states.

We are particularly interested in Gaussian channels. An overview can be found in [EW07].

Since a Gaussian channel sends Gaussian states to Gaussian states and those are completely specified by the displacement vector and the covariance matrix, we can consider Gaussian channels on the level of covariance matrices:

Theorem 2.2.12. *An n-mode Gaussian channel $\mathcal{T}$ is given by a triple (X, Y, v) with $v \in \mathbb{R}^{2n}$ and $X, Y \in \mathbb{R}^{2n\times 2n}$ fulfiling*

$$Y \geq i\sigma_{2n} - iX^T\sigma_{2n}X \tag{2.17}$$

such that for any Gaussian state ρ with covariance matrix γ and displacement d we have

$$\gamma \mapsto X^T\gamma X + Y \qquad d \mapsto Xd + v. \tag{2.18}$$

For some applications, it can be beneficial to also study the set of completely positive maps and not just the (subclass) of quantum channels. A complete description of Gaussian completely positive maps and the yet bigger class of Gaussian positive maps is found in [GIC02, Fiu02]. The set is most commonly constructed using the Choi-Jamiolkowski isomorphism.

We have defined channels mostly as tracing out an environment. Clearly, it should be possible to add a (potential) environment to get a unitary channel again. This is the physical interpretation of what is known as *Stinespring dilation theorem*. For Gaussian channels, this was studied in detail in [CEGH08, CEGH11]. One of the main results is:

Theorem 2.2.13 ([CEGH08], Theorem 1)**.** *For any Gaussian bosonic channel (X, Y, v) there exists $l \leq 2n$ with $S \in Sp(2n + 2l)$ and $\gamma_E \geq i\sigma_{2l}/2$ such that*

$$(S \operatorname{diag}(\gamma, \gamma_E) S^T)_{2n\times 2n} = X^T \gamma X + Y \qquad \forall \gamma \geq i\sigma_{2n}, \tag{2.19}$$

where $(\cdot)_{2n\times 2n}$ denotes the restriction to the upper left $2n \times 2n$ principal submatrix.

For some applications, it is interesting to study these dilations for subclasses of channels, for instance those channels that have a pure or squeezed environment.

Many subclasses of quantum channels have been described in the literature, either because they are physically interesting or because they have a mathematical structure that can be exploited. Let us mention two such subclasses:

Gaussian unitary channels Closed system dynamics are always unitary following the fact that a Hamiltonian is (by definition) a self-adjoint operator which generates a unitary one-parameter time evolution via Stone's theorem (cf. [BEH08], Section 5.9). *Gaussian unitaries* are the subset of unitary channels that are Gaussian channels. We have the following characterisation (cf. [DVV77]):

Proposition 2.2.14. *A Gaussian unitary operator is completely specified by a tuple (S, v) with $S \in Sp(2n)$ and $v \in \mathbb{R}^{2n}$ which acts on the Weyl operators as*

$$W_\xi \mapsto \exp(i\xi \cdot v) W_{S\xi}. \tag{2.20}$$

In other words, a Gaussian unitary consists of a symplectic transformation of the modes (which translates to a transformation $\gamma \mapsto S^T \gamma S$ of the covariance matrix and $d \mapsto Sd$ on the displacement of a quantum state) and a displacement.

In quantum optics, there are several optical instruments which implement symplectic matrices. Using the Euler decomposition Proposition 2.1.6, one can implement any Gaussian unitary with the help of only single-mode squeezers (implementing operations in $Z(n)$), beam splitters and phase shifters (implementing operations in $K(n)$ using [RFP10]) and Weyl displacements. Other implementations are given by different normal forms of unitary matrices such as the Sinkhorn normal form [DVDB14b].

Gauge-Covariant and Contravariant Gaussian channels Similar to other theories, one can consider the transformation $a_k \mapsto \exp(i\phi)a_k$ on the creation (and annihilation) operators. This *gauge transformation* (a global phase) is given by the group of symplectic transformations $\{\exp(i\sigma_{2n}\phi)|\phi \in [0, 2\pi)\}$ and can also be defined on the level of the Weyl system (cf. [GHGP15]):

$$W(\exp(i\sigma_{2n}\phi)\xi) = U_\phi^\dagger W(\xi) U_\phi.$$

Using this transformation, one can define gauge invariant states or channels and study normal forms for these states (as in [HSH99]). We want to study a slightly broader class:

Definition 2.2.15. Let $\mathcal{T}$ be an n-mode Gaussian quantum channel. Then $\mathcal{T}$ is called gauge *covariant* (*contravariant*) if it fulfils the relation

$$\mathcal{T}\left(\exp\left(i\phi\sum_{i=1}^n a_i^\dagger a_i\right)\rho\exp\left(-i\phi\sum_{i=1}^n a_i^\dagger a_i\right)\right) = \exp\left(\pm i\phi\sum_{i=1}^n a_i^\dagger a_i\right)\mathcal{T}(\rho)\exp\left(\mp i\phi\sum_{i=1}^n a_i^\dagger a_i\right). \tag{2.21}$$

Gauge covariant channels have a very simple description in terms of maps on covariance matrices:

Proposition 2.2.16 (for instance [LGW13])**.** *An n-mode Gaussian channel $\mathcal{T}$ is gauge-covariant if and only if the corresponding triple (X, Y, z) fulfils $[X, \sigma_{2n}] = [Y, \sigma_{2n}] = 0$.*

Proof. The proof rests on the observation that $[X, \sigma_{2n}] = 0$ if and only if $[X, \exp(i\phi\sigma_{2n})] = 0$, which in turn implies that the channel commutes with the unitary group $\exp(i\phi a^\dagger a)$ if and only if (X, Y) commute with σ_{2n}. But this is true if and only if the channel is covariant. □

Any matrix $A \in \mathbb{R}^{2n\times 2n}$ which commutes with σ_{2n} is of the form

$$A = \begin{pmatrix} B & C \\ -C & B \end{pmatrix}$$

and can therefore be written as a matrix $\hat{A} := B + iC \in \mathbb{C}^{n\times n}$. The corresponding condition for complete positivity of a channel then translates to (cf. [HHW10])

$$\hat{Y} \geq \pm(\mathbb{1}_n - \hat{X}^\dagger\hat{X}).$$

The channels are called *quantum limited*, if equality is attained. In particular:

1. The *quantum limited attenuator* is a map with $\hat{X}^\dagger\hat{X} \leq \mathbb{1}$ and $\hat{Y} = (\mathbb{1}_n - \hat{X}^\dagger\hat{X})$.
2. The *quantum limited amplifier* is a map with $\hat{X}^\dagger\hat{X} \geq \mathbb{1}$ and $\hat{Y} = (\mathbb{1}_n - \hat{X}^\dagger\hat{X})$.
3. The *quantum limited gauge contravariant channels* are maps with $\hat{Y} = (\mathbb{1}_n + \hat{X}^\dagger\hat{X})$

This characterisation paves the way to a normal form of gauge covariant (and contravariant) channels, which states:

Theorem 2.2.17 (cf. [Hol15, GPNBL$^+$12])**.** *Any gauge covariant channel is a concatenation of a quantum limited amplifier and a quantum limited attenuator.*

Any gauge contravariant channel is a concatenation of a quantum limited amplifier and a quantum limited contravariant channel.

With these two examples, let us return to one of the major questions in quantum information: How well can one communicate with a given channel? The task is vaguely defined for the reason that multiple communication tasks can be evaluated: For instance, we can consider communication of bitstrings encoded into quantum states or communication of (instances of) quantum states themselves.

One of the most important capacities is the *classical capacity* (sending classical information over a quantum channel with asymptotically zero error) for which the following formula was proved in [SW97, Hol98]

$$C(\mathcal{T}) = \lim_{n\to\infty} \sup_{\{p_j;\rho_j\}} \frac{1}{n} \left(S\left(\mathcal{T}^{\otimes n}\left(\sum_j p_j \rho_j \right) \right) - \sum_j p_j S(\mathcal{T}^{\otimes n}(\rho_j)) \right).$$

The maximisation over all ensembles is a serious problem when computing the capacity. Since this is often the case in classical information theory, it was conjectured in [HW01] that the minimisation can be restricted to Gaussian states only. It turns out that in many cases, this reduces to the conjecture that Gaussian inputs minimise the entropy of the output. For those channels, where the limit can be dropped because the maximum does not depend on the number of channel uses (so called *additive channels*) the capacity could then easily be computed analytically.

This conjecture remained completely open for over a decade despite considerable effort (see [GHGP15] and links in the introduction), but was recently solved in [GHGP15, GHM15, MGH14] for gauge covariant and contravariant channels, which also led to capacity calculations such as [GGPCH14] and a flurry of further activity. The proof heavily relies on the normal form for gauge covariant and contravariant channels in Theorem 2.2.17 and the fact that the conjecture was known to be true for quantum limited attenuators [GGL$^+$04].

It is still open to extend these results to all channels, for example the channel defined by a beam splitter and a squeezed environment. One approach could be to use similar normal forms and reduce the effort to a specific type of channel.

2.3 Resources and Resource Theories in quantum information

Most of the basic tasks of quantum information theory such as quantum teleportation, entanglement sharing, or channel coding require the preparation and transformation of specific states. In principle, preparing a quantum state is always possible if one has full control over the system, but in reality this is not always the case.

For example, if one wants to create an entangled state between two labs, implementing operations which need access to both labs is impossible in many cases. As another example, creating a state with a high amount of squeezing is difficult if not impossible. The goal must therefore be to create certain states despite these problems and use them as *resources*. The abstract mathematical framework of resources and state interconvertibility is then called a *resource theory*.

Although they were only studied peripherally in the papers underlying this thesis, we will study abstract mathematical frameworks for resource theories because they provide motivation, context, and an interesting aspect of further study for my research about squeezing.

2.3.1 Basic concepts of resource theories

We will describe resource theories similar to the description in [BaG15].

Resource theories are defined by two sets:

1. A set $\mathcal{S}_{\text{free}} \subset \mathcal{S}_1(\mathcal{H})$ of *free states*. Those are states which can easily be created. Free states can for instance include the vacuum.

2. A set $\mathcal{O}_{\text{op}} \subseteq CP(\mathcal{S}_1(\mathcal{H}))$ of *allowed operations*. Usually, this set will contain a subset of completely positive operations ($CP(\mathcal{S}_1(\mathcal{H}))$) restricted by experimental realities.

In addition all states not in the set $\mathcal{R} := \{\mathcal{T}(\rho) | \rho \in \mathcal{S}_{\text{free}}, \mathcal{T} \in \mathcal{O}_{\text{op}}\}$ are called *resource states*. These are the states which cannot be prepared with free states and allowed operations only. Usually the distinction between free states and resource states is given by a *resource* which is quantified in some way.

Goals of a resource theory include the solution to the following problems:

1. Classify the set of resource states.

2. For any given $\sigma \notin \mathcal{R}$ find $\rho \in \mathcal{S}_{\text{free}}, \rho'$ a resource state, and $\mathcal{T} \in \mathcal{O}_{\text{op}}$ such that $\sigma = \mathcal{T}(\rho \otimes \rho')$ and ρ' is as cheap as possible.

3. Find a subset $\mathcal{S}_{\text{resource}} \subset \mathcal{R}$ which has a simple description and still solves the second task approximately. If not prespecified, find functions which describe the cost of the resource.

Clearly, the framework is extremely general and it is thus no surprise that it has been applied to many different areas of quantum information and beyond. More recently, a general study of resource theories was initiated from a mathematical perspective, for instance in [BaG15, Fri15, CFS16].

The study of resource theories is a very active area of research. We will only study entanglement and squeezing in more detail because they are relevant for the thesis. Many other resource theories exist such as "thermal operations" in quantum thermodynamics [HO13, NGP15], complexity [ICK$^+$16], asymmetry [MS13], reference frames [GS08], coherence [CG16, SAP16], or knowledge [dRKR15].

2.3.2 Entanglement

We start with a short discussion of the best known resource in quantum information: Entanglement. For any number n of parties we consider quantum states $\rho \in \mathcal{S}_1(\mathcal{H}_1 \otimes \ldots, \mathcal{H}_n)$ for some Hilbert spaces $\mathcal{H}_j$. A state is called *separable* if it is a convex linear combination of product states $\rho = \rho_1 \otimes \ldots, \rho_n$ with $\rho_j \in \mathcal{S}_1(\mathcal{H}_j)$. Any other state is called *entangled*. The resource theory of entanglement now consists of

1. The set of separable state S_{SEP} as free states.

2. The set LOCC of local operations with classical communication. These are all operations which can be implemented if we assume that every party sits in a lab spatially separated from the others.

The resource states then correspond to the entangled states and functions measuring entanglement are usually called *entanglement monotones* (cf. [Vid00, VW02]). Multipartite entanglement can be a very rich subject (cf. [DVC00, VDDMV02]), so we restrict to $n = 2$, $\mathcal{H}_1 = \mathcal{H}_2 =: \mathcal{H}$ and pure states to give an idea of possible questions and results. The corresponding entangled states are called *bipartite entangled states*. A good overview can be found in [HHHH09].

For pure states, it turns out that the class of entangled states is captured by one state alone, the so called *maximally entangled state* $|\Phi\rangle$ in the following sense:

Theorem 2.3.1 ([Nie99]). *Let $|\psi\rangle, |\phi\rangle \in \mathcal{H}\otimes\mathcal{H}$ be two (normalised) quantum states, then there exists an* LOCC *operation $\mathcal{T}$ such that $\mathcal{T}(|\psi\rangle\langle\psi|) = |\phi\rangle\rangle\phi|$ if and only if for all $k = 1, \ldots, \dim(\mathcal{H})$*

$$\sum_{j=1}^{k} \lambda_k(\mathrm{tr}_2(|\phi\rangle\langle\phi|)) \geq \sum_{j=1}^{k} \lambda_k(\mathrm{tr}_2(|\psi\rangle\langle\psi|)) \tag{2.22}$$

where λ_k denotes the k-th largest eigenvalue.

The condition in Equation (2.22) is also known as *majorisation* (cf. [Bha96]) and the maximally entangled state is then naturally given by the state $|\Phi\rangle$ where the reduced density matrix is proportional to the identity.

Two other important questions concerning the theory are (see [BDSW96]):

1. Given a state $\omega^{\otimes m}$ where $\omega = |\Phi\rangle\langle\Phi|$, how many copies n of a given state $|\psi\rangle$ can we produce using LOCC operations and additional separable states? The ratio n/m is a different entanglement monotone called *entanglement of formation*. For qubits, the entanglement of formation tends to $S(\mathrm{tr}_2(\rho))$.

2. Given any pure state $\rho^{\otimes m}$, how many copies n of the maximally entangled states of some dimension d can we produce using only LOCC operations and additional separable states? The ratio n/m is called *distillable entanglement*. For qubits, the distillable entanglement tends to $1/S(\mathrm{tr}_2(\rho))$.

For mixed states, not all states can be distilled to maximally entangled states. For systems of dimension $3 \otimes 3$ or larger, the complete classification is still open which is known as the NPT-bound entanglement problem [Cla06].

The entanglement of formation for mixed states is differently defined than for pure states, since a mixed state can be created as a mixture of pure quantum states in many different ways. The definition therefore amounts to a convex roof (cf. [BDSW96])

$$E_F(\rho) = \inf\left\{ \sum_i r_i S(\mathrm{tr}_2(\rho_i)) \middle| \sum_i r_i \rho_i = \rho \right\}. \tag{2.23}$$

Computing the entanglement of formation has been a major challenge. The seemingly simpler problem whether for any states $\rho = \rho^{(1)} \otimes \rho^{(2)}$ we have $E_F(\rho) = E_F(\rho^{(1)}) + E_F(\rho^{(2)})$ was an open problem for a long time. It was settled in the negative in [Han09] although a concrete example is still missing.

For the class of Gaussian states, one is usually interested in the restriction of LOCC to the class GLOCC of Gaussian local operations and classical communications. Remarkably, distillation of entanglement is entirely impossible using GLOCC operations as shown by [GIC02, ESP02]. In contrast to this the determination of the entanglement of formation is still open. Similarly to the conjectures about channel capacities described in Section 2.2.3, it might a good guess that the minimum is attained on Gaussian states. For the corresponding measure, the *Gaussian entanglement of formation*, we have:

Theorem 2.3.2 ([WGK$^+$04], Proposition 1)**.** *Given a Gaussian state ρ with covariance matrix γ, the Gaussian entanglement of formation is given by*

$$E_G(\rho) = \inf\{S((\gamma_p)_{\mathrm{red}})|\gamma \geq \gamma_p = 2S^T S, S \in Sp(2n)\}. \tag{2.24}$$

γ_{red} denotes the covariance matrix of the reduced state of ρ and S is the entropy.

The measure is therefore a minimisation over all pure Gaussian states with covariance matrices γ_p, similar to the squeezing measures we introduce in Paper I. It is only known to be equal to the entanglement of formation for a certain class of Gaussian states [GWK$^+$03].

2.3.3 Squeezing

For a Gaussian state, entanglement is linked to squeezing [WEP03, LGW13]. For instance, we have the following theorem:

Theorem 2.3.3 ([WEP03], Proposition 2)**.** *Let $\gamma \mapsto K^T\gamma K$ be a passive transformation (i.e. $K \in Sp(2n) \cap O(2n)$) acting on a Gaussian state of $n \geq 2$ modes with covariance matrix γ. The maximum attainable amount of entanglement obtained for an arbitrary two-mode subsystem of $K^T\gamma K$ is then given by*

$$E_N = \max\{0, -\log_2(4\lambda_1\lambda_2)\} \tag{2.25}$$

where λ_1, λ_2 are the two smallest eigenvalues of γ. Here, E_N denotes the logarithmic negativity, which is an entanglement measure for mixed Gaussian states (cf. [VW02]).

Clearly, the state has zero entanglement if it is not squeezed.

The fact that squeezing can be considered a resource was first noted in [Bra05], where the author proved that a state could not be squeezed without squeezers. Based on this approach, the following resource theory seems natural:

1. The free states are given by all Gaussian states, which are unsqueezed, i.e. the covariance matrix fulfils $\gamma \geq \mathbb{1}$.

2. The allowed operations include drawing unsqueezed ancillary states, making measurements, adding noise, performing beam splitters and phase shifters as well as Weyl displacements.

Resource states are all squeezed states. We now only need a measure for squeezing. A simple measure, the smallest eigenvalue of the covariance matrix, was proposed in [KGLC03, Lee88]. Analogously to the Gaussian entanglement of formation one can now also consider "squeezing of formation", which leads us to postulate a measure similar to Theorem 2.3.2. Indeed, this is one possible starting point for considering the measures in Paper I.

Studying squeezing as a resource is interesting not only because it is linked to entanglement, but mostly because it is considered a hard task. The problem is twofold: Squeezed states decohere very quickly and single-mode squeezing becomes infeasible for large amount of squeezing. Current upper limits are reported in [AGML15].

Bibliography

[Aar05] Scott Aaronson. Quantum computing and hidden variables. *Phys. Rev. A*, 71, 03 2005. 6

[AB15] Ole Andersson and Ingemar Bengtsson. Clifford-tori and unbiased vectors, 2015. arXiv:1506.09062 [quant-ph]. 11

[ADMS95a] Arvind, B. Dutta, N. Mukunda, and R. Simon. The real symplectic groups in quantum mechanics and optics. *Pramana*, 45(6):471–497, 1995. 12, 13

[ADMS95b] Arvind, B. Dutta, N. Mukunda, and R. Simon. Two-mode quantum systems: Invariant classification of squeezing transformations and squeezed states. *Physical review. A*, 52(2):1609—1620, 08 1995. 19

[AGML15] Ulrik L. Andersen, Tobias Gehring, Christoph Marquardt, and Gerd Leuchs. 30 years of squeezed light generation, 2015. arXiv:1511.03250v2 [quant-ph]. 30

[ARL14] Gerardo Adesso, Sammy Ragy, and Antony R. Lee. Continuous variable quantum information: Gaussian states and beyond. *Open Systems &; Information Dynamics*, 21(01n02):1440001, 2014. 14, 16

[Arn89] Vladimir I. Arnol'd. *Mathematical Methods of Classical Mechanics*. Springer, 1989. 17

[BaG15] Fernando G. S. L. Brandão and Gilad Gour. Reversible framework for quantum resource theories. *Phys. Rev. Lett.*, 115:070503, 08 2015. 26, 27

[BB84] Charles H. Bennett and G. Brassard. Quantum cryptography: Public key distribution and coin tossing. In *International Conference on Computer System and Signal Processing, IEEE, 1984*, pages 175–179, 1984. 14

[BDSW96] Charles H. Bennett, David P. DiVincenzo, John A. Smolin, and William K. Wootters. Mixed-state entanglement and quantum error correction. *Phys. Rev. A*, 54:3824–3851, 10 1996. 28

[BEH08] Jiří Blank, Pavel Exner, and Miloslav Havlíček. *Hilbert Space Operators in Quantum Physics*. Springer, 2 edition, 2008. 16, 17, 23

[BEP04] Paul Biran, Michael Entov, and Leonid Polterovich. Calabi quasimorphisms for the symplectic ball. *Communications in Contemporary Mathematics*, 06(05):793–802, 2004. 11

[Bha96] Rajendra Bhatia. *Matrix Analyis*. Springer, 1996. 4, 28

[BJ15] Rajendra Bhatia and Tanvi Jain. On symplectic eigenvalues of positive definite matrices. *Journal of Mathematical Physics*, 56(11), 2015. 4

[Bra05] Samuel L. Braunstein. Squeezing as an irreducible resource. *Phys. Rev. A*, 71:055801, 05 2005. 29

[BvL05] Samuel L. Braunstein and Peter van Loock. Quantum information with continuous variables. *Rev. Mod. Phys.*, 77:513–577, 06 2005. 14

[CEGH08] Filippo Caruso, Jens Eisert, Vittorio Giovannetti, and Alexander S. Holevo. Multi-mode bosonic gaussian channels. *New Journal of Physics*, 10(8):083030, 2008. 23

[CEGH11] Filippo Caruso, Jens Eisert, Vittorio Giovannetti, and Alexander S. Holevo. Optimal unitary dilation for bosonic gaussian channels. *Phys. Rev. A*, 84:022306, 08 2011. 23

[CFS16] Bob Coecke, Tobias Fritz, and Robert W. Spekkens. A mathematical theory of resources. *Information and Computation*, 2016. 27

[CFZ14] Thomas Curtright, David B. Fairlie, and Cosmas K. Zachos. *A concise treatise on quantum mechanics in phase space*. World Scientific, 2014. 18

[CG16] Eric Chitambar and Gilad Gour. Critical examination of incoherent operations and a physically consistent resource theory of quantum coherence. *Phys. Rev. Lett.*, 117:030401, 07 2016. 27

[Cho04] Cheol-Hyun Cho. Holomorphic discs, spin structures, and Floer cohomology of the Clifford torus. *Int. Math. Res. Not.*, 2004(35):1803–1843, 2004. 11

[Cla06] Lieven Clarisse. The distillability problem revisited. *Quantum Info. Comput.*, 6(6):539–560, 2006. 28

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. 13

[Dar82] Gaston Darboux. Sur le problème de pfaff. *Bull. Sci. Math.*, 6:14–36, 49–68, 1882. 10

[Dev05] Igor Devetak. The private classical capacity and quantum capacity of a quantum channel. *IEEE Transactions on Information Theory*, 51(1):44–55, 01 2005. 14

[dG06] Maurice A. de Gosson. *Symplectic Geometry and Quantum Mechanics*. Operator Theory: Advances and Applications / Advances in Partial Differential Equations. Birkhäuser Basel, 2006. 10

[dGL09] Maurice de Gosson and Franz Luef. Symplectic capacities and the geometry of uncertainty: The irruption of symplectic topology in classical and quantum mechanics. *Physics Reports*, 484(5):131 – 179, 2009. 19

[dRKR15] Lidia del Rio, Lea Kraemer, and Renato Renner. Resource theories of knowledge, 2015. arXiv:1511.08818 [quant-ph]. 27

[DVC00] W. Dür, G. Vidal, and J. I. Cirac. Three qubits can be entangled in two inequivalent ways. *Phys. Rev. A*, 62:062314, 11 2000. 27

[DVDB14a] Alexis De Vos and Stijn De Baerdemacker. Scaling a unitary matrix. *Open Systems & Information Dynamics*, 21(4), 2014. 5, 11

[DVDB14b] Alexis De Vos and Stijn De Baerdemacker. The synthesis of a quantum circuit. In *Proceedings of the 11th International Workshop of Boolean Problems*, pages 129–136. Freiberg University of Mining and Technology, 2014. 11, 23

[DVDB16] Alexis De Vos and Stijn De Baerdemacker. The birkhoff theorem for unitary matrices of prime dimension. *Linear Algebra and its Applications*, 493:455 – 468, 2016. 11

BIBLIOGRAPHY

[DVV77] B. Demoen, P. Vanheuverzwijn, and A. Verbeure. Completely positive maps on the ccr-algebra. *Letters in Mathematical Physics*, 2(2):161–166, 1977. 23

[Eke91] Artur K. Ekert. Quantum cryptography based on bell's theorem. *Phys. Rev. Lett.*, 67:661–663, 08 1991. 14

[ESP02] J. Eisert, S. Scheel, and M. B. Plenio. Distilling Gaussian states with Gaussian operations is impossible. *Phys. Rev. Lett.*, 89:137903, 09 2002. 29

[EW07] Jens Eisert and Michael M. Wolf. Gaussian quantum channels. In N. J. Cerf, G. Leuchs, and E. S. Polzik, editors, *Quantum Information with continuous variables of atoms and light*. World Scientific, USA, 2007. 22

[Fiu02] Jaromír Fiurášek. Gaussian transformations and distillation of entangled gaussian states. *Phys. Rev. Lett.*, 89:137904, 09 2002. 22

[FO99] Kenji Fukaya and Kaoru Ono. Arnold conjecture and Gromov–Witten invariant. *Topology*, 38(5):933 – 1048, 1999. 11

[FR15] Hartmut Führ and Ziemowit Rzeszotnik. On biunimodular vectors for unitary matrices. *Linear Algebra and its Applications*, 484(0):86 – 129, 2015. 11

[Fri15] T. Fritz. Resource convertibility and ordered commutative monoids. *Mathematical Structures in Computer Science*, pages 1–89, 2015. 27

[GGL+04] V. Giovannetti, S. Guha, S. Lloyd, L. Maccone, J. H. Shapiro, and H. P. Yuen. Classical capacity of the lossy bosonic channel: The exact solution. *Phys. Rev. Lett.*, 92:027902, 01 2004. 25

[GGPCH14] Vittorio Giovannetti, Raul Garcia-Patron, N. J. Cerf, and Alexander S. Holevo. Ultimate communication capacity of quantum optical channels by solving the Gaussian minimum-entropy conjecture. *Nature Photonics*, 8:796–800, 2014. 2, 5, 25

[GHGP15] Vittorio Giovannetti, Alexander S. Holevo, and Raul García-Patrón. A solution of gaussian optimizer conjecture for quantum channels. *Communications in Mathematical Physics*, 334(3):1553–1571, 2015. 24, 25

[GHM15] Vittorio Giovannetti, Alexander S. Holevo, and Andrea Mari. Majorization and additivity for multimode bosonic gaussian channels. *Theoretical and Mathematical Physics*, 182(2):284–293, 2015. 25

[GIC02] Géza Giedke and J. Ignacio Cirac. Characterization of Gaussian operations and distillation of Gaussian states. *Phys. Rev. A*, 66:032316, 09 2002. 22, 29

[GPNBL+12] Raúl García-Patrón, Carlos Navarrete-Benlloch, Seth Lloyd, Jeffrey H. Shapiro, and Nicolas J. Cerf. Majorization theory approach to the gaussian channel minimum entropy conjecture. *Phys. Rev. Lett.*, 108:110505, 03 2012. 25

[Gro85] Mikhail L. Gromov. Pseudo holomorphic curves in symplectic manifolds. *Inventiones Mathematicae*, 82:307–347, 1985. 10

[GS08] Gilad Gour and Robert W. Spekkens. The resource theory of quantum reference frames: manipulations and monotones. *New Journal of Physics*, 10(3), 2008. 27

[GWK+03] Géza Giedke, Michael M. Wolf, Ole Krüger, Reinhard. F. Werner, and J. Ignacio Cirac. Entanglement of formation for symmetric gaussian states. *Phys. Rev. Lett.*, 91:107901, 9 2003. 29

[Han09] Mathew B. Hanstings. Superadditivity of communication capacity using entangled inputs. *Nature Physics*, 5:255–257, 2009. 28

[Hel01] Sigurdur Helgason. *Differential geometry, Lie groups, and symmetric spaces*. American Mathematical Society, 2001. 12, 13

[hf11] Joel Fine (http://mathoverflow.net/users/380/joel fine). What is a lagrangian submanifold intuitively? MathOverflow, 2011. URL:http://mathoverflow.net/q/60519 (version: 2011-04-04). 11

[HHHH09] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki. Quantum entanglement. *Rev. Mod. Phys.*, 81:865–942, 06 2009. 27

[HHW10] Teiko Heinosaari, Alexander S. Holevo, and Michael M. Wolf. The semigroup structure of gaussian channels. *Quantum Info. Comput.*, 10(7):619–635, 7 2010. 24

[HO13] Michal Horodecki and Jonathan Oppenheim. Fundamental limitations for quantum and nanoscale thermodynamics. *Nature Communications*, 4, 2013. 27

[Hol98] Alexander S. Holevo. The capacity of the quantum channel with general signal states. *IEEE Transactions on Information Theory*, 44(1):269–273, 01 1998. 14, 25

[Hol11] Alexander S. Holevo. *Probabilistic and Statistical Aspects of Quantum Theory*. Edizioni della Normale, 2011. 15, 18

[Hol15] Alexander S. Holevo. Gaussian optimizers and the additivity problem in quantum information theory. *Russian Mathematical Surveys*, 70(2):331, 2015. 25

[HSH99] A. S. Holevo, M. Sohma, and O. Hirota. Capacity of quantum gaussian channels. *Phys. Rev. A*, 59:1820–1828, 03 1999. 21, 24

[HW01] Alexander S. Holevo and Reinhard F. Werner. Evaluating capacities of bosonic Gaussian channels. *Phys. Rev. A*, 63:032312, 02 2001. 25

[ICK+16] Raban Iten, Roger Colbeck, Ivan Kukuljan, Jonathan Home, and Matthias Christandl. Quantum circuits for isometries. *Phys. Rev. A*, 93:032318, 03 2016. 27

[KGLC03] Barbara Kraus, Géza Giedke, Maciej Lewenstein, and Juan Ignacio Cirac. Entanglement properties of Gaussian states. *Fortschritte der Physik*, 51(4-5):305–312, 2003. 4, 30

[KJR14] Kamil Korzekwa, David Jennings, and Terry Rudolph. Operational constraints on state-dependent formulations of quantum error-disturbance trade-off relations. *Phys. Rev. A*, 89:052108, 05 2014. 5, 11

[Lee88] C.T. Lee. Wehrl's entropy as a measure of squeezing. *Optics Communications*, 66(1):52 – 54, 1988. 30

[LGW13] Daniel Lercher, Géza Giedke, and Michael M. Wolf. Standard super-activation for gaussian channels requires squeezing. *New Journal of Physics*, 15(12):123003, 2013. 24, 29

[Lou00] Rodney Loudon. *The Quantum Theory of Light*. Oxford University Press, 2000. 16

[LT98] Gang Liu and Gang Tian. Floer homology and Arnold conjecture. *J. Differential Geom.*, 49(1):1–74, 1998. 11

[MGH14] Andrea Mari, Vittorio Giovannetti, and Alexander S. Holevo. Quantum state majorization at the output of bosonic Gaussian channels. *Nature Communications*, 5, 2014. 25

BIBLIOGRAPHY

[MS98] Dusa McDuff and Dietmar Salamon. *Introduction to Symplectic Topology*. Oxford Science Publications, 1998. 10, 11

[MS13] Iman Marvian and Robert W. Spekkens. Extending noether's theorem by quantifying the asymmetry of quantum states. *Nature Communications*, 5, 2013. 27

[NC00] Michael Nielsen and Isaac Chuang. *Quantum Computation and Quantum Information*. Camebridge University Press, 2000. 14

[NGP15] Miguel Navascués and Luis Pedro García-Pintos. Nonthermal quantum channels as a thermodynamical resource. *Phys. Rev. Lett.*, 115:010405, 07 2015. 27

[Nie99] Michael A. Nielsen. Conditions for a class of entanglement transformations. *Phys. Rev. Lett.*, 83:436–439, 07 1999. 28

[NO86] Francis J. Narcowich and R. F. O'Connell. Necessary and sufficient conditions for a phase-space function to be a wigner distribution. *Phys. Rev. A*, 34:1–6, 07 1986. 18

[Oli12] S. Olivares. Quantum optics in the phase space. *The European Physical Journal Special Topics*, 203(1):3–24, 2012. 18, 20

[PRC+15] Zbigniew Puchała, Łukasz Rudnicki, Krzysztof Chabuda, Mikołaj Paraniak, and Karol Życzkowski. Certainty relations, mutual entanglement, and nondisplaceable manifolds. *Phys. Rev. A*, 92:032109, 09 2015. 11

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 23

[Sal99] Dietmar Salamon. Lectures on Floer homology. *Symplectic geometry and topology (Park City, UT, 1997)*, 7:143–229, 1999. 11

[SAP16] Alexander Streltsov, Gerardo Adesso, and Martin B. Plenio. Quantum coherence as a resource, 2016. arXiv:1609.02439 [quant-ph]. 27

[Sch95] Benjamin Schumacher. Quantum coding. *Phys. Rev. A*, 51:2738–2747, 04 1995. 14

[SCS99] R. Simon, S. Chaturvedi, and V. Srinivasan. Congruences and canonical forms for a positive matrix: Application to the schweinler–wigner extremum principle. *Journal of Mathematical Physics*, 40(7):3632–3642, 1999. 20

[Sha48] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 10 1948. 13

[Son98] Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer, New York, 2 edition, 1998. 4

[SW97] Benjamin Schumacher and Michael D. Westmoreland. Sending classical information via noisy quantum channels. *Phys. Rev. A*, 56:131–138, 07 1997. 14, 25

[SY08] Graeme Smith and Jon Yard. Quantum communication with zero-capacity channels. *Science*, 321(1812), 2008. 2

[SZ97] Marlan O. Scully and M. Suhail Zubairy. *Quantum Optics*. Cambridge University Press, 1997. 16

[VDDMV02] F. Verstraete, J. Dehaene, B. De Moor, and H. Verschelde. Four qubits can be entangled in nine different ways. *Phys. Rev. A*, 65:052112, 04 2002. 27

[Vid00] Guifré Vidal. Entanglement monotones. *Journal of Modern Optics*, 47(2-3):355–376, 2000. 27

[VW02] Guifre Vidal and Reinhard F. Werner. Computable measure of entanglement. *Phys. Rev. A*, 65:032314, 02 2002. 27, 29

[WEP03] Michael M. Wolf, Jens Eisert, and Martin B. Plenio. Entangling power of passive optical elements. *Phys. Rev. Lett.*, 90:047904, 01 2003. 29

[WGK+04] M. M. Wolf, G. Giedke, O. Krüger, R. F. Werner, and J. I. Cirac. Gaussian entanglement of formation. *Phys. Rev. A*, 69:052320, 05 2004. 29

[Wig32] Eugene Wigner. On the quantum correction for thermodynamic equilibrium. *Phys. Rev.*, 40:749–759, 06 1932. 18

[Wil36] John Williamson. On the algebraic problem concerning the normal forms of linear dynamical systems. *American Journal of Mathematics*, 58(1):pp. 141–163, 1936. 19

[Wol08] Michael M. Wolf. Not-so-normal mode decomposition. *Phys. Rev. Lett.*, 100:070505, 02 2008. 21

[WPGP+12] Christian Weedbrook, Stefano Pirandola, Raul Garcia-Patron, Nicolas J Cerf, Timothy C Ralph, Jeffrey H Shapiro, and Seth Lloyd. Gaussian quantum information. *Reviews of Modern Physics*, 84(2):621, 2012. 14

An operational measure for squeezing

M. Idel, D. Lercher and M. M. Wolf February 12, 2017

A quantum state ρ is squeezed if its covariance matrix γ_ρ has an eigenvalue $\lambda < 1$ (leaving out factors of $1/2$). Squeezing is hard in practice and therefore a useful resource [1]. In order to study the resource, we need to have operational measures (as in the case of entanglement). So far, the only measure of squeezing in the literature is the minimum eigenvalue of the covariance matrix [2]. Clearly, this is not an operational measure for most tasks such as preparation of a squeezed state, because the two quantum states with covariance matrix $\gamma_1 = \mathrm{diag}(s, 1, s^{-1}, 1)$ and $\gamma_2 = \mathrm{diag}(s, s, s^{-1}, s^{-1})$ have the same minimum eigenvalue, but the squeezing cost for the second should be twice the amount of the first, as it is squeezed in both modes. We define measures covering preparation costs for squeezed states.

1 Operational squeezing

Before considering the case of a squeezed state, we can ask: For any Gaussian unitary operation given by the symplectic transformation S, what is the least amount of single-mode squeezing necessary to implement it? If $s_i^{\downarrow}$ denotes the decreasingly ordered singular values, we prove:

Theorem 1.1. *Let S be a symplectic matrix. Then*

$$F(S) = \sum_{i=1}^{n} \log(s_i^{\downarrow})(S) \tag{1}$$

is the minimal amount of single-mode squeezing required among

1. *all products $S = S_1 \cdots S_n$, where $S_i \in SO(2n) \cap Sp(2n)$ or S_i is a single-mode squeezer with squeezing parameter s and associated cost $\log(s)$,*
2. *all rectifiable, almost everywhere differentiable paths $\gamma : [0, 1] \to Sp(2n)$ with $\gamma(0) = \mathbb{1}$ and $\gamma(1) = S$, where the costs are quantified according to the coefficient of the active generators.*

In both cases, the minimum is given by the Euler decomposition.

The costs for paths on $Sp(2n)$ are chosen such that they are consistent with the choice for single-mode squeezers. A detailed construction is given in the paper. The proof for paths on Lie groups relies on results from the theory of ordinary differential equations. In particular it follows from perturbation theory of the propagator of the ODE given by the path.

2 State preparation with minimal squeezing

In the main part of the paper, we answer the following question: If we can freely draw any state with covariance matrix $\gamma \geq \mathbb{1}$, freely add ancillas, add noise, create convex combinations and perform passive operations, measurements and Weyl displacements, what is the minimal amount of single-mode squeezing needed to create a given state ρ?

Theorem 2.1. *Let $\gamma \in \mathbb{R}^{2n\times 2n}$ be the covariance matrix of a quantum state. If we measure the amount of any single mode squeezer by $\log(s)$, where s is the squeezing parameter, then the minimal amount of squeezing necessary in order to create the state ρ using the free operations outlined above is given by*

$$G(\gamma) = \inf\left\{\sum_{i=1}^{n} \log(s_i^{\downarrow})(S) \middle| \gamma \geq S^T S, S \in Sp(2n)\right\}. \tag{2}$$

The measure G is convex, lower semi-continuous and subadditive.

The proof of convexity uses the Cayley transform for matrices. On the basis of the convexity of G, we provide a program to calculate G numerically. Moreover, we provide upper and lower bounds. The proof that the measure is operational relies on convexity as well as Cauchy's interlacing theorem.

Finally, we give a short discussion on squeezing as a resource. We prove that our measure remains the same if we consider single-mode squeezed states as "resource states" similar to the entanglement of formation in entanglement theory. Consequently, our measure could also be dubbed "squeezing of formation".

3 Legal statement

The project was started by Daniel Lercher and Michael Wolf. I continued the project when the first part of Theorem 1.1 as well as the definition, convexity and some properties of G had already been sketched. Based on the idea to use propagators (by Daniel Lercher), I worked out the proof of the second part of Theorem 1.1 and filled in all gaps in the sketches of the first part. All other parts (especially Sections 5-7 in the paper) are my work with guiding input by Michael Wolf.

References

[1] Samuel L. Braunstein and Peter van Loock. Quantum information with continuous variables. *Rev. Mod. Phys.*, 77:513–577, 06 2005.

[2] Barbara Kraus, Klemens Hammerer, Géza Giedke, and J. Ignacio Cirac. Entanglement generation and Hamiltonian simulation in continuous-variable systems. *Phys. Rev. A*, 67:042314, 04 2003.

IOP Publishing Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **49** (2016) 445304 (45pp) doi:10.1088/1751-8113/49/44/445304

An operational measure for squeezing

Martin Idel, Daniel Lercher and Michael M Wolf

Zentrum Mathematik, Technische Universität München, Germany

E-mail: martin.idel@tum.de

Received 5 July 2016, revised 26 August 2016
Accepted for publication 14 September 2016
Published 13 October 2016

Abstract

We propose and analyse a mathematical measure for the amount of squeezing contained in a continuous variable quantum state. We show that the proposed measure operationally quantifies the minimal amount of squeezing needed to prepare a given quantum state and that it can be regarded as a squeezing analogue of the 'entanglement of formation'. We prove that the measure is convex and subadditive and we provide analytic bounds as well as a numerical convex optimisation algorithm for its computation. By example, we then show that the amount of squeezing needed for the preparation of certain multi-mode quantum states can be significantly lower than naive state preparation suggests.

Keywords: squeezing, continuous variable quantum information, operational measure, Euler decomposition, bosonic systems

(Some figures may appear in colour only in the online journal)

1. Introduction

The interplay between quantum optics and the field of quantum information processing, in particular via the subfield of continuous variable quantum information, has been developing for several decades and is interesting also due to its experimental success (see [KL10] for a thorough introduction).

Coherent bosonic states and the broader class of Gaussian bosonic states, quantum states whose Wigner function is characterised by its first and second moments, are of particular interest in the theory of continuous variable quantum information. Their interest is also due to the fact that modes of light in optical experiments behave like Gaussian coherent states.

For any bosonic state, its matrix of second moments, the so called covariance matrix, must fulfil Heisenberg's uncertainty principle in all modes. If the state possesses a mode, where despite this inequality $\Delta x \Delta p \geqslant \hbar/2$ either Δx or Δp is strictly smaller than $\sqrt{\hbar/2}$, it is called *squeezed*. The production of squeezed states is experimentally possible, but it

requires the use of nonlinear optical elements [Bra05], which are more difficult to produce and handle than the usual linear optics (i.e. beam splitters and phase shifters). Nevertheless, squeezed states play a crucial role in many experiments in quantum information processing and beyond. Therefore, it is natural both theoretically and practically to investigate the amount of squeezing which is necessary to create an arbitrary quantum state.

As a qualitative answer, squeezing is known to be an irreducible resource with respect to linear quantum optics [Bra05]. In the Gaussian case, it is also known to be closely related to entanglement of states [WEP03] and the non-additivity of quantum channel capacities [LGW13]. In addition, quantitative measures of squeezing have been provided on multiple occasions [Kra+03, Lee88], yet none of these measures are operational for more than a single mode in the sense that they do not measure the minimal amount of squeezing necessary to prepare a given state.

The goal of this paper is therefore twofold: first, we define and study operational squeezing measures, especially measures quantifying the amount of squeezing needed to prepare a given state. Second, we reinvestigate in how far squeezing is a resource in a mathematically rigorous manner and study the resulting resource theory by defining preparation measures.

In order to give a brief overview of the results, we assume the reader is familiar with standard notation of the field, which is also gathered in section 2. In particular, let γ denote covariance matrices. A squeezed state is a state where at least one of the eigenvalues of γ is smaller than one.

To obtain operational squeezing measures, we first study operational squeezing in section 3: suppose we want to implement an operation on our quantum state corresponding to some unitary U. Any such unitary can be implemented as the time-evolution of Hamiltonians. Recall that any quantum-optical Hamiltonian can be split into 'passive' and 'active' parts, where the passive parts are implementable by linear optics and the active parts require nonlinear media. We assume that the active transformations available are single-mode squeezers with Hamiltonian

$$H_{\text{squeeze},j} = \mathrm{i}\frac{\hbar}{2}(a_j^2 - a_j^{\dagger\,2}),$$

where the j denotes squeezing in the jth mode. We therefore consider any Hamiltonian of the form

$$H = H_{\text{passive}}(t) + \sum_k c_k(t) H_{\text{squeeze},j}, \tag{1}$$

where c_k are complex coefficients, which can be seen as the interaction strength of the medium and H_{passive} is an arbitrary passive Hamiltonian. Then, a natural measure of the squeezing costs to implement this Hamiltonian would be given by

$$f_{\text{squeeze}}(H) = \int \sum_k |c_k(t)|\,\mathrm{d}t.$$

Our squeezing measure for the operation U is then defined as the mimimum of $f_{\text{squeeze}}(H)$ for all Hamiltonians implementing the operation U of the form (1). With this definition, we have an operational measure answering the question: given an operation U, what is the most efficient way (in terms of squeezing) to implement it using passive operations and single-mode squeezers?

Instead of working with the generators, which are unbounded operators and therefore introduce a lot of analytic problems, we will work on the level of Wigner functions and therefore with the symplectic group. The unitary U then corresponds to a symplectic matrix S and we prove that the most efficient way to implement it is by using the Euler decomposition, also known as Bloch–Messiah decomposition. We show this result first in the case where the

functions c_i are step functions and later on in the more general case of measurable c (section 3.2). In particular, the result implies that the minimum amount of squeezing to implement the symplectic matrix $S \in \mathbb{R}^{2n\times 2n}$ is given by

$$F(S) := \sum_{i=1}^{n} \log s_i^{\downarrow}(S), \tag{2}$$

where $s_i^{\downarrow}$ denotes the ith singular value of S ordered decreasingly.

With this in mind, we define a squeezing measure for preparation procedures where one starts out with a covariance matrix of an unsqueezed state and then performs symplectic (and possibly other) operations to obtain the state. More precisely, we define

$$G(\gamma) := \inf\left\{ \sum_{j=1}^{n} \log s_j^{\downarrow}(S) \,\middle|\, \gamma \geqslant S^T S, \;\; S \in Sp(2n) \right\}. \tag{3}$$

One of the main results of this paper, which will be proven in section 5, is that this measure is indeed operational in that it quantifies the minimal amount of single-mode squeezing necessary to prepare a state with covariance matrix γ, using linear optics with single-mode squeezers, ancillas, measurements, convex combinations and addition of classical noise.

We also define a second squeezing measure, which is a squeezing-analogue of the entanglement of formation, the 'squeezing of formation', i.e. the amount of single-mode squeezed resource states needed to prepare a given state using only passive operations and adding of noise. This is done in section 5.3, where we also prove that this measure is equal to G.

In addition, we prove several structural facts about G in section 4. In particular, G is convex, lower semicontinuous everywhere, continuous on the interior and subadditive. Moreover, we show

$$\frac{1}{2} \sum_{\lambda_j<1}^{n} \log(\lambda_j(\gamma)) \leqslant G(\gamma)$$

with the eigenvalues λ_j of γ. Equality in this lower bound is usually not achievable, albeit numerical tests have shown that the bound is often very good.

The measure would lose a lot of its appeal, if it could not be computed. Although we cannot give an efficient analytical formula for more than one mode, we provide a numerical algorithm to obtain G for any state. To demonstrate that this works in principle, we calculate G approximately for a state studied in [MK08] (section 6). The calculations also demonstrate that the preparation procedure obtained from minimising G can greatly lower the squeezing costs when compared to naive preparation procedures. Finally, we critically discuss the flexibility and applicability of our measures in section 7. We believe that while we managed to give reasonable measures and interesting tools to study the resource theory of squeezing from a theoretical perspective, G might not reflect the experimental reality in all parts. In particular, it becomes extraordinarily difficult to achieve high squeezing in a single mode [And+15], which is not reflected by taking the logarithm of the squeezing parameter. We show that this shortcoming can be easily corrected for a broad class of cost functions. In addition, the form of the active part of the Hamiltonian (1) might not reflect the form of the Hamiltonian in the lab. This cannot be corrected as easily but in any case, our measure will give a lower bound.

2. Preliminaries

In this section, we collect basic notions from continuous variable quantum information and symplectic linear algebra that we need later on. For a broader overview, we refer to [ARL14, BL05].

2.1. Phase space in quantum physics

Consider a bosonic system with n-modes, each of which is characterised by a pair of canonical variables $\{Q_k, P_k\}$. Setting $R = (Q_1, P_1, \ldots, Q_n, P_n)^T$ the canonical commutation relations (CCRs) take on the form $[R_k, R_l] = \mathrm{i}\sigma_{kl}$ with the standard symplectic form

$$\sigma = \bigoplus_{i=1}^{n} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Since it will sometimes be convenient, we also introduce another basis of the canonical variables: let $\tilde{R} = (Q_1, Q_2, \ldots, Q_n, P_1, P_2, \ldots, P_n)^T$, then the symplectic CCRs take on the form $[\tilde{R}_k, \tilde{R}_l] = \mathrm{i}J_{kl}$ with the symplectic form

$$J = \begin{pmatrix} 0 & \mathbb{1}_n \\ -\mathbb{1}_n & 0 \end{pmatrix}.$$

Clearly, J and σ differ only by a permutation, since R and $\tilde{R}$ differ only by a permutation. From functional analysis, it is well-known that the operators Q_k and P_k cannot be represented by bounded operators on a Hilbert space. In order to avoid complications associated to unbounded operators, it is usually easier to work with a representation of the CCR-relations on some Hilbert space $\mathcal{H}$, instead. The standard representation is known as the *Schrödinger representation* and defines the *Weyl system*, a family of unitaries W_ξ with $\xi \in \mathbb{R}^{2n}$ and

$$W_\xi \coloneqq \exp(\mathrm{i}\xi\sigma R), \qquad \xi \in \mathbb{R}^{2n}$$

fulfiling the Weyl relations $W_\xi W_\eta = \exp^{-\mathrm{i}/2\xi\sigma\eta} W_{\xi+\eta}$ for all ξ, η. Such a system is unique up to isomorphism under further assumptions of continuity and irreducibility as obtained by the Stone–von Neumann theorem. Given W_ξ it is important to note that

$$W_\xi R_k W_\xi^* = R_k + \xi_k \mathbb{1} \qquad \forall\, \xi \in \mathbb{R}^{2n}. \tag{4}$$

In this paper, we will not use many properties of the Weyl system, since instead, we can work with the much simpler *moments* of the state: given a quantum state $\rho \in \mathcal{S}_1(L^2(\mathbb{R}^{2n}))$ (trace-class operators on L^2), its first and second centred moments are given by

$$d_k \coloneqq \mathrm{tr}(\rho R_k), \tag{5}$$

$$\gamma_{kl} \coloneqq \mathrm{tr}(\rho\, \{R_k - d_k\mathbb{1}, R_l - d_l\mathbb{1}\}_+) \tag{6}$$

with $\{\cdot,\cdot\}_+$ the regular anticommutator. We will write Γ instead of γ for the covariance matrix, if we work with $\tilde{R}$ instead of R. Again, a simple permutation relates the two.

An important question one can ask is when a matrix γ can occur as a covariance matrix of a quantum state. The answer is given by Heisenberg's principle, which here takes the form of a matrix inequality:

Proposition 2.1. *Let $\gamma \in \mathbb{R}^{2n\times 2n}$, then there exists a quantum state ρ with covariance matrix γ if and only if*

$$\gamma \geqslant \mathrm{i}\sigma,$$

where $\geqslant$ *denotes the standard partial order on matrices (i.e.* $\gamma \geqslant \mathrm{i}\sigma$ *if* $\gamma - \mathrm{i}\sigma$ *is positive semidefinite). Note that we leave out the usual factor of* $\hbar/2$ *to simplify notation.*

Another question one might ask is when a covariance matrix belongs to a pure quantum state. This question cannot be answered without more information about the higher order terms If we however require the state to be uniquely determined by its first and second moments, i.e. if we consider the so called *Gaussian states*, we have an answer (see [ASI04]):

Proposition 2.2. *Let* ρ *be an n-mode Gaussian state (i.e. completely determined by its first and second moments), then* ρ *is pure if and only if* $\det(\gamma_\rho) = 1$.

2.2. The linear symplectic group and squeezing

A very important set of operations on a quantum system are those, that leave the CCRs invariant, i.e. linear transformations S such that $[SR_k, SR_l] = \mathrm{i}\sigma_{kl}$. Such transformations are called *symplectic transformations*.

Definition 2.3. Given a symplectic form σ on $\mathbb{R}^{2n\times 2n}$, the set of matrices $S \subset \mathbb{R}^{2n\times 2n}$ such that $S^T\sigma S = \sigma$ is called the *linear symplectic group* and is denoted by $Sp(2n, \mathbb{R}, \sigma)$.

We will usually drop both σ and $\mathbb{R}$ in the description of the symplectic group since this will be clear from the context. The linear symplectic group is a Lie group and as such contains a lot of structure. For more information on the linear symplectic group and its connection to physics, we refer the reader to [Gos06, MS98] chapter 2. An overview for physicists is also found in [Arv+95a]. All of the following can be found in that paper:

Definition 2.4. Let $O(2n, \mathbb{R})$ be the real orthogonal group, Then we define the following three subsets of $Sp(2n)$:

$$\begin{aligned} K(n) &:= Sp(2n, \mathbb{R}) \cap O(2n, \mathbb{R}), \\ Z(n) &:= \{\mathbb{1}_{2(j-1)} \oplus \mathrm{diag}(s_i, s_i^{-1}) \oplus \mathbb{1}_{2(n-(j+1))} | s \geqslant 0, j = 1, \ldots, n\}, \\ \Pi(n) &:= \{S \in Sp(2n, \mathbb{R}) | S \geqslant 0\}. \end{aligned}$$

The first subset is the *maximally compact subgroup* of $Sp(2n)$, the second subset is the subset of *single-mode-squeezers*. It generates the multiplicative subgroup $\mathcal{A}(2n)$, a maximally abelian subgroup of $Sp(2n)$. The third set is the set of positive definite symplectic matrices.

In addition, since $Sp(2n)$ is a Lie group, it possesses a *Lie algebra*. Let us collect a number of relevant facts about the Lie algebra and some subsets:

Proposition 2.5. *The Lie algebra* $\mathfrak{sp}(2n)$ *of* $Sp(2n)$ *is given by*

$$\mathfrak{sp}(2n) := \{T \in \mathbb{R}^{2n\times 2n} | \sigma T + T\sigma = 0\}$$

together with the commutator as Lie bracket. Certain other Lie algebras or subsets of Lie algebras are of relevance to us:

(1) $\mathfrak{so}(2n) := \{A \in \mathbb{R}^{2n\times 2n} | A + A^T = 0\}$ *the Lie algebra of* $SO(2n)$.

(2) $\mathfrak{k}(n) := \{A \in \mathbb{R}^{2n\times 2n} | A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}, a = -a^T, b = b^T\}$ *the Lie algebra of* $K(n)$.

(3) $\pi(n) := \{A \in \mathbb{R}^{2n\times 2n} | A = \begin{pmatrix} a & b \\ b & -a \end{pmatrix}, a = a^T, b = b^T\}$ *the subspace of the Lie algebra* $\mathfrak{sp}(2n)$ *corresponding to* $\Pi(n)$.

Since the Lie algebra is a vector space, it is spanned by a set of vectors, the *generators*. A standard decomposition is given by taking the generators of $\mathfrak{k}(n)$, the so called *passive transformations* as one part and the generators of $\pi(n)$, the so called *active transformations* as the other part. That these two sets together determine the Lie algebra completely can be seen with the *polar decomposition*:

Proposition 2.6 (**Polar decomposition** [Arv+95a]). *For every symplectic matrix* $S \in Sp(2n)$ *there exists a unique* $U \in K(n)$ *and a unique* $P \in \Pi(n)$ *such that* $S = UP$.

A basis for the Lie algebras $\mathfrak{k}(n)$ and $\pi(n)$ therefore characterises the complete Lie algebra $\mathfrak{sp}(2n)$. Elements of the Lie algebras are also called *generators* and a basis of generators therefore fixes the Lie algebra. Via the polar decomposition, this implies that they also generate the whole Lie group. We will need a set of generators $g_{ij}^{(p)} \in \mathfrak{k}(n)$ and $g_{ij}^{(a)} \in \pi(n)$ later on, which we will fix via the metaplectic representation:

Proposition 2.7 (**Metaplectic representation** [Arv+95a]). *Let* W_ξ *be the continuous irreducible Weyl system defined above and let* $S \in Sp(2n)$. *Then there exists an up to a phase unique unitary* U_S *with*

$$\forall \xi: \quad U_S W_\xi U_S^\dagger = W_{S\xi}.$$

Since we have the liberty of a phase, this is not really a representation of the symplectic group, but of its two-fold cover, the metaplectic group. We can also study the generators of this representation, which are given by $1/2\{R_k, R_l\}_+$.

For the reader familiar with annihilation and creation operators, if we denote by a_i, $a_i^\dagger$ the annihilation and creation operators of the n bosonic modes, the generators of the metaplectic representation are given by

$$G_{ij}^{p(1)} := \mathrm{i}(a_j^\dagger a_i - a_i^\dagger a_j) \qquad G_{ij}^{p(2)} := a_j^\dagger a_j + a_j^\dagger a_i, \tag{7}$$

$$G_{ij}^{a(3)} := \mathrm{i}(a_j^\dagger a_i^\dagger - a_i a_j) \qquad G_{ij}^{a(4)} := a_i^\dagger a_j^\dagger + a_i a_j, \tag{8}$$

where the p stands for 'passive' and the a for 'active'. The passive generators are also frequently called *linear transformations* in the literature (see [Kok+07]). We can now define a set of generators of the symplectic group $Sp(2n)$ by using the set of metaplectic generators G_{ij} above and take corresponding generators g_{ij} in the Lie algebra $\mathfrak{sp}(2n)$ in a consistent way. As one would expect from the name, the passive metaplectic generators correspond to a set of passive generators of $\mathfrak{k}(n)$ and the set of active metaplectic generators corresponds to a set of active generators of $\pi(n)$. The details of the correspondence are irrelevant (they are explicitly spelled out in equation (6.6b) in [Arv+95a]), except for the fact that the set $G_{ii}^{a(3)}$, $i = 1,\ldots,n$ corresponds to the generators $g_{ii}^{a(3)}$ generating matrices in Z_n.

Given a Hamiltonian, the associated time evolution corresponds to a path on the Lie group: for a (sufficiently regular) path $\gamma: [0, 1] \to Sp(2n)$ we can find a function $A(t) \in \mathfrak{sp}(2n)$ such that

$$\gamma'(t) = A(t)\gamma(t). \tag{9}$$

Instead of directly studying Hamiltonians with time-dependent coefficients as in equation (1), it is equivalent to study functions $A : [0, 1] \to \mathfrak{sp}(2n)$.

There are a number of decompositions of the Lie group and its subgroup in addition to the polar decomposition. We will mostly be concerned with the so called *Euler decomposition* (sometimes called *Bloch–Messiah* decomposition) and *Williamson's decomposition*:

Proposition 2.8 (**Euler decomposition** [Arv+95a]). *Let $S \in Sp(2n)$, then there exist $K, K' \in K(n)$ and $A \in \mathcal{A}(n)$ such that $S = KAK'$.*

Proposition 2.9 (**Williamson's theorem** [Wil36]). *Let $M \in \mathbb{R}^{2n\times 2n}$ be a positive definite matrix, then there exists a symplectic matrix $S \in Sp(2n, \mathbb{R})$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that*

$$M = S^T\tilde{D}S,$$

where $\tilde{D} = \mathrm{diag}(D, D)$ is diagonal. The entries of D are also called symplectic eigenvalues.

In particular, for $M \in \Pi(n)$, this implies that M has a symplectic square root. Since covariance matrices are always positive definite, this implies also that a Gaussian state is pure if and only if its covariance matrix is symplectic. Heisenberg's uncertainty principle has also a Williamson version:

Corollary 2.10. *A positive definite matrix M is a covariance matrix of a quantum state if and only if all symplectic eigenvalues are larger or equal to one.*

The proof is simple and therefore omitted.

2.3. Quantum optical operations and squeezing

We have already noted that an important class of operations are those, which leave the CCR-relations invariant, namely the symplectic transformations. Given a quantum state ρ, the action of the symplectic group on the canonical variables R descends to a subgroup of unitary transformations on ρ via the metaplectic representation (see [Arv+95b]). Its action on the covariance matrix γ_ρ of ρ is even easier: Given $S \in Sp(2n)$,

$$\gamma_\rho \mapsto S^T\gamma_\rho S. \tag{10}$$

In quantum optics, symplectic transformations can be implemented by the means of

(1) beam splitters and phase shifters, implementing operations in $K(n)$ ([Rec+94])
(2) single-mode squeezers, implementing operations in $Z(n)$.

Via the Euler decomposition, this implies that any symplectic transformation can be implemented (approximately) by a combination of those three elements.

Definition 2.11. An n-mode bosonic state ρ is called *squeezed*, if its covariance matrix γ_ρ possesses an eigenvalue $\lambda < 1$.

Especially in the early literature, squeezing is usually defined differently: a state ρ is squeezed if there exists a unitary transformation $K \in K(n)$ such that $K^T\gamma_\rho K$ has a diagonal entry smaller than one. This again comes from the physical definition of squeezed states being

states where the Heisenberg uncertainty relations are satisfied with equality for at least one mode. These definitions however are well-known to be equivalent (see [SMD94]).

3. An operational squeezing measure for symplectic transformations

Throughout this section, we will always use σ as our standard symplectic form.

3.1. Definition and basic properties

We will now define a first operational squeezing measure for symplectic transformations, which will later be used to define a measure for operational squeezing.

Definition 3.1. Define the function $F : \mathbb{R}^{2n\times 2n} \to \mathbb{R}$

$$F(A) = \sum_{i=1}^{n} \log(s_i^{\downarrow}(A)), \tag{11}$$

where $s_i^{\downarrow}$ are the decreasingly ordered singular values of A.

Note that we sum only over half of the singular values. Restricting this function to symplectic matrices will yield an operational squeezing measure for symplectic transformations: recall that the symplectic group is generated by symplectic orthogonal matrices and single-mode squeezers. The orthogonal matrices are easy to implement and therefore will be considered a free resource. The squeezers have singular values s and s^{-1} and they are experimentally hard to implement and should therefore be assigned a cost that depends on the squeezing parameter s. Using this, the amount of squeezing seems to be characterised by the largest singular values. Here, we quantify the amount of squeezing by a cost $\log(s)$, which can be seen as the interaction strength of the Hamiltonian needed to implement the squeezing.

Let us make this more precise: define the map

$$\begin{aligned} \Delta : Sp(2n) &\to \bigcup_{m\in\mathbb{N}} Sp(2n)^{\times m} \\ S &\mapsto \bigcup_{m\in\mathbb{N}} \{(S_1,\ldots,S_m) | S = S_1\cdots S_m,\ S_i \in K(n)\cup Z(n)\}. \end{aligned}$$

The image of Δ for a given symplectic matrix contains all possible ways to construct S as a product of matrices from $K(n)$ or $Z(n)$. We define:

Definition 3.2. Let $\overline{F} : Sp(2n) \to \mathbb{R}$ be a map defined via

$$\overline{F}(S) := \log \inf\left\{\prod_{i=1}^{m} s_1^{\downarrow}(S_i) \mid (S_1,\ldots,S_m) \in \Delta(S)\right\}. \tag{12}$$

Proposition 3.3. *If $S \in Sp(2n)$ then $F(S) = \overline{F}(S)$.*

Proof. Let $S = KAK'$ be the Euler decomposition of S with $K, K' \in K(n)$ and $A \in \mathcal{A}(n)$. Assume without loss of generality that $A = \mathrm{diag}(a_1, a_1^{-1},\ldots,a_n, a_n^{-1})$ and $a_1 \geqslant a_2 \geqslant \ldots \geqslant a_n \geqslant 1$ and define $A_i = \mathrm{diag}(1,\ldots,1, a_i, a_i^{-1}, 1,\ldots,1)$. By construction $A = A_1\cdots A_n$ and $A_i \in Z(n)$. Since $K, K' \in K(n)$, $(K, A_1,\ldots,A_n, K') \in \Delta(S)$. Using that $s_i^{\downarrow}(K) = s_i^{\downarrow}(K') = 1$ and the fact that the Euler decomposition is actually equivalent to the

singular value decomposition of S, we obtain:

$$\overline{F}(S) \leqslant \log\left(s_1^{\downarrow}(K)\prod_{i=1}^{n} s_1^{\downarrow}(A_i)s_1^{\downarrow}(K')\right) = \log\prod_{i=1}^{n} s_i^{\downarrow}(S) = F(S).$$

Conversely, consider $(S_1,\ldots,S_m) \in \Delta(S)$. Using that by definition for each $S_j \in K(n) \cup Z(n)$ we have $\prod_{i=1}^{n} s_i^{\downarrow}(S_j) = s_1^{\downarrow}(S_j)$, we conclude:

$$F(S) = \log\left(\prod_{i=1}^{n} s_i^{\downarrow}(S)\right) \overset{(*)}{\leqslant} \log\left(\prod_{j=1}^{m}\prod_{i=1}^{n} s_i^{\downarrow}(S_j)\right) = \log\left(\prod_{j=1}^{m} s_1^{\downarrow}(S_j)\right),$$

where in $(*)$ we used a special case of a theorem by Gel'fand and Naimark ([Bha96], theorem III.4.5 and equation (III.19)). Taking the infimum on the right-hand side gives $F(S) \leqslant \overline{F}(S)$. □

Let us write the last observation in $(*)$ as a small lemma for later use:

Lemma 3.4. *Let S, $S' \in Sp(2n)$. Then $F(SS') \leqslant F(S) + F(S')$.*

3.2. Lie algebraic definition

Up to now, we have only considered products of symplectic matrices, which would correspond to a chain of beam splitters, phase shifters and single-mode squeezers. The goal of this section is to prove that one cannot improve the results with arbitrary paths on $Sp(2n)$, corresponding to general Hamiltonians of the form of equation (1) as described in section 2.

Let $\mathcal{C}^r(S)$ be the set of absolutely continuous paths $\alpha : [0, 1] \to Sp(2n)$ with a derivative which is bounded almost everywhere such that $\alpha(0) = \mathbb{1}$ and $\alpha(1) = S$. Such paths seem to capture most if not all physically relevant cases.

Recall the set of generators g of $\mathfrak{sp}(2n)$ defined in section 2 and order them in a single vector. Usin equation (9), any $\alpha \in \mathcal{C}^r(S)$ corresponds to a $A \in L^{\infty}([0, 1], \mathfrak{sp}(2n))$. Since the generators g form a basis, we can write $A(t) = c_\alpha(t) \cdot g$ with a function $c_\alpha \in L^{\infty}([0, 1], \mathfrak{sp}(2n))$. Both A or c_α together with the condition $\alpha(0) = \mathbb{1}$ uniquely define α.

The goal of this section is to prove that this does not give us any better way to avoid squeezing:

Theorem 3.5. *For any $S \in Sp(2n)$, we have*

$$F(S) = \inf\left\{\int_0^1 \|\vec{c}_\alpha^{\,a}(t)\|_1 \,\mathrm{d}t \,\middle|\, \alpha \in \mathcal{C}^r(S),\ \dot{\alpha}(t) = (\vec{c}_\alpha^{\,p}(t)g^p(\alpha(t)),\ \vec{c}_\alpha^{\,a}(t)g^a(\alpha(t)))^T\right\}, \tag{13}$$

where we introduced the notation $\vec{c}$ to clarify that $g^{p/a}$ are actually vectors containing a set of generators each, and the coefficients might differ for each of these generators.

The proof of this theorem is quite lengthy in details, thus we split it up into several lemmata. The general idea is easy to relate: we first show that paths corresponding to products of symplectic matrices of type $Z(n)$ or $K(n)$ produce the same outcome in (13) and (12). We then use an approximation argument: given any path, we can approximate it by a path of products of symplectic matrices to arbitrary precision.

To start, we prove the following lemma:

Lemma 3.6. *Let* $A \in \mathfrak{sp}(2n)$ *and write* $A = 1/2(A + A^T) + 1/2(A - A^T) =: A_+ + A_-$. *Then* $A_+ \in \pi(2n)$ *and* $A_- \in \mathfrak{k}(n)$ *and we have* $F(\exp(A)) \leqslant F(\exp(A_+))$.

Proof. First note that F is continuous in S since the singular values are. Using the Trotter-formula, we obtain:

$$\begin{aligned} F(\exp(A)) &= F(\lim_{n\to\infty}(\exp(A_+/n)\exp(A_-/n))^n) \\ &\overset{(*)}{\leqslant} \lim_{n\to\infty}(nF(\exp(A_+/n)) + nF(\exp(A_-/n))) \\ &= \lim_{n\to\infty} nF(\exp(A_+/n)) = F(\exp(A_+)), \end{aligned}$$

where we used that $F(\exp(A_-)) = 0$ since $A_- \in \mathfrak{k}(n)$ and in $(*)$, we used a version of a theorem by Gel'fand and Naimark again (see [Bha96], equation (III.20)). □

Let us define yet another version of F which we call $\hat{F}$ in the following way:

$$\begin{aligned} C^N(S) &:= \left\{ (\vec{c}_1^a, \vec{c}_1^p, \ldots, \vec{c}_N^a, \vec{c}_N^p) \;\middle|\; S = \prod_{j=1}^{N} \exp(\vec{c}_j^a g^a + \vec{c}_j^p g^p),\ \vec{c}_j \in \mathbb{R}^{4n^2} \right\}, \\ C(S) &:= \bigcup_{N\in\mathbb{N}} C^N(S), \\ \hat{F}(S) &:= \inf\left\{ \sum_i \|\vec{c}_i^a\|_1 \;\middle|\; \vec{c} \in C(S) \right\}. \end{aligned}$$

This definition is of course reminiscent of the definition of $\overline{F}$ in equation (12):

Lemma 3.7. *For* $S \in Sp(2n)$, *we have* $\hat{F}(S) = \overline{F}(S)$.

Proof. To prove $\hat{F} \leqslant \overline{F}$, consider the Euler decomposition $S = K_1 A_1 \ldots A_n K_2$ with $A_i \in Z(n)$ and $K_1, K_2 \in Sp(2n)$. Since $K(n)$ is compact, the exponential map is surjective and there exist $\vec{c}_1^p$ and $\vec{c}_2^p$ such that $\exp(\vec{c}_1^p g^p) = K_1$ and $\exp(\vec{c}_2^p g^p) = K_2$. Recall that we ordered the vector g^a in such a way that the generators g_i^a generate the matrices in $Z(n)$ for $i = 1,\ldots,n$, hence we know that there exist $\vec{c}_i^a = (0,\ldots,0,(\vec{c}_i^a)_{(i)},0,\ldots,0)$ for $i = 1,\ldots,n$ such that

$$S = \exp(\vec{c}_1^p g^p)\prod_{i=1}^{n} \exp(\vec{c}_i^a g^a)\exp(\vec{c}_2^p g^p).$$

This implies

$$\hat{F}(S) \leqslant \sum_i \|\vec{c}_i^a\|_1 = \sum_i F(\exp(\vec{c}_i^a g^a)) = \sum_i F(\exp((\vec{c}_i^a)_{(i)} g_i^a)) = \sum_i \log s_1(A_i) = \overline{F}(S).$$

Here we used that $(\vec{c}_i^a)_{(i)}$ is also the largest singular value of $\exp((\vec{c}_i^a)_{(i)} g_i^a) \in Z(n)$, as $F(\exp((\vec{c}_i^a)_{(i)} g_i^a)) = (\vec{c}_i^a)_{(i)}$ by normalisation of g.

For the other direction $\hat{F} \geqslant \overline{F}$, let S be arbitrary. Let $c \in C(S)$ and consider each vector $\vec{c}_i$ separately. We drop the index i for readability, since we need to consider the entries of the vector $\vec{c}_i$. To make the distinction clear, we denote the jth entry of the vector $\vec{c}$ by $\vec{c}_{(j)}$. Recall that the active generators are exactly those generating the positive matrices. Then:

$$F(\exp(\vec{c}\,g)) \overset{\text{Lemma 3.6}}{\leqslant} F(\exp(\vec{c}^{\,a}g^a)) = \lim_{n\to\infty} F((\prod_i \exp(\vec{c}^{\,a}_{(i)}\, g_i^{\,a}/n))^n)$$
$$\leqslant \sum_i F(\exp(\vec{c}^{\,a}_{(i)}\, g_i^{\,a})) = \sum_i |\vec{c}^{\,a}_{(i)}| = \|\vec{c}^{\,a}\|_1,$$

where we basically redid the calculations we used to prove lemma 3.6, using the continuity of F and the Trotter formula from matrix analysis. Until now, we have considered only one $\vec{c}_i$ of $c \in C(S)$. Now, if we define $S_i = \exp(\vec{c}_i g)$, then we have $\prod_i S_i = S$ and hence, using lemma 3.4, we find:

$$\overline{F}(S) \overset{\text{Lemma 3.4}}{\leqslant} \sum_i F(S_i) \leqslant \sum_i F(\exp(\vec{c}_i g)) \leqslant \sum_i \|\vec{c}_i^{\,a}\|_1 \qquad \forall\, c \in C(S).$$

But this means $\overline{F}(S) \leqslant \hat{F}(S)$, as we claimed. □

We can now prove the first half of the theorem:

Lemma 3.8. *For $S \in Sp(2n)$ we have*

$$F(S) \geqslant \inf\left\{\int_0^1 \|\vec{c}^{\,a}_\alpha(t)\|_1 \,\mathrm{d}t \middle| \alpha \in \mathcal{C}^r(S),\ \dot{\alpha}(t)\right\}.$$

Proof. Let $S \in Sp(2n)$ and consider the Euler decomposition $S = K_1 S_1 \cdots S_n K_2$. We can define a function $A : [0, 1] \to \mathfrak{sp}(2n)$ via:

$$A(t) := \begin{cases} (n+2)\cdot \vec{c}_1^{\,p} g^p & t \in [0,\ 1/(n+2)), \\ (n+2)\cdot (\vec{c}^{\,a}_{i+1})_{(i)} g_i^{\,a} & t \in [i/(2(n+2)),\ (i+1)/(n+2)),\ i = 1,\ldots,n, \\ (n+2)\cdot \vec{c}^{\,p}_{n+2} g^p & t \in [(n+1)/(n+2),\ 1], \end{cases} \tag{14}$$

where $(\vec{c}_1^{\,p},\ 0,\ 0,\ \vec{c}_2^{\,a},\ldots,0,\ \vec{c}^{\,a}_{n+1},\ \vec{c}^{\,p}_{n+2},\ 0)$ denotes the element in $C^{n+2}(S)$ for the Euler decomposition and vector indices are denoted by a subscript $_{(i)}$ as before. Let $U(s, t)$ be the propagator corresponding to A, then for $t \in [0,\ 1/(n+2))$ according to proposition A.1, since A does not depend on t on this interval, it is given by $U(t, s) = \exp((t-s)A)$. In particular, $U(1/(n+2),\ 0) = \exp(\vec{c}^{\,p}_{n+2} g^p) = K_2$.

Iterating the procedure above, using $U(0,\ 1) = U(0,\ 1/(n+2)) \cdots U((n+1)/(n+2),\ 1)$, we can see that by construction, $U(0,\ 1) = K_1 S_1 \cdots S_n K_2 = S$. Hence A defines a continuous path on $Sp(2n)$ via $U(s, t)$. We can calculate:

$$\int_0^1 \|\vec{c}^{\,a}(t)\|_1 \,\mathrm{d}t = \sum_{i=1}^n \int_{i/(n+2)}^{(i+1)/(n+2)} |(n+2)\cdot (\vec{c}^{\,a}_{i+1})_{(i)}|\,\mathrm{d}t$$
$$= \sum_{i=1}^n |(\vec{c}^{\,a}_{i+1})_{(i)}| \overset{\text{Lemma 3.7}}{=} F(S),$$

where we used that the integral over the interval $[0,\ 1/(n+2))$ and $[(n+1)/(n+2),\ 1]$ is empty due to the fact that all active components are zero. In the last step, we used that for the Euler decomposition, which takes the minimum in $\hat{F}$, this value is exactly $\sum_i |(\vec{c}^{\,a}_{i+1})_{(i)}| = \sum_i \|\vec{c}^{\,a}_{i+1}\|_1$, since $(\vec{c}^{\,a}_{i+1})_{(j)} = 0$ for $j \neq i$. Taking the infimum on the left-hand side only decreases the value. □

For the other direction, we need some facts about ordinary differential equations that are collected in appendix A.

Lemma 3.9. *For $S \in Sp(2n)$ we have*

$$F(S) \leqslant \inf\left\{\int_0^1 \|\vec{c}_\alpha^a(t)\|_1 \, \mathrm{d}t \mid \alpha \in \mathcal{C}^r(S), \dot{\alpha}(t)\right\}. \tag{15}$$

Proof. Let $S \in Sp(2n)$ be arbitrary. Combining the proof of lemma 3.8 with proposition 3.3 and lemma 3.7 we have already proved:

$$F(S) = \inf\left\{\int_0^1 \|\vec{c}_\alpha^a(t)\|_1 \, \mathrm{d}t \,\middle|\, \alpha \in \mathcal{C}^r(S), \right.$$
$$\left. \dot{\alpha}(t) = (\vec{c}_\alpha^p(t) g^p(\alpha(t)), \vec{c}_\alpha^a(t) g^a(\alpha(t)))^T, \vec{c} \text{ step fct.}\right\}.$$

The only thing left to prove is that we can drop the step-function assumption. This will be done by a standard approximation argument: let $\tilde{F}(S)$ denote the right-hand side of equation (15). Let $\varepsilon > 0$ and consider an arbitrary $A \in L^\infty$ such that

$$\left|\int_0^1 \|\vec{c}_\alpha^a(t)\|_1 \, \mathrm{d}t - \tilde{F}(S)\right| < \varepsilon \tag{16}$$

i.e. A corresponds to a path that is close to the infimum in the definition of $\tilde{F}$. We can now approximate c_α by step-functions $c_{\alpha'}$ (corresponding to a function A', see lemma A.2) such that

$$\left|\int_0^1 \|c_\alpha(t) - c_{\alpha'}\|_1 \, \mathrm{d}t\right| < \varepsilon. \tag{17}$$

Using the fact that the propagators U_A, $U_{A'}$ are differentiable almost everywhere (proposition A.1) and absolutely continuous when one entry is fixed, we can define a function $f(s) := U_A(0, s) U_{A'}(s, t)$, which is also differentiable almost everywhere. Furthermore, the fundamental theorem of calculus holds for $f(s)$ (see [Rud87], theorems 6.10 and 7.8).

$$\frac{\mathrm{d}}{\mathrm{d}s} f(s) = -U_A(0, s) A(s) U_{A'}(s, t) + U_A(0, s) A'(s) U_{A'}(s, t)$$

almost everywhere, which implies:

$$U_{A'}(0, t) - U_A(0, t) = f(t) - f(0) = \int_0^t \frac{\mathrm{d}}{\mathrm{d}s} f(s) \mathrm{d}s$$
$$= \int_0^t U_A(0, s)(A'(s) - A(s)) U_{A'}(s, t) \mathrm{d}s.$$

Since U and g are bounded in $\|\cdot\|_\infty$, we obtain

$$\|U_{A'}(0, t) - U_A(0, t)\|_1 \leqslant M \int_0^t \|c_{\alpha'}(s) - c_\alpha(s)\|_1 \, \mathrm{d}s \leqslant M\varepsilon. \tag{18}$$

M can explicitly be computed by the bounds given in proposition A.1.

Up to now, we have taken a path α to S close to the infimum and approximated it by a path α'. It is immediate by equations (16) and (17) that

$$\left| \int_0^1 \|c_{\alpha'}(t)\|_1 \, \mathrm{d}t - \tilde{F}(S) \right| < 2\varepsilon. \tag{19}$$

Since $c_{\alpha'} \in C^N(S')$ for some $N \in \mathbb{N}$ and $S' = U_{A'}(0, 1)$, we would be done if $S' = S$. To remedy this, we want to extend α' to a path $\tilde{\alpha}$ such that it ends at S. This is where equation (18) enters: set $\tilde{S} \coloneqq U_{A'}(0, 1)^{-1} U_A(0, 1)$, then

$$\|\tilde{S} - \mathbb{1}\|_1 \leqslant \frac{M}{\|S\|_1} \varepsilon \tag{20}$$

hence $\tilde{S} \approx \mathbb{1}$ for ε small enough. Using the polar decomposition, we can write $\tilde{S} = \exp(\vec{c}^{\,p}_{N+1}) \exp(\vec{c}^{\,a}_{N+2})$. A quick calculation yields

$$\| \log \tilde{S} \|_1 \leqslant n \log \left(\varepsilon \frac{M}{\|S\|_1} \right) \leqslant n\varepsilon \frac{M}{\|S\|_1} =: C\varepsilon. \tag{21}$$

This lets us construct a new $\tilde{A} : [0, 2] \to \mathfrak{sp}(2n)$:

$$t \mapsto \begin{cases} A'(t) & t \in [0, 1], \\ 2 \cdot \vec{c}^{\,p}_{N+1} g^p & t \in (1, 3/2), \\ 2 \cdot \vec{c}^{\,a}_{N+2} g^a & t \in (3/2, 2]. \end{cases}$$

By construction, for the corresponding propagator we have $U_{\tilde{A}}(0, 2) = S$ and $\tilde{\alpha}$ is a feasible path for $\tilde{F}(S)$ (at least after reparameterisation) fulfiling:

$$\begin{aligned} \left| \int_0^2 \|c^a_{\tilde{\alpha}}(t)\|_1 \, \mathrm{d}t - \tilde{F}(S) \right| &\leqslant \left| \int_0^1 \|c^a_{\tilde{\alpha}}(t)\|_1 \, \mathrm{d}t - \tilde{F}(S) \right| + \left| \int_0^1 \|c^a_{\tilde{\alpha}}(t)\|_1 \, \mathrm{d}t \right| \\ &\overset{(19)+(21)}{\leqslant} (2 + C)\varepsilon. \end{aligned}$$

Since, $c_{\tilde{\alpha}} \in C^{N+2}(S)$, $\tilde{\alpha}$ is a valid path for $\hat{F}(S)$, which implies that for any $\epsilon > 0$, choosing $\varepsilon \coloneqq \epsilon/(2 + C)$, we have seen:

$$\hat{F}(S) < \tilde{F}(S) + \epsilon. \tag{22}$$

For $\epsilon \to 0$, $\hat{F}(S) \leqslant \tilde{F}(S)$, which implies the lemma via lemma 3.7. □

4. A mathematical measure for squeezing of arbitrary states

Throughout this section, for convenience, we will switch to using J as symplectic form. Having defined the measure F, we will now proceed to define a squeezing measure for creating an arbitrary (mixed) state:

Definition 4.1. Let ρ be an n-mode bosonic quantum state with covariance matrix Γ. We then define:

$$G(\rho) \equiv G(\Gamma) \coloneqq \inf\{F(S) | \Gamma \geqslant S^T S, \, S \in Sp(2n)\}. \tag{23}$$

Note that G is always finite: for any given covariance matrix Γ, by Williamson's theorem and corollary 2.10, we can find $S \in Sp(2n)$ and $\tilde{D} \geqslant \mathbb{1}$ such that $\Gamma = S^T \tilde{D} S \geqslant S^T S$. Furthermore G is also non-negative since F is non-negative for symplectic S. We will prove in section 5 that this is indeed an operational measure.

4.1. Different reformulations of the measure

We will now give several reformulations of the squeezing measure and prove some of its properties. In particular, G is convex and one of the crucial steps towards proving convexity of G is given by a reformulation of G with the help of the Cayley transform. For the reader unfamiliar with the Cayley transform, a definition and basic properties are provided in appendix B.

Proposition 4.2. *Let* $\Gamma \geqslant \mathrm{i}J$ *and* $\Gamma \in \mathbb{R}^{2n\times 2n}$ *symmetric. Then:*

$$G(\Gamma) = \inf\{F(\Gamma_0^{1/2}) | \Gamma \geqslant \Gamma_0 \geqslant \mathrm{i}J\} \tag{24}$$

$$= \inf\left\{\frac{1}{2}\sum_{i=1}^{n}\log\left(\frac{1+s_i(A+\mathrm{i}B)}{1-s_i(A+\mathrm{i}B)}\right) \mid \mathcal{C}^{-1}(\Gamma) \geqslant H,\, H\in\mathcal{H}\right\}, \tag{25}$$

where $\mathcal{H}$ *is defined via:*

$$\mathcal{H} = \left\{H = \begin{pmatrix} A & B \\ B & -A\end{pmatrix} \in \mathbb{R}^{2m\times 2m} \mid A^T = A,\, B^T = B,\, \mathrm{spec}(H)\subset(-1,1)\right\}. \tag{26}$$

Proof. First note that the infimum in all three expressions is actually attained. We can see this most easily in the definition (23): the matrix inequalities $\Gamma \geqslant S^TS\,(\geqslant \mathrm{i}J)$ imply that the set of feasible S in the minimisation is compact, hence its minimum is attained. To see (23) = (24), first note that (24) $\leqslant$ (23) since any $S\in Sp(2n)$ also fulfils $S^TS \geqslant \mathrm{i}J$, hence $\Gamma \geqslant S^TS \geqslant \mathrm{i}J$. For equality, note that for any $\Gamma \geqslant \Gamma_0 \geqslant \mathrm{i}J$, using Williamson's theorem we can find $S\in Sp(2n)$ and a diagonal $\tilde{D}\geqslant \mathbb{1}$ (via corollary 2.10) such that $\Gamma_0 = S^TDS \geqslant S^TS \geqslant \mathrm{i}J$. But since $F(\Gamma_0^{1/2}) \geqslant F((S^TS)^{1/2}) = F(S)$ via the Weyl monotonicity principle, the infimum is achieved on symplectic matrices.

Finally, let us prove equality with (25). First observe that we can replace $Sp(2n)$ by $\mathcal{H}$ using proposition B.1(4).

Using the fact that $s_i^{\downarrow}(S) = \lambda_i^{\downarrow}(S^TS)^{1/2} = \lambda_i^{\downarrow}(\mathcal{C}(H))^{1/2}$ and the fact that H is diagonalised by the same unitary matrices as $\mathcal{C}(H) = (\mathbb{1}+H)\cdot(\mathbb{1}-H)^{-1}$ whence its eigenvalues are

$$\lambda_i^{\downarrow}(\mathcal{C}(H)) = \frac{1+\lambda_i^{\downarrow}(H)}{1-\lambda_i^{\downarrow}(H)},$$

we have:

$$\inf\{F(S)|\Gamma\geqslant S^TS,\, S\in Sp(2n)\} = \inf\left\{\log\prod_{i=1}^{n}\left(\frac{1+\lambda_i^{\downarrow}(H)}{1-\lambda_i^{\downarrow}(H)}\right)^{\frac{1}{2}} |\Gamma\geqslant\mathcal{C}(H),\, H\in\mathcal{H}\right\}.$$

Next we claim $\lambda_i^{\downarrow}(H) = s_i^{\downarrow}(A+\mathrm{i}B)$ for $i=1,\ldots,n$. To see this note:

$$\frac{1}{2}\begin{pmatrix}\mathbb{1} & \mathrm{i}\mathbb{1}\\ \mathbb{1} & -\mathrm{i}\mathbb{1}\end{pmatrix}\cdot\begin{pmatrix}A & B\\ B & -A\end{pmatrix}\cdot\begin{pmatrix}\mathbb{1} & \mathbb{1}\\ -\mathrm{i}\mathbb{1} & \mathrm{i}\mathbb{1}\end{pmatrix} = \begin{pmatrix}0 & A+\mathrm{i}B\\ A-\mathrm{i}B & 0\end{pmatrix}. \tag{27}$$

The singular values of the matrix on the right-hand side of equation (27) are the eigenvalues of $\mathrm{diag}((A+\mathrm{i}B)^{\dagger}(A+\mathrm{i}B),\,(A+\mathrm{i}B)(A+\mathrm{i}B)^{\dagger})^{1/2}$, which are the singular values of $A+\mathrm{i}B$ with double multiplicity. From the structure of H, it is immediate that the eigenvalues of the right-hand side of equation (27) and thus of H come in pairs $\pm s_i(A+\mathrm{i}B)$. Hence

$\lambda_i^{\downarrow}(H) = s_i^{\downarrow}(A + \mathrm{i}B)$ for $i = 1,\ldots,n$ and we have:

$$\begin{aligned}&\inf\{F(S)|\Gamma \geqslant S^T S,\ S \in Sp(2n)\}\\ &= \inf\left\{\frac{1}{2}\sum_{i=1}^{n}\log\left(\frac{1 + s_i(A + \mathrm{i}B)}{1 - s_i(A + \mathrm{i}B)}\right) \mid \Gamma \geqslant \mathcal{C}(H),\ H \in \mathcal{H}\right\}.\end{aligned}$$

To see that that the right-hand side equals (25), we only need to use the fact that $\Gamma \geqslant \mathcal{C}(H) \Leftrightarrow \mathcal{C}^{-1}(\Gamma) \geqslant H$ for all $H \in \mathcal{H}$ and $\Gamma \geqslant \mathrm{i}J$ since the Cayley transform and its inverse are operator monotone. □

4.2. Convexity

The reformulation (25) will allow us to prove:

Theorem 4.3. *G is convex on the set of covariance matrices $\{\Gamma \in \mathbb{R}^{2n\times 2n} | \Gamma \geqslant \mathrm{i}J\}$.*

The crucial part of the proof is the following lemma:

Lemma 4.4. *Consider the map $f : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \to \mathbb{R}$:*

$$f(A, B) = \frac{1}{2}\sum_{i=1}^{n}\log\left(\frac{1 + s_i(A + \mathrm{i}B)}{1 - s_i(A + \mathrm{i}B)}\right). \tag{28}$$

If we restrict f to symmetric matrices A and B such that $s_i(A + \mathrm{i}B) < 1$ for all $i = 1,\ldots,n$, f is jointly convex in A, B, i.e.

$$f(tA + (1 - t)A', tB + (1 - t)B') \leqslant tf(A, B) + (1 - t)f(A', B') \qquad \forall t \in [0, 1].$$

Proof. Let $\tilde{A} := tA + (1 - t)A'$ and $\tilde{B} := tB + (1 - t)B'$. Note that $\tilde{A}$ and $\tilde{B}$ are also symmetric, and the largest singular value of $\tilde{A} + \mathrm{i}\tilde{B}$ fulfils $s_1^{\downarrow}(\tilde{A} + \mathrm{i}\tilde{B}) \leqslant ts_1^{\downarrow}(A + \mathrm{i}B) + (1 - t)s_1^{\downarrow}(A' + \mathrm{i}B')$. Therefore, the singular values of any convex combination of $A + \mathrm{i}B$ and $A' + \mathrm{i}B'$ also lie in the interval $[0, 1)$. This makes our restriction well-defined under convex combinations.

For any $j = 1,\ldots,n$, by Thompson's theorem (see [Tho76]), which states that for every complex A, B, there exist unitary matrices U, V such that $|A + B| \leqslant U|A|U^* + V|B|V^*$, we have

$$s_j(\tilde{A} + \mathrm{i}\tilde{B}) = \lambda_j(|\tilde{A} + \mathrm{i}\tilde{B}|) \leqslant \lambda_j(U|t(A + \mathrm{i}B)|U^* + V|(1 - t)(A' + \mathrm{i}B')|V^*).$$

Using Lidskii's theorem ([Bha96], chapter III with explicit formulation in exercise III.4.3), we have

$$\begin{aligned}s_j(\tilde{A} + \mathrm{i}\tilde{B}) &\leqslant \lambda_j(U|t(A + \mathrm{i}B)|U^*) + \sum_{\pi} p_\pi \lambda_{\pi(j)}(V|(1 - t)(A' + \mathrm{i}B')|V^*)\\ &\overset{(*)}{=} \lambda_j(|t(A + \mathrm{i}B)|) + \sum_{\pi} p_\pi \lambda_{\pi(j)}(|(1 - t)(A' + \mathrm{i}B')|)\\ &= t\lambda_j(|A + \mathrm{i}B|) + (1 - t)\sum_{\pi} p_\pi \lambda_{\pi(j)}(|A' + \mathrm{i}B'|)\end{aligned} \tag{29}$$

with $p_\pi \geqslant 0$ and $\sum_\pi p_\pi = 1$. In $(*)$, we used that unitaries do not change the spectrum. Now each summand in equation (28) is the Cayley transform of a singular value. We can use the

log-convexity of the Cayley-transform to prove the joint convexity of f:

$$\begin{aligned}
f(\tilde{A}, \tilde{B}) &= \sum_{i=1}^{n} \log \mathcal{C}[s_i(\tilde{A} + \mathrm{i}\tilde{B})] \\
&\leqslant \sum_{i=1}^{n} \log \mathcal{C}\left[t\lambda_i(|A + \mathrm{i}B|) + (1-t)\sum_{\pi} p_\pi \lambda_{\pi(i)}(|A' + \mathrm{i}B'|) \right] \\
&\leqslant \sum_{i=1}^{n} \left(t \log \mathcal{C}[\lambda_i(|A + \mathrm{i}B|)] + (1-t)\sum_{\pi} p_\pi \log \mathcal{C}[\lambda_{\pi(i)}(|A' + \mathrm{i}B'|)] \right) \\
&= \sum_{i=1}^{n} t \log \mathcal{C}[\lambda_i(|A + \mathrm{i}B|)] + (1-t)\sum_{\pi} p_\pi \left(\sum_{i=1}^{n} \log \mathcal{C}[\lambda_{\pi(i)}(|A' + \mathrm{i}B'|)] \right) \\
&\leqslant t\sum_{i=1}^{n} \log \mathcal{C}[\lambda_i(|A + \mathrm{i}B|)] + (1-t)\sum_{\pi} p_\pi \cdot \max_{\pi} \left(\sum_{i=1}^{n} \log \mathcal{C}[\lambda_{\pi(i)}(|A' + \mathrm{i}B'|)] \right) \\
&\overset{(**)}{=} t\sum_{i=1}^{n} \log \mathcal{C}[\lambda_i(|A + \mathrm{i}B|)] + (1-t)\sum_{i=1}^{n} \log \mathcal{C}[\lambda_i(|A' + \mathrm{i}B'|)] \\
&= tf(A, B) + (1-t)f(A', B'),
\end{aligned}$$

where in $(**)$ we use that the sum of all eigenvalues is of course not dependent on the order of the eigenvalues. □

This lemma will later allow us to calculate G as a convex programme.

Proof of theorem 4.3. We can now finish the proof of the convexity of G.

First note that using the definition of f in lemma 4.4 we can reformulate (25) to

$$G(\Gamma) = \inf\{f(A, B) | \mathcal{C}^{-1}(\Gamma) \geqslant H, H \in \mathcal{H}\}. \tag{30}$$

Let $\Gamma \geqslant \mathrm{i}J$, $\Gamma' \geqslant \mathrm{i}J$ be two covariance matrices and let H, $H' \in \mathcal{H}$ be the matrices that attain the minimum of $G(\Gamma)$, $G(\Gamma')$ respectively. Then, in particular, $tH + (1-t)H' \in \mathcal{H}$. Furthermore, since $\mathcal{C}^{-1}(\Gamma) \geqslant H$ and $\mathcal{C}^{-1}(\Gamma') \geqslant H'$ we have

$$\mathcal{C}^{-1}(t\Gamma + (1-t)\Gamma') \overset{(*)}{\geqslant} t\mathcal{C}^{-1}(\Gamma) + (1-t)\mathcal{C}^{-1}(\Gamma') \geqslant tH + (1-t)H',$$

where we used the operator concavity of $\mathcal{C}^{-1}$ in $(*)$. This means that $tH + (1-t)H'$ is a feasible matrix for the minimisation in G, which implies using equation (30)

$$G(t\Gamma + (1-t)\Gamma') \leqslant f(tA + (1-t)A', tB + (1-t)B').$$

The convexity now follows directly from lemma 4.4 and the fact that we chose H and H' to attain $G(\Gamma)$ and $G(\Gamma')$. □

4.3. Continuity properties

From the convexity of G on the set of covariance matrices, it follows from general arguments in convex analysis that G is continuous on the interior of the set of covariance matrices (see [Roc97], theorem 10.1). What more can we say about the boundary?

Theorem 4.5. *G is lower semicontinuous on the set of covariance matrices $\{\Gamma \in \mathbb{R}^{2n\times 2n} | \Gamma \geqslant \mathrm{i}J\}$ and continuous on its interior. Moreover, $G(\Gamma + \varepsilon \mathbb{1}) \to G(\Gamma)$ for $0 < \varepsilon \to 0$ for any $\Gamma \geqslant \mathrm{i}J$.*

The ultimate goal is to extend continuity from the interior to the exterior, which we do not know how to do at present. The proof will need a few notions from set-valued analysis that we review in appendix C.

Proof of theorem 4.5. As already observed, G is continuous on the interior. Let $\Gamma_0 \geqslant \mathrm{i}J$ be arbitrary and suppose

$$\mathcal{A}(\Gamma) := \{\hat{\Gamma} | (\Gamma - 2\|\Gamma_0\|\mathbb{1}) \leqslant \hat{\Gamma} \leqslant \Gamma\}.$$

By definition, $\mathcal{A}$ is compact and convex for any Γ. Moreover, it defines a set-valued function on the set of covariance matrices with non-empty values. Let $\varepsilon > 0$, then for all $\Gamma \geqslant \mathrm{i}J$ with $\|\Gamma - \Gamma_0\| < \varepsilon$, we have that for any $\hat{\Gamma} \in \mathcal{A}(\Gamma)$, $\tilde{\Gamma} := \hat{\Gamma} + (\Gamma - \Gamma_0) \in \mathcal{A}(\Gamma_0)$ and $\|\hat{\Gamma} - \tilde{\Gamma}\| < \varepsilon$. This is the condition in lemma C.2 hence the set-valued function defined by $\mathcal{A}$ is upper semicontinuous at Γ_0, which implies that $\mathcal{A}(\Gamma) \cap \{X | \mathrm{i}J \leqslant X\}$ is also upper semicontinuous by proposition C.3. If ε is small enough (e.g. $\varepsilon < 1$), this implies

$$\mathcal{A}(\Gamma) \bigcap \{X | \mathrm{i}J \leqslant X\} = \{X | \mathrm{i}J \leqslant X \leqslant \Gamma\} =: \mathcal{G}(\Gamma),$$

hence this set is upper semicontinuous at Γ_0.

Since F is continuous on positive definite matrices, it is absolutely continuous if we restrict to a small neighbourhood of the covariance matrix Γ_0. This means that for every $\varepsilon > 0$ there exists an $\epsilon > 0$ such that

$$F(\tilde{\Gamma}) - \varepsilon < F(\hat{\Gamma}) < F(\tilde{\Gamma}) + \varepsilon \tag{31}$$

for all $\|\tilde{\Gamma} - \hat{\Gamma}\| < \epsilon$ and all $\tilde{\Gamma}, \hat{\Gamma} \in \bigcup_{\|\Gamma - \Gamma_0\| < 1} \mathcal{G}(\Gamma)$.

Assuming without loss of generality that $\|\Gamma - \Gamma_0\| < 1$, the set $\mathcal{G}(\Gamma)$ is exactly the set for the minimisation in the definition of G. The upper semicontinuity of $\mathcal{G}(\Gamma)$ implies by lemma C.2 that for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $\|\Gamma - \Gamma_0\| < \delta$ we have: for all $\hat{\Gamma} \in \mathcal{G}(\Gamma)$ there exists a $\tilde{\Gamma} \in \mathcal{G}(\Gamma_0)$ such that $\|\hat{\Gamma} - \tilde{\Gamma}\| < \epsilon$. In particular, this is true for all minimisers $\hat{\Gamma}$ with $G(\Gamma) = F(\hat{\Gamma}^{1/2})$, where $\hat{\Gamma}$ and $\tilde{\Gamma} \in \bigcup_{\|\Gamma - \Gamma_0\| < 1} \mathcal{G}(\Gamma)$. Using equation (31) we obtain: for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $\|\Gamma - \Gamma_0\| < \delta$, we have a pair $\hat{\Gamma}$, $\tilde{\Gamma}$ with $\hat{\Gamma} \in \mathcal{G}(\Gamma)$ minimising $G(\Gamma)$ and $\tilde{\Gamma} \in \mathcal{G}(\Gamma_0)$ such that

$$F(\tilde{\Gamma}) - \varepsilon < F(\hat{\Gamma}) = G(\Gamma).$$

This implies that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$G(\Gamma_0) \leqslant G(\Gamma) + \varepsilon$$

for all $\|\Gamma - \Gamma_0\| < \delta$.

Taking the limit inferior on both sides implies that G is lower semicontinuous at Γ_0. Upper semicontinuity would follow for instance if $\mathcal{G}(\Gamma_0)$ is also lower semicontinuous.

Finally, let us prove that $G(\Gamma_0 + \varepsilon\mathbb{1}) \to 0$ for $\varepsilon \to 0$. To see this, consider the closed sets

$$C_n := \bigcup_{0 \leqslant \xi \leqslant 1/n} \mathcal{G}(\Gamma_0 + \xi\mathbb{1})$$

for any $n \in \mathbb{N}$. It is easy to see that $C_{n+1} \subseteq C_n$ and that $\bigcap_{n\in\infty} C_n = \mathcal{G}(\Gamma_0)$. Moreover, C_1 is compact. Now let Γ_n be the sequence of minimisers for $G(\Gamma_0 + 1/n\mathbb{1})$, then $\Gamma_n \in C_n$ for all $n \in \mathbb{N}$. By compactness, a subsequence will converge to

$$\Gamma \in \bigcap_{n\in\infty} C_n = \mathcal{G}(\Gamma_0).$$

Therefore, $G(\Gamma_0) \leqslant \lim_{\varepsilon\to 0} G(\Gamma_0 + \varepsilon\mathbb{1})$, but since $\Gamma_0 \leqslant \Gamma_0 + \varepsilon\mathbb{1}$ for all $\varepsilon > 0$ we also have $G(\Gamma) \geqslant \lim_{\varepsilon\to 0} G(\Gamma_0 + \varepsilon\mathbb{1})$. □

4.4. Additivity properties

Now we consider additivity properties of G. We switch our basis again and use γ and σ.

Proposition 4.6. *For any covariance matrices $\gamma_A \in \mathbb{R}^{2n_1\times 2n_1}$ and $\gamma_B \in \mathbb{R}^{2n_2\times 2n_2}$, we have*

$$\frac{1}{2}(G(\gamma_A) + G(\gamma_B)) \leqslant G(\gamma_A \oplus \gamma_B) \leqslant G(\gamma_A) + G(\gamma_B).$$

In particular, G is subadditive.

Proof. For subadditivity, let $S^T S \leqslant \gamma_A$ and $S'^T S' \leqslant \gamma_B$ obtain the minimum in $G(\gamma_A)$ and $G(\gamma_B)$ respectively. Then $S \oplus S'$ is symplectic and $(S \oplus S')^T (S \oplus S') \leqslant \gamma_A \oplus \gamma_B$ hence, $G(\gamma_A \oplus \gamma_B) \leqslant G(A) + G(B)$.

To prove the lower bound, we need the following equation that we will only prove later on (see equation (46)):

$$a \geqslant 1 : \qquad G(\gamma_A) \leqslant G(\gamma_A \oplus a\mathbb{1}_{n_2}). \tag{32}$$

Assuming this inequality, let $a \geqslant 1$ be such that $a\mathbb{1}_{n_2} \geqslant \gamma_B$, then

$$G(\gamma_A \oplus a\mathbb{1}_{n_2}) \leqslant G(\gamma_A \oplus \gamma_B)$$

hence $G(\gamma_A) \leqslant G(\gamma_A \oplus \gamma_B)$ and since we can do the same reasoning for γ_B, we have $G(\gamma_A) + G(\gamma_B) \leqslant 2G(\gamma_A \oplus \gamma_B)$. □

We do not know whether G is also superadditive, which would make it additive. At present, we can only prove:

Corollary 4.7. *Let $\gamma_A \in \mathbb{R}^{2n_1\times 2n_1}$ and $\gamma_B \in Sp(2n_2)$, be two covariance matrices (i.e. γ_B is a covariance matrix of a pure state). Then G is additive.*

Proof. Subadditivity has already been proven in the lemma. For superadditivity, we use the second reformulation of the squeezing measure in equation (24): note that there is only one matrix $\gamma_B \geqslant \gamma \geqslant \mathrm{i}\sigma$, namely γ_B itself. Now write

$$\gamma_A \oplus \gamma_B \geqslant \begin{pmatrix} \tilde{A} & C \\ C^T & \tilde{B} \end{pmatrix} \geqslant \mathrm{i}\sigma$$

for $\tilde{A} \in \mathbb{R}^{2n_1\times 2n_1}$ and $\tilde{B} \in \mathbb{R}^{2n_2\times 2n_2}$. Then in particular $\gamma_B - \tilde{B} \geqslant 0$, but also $\tilde{B} \geqslant \mathrm{i}\sigma$, hence $\gamma_B \geqslant \tilde{B} \geqslant \mathrm{i}\sigma$ and therefore $\tilde{B} = \gamma_B$. But then

$$\gamma_A \oplus \gamma_B - \begin{pmatrix} \tilde{A} & C \\ C^T & \tilde{B} \end{pmatrix} = \begin{pmatrix} \gamma_A - \tilde{A} & C \\ C^T & 0 \end{pmatrix}$$

hence also $C = 0$ and the matrix that takes the minimum in $G(\gamma_A \oplus \gamma_B)$ must be block-diagonal. Then $\gamma_A \oplus \gamma_B \geqslant \tilde{A} \oplus \gamma_B \geqslant 0$ and $\tilde{A}$ is in the feasible set of $G(\gamma_A)$. □

Corollary 4.8. *For any covariance matrices* $\gamma_A \in \mathbb{R}^{2n_1\times 2n_1}$ *and* $\gamma_B \in \mathbb{R}^{2n_2\times 2n_2}$,

$$G(\gamma_A) + G(\gamma_B) \leqslant 2G\left(\begin{pmatrix}\gamma_A & C\\ C^T & \gamma_B\end{pmatrix}\right).$$

If G is superadditive, then this inequality holds without the factor of two.

Proof.

$$G(\gamma_A) + G(\gamma_B) \leqslant 2G\left(\begin{pmatrix}\gamma_A & 0\\ 0 & \gamma_B\end{pmatrix}\right) = 2G\left(\frac{1}{2}\begin{pmatrix}\gamma_A & C\\ C^T & \gamma_B\end{pmatrix} + \frac{1}{2}\begin{pmatrix}\gamma_A & -C\\ -C^T & \gamma_B\end{pmatrix}\right)$$
$$\overset{(*)}{\leqslant} G\left(\begin{pmatrix}\gamma_A & C\\ C^T & \gamma_B\end{pmatrix}\right) + G\left(\begin{pmatrix}\gamma_A & -C\\ -C^T & \gamma_B\end{pmatrix}\right) \overset{(**)}{=} 2G\left(\begin{pmatrix}\gamma_A & C\\ C^T & \gamma_B\end{pmatrix}\right).$$

Here we used proposition 4.6 and then convexity of G in $(*)$. Finally, in $(**)$ we used that for every

$$\begin{pmatrix}\gamma_A & C\\ C^T & \gamma_B\end{pmatrix} \geqslant \begin{pmatrix}S_A & \tilde{C}\\ \tilde{C}^T & S_B\end{pmatrix}\begin{pmatrix}S_A & \tilde{C}\\ \tilde{C}^T & S_B\end{pmatrix}^T \in Sp(2(n_1+n_2)) \tag{33}$$

we also have:

$$\begin{pmatrix}\gamma_A & -C\\ -C^T & \gamma_B\end{pmatrix} \geqslant \begin{pmatrix}S_A & -\tilde{C}\\ -\tilde{C}^T & S_B\end{pmatrix}\begin{pmatrix}S_A & -\tilde{C}\\ -\tilde{C}^T & S_B\end{pmatrix}^T \in Sp(2(n_1+n_2)) \tag{34}$$

and vice versa. Since the two matrices on the right-hand side of equations (33) and (34) have equal spectrum, the two squeezing measures of the matrices on the left-hand side need to be equal. □

4.5. Bounds

Let us give a few simple bounds on G.

Proposition 4.9 (**Spectral bounds**). *Let* $\Gamma \geqslant \mathrm{i}J$ *be a valid covariance matrix and* $\lambda^{\downarrow}(\Gamma)$ *be the vector of eigenvalues in decreasing order. Then:*

$$-\frac{1}{2}\sum_{\lambda_i^{\downarrow}(\Gamma)<1}\log(\lambda_i^{\downarrow}(\Gamma)) \leqslant G(\Gamma) \leqslant \frac{1}{2}\sum_{i=1}^{n}\log\lambda_i^{\downarrow}(\Gamma) = F(\Gamma^{1/2}). \tag{35}$$

Proof. According to the Euler decomposition, a symplectic positive definite matrix has positive eigenvalues that come in pairs s, s^{-1} and we can find $O \in SO(2n)$ such that for any $S^TS \leqslant \Gamma$

$$O^T\Gamma O \geqslant \mathrm{diag}(s_1,\ldots,s_n,\, s_1^{-1},\ldots,s_n^{-1}).$$

But then, $\lambda_k^{\downarrow}(\Gamma) \geqslant \lambda_k^{\downarrow}(\mathrm{diag}(s_1,\ldots,s_n,\, s_1^{-1},\ldots,s_n^{-1}))$ via the Weyl inequalities $\lambda_i^{\downarrow}(A) \geqslant \lambda_i^{\downarrow}(B)$ for all i and $A - B \geqslant 0$ (see [Bha96], theorem III.2.3). This implies:

$$G(\Gamma) \leqslant \sum_{i=1}^{n} \log(\max\{s_i, s_i^{-1}\}) \leqslant \sum_{i=1}^{n} \log \lambda_i^{\downarrow}(\Gamma)^{1/2}.$$

For the lower bound, given an optimal matrix S with eigenvalues s_i, we have

$$G(\Gamma) = \sum_{i} \max\{s_i, s_i^{-1}\}.$$

If $S^T S = O^T \mathrm{diag}(s_1^2, \ldots, s_n^2, s_1^{-2}, \ldots, s_n^{-2}) O$ with $O \in SO(2n)$ is the diagonalisation of $S^T S$, we can write:

$$O^{-T} \Gamma^{-1} O^{-1} \leqslant \mathrm{diag}(s_1^2, \ldots, s_n^2, s_1^{-2}, \ldots, s_n^{-2})$$

and again by Weyl's inequalites, we can find for all $k \leqslant n$:

$$-\frac{1}{2} \sum_{i=2n-k+1}^{2n} \log(\lambda_i^{\downarrow}(\Gamma)) \leqslant \frac{1}{2} \sum_{i=1}^{k} \log \lambda_i^{\downarrow}(\mathrm{diag}(s_1^2, \ldots, s_n^2, s_1^{-2}, \ldots, s_n^{-2})) \leqslant G(\Gamma). \tag{36}$$

Now, $-\frac{1}{2}\sum_{i=2n-k+1}^{2n} \lambda_i^{\downarrow}(\Gamma)$ can be upper bounded by restricting to eigenvalues $\lambda_i^{\downarrow}(\Gamma) < 1$. This implies

$$-\frac{1}{2} \sum_{\lambda_i^{\downarrow}(\Gamma)<1} \log(\lambda_i^{\downarrow}(\Gamma)) \leqslant G(\Gamma)$$

using that the number of eigenvalues $\lambda_i(\Gamma) < 1$ can at most be n (hence $k \leqslant n$ in the inequality of equation (36)), since $\Gamma \geqslant S^T S$ and $S^T S$ has at least n eigenvalues bigger than one. □

Numerics suggest that the lower bound is often very good for low dimensions. In fact, it can sometimes be achieved:

Proposition 4.10. *Let $\Gamma \geqslant \mathrm{i}J$ be a covariance matrix, then G achieves the lower bound in equation (35) if there exists an orthonormal eigenvector basis $\{v_i\}_{i=1}^{2n}$ of Γ with $v_i^T J v_j = \delta_{i,n+j}$. Conversely, if G achieves the lower bound, then $v_i^T J v_j = 0$ for all normalised eigenvectors v_i, v_j of Γ with λ_i, $\lambda_j < 1$.*

Proof. Suppose that the lower bound in equation (35) is achieved. Via Weyl's inequalities (see [Bha96] theorem III.2.3), for all $S^T S \leqslant \Gamma$ in the definition of G we have $\lambda_i^{\downarrow}(S^T S) \leqslant \lambda_i^{\downarrow}(\Gamma)$. For the particular S achieving G, this implies that for all $\lambda_i^{\downarrow}(\Gamma) < 1$ we have $\lambda_i^{\downarrow}(S^T S) = \lambda_i^{\downarrow}(\Gamma)$. But then $\Gamma \geqslant S^T S$ implies that $S^T S$ and Γ share all eigenvectors to the smallest eigenvalue. Iteratively, every eigenvector of Γ with $\lambda_i(\Gamma) < 1$ must be an eigenvector of $S^T S$ with the same eigenvalue.

Since the matrix diagonalising $S^T S$ also diagonalises $\mathcal{C}^{-1}(S^T S)$, the eigenvectors of the two matrices are the same. Now, since $\mathcal{C}^{-1}(S^T S) \in \mathcal{H}$ by reformulation (25), for any eigenvector v_i of any eigenvalue $\mathcal{C}^{-1}(\lambda_i) < 0$, Jv_i is also an eigenvector of $\mathcal{C}^{-1}(S^T S)$ to the eigenvalue $-\mathcal{C}^{-1}(\lambda_i)$, implying $v_i^T J v_j = 0$ for all i, j. By definition, this means that $\{v_i, Jv_j\}$ forms a symplectic basis. Above, we already saw that the eigenvectors of Γ for $\lambda_i(\Gamma) < 1$ are also eigenvalues of $S^T S$, hence $v_i^T J v_j = 0$ for all i such that $\lambda_i(\Gamma) < 1$.

Conversely, suppose we have an orthonormal basis $\{v_i\}_{i=1}^{2n}$ such that $v_i^T J v_j = \delta_{i,j+n}$ (modulo $2n$ if necessary) for all eigenvectors of Γ, i.e. Γ is diagonalisable by a symplectic orthonormal matrix $\tilde{O} \in U(n)$. Then

$$\tilde{O}\Gamma\tilde{O}^T = \mathrm{diag}(\lambda_1, \ldots, \lambda_{2n}).$$

Since $\Gamma \geqslant \mathrm{i}J$ we have $\lambda_i \lambda_{2i} \geqslant 1$. Assume that $\lambda_i \geqslant \lambda_{n+i}$ for all $i = 1,\ldots,n$ and the λ_{n+i} are ordered in decreasing order. Then $\lambda_{n+r} < 1 \leqslant \lambda_{n+r-1}$ for some $r \leqslant n$ and

$$S^T S = \tilde{O}^T \mathrm{diag}(1, \ldots, 1,\ \lambda_r^{-1}, \ldots, \lambda_n^{-1},\ 1, \ldots, 1,\ \lambda_{n+r}, \ldots, \lambda_{2n})\tilde{O}$$

fulfils $S^T S \leqslant \Gamma$ and obviously achieves the lower bound in equation (35). □

In contrast to this, the upper bound can be arbitrarily bad. For instance, consider the thermal state $\Gamma = (2N + 1) \cdot \mathbb{1}$ for increasing N. It can easily be seen that $G(\Gamma) = 0$, since $\Gamma \geqslant \mathbb{1} \in \Pi(n)$ and $F(\mathbb{1}) = 0$, hence $G(\Gamma) \leqslant 0$. However, the upper bound in equation (35) is $n/2 \log(2N + 1) \to \infty$ for $N \to \infty$, therefore arbitrarily bad.

We can achieve better upper bounds by using Williamson's normal form:

Proposition 4.11 (Williamson bounds). *Let $\Gamma \in \mathbb{R}^{2n\times 2n}$ be such that $\Gamma \geqslant \mathrm{i}J$ and consider its Williamson normal form $\Gamma = S^T D S$. Then:*

$$F(S) - \log(\sqrt{\det(\Gamma)}) \leqslant G(\Gamma) \leqslant F(S). \tag{37}$$

Proof. Since $D \geqslant \mathbb{1}$ via $\Gamma \geqslant \mathrm{i}J$, the upper bound follows directly from the definition. Also, $F(S) \leqslant F(\Gamma^{1/2})$, which makes this bound trivially better than the spectral upper bound in equation (35).

The lower bound follows from:

$$\begin{aligned} G(\Gamma) &\overset{(36)}{\geqslant} \frac{1}{2}\log\left(\prod_{i=n+1}^{2n} \lambda_i^{\downarrow}(\Gamma)^{-1}\right) = \frac{1}{2}\log\frac{\prod_{i=1}^{n}\lambda_i^{\downarrow}(\Gamma)}{\prod_{i=1}^{2n}\lambda_i^{\downarrow}(\Gamma)} \\ &= F(\Gamma^{1/2}) - \log(\det(\Gamma)^{1/2}) \geqslant F((S^T S)^{1/2}) - \log(\sqrt{\det(\Gamma)}) \\ &= F(S) - \log(\sqrt{\det(\Gamma)}) \end{aligned}$$

using Weyl's inequalities once again, implying that since $S^T S \leqslant \Gamma$, we also have $F(S)^2 = F(S^T S) \leqslant F(\Gamma)$. □

The upper bound here can also be arbitrarily bad. One just has to consider $\Gamma := S^T(N \cdot \mathbb{1})S$ with $S^2 = \mathrm{diag}(N - 1, \ldots, N - 1, (N - 1)^{-1}, \ldots, (N - 1)^{-1}) \in Sp(2n)$. Then $\Gamma \geqslant \mathbb{1}$, i.e. $G(\Gamma) = 0$, but $F(S) \to \infty$ for $N \to \infty$.

Proposition 4.12. *Let $\Gamma \geqslant \mathrm{i}J$ be a covariance matrix. Then*

$$G(\Gamma) \geqslant \frac{1}{4}\inf\{\|\gamma_0\|_1 | \log\Gamma \geqslant \gamma_0,\ \gamma_0 \in \pi(n)\}, \tag{38}$$

where $\pi(n)$ was defined in proposition 2.5 as the Lie algebra of the positive semidefinite symplectic matrices. This infimum can be computed efficiently as a semidefinite programme.

Proof. Recall that the logarithm is operator monotone on positive definite matrices. Using this, we have:

$$\begin{aligned} G(\Gamma) &= \log \inf\left\{\prod_{i=1}^{n} \lambda_i^{\downarrow}(S^T S)^{1/2} | \Gamma \geqslant S^T S\right\} \\ &\geqslant \inf\left\{\sum_{i=1}^{n} \log \lambda_i^{\downarrow}(\exp(\gamma_0))^{1/2} \,|\log \Gamma \geqslant \gamma_0,\, \gamma_0 \in \pi(n)\right\} \\ &= \inf\left\{\frac{1}{2}\sum_{i=1}^{n} \lambda_i^{\downarrow}(\gamma_0) | \log \Gamma \geqslant \gamma_0,\, \gamma_0 \in \pi(n)\right\} \\ &= \inf\left\{\frac{1}{4}\sum_{i=1}^{2n} s_i^{\downarrow}(\gamma_0) | \log \Gamma \geqslant \gamma_0,\, \gamma_0 \in \pi(n)\right\}. \end{aligned}$$

The last step is valid, because the eigenvalues of matrices $\gamma_0 \in \pi(n)$ come in pairs $\pm\lambda_i$. Since the sum of all the singular values is just the trace-norm, we are done.

It remains to see that this can be computed by a semidefinite programme. First note that since the matrices $H \in \pi(n)$ are those symmetric matrices with $HJ + JH = 0$, the constraints are already linear semidefinite matrix inequalities. The trace norm is an SDP by standard reasoning [RFP10, VB96]:

$$\|\gamma_0\|_1 = \min\left\{\frac{1}{2}\mathrm{tr}(A + B) \left| \begin{pmatrix} A & \gamma_0 \\ \gamma_0 & B \end{pmatrix} \geqslant 0\right.\right\}$$

which is clearly a semidefinite programme. □

Numerics for small dimensions suggest that this bound is mostly smaller than the spectral lower bounds.

5. An operational definition of the squeezing measure

We claim that G answers the question: given a state, what is the minimal amount of single-mode squeezers needed to prepare it? In other words, it quantifies the amount of squeezing needed for the preparation of a state.

5.1. Operations for state preparation and an operational measure for squeezing

We first specify the preparation procedure. Since we want to quantify squeezing, it seems natural that we allow to freely draw states from the vacuum or a thermal bath to start with. Furthermore, we can perform an arbitrary number of the following operations for free:

(1) Add ancillary states also from a thermal bath or the vacuum.
(2) Add Gaussian noise.
(3) Implement any gate from linear optics.
(4) Perform Weyl-translations of the state.
(5) Perform selective or non-selective Gaussian measurements such as homodyne or heterodyne detection.
(6) Forget part of the state.
(7) Create convex combinations of ensembles.

In addition, the following operation comes with an associated cost:

(8) Implement single-mode squeezers at a cost of $\log(s)$, where s is the squeezing parameter.

All these operations are standard operations in quantum optics and they should capture all important Gaussian operations except for squeezing.

It is well-known that all of these operations are captured by the following set of operations on the covariance matrix (for a justification, see appendix D):

(O0) We can always draw N-mode states with $\gamma \in \mathbb{R}^{2N\times 2N}$ for any dimension N from the vacuum $\gamma = \mathbb{1}$ or a bath fulfiling $\gamma \geqslant \mathbb{1}$.
(O1) We can always add ancillary modes from the vacuum $\gamma_{\text{anc}} = \mathbb{1}$ or a bath $\gamma \geqslant \mathbb{1}$ and consider $\gamma \oplus \gamma_{\text{anc}}$.
(O2) We can freely add noise with $\gamma_{\text{noise}} \geqslant 0$ to our state, which is simply added to the covariance matrix of a state.
(O3) We can perform any beam splitter or phase shifter and in general any operation $S \in K(n)$, which translates to a map $\gamma \mapsto S^T\gamma S$ on covariance matrices of states.
(O4) We can perform any single-mode squeezer $S = \text{diag}(1, \ldots, 1, s, s^{-1}, 1 \ldots, 1)$ for some $s \in \mathbb{R}_+$.
(O5) We can perform any Weyl-translation leaving the covariance matrix invariant.
(O6) Given two states with covariance matrices γ_1 and γ_2, we can always take their convex combination $p\gamma_1 + (1-p)\gamma_2$ for any $p \in [0, 1]$.
(O7) At any point, we can perform a selective measurement of the system corresponding to a projection into a finitely or infinitely squeezed state. Given a state with covariance matrix $\gamma = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, this maps
$$\gamma \mapsto A - C(B - \gamma_G)^{\text{MP}}C^T,$$
where $^{\text{MP}}$ denotes the Moore–Penrose pseudoinverse.

Only operation (O4) comes at a cost of $\log(s)$, all other operations are free.

We are now ready to state our main theorem, which states that the minimal squeezing cost for any possible preparation procedure consisting of operations (1)–(8). is given by G.

Theorem 5.1. *Let ρ be a quantum state with covariance matrix γ. Consider arbitrary sequences*

$$\vec{\gamma}_N := \gamma_0 \to \gamma_1 \to \cdots \to \gamma_N,$$

where γ_0 fulfils (O0) and every arrow corresponds to an arbitrary operation (O1)–(O5) or (O7). Using (O6), we can merge two sequences $\vec{\gamma}_{N_1}$ and $\vec{\gamma}_{N_2}$ to one resulting tree with $\gamma_{N_1+N_2+1} = \lambda\gamma_{N_1} + (1-\lambda)\gamma_{N_2}$ for some $\lambda \in (0, 1)$. Iteratively, we can construct trees of any depth and width using operations (O1)–(O7).

Let $\mathfrak{O}_N(\gamma)$ be the set of such trees with N operations ending with γ (i.e. $\gamma_N = \gamma$). Let $\mathfrak{O}(\gamma) = \bigcup_{N=1}^{\infty} \mathfrak{O}_N(\gamma)$.

Furthermore, for any tree $\hat{\gamma} \in \mathfrak{O}_N(\gamma)$, let $\vec{s} = \{s_i\}_{i=1}^M$ be the sequence of the largest singular values of any single-mode squeezer (O4) implemented along the sequence (in particular, $M \leqslant N$). Then

$$G(\gamma) = \inf\left\{\sum_i \log s_i | s_i \in \vec{s},\ \hat{\gamma} \in \mathfrak{O}(\gamma)\right\}. \tag{39}$$

5.2. Proof of the main theorem

Since we consider many different operations, the proof is rather lengthy, where the main difficulties will be in showing that measurements do not squeeze. In order to increase readability, the proof will be split into several lemmata.

Lemma 5.2. *Let $\gamma \in \mathbb{R}^{2n\times 2n}$ be a covariance matrix, $\gamma_0 \geqslant \mathbb{1}$, let $N \in \mathbb{N}$ and*

$$\gamma_0 \to \gamma_1 \to \cdots \to \gamma_N = \gamma \tag{40}$$

be any sequence of actions (O1)–(O5) or (O7). If we denote the cost (sum of the logarithm of the largest singular values of any symplectic matrix involved) of this sequence by c, then one can replace this sequence by:

$$\begin{aligned}\gamma_0 &\overset{(O1)}{\to} \gamma_0 \oplus \gamma_{\text{anc}} \overset{(O2)}{\to} \gamma_0 \oplus \gamma_{\text{anc}} + \gamma_{\text{noise}} \overset{(O3),\,(O4)}{\to} S^T(\gamma_0 \oplus \gamma_{\text{anc}} + \gamma_{\text{noise}})S \\ &\overset{(O7)}{\to} \mathcal{M}(S^T(\gamma_0 \oplus \gamma_{\text{anc}} + \gamma_{\text{noise}})S)\end{aligned} \tag{41}$$

with $\gamma_{\text{anc}} \geqslant \mathbb{1}$, $\gamma_{\text{noise}} \geqslant 0$, $S \in Sp(2n)$ and $\mathcal{M}$ a partial Gaussian measurement of type specified in (O7). For this sequence, $c \geqslant F(S)$.

Proof. We prove the proposition by proving that given any chain $\gamma_0 \to \gamma_1 \to \cdots \to \gamma_N = \gamma$ as in (40), we can interchange all operations and obtain a chain as in equation (41). For readability, we will not always specify the size of the matrices and we will assume that $\gamma \geqslant i\sigma$, $\gamma_{\text{anc}} \geqslant \mathbb{1}$, $\gamma_{\text{noise}} \geqslant 0$, and S a symplectic matrix, whenever the symbols arise:

(1) We can combine any sequence $\gamma_i \to \gamma_{i+1} \to \cdots \to \gamma_{i+m}$ for some $m \in \mathbb{N}$ where each of the arrows corresponds to a symplectic transformation S_j, $j = 1,\ldots,m$ as in (O3) or (O4), into a single symplectic matrix $S \in Sp(2n)$ such that $\gamma_{i+m} = S^T\gamma_i S$. Furthermore lemma 3.4 implies $F(S) \leqslant \sum_i s_1^{\downarrow}(S_i)$, hence this recombination of steps only lowers the amount of squeezing.

(2) Any sequence $\gamma \to S^T\gamma S \to S^T\gamma S + \gamma_{\text{noise}}$ can be converted into a sequence $\gamma \to S^T(\gamma + \tilde{\gamma}_{\text{noise}})S$ with the same S and hence the same costs by setting $\tilde{\gamma}_{\text{noise}} := S^{-T}\gamma_{\text{noise}}S^{-1} \geqslant 0$.

(3) Any sequence $\gamma \to S^T\gamma S \to S^T\gamma S \oplus \gamma_{\text{anc}}$ can be converted into a sequence $\gamma \to \gamma \oplus \gamma_{\text{anc}} \to \tilde{S}^T(\gamma \oplus \gamma_{\text{anc}})\tilde{S}$ by setting $\tilde{S} = S \oplus \mathbb{1}$ with $\mathbb{1}$ of the same dimension as γ_{anc}. Since we only add the identity, we have $F(\tilde{S}) = \sum_i \log s_i^{\downarrow}(\tilde{S}) = F(S)$ and the costs do not increase.

(4) Any sequence $\gamma \to \gamma + \gamma_{\text{noise}} \to (\gamma + \gamma_{\text{noise}}) \oplus \gamma_{\text{anc}}$ can be converted into a sequence $\gamma \to \gamma \oplus \gamma_{\text{anc}} \to \gamma \oplus \gamma_{\text{anc}} + \tilde{\gamma}_{\text{noise}}$ by setting $\tilde{\gamma}_{\text{noise}} = \gamma_{\text{noise}} \oplus 0 \geqslant 0$, which is again a valid noise matrix. As no operation of type (O4) is involved, the squeezing costs do not change.

In a next step we consider measurements. We will only consider homodyne detection, since the proof is exactly the same for arbitrary Gaussian measurements of type (O7). Given a covariance matrix γ, we assume a decomposition

$$\gamma = \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}; \qquad \mathcal{M}(\gamma) = A - C(\pi B\pi)^{\text{MP}}C^T$$

as in the definition of (O7) with $\pi = \text{diag}(1, 0)$.

(5) Any sequence $\gamma \to \mathcal{M}(\gamma) \to S^T\mathcal{M}(\gamma)S$ can be converted into a sequence $\gamma \to \tilde{S}^T\gamma\tilde{S} \to \mathcal{M}(\tilde{S}^T\gamma\tilde{S})$ by setting $\tilde{S} = S \oplus \mathbb{1}_2$. To see this, write $S^T\mathcal{M}(\gamma)S = S^TAS - S^TC(\pi B\pi)^{\mathrm{MP}}C^TS$ and

$$\mathcal{M}\left(\begin{pmatrix} S & 0 \\ 0 & \mathbb{1} \end{pmatrix}^T \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}\begin{pmatrix} S & 0 \\ 0 & \mathbb{1} \end{pmatrix}\right) = \mathcal{M}\left(\begin{pmatrix} S^TAS & S^TC \\ C^TS & B \end{pmatrix}\right) = S^TAS - S^TC(\pi B\pi)^{\mathrm{MP}}C^TS$$

hence the final covariance matrices are the same. By the same reasoning as in (3), the costs are equivalent.

(6) Any sequence $\gamma \to \mathcal{M}(\gamma) \to \mathcal{M}(\gamma) + \gamma_{\mathrm{noise}}$ can be converted into a sequence $\gamma \to \gamma + \tilde{\gamma}_{\mathrm{noise}} \to \mathcal{M}(\gamma + \tilde{\gamma}_{\mathrm{noise}})$ by setting $\tilde{\gamma}_{\mathrm{noise}} = \gamma_{\mathrm{noise}} \oplus 0$, with 0 on the last mode being measured. Since no symplectic matrices are involved, the costs are equivalent.

(7) Any sequence $\gamma \to \mathcal{M}(\gamma) \to \mathcal{M}(\gamma) \oplus \gamma_{\mathrm{anc}}$ can be changed into a sequence $\gamma \to \gamma \oplus \gamma_{\mathrm{anc}} \to \tilde{\mathcal{M}}(\gamma \oplus \gamma_{\mathrm{anc}})$, where the measurement $\tilde{\mathcal{M}}$ measures the last mode of γ, i.e.

$$\tilde{M}\left(\begin{pmatrix} A & C & 0 \\ C^T & B & 0 \\ 0 & 0 & \gamma_{\mathrm{anc}} \end{pmatrix}\right) = (A \oplus \gamma_{\mathrm{anc}}) - (C \oplus 0)(\pi B\pi)^{\mathrm{MP}}(C \oplus 0)^T.$$

Clearly, the resulting covariance matrices of the two sequences are the same and the costs are equivalent.

We can now easily prove the lemma. Let $\gamma_0 \to \cdots \to \gamma_n$ be an arbitrary sequence with operations of type (O1)–(O5) or (O7). We can first move all measurements to the right of the sequence, i.e. we first perform all operations of type (O1)–(O5) and then all measurements. This is done using the observations above. Note also that this step is similar to the quantum circuit idea to 'perform all measurements last' (see [NC00], chapter 4).

Similarly, we can combine operations of type (O3) and (O4) and rearrange the other operations to obtain a new sequence as in equation (41) with at most the costs of the sequence $\gamma_1 \to \cdots \to \gamma_m$ we started with. □

We can now slowly work towards theorem 5.1:

Lemma 5.3. *Let $\gamma \in \mathbb{R}^{2n\times 2n}$ be a covariance matrix, then*

$$G(\gamma) = \inf\{F(S) | \gamma = S^T(\gamma_0 \oplus \gamma_{\mathrm{anc}} + \gamma_{\mathrm{noise}})S,\ S \in Sp(2n),\ \gamma_0 \oplus \gamma_{\mathrm{anc}} \geqslant \mathbb{1},\ \gamma_{\mathrm{noise}} \geqslant 0\}. \tag{42}$$

Proof. First note that for any $\gamma \geqslant i\sigma$, we can find $S \in Sp(2n)$, $\gamma_0 \in \mathbb{R}^{2n\times 2n}$ with $\gamma_0 \geqslant \mathbb{1}$ and $\gamma_{\mathrm{noise}} \in \mathbb{R}^{2n\times 2n}$ with $\gamma_{\mathrm{noise}} \geqslant 0$ such that $\gamma = S^T(\gamma_0 + \gamma_{\mathrm{noise}})S$ by using Williamson's theorem, hence the feasible set is never empty. The lemma is immediate by observing that for any $\gamma = S^T(\gamma_0 \oplus \gamma_{\mathrm{anc}} + \gamma_{\mathrm{noise}})S$ since $(\gamma_0 \oplus \gamma_{\mathrm{anc}} + \gamma_{\mathrm{noise}}) \geqslant \mathbb{1}$ we have $\gamma \geqslant S^TS$ and conversely, for any $\gamma \geqslant S^TS$, defining $\gamma_0 := S^{-T}\gamma S^{-1} \geqslant \mathbb{1}$, we have $\gamma = S^T\gamma_0 S$. □

As an intermediate step we introduce the following notation:

$$\tilde{G}(\gamma) \coloneqq \inf\{F(S)|\gamma = \mathcal{M}(S^T(\gamma_0 \oplus \gamma_{\text{anc}} + \gamma_{\text{noise}})S),\ S \in Sp(2n),\ \gamma_0 \oplus \gamma_{\text{anc}} \geqslant \mathbb{1}_{2n},\ \gamma_{\text{noise}} \geqslant 0,\ \mathcal{M} \text{ measurement}\}. \tag{43}$$

Then we have:

Lemma 5.4. *For $\gamma \in \mathbb{R}^{2n\times 2n}$ a covariance matrix, we have*

$$\tilde{G}(\gamma) = \inf\{F(\hat{\gamma}^{1/2})|\gamma = \mathcal{M}(\tilde{\gamma}),\ \tilde{\gamma} \geqslant \hat{\gamma} \geqslant \mathrm{i}\sigma,\ \mathcal{M} \text{ measurement}\}. \tag{44}$$

Proof. This follows from lemma 5.3:

$$\begin{aligned}
\tilde{G}(\gamma) &= \inf\{F(S)|\gamma = \mathcal{M}(S^T(\gamma_0 \oplus \gamma_{\text{anc}} + \gamma_{\text{noise}})S)\} \\
&= \inf\{F(S)|\gamma = \mathcal{M}(\tilde{\gamma}),\ \tilde{\gamma} = S^T(\gamma_0 \oplus \gamma_{\text{anc}} + \gamma_{\text{noise}})S \geqslant \mathrm{i}\sigma\} \\
&\overset{\text{Lemma 5.3}}{=} \inf\{G(\tilde{\gamma})|\gamma = \mathcal{M}(\tilde{\gamma}),\ \tilde{\gamma} \geqslant \mathrm{i}\sigma\} \\
&\overset{\text{Prop. 4.2}}{=} \inf\{F(\hat{\gamma}^{1/2})|\gamma = \mathcal{M}(\tilde{\gamma}),\ \tilde{\gamma} \geqslant \hat{\gamma} \geqslant \mathrm{i}\sigma\}
\end{aligned} \tag{45}$$

by taking the infimum over all measurements last. □

Note here, that equation (45) together with the following proposition 5.5 finishes the proof of proposition 4.6 via:

$$G(\gamma) = \inf\{G(\tilde{\gamma})|\gamma = \mathcal{M}(\tilde{\gamma}),\ \tilde{\gamma} \geqslant \mathrm{i}\sigma\} \leqslant G(\gamma \oplus a\mathbb{1}_{n_2}) \tag{46}$$

for $a \geqslant 1$, using that measuring the last modes we obtain $\mathcal{M}(\gamma \oplus a\mathbb{1}_{n_2}) = \gamma$ and therefore, $\gamma \oplus a\mathbb{1}_{n_2}$ is in the feasible set of $\tilde{G}(\gamma) = G(\gamma)$.

Proposition 5.5. *For $\gamma \in \mathbb{R}^{2n\times 2n}$ a covariance matrix we have*

$$\tilde{G}(\gamma) = G(\gamma).$$

This proposition shows that G is operational if we exclude convex combinations (and therefore also non-selective measurements).

Proof. Using lemma 5.4, the proof of this proposition reduces to the question whether:

$$\inf\{F(\hat{\gamma}^{1/2})|\tilde{\gamma} \geqslant \hat{\gamma} \geqslant \mathrm{i}\sigma,\ \mathcal{M}(\tilde{\gamma}) = \gamma\} = \inf\{F(\overline{\gamma}^{1/2})|\gamma \geqslant \overline{\gamma} \geqslant \mathrm{i}\sigma\}. \tag{47}$$

Since we do not need to use measurements, $\leqslant$ is obvious.

Let $\tilde{\gamma} \geqslant \hat{\gamma} \geqslant \mathrm{i}\sigma$ for some $\mathcal{M}(\tilde{\gamma}) = \gamma$. Our first claim is that

$$\gamma \geqslant \mathcal{M}(\hat{\gamma}) \geqslant \mathrm{i}\sigma \tag{48}$$

$\mathcal{M}(\hat{\gamma}) \geqslant \mathrm{i}\sigma$ is clear from the fact that $\hat{\gamma}$ is a covariance matrix and a measurement takes states to states. $\gamma \geqslant \mathcal{M}(\hat{\gamma})$ is proved using *Schur complements*. Let $\mathcal{M}$ be a Gaussian measurement as in equation (68) with $\gamma_G = \mathrm{diag}(d, 1/d)$ with $d \in \mathbb{R}^+$. It is well-known that

$$\mathcal{M}(\gamma) = (\mathbb{1} \oplus \mathrm{diag}(1/d, d)\gamma(\mathbb{1} \oplus \mathrm{diag}(1/d, d)) + 0 \oplus \mathbb{1}_2)^S,$$

where ^{S} denotes the Schur complement of the block in the lower-right corner of the matrix. For homodyne measurements, we take the limit $d \to \infty$. Since for any $\tilde{\gamma} \geqslant \hat{\gamma} \geqslant 0$, the Schur complements of the lower right block fulfil $\tilde{\gamma}^S \geqslant \hat{\gamma}^S \geqslant 0$ (see [Bha07], exercise 1.5.7), we have $\gamma \geqslant \mathcal{M}(\hat{\gamma})$ as claimed in equation (48).

Next, we claim

$$F(\mathcal{M}(\hat{\gamma})^{1/2}) \leqslant F(\hat{\gamma}^{1/2}). \tag{49}$$

To prove this claim, note that via the monotonicity of the exponential function on $\mathbb{R}$, it suffices to prove

$$\prod_{j=1}^{m} s_j^{\downarrow}(\mathcal{M}(\hat{\gamma})) \leqslant \prod_{j=1}^{n} s_j^{\downarrow}(\hat{\gamma})$$

when we assume $\hat{\gamma} \in \mathbb{R}^{2n\times 2n}$ and $\mathcal{M}(\hat{\gamma}) \in \mathbb{R}^{2m\times 2m}$ with $m \leqslant n$. If we write

$$\hat{\gamma} = \begin{pmatrix} \hat{A} & \hat{C} \\ \hat{C}^T & \hat{B} \end{pmatrix}$$

then the state after measurement is given by $\mathcal{M}(\hat{\gamma}) = \hat{A} - \hat{C}(\hat{B} + \mathrm{diag}(d, 1/d))^{-1}\hat{C}^T$ or the limit $d \to \infty$ for homodyne measurements. In any case $\hat{C}(\hat{B} + \mathrm{diag}(d, 1/d))^{-1}\hat{C}^T \geqslant 0$ and $\mathcal{M}(\hat{\gamma}) \leqslant \hat{A}$ and therefore, by Weyl's inequalities, also

$$\prod_{j=1}^{m} s_j^{\downarrow}(\mathcal{M}(\hat{\gamma})) \leqslant \prod_{j=1}^{m} s_j^{\downarrow}(\hat{A}).$$

Now we use Cauchy's interlacing theorem (see [Bha96], corollary III.1.5): as $\hat{A}$ is a submatrix of $\hat{\gamma}$, we have $\lambda_i^{\downarrow}(\hat{A}) \leqslant \lambda_i^{\downarrow}(\hat{\gamma})$ for all $i = 1, \ldots, 2m$. Since at least m eigenvalues of $\hat{A}$ are bigger or equal one and at least n eigenvalues of $\hat{\gamma}$ are bigger or equal one, this implies

$$\prod_{j=1}^{m} s_j^{\downarrow}(\hat{A}) = \prod_{j=1}^{m} \lambda_j^{\downarrow}(\hat{A}) \leqslant \prod_{j=1}^{m} \lambda_j^{\downarrow}(\hat{\gamma}) \leqslant \prod_{j=1}^{n} \lambda_j^{\downarrow}(\hat{\gamma}) = \prod_{j=1}^{n} s_j^{\downarrow}(\hat{\gamma}). \tag{50}$$

In particular, this proves equation (49).

We can then complete the proof: let $\tilde{\gamma} \geqslant \hat{\gamma} \geqslant \mathrm{i}\sigma$ for some $\mathcal{M}(\tilde{\gamma}) = \gamma$ in equation (47). We have just seen that this implies $\gamma \geqslant \mathcal{M}(\hat{\gamma}) \geqslant \mathrm{i}\sigma$ via equation (48) and furthermore that $F(\hat{\gamma}^{1/2}) \geqslant F(\mathcal{M}(\hat{\gamma})^{1/2})$ via equation (49). But this means that we have found $\overline{\gamma} \coloneqq \mathcal{M}(\hat{\gamma})$ such that $\gamma \geqslant \overline{\gamma} \geqslant \mathrm{i}\sigma$. Hence $\overline{\gamma}$ is in the feasible set of the right-hand side of (47) and $F(\tilde{\gamma}^{1/2}) \geqslant F(\overline{\gamma}^{1/2})$, which implies $\geqslant$ in equation (47). □

Finally, we can prove theorem 5.1 by also covering convex combinations:

Proof. Let $\gamma \in \mathbb{R}^{2n\times 2n}$ be a covariance matrix. First consider only sequences $\vec{\gamma}$: we replace any sequence by the special type of sequence of lemma 5.3. For these sequences, we have seen that the minimum cost is given by $G(\gamma)$ in proposition 5.5.

However, we explicitly excluded convex combinations (O6) by considering only sequences and not trees $\hat{\gamma}$: consider a tree of operations (O1)–(O7) which has γ at its root and $\gamma_0 = \mathbb{1}$ as leaves. Let us consider any node closest to the leaves. At such a node, we start with two covariance matrices γ_1 and γ_2 that were previously constructed without using convex combinations and with costs $G(\gamma_1)$ and $G(\gamma_2)$. The combined matrix would be $\tilde{\gamma} \coloneqq \lambda\gamma_1 + (1-\lambda)\gamma_2$ for some $\lambda \in (0, 1)$ and the costs would be $\lambda G(\gamma_1) + (1-\lambda)G(\gamma_2)$.

By convexity of G (see theorem 4.3):

$$G(\lambda\gamma_1 + (1-\lambda)\gamma_2) \leqslant \lambda G(\gamma_1) + (1-\lambda)G(\gamma_2)$$

which means that we can find a sequence (without any convex combinations) producing $\lambda\gamma_1 + (1-\lambda)\gamma_2$ which is cheaper than first producing γ_1 and γ_2 and then taking a convex combination. Iteratively, this means we can eliminate every node from the tree and replace the tree by a sequence of operations (O1)–(O5) and (O7), which is cheaper than the tree and trees do not matter. □

5.3. The squeezing measure as a resource measure

We have now seen that the measure G can be interpreted as a measure of the amount of single-mode squeezing needed to create a state ρ. Let us now take a different perspective, which is the analogue of the entanglement of formation for squeezing: consider covariance matrices of the form

$$\gamma_s := \begin{pmatrix} s & 0 \\ 0 & s^{-1} \end{pmatrix}. \tag{51}$$

These are single-mode squeezed states with squeezing parameter $s \geqslant 1$. We will now allow these states as *resources* and ask the question: given a (Gaussian) state ρ with covariance matrix γ, what is the minimal amount of these resources needed to construct γ, if we can freely transform the state by the same operations as before excluding squeezing ((O1)–(O7) excluding (O4)).

The corresponding measure is once again G:

Theorem 5.6. *Let ρ be an n-mode state with covariance matrix $\gamma \in \mathbb{R}^{2n\times 2n}$. Then*

$$G(\gamma) = \inf\left\{\sum_{i=1}^{m}\frac{1}{2}\log(s_m) \middle| \gamma = \mathcal{T}\left(\bigoplus_{i=1}^{m}\gamma_{s_i}\right)\right\}, \tag{52}$$

where $\mathcal{T}\colon \mathbb{R}^{2m\times 2m} \to \mathbb{R}^{2n\times 2n}$ is a combination of the operations (1)–(6) above.

Proof. $\leqslant$: Note that for any feasible $S \in Sp(2n)$ in $G(\gamma)$, i.e. any S with $S^TS \leqslant \gamma$, we can find $O \in Sp(2n)\cap O(2n)$ and $D = \bigoplus_{i=1}^{n}\gamma_{s_i}$ with $S^TS = O^TDO$ via the Euler decomposition. Using that the Euler decomposition minimises F, we have $F(S) = \frac{1}{2}F(D) = \sum_{i=1}^{n}\frac{1}{2}\log(s_i)$. But then, since we can find $\gamma_{\text{noise}} \geqslant 0$ such that $\gamma = O^T\bigoplus_{i=1}^{n}\gamma_{s_i}O + \gamma_{\text{noise}}$, we have that D is a feasible resource state to produce γ. This implies $G^{\text{resource}}(\gamma) \leqslant G(\gamma)$.

$\geqslant$: For the other direction, the proof proceeds exactly as the proof of theorem 5.1. First, we exclude convex combinations. Then, we realise that we can change the order of the different operations (even if we include adding resource states during any stage of the preparation process) according to lemma 5.2, making sure that any preparation procedure can be implemented via:

$$\gamma = \mathcal{M}\left(O\left(\bigoplus_{i=1}^{m}\gamma_{s_i} \oplus \mathbb{1}_{2m'} + \gamma_{\text{noise}}\right)O^T\right),$$

where $O \in Sp(2m+2m')\cap O(2m+2m')$, $\gamma_{\text{noise}} \in \mathbb{R}^{2m+2m'\times 2m+2m'}$ with $\gamma_{\text{noise}} \geqslant 0$ and $\mathcal{M}$ a measurement. Now the only difference to proof of 5.1 is that we had the vacuum $\mathbb{1}$ instead of $\bigoplus_{i=1}^{m}\gamma_{s_i} \oplus \mathbb{1}_{2m'}$ and an arbitrary symplectic matrix S instead of O, but the two ways of writing the maps are completely interchangeable, so that the proof proceeds as in theorem 5.1. □

We could call this measure the '(Gaussian) squeezing of formation', as it is the analogue to the Gaussian entanglement of formation. Note also that the measure is similar to the Gaussian entanglement of formation as defined in [Wol+04]. One natural further question would be whether 'distillation of squeezing' is possible with Gaussian operations. It is impossible in some sense for the minimal eigenvalue via [Kra+03], while it is possible and has been investigated for non-Gaussian states in many papers (see [Fil13, Hee+06] and references therein). In our case, it is not immediately clear whether extraction of single-mode squeezed states with less squeezing is possible or not. This could be investigated in future work.

6. Calculating the squeezing measure

We have seen that the measure G is operational. However, to be useful, we need a way to compute it.

6.1. Analytical solutions

Proposition 6.1. *Let $n = 1$, then $G(\Gamma) = -\frac{1}{2}\min_i \log(\lambda_i(\Gamma))$ for all $\Gamma \in \mathbb{R}^{2n\times 2n}$.*

Proof. Note that this is the lower bound in proposition 4.9, hence $-\frac{1}{2}\min_i \log(\lambda_i(\Gamma)) \leqslant G(\Gamma)$. Now consider the diagonalisation $\Gamma = O\,\mathrm{diag}(\lambda_1, \lambda_2)O^T$ with $O \in SO(2)$ and assume $\lambda_1 \geqslant \lambda_2$. Then, $\lambda_2^{-1} \leqslant \lambda_1$ since otherwise, $\Gamma \not\geqslant \mathrm{i}J$.

Consider $\mathrm{diag}(\lambda_1, \lambda_2) \geqslant O^{-T}S^TSO^{-1}$ for some $S \in Sp(2)$ with eigenvalues $s \geqslant 1$ and s^{-1}. Since $\mathrm{diag}(\lambda_1, \lambda_2) \geqslant O^{-T}S^TSO^{-1}$, this implies in particular that $s^{-1} \leqslant \lambda_2$ by Weyl's inequality. Since $F(S^TS) = \log s$, in order to minimise $F(S)$ over $S^TS \leqslant \Gamma$, we need to maximize s^{-1}. Setting $s^{-1} = \lambda_2$ we obtain $s = \lambda_2^{-1} \leqslant \lambda_1$ and $\mathrm{diag}(\lambda_1, \lambda_2) \geqslant \mathrm{diag}(s, s^{-1})$. Since $SO(2) = K(1)$, $S^TS \coloneqq O^T\,\mathrm{diag}(\lambda_1, \lambda_2)O \leqslant \Gamma$ is the minimising matrix in G and $G(\Gamma) = F(S) = \frac{1}{2}\log\lambda_2^{-1}$. □

Proposition 6.2. *Let ρ be a pure, Gaussian state with covariance matrix $\Gamma \in \mathbb{R}^{2n\times 2n}$. Then $G(\Gamma) = F(\Gamma^{1/2})$.*

Proof. From proposition 2.2, we know that $\det(\Gamma) = 1$ in particular. Therefore, the bounds in proposition 4.11 are tight and $G(\Gamma) = F(\Gamma^{1/2})$. □

6.2. Numerical calculations using Matlab

The crucial observation to numerically find the optimal squeezing measure is given in lemma 4.4: if we use G in the form of equation (25), we know that the function to be minimised is convex on $\mathcal{H}$. In general, convex optimisation with convex constraints is efficiently implementable and there is a huge literature on the topic (see [BV04] for an overview).

In our case, a certain number of problems occur when performing convex optimisation:

(1) The function f in equation (28) is highly nonlinear. It is also not differentiable at eigenvalue crossings of $A + \mathrm{i}B$ or $H \in \mathcal{H}$. In particular, it is not differentiable when one of the eigenvalues becomes zero, which is to be expected at the minimum.

(2) While the constraints $\mathcal{C}^{-1}(\gamma) \geqslant H$ and $\mathbb{1} > H > -\mathbb{1}$ are linear in matrices, they are nonlinear in simple parameterisations of matrices.
(3) For γ on the boundary of the set of allowed density operators, the set of feasible solutions might not have an inner point.

The first and second problem imply that most optimisation methods are unsuitable, as they are either gradient-based or need more problem structure. It also means that there is no guarantee for good stability of the solutions. The third problem implies that interior point methods become unsuitable on the boundary, which limits applications. For instance, our example of the next section (see equation (53)) lies on the boundary. As a proof of principle implementation, we used the MATLAB-based solver SOLVOPT (for details see the manual [KK97]). We believe our implementation could be made more efficient and more stable, but it seems to work well in most cases for less than ten modes. More information on the programme is provided in appendix E.

6.3. Squeezing-optimal preparation for certain three-mode separable states

Let us now work with a particular example that has been studied in the quantum information literature. In [MK08], Mišta Jr and Korolkova define the following three-parameter group of three-mode states where the modes are labelled A, B, C:

$$\gamma = \gamma_{AB} \oplus \mathbb{1}_C + x(q_1 q_1^T + q_2 q_2^T) \tag{53}$$

with

$$\begin{aligned} \gamma_{AB} &= \begin{pmatrix} e^{2d}a & 0 & -e^{2d}c & 0 \\ 0 & e^{-2d}a & 0 & e^{-2d}c \\ -e^{2d}c & 0 & e^{2d}a & 0 \\ 0 & e^{-2d}c & 0 & e^{-2d}a \end{pmatrix}, \\ q_1 &= (0,\ \sin\phi,\ 0,\ -\sin\phi,\ \sqrt{2},\ \sqrt{2})^T, \\ q_2 &= (\cos\phi,\ 0,\ \cos\phi,\ 0\sqrt{2},\ \sqrt{2})^T, \end{aligned}$$

where $a = \cosh(2r)$, $c = \sinh(2r)$, $\tan\phi = e^{-2r}\sinh(2d) + \sqrt{1 + e^{-4r}\sinh^2(2d)}$. The remaining parameters are $d \geqslant r > 0$ and $x \geqslant 0$. For

$$x = x_{\text{sep}} \geqslant \frac{2\sinh(2r)}{e^{2d}\sin^2\phi + e^{-2d}\cos^2\phi}$$

the state becomes fully separable [MK08]. The state as such is a special case of a bigger family described in [Gie+01]. In [MK08], it was used to entangle two systems at distant locations using fully separable mediating ancillas (here the system labelled C). Therefore, Mišta Jr and Korolkova considered also an LOCC procedure to prepare the state characterised by (53). For our purposes, this is less relevant and we allow for arbitrary preparations of the state. This was also done in [MK08] by first preparing modes A and B each in a pure squeezed-state with position quadratures $e^{2(d-r)}$ and $e^{2(d+r)}$. A vacuum mode in C was added and $x(q_1 q_1^T + q_2 q_2^T)$ was added as random noise. Therefore, the squeezing needed to produce this state in this protocol is given by

$$c = \frac{1}{2}\log(e^{2(d-r)} \cdot e^{2(d+r)}) = 2d. \tag{54}$$

We numerically approximated the squeezing measure for γ_{ABC}, choosing $x = x_{\text{sep}}$, which leaves a two-parameter family of states. We chose parameters d and r according to

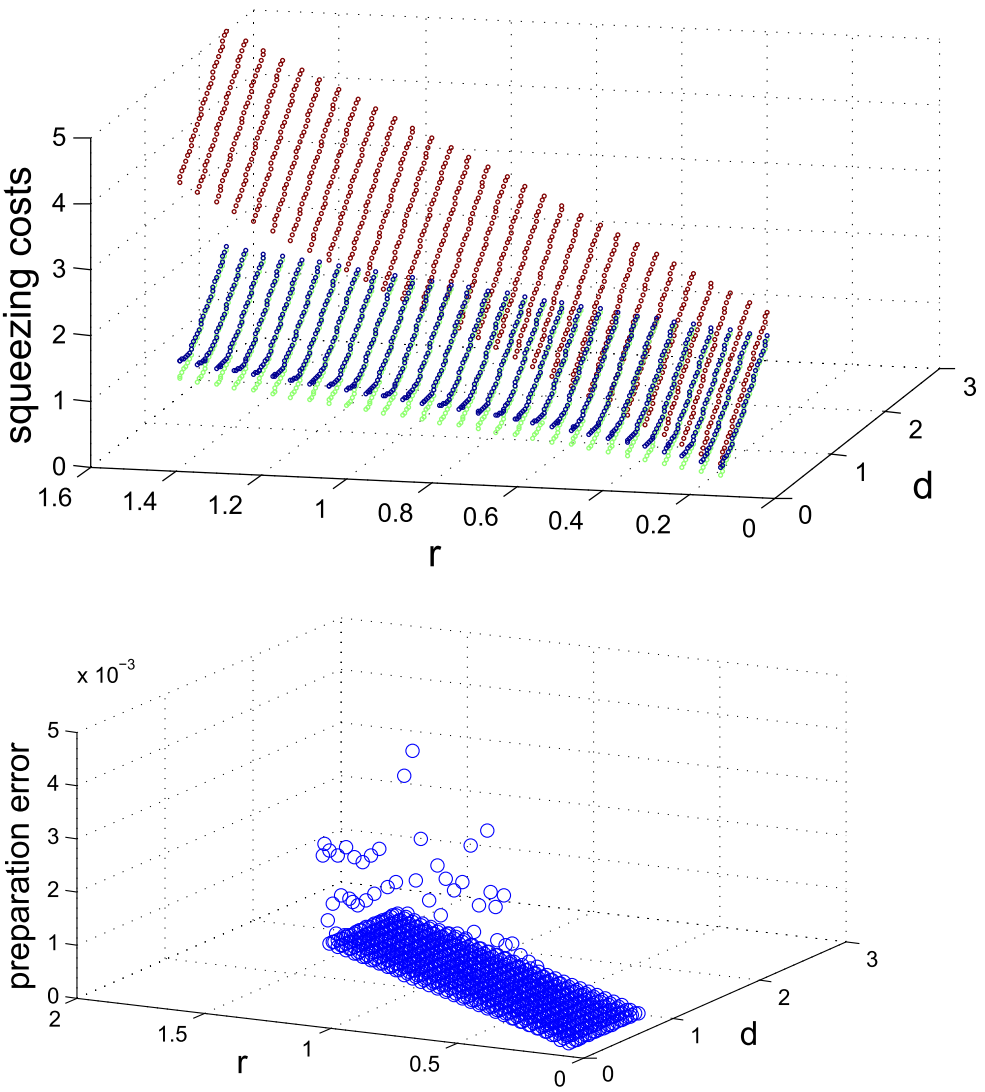

Figure 1. Results of numerical calculations (formulas for d and r in equation (55)). On the upper figure, the lower range of points (green in the online version) are the best lower bound, the middle points (blue) denote the value of the objective function at the minimum found by SOLVOPT and the upper points (red) denote the squeezing costs of the preparation protocol of [MK08] (equation (54)). The lower figure shows the preparation error. It is mostly below 10^{-6}.

$$r = 0.1 + j \cdot 0.05, \qquad d = r + i \cdot 0.03 \tag{55}$$

with $i, j \in \{1, \ldots, 30\}$ for a total of 900 data points. Since the algorithm is not an interior point algorithm as described above, to check the result, we reprepared the state in the following way:

(1) Let S be the symplectic matrix at the value optimum found by SOLVOPT for a covariance matrix γ_{ABC}.
(2) Calculate $S^{-T}\gamma_{ABC}S^{-1}$ and calculate its lowest eigenvalue λ_{2n}.

(3) Define $\tilde{\gamma} \coloneqq S^{-T}\gamma_{ABC}S^{-1} + (1 - \min\{1, \lambda_{2n}\})\mathbb{1} \geqslant \mathbb{1}$. Calculate the largest singular value of $S^T\tilde{\gamma}S - \gamma$.

If S was a feasible point, then $S^T\tilde{\gamma}S = \gamma$. Since it is obvious how to prepare $\tilde{\gamma}$ with operations specified in section 5, the largest singular value of $S^T\tilde{\gamma}S - \gamma$ is an indicator of how well we can approximate the state we want to prepare by a state with comparably low squeezing costs.

The results of the numerical computation are shown in figure 1. We computed the minimum both with the help of numerical and analytical subgradients and took the value with a better approximation error. At rare occasions, one algorithm failed to obtain a minimum. Possible reasons for this are discussed in appendix E. The optimal values computed by the algorithm are close to the lower bound and a lot better than the upper bound and the costs obtained by equation (54). One can easily see that γ_{ABC} cannot achieve the spectral lower bound as the assumptions of lemma 4.10 are not met.

7. Discussion of modifications to allowed operations

In experiments, squeezing of a state is most commonly measured by the logarithm of the smallest eigenvalue (up to a constant) and the unit is usually referred to as decibel (dB) [Lvo15]. We know of no operational interpretation for this measure that is similar to the interpretation given in section 5 and the measure is not natural for multimode states.

In contrast, G is a natural measure for multimode states. However, squeezing is not just experimentally challenging, it gets much harder if we want to achieve a larger amount of single-mode squeezing. Currently, the highest amount of squeezing obtained in quantum optical systems seems to be about 13 dB (see [And+15]). In other words, the two states ρ and ρ' with covariance matrices

$$\gamma = \operatorname{diag}(s, s^{-1}, s, s^{-1}), \qquad \gamma' = \operatorname{diag}(s^2, s^{-2}, 1, 1) \tag{56}$$

will not be equally hard to prepare although $G(\gamma) = G(\gamma')$. This is due to the fact that we quantified the cost of a single-mode squeezer by $\log s$.

To amend this, one could propose an easy modification to the definition of F in equation (11):

$$F_g(\gamma) = \sum_{i=1}^{n} \log(g(s_i^{\downarrow}(S))) \tag{57}$$

by inserting another function $g : \mathbb{R} \to \mathbb{R}$ to make sure that for the corresponding measure $G_g(\rho) \equiv G_g(\gamma)$, we have $G_g(\gamma) \neq G_g(\gamma')$ in equation (56). We pose the following natural restrictions on g:

- We need $g(1) = 1$ since $G_g(\rho)$ should be zero for unsqueezed states.
- Squeezing should get harder with larger parameter, hence g should be monotonously increasing.
- For simplicity, we assume g to be differentiable.

Let us first consider squeezing operations and the measure F_g. We proved in proposition 3.3 and theorem 3.5 that F is minimised by the Euler decomposition. A crucial part was given by lemma 3.4. In order to be useful for applications, we must require the same to be true for F_g, i.e.

$$\sum_{i=1}^{n} \log(g(s_i^{\downarrow}(SS'))) \leqslant \sum_{i=1}^{n} [\log(g(s_i^{\downarrow}(S))) + \log(g(s_i^{\downarrow}(S')))].$$

This puts quite strong restraints on g: considering $n = 1$ and assuming that S and S' are diagonal with ordered singular values, this implies that g must fulfill $g(xy) \leqslant g(x)g(y)$ for $x, y \geqslant 1$. This submultiplicativity restraint rules out all interesting classes of functions: Assume for instance that $g(2) = c$, then $g(2^n) \leqslant c^n$, where equality is attained if $g(x) = c \cdot x$. Therefore, all submultiplicative functions $g(x)$ for $x \geqslant 1$ must lie below $g(x) = c \cdot x$ at least periodically. Hence, lemma 3.4 does not hold if we consider increasingly growing functions g. This implies that one could make the measure arbitrarily small by splitting the single-mode squeezer into many successive single-mode squeezers with smaller squeezing parameter, which does not reflect experimental reality.

A way to circumvent the failure of lemma 3.4 would be to work with the 'squeezing of formation' measure. Likewise, one could require that there was only one operation of type (O4) as specified in section 5 in any preparation procedure. In that case we have:

Proposition 7.1. *If* $g : \mathbb{R} \to \mathbb{R}$ *fulfils*

(1) $\log \circ g \circ \mathcal{C}$ *is convex on* $(1, \infty)$,
(2) $\log(g(\exp(t)))$ *is convex and monotone increasing in* t,

then the squeezing of formation measure G_g *is still operational, i.e. theorem 5.1 still holds.*

Proof. The first condition replaces the log-convexity of the Cayley transform in the proof of theorem 4.3, making the measure convex. Using [Bha96], II.3.5 (v), the second condition makes sure that equation (50) still holds. The second condition can probably be relaxed while the proof of theorem 5.1 is still applicable. A function g fulfilling these prerequisites is $g(x) = \exp(x)$, which would correspond to a squeezing cost increasing linearly in the squeezing parameter. One could even introduce a cutoff after which g would be infinite. □

A simpler way to reflect the problems of equation (56) would be to consider the measures G and G_{minEig} together (calculating G_{minEig} of both the state and the minimal preparation procedure in G).

Another problem is associated with the form of the Hamiltonian (1). In the lab, the Hamiltonians that can be implemented might not be single-mode squeezers, but other operations such as symmetric two-mode squeezers (e.g. [SZ97], chapter 2.8). It is clear how to define a measure G' for these kinds of squeezers. Using the Euler decomposition, G is a lower bound to G', but we did not investigate this any further.

Acknowledgments

MI thanks Konstantin Pieper for discussions about convex optimisation and Alexander Müller-Hermes for discussions about MATLAB. MI is supported by the Studienstiftung des deutschen Volkes.

Appendix A. Preliminaries for the proof of theorem 3.5

Let us collect facts about ordinary differential equations needed in the proof:

Proposition A.1. *Consider the following system of differential equations for* $x : [0, 1] \to \mathbb{R}^{2n}$:

$$\begin{aligned} \dot{x}(t)^T &= x(t)^T A(t) \qquad \forall\, t \in [0, 1], \\ x(s) &= x_s \qquad \text{for some } x_s \in \mathbb{R}^{2n},\ s \in [0, 1], \end{aligned} \tag{58}$$

where $A \in L^\infty([0, 1], \mathfrak{sp}(2n))$. *Then this system has a unique solution, which is linear in* x_s *and defined on all of* $[0, 1]$ *such that we can define a map*

$$\forall\, s, t \in [0, 1]: \quad (s, t) \mapsto U(s, t) \in \mathcal{B}(\mathbb{R}^{2n})$$

via $x(t)^T = x_s^T U(s, t)$ *called the* propagator *of (58) that fulfils:*

(1) U *is continuous and differentiable almost everywhere.*
(2) $U(s, \cdot)$ *is absolutely continuous in* t.
(3) $U(t, t) = \mathbb{1}$ *and* $U(s, r)U(r, t) = U(s, t)$ *for all* $s, t \in [0, 1]$.
(4) $U(s, t)^{-1} = U(t, s)$ *for all* $s, t \in [0, 1]$.
(5) U *is the unique generalised (i.e. almost everywhere) solution to the initial value problem*

$$\begin{aligned} \partial_t U(s, t) - U(s, t)A(t) &= 0 \\ U(s, s) &= \mathbb{1} \end{aligned} \tag{59}$$

on $C([0, 1]^2, \mathbb{R}^{2n\times 2n})$.
(6) If $A(t) = A$ *does not depend on* t, *then* $S(r) = \exp(rA)$ *solves equation (59) with* $U(s, t) := S(t - s)$.
(7) for all $s, t \in [0, 1]$:

$$\|U(s, t)\|_\infty \leqslant \exp\left(\int_s^t \|A(\tau)\|_1 \, \mathrm{d}\tau\right).$$

(8) $U(s, t) \in Sp(2n)$ *for all* $t, s \in [0, 1]$ *and* $\gamma(t) = U(0, t)$ *fulfills equation (9) with* $\gamma(0) = \mathbb{1}$.

Proof. The proof of this (except for the part about $U(s, t) \in Sp(2n)$) can be found in [Son98] (theorem 55 and lemma C.4.1) for the transposed differential equation $\dot{x}(t) = A(t)x(t)$.

For the last part, note that since $U(s, s) = \mathbb{1} \in Sp(2n)$, we have $U(s, s)^T J U(s, s) = J$. We can now calculate almost everywhere:

$$\partial_t(U(t, s)^T J U(t, s)) = -U(t, s)^T (A^T(t)J - JA(t))U(t, s) = 0$$

since $A(t) \in \mathfrak{sp}(2n)$ and therefore $A^T(t)J - JA(t) = 0$.

But this implies $U(t, s)^T J U(t, s) = J$, hence U is symplectic. Obviously, $U(0, t)$ solves equation (9). □

We will also need another well-known lemma from functional analysis:

Lemma A.2. *Let* $A : [0, 1] \to \mathfrak{sp}(2n)$, $A \in L^\infty([0, 1], \mathbb{R}^{2n\times 2n})$. *Then* A *can be approximated in* $\|\cdot\|_1$*-norm by step-functions, which we can assume to map to* $\mathfrak{sp}(2n)$ *without loss of generality.*

The approximation by step-function can be found e.g. in [Rud87] (chapter 2, exercise 24).

Appendix B. The Cayley trick for matrices

In this appendix, we give an introduction to the Cayley-transform. The definition and properties needed in the main text are summarised by the following proposition:

Proposition B.1. *Define the Cayley transform and its inverse via:*

$$\begin{aligned} &\mathcal{C}: \{H \in \mathbb{R}^{n\times n} | \mathrm{spec}(H) \cap \{+1\} = \varnothing\} \to \mathbb{R}^{n\times n} \\ &\quad H \mapsto \frac{\mathbb{1}+H}{\mathbb{1}-H}, \end{aligned} \tag{60}$$

$$\begin{aligned} &\mathcal{C}^{-1}\{S \in \mathbb{R}^{n\times n} | \mathrm{spec}(H) \cap \{-1\} = \varnothing\} \to \mathbb{R}^{n\times n} \\ &\quad S \mapsto \frac{S-\mathbb{1}}{S+\mathbb{1}} \end{aligned} \tag{61}$$

$\mathcal{C}$ is a diffeomorphism onto its image with inverse $\mathcal{C}^{-1}$. Furthermore, it has the following properties:

(1) $\mathcal{C}$ is operator monotone and operator convex on matrices A with $\mathrm{spec}(A) \subset (-1, 1)$.
(2) $\mathcal{C}^{-1}$ is operator monotone and operator concave on matrices A with $\mathrm{spec}(A) \subset (-1, \infty)$.
(3) $\mathcal{C}: \mathbb{R} \to \mathbb{R}$ with $\mathcal{C}(x) = (1+x)/(1-x)$ is log*-convex on* $[0, 1)$.
(4) For $n = 2m$ even, $H \in \mathbb{R}^{2m\times 2m}$ and $H \in \mathcal{H}$ if and only if $\mathcal{C}(H) \in Sp(2m, \mathbb{R})$ and $\mathcal{C}(H) \geqslant \mathrm{i}J$.

where $\mathcal{H}$ is defined via:

$$\mathcal{H} = \left\{ H = \begin{pmatrix} A & B \\ B & -A \end{pmatrix} \in \mathbb{R}^{2m\times 2m} \;\middle|\; A^T = A,\, B^T = B,\, \mathrm{spec}(H) \subset (-1, 1),. \right\}$$

The definition and the fact that this maps the upper half plane of positive definite matrices to matrices inside the unit circle is present in [AG88] (I.4.2) and [MS98] (proposition 2.51, proof 2). Since no proof is given in the references and they do not cover the whole proposition, we provide them here.

We start with well-definedness:

Lemma B.2. *$\mathcal{C}$ and $\mathcal{C}^{-1}$ are well-defined and inverses of each other. Moreover, $\mathcal{C}$ is a diffeomorphism onto its image* $\mathrm{dom}(\mathcal{C}^{-1})$.

Proof. If $\mathrm{spec}(H) \cap \{+1\} = \varnothing$, then $\mathbb{1} - H$ is invertible and $H \mapsto (\mathbb{1}+H)/(\mathbb{1}-H)$ is well-defined, as $[\mathbb{1}+H, \mathbb{1}-H] = 0$. Now let $H \in \mathbb{R}^{m\times m}$ be such that $\mathrm{spec}(H) \cap \{+1\} = \varnothing$. We will show that $\mathcal{C}(H)$ contains no eigenvalue -1. To see this, let

$$H = T \bigoplus_i J(n_i, \lambda_i) T^{-1} \tag{62}$$

be the Jordan normal form with block sizes n_i and eigenvalues λ_i. Let us here consider the complex Jordan decomposition, i.e. λ_i are allowed to be complex. Then:

$$\mathbb{1} + H = T \bigoplus_i J(n_i, 1+\lambda_i) T^{-1}, \quad \mathbb{1} - H = T \bigoplus_i J(n_i, 1-\lambda_i) T^{-1} \tag{63}$$

and thus

$$\mathcal{C}(H) = T \bigoplus_i J(n_i, 1 + \lambda_i) \cdot J(n_i, 1 - \lambda_i)^{-1} T^{-1}.$$

For the inverse of the Jordan blocks, we can use the well-known formula:

$$\begin{pmatrix} 1-\lambda_i & 1 & \dots & 0 \\ 0 & 1-\lambda_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1-\lambda_i \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{1-\lambda_i} & \frac{-1}{(1-\lambda_i)^2} & \cdots & \frac{(-1)^{n_i-1}}{(1-\lambda_i)^{n_i}} \\ 0 & \frac{1}{1-\lambda_i} & \dots & \frac{(-1)^{n_i-2}}{(1-\lambda_i)^{n_i-1}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{1-\lambda_i} \end{pmatrix}.$$

In particular, this is still upper triangular. Then $J(n_i, 1 + \lambda_i)J(n_i, 1 - \lambda_i)^{-1}$ is still upper triangular with diagonal entries $(1 + \lambda_i)/(1 - \lambda_i)$. Since $(1 + \lambda_i)/(1 - \lambda_i) \neq -1$ for all $\lambda_i \in \mathbb{C}$, we find that $J(n_i, 1 + \lambda_i)J(n_i, 1 - \lambda_i)^{-1}$ cannot have eigenvalue -1 for any i, hence $\mathrm{spec}(\mathcal{C}(H)) \cap \{-1\} \neq \varnothing$.

Finally, we observe:

$$\mathcal{C}^{-1}\mathcal{C}(H) = \frac{\frac{\mathbb{1}+H}{\mathbb{1}-H} - \mathbb{1}}{\frac{\mathbb{1}+H}{\mathbb{1}-H} + \mathbb{1}} = \frac{\mathbb{1} + H - \mathbb{1} + H}{\mathbb{1} + H + \mathbb{1} - H} = H.$$

Moreover, set $f_1(A) = -2A - \mathbb{1}$ for all matrices $A \in \mathbb{R}^{m\times m}$, $f_2(A) = A^{-1}$ for all invertible matrices $A \in \mathbb{R}^{m\times m}$ and $f_3(A) = A - \mathbb{1}$ for all matrices $A \in \mathbb{R}^{m\times m}$. Then we have

$$f_1 \circ f_2 \circ f_3(H) = f_1 \circ f_2(H - \mathbb{1}) = f_1\left(\frac{1}{H - \mathbb{1}}\right) = -\frac{2}{H - \mathbb{1}} - \mathbb{1} = \mathcal{C}(H). \tag{64}$$

Since f_i are differentiable for all $i = 1, 2, 3$, we have that $\mathcal{C}$ is invertible.

The same considerations with a few signs reversed also lead us to conclude that $\mathcal{C}^{-1}$ is well-defined and indeed the inverse of $\mathcal{C}$. We can similarly decompose $\mathcal{C}^{-1}$ to show that it is differentiable, making $\mathcal{C}$ a diffeomorphism. Here, we define $g_1(A) = 2A + \mathbb{1}$ for all $A \in \mathbb{R}^{m\times m}$, $g_2(A) = A^{-1}$ for all invertible $A \in \mathbb{R}^{m\times m}$ and $g_3(A) = A + \mathbb{1}$ for all $A \in \mathbb{R}^{m\times m}$. A quick calculation shows

$$g_1 \circ g_2 \circ g_3(S) = \mathcal{C}^{-1}(S). \tag{65}$$

□

Denote by $\mathcal{H}$ the set

$$\mathcal{H} := \left\{ H = \begin{pmatrix} A & B \\ B & -A \end{pmatrix} \middle| A \in \mathbb{R}^{2n\times 2n} A^T = A,\, B^T = B,\, -\mathbb{1} < H < \mathbb{1} \right\} \tag{66}$$

where $H < \mathbb{1}$ means that $\mathbb{1} - H$ is positive definite (not just positive semidefinite). We can then prove the Cayley trick:

Proposition B.3. *Let $H \in \mathbb{R}^{2n\times 2n}$. Then $H \in \mathcal{H} \Leftrightarrow (\mathcal{C}(H) \in Sp(2n) \wedge \mathcal{C}(H) \geqslant \mathrm{i}J)$.*

Proof. Note that for $H \in \mathcal{H}$, $1 \notin \mathrm{spec}(H)$, hence $\mathcal{C}(H)$ is always well-defined. $\mathcal{C}(H) = (\mathbb{1} + H)(\mathbb{1} - H)^{-1} \geqslant 0$, since $\mathbb{1} + H \geqslant 0$ and $(\mathbb{1} - H)^{-1} \geqslant 0$ as $-\mathbb{1} < H < \mathbb{1}$. Observe:

$$HJ = \begin{pmatrix} A & B \\ B & -A \end{pmatrix}\begin{pmatrix} 0 & \mathbb{1} \\ -\mathbb{1} & 0 \end{pmatrix} = \begin{pmatrix} -B & A \\ A & B \end{pmatrix} = -\begin{pmatrix} 0 & \mathbb{1} \\ -\mathbb{1} & 0 \end{pmatrix}\begin{pmatrix} A & B \\ B & -A \end{pmatrix} = -JH.$$

Then we can calculate:

$$\begin{aligned}(\mathbb{1}+H)\cdot(\mathbb{1}-H)^{-1}J &= -(\mathbb{1}+H)\cdot(J(\mathbb{1}-H))^{-1} = -(\mathbb{1}+H)\cdot((\mathbb{1}+H)J)^{-1} \\ &= (\mathbb{1}+H)J(\mathbb{1}+H)^{-1} = J(\mathbb{1}-H)\cdot(\mathbb{1}+H)^{-1},\end{aligned}$$

hence $\mathcal{C}(H)J = J\mathcal{C}(H)^{-1}$ and as $\mathcal{C}(H)$ is Hermitian, we have $\mathcal{C}(H)^T J\mathcal{C}(H) = J$ and $\mathcal{C}(H)$ is symplectic. Via corollary 2.10, as $\mathcal{C}(H)$ is symplectic and positive definite, we can conclude that $\mathcal{C}(H) \geqslant \mathrm{i}J$.

Conversely, let $S \in Sp(2n)$ and $S \geqslant \mathrm{i}J$. Then $S \geqslant -\mathrm{i}J$ by complex conjugation and $S \geqslant 0$ after averaging the two inequalities. Since any element of $Sp(2n)$ is invertible, this implies $S > 0$. From this we obtain:

$$\begin{aligned}\frac{S-\mathbb{1}}{S+\mathbb{1}} &> -\mathbb{1} \qquad \text{as } S+\mathbb{1} > \mathbb{1}, \\ \frac{S-\mathbb{1}}{S+\mathbb{1}} &< \mathbb{1} \qquad \text{always.}\end{aligned}$$

Write $(S-\mathbb{1})\cdot(S+\mathbb{1})^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$. As S is Hermitian, $A^T = A$ and $C = B^T$, $D^T = D$.

We have on the one hand

$$\begin{aligned}\frac{S-\mathbb{1}}{S+\mathbb{1}}J &= (S-\mathbb{1})\cdot(-S^{-T}J - J)^{-1} = (S-\mathbb{1})(-J)^{-1}(S^{-T}+\mathbb{1})^{-1} \\ &= (SJ-J)\cdot(S^{-T}+\mathbb{1})^{-1} = J(S^{-T}-\mathbb{1})S^T S^{-T}(S^{-T}+\mathbb{1})^{-1} \\ &= -J\frac{S-\mathbb{1}}{S+\mathbb{1}}\end{aligned}$$

and on the other hand

$$\begin{aligned}\begin{pmatrix} A & B \\ B^T & D \end{pmatrix}J &= \begin{pmatrix} -B & A \\ -D & B^T \end{pmatrix}, \\ -J\begin{pmatrix} A & B \\ B^T & D \end{pmatrix} &= \begin{pmatrix} -B^T & -D \\ A & B \end{pmatrix}.\end{aligned}$$

Put together this implies $B = B^T$ and $D = -A$, hence $\mathcal{C}^{-1}(S) \in \mathcal{H}$, which is what we claimed. □

Proposition B.4. *The Cayley transform $\mathcal{C}$ is operator monotone and operator convex on the set of $A = A^T \in \mathbb{R}^{m\times m}$ with $\mathrm{spec}(A) \subset (-1, 1)$. $\mathcal{C}^{-1}$ is operator monotone and operator concave on the set of $A = A^T \in \mathbb{R}^{m\times m}$ with $\mathrm{spec}(A) \subset (-1, \infty)$.*

Proof. Recall equation (64) and the definition of f_1, f_2, f_3. f_1 and f_3 are affine and thus for all $X \geqslant Y$: $f_3(X) \geqslant f_3(Y)$ and $f_1(X) \leqslant f_1(Y)$. For $X \geqslant Y \geqslant 0$, we also have $f_2(Y) \geqslant f_2(X) \geqslant 0$ since matrix inversion is antimonotone. Now let $-\mathbb{1} \leqslant Y \leqslant X \leqslant 1$, then $-2\mathbb{1} \leqslant f_3(Y) \leqslant f_3(X) \leqslant 0$ and $-1/2\mathbb{1} \geqslant f_2 \circ f_3(Y) \geqslant f_2 \circ f_3(X) \geqslant 0$ and finally $\mathcal{C}(X) \geqslant \mathcal{C}(Y) \geqslant 0$, proving monotonicity of $\mathcal{C}$. Similarly, one can prove that $\mathcal{C}^{-1}$ is monotonous using equation (65).

For the convexity of $\mathcal{C}$, we note that since f_1, f_3 are affine they are both convex and concave. It is well-known that $1/x$ is operator convex for positive definite and operator concave for negative definite matrices (to prove this, consider convexity/concavity of the

functions $\langle \psi, X^{-1}\psi \rangle$ for all ψ). It follows that for $-\mathbb{1} \leqslant H \leqslant \mathbb{1}$ we have $f_3(x) \leqslant 0$, hence $f_2 \circ f_3$ is operator concave on $-\mathbb{1} \leqslant H \leqslant \mathbb{1}$. As $f_1(A) = -2A - \mathbb{1}$, this implies that $\mathcal{C} = f_1 \circ f_2 \circ f_3$ is operator convex.

For the concavity of $\mathcal{C}^{-1}$, recall equation (65) and the definitions of g_1, g_2, g_3. Then, given $-\mathbb{1} \leqslant X$, we have $g_3(X)$ is positive definite and concave as an affine map. g_2 is concave on positive definite matrices, as $1/x$ is convex and (-1) is order-reversing, hence $-1/x$ is concave on positive definite matrices. Since g_1 is concave as an affine map, $g_1 \circ g_2 \circ g_3 = \mathcal{C}^{-1}$ is operator concave for all $-\mathbb{1} \leqslant X$. □

Lemma B.5. $\mathcal{C} : \mathbb{R} \to \mathbb{R}$ *is* log-*convex on* $[0, 1)$.

Proof. We need to see that the function $h(x) = \log \frac{1+x}{1-x}$ is convex for $x \in [0, 1)$. Since h is differentiable on $[0, 1)$, this is true iff the second derivative is non-negative:

$$h''(x) = \frac{4x}{(1 - x^2)^2}$$

is clearly positive on $[0, 1)$ and h is therefore log-convex. □

Appendix C. Continuity of set-valued functions

Here, we provide some definitions and lemmata from set-valued analysis for the reader's convenience. This branch of mathematics deals with functions $f : X \to 2^Y$ where X and Y are topological spaces and 2^Y denotes the power set of Y.

In order to state the results interesting to us we define:

Definition C.1. Let $X, Y \subseteq \mathbb{R}^{n\times m}$ and $f : X \to 2^Y$ be a set-valued function. Then we say that a function is *upper semicontinuous* (often also called *upper hemicontinuous* to distinguish it from other notions of continuity) at $x_0 \in X$ if for all open neighbourhoods Q of $f(x_0)$ there exists an open neighbourhood W of x_0 such that $W \subseteq \{x \in X | f(x) \subset Q\}$. Likewise, we call it *lower semicontinuous* (often called *lower hemicontinuous*) at a point x_0 if for any open set V intersecting $f(x_0)$, we can find a neighbourhood U of x_0 such that $f(x) \cap V \neq \varnothing$ for all $x \in U$.

Note that the definitions are valid in all topological spaces, but we only need the case of finite dimensional normed vector spaces. Using the metric, we can give the following characterisation of upper semicontinuity:

Lemma C.2. *Let* $X, Y \subseteq \mathbb{R}^{n\times m}$ *and* $f : X \to 2^Y$ *be a set-valued function such that* $f(x)$ *is compact for all* x. *Then* f *is upper semicontinuous at* x_0 *if and only if for all* $\varepsilon > 0$ *there exists a* $\delta > 0$ *such that for all* $x \in X$ *with* $\|x - x_0\| < \delta$ *we have: for all* $y \in f(x)$ *there exists a* $\tilde{y} \in f(x_0)$ *such that* $\|y - \tilde{y}\| < \varepsilon$.

Proof. $\Rightarrow$: Let f be lower semicontinuous at x_0. For any $\varepsilon > 0$ the set

$$B(\varepsilon, f(x_0)) \bigcup_{y \in f(x_0)} \{\hat{y} \in Y | \|y - \hat{y}\| < \varepsilon\} \tag{67}$$

is an open neighbourhood of $f(x_0)$. Hence there exists an open neighbourhood W of x_0, which contains a ball of radius $\delta > 0$ such that $B_\delta(x_0) \subseteq W \subseteq \{x \in X | f(x) \subset B(\varepsilon, f(x_0))\}$. Clearly this implies the statement.

$\Leftarrow$: Let Q be a neighbourhood of $f(x_0)$. Since $f(x_0)$ is compact this implies that there is a $\varepsilon > 0$ such that $B(\varepsilon, f(x_0)) \subseteq Q$ where this set is defined as in equation (67). If this were not the case, for every $n \in \mathbb{N}$ there must be a $y_n \in Y \backslash Q$ such that $\inf_{\hat{y} \in f(x_0)} \| y_n - \hat{y} \| < 1/n$. Since by construction this implies that $y_n \in B(1, f(x_0))$, which is compact, a subsequence of these y_n must converge to y. As $Y \backslash Q$ is closed as Q is open, $y \in Y \backslash Q$. However, $\inf_{\hat{y} \in f(x_0)} \| y - \hat{y} \| = 0$ by construction and since $f(x_0)$ is compact, the infimum is attained, which implies $y \in f(x_0)$. This contradicts the fact that Q is a neighbourhood of $f(x_0)$.

Hence we know that for any open Q containing $f(x_0)$ there exists a $\varepsilon > 0$ such that $B(\varepsilon, f(x_0)) \subseteq Q$. By assumption, this implies that there exists a $\delta > 0$ such that $B_\delta(x_0) \subseteq \{x \in X | f(x) \subset B(\varepsilon, f(x_0))\}$. Since clearly $\{x \in X | f(x) \subset B(\varepsilon, f(x_0))\} \subseteq \{x \in X | f(x) \subset Q\}$ we can choose $W := B_\delta(x_0)$ to finish the proof. □

This second characterisation is sometimes called *upper Hausdorff semicontinuity* and it can equally be defined in any metric space. Clearly, the notions can differ for set-valued functions with non-compact values or in spaces which are not finite dimensional. With these two definitions, we can state the following classic result:

Proposition C.3 ([DR79]). *Let Y be a complete metric space, X a topological space and $f : X \to 2^Y$ a compact-valued set-valued function. The following statements are equivalent:*

- *f is upper semicontinuous at x_0.*
- *for each closed $K \subseteq X$, $K \cap f(x_0)$ is upper semicontinuous at x_0.*

An interesting question would be whether the converse is also true. Even if $f(x)$ is always convex, this need not be the case if $K \cap f(x_0)$ has empty interior as simple counterexamples can show. In case the interior is non-empty, another classic results guarantees a converse in many cases:

Proposition C.4 ([Mor75]). *Let X be a compact interval and Y a normed space. Let $f : X \to 2^Y$ and $g : X \to 2^Y$ be two convex-valued set-valued functions. Suppose that $\operatorname{diam}(f(t) \cap g(t)) < \infty$ and $f(t) \cap \operatorname{int}(g(t)) \neq \varnothing$ for all t. Then if f, g are continuous (in the sense above) so is $f \cap g$.*

Appendix D. Reduction of the set of necessary operations for state preparation

In this section, we give a justification of why the operations (O0)–(O7) are enough to implement all operations described in section 5. All of this is known albeit scattered throughout the literature, hence we collect it here.

In order to prepare a state, we could start with the vacuum $\gamma = \mathbb{1}$ or alternatively a thermal state for some bath ($\gamma = (1/2 + N)\mathbb{1}$ with photon number N, see e.g. [Oli12]). Of course, we should be able to draw arbitrary ancillary modes of this system, too. The effect of Gaussian noise on the covariance matrix is given in [Lin00]. Since for any $\gamma \geqslant \mathbb{1}$ we can decompose it as $\gamma = \mathbb{1} + \gamma_{\text{noise}}$, this implies that the operations (O0)–(O2) are enough to implement all operations 1. and 2.

As with other squeezing measures, passive transformations should not change the squeezing measure, while single-mode-squeezers are not free. The effect of symplectic transformations on the covariance matrix has already been observed in equation (10), hence (O3) and (O4) implement operations (3) and (8).

Since we have the Weyl-system at our disposal, we can also consider its action on a quantum state (translation in phase space). Direct computation shows that it does not affect the covariance matrix. Including it as operation (O5) is beneficial if we consider a convex combination of states. In an experiment, this can be done by creating ensembles of the states of the convex combination and creating another ensemble where the ratio of the different states is that of the convex combination. On the level of covariance matrices, we have the following lemma:

Lemma D.1. *Let ρ and ρ' be two states with displacement d^ρ and $d^{\rho'}$ and (centred) covariance matrices γ^ρ and $\gamma^{\rho'}$. For $\lambda \in (0, 1)$, the covariance matrix of $\tilde{\rho} \coloneqq \lambda\rho + (1 - \lambda)\rho'$ is given by:*

$$\gamma^{\tilde{\rho}} = \lambda\gamma^\rho + (1 - \lambda)\gamma^{\rho'} + 2\lambda(1 - \lambda)(d^\rho - d^{\rho'})(d^\rho - d^{\rho'})^T$$

A proof of this statement can be found in [WW01] (in the proof of proposition 1). Note that for *centralised* states with $d^\rho = 0$ and $d^{\rho'} = 0$, a convex combination of states translates to a convex combination of covariance matrices. Since in particular, $2\lambda(1 - \lambda)(d^\rho - d^{\rho'})(d^\rho - d^{\rho'})^T \geqslant 0$, any convex combination of ρ and ρ' is on the level of covariance matrices equivalent to

- centring the states (no change in the covariance matrices),
- taking a convex combination of the states (resulting in a convex combination of covariance matrices),
- performing a Weyl translation to undo the centralization in the first step (no change in the covariance matrix).
- Adding noise $2\lambda(1 - \lambda)(d^\rho - d^{\rho'})(d^\rho - d^{\rho'})^T \geqslant 0$.

This implies that the effect of any convex combination of states (operation 4) on the covariance matrix can equivalently be obtained from operations (O2), (O5) and (O6). Finally, we consider measurements. *Homodyne detection* is the measurement of Q or P in one of the modes, which corresponds to the measurement of an infinitely squeezed pure state in lemma D.2. A broader class of measurements known as *heterodyne detection* measures arbitrary coherent states [Wee+12]. Let us focus our attention on the even broader class of projections onto Gaussian pure states.

Lemma D.2. *Let ρ be an $(n + 1)$-mode quantum state with covariance matrix γ and $|\gamma_G, d\rangle\langle\gamma_G, d|$ be a pure single-mode Gaussian state with covariance matrix $\gamma_G \in \mathbb{R}^{2\times 2}$ and displacement d. Let*

$$\gamma = \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}, \qquad B \in \mathbb{R}^{2\times 2}$$

then the selective measurement of $|\gamma_G, d\rangle$ in the last mode results in a change of the covariance matrix of ρ according to:

$$\gamma' = A - C(B - \gamma_G)^{\mathrm{MP}}C^T, \tag{68}$$

where $^{\mathrm{MP}}$ *denotes the Moore–Penrose pseudoinverse.* Homodyne detection *corresponds to the case where* γ_G *is an infinitely squeezed state.*

This can most easily be seen on the level of Wigner functions, as demonstrated in [ESP02, GIC02]. The generalisation to multiple modes is straightforward.

Since the covariance matrix of a Gaussian pure state is a symplectic matrix (see proposition 2.2), using the Euler decomposition we can implement a selective Gaussian measurement by

(1) a passive symplectic transformation $S \in K(n+1)$,
(2) a measurement in the Gaussian state $\mathrm{diag}(d, 1/d)$ for some $d \in \mathbb{R}_+$ according to lemma D.2.

A non-selective measurement (forgetting the information obtained from measurement) would then be a convex combination of such projected states. A measurement of a multi-mode state can be seen as successive measurements of single-mode states since the Gaussian states we measure are diagonal.

For homodyne detection, since an infinitely squeezed single-mode state is given by the covariance matrix $\lim_{d\to\infty}\mathrm{diag}(1/d, d)$, we have

$$\gamma' = \lim_{d\to\infty} (A - C(B - \mathrm{diag}(1/d, d))^{-1}C^T) = A - C(\pi B\pi)^{\mathrm{MP}}C^T, \qquad (69)$$

where $\pi = \mathrm{diag}(1, 0)$ is a projection and $^{\mathrm{MP}}$ denotes the Moore–Penrose-pseudoinverse. It has been shown (see [Wee+12] E.2 and E.3 as well as [ESP02, GIC02]) that any (partial or total) Gaussian measurement is a combination of passive transformations, discarding subsystems, projection onto Gaussian states and homodyne detection.

Therefore, we should also allow to discard part of the system, i.e. taking the partial trace. However, this can be expressed as a combination of operations (O1)–(O6) and homodyne detection:

Lemma D.3. *Given a covariance matrix* $\gamma = \begin{pmatrix} A & C \\ C^T & B \end{pmatrix}$ *a partial trace on the second system translates to a map* $\gamma \mapsto A$. *The partial trace can then be implemented by measurements and adding noise.*

Proof. When measuring the modes B, we note that since $C(\pi B\pi)^{\mathrm{MP}}C^T \geqslant 0$ in equation (69), a partial trace is equivalent to first performing a homodyne detection on the B-modes of the system and then adding noise. □

Given the discussion above, lemmas D.2 and D.3 put together imply: on the level of covariance matrices, in order to allow for general Gaussian measurements, it suffices to consider Gaussian measurements of the state $|\gamma_d, 0\rangle\langle\gamma_d, 0|$ with covariance matrix $\gamma_d = \mathrm{diag}(1/d, d)$ for $d \in \mathbb{R}_+ \cup \{+\infty\}$. All Gaussian measurements are then just combinations of these special measurements and operations (O1)–(O6).

Appendix E. Numerical implementation and documentation

Here, we provide a short documentation to the programme written in MATLAB, Version R2014a, and used for the numerical computations in section 6. The source Code can be found at GitHub https://github.com/Martin-Idel/operationalsqueezing.

The programme tries to minimise the function f defined in equation (28) over the set $\mathcal{H}$. Throughout, suppose we are given a covariance matrix γ.

Let us first describe the implementation of f: as parameterisation of $\mathcal{H}$, we choose the simplest parameterisation such that for matrices with symplectic eigenvalues larger than one, the set of feasible points has non-empty interior: we parameterise A, B via matrix units E_i, E_{jk} with $i \in \{1,\ldots,n\}$, $k \in \{1,\ldots,n-1\}$ and $j < k$, where $(E_i)_{jk} = \delta_{ij}\delta_{ik}$ and $(E_{jk})_{lm} = \delta_{jl}\delta_{km} + \delta jm\delta_{kl}$. This parameterisation might not be very robust, but it is good enough for our purpose. Instead of working with complex parameters, we compute $s_i(A + \mathrm{i}B)$ as $\lambda_i^{\downarrow}(H)$ for the matrix

$$H = \begin{pmatrix} A & B \\ B & -A \end{pmatrix}. \tag{70}$$

The evaluation of f is done in function OBJECTIVE.M. Since f is not convex for (A, B) with the corresponding H having eigenvalues $\geqslant 1$ or $\leqslant -1$, the function first checks, whether this constraint is satisfied and outputs a value that is 10^7-times larger than the value of the objective function at the starting point otherwise.

The constraints are implemented in function MAXRESIDUAL.M. Via symmetry, it is enough to check that for any H tested, $\lambda_{2n}^{\downarrow}(H) \geqslant 1$. The second constraint is given by $\mathcal{C}^{-1}(\gamma) \geqslant H$ and this is tested by computing the smallest eigenvalue of the difference.

The function which is most important for users is MINIMUM.M, which takes a covariance matrix $\Gamma \geqslant \mathrm{i}J$, its dimensions n and a number of options as arguments and outputs the minimum. Note that the programme checks whether the covariance matrix is valid. For the minimisation, we use the MATLAB-based solver SOLVOPT ([KK97], latest version 1.1). SOLVOPT uses a subgradient based method and the method of exact penalization to compute (local) minima. For convex programming, any minimum found by the solver is therefore an absolute minimum. In order to work, the objective function may not be differentiable on a set of measure zero and it is allowed to be non-differentiable at the minimum. Since f is differentiable for all H with non-degenerate eigenvalues, this condition is met. In addition, SOLVOPT needs f to be defined everywhere, as it is not an interior point method. Since f is well-defined but not convex for $H \notin \mathcal{H}$ and $\mathrm{spec}(H) \cup \{1\} = \varnothing$, we remedy this by changing the output of OBJECTIVE.M to be very large when $H \notin \mathcal{H}$ as described above. Constraints are handled via the method of exact penalisation. We used SOLVOPT's algorithm to compute the penalisation functions on its own.

It is possible (and for speed purposes advisable) to implement analytical gradients of both the objective and the constraint functions. Following [Mag85], for diagonalisable matrices A with no eigenvalue multiplicities, the derivative of an eigenvalue $\lambda_i(A)$ is given by:

$$\partial_E \lambda_i(A) = v_i(A)^T \partial_E A v_i(A), \tag{71}$$

where $v_i(A)$ is the eigenvector corresponding to $\lambda_i(A)$ and $\partial_v(A) = \lim_{h\to 0}(A + hE - A)/h = E$. Luckily, if A is not differentiable, this provides at least one subgradient. An easy calculation shows that a subgradient of the objective function f for matrices H with $-\mathbb{1} < H < \mathbb{1}$ in the parameterisation of the matrix units E_{ij} is given by

$$(\nabla f)_i = \sum_{j=1}^{n} \frac{\partial_i \lambda_j^{\downarrow}(H)}{(1 + \lambda_j(H))(1 - \lambda_j(H))^2} = \sum_{j,k=1}^{n} \frac{v_{j,k}^T F(i) v_{k,j}}{(1 + \lambda_j(H))(1 - \lambda_j(H))^2} \tag{72}$$

with F being the matrices corresponding to the chosen parameterisation. The gradient of the constraint function is very similar and given by equation (71) for $A = \gamma - H$ or $A = 2\mathbb{1} - H$ depending on which constraint is violated. This is implemented in functions OBJECTIVEGRAD.M and MAXRESIDUALGRAD.M.

SOLVOPT needs a starting point. Given Γ, via Williamson's theorem, $\Gamma = S^T D S \geqslant S^T S$, hence $S^T S$ provides a good starting point. The function WILLIAMSON.M computes the Williamson normal form for γ and returns S, D and $S^T S$, the latter of which is used as starting point. It computes S and D essentially by computing the Schur decomposition of $\Gamma^{-1/2} J \Gamma^{-1/2}$ (in the σ-basis instead of the J-basis). S is then given by $S^T = \gamma_{1/2} K D^{-1/2}$ (see the proof of [SCS99]), where K is the Schur transformation matrix.

A number of comments are in order:

(1) All functions use global variables instead of function handles. This is required by the fact that SOLVOPT has not been adapted to the use of function handles. The user should therefore always reset all variables before running the programme.
(2) SOLVOPT is not an interior point method, i.e. the results can at times violate constraints. We use the default value for the accuracy of constraints, which is 10^{-8} and can be modified by option six. The preparation error should be of the same order than the accuracy of constraints as long as the largest eigenvalue of the minimising symplectic matrix is of order one.
(3) For our numerical tests, we used bounds on the minimal step-size and the minimal error in f (SOLVOPT options two and three) of the order 10^{-6} and 10^{-8}, which seemed sufficient.
(4) All functions called by SOLVOPT (the functions OBJECTIVE.M, OBJECTIVEGRAD.M, MAXRESIDUAL,M, MAXRESIDUALGRAD.M and XTOH.M) are properly vectorised to ensure maximal speed.

Finally, BOUNDS.M contains all lower- and upper bounds described in section 4.5. The semidefinite programme was solved using CVX (version SDPT3 4.0), a toolbox developed in MATLAB for disciplined convex programming including semidefinite programming [GB08], [GB14]. The third bound is not described in section 4.5—it is an iteration of corollary 4.8 assuming superadditivity, hence in principle it could be violated. If it were violated, this would immediately disprove superadditivity, which has never been observed in our tests.

Issues and further suggestions: It occurs sometimes that the algorithm does not converge to a minimum inside or near the feasible set. We believe that this is due to instabilities in the parameterisation and implementation. The behaviour can occur while using numerical as well as analytical subgradients, although it occurs more often with analytical ones. For every example where we could observe a failure with either numerical or analytical subgradients, one other method (using numerical subgradients, using analytical subgradients or a mixture thereof) worked fine. In cases of failure, the routine issued several warnings and the result usually lies below the lower bound. A different type of implementation might lead to an algorithm that is more stable, but we did not pursue this any further. It might also be worth to consider trying to compute the penalty function analytically.

In terms of performance times, the algorithm is generally fast for small numbers of modes. When analytical subgradients are not implemented, the performance bottleneck is given by the functions XTOH.M, which is called most often. When analytical subgradients are provided, the performance is naturally much faster. This is particularly important when the number of modes increases. While for five modes, the calculation is done within seconds, already for ten modes and depending on the matrix, it can take a minute on a usual laptop (the algorithm now takes the most amount of time for eigenvalue computations, which seems unavoidable). For even larger matrices, it might be advisable to switch from using the Matlab function EIG to EIGF, but for our examples this did not lead to a time gain.

References

[AG88] Arnol'd V I and Givental' A B 1988 Symplectic geometry *Dynamical Systems IV* (Berlin: Springer)

[And+15] Andersen U L *et al* 2016 30 years of squeezed light generation *Phys. Scr.* **91** 053001

[ARL14] Adesso G, Ragy S and Lee A R 2014 Continuous variable quantum information: Gaussian states and beyond *Open Syst. Inf. Dyn.* **21** 1440001

[Arv+95a] Arvind *et al* 1995 The real symplectic groups in quantum mechanics and optics *Pramana* **45** 471–97

[Arv+95b] Arvind *et al* 1995 Two-mode quantum systems: invariant classification of squeezing transformations and squeezed states *Phys. Rev.* A **52** 1609–20

[ASI04] Adesso G, Serafini A and Illuminati F 2004 Extremal entanglement and mixedness in continuous variable systems *Phys. Rev.* A **70** 022318

[Bha07] Bhatia R 2007 *Positive Definite Matrices* (Princeton, NJ: Princeton University Press)

[Bha96] Bhatia R 1996 *Matrix Analysis* (Berlin: Springer)

[BL05] Braunstein S L and van Loock P 2005 Quantum information with continuous variables *Rev. Mod. Phys.* **77** 513–77

[Bra05] Braunstein S L 2005 Squeezing as an irreducible resource *Phys. Rev.* A **71** 055801

[BV04] Boyd S and Vandenberghe L 2004 *Convex Optimization* (New York: Cambridge University Press) ISBN 0521833787

[DR79] Dolecki S and Rolewicz S 1979 Metric characterizations of upper semicontinuity *J. Math. Anal. Appl.* **69** 146–52

[ESP02] Eisert J, Scheel S and Plenio M B 2002 Distilling Gaussian states with Gaussian operations is impossible *Phys. Rev. Lett.* **89** 137903

[Fil13] Filip R 2013 Distillation of quantum squeezing *Phys. Rev.* A **88** 063837

[GB14] Grant M and Boyd S 2014 CVX: Matlab Software for Disciplined Convex Programming, version 2.1 (http://cvxr.com/cvx)

[GB08] Grant M and Boyd S 2008 Graph implementations for nonsmooth convex programs *Recent Advances in Learning and Control (Lecture Notes in Control and Information Sciences)* ed V Blondel, S Boyd and H Kimura (Springer) pp 95–110

[GIC02] Giedke G and Cirac J I 2002 Characterization of Gaussian operations and distillation of Gaussian states *Phys. Rev.* A **66** 032316

[Gie+01] Giedke G *et al* 2001 Separability properties of three-mode gaussian states *Phys. Rev.* A **64** 052303

[Gos06] de Gosson M A 2006 *Symplectic Geometry and Quantum Mechanics* (*Operator Theory: Advances and Applications/Advances in Partial Differential Equations*) (Basel: Birkhäuser) ISBN 9783764375751

[Hee+06] Heersink J *et al* 2006 Distillation of squeezing from non-gaussian quantum states *Phys. Rev. Lett.* **96** 253601

[KK97] Kuntsevich A and Kappel F 1997 SolvOpt: the solver for local nonlinear optimization problems (manual) (Institute for Mathematics, Karl-Franzens University of Graz)

[KL10] Kok P and Lovett B W 2010 *Introduction to Optical Quantum Information Processing* (Cambridge: Cambridge University Press)

[Kok+07] Kok P *et al* 2007 Linear optical quantum computing with photonic qubits *Rev. Mod. Phys.* **79** 135–74

[Kra+03] Kraus B *et al* 2003 Entanglement generation and Hamiltonian simulation in continuous-variable systems *Phys. Rev.* A **67** 042314

[Lee88] Lee C T 1988 Wehrl's entropy as a measure of squeezing *Opt. Commun.* **66** 52–4

[LGW13] Lercher D, Giedke G and Wolf M M 2013 Standard super-activation for gaussian channels requires squeezing *New J. Phys.* **15** 123003

[Lin00] Lindblad G 2000 Cloning the quantum oscillator *J. Phys. A: Math. Gen.* **33** 5059

[Lvo15] Lvovsky A I 2015 Squeezed light *Photonics Volume 1: Fundamentals of Photonics and Physics* ed D Andrews (New York: Wiley) pp 121–164

[Mag85] Magnus J R 1985 On differentiating eigenvalues and eigenvectors *Econometric Theor.* **1** 179–91

[MK08] Mišta L and Korolkova N 2008 Distribution of continuous-variable entanglement by separable gaussian states *Phys. Rev.* A **77** 050302

[Mor75] Moreau J J 1975 Intersection of moving convex sets in a normed space *Math. Scand.* **36** 159–73
[MS98] McDuff D and Salamon D 1998 *Introduction to Symplectic Topology* (Oxford: Oxford University Press)
[NC00] Nielsen M and Chuang I 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press) (doi:10.1017/CBO9780511976667)
[Oli12] Olivares S 2012 Quantum optics in the phase space *Eur. Phys. J. Spec. Top.* **203** 3–24
[Rec+94] Reck M *et al* 1994 Experimental realization of any discrete unitary operator *Phys. Rev. Lett.* **73** 58–61
[RFP10] Recht B, Fazel M and Parrilo P A 2010 Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization *SIAM Rev.* **52** 471–501
[Roc97] Rockafellar R T 1997 *Convex Analysis* (Princeton, NJ: Princeton University Press) ISBN 9780691015866
[Rud87] Rudin W 1987 *Real and Complex Analysis* (*Mathematics Series*) (New York: McGraw-Hill) ISBN 9780070542341
[SCS99] Simon R, Chaturvedi S and Srinivasan V 1999 Congruences and canonical forms for a positive matrix: application to the Schweinler–Wigner extremum principle *J. Math. Phys.* **40** 3632–42
[SMD94] Simon R, Mukunda N and Dutta B 1994 Quantum-noise matrix for multimode systems: $U(n)$ invariance, squeezing, and normal forms *Phys. Rev.* A **49** 1567–83
[Son98] Sontag E D 1998 *Mathematical Control Theory: Deterministic Finite Dimensional Systems* 2nd edn (New York: Springer)
[SZ97] Scully M O and Zubairy M S 1997 *Quantum Optics* (Cambridge: Cambridge University Press) ISBN 9780521435956
[Tho76] Thompson R C 1976 Convex and concave functions of singular values of matrix sums *Pac. J. Math.* **66** 285–90
[VB96] Vandenberghe L and Boyd S 1996 Semidefinite programming *SIAM Rev.* **38** 49–95
[Wee+12] Weedbrook C *et al* 2012 Gaussian quantum information *Rev. Mod. Phys.* **84** 621
[WEP03] Wolf M M, Eisert J and Plenio M B 2003 Entangling power of passive optical elements *Phys. Rev. Lett.* **90** 047904
[Wil36] Williamson J 1936 On the algebraic problem concerning the normal forms of linear dynamical systems *Am. J. Math.* **58** 141–63
[Wol+04] Wolf M M *et al* 2004 Gaussian entanglement of formation *Phys. Rev.* A **69** 052320
[WW01] Werner R F and Wolf M M 2001 Bound entangled Gaussian states *Phys. Rev. Lett.* **86** 3658–61

Your Paper has now been accepted for publication

Journal of Physics A: Mathematical and Theoretical <onbehalfof+jphysa+iop.org@manuscriptcentral.com>

Mi 14.09.2016 17:43

An:Idel, Martin <martin.idel@tum.de>; daniel.lercher@ma.tum.de <daniel.lercher@ma.tum.de>; Wolf, Michael Marc <m.wolf@tum.de>;

Dear Mr Idel,

Re: "An operational measure for squeezing" by Idel, Martin; Lercher, Daniel; Wolf, Michael
Article reference: JPhysA-106250.R1

We are pleased to tell you that we have now formally accepted your Paper. We have everything we need to proceed to publish your Paper in Journal of Physics A: Mathematical and Theoretical.

We will contact you again soon when proofs of your article are ready for final approval. Please return your article proofs by the date given to enable us to publish the final version of record as soon as possible.

All articles published by IOP Publishing are available online to readers at http://iopscience.org/. For more information, please contact our Customer Services department at custserv@iop.org. For advice on complying with US funder requirements, please go to http://iopscience.iop.org/info/page/chorus.

Thank you for choosing to publish in Journal of Physics A: Mathematical and Theoretical. We look forward to publishing your Paper.

Yours sincerely

Lucy Joy

On behalf of the IOP peer-review team:
Sarah Whitehouse - Editor
Eimear O'Callaghan, Phil Brown and Thomas Farrell - Associate Editors
Kayleigh Parsons and Lucy Joy - Editorial Assistants
jphysa@iop.org

IOP Publishing
Temple Circus, Temple Way, Bristol
BS1 6HG, UK

www.iopscience.org/jphysa

Letter reference: DRWA03

Re: Permission of use in PhD thesis

LÖSCHEN ANTWORTEN ALLEN ANTWORTEN WEITERLEITEN •••

Kathryn Shaw <Kathryn.Shaw@iop.org> im Auftrag von Per **Als ungelesen markieren**

Fr 23.09.2016 15:09

An: Idel, Martin;

Dear Martin Idel,

Thank you for your email and for taking the time to seek this permission.

When you transferred the copyright in your article to IOP, we granted back to you certain rights, including the right to include the Final Published Versionof the article within any thesis or dissertation. Please note you may need to obtain separate permission for any third party content you included within your article.

Please include citation details, "© IOP Publishing. Reproduced with permission. All rights reserved" and for online use, a link to the Version of Record.

The only restriction is that if, at a later date, your thesis were to be published commercially, further permission would be required.

Please let me know if you have any further questions.

In the meantime, I wish you the best of luck with the completion of your dissertation.

Kind regards,

Kathryn Shaw

Copyright & Permissions Team
Gemma Alaway – Rights & Permissions Adviser
Kathryn Shaw - Editorial Assistant

Contact Details
E-mail: permissions@iop.org

For further information: http://iopscience.iop.org/page/copyright

Please see our Author Rights Policy http://ioppublishing.org/author-rights/

Please note: We do not provide signed permission forms as a separate attachment. Please print this email and provide it to your institution as proof of permission.

Please note: Any statements made by IOP Publishing to the effect that authors do not need to get permission to use any content are not intended to constitute any sort of legal advice. Authors must make their own decisions as

Perturbation Bounds for Williamson's Symplectic Normal Form

M. Idel, S. Soto Gaona and M. M. Wolf September 26, 2016

Williamson's normal form defines a diagonalisation of positive semidefinte matrices via symplectic conjugation. The corresponding diagonal entries are therefore also called the "symplectic spectrum". The normal form and the symplectic spectrum is very important in continuous variable quantum information as symplectic conjugation corresponds to the implementation of Gaussian unitaries which include many important operations like beam-splitters and phase shifters on light packets. Moreover, the symplectic spectrum can be interpreted as "mixedness" of the quantum state and completely determines its entropy in the case of Gaussian states. Therefore it is interesting to consider perturbation bounds for the normal form. Such bounds have recently been published in [2, 3].

1 Stability of symplectic eigenvalues

Based on the usual perturbation theory, we provide the following bound:

Theorem 1.1. *Let $M, M' \in \mathbb{R}^{2n\times 2n}$ be two positive definite matrices and $\tilde{D}, \tilde{D}'$ their Williamson diagonalisations. Then*

$$\|\tilde{D} - \tilde{D}'\| \leq (\kappa(M)\kappa(M'))^{1/2}\|M - M'\| \tag{1}$$

for every unitarily invariant norm $\|\cdot\|$, where $\kappa(M)$ is the condition number of M.

Clearly, the constant can become arbitrarily bad. We also investigate whether this can be amended and provide a counterexample, stating that the constant must somehow depend on the norms of the matrices involved.

2 Stability of the diagonalising matrix

Similar to usual matrix theory we then consider the stability of the diagonalising matrix $S \in Sp(2n)$, where $S^T M S$ is diagonal. This corresponds to the stability of eigenvectors. Using results from classical perturbation theory of Hermitian matrices, we can show that S is only stable if the symplectic spectrum is nondegenerate as in the case of the usual spectrum. In addtion, we prove that the stability of the eigenspaces leads to the following theorem:

Theorem 2.1. *Let $M \in \mathbb{R}^{2n\times 2n}$ be a positive definite matrix with Williamson decomposition $M = S^{-T}\tilde{D}S^{-1}$.*

Let E be any symmetric matrix with $\|E\| = 1$. Assume that $\varepsilon > 0$ is small enough such that $M_\varepsilon := M + \varepsilon E$ is still positive definite and let $M_\varepsilon = S_\varepsilon^{-T} \tilde{D}_\varepsilon S_\varepsilon^{-1}$ be its Williamson decomposition.

Then for any ε such that M_ε is positive definite and

$$0 < \varepsilon < \min \left\{ \frac{\|M\|}{(6\kappa(M))^{4/3}}, \frac{1}{2\|M\|}, \|M\| \right\} \tag{2}$$

with the condition number κ, then

$$\|S^{-T}S^{-1} - S_\varepsilon^{-T} S_\varepsilon^{-1}\| \leq 9\pi n^3 \kappa(M)^2 \|M^{-1}\|^{1/4} \varepsilon^{1/4}. \tag{3}$$

The proof is technically involved, but easy to relate: The idea is to divide the spectrum into parts that are close by (signifying "eigenspaces" of the symplectic spectrum) and then use results from eigenvalue perturbation theory on the question of the stability of eigenspaces (see for instance [1], Chapters IX and X) to prove the stability of $S^{-T}S^{-1}$. The tricky part is to choose the correct scaling: The estimates for the stability of eigenspaces behave like $1/\delta$, where δ is the spectral gap. Naive choices of dividing the spectrum blow up if δ becomes too small.

3 Applications

We apply the results to provide perturbation bounds for the entropy of a Gaussian state and we also note that Theorem 2.1 is needed to ensure that the algorithm in the paper "An operational measure for squeezing" works as planned.

4 Legal statement

The project was proposed by Michael M. Wolf for the Bachelor thesis of Sebastián Soto Gaona whom I advised during his writing process. The result of the thesis was a weaker version of Theorem 1.1. All other results have primarily been my work with advice by Michael Wolf and proofreading by Sebastián Soto Gaona.

References

[1] Rajendra Bhatia. *Matrix Analyis.* Springer, 1996.

[2] Rajendra Bhatia and Tanvi Jain. On symplectic eigenvalues of positive definite matrices. *Journal of Mathematical Physics*, 56(11), 2015.

[3] Robert König. The conditional entropy power inequality for gaussian quantum states. *Journal of Mathematical Physics*, 56(2), 2015.

Perturbation Bounds for Williamson's Symplectic Normal Form

Martin Idel, Sebatián Soto Gaona, Michael M. Wolf

Zentrum Mathematik, Technische Universität München

Abstract

Given a real-valued positive semidefinite matrix, Williamson proved that it can be diagonalised using symplectic matrices. The corresponding diagonal values are known as the symplectic spectrum. This paper is concerned with the stability of Williamson's decomposition under perturbations. We provide norm bounds for the stability of the symplectic eigenvalues and prove that if S diagonalises a given matrix M to Williamson form, then S is stable if the symplectic spectrum is nondegenerate and S^TS is always stable. Finally, we sketch a few applications of the results in quantum information theory.

1. Introduction

It is well-known that the eigenvalues of a Hermitian matrix are stable under small perturbations. This lies at the heart of perturbation theory, which in turn is important in numerical analysis as well as most areas of theoretical physics. Many classic books have been fully devoted to (spectral) perturbation theory in finite [Wil65] or infinite [Kat95] dimensions and the results are well-known today.

Next to the general notion of eigenvalues of a linear operator, in symplectic linear algebra there exists the notion of *symplectic eigenvalues*: Given a positive semidefinite matrix $M \in \mathbb{R}^{2n\times 2n}$, Williamson [Wil36] showed that there exists a symplectic basis such that M is diagonal. The diagonal entries form the symplectic spectrum. The symplectic spectrum plays an important role in continuous variable quantum information, since for Gaussian states, the spectrum of the density matrix (which defines for instance the von Neumann entropy) can be obtained from the symplectic spectrum [HSH99]. The literature on symplectic eigenvalue perturbation is not as rich as for the usual eigenvalue problem. First results concerning perturbations for matrices in Williamson normal form can be found in [SEW05]. A more general approach was published in [Kön15], and a bound similar to usual matrix perturbation bounds in the literature for matrix analysis has appeared recently in [BJ15]. The bounds in the present paper improve on the last result and also consider the Williamson analoga of eigenvectors and eigenspaces.

To summarise the results, let $M = S^TDS$ be the Williamson decomposition of a positive definite matrix. Since the symplectic spectrum is ultimately defined through the

usual eigenvalue spectrum of a diagonalisable matrix, the immediate intuition is that the spectrum should be stable, while the diagonalising matrix S will not be stable when symplectic eigenvalues are degenerate. Likewise, there is a chance that $S^T S$ is stable, because this would encode the information about "symplectic eigenspaces" and eigenspaces are generally stable [Bha96]. In accordance with this intuition, we find:

- The symplectic spectrum is stable and we derive norm bounds for all unitarily invariant norms, improving the bounds in [BJ15].
- The diagonalising matrix S is stable as long as no eigenvalue crossings occur and we derive a norm bound depending on the smallest gap in the spectrum. We also give a counterexample for the stability of matrices with degenerate eigenvalues.
- $S^T S$ is stable and we derive norm bounds for the operator norm.

These results can be useful for proving continuity and approximation results at least in the context of continuous variable quantum information. We sketch a few applications in the last section. For the reader's convenience, we recall the most important theorems and a number of small lemmata with simple calculations in Appendix A.

2. Notation and Williamson's normal form

Throughout this paper, let $\sigma \in \mathbb{R}^{2n\times 2n}$ be the standard symplectic form defined as

$$\sigma = \begin{pmatrix} 0 & \mathbb{1}_n \\ -\mathbb{1}_n & 0 \end{pmatrix}. \tag{1}$$

Furthermore, denote by $Sp(2n)$ the group of $2n \times 2n$ real symplectic matrices, by $O(n) \subseteq \mathbb{R}^{n\times n}$ the group of real orthogonal matrices, and by $U(n) \subseteq \mathbb{C}^{n\times n}$ the group of unitary matrices. Let us now define the symplectic spectrum through Williamson's theorem:

Theorem 2.1 (Williamson [Wil36])**.** *Let $M \in \mathbb{R}^{2n\times 2n}$ be a positive definite matrix. Then there exists a nonnegative diagonal matrix $D \in \mathbb{R}^{n\times n}$ and a symplectic matrix $S \in Sp(2n)$ such that*

$$S^T M S = \mathrm{diag}(D, D). \tag{2}$$

We can assume without loss of generality that $D_{11} \geq D_{22} \geq \ldots \geq D_{nn} > 0$. The entries of D are sometimes called the symplectic eigenvalues *of M and they are the positive eigenvalues of $i\sigma M$.*

The theorem can be extended to the case of positive semidefinite matrices M. In this case, $S^T M S = \mathrm{diag}(D_1, D_2)$ where D_1 and D_2 contain zeroes.

Proof. One proof that covers also the case of semidefinite matrices can be found in [Gos06]. We sketch the proof of [SCS99] and [SG15]:

Let $\mathrm{diag}(D, D) =: \tilde{D}$. Using that D and M are positive definite, consider an ansatz of the form $S = M^{-1/2} K \tilde{D}^{1/2}$, where $K \in O(2n)$. By construction $S^T M S = \tilde{D}$ and we only need to check that K can be chosen such that S is symplectic. This is equivalent to

$$K^T (M^{-1/2} \sigma M^{-1/2}) K = \begin{pmatrix} 0 & D^{-1} \\ -D^{-1} & 0 \end{pmatrix}.$$

Using that $(M^{-1/2}\sigma M^{-1/2})^T = -M^{-1/2}\sigma M^{-1/2}$, we know that we can indeed find an orthogonal K achieving this construction. The idea is that $iM^{1/2}\sigma M^{1/2}$ is a Hermitian matrix and therefore diagonalisable by a unitary matrix $U \in U(2n)$. It is easy to see that the eigenvalues come in pairs $\pm\lambda_j$ with eigenvectors $x_j \pm iv_j$ for $j = 1, \ldots, n$ and $x_j, y_j \in \mathbb{R}^{2n}$. One can then show that K is given by $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ using that $\sigma x_j = y_j$ and $\sigma y_j = -x_j$. □

The goal of the main part of this paper is to consider the stability of the symplectic eigenvalues, the diagonalising matrix S and the matrix $S^T S$.

3. Stability of the symplectic eigenvalues

Let us first consider the stability of D:

Theorem 3.1. *Let $M, M' \in \mathbb{R}^{2n \times 2n}$ be two positive definite matrices and $\tilde{D}, \tilde{D}'$ their Williamson diagonalisations as in Theorem 2.1. Then*

$$\|\tilde{D} - \tilde{D}'\| \leq (\kappa(M)\kappa(M'))^{1/2} \|M - M'\| \tag{3}$$

for every unitarily invariant norm $\|\cdot\|$, where $\kappa(M)$ is the condition number of M.

Proof. First note that by Williamson's theorem, $i\sigma M$ is diagonalisable. This can be seen via $S^{-1}(i\sigma M)S = i\sigma S^T M S = i\sigma\,\mathrm{diag}(D, D)$ and the fact that the latter has eigenvalues $\pm D_{jj}$ with eigenvectors $(0, \ldots, 0, 1, 0, \ldots, 0, \pm i, 0, \ldots, 0)^T$, where 1 and $\pm i$ are at positions j and $n + j$. Let T be the matrix diagonalising $i\sigma\,\mathrm{diag}(D, D)$. Hence $i\sigma M$ is diagonalisable by ST with real eigenvalues and the eigenvalues are given by $\pm D_{ii}$ for $i = 1, \ldots n$.
Using Lemma A.2 ([Bha96] (Theorem VIII.3.9)), we obtain directly:

$$\|\tilde{D} - \tilde{D}'\| \leq (\kappa(ST)\kappa(S'T'))^{1/2} \|i\sigma M - i\sigma M'\| \tag{4}$$

$$= (\kappa(ST)\kappa(S'T'))^{1/2} \|M - M'\| \tag{5}$$

for all unitarily invariant norms. We also used $\|i\sigma N\| = \|N\|$ for all Hermitian N and all unitarily invariant norms, since $i\sigma$ is a unitary matrix.
Since $i\sigma\,\mathrm{diag}(D, D)$ is Hermitian, its eigenvectors are orthogonal and we can choose $T \in U(2n)$. Therefore, since $\kappa(A) = \|A^{-1}\|_\infty \|A\|_\infty$ for all invertible matrices, $\kappa(ST) = \kappa(S)$ as the operator norm $\|\cdot\|_\infty$ is unitarily invariant.

Let us now proceed as in [SG15]: We can write $S = M^{-1/2}K\tilde{D}^{1/2}$ with an orthogonal matrix $K \in O(2n)$ using the proof of Williamson's theorem. Then

$$\begin{aligned}\kappa(S) &= \|S\|_\infty \|S^{-1}\|_\infty = \|M^{-1/2}K\tilde{D}^{1/2}\|_\infty \|\tilde{D}^{-1/2}K^T M^{1/2}\|_\infty \\ &\leq \|M^{-1/2}\|_\infty \|\tilde{D}^{1/2}\|_\infty \|\tilde{D}^{-1/2}\|_\infty \|M^{1/2}\|_\infty = \kappa(M^{1/2})\kappa(\tilde{D}^{1/2})\end{aligned}$$

Furthermore,

$$\|\tilde{D}\|_\infty \leq \max\{s(i\sigma M)\} = \|i\sigma M\|_\infty = \|M\|_\infty \tag{6}$$

$$\|\tilde{D}^{-1}\|_\infty \leq \max\{s(i\sigma M^{-1})\} = \|i\sigma M^{-1}\|_\infty = \|M^{-1}\|_\infty \tag{7}$$

where $s(A)$ denotes the vector of singular values of A. Using the C^*-property of the operator norm we obtain $\|(\tilde{D}^{1/2})^2\|_\infty = \|\tilde{D}^{1/2}\|_\infty^2$ and hence

$$\kappa(S) \leq \kappa(M^{1/2})\kappa(\tilde{D}^{1/2}) \leq \kappa(M^{1/2})\kappa(M^{1/2}) = \kappa(M).$$

Since the same is true for M' this completes the proof. □

In [BJ15], the authors provide a different bound of this type, which for the operator norm reads

$$\|\tilde{D} - \tilde{D}'\|_\infty \leq (\|M\|_\infty^{1/2} + \|M'\|_\infty^{1/2})\|M - M'\|_\infty^{1/2}. \tag{8}$$

Note that the scaling in $\|M - M'\|_\infty$ is better in Theorem 3.1 than in Bhatia and Jain's bound [BJ15].
One natural question is, whether there is hope to improve a lot on this inequality. In particular, let us ask the question whether an inequality of the type

$$\|\tilde{D} - \tilde{D}'\| \leq c\|M - M'\| \tag{9}$$

holds for some constant $c \in \mathbb{R}$ independent of M, M' and some unitarily invariant norm. The answer is "no":

Proposition 3.2. *Consider the following matrices:*

$$M = \operatorname{diag}(x, 1) \qquad E = \begin{pmatrix} 2 & -5 \\ -5 & -2 \end{pmatrix} \tag{10}$$

Let $M_\varepsilon := M + \varepsilon \cdot E$ for $\varepsilon > 0$. Then, for all $0 < \varepsilon < 1/10$ and for all $c > 0$ there exists an $x_0 \geq 1$ such that for all $x \geq x_0$ we have

$$\|\tilde{D} - \tilde{D}_\varepsilon\| > c\|M - M_\varepsilon\| \tag{11}$$

for all unitarily invariant norms, thereby showing that c must depend on M and M' in equation (9).

Proof. First note that $M_\varepsilon > 0$ for $x \geq 1$ and $\varepsilon < 1/10$, since trace and determinant are both positive. Note that $\|M - M_\varepsilon\| = \varepsilon\|E\|$. Now, since $\mathrm{tr}(E) = 0$, the singular values of E are both the same and given by $s(E) = \sqrt{29}$.
Since $\tilde{D}$ is two-dimensional and M, M_ε are invertible, $\tilde{D}$ and $\tilde{D}_\varepsilon$ are multiples of the identity. Any unitarily invariant norm is a so called *gauge function* of the singular values (see [Bha96], chapter IV). Since the matrices $\tilde{D}$ and $\tilde{D}_\varepsilon$ have only one singular value (excluding multiplicities), this implies that we can prove the statement for all unitarily invariant norm by proving it for the operator norm only.
Now we need to calculate the singular value of $\tilde{D}$ and $\tilde{D}_\varepsilon$, which is the positive eigenvalue of $i\sigma M$ and $i\sigma M_\varepsilon$ respectively. The characteristic polynomials are

$$\chi(i\sigma M) = \lambda^2 - x \tag{12}$$
$$\chi(i\sigma M_\varepsilon) = \lambda^2 + 5^2\varepsilon^2 - (1 - 2\varepsilon)(x + 2\varepsilon) \tag{13}$$

Note that for $x \geq 1$ and ε small enough ($\varepsilon < 1/10$ is sufficient), we have $\lambda_1(i\sigma M) = \sqrt{x} \geq \sqrt{x - 2\varepsilon(x-1) - 29\varepsilon^2} = \lambda_1(i\sigma M_\varepsilon)$. Therefore, we have:

$$c\|M - M_\varepsilon\|_\infty - \|\tilde{D} - \tilde{D}_\varepsilon\|_\infty < 0 \tag{14}$$
$$\Leftrightarrow \quad \sqrt{29}c\varepsilon - |\sqrt{x} - \sqrt{x - 2\varepsilon(x-1) - 29\varepsilon^2}| \leq 0 \tag{15}$$
$$\Leftrightarrow \quad 2\sqrt{29x}c\varepsilon \leq 29\varepsilon^2(1 + c^2) + 2\varepsilon(x - 1) \tag{16}$$

if we assume $x \geq 1$ and $\varepsilon < 1/10$. For $c = 1$, if $x \geq 33$, we have $\sqrt{29x} < (x - 1)$ and therefore (independent of ε) $\|\tilde{D} - \tilde{D}_\varepsilon\|_\infty \geq \|M - M_\varepsilon\|_\infty$. Similarly, for any $c > 0$ we can find $x_0 \geq 0$, such that for all $x \geq x_0$ equation (16) is satisfied, since $2(x-1) - \sqrt{29x}c \to +\infty$. □

A further evaluation shows that if one sets $c = \kappa(M)^{1/2}\kappa(M_\varepsilon)^{1/2} \approx x$ then $c\|M - M_\varepsilon\|_\infty$ scales as $x\varepsilon$ to lowest order in ε (for $x \geq 1$), while $\|D - D_\varepsilon\|_\infty$ scales as $\sqrt{x}\varepsilon$. Therefore, the example above does not attain the bound. Whether the bound $c = \kappa(M)^{1/2}\kappa(M_\varepsilon)^{1/2}$ is optimal can therefore not be determined by this counterexample. However, the scaling of c in $x = \|M\|_\infty$ can only be improved by at most a square root.

4. Stability of the diagonalising matrix S

Next, we will analyse stability of the matrix S. General wisdom from usual diagonalisation of Hermitian matrices tells us that this should hold at least when the eigenvalues are simple:

Proposition 4.1. *Let $M \in \mathbb{R}^{2n\times 2n}$ be a positive definite matrix such that all eigenvalues of $i\sigma M$ are nondegenerate and let $S \in Sp(2n)$ be the matrix diagonalising M in Williamson's theorem. Let E be a symmetric matrix with $\|E\|_\infty = 1$, $\varepsilon > 0$ such that $M_\varepsilon := M + \varepsilon E$ is positive definite. Then we have:*

For $\varepsilon > 0$ small enough, the diagonalising matrix $S_\varepsilon \in Sp(2n)$ of M_ε can be chosen in such a way that

$$\|S - S_\varepsilon\|_\infty < 4\left(\sqrt{\kappa(M)} + \frac{\sqrt{n^3\|M\|_\infty/\|M^{-1}\|_\infty}}{2\delta}\right)\|M^{-1/2}\|_\infty\sqrt{\varepsilon}. \tag{17}$$

where $\delta := \min_{i\neq j}|\lambda_i(i\sigma M) - \lambda_j(i\sigma M)|$ and $\kappa(M)$ is the condition number.

Proof. First observe that $\|S - S_\varepsilon\|_\infty = O(\sqrt{\varepsilon})$ cannot be true for all choices of S and S_ε if those matrices are not unique, which occurs whenever there exists a matrix $O \in Sp(2n) \cap O(2n)$ that commutes with M or M'.
We consider the construction of S in the proof of Theorem 2.1 as $S = M^{-1/2}K\tilde{D}^{1/2}$ where $\tilde{D} = \mathrm{diag}(D, D)$. The stability of S depends thus on the stability of K. Since the eigenvalues are simple, it is known that the eigenvectors are analytic functions in ε ([Wil65], chapter 2, section 5). Let $x_i(\varepsilon)$ denote the normalised eigenvectors of $i(M + \varepsilon E)^{-1/2}\sigma(M + \varepsilon E)^{-1/2}$. Then, using Lemma A.1, there exists some constant c_{vec} such that for all $\varepsilon < c_{\mathrm{vec}}$ and all i we have

$$\|x_i - x_i(\varepsilon)\|_2 \leq \frac{2n}{\min_{i\neq j}|\lambda_i(i\sigma M) - \lambda_j(i\sigma M)|}\varepsilon. \tag{18}$$

Note that $\lambda_j(iM^{1/2}\sigma M^{1/2}) = \lambda_j(i\sigma M)$, $iM^{1/2}\sigma M^{1/2}$ is Hermitian and $\|E\|_\infty \leq 1$ fulfiling assumptions of Lemma A.1.
We know that the parallelogram law holds for the vector norm $\|\cdot\|_2$:

$$\begin{aligned}\|\Re(x_i) + i\Im(x_i) - \Re(x_i(\varepsilon)) - i\Im(x_i(\varepsilon))\|_2^2 + \|\Re(x_i) - i\Im(x_i) - \Re(x_i(\varepsilon)) + i\Im(x_i(\varepsilon))\|_2^2 \\ = 2\|\Re(x_i) - \Re(x_i(\varepsilon))\|_2^2 + 2\|\Im(x_i) - \Im(x_i(\varepsilon))\|_2^2\end{aligned}$$

Furthermore, we know that K consists of the real and imaginary parts of eigenvectors of $M^{-1/2}\sigma M^{-1/2}$ and that when $x_i = \Re(x_i) + i\Im(x_i)$ is an eigenvector to the eigenvalue λ_i, then $x_i' = \Re(x_i) - i\Im(x_i)$ is the eigenvector to the eigenvalue $-\lambda_i$. Thus we can find K_ε such that:

$$\begin{aligned}\|K - K_\varepsilon\|_\infty &\leq \left(\sum_{i=1}^{2n}\|K_i - (K_\varepsilon)_i\|_2^2\right)^{1/2} \\ &= \left(\sum_{i=1}^{n}\|\Re(x_i) - \Re(x_i(\varepsilon))\|_2^2 + \sum_{i=1}^{n}\|\Im(x_i) - \Im(x_i(\varepsilon))\|_2^2\right)^{1/2} \\ &= \left(\frac{1}{2}\sum_{i=1}^{n}\|x_i - x_i(\varepsilon)\|_2^2 + \frac{1}{2}\sum_{i=1}^{n}\|x_i' - x_i'(\varepsilon)\|_2^2\right)^{1/2} \\ &\overset{\text{Lemma A.1}}{\leq} \left(\sum_{i=1}^{n}\frac{4n^2}{\min_{i\neq j}|\lambda_i(i\sigma M) - \lambda_j(i\sigma M)|^2}\varepsilon^2\right)^{1/2}\end{aligned}$$

$$\leq \frac{2n^{3/2}}{\min_{i\neq j}|\lambda_i(i\sigma M)-\lambda_j(i\sigma M)|}\varepsilon$$

where K_i denotes the i-th column of K. Here, we used the fact that $\|A\|_\infty \leq \|A\|_F$ for the Frobenius norm, which is equivalent to the right hand side of the first inequality. But then, using equation (6), Lemma A.4, A.5 and our stability result Theorem 3.1 we obtain:

$$\begin{aligned}
\|S-S_\varepsilon\|_\infty &= \|M^{-1/2}K\tilde{D}^{1/2} - M_\varepsilon^{-1/2}K_\varepsilon\tilde{D}_\varepsilon^{1/2}\|_\infty \\
&\leq \|M^{-1/2}-M_\varepsilon^{-1/2}\|_\infty\|\tilde{D}^{1/2}\|_\infty + \|M_\varepsilon^{-1/2}\|_\infty\|K-K_\varepsilon\|_\infty\|\tilde{D}^{1/2}\|_\infty \\
&\quad + \|M_\varepsilon^{-1/2}\|_\infty\|\tilde{D}^{1/2}-\tilde{D}_\varepsilon^{1/2}\|_\infty \\
&\leq \|M^{1/2}\|_\infty\|M^{-1/2}\|_\infty\|M_\varepsilon^{-1/2}\|_\infty\varepsilon^{1/2} \qquad (19)\\
&\quad + \|M_\varepsilon^{-1/2}\|_\infty(\kappa(M)\kappa(M_\varepsilon))^{1/4}\varepsilon^{1/2} \qquad (20)\\
&\quad + \|M_\varepsilon^{-1/2}\|_\infty\|M^{1/2}\|_\infty\frac{2n^{3/2}}{\min_{i\neq j}|\lambda_i(i\sigma M)-\lambda_j(i\sigma M)|}\varepsilon. \qquad (21)
\end{aligned}$$

We can now use Lemma A.6 and A.7 to obtain for $\varepsilon < 1/(2\|M^{-1}\|_\infty)$ and $\varepsilon < \|M\|_\infty$:

$$\|M_\varepsilon^{-1/2}\|_\infty \leq 2\|M^{-1/2}\|_\infty, \qquad \kappa(M_\varepsilon)\leq 4\kappa(M).$$

By assumption

$$\min_{i\neq j}|\lambda_i(i\sigma M)-\lambda_j(i\sigma M)| \geq \delta.$$

Hence we have for the summands in (19) - (21)

$$\begin{aligned}
\|M^{1/2}\|_\infty\|M^{-1/2}\|_\infty\|M_\varepsilon^{-1/2}\|_\infty\varepsilon^{1/2} &\leq \sqrt{2}\kappa(M)^{1/2}\|M^{-1/2}\|_\infty\varepsilon^{1/2} \qquad (22)\\
\|M_\varepsilon^{-1/2}\|_\infty(\kappa(M)\kappa(M_\varepsilon))^{1/4}\varepsilon^{1/2} &\leq \sqrt{2}\|M^{-1/2}\|_\infty(4\kappa(M)^2)^{1/4}\varepsilon^{1/2} \qquad (23)\\
\|M_\varepsilon^{-1/2}\|_\infty\|M^{1/2}\|_\infty\frac{2n^{3/2}}{\min_{i\neq j}|\lambda_i-\lambda_j|}\varepsilon &\leq (2\|M^{-1}\|_\infty\varepsilon)^{1/2}\|M^{1/2}\|_\infty\frac{2n^{3/2}}{\delta}\varepsilon^{1/2} \\
&\leq \|M^{1/2}\|_\infty\frac{2n^{3/2}}{\delta}\varepsilon^{1/2} \qquad (24)
\end{aligned}$$

where we used $\|M^{-1}\|_\infty\varepsilon < 1/2$ by assumption on ε.
Put together, this implies that for all $0<\varepsilon<\min\{1/(2\|M^{-1}\|_\infty), \|M\|_\infty, c_{\text{vec}}\}$, we can find S_ε diagonalising M_ε:

$$\|S-S_\varepsilon\|_\infty < 2\left((1+1/\sqrt{2})\kappa(M)^{1/2} + \frac{n^{3/2}\|M\|_\infty^{1/2}}{\delta\|M^{-1}\|_\infty^{1/2}}\right)\|M^{-1/2}\|_\infty\varepsilon^{1/2}. \qquad (25)$$

The constant is probably not optimal. □

However, this will not be true in general if we have eigenvalue crossings. To see this, consider the following counterexample:

$$M = \begin{pmatrix} 1 & \varepsilon & 0 & 0 \\ \varepsilon & 1 & 0 & 0 \\ 0 & 0 & 1 & \varepsilon \\ 0 & 0 & \varepsilon & 1 \end{pmatrix} \quad M' = \mathrm{diag}(1+\varepsilon, 1-\varepsilon, 1+\varepsilon, 1-\varepsilon) \tag{26}$$

By Williamson's theorem, for two matrices S_1, S_2 diagonalising M, we have $S_1^{-1}S_2 \in Sp(2n) \cap O(2n)$ and $[S_1^{-1}S_2, M] = 0$. Hence we need to consider the commutants of M and M'. For $\varepsilon > 0$, an easy computation shows that $[M, O] = 0$ or $[M', O] = 0$ if and only if

$$O = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad A, B, C, D \in \mathbb{R}^{2\times 2}, [A, E] = [B, E] = [C, E] = [D, E] = 0,$$

where $E \in \mathbb{R}^{2\times 2}$ denotes the upper left block in M or M'. This reduces the problem to finding the commutant of the upper left blocks in M. A simple computation shows that these are independent of ε. More precisely,

$$\begin{aligned} [A, \mathrm{diag}(1+\varepsilon, 1-\varepsilon)] = 0 \quad &\Leftrightarrow \quad A = \mathrm{diag}(a, b),\ a, b \in \mathbb{R} \\ [A, \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}] = 0 \quad &\Leftrightarrow \quad A = \begin{pmatrix} a & b \\ b & a \end{pmatrix},\ a, b \in \mathbb{R}. \end{aligned}$$

But then, the commutant of M and M' are independent of $\varepsilon > 0$ and so is the intersection of the commutant with $Sp(2n) \cap O(2n)$. Since this intersection is a closed set (commutants are closed), this implies that any matrix S diagonalising $i\sigma M$ with $\|S\|_\infty = 1$ and any matrix S' diagonalising $i\sigma M'$ either fulfil $\|S - S'\|_\infty > C$ for some fixed constant $C > 0$ independent of ε, or there is a matrix S diagonalising both $i\sigma M$ and $i\sigma M'$. Since $[i\sigma M, i\sigma M'] \neq 0$, the two matrices cannot be diagonalised simultaneously, whence $\|S - S'\|_\infty$ cannot become arbitrarily close to zero.

5. Stability of the matrix $S^{-T}S^{-1}$

We have seen that S need not be stable when eigenvalue crossings occur, because eigenvectors need not be stable. However, it turns out that $S^{-T}S^{-1}$ is still stable because it contains only the (real parts of) projections onto the eigenspaces, which are stable according to general wisdom. In this section, $\|\cdot\|$ will always denote the norm $\|\cdot\|_\infty$ in order not to clutter the text with notation.

Theorem 5.1. *Let $M \in \mathbb{R}^{2n\times 2n}$ be a positive definite matrix with Williamson decomposition $M = S^{-T}\tilde{D}S^{-1}$.*
Let E be any symmetric matrix with $\|E\| = 1$. Assume that $\varepsilon > 0$ is small enough such that $M_\varepsilon := M + \varepsilon E$ is still positive definite and let $M_\varepsilon = S_\varepsilon^{-T}\tilde{D}_\varepsilon S_\varepsilon^{-1}$ be its Williamson decomposition.

Then for any ε such that M_ε is positive definite and

$$0 < \varepsilon < \min\left\{\frac{\|M\|}{(6\kappa(M))^{4/3}}, \frac{1}{2\|M\|}, \|M\|\right\} \tag{27}$$

with the condition number κ, then

$$\|S^{-T}S^{-1} - S_\varepsilon^{-T}S_\varepsilon^{-1}\| \leq 9\pi n^3 \kappa(M)^2 \|M^{-1}\|^{1/4}\varepsilon^{1/4}. \tag{28}$$

The inequality can be improved by a more careful analysis of the prefactors.

Proof. From the proof of Theorem 2.1 we know $S = M^{-1/2}K\tilde{D}^{1/2}$ and therefore

$$S^{-T}S^{-1} = M^{1/2}K\tilde{D}^{-1}K^TM^{1/2},$$

where $\tilde{D} = \operatorname{diag}(d_1, \dots, d_n, d_1, \dots, d_n)$ with $d_i > 0$ and $K \in O(2n)$ is given by

$$K = (v_1^{\Re}, \dots, v_n^{\Re}, v_1^{\Im}, \dots, v_n^{\Im}). \tag{29}$$

Here, $v_i = v_i^{\Re} + iv_i^{\Im}$ are the eigenvectors of $iM^{1/2}\sigma M^{1/2}$ corresponding to d_i. We have

$$\begin{aligned}
&\|S^{-T}S^{-1} - S_\varepsilon^{-T}S_\varepsilon^{-1}\| \\
&\quad= \|M^{1/2}K\tilde{D}^{-1}K^TM^{1/2} - M_\varepsilon^{1/2}K_\varepsilon\tilde{D}_\varepsilon^{-1}K_\varepsilon^TM_\varepsilon^{1/2}\| \\
&\quad\leq \|M^{1/2}K\tilde{D}^{-1}K^TM^{1/2} - M_\varepsilon^{1/2}K\tilde{D}^{-1}K^TM^{1/2}\| \\
&\qquad+ \|M_\varepsilon^{1/2}K\tilde{D}^{-1}K^TM^{1/2} - M_\varepsilon^{1/2}K_\varepsilon\tilde{D}_\varepsilon^{-1}K_\varepsilon^TM^{1/2}\| \\
&\qquad+ \|M_\varepsilon^{1/2}K_\varepsilon\tilde{D}_\varepsilon^{-1}K_\varepsilon^TM^{1/2} - M_\varepsilon^{1/2}K_\varepsilon\tilde{D}_\varepsilon^{-1}K_\varepsilon^TM_\varepsilon^{1/2}\| \\
&\quad\leq \|M^{1/2} - M_\varepsilon^{1/2}\|\|M^{1/2}\|\|\tilde{D}^{-1}\| + \|M^{1/2} - M_\varepsilon^{1/2}\|\|M_\varepsilon^{1/2}\|\|\tilde{D}_\varepsilon^{-1}\| \\
&\qquad+ \|M_\varepsilon^{1/2}\|\|M^{1/2}\|\|K\tilde{D}^{-1}K^T - K_\varepsilon\tilde{D}_\varepsilon^{-1}K_\varepsilon^T\| \\
&\quad\leq (\|M^{1/2}\|\|\tilde{D}^{-1}\| + \|M_\varepsilon^{1/2}\|\|\tilde{D}_\varepsilon^{-1}\|)\|M^{1/2} - M_\varepsilon^{1/2}\| && (30) \\
&\qquad+ \|M_\varepsilon^{1/2}\|\|M^{1/2}\|\|\tilde{D}^{-1} - \tilde{D}_\varepsilon^{-1}\| && (31) \\
&\qquad+ \|M_\varepsilon^{1/2}\|\|M^{1/2}\|\|K\tilde{D}^{-1}K^T - K_\varepsilon\tilde{D}^{-1}K_\varepsilon^T\| && (32)
\end{aligned}$$

We deal with each term separately, where the hardest term is the last.
Term (30): Using (7) we have $\|\tilde{D}^{-1}\| \leq \|M^{-1}\|$. For $\|M^{-1}\|\varepsilon < 1/2$ and $\varepsilon < \|M\|$, Lemma A.4 and A.6 imply

$$\begin{aligned}
&(\|M^{1/2}\|\|\tilde{D}^{-1}\| + \|M_\varepsilon^{1/2}\|\|\tilde{D}_\varepsilon^{-1}\|)\|M^{1/2} - M_\varepsilon^{1/2}\| \\
&\qquad\leq (\|M^{1/2}\|\|M^{-1}\| + 4\|M^{1/2}\|\|\tilde{M}^{-1}\|)\varepsilon^{1/2} \\
&\qquad\leq 5\kappa(M)^{1/2}\|M^{-1}\|^{1/2}\varepsilon^{1/2}.
\end{aligned} \tag{33}$$

Term (31): Since $\tilde{S}^{-1}$ diagonalises $M^{-1} \geq 0$, Theorem 3.1 implies

$$\|\tilde{D}^{-1} - \tilde{D}_\varepsilon^{-1}\| \leq (\kappa(M)\kappa(M_\varepsilon))^{1/2}\|M^{-1} - M_\varepsilon^{-1}\|$$

$$\overset{\text{Lemma A.5,A.7}}{\leq} 4\kappa(M)\|M^{-1}\|^2\varepsilon$$

for $\|M^{-1}\|\varepsilon < 1/2$ and with $\|M_\varepsilon\| \leq \|M\| + \varepsilon \leq 2\|M\|$. Plugging this into (31) and using $\varepsilon < \|M\|$ we obtain

$$\|M_\varepsilon^{1/2}\|\|M^{1/2}\|\|\tilde{D}^{-1} - \tilde{D}_\varepsilon^{-1}\| \leq 4\kappa(M)^{3/2}\|M^{-1}\|\varepsilon \tag{34}$$

Term (32): The interesting part is $\|K\tilde{D}^{-1}K^T - K_\varepsilon\tilde{D}^{-1}K_\varepsilon^T\|$. We start by observing:

$$K\tilde{D}^{-1}K^T = \sum_{i=1}^n d_i^{-1}v_i^\Re v_i^{\Re T} + \sum_{i=1}^n d_i^{-1}v_i^\Im v_i^{\Im T} = \sum_{i=1}^n d_i^{-1} \sum_{j\in\{k|d_k=d_i,k=1,\ldots,n\}} (v_j^\Re v_j^{\Re T} + v_j^\Im v_j^{\Im T}). \tag{35}$$

Furthermore,

$$\sum_{j\in\{k|d_k=d_i,k=1,\ldots,n\}} (v_j^\Re v_j^{\Re T} + v_j^\Im v_j^{\Im T}) \quad = \Re \sum_{j\in\{k|d_k=d_i,k=1,\ldots,n\}} (v_j^\Re + iv_j^\Im)(v_j^\Re + iv_j^\Im)^* = \Re(P_M(d_i)) \tag{36}$$

where $\Re$ denotes the real part of the expression and $P_M(d_i)$ denotes the spectral projection onto the eigenvalue subspace of the eigenvalue d_i of $iM^{1/2}\sigma M^{1/2}$. We wish to apply general knowledge about the stability of eigenspaces. For convenience, the relevant theorem ([Bha96] Theorem VII.3.2) is stated in Lemma A.3.
In order to apply it, we need to consider the spectrum of $iM_\varepsilon^{1/2}\sigma M_\varepsilon^{1/2}$: By construction, $\text{spec}(iM^{1/2}\sigma M^{1/2}) = \{\pm d_1, \ldots, \pm d_n\}$. Denote the positive eigenvalues as spec_+, then we can write $\text{spec}_+(iM^{1/2}\sigma M^{1/2}) = \bigcup_{j=1}^k S_j$ where all S_j contain d_i with multiplicities, fulfil $\text{dist}(S_j, S_k) := \min\{|d-e| \mid d \in S_j, e \in S_k\} > \|M\|^{3/4}\varepsilon^{1/4}$ and k is maximal.
The stability of the symplectic spectrum implies:

$$\begin{aligned}\|\tilde{D} - \tilde{D}_\varepsilon\| &\overset{\text{Theorem 3.1}}{<} (\kappa(M)\kappa(M_\varepsilon))^{1/2}\varepsilon \\ &\overset{\text{Lemma A.7}}{\leq} 2\kappa(M)\varepsilon = \frac{6\kappa(M)}{\|M\|^{3/4}}\varepsilon^{3/4}\frac{\|M\|^{3/4}\varepsilon^{1/4}}{3} \\ &\overset{\text{Assumption (27)}}{<} \|M\|^{3/4}\varepsilon^{1/4}/3\end{aligned} \tag{37}$$

hence if we set $d_i := \tilde{D}_{ii}$ and $e_i := (\tilde{D}_\varepsilon)_{ii}$, then we have

$$|d_i - e_i| < \|M\|^{3/4}\varepsilon^{1/4}/3 \quad \forall i = 1, \ldots, n. \tag{38}$$

We can now define the multisets $R_j := \{e_i | d_i \in S_j\}$ for every S_j and make the following observations:

1. The diameter of S_j does not exceed $\|M\|^{3/4}\varepsilon^{1/4}|S_j|$.
2. $|R_j| = |S_j|$ for every $j = 1, \ldots, k$.
3. $\text{dist}(R_i, R_j) > 1/3\|M\|^{3/4}\varepsilon^{1/4}$ for $i \neq j$.

Observation 1 follows from the maximal number of S_j: If the diameter was larger, by the pidgeon-hole principle we could divide S_j into two sets with distance larger than $\|M\|^{3/4}\varepsilon^{1/4}$.
Observation 2 follows, since any $e_i \in R_j$ is at most $1/3\|M\|^{3/4}\varepsilon^{1/4}$ away from some $d_i \in S_j$ and since the distance of S_j and S_k is at least $\|M\|^{3/4}\varepsilon^{1/4}$, $|e_i - d_k| > 2/3\|M\|^{3/4}$ for any $d_k \in S_k$ with $k \neq j$. Incidentally, this also proves Observation 3.
Now let e_i be as defined with equation (38). Using equation (36), we see

$$\|K\tilde{D}^{-1}K^T - K_\varepsilon\tilde{D}^{-1}K_\varepsilon^T\| = \left\|\sum_{j=1}^{k}\sum_{d_i\in S_j} d_i^{-1}(\Re(P_M(d_i)) - \Re(P_{M_\varepsilon}(e_i)))\right\|$$

For every set S_j, pick a value $d_{S_j} \in S_j$ and we have:

$$\begin{aligned}\|K\tilde{D}^{-1}K^T - K_\varepsilon\tilde{D}^{-1}K_\varepsilon^T\| = \Bigg\|&\sum_{j=1}^{k} d_{S_j}^{-1}(\Re(P_M(S_j)) - \Re(P_{M_\varepsilon}(R_j)))\\ &+ \sum_{j=1}^{k}\sum_{d_i\in S_k}(d_i^{-1} - d_{S_j}^{-1})\sum_{d_i\in S_j}(\Re(P_M(d_i)) - \Re(P_{M_\varepsilon}(e_i)))\Bigg\|\\ \leq &\sum_{j=1}^{k} d_{S_j}^{-1}\|\Re(P_M(S_j)) - \Re(P_{M_\varepsilon}(R_j))\| \qquad (39)\\ &+ \sum_{j=1}^{k}\sum_{d_i\in S_j}\frac{|d_i - d_{S_j}|}{d_i d_{S_j}}\|\Re(P_M(d_i)) - \Re(P_{M_\varepsilon}(e_i))\| \qquad (40)\end{aligned}$$

Recall that the real part of an operator T is defined as $\Re(T) := (T + T^*)/2$. Since it is clearly linear for all matrices T, using that $\|T + T^*\| \leq 2\|T\|$ and the fact that every unitarily invariant norm fulfils $\|T^*\| = \|T\|$ implies

$$\|\Re(P_M(S_i)) - \Re(P_{M_\varepsilon}(R_j))\| \leq \|P_M(S_i) - P_{M_\varepsilon}(R_j)\| \quad \forall i, j.$$

Now we can apply Lemma A.3 to the term (39): Let $P_M^c(S_j)$ and $P_{M_\varepsilon}^c(R_i)$ be complementary orthogonal projections such that in particular $P_M(S_j) + P_M^c(S_j) = \mathbb{1}$. Then we have for every $j = 1, \ldots, k$:

$$\begin{aligned}\|P_M(S_j) - P_{M_\varepsilon}(R_j)\| &= \|P_M(S_j)(P_{M_\varepsilon}(R_j) + P_{M_\varepsilon}^c(R_j)) - (P_M(S_j) + P_M^c(S_j))P_{M_\varepsilon}(R_j)\|\\ &= \|P_M(S_j)P_{M_\varepsilon}^c(R_j) - P_M^c(S_j)P_{M_\varepsilon}(R_j)\|\\ &\leq \frac{3\pi}{2\varepsilon^{1/4}\|M\|^{3/4}}\|iM^{1/2}\sigma M^{1/2} - iM_\varepsilon^{1/2}\sigma M_\varepsilon^{1/2}\|\\ &\leq \frac{3\pi}{2\varepsilon^{1/4}\|M\|^{3/4}}(\|M^{1/2}\| + \|M_\varepsilon^{1/2}\|)\|M^{1/2} - M_\varepsilon^{1/2}\|.\end{aligned}$$

Here, we used Observation 3 of the decomposition, which gives a lower bound on $\operatorname{dist}(S_j, R_i)$ for $i \neq j$.

For the term (40) we use that the norm of the difference of two projections never exceeds one:

$$\sum_{j=1}^{k}\sum_{d_i\in S_j}\frac{|d_i-d_{S_j}|}{d_i d_{S_j}}\|\Re(P_M(d_i))-\Re(P_{M_\varepsilon}(e_i))\|\leq\sum_{j=1}^{k}\sum_{d_i\in S_j}\frac{|d_i-d_{S_j}|}{d_i d_{S_j}}. \tag{41}$$

We can now use the Observation 1 and the fact that for any S_j, $|S_j|\leq n$. This gives an upper bound to $|d_i-d_{S_j}|$. Furthermore, $|d_i d_{S_j}|\geq|d_{\min}|^2=1/\|\tilde{D}^{-1}\|^2$ and hence,

$$\sum_{j=1}^{k}\sum_{d_i\in S_j}\frac{|d_i-d_{S_j}|}{d_i d_{S_j}}\leq\sum_{j=1}^{k}n^2\|\tilde{D}^{-1}\|^2\|M\|^{3/4}\varepsilon^{1/4}\leq n^3\|\tilde{D}^{-1}\|^2\|M\|^{3/4}\varepsilon^{1/4}. \tag{42}$$

In total, we obtain

$$\begin{aligned}\|K\tilde{D}^{-1}K^T-K_\varepsilon\tilde{D}^{-1}K_\varepsilon^T\|\leq\sum_{j=1}^{k}d_i^{-1}\frac{3\pi}{2\varepsilon^{1/4}\|M\|^{3/4}}(\|M^{1/2}\|+\|M_\varepsilon^{1/2}\|)\|M^{1/2}-M_\varepsilon^{1/2}\|\\+n^3\|\tilde{D}^{-1}\|^2\|M\|^{3/4}\varepsilon^{1/4}.\end{aligned} \tag{43}$$

Now, we know that $\sum_{i=1}^{k}d_i^{-1}\leq\|\tilde{D}^{-1}\|_1\leq n\|M^{-1}\|$ by equation (7) and the fact that $\|\mathbb{1}\|_1=n$. Using Lemmata A.4, A.6 and A.7, we can now fully evaluate the term (32):

$$\begin{aligned}&\|M_\varepsilon^{1/2}\|\|M^{1/2}\|\|K\tilde{D}^{-1}K^T-K_\varepsilon\tilde{D}^{-1}K_\varepsilon^T\|\\&\qquad\leq 2\|M\|\left(n\|M^{-1}\|\frac{9\pi}{2\varepsilon^{1/4}\|M\|^{3/4}}\cdot\|M\|^{1/2}\varepsilon^{1/2}+n^3\|M^{-1}\|^2\|M\|^{3/4}\varepsilon^{1/4}\right)\\&\qquad\leq 9\pi n\kappa(M)^{3/4}\|M^{-1}\|^{1/4}\varepsilon^{1/4}+n^3\kappa(M)^{7/4}\|M^{-1}\|^{1/4}\varepsilon^{1/4}.\end{aligned} \tag{44}$$

Finally, we can put everything together by substitution (33), (34) and (44) into (30)-(32). Using $\|M^{-1}\|\varepsilon<\|M^{-1}\|^{1/4}\varepsilon^{1/4}<1/2$, we obtain:

$$\|S^{-T}S^{-1}-S_\varepsilon^{-T}S_\varepsilon^{-1}\|\leq 9\pi n^3\kappa(M)^2\|M^{-1}\|^{1/4}\varepsilon^{1/4}. \tag{45}$$

The constant is not optimal. □

6. Applications

Let us now sketch a few applications of the theorems to quantum information theory (an overview can be found in [ARL14]). The basic object in quantum mechanics are the quantum state of a system. In the case of systems consisting of n bosonic modes (such systems are considered especially in quantum optics), an important set of states are the so called *Gaussian states*. They can be characterised by their first and second moments, which correspond to a vector $d\in\mathbb{R}^{2n}$ and a covariance matrix $\gamma\in\mathbb{R}^{2n\times 2n}$. Necessary and sufficient conditions for γ to be the covariance matrix of a quantum state are given as $\gamma\geq i\sigma$ by Heisenberg's inequality. *Pure states* then correspond to symplectic positive

definite matrices. Given two systems of n-modes and a state on two systems given by $\gamma_{AB} \in \mathbb{R}^{4n\times 4n}$, one can consider the *reduced state* of the quantum system, which is given by the upper left $2n \times 2n$-submatrix of γ_{AB}.
An important quantity in quantum information is the entropy of entanglement. As proven in [HSH99], the *entanglement entropy* for Gaussian states is a continuous function of the symplectic spectrum of the reduced state of a system. Given a Gaussian quantum state with covariance matrix γ, it is given by

$$H(\gamma) = \sum_{k=1}^{n} \left(g\left(\frac{d_k+1}{2}\right) - g\left(\frac{d_k-1}{2}\right)\right) \tag{46}$$

where $g(x) = x\log(x)$ and the d_k are the symplectic eigenvalues.
An easy corollary of Theorem 3.1 is the following norm bound on the entropy difference:

Corollary 6.1. *For Gaussian states characterised by (γ_{AB}, d), the entropy of entanglement is continuous in γ_{AB}. Furthermore, for two states γ and $\tilde{\gamma}$ in the interior of the set of covariance matrices, the entropy difference is bounded by*

$$|H(\gamma) - H(\tilde{\gamma})| \leq (\kappa(\gamma)\kappa(\tilde{\gamma})^{1/2})(1 + \log(\max(\|\gamma\|_\infty, (\|\gamma^{-1}\|_\infty^{-1} - 1)/2)))\|\gamma - \tilde{\gamma}\|_1$$

Proof. The entropy is continuous, since g is continuous and the symplectic eigenvalues are continuous.
Let d_k be the entries of D, the Williamson diagonalisation of γ (likewise $\tilde{d}_k$), which implies that $\frac{d_k+1}{2} \geq 1$ always and $\frac{d_k-1}{2} \geq 0$. For $x > 0$ we have

$$\begin{aligned} |x\log(x) - y\log(y)| &= |x\log(x) - y\log(x) + y\log(x) - y\log(y)| \\ &= |(x-y)\log(x) + y\log(1 + (x-y)/y)| \leq |\log(x)||x-y| + |x-y| \end{aligned}$$

For $x = 0$, the upper bound is clearly also true, since $x\log(x) = 0$. Using that $\log((d_k + 1)/2) \leq \log(d_k)$, we obtain

$$\begin{aligned} &\left| \left(g\left(\frac{d_k+1}{2}\right) - g\left(\frac{d_k-1}{2}\right)\right) - \left(g\left(\frac{\tilde{d}_k+1}{2}\right) - g\left(\frac{\tilde{d}_k-1}{2}\right)\right)\right| \\ &\qquad \leq (1 + \log(d_k))|d_k - \tilde{d}_k| + (1 + \log((d_k-1)/2))|d_k - \tilde{d}_k| \end{aligned}$$

Taking the sum and noting that $d_k \leq \|D\|_\infty$ by assumption and $\min_k d_k \leq \|D^{-1}\|_\infty^{-1}$, we have

$$|H(\gamma) - H(\tilde{\gamma})| \leq (1 + \log(\|D\|_\infty))\|D - \tilde{D}\|_1 + (1 + \log((1/\|D^{-1}\|_\infty - 1)/2))\|D - \tilde{D}\|_1.$$

The rest then follows by using Theorem 3.1. □

Note that the bound becomes arbitrarily bad if D has eigenvalues close to one. This is to be expected, since the function $x\log(x)$ is not uniformly continuous at 0 and hence cannot be norm bounded with a constant independent of x.

Another interesting measure in quantum information is the Gaussian entanglement of formation. This is a measure to quantify the amount of entanglement needed to prepare a state of two systems under so-called LOCC operations (local quantum operations on each part of the systems and classical communication between the parts). It was shown in [Wol+04] that this measure can be written as

$$E_{\text{form}}(\gamma_{AB}) = \min\{H(\gamma_p)|\gamma_{AB} \geq S^T S, S \in Sp(2n)\}$$

where γ_p is the reduced state of $S^T S$ and $H(\cdot)$ denotes the entropy of entanglement. Using the methods of [ILW16] and the stability results of this paper, one can now prove:

Proposition 6.2. *The Gaussian entanglement of formation is continuous on the interior of the set of covariance matrices.*

Sketch of the proof. This can be proven in two ways. For γ_{AB} in the interior of the set of covariance matrices, one can either use set-valued analysis as in the proof of Theorem 4.4 in [ILW16] to prove that the set $\{\gamma_{AB} \geq \gamma_0 \geq i\sigma\}$ is convex and varies continuously with γ_{AB}. Then the result follows from Corollary 6.1.
Equivalently, one can use Theorem 5.1 to show that for any $\varepsilon > 0$ and any symmetric E small enough, for any $S^T S \leq \gamma_{AB}$ there exists $S_\varepsilon^T S_\varepsilon \leq \gamma_{AB} + \varepsilon E$, such that their norm difference is small and then apply Corollary 6.1. □

As a last application, let us mention that the stability of $S^T S$ as in Theorem 5.1 is implicitly useful in [ILW16]: There, we provide a program to compute an operational measure for squeezing. Given a covariance matrix γ of a state to be constructed, the program first computes Williamson's normal form and takes $S^T S$ as a starting point. If $S^T S$ was not continuous in the covariance matrix, this would imply that rounding errors in γ could result in the corresponding $S^T S$ not being a feasible point for the program. Theorem 5.1 asserts that this problem cannot occur.

References

[ARL14] Gerardo Adesso, Sammy Ragy, and Antony R. Lee. "Continuous Variable Quantum Information: Gaussian States and Beyond". In: *Open Systems & Information Dynamics* 21.01n02 (2014), p. 1440001. DOI: 10.1142/S1230161214400010.

[Bha96] Rajendra Bhatia. *Matrix Analysis*. Springer, 1996. ISBN: 978-1-4612-0653-8.

[BJ15] Rajendra Bhatia and Tanvi Jain. "On symplectic eigenvalues of positive definite matrices". In: *Journal of Mathematical Physics* 56.11, 112201 (2015). DOI: 10.1063/1.4935852.

[Gos06] Maurice A. de Gosson. *Symplectic Geometry and Quantum Mechanics*. Operator Theory: Advances and Applications / Advances in Partial Differential Equations. Birkhäuser Basel, 2006. ISBN: 9783764375751.

[Hag89] William W. Hager. "Updating the Inverse of a Matrix". In: *SIAM Review* 31.2 (1989), pp. 221–239. DOI: 10.1137/1031049.

[HSH99] A. S. Holevo, M. Sohma, and O. Hirota. "Capacity of quantum Gaussian channels". In: *Phys. Rev. A* 59 (3 Mar. 1999), pp. 1820–1828. DOI: 10.1103/PhysRevA.59.1820.

[ILW16] Martin Idel, Daniel Lercher, and Michael M. Wolf. *An operational measure for squeezing*. arXiv:1607.00873v1 [math-ph]. 2016.

[Kat95] Tosio Kato. *Perturbation Theory for Linear Operators*. 2nd ed. Classics in Mathematics. Springer, 1995.

[Kön15] Robert König. "The conditional entropy power inequality for Gaussian quantum states". In: *Journal of Mathematical Physics* 56.2, 022201 (2015). DOI: 10.1063/1.4906925.

[SCS99] R. Simon, S. Chaturvedi, and V. Srinivasan. "Congruences and canonical forms for a positive matrix: Application to the Schweinler–Wigner extremum principle". In: *Journal of Mathematical Physics* 40.7 (1999), pp. 3632–3642. DOI: 10.1063/1.532913.

[SEW05] A. Serafini, J. Eisert, and M. M. Wolf. "Multiplicativity of maximal output purities of Gaussian channels under Gaussian inputs". In: *Phys. Rev. A* 71 (1 Jan. 2005), p. 012320. DOI: 10.1103/PhysRevA.71.012320.

[SG15] Sebastián Soto Gaona. "Stability problems related to Williamson's symplectic normal form". Bachelor's thesis. Technische Universität München, 2015.

[Wil36] John Williamson. "On the Algebraic Problem Concerning the Normal Forms of Linear Dynamical Systems". In: *American Journal of Mathematics* 58.1 (1936), pp. 141–163. DOI: 10.2307/2371062.

[Wil65] James Hardy Wilkinson. *The algebraic eigenvalue problem*. Vol. 87. Clarendon Press Oxford, 1965.

[Wol+04] M. M. Wolf et al. "Gaussian entanglement of formation". In: *Phys. Rev. A* 69 (5 May 2004), p. 052320. DOI: 10.1103/PhysRevA.69.052320.

A. Some useful lemmata

In this appendix, we will first review the main nontrivial theorems we use in the main text to establish our results. In addition, we give a number of small lemmata that include calculations that are frequently used in the main text to eliminate ε-dependencies of the constants. Since these are not very important for the gist of the argument, they are collected here in order not to further clutter the main text.

Lemma A.1 ([Wil65] Chapter 2 Section 10)**.** *Let A be a Hermitian matrix with nondegenerate spectrum and B be a perturbation with $\|B\|_\infty \leq 1$. Then there exists a number $c_{\text{vec}} > 0$ such that for all $c_{\text{vec}} > \varepsilon > 0$ we have*

$$\|x_i - x_i(\varepsilon)\|_2 \leq \frac{2n}{\min_{i\neq j} |\lambda_i - \lambda_j|}\varepsilon \tag{47}$$

where x_i denotes the i-th eigenvector of A and $x_i(\varepsilon)$ the i-th eigenvalue of $A + \varepsilon B$.

Proof. Since this is not the exact formulation of the section in [Wil65], a few words on how this theorem is related to what is written there: The section computes the first order term in the perturbative expansion of an eigenvector $x_i(\varepsilon)$. Since Hermitian eigenvectors are orthogonal the first order term of the eigenvector expansion of x_1 as in (10.2) of [Wil65] reads

$$\leq \varepsilon \left(\frac{\beta_{21} x_2}{\lambda_1 - \lambda_2} + \ldots + \frac{\beta_{n1} x_n}{\lambda_n - \lambda_2} \right),$$

where $|\beta_{21}| \leq \|B\|_\infty \leq 1$ are some numbers and λ_i are the eigenvalues of A. Hence for ε small enough, we have

$$\begin{aligned}\|x_i - x_i(\varepsilon)\|_2 &\leq \left(\frac{\|x_2\|_2}{\lambda_1 - \lambda_2} + \ldots + \frac{\|x_n\|_2}{\lambda_n - \lambda_2}\right)\varepsilon + \mathcal{O}(\varepsilon^2) \\ &\leq \frac{n}{\min_{i\neq j}|\lambda_i - \lambda_j|}\varepsilon + \mathcal{O}(\varepsilon^2).\end{aligned}$$

For ε small enough, this then implies the bound in the theorem. □

Lemma A.2 ([Bha96] Theorem VIII.3.9)**.** *Let A, B be any two matrices such that $A = SD_1S^{-1}$, $B = TD_2T^{-1}$, where S, T are invertible matrices and D_1, D_2 are real diagonal matrices. Then*

$$\|\mathrm{Eig}^{\downarrow}(A) - \mathrm{Eig}^{\downarrow}(B)\| \leq (\kappa(S)\kappa(T))^{1/2}\|A - B\| \tag{48}$$

for every unitarily invariant norm. Here, κ is the condition number and $\mathrm{Eig}^{\downarrow}$ denotes the (ordered) set of eigenvalues.

Lemma A.3 ([Bha96] Theorem VII.3.2)**.** *Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian operators and let S_1, S_2 be any two subsets of $\mathbb{R}$ such that $\mathrm{dist}(S_1, S_2) = \delta > 0$. Let $E = P_A(S_1)$ ($F = P_B(S_2)$) be the spectral projection onto the space spanned by the eigenvectors of A (B) corresponding to eigenvalues in S_1. Then, for every unitarily invariant norm,*

$$\|EF\| \leq \frac{\pi}{2\delta}\|A - B\| \tag{49}$$

Lemma A.4. *Let $A, B \in \mathbb{C}^{n\times n}$ be positive semidefinite operators. Then, for every unitarily invariant norm,*

$$\|A^{1/2} - B^{1/2}\| \leq \|A - B\|_\infty^{1/2}\|\mathbb{1}\| \tag{50}$$

Proof. This follows directly from the proof of Theorem X.1.1 in [Bha96] using that the square root function is operator monotone on positive semidefinite matrices and $0^{1/2} = 0$. □

Lemma A.5. *Let $A, B \in \mathbb{C}^{n\times n}$ be positive definite operators. Then for every unitarily invariant norm,*

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\|B^{-1}\|\|A - B\| \tag{51}$$

Proof. Calculate:

$$\begin{aligned}\|A^{-1} - B^{-1}\| = \|A^{-1}(\mathbb{1} - AB^{-1})BB^{-1}\| &\leq \|A^{-1}\|\|B^{-1}\|\|(\mathbb{1} - AB^{-1})B\| \\ &\leq \|A^{-1}\|\|B^{-1}\|\|A - B\|\end{aligned}$$

□

Lemma A.6. *Let M be an invertible matrix, E a matrix with $\|E\|_\infty = 1$ and $\|M^{-1}\|_\infty \leq \frac{1}{2\varepsilon}$, then*

$$\|(M+\varepsilon E)^{-1}\|_\infty \leq 2\|M^{-1}\|_\infty \tag{52}$$

Proof. Using the Woodbury formula (which was not found by Woodbury [Hag89]), we have:

$$(M+\varepsilon E)^{-1} = M^{-1} - M^{-1}(I+\varepsilon EM^{-1})^{-1}\varepsilon EM^{-1}.$$

Since $\|M^{-1}\|_\infty \leq \frac{1}{2\varepsilon}$, the Neuman series of $(I+\varepsilon EM^{-1})^{-1}$ converges and we have $(I+\varepsilon EM^{-1})^{-1} = \sum_{n=0}^{\infty} \varepsilon^n (EM^{-1})^n$, and hence $\|(I+\varepsilon EM^{-1})^{-1}\|_\infty \leq \sum_{n=0}^{\infty} \varepsilon^n \|M^{-1}\|_\infty^n \leq 2$, which implies:

$$\|(M+\varepsilon E)^{-1}\|_\infty \leq \|M^{-1}\|_\infty + \|M^{-1}\|_\infty \cdot 2\varepsilon \|EM^{-1}\|_\infty.$$

Finally, since $\varepsilon\|M^{-1}\|_\infty \leq 1/2$ by assumption, we have $\|M^{-1}\|_\infty \cdot 2\varepsilon\|EM^{-1}\|_\infty \leq \|M^{-1}\|_\infty \cdot 2\varepsilon\|M^{-1}\|_\infty \leq \|M^{-1}\|_\infty$. □

Lemma A.7. *Let $M, E \in \mathbb{R}^{n\times n}$, $M \geq 0$ and $\|E\|_\infty = 1$. If $\|M^{-1}\|_\infty \leq \frac{1}{2\varepsilon}$ and $\varepsilon < \|M\|_\infty$, we have*

$$\kappa(M+\varepsilon E) \leq 4\kappa(M). \tag{53}$$

Proof. We use $\kappa(M+\varepsilon E) = \|M+\varepsilon E\|_\infty \|(M+\varepsilon E)^{-1}\|_\infty$ by definition and apply Lemma A.6 to obtain:

$$\kappa(M+\varepsilon E) \leq 2\|M+\varepsilon E\|_\infty \|M^{-1}\|_\infty$$

Using $\|M+\varepsilon E\|_\infty \leq \|M\| + \varepsilon \leq 2\|M\|$ finishes the proof. □

On additive Gaussian quantum channels

M. Idel, R. König September 26, 2016

Quantum channels, i.e. completely positive trace preserving maps, provide the abstract framework to study information transmission. Since the set of channels is too big to be useful it is natural to study subclasses of channels. One such subclass is given by the set of additive channels, channels of the form

$$\mathcal{E}(\rho) = \mathrm{tr}_E \left(U_\lambda(\rho \otimes \rho_E)U_\lambda^*\right) \ , \tag{1}$$

where U is a beam splitter with some transmissivity. Such channels arise naturally when the environment is considered a "resource". We study the set of additive quantum channels from the viewpoint of squeezing. It is then natural to consider the broader class of channels with squeezed environment and a unitary U which cannot squeeze but is not necessarily a beam splitter. We provide a classification of all such channels and connect it to other classes of channels such as covariant channels.

1 Dilation theorem

Given a Gaussian channel (X, Y) which transforms the covariance matrix γ of a Gaussian state as $\gamma \mapsto X^T\gamma X + Y$, we first ask the question which of those channels can be implemented in the form of Equation (1) with passive unitary U. We call those channels *passively dilatable*:

Theorem 1.1. *Let $\Phi_{X,Y}$ be an n-mode Gaussian channel and suppose Y and X have full rank. The following conditions are equivalent:*

(i) There exists a passive dilation with l environment modes and $S \in Sp(2(n+l)) \cap O(2(n+l))$.

(ii) The matrices X, Y satisfy $\mathbb{1}_{2n} - XX^T > 0$, $[X, \sigma_{2n}] = 0$ and $l \geq n$.

We can leave out the condition that X, Y have full rank by introducing some further technicalities. The proof of this theorem (in the full version) heavily relies on properties of the Moore-Penrose pseudoinverse and the structure of matrices in the group $Sp(2n \cap O(2n)$. Some methods are similar to methods in earlier papers studying general dilations in [1, 2].

We also prove uniqueness up to a passive rotation for *minimal dilations*, i.e. those dilations where l is as small as possible. This is in line with the usual uniqueness statements of Stinespring's theorem. In addition, we can also classify channels where the environment is not squeezed. Those are channels that fulfil $[Y, \sigma_{2n}] = 0$ in addition to all other properties of the equivalence.

2 Normal form

Given the dilation result and using the singular value decomposition we obtain the following characterisation of all passively dilatable channels:

Theorem 2.1. *Let $\Phi : \mathcal{B}(A_1 \dots A_n) \to \mathcal{B}(A_1 \cdots A_n)$ be a passively dilatable n-mode Gaussian channel. Then there is an n-mode Gaussian state $\rho_E = \rho_{E_1 \cdots E_n}$, n-mode Gaussian unitaries V, W and transmissivities $\lambda = (\lambda_1, \dots, \lambda_n) \in [0,1]^n$ such that for the multi-mode beam splitter $U_\lambda = U_{\lambda_1}^{A_1 E_1} \otimes \cdots \otimes U_{\lambda_n}^{A_n E_n}$, we have*

$$\Phi(\rho) = V \left(\mathrm{tr}_E \, U_\lambda (W \rho W^* \otimes \rho_E) U_\lambda^*\right) V^* \qquad \text{for all states } \rho. \tag{2}$$

The theorem tells us that passively dilatable channels and additive channels are the same (modulo Gaussian unitaries).

3 Open questions and connections to other work

Passively dilatable or additive Gaussian channels are interesting for the resource theory of squeezing. One could for instance quantify the amount of squeezing in the environment by one of the available squeezing measures and ask for the maximum output squeezing provided that the input is not squeezed. This could then provide bounds on the output entanglement and maybe relate to channel capacities. Another possible application and one of the primary motivations was to quantify the amount of squeezing necessary for superactivation, which is known to be nonzero [3]. So far however we have no nontrivial new results.

4 Legal statement

The initial idea was proposed by Robert König. In all parts of this work I was significantly involved.

References

[1] Filippo Caruso, Jens Eisert, Vittorio Giovannetti, and Alexander S. Holevo. Multi-mode bosonic gaussian channels. *New Journal of Physics*, 10(8):083030, 2008.

[2] Filippo Caruso, Jens Eisert, Vittorio Giovannetti, and Alexander S. Holevo. Optimal unitary dilation for bosonic gaussian channels. *Phys. Rev. A*, 84:022306, 08 2011.

[3] Daniel Lercher, Géza Giedke, and Michael M. Wolf. Standard super-activation for gaussian channels requires squeezing. *New Journal of Physics*, 15(12):123003, 2013.

On additive Gaussian quantum channels

Martin Idel*[1] and Robert König[1,2]

[1] *Zentrum Mathematik, Technische Universität München, 85748 Garching, Germany*

[2] *Institute for Advanced Studies, Technische Universität München, 85748 Garching, Germany*

August 22, 2016

Abstract

We give necessary and sufficient conditions for a Gaussian quantum channel to have a dilation involving a passive, i.e., number-preserving unitary. We then establish a normal form of such channels: any passively dilatable channel is the result of applying passive unitaries to the input and output of a Gaussian additive channel. The latter combine the state of the system with that of the environment by means of a multi-mode beamsplitter.

1 Introduction

It is a fortunate fact of nature that many physical systems are well-described by a quadratic approximation. Harmonic oscillators are ubiquitous in physics, and are the basis for our understanding of a variety of phenomena in the domain of classical mechanics, electrodynamics, solid state physics, quantum field theory and gravity. Gaussian processes are also essential in probability theory and information theory as a source of non-trivial yet exactly solvable scenarios of interest. Arguably one of the most promiment examples is Shannon's capacity formula for the additive white Gaussian noise (AWGN) channel [10]. The latter constitutes a realistic model for fiberoptic communication. It transforms an analog input signal X (modeled by a random variable on $\mathbb{R}^n$) into the output $Y = X + Z$ by adding an independent centered unit-variance Gaussian random variable Z representing the noise. More generally, Z may be replaced by an arbitrary random variable Z, in which case we refer to this as an additive noise channel.

In quantum mechanics, Gaussian states arise naturally as thermal states of Hamiltonians which are quadratic in the mode operators of a bosonic system. The latter provide

*martin.idel@tum.de

an accurate description of many systems of interest. Restricting to such Hamiltonians, Gaussian channels result whenever a system interacts with an environment in a Gaussian state. A typical example is a channel of the form

$$\mathcal{E}(\rho) = \mathrm{tr}_E \left(U_\lambda (\rho \otimes \rho_E) U_\lambda^* \right) \ , \tag{1}$$

where U_λ is a beamsplitter with transmissivity $\lambda \in [0, 1]$, and ρ_E is a Gaussian state of the environment (see Example 3.2 below). This channel constitutes a natural quantum counterpart of the classical additive noise channel, and, correspondingly, we refer to it as a (quantum) *additive Gaussian noise channel*. In the special case where ρ_E is the thermal state of the harmonic oscillator Hamiltonian, it is also called a thermal noise channel (and is the counterpart of the AWGN channel).

The channel (1) also arises naturally from the viewpoint of resources in e.g., quantum optics. The unitary U_λ obeys a special property: it cannot generate squeezing. More generally, a unitary U acting jointly on n modes of a system and l environment modes is called *passive* if it commutes with the total number operator $\hat{N} = \sum_{k=1}^{n+l} a_k^* a_k$. Here $a_k = (Q_k + iP_k)/\sqrt{2}$ is the usual annihilation operator associated with the k-th mode. The unitary U_λ describing the beamsplitter is an example of such a passive unitary. In fact, a Gaussian unitary is passive if and only if it is the composition of beamsplitters and phase shifters [9]. Thus passive Gaussian unitary operations are experimentally easy to implement.

Considering squeezing as a resource, it is natural to try to separate preexisting squeezing (in the form of a potentially squeezed state of the environment) from evolutions generating squeezing. One is then led to consider the class of *passively dilatable channels*: these are channels possessing a dilation with a passive unitary. Motivated by the decomposition [9] of passive Gaussian unitaries, we ask if passively dilatable channels also have a special structure. The main result of our paper is such a normal form: we establish a close connection between additive channels and the class of passively dilatable channels. That is, any passively dilatable channel is the composition of (i) a passive unitary applied to the input, (ii) an additive Gaussian noise channel and (iii) a passive unitary applied to the output.

Our result thus provides an alternative characterization of quantum additive channels as canonical examples of non-unitary channels which do not generate squeezing. It is a further manifestation, but in a non-unitary context, of the well-known fact that non-linear optical elements are generally required for the generation of squeezed states [1]. We refer to [6] for a recent study of the operational quantification of squeezing, and a more detailed discussion of its role in quantum optics.

Our work also establishes simple necessary and sufficient criteria for deciding when a given passively dilatable channel has a dilation with l environment modes. Our considerations cover all cases, including rank-deficient ones. Using these criteria, we compute the minimal number of required environment modes for a passive dilation to exist. These

results are similar, in spirit, to those of [2, 3], but in contrast to the latter, geared towards characterizing non-squeezing resources. Specifically, [2] constructs a unitary dilation of an arbitrary Gaussian quantum channel, and presents a number of applications to weak degradability. In [3], the minimal number of environment modes required to provide a unitary Gaussian dilation with pure state environment is identified, and bounds for the case of mixed state enviroments are given (see Remarks 3.2 and 3.3 below).

2 Preliminaries

We begin by introducing some of the basic relevant terminology associated with continuous variable quantum information (for longer reviews of the material see for instance [4, 11]). This will also serve to introduce our notation.

2.1 Gaussian states and operations

We consider n-mode bosonic systems with n pairs of quadratures (or modes) given by $R = (Q_1, P_1, Q_2, P_2, \ldots, Q_n, P_n)$, or, equivalently, the annihilation and creation operators

$$a_k = \frac{1}{\sqrt{2}}(Q_k + iP_k) \qquad \text{and} \qquad a_k^* = \frac{1}{\sqrt{2}}(Q_k - iP_k)$$

for $k = 1, \ldots, n$. The commutators

$$[R_j, R_k] = i\sigma_{jk}\mathsf{id} \tag{2}$$

are given by the standard symplectic form

$$\sigma := \bigoplus_{i=1}^{n} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

To simplify notation, it is often convenient to work in the permuted basis $(Q_1, \ldots, Q_m, P_1, \ldots, P_m, Q_{m+1}, \ldots, Q_{m+l}, P_{m+1}, \ldots, P_{m+l})$, where σ takes the form $\sigma = \sigma_{2m} \oplus \sigma_{2l}$ with

$$\sigma_{2k} := \begin{pmatrix} 0_{k\times k} & \mathbb{1}_k \\ -\mathbb{1}_k & 0_{k\times k} \end{pmatrix}.$$

For concreteness, we will henceforth assume that the CCR-relations (2) are realized by unbounded operators acting on the tensor product $\mathcal{H}^{\otimes n}$ where $\mathcal{H} \cong L^2(\mathbb{R})$ is the Hilbert space associated with a single mode. When convenient, we will also use the notation $\mathcal{H}_{A_1\cdots A_n} = \mathcal{H}^{\otimes n}$ to denote multipartite Hilbert spaces.

An important subset of states is given by the *Gaussian states*: such a state ρ is fully characterised by its first and second moments

$$d_k = \mathrm{tr}(\rho R_k) \qquad \text{and} \qquad \gamma_{k\ell} = \mathrm{tr}(\rho\{R_k - d_k\mathsf{id}, R_\ell - d_\ell\mathsf{id}\}) \ ,$$

where $\{A, B\} = AB + BA$ denotes the anticommutator. Here $d \in \mathbb{R}^{2n}$ is the *displacement vector*, whereas the symmetric matrix $\gamma = \gamma^T \in \mathbb{R}^{2n\times 2n}$ is referred to as the *covariance matrix*. By Heisenberg's uncertainty principle, the covariance matrix of any state satisfies the operator inequality

$$\gamma \geq i\sigma_{2n} \ . \tag{3}$$

Conversely, any pair (d, γ) with $d \in \mathbb{R}^{2n}$ and $\gamma = \gamma^T \in \mathbb{R}^{2n\times 2n}$ satisfying (3) uniquely defines a Gaussian n-mode state. As a consequence, we may identify the set of Gaussian states with the set of such pairs.

2.2 Gaussian operations

A quantum operation (or channel) acting on an n-mode system is described by a completely positive trace-preserving map $\Phi : \mathcal{B}(\mathcal{H}^{\otimes n}) \to \mathcal{B}(\mathcal{H}^{\otimes n})$. Here $\mathcal{B}(\mathcal{H}^{\otimes n})$ is the set of bounded linear operators on $\mathcal{H}^{\otimes n}$. Again, the subset of *Gaussian channels* is distinguished by the property that such channels map Gaussian states to Gaussian states. Such a channel is completely characterized by its action on Gaussian states, and the latter has a convenient description: for a Gaussian state ρ with displacement vector d and covariance matrix γ, the Gaussian state $\Phi(\rho)$ resulting from application of the channel is described by the pair (d', γ') obtained from the map

$$\begin{aligned} \gamma &\mapsto X\gamma X^T + Y \\ d &\mapsto Xd + v \ , \end{aligned}$$

where the matrices $X, Y \in \mathbb{R}^{2n\times 2n}$ and the vector $v \in \mathbb{R}^{2n}$ determine the action of the channel. Clearly, $Y = Y^T$ has to be symmetric for this to map covariance matrices to covariance matrices. The map is completely positive if and only if[1] (cf. [4])

$$Y \geq i\sigma_{2n} - iX\sigma_{2n}X^T \ . \tag{4}$$

Conversely, and similarly as for Gaussian states, any triple (X, Y, v) with $Y = Y^T$ symmetric, (X, Y) satisfying (4) and $v \in \mathbb{R}^{2n}$ arbitrary uniquely determines a Gaussian n-mode channel. We will thus identify the set of Gaussian channels with the set of such triples.

In fact, the displacement vector $v \in \mathbb{R}^{2n}$ has no influence on operational properties of the channel such as capacities since it can be changed arbitrarily by applying a displacement operator (a Gaussian unitary) to the output of the channel (see e.g., [4]). In contrast, the matrices (X, Y) determine all important characteristics of the channel. As a consequence, we will henceforth assume that $v = 0$ (as in [2, 3]), and write $\Phi_{X,Y} : \mathcal{B}(\mathcal{H}^{\otimes n}) \to \mathcal{B}(\mathcal{H}^{\otimes n})$ for the Gaussian channel determined by the pair (X, Y).

[1]Note that in [4], the condition is stated with a minus sign, but since $\sigma^T = -\sigma$ and Y is symmetric, this conditions is equivalent.

2.3 Gaussian unitaries and passive unitaries

A Gaussian unitary channel is one of the form $\Phi_{X,0}$ (i.e., $Y = 0$). For such channels, the constraint (4) implies that X preserves the symplectic form (i.e., $X\sigma_{2n}X^T = \sigma_{2n}$), i.e., X is an element of

$$Sp(2n) = \{S \in \mathbb{R}^{2n\times 2n} \mid S\sigma_{2n}S^T = \sigma_{2n}\} \ ,$$

the group of real symplectic matrices. It can be shown that any element $S \in Sp(2n)$ defines a unitary U_S on $\mathcal{H}^{\otimes n}$ such that

$$\Phi_{S,0}(\rho) = U_S \rho U_S^* \ .$$

Furthermore, $S \mapsto U_S$ defines a representation called the metaplectic representation of $Sp(2n)$.

In more physical terms, a Gaussian unitary describes the evolution (for a fixed amount of time) generated by a Hamiltonian H which is quadratic in the creation- and annihilation operations, i.e., one that has the form

$$H = \sum_{j,k=1}^{n} h_{j,k} a_j^* a_k + h.c. \tag{5}$$

A Hamiltonian of the form (5) which commutes with the total number operator, i.e., satisfies

$$[H, \sum_{j=1}^{n} a_j^* a_j] = 0$$

is called *passive*. A passive Hamiltonian generates Gaussian unitaries which are associated with orthogonal symplectic matrices $S \in Sp(2n) \cap O(2n)$, where

$$O(2n) = \{O \in \mathbb{R}^{2n\times 2n} \mid OO^T = \mathbb{1}_{2n}\} \ .$$

We call such Gaussian unitaries *passive*. It can be shown that passive Gaussian unitaries can be realized using beamsplitters and phase shifters only [9].

2.4 On the orthogonal symplectic group

Let us collect a few facts about the group $Sp(2n) \cap O(2n)$. Crucially, there is an isomorphism $U(n) \cong Sp(2n) \cap O(2n)$ between this group and the group

$$U(n) = \{U \in \mathbb{C}^{n\times n} \mid U^\dagger U = \mathbb{1}_n\}$$

of unitary $n \times n$ matrices. For our purposes, it will be convenient to write out this isomorphism for the case of $n + l$ modes (associated with a system and its environment), as follows:

Lemma 2.1. *The map*

$$\begin{array}{rcl} \phi : U(n+l) & \to & Sp(2(n+l)) \cap O(2(n+l)) \\ U = \begin{pmatrix} u_1 & u_2 \\ u_3 & u_4 \end{pmatrix} & \mapsto & \phi(U) = S = \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix} \end{array} \qquad \text{where} \quad s_i = \begin{pmatrix} \operatorname{Re}(u_i) & \operatorname{Im}(u_i) \\ -\operatorname{Im}(u_i) & \operatorname{Re}(u_i) \end{pmatrix}$$

and where $u_1 \in \mathbb{C}^{n\times n}, u_2 \in \mathbb{C}^{n\times l}, u_3 \in \mathbb{C}^{l\times n}, u_4 \in \mathbb{C}^{l\times l}$ *is an isomorphism.*

Proof. The existence of the isomorphism is well-known (see [8]). We need to show that S is symplectic:

$$S\sigma_{2(n+l)}S^T = \begin{pmatrix} s_1\sigma_{2n}s_1^T + s_2\sigma_{2l}s_2^T & s_1\sigma_{2n}s_3^T + s_2\sigma_{2l}s_4^T \\ s_3\sigma_{2n}s_1^T + s_4\sigma_{2l}s_2^T & s_3\sigma_{2n}s_3^T + s_4\sigma_{2l}s_4^T \end{pmatrix}$$

Note that

$$s_i\sigma_{2n}s_j^T = \begin{pmatrix} \operatorname{Re}(u_i)\operatorname{Im}(u_j)^T - \operatorname{Im}(u_i)\operatorname{Re}(u_j)^T & \operatorname{Re}(u_i)\operatorname{Re}(u_j)^T + \operatorname{Im}(u_i)\operatorname{Re}(u_j)^T \\ -(\operatorname{Re}(u_i)\operatorname{Re}(u_j)^T + \operatorname{Im}(u_i)\operatorname{Re}(u_j)^T) & \operatorname{Re}(u_i)\operatorname{Im}(u_j)^T - \operatorname{Im}(u_i)\operatorname{Re}(u_j)^T \end{pmatrix}$$

and that

$$\begin{aligned} \operatorname{Re}(u_i)\operatorname{Re}(u_j)^T + \operatorname{Im}(u_i)\operatorname{Re}(u_j)^T &= \operatorname{Re}(u_i u_j^\dagger) \\ \operatorname{Re}(u_i)\operatorname{Im}(u_j)^T - \operatorname{Im}(u_i)\operatorname{Re}(u_j)^T &= \operatorname{Im}(u_i u_j^\dagger). \end{aligned}$$

Therefore, since U is unitary, it follows that $S\sigma_{2(n+l)}S^T = \sigma_{2(n+l)}$. Similarly,

$$SS^T = \begin{pmatrix} s_1s_1^T + s_2s_2^T & s_1s_3^T + s_2s_4^T \\ s_3s_1^T + s_4s_2^T & s_3s_3^T + s_4s_4^T \end{pmatrix} = \mathbb{1}_{2n}$$

using the unitarity of U. To prove that this is an isomorphism, one then has to consider the inverse map. This is well-defined because any matrix $S \in Sp(2n) \cap O(2n)$ is of the form of the image of the map ϕ in (2.1) (see [8]). □

The following lemma will be an important tool in what follows.

Lemma 2.2. *For any matrix* $X \in \mathbb{R}^{2n\times 2n}$, $[X, \sigma_{2n}] = 0$ *if and only if* X *has the form*

$$X = \begin{pmatrix} A & B \\ -B & A \end{pmatrix}$$

for some matrices $A, B \in \mathbb{R}^{n\times n}$. *In particular, any matrix* $X \in Sp(2n)\cap O(2n)$ *commutes with* σ_{2n}.

Furthermore, any eigenvalue of a matrix of form (2.2) has even multiplicity.

In fact, it can be shown (see [8]) that any two of the three properties $X\sigma_{2n}X^T = \sigma$, $[X, \sigma_{2n}] = 0$ and $XX^T = \mathbb{1}_{2n}$ implies the third, a feature known as the 2-out-of-3 property.

Proof. The proof is straightforward. The fact that this holds for $X \in Sp(2n) \cap O(2n)$ is clear from Lemma 2.1 (specialized to $l = 0$).

For the eigenvalue multiplicity, note that if $v \equiv (v_1, v_2)^T$ with $v_1, v_2 \in \mathbb{R}^n$ is an eigenvector to the eigenvalue λ of X, then $\sigma_{2n}v = (v_2, -v_1)^T$ is an eigenvector to the same eigenvalue and $\sigma_{2n}v \perp v$. Now, if $\{v, \sigma_{2n}v\}^\perp$ contains another eigenvalue $w \in \mathbb{R}^{2n}$ with eigenvalue λ, then $\sigma_{2n}w$ is again an eigenvector of X with eigenvalue λ. We claim that $\{v, \sigma_{2n}v, w, \sigma_{2n}w\}$ is an orthonormal set of eigenvectors to eigenvalue λ. By construction, we have $w \perp \{v, \sigma_{2n}v\}$ and $\sigma_{2n}w \perp w$. Finally, $\sigma_{2n}w \perp v$ as $\langle \sigma_{2n}w, v\rangle = -\langle w, \sigma_{2n}v\rangle = 0$. Iteratively, we can construct an orthonormal basis of every eigenspace, which will necessarily have even multiplicity. □

The next lemma is an extension theorem for orthogonal symplectic matrices:

Lemma 2.3. *Assume that $s_1 \in \mathbb{R}^{2n\times 2n}$ and $s_2 \in \mathbb{R}^{2n\times 2l}$ satisfy*

$$\begin{aligned} s_1\sigma_{2n}s_1^T + s_2\sigma_{2l}s_2^T &= \sigma_{2n} \\ s_1 s_1^T + s_2 s_2^T &= \mathbb{1}_{2n} \ . \end{aligned} \tag{6}$$

Then there are $s_3 \in \mathbb{R}^{2l\times 2n}$ and $s_4 \in \mathbb{R}^{2l\times 2l}$ such that

$$S = \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix} \in Sp(2(n+l)) \cap O(2(n+l)) \ . \tag{7}$$

Furthermore, if S is of the form (7) and

$$S' = \begin{pmatrix} s_1 & s_2 \\ s_3' & s_4' \end{pmatrix} \in Sp(2(n+l)) \cap O(2(n+l)) \ ,$$

then there is an orthogonal symplectic matrix $o \in Sp(2l) \cap O(2l)$ such that

$$S' = \begin{pmatrix} \mathbb{1}_{2n} & 0_{2n\times 2l} \\ 0_{2l\times 2n} & o \end{pmatrix} S \ . \tag{8}$$

Proof. This is essentially saying that one can always extend suitable matrices to orthogonal symplectic matrices. It is clear by symplectic Gram-Schmidt (see [8]) that it is always possible to find s_3, s_4 to construct a symplectic matrix S, which however is not necessarily orthogonal. Therefore, we take the isomorphism to unitary matrices: Since s_1 and s_2 satisfy the relations (6), we can choose u_1, u_2 from the isomorphism in Lemma 2.1. In particular, the matrix $V := \begin{pmatrix} u_1 & u_2 \end{pmatrix}$ fulfills $VV^\dagger = \mathbb{1}_{2n}$, hence we can extend it to a unitary matrix U and use the isomorphism again to find s_3 and s_4. The corresponding S is now orthogonal symplectic by construction.

For the second statement, let $S, S' \in Sp(2(n+l)) \cap O(2(n+l))$ be given by (7) and (8), respectively. Then

$$S'S^T = \begin{pmatrix} \mathbb{1}_{2n} & s_1 s_3^T + s_2 s_4^T \\ s_3' s_1^T + s_4' s_2^T & s_3' s_3^T + s_4' s_4^T \end{pmatrix} \tag{9}$$

by the orthogonality relation (6). But $S'S^T \in Sp(2(n+l)) \cap O(2(n+l))$, hence it follows that

$$S'S^T = \begin{pmatrix} \mathbb{1}_{2n} & 0_{2n\times 2l} \\ 0_{2l\times 2n} & o \end{pmatrix} =: O$$

for some $o \in Sp(2l) \cap O(2l)$. Combining this with (9) immediately gives $O^T S' S^T = \mathbb{1}_{2(n+l)}$. The claim follows by left- and right-multiplying the latter identity with O and S, respectively. □

2.5 Dilations of Gaussian channels

Consider the Gaussian n-mode channel $\Phi_{X,Y}$ as defined in Section 2.1. It is well-known (see [2]) that one can find a Gaussian state ρ_E of $l \leq n$ environment modes and a Gaussian unitary matrix U acting on $n+l$ modes such that $\Phi_{X,Y}$ can be written as

$$\Phi(\rho) = \mathrm{tr}_E(U(\rho \otimes \rho_E)U^*) \ . \tag{10}$$

Note that we do not demand ρ_E to be a pure state (if it is, this is referred to as the Stinespring representation, see Remark 3.3 below). In Eq. (10), $U = U_S$ is the image under the metaplectic representation of a symplectic matrix $S \in Sp(2(n+l))$.

The relationship between S and (X, Y) is obtained by analyzing the action on covariance matrices: if the l-mode Gaussian state ρ_E has covariance matrix γ_E, then the channel's action is given by

$$\gamma \mapsto (S(\gamma \oplus \gamma_E)S^T)_{2n\times 2n} = X\gamma X^T + Y$$

where $(\cdot)_{2n\times 2n}$ means that we restrict to the upper left block of the size $2n \times 2n$. More precisely, writing

$$S = \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix}$$

with $s_1 \in \mathbb{R}^{2n\times 2n}$ and $s_4 \in \mathbb{R}^{2l\times 2l}$, we have

$$\begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix} \begin{pmatrix} \gamma & 0_{2n\times 2l} \\ 0_{2l\times 2n} & \gamma_E \end{pmatrix} \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix}^T = \begin{pmatrix} s_1\gamma s_1^T + s_2\gamma_E s_2^T & * \\ * & * \end{pmatrix}$$

and therefore

$$X\gamma X^T + Y = s_1\gamma s_1^T + s_2\gamma_E s_2^T \tag{11}$$

for all covariance matrices γ. Thus the pair (X, Y) and (s_1, s_2, γ_E) are related by

$$X = s_1 \qquad \text{and} \qquad Y = s_2\gamma_E s_2^T \ . \tag{12}$$

3 Passively dilatable Gaussian channels

Given a Gaussian channel $\Phi_{X,Y}$, we ask if there is passive unitary associated with an element $S \in Sp(2n) \cap O(2n)$ and an (arbitrary) state ρ_E of the environment constituting a dilation of the channel. We shall call any channel with this property *passively dilatable*. Our main result is the following.

Theorem 3.1. *Let $\Phi_{X,Y}$ be an n-mode Gaussian channel. The following conditions are equivalent:*

(i) There exists a passive dilation with l environment modes and $S \in Sp(2(n+l)) \cap O(2(n+l))$.

(ii) The matrices X, Y satisfy $\mathbb{1}_{2n} - XX^T \geq 0$, $[X, \sigma_{2n}] = 0$, $\ker(Y) = \ker(\mathbb{1}_{2n} - XX^T)$ and $2l \geq \operatorname{rank}(\mathbb{1}_{2n} - XX^T)$.

We defer the proof of this theorem to Section 3.2, and first discuss some examples.

Remark 3.1. *Note that if $[X, \sigma_{2n}] = 0$, then $\operatorname{rank}(\mathbb{1}_{2n} - XX^T)$ is even (see Lemma 2.2) and therefore also $\operatorname{rank}(Y)$.*

Example 3.1. *Consider the* classical noise channel *given by $X = \mathbb{1}$ and $Y \geq 0$, $Y \neq 0$. According to Theorem 3.1, this channel is not passively dilatable because the condition $\ker(Y) = \ker(\mathbb{1} - XX^T)$ is not met. A dilation of this channel with two environment modes is given in [5].*

Example 3.2. *Let U_λ be the two-mode beamsplitter with transmissivity $\lambda \in [0, 1]$, i.e., the Gaussian unitary given by the symplectic matrix*

$$S_\lambda = \begin{pmatrix} \sqrt{\lambda} I_2 & \sqrt{1-\lambda} I_2 \\ \sqrt{1-\lambda} I_2 & -\sqrt{\lambda} I_2 \end{pmatrix}$$

with respect to the ordering (Q_1, P_1, Q_2, P_2) of the modes. Let ρ_E be a one-mode Gaussian state with covariance matrix γ_E. Consider a channel of the form

$$\Phi(\rho) = \operatorname{tr}_E U_\lambda(\rho \otimes \rho_E)U_\lambda^* \ . \tag{13}$$

We call this an additive *Gaussian channel.*

Since U_λ is passive, Φ is clearly passively dilatable. To see that the conditions of the theorem are satisfied, observe that

$$X = \sqrt{\lambda}\mathbb{1}_2 \qquad \text{and} \qquad Y = (1-\lambda)\gamma_E \ .$$

Assume that $\lambda \in]0,1[$. Then it is easily verified (using the fact that covariance matrices are positive definite) that the conditions of (ii) *are satisfied for any $l \geq 1$. In particular, the theorem implies that there is a dilation with l modes for all $l \geq 1$. This is consistent with expression* (13)*. The theorem also implies that at least one environment mode is necessary.*

On the other hand, assume that $\lambda = 1$. Then the conditions of (ii) *are satified for any $l \geq 0$, implying the existence of a dilation with no environment modes. Indeed, in this case, the channel is simply the identity channel, with trivial dilation $\Phi(\rho) = \rho$ for all states ρ.*

Finally, consider the case where $\lambda = 0$. Here the conditions (ii) *apply with $l \geq 1$, which is also consistent with* (13)*.*

In most cases, the theorem can be stated in a simpler fashion.

Corollary 3.2. *Let $\Phi_{X,Y}$ be an n-mode Gaussian channel such that X, Y and $\mathbb{1}_{2n} - XX^T$ have full rank. Then there exists a passive dilation with n modes if and only if $\mathbb{1}_{2n} - XX^T \geq 0$ and $[X, \sigma_{2n}] = 0$.*

In fact, we remark that this Corollary can be shown directly by constructing an orthogonal symplectic unitary from $s_1 = X$, $s_2 = (\mathbb{1}_{2n} - XX^T)^{1/2}$ and using the covariance matrix $\gamma_E = s_2^{-1} Y (s_2^{-1})^T$.

3.1 General observations about dilations

We can now make a first step towards proving the theorem:

Lemma 3.3. *Let $\Phi_{X,Y}$ be an n-mode Gaussian channel. Using the notation of equation (10), such a Gaussian channel can be passively dilated with l environment modes if and only if there exists a tuple (s_2, γ_E) with $s_2 \in \mathbb{R}^{2n \times 2l}$, $\gamma_E \in \mathbb{R}^{2l \times 2l}$ and $\gamma_E \geq i\sigma_{2l}$ such that*

$$\begin{aligned} s_2 \sigma_{2l} s_2^T &= \sigma_{2n} - X \sigma_{2n} X^T =: \Sigma \\ s_2 s_2^T &= \mathbb{1}_{2n} - XX^T =: \hat{\Sigma} \\ s_2 \gamma_E s_2^T &= Y \end{aligned} \tag{14}$$

There is a dilation with $s_1 = X$.

Proof. Given a passive dilation of the channel with a matrix $S \in Sp(2(n+l)) \cap O(2(n+l))$, we know that $X\gamma X^T + Y = s_1 \gamma s_1^T + s_2 \gamma_E s_2^T$ by equation (11). Therefore, it must hold that $s_1 = X$ and $s_2 \gamma_E s_2^T = Y$, which is the third equation of (14). In addition, we need that S is symplectic and orthogonal, which means that the following conditions must always hold:

$$\begin{aligned} s_1 \sigma_{2n} s_1^T + s_2 \sigma_{2l} s_2^T &= \sigma_{2n} \\ s_1 s_1^T + s_2 s_2^T &= \mathbb{1}_{2n} \\ s_1 \sigma_{2n} s_3^T + s_2 \sigma_{2l} s_4^T &= 0 \\ s_1 s_3^T + s_2 s_4^T &= 0 \\ s_3 \sigma_{2n} s_3 + s_4 \sigma_{2l} s_4^T &= \sigma_{2l} \\ s_3 s_3^T + s_4 s_4^T &= \mathbb{1}_{2n} \end{aligned}$$

If we plug in $s_1 = X$, the first two conditions are exactly equations (6) so that it is necessary to satisfy system (14) in order to have a passive dilation.

Conversely, using Lemma 2.3, having a solution to (14), we can always choose s_3 and s_4 to extend S to an orthogonal symplectic matrix. □

This lemma implies that proving Theorem 3.1 is equivalent to characterizing the solvability of the system of equations (14). From the fact that $s_2 s_2^T$ is positive semidefinite, it is immediately clear that the system can only be solvable if $\hat{\Sigma} \geq 0$, which is one of the conditions stated in Theorem 3.1. To recover the other conditions, we will need the next lemma:

Lemma 3.4. *In the notation of Lemma 3.3, for any passive dilation of an n-mode passively dilatable Gaussian channel $\Phi_{X,Y}$ we have $\Sigma = \sigma_{2n}\hat{\Sigma}$ and both Σ and $\hat{\Sigma}$ commute with σ_{2n}.*

Proof. By definition, we need $s_2 \sigma_{2l} s_2^T = \Sigma$ and $s_2 s_2^T = \hat{\Sigma}$. Since s_2 is derived from an orthogonal symplectic matrix, it is of the form (see Lemma 2.1)

$$s_2 = \begin{pmatrix} \operatorname{Re}(u_2) & \operatorname{Im}(u_2) \\ -\operatorname{Im}(u_2) & \operatorname{Re}(u_2) \end{pmatrix}.$$

Setting

$$\begin{aligned} \mu &:= \operatorname{Re}(u_2) \operatorname{Re}(u_2)^T + \operatorname{Im}(u_2) \operatorname{Im}(u_2)^T \\ \nu &:= \operatorname{Im}(u_2) \operatorname{Re}(u_2)^T - \operatorname{Re}(u_2) \operatorname{Im}(u_2)^T, \end{aligned}$$

we obtain:

$$\begin{aligned} s_2 s_2^T &= \begin{pmatrix} \mu & \nu \\ -\nu & \mu \end{pmatrix} \stackrel{!}{=} \hat{\Sigma} \\ s_2 \sigma_{2l} s_2^T &= \begin{pmatrix} -\nu & \mu \\ -\mu & -\nu \end{pmatrix} \stackrel{!}{=} \Sigma \end{aligned}$$

Since Σ and $\hat{\Sigma}$ are of the form specified in Lemma 2.2, they commute with σ_{2n}. □

3.2 Proof of Theorem 3.1

3.2.1 Characterization of passively dilatable channels ((i)⇒(ii))

We begin by proving the first part of Theorem 3.1, namely that the stated conditions are necessary:

Lemma 3.5. *Let $\Phi_{X,Y}$ be an n-mode Gaussian channel. The conditions $\mathbb{1}_{2n} - XX^T \geq 0$, $[X, \sigma_{2n}] = 0$ and $2l \geq \operatorname{rank}(\mathbb{1}_{2n} - XX^T)$ are necessary for the existence of a passive dilation of the channel with $2l$ environment modes.*

Proof. By Lemma 3.3, in order for a dilation to exist, the system of equations (14) must be satisfied. In particular, $s_2 s_2^T = \mathbb{1}_{2n} - XX^T$. Due to the fact that $s_2 s_2^T$ is positive semidefinite, $\mathbb{1}_{2n} - XX^T$ must be positive semidefinite. In addition, if $s_2 \in \mathbb{R}^{2n \times 2l}$, then $\operatorname{rank}(s_2 s_2^T) \leq 2l$, which implies that $s_2 s_2^T = \mathbb{1}_{2n} - XX^T$ can only have a solution if $\operatorname{rank}(\mathbb{1}_{2n} - XX^T) \leq 2l$. Finally, for a passive dilation we have $S \in Sp(2n) \cap O(2n)$ by definition. The 2-out-of-3 property of the unitary group (Lemma 2.2) then implies $[S, \sigma_{2(n+l)}] = 0$ and therefore $[s_1, \sigma_{2n}] = 0$. Hence $[X, \sigma_{2n}] = 0$ is a necessary condition as $X = s_1$. □

Lemma 3.6. *Let $\Phi_{X,Y}$ be a Gaussian channel. The condition $\ker(\mathbb{1}_{2n} - XX^T) = \ker(Y)$ is necessary for the existence of a passive dilation of the channel.*

Proof. We suppose that we have found (s_2, γ_E) such that $s_2 s_2^T = \mathbb{1}_{2n} - XX^T$ and $\gamma_E \geq i\sigma_{2l}$ such that $s_2 \gamma_E s_2^T = Y$. First note that for every $y \in \ker(s_2^T)$ we have $s_2 \gamma_E s_2^T y = 0$, hence $y \in \ker(Y)$ or $\ker(s_2^T) \subseteq \ker(Y)$. Now, on the other hand

$$\operatorname{rank}(s_2^T) \geq \operatorname{rank}(s_2 \gamma_E s_2^T) \geq \operatorname{rank}(s_2^+ s_2 \gamma_E s_2^T s_2^{+T})$$

with the pseudoinverse s_2^+ (see Appendix A for definition and basic properties), using that the rank of a product of matrices is always smaller than the rank of its factors. Now note that $s_2^+ s_2 = Q$ is the orthogonal projection onto the range of s_2^T. Since $\gamma_E \geq i\sigma_{2l}$, one can easily see that $\gamma_E \geq 0$ has full rank, which means that there is $\varepsilon > 0$ such that $\gamma_E \geq \varepsilon \mathbb{1}_{2l}$. Then we have that $Q\gamma_E Q \geq \varepsilon Q^2 = \varepsilon Q$, hence

$$\operatorname{rank}(s_2^+ s_2 \gamma_E s_2^T s_2^{+T}) \geq \operatorname{rank}(Q) = \operatorname{rank}(s_2^T)$$

But then, $\operatorname{rank}(Y) = \operatorname{rank}(s_2 \gamma_E s_2^T) = \operatorname{rank}(s_2^T)$ and therefore $\ker(Y) = \ker(s_2^T)$. Finally, since $\ker(s_2) = \operatorname{im}(s_2^T)^\perp$, $\ker(s_2 s_2^T) = \ker(s_2^T)$ and hence $\ker(\mathbb{1}_{2n} - XX^T) = \ker(Y)$ is a necessary condition. □

Lemmas 3.5 and 3.6 show that the conditions stated in Theorem 3.1 are necessary for a passive dilation to exist. This proves the implication (i)⇒(ii).

3.2.2 Existence of unitary dilations ((ii)⇒(i))

We now consider the converse direction, i.e., we assume that (X, Y) satisfy the conditions stated in (ii) of Theorem 3.1 and show that these are sufficient to imply the existence of a passive dilation (as in (i)).

Lemma 3.7. *Let $\Phi_{X,Y}$ be an n-mode Gaussian channel satisfying $2l \geq \mathrm{rank}(\mathbb{1} - XX^T)$, $\mathbb{1}_{2n} - XX^T \geq 0$, $\ker(\mathbb{1}_{2n} - XX^T) = \ker(Y)$ and $[\sigma_{2n}, X] = 0$. Then there is a passive dilation with l environment modes.*

Proof. From the spectral theorem, it is known that if $[A, B] = 0$ and A is normal, then also $[P_{\lambda(A)}, B] = 0$ for any spectral projection $P_{\lambda(A)}$ of A and therefore $[A^{1/2}, B] = 0$, where $A^{1/2}$ denotes the unique positive square root of A. Define $\hat{\Sigma} = \mathbb{1} - XX^T \geq 0$ and $\Sigma = \sigma_{2n} - X\sigma_{2n}X^T$. Using $[\sigma_{2n}, X] = 0$, we have $\sigma_{2n}\hat{\Sigma} = \Sigma$ and $\sigma_{2n}\hat{\Sigma} = \hat{\Sigma}\sigma_{2n}$, i.e. $\hat{\Sigma}$ commutes with σ_{2n}. Therefore

$$[\hat{\Sigma}^{1/2}, \sigma_{2n}] = 0 \tag{15}$$

and thus (see Lemma 2.2) the matrix $\hat{\Sigma}^{1/2}$ is of the form

$$\hat{\Sigma}^{1/2} = \begin{pmatrix} \mu & \nu \\ -\nu & \mu \end{pmatrix} \tag{16}$$

and

$$\Sigma = \hat{\Sigma}^{1/2}\sigma_{2n}\hat{\Sigma}^{1/2} \ . \tag{17}$$

Furthermore, by definition of the square root (and since $\hat{\Sigma}$ is symmetric), we have

$$(\hat{\Sigma}^{1/2})^T = \hat{\Sigma}^{1/2} \ . \tag{18}$$

We divide the proof into three cases:

1. Consider the case where $l = n$. We proceed by constructing a pair (s_2, γ_E) satisfying the conditions of Lemma 3.3, implying the existence of a passive dilation of (X, Y).

   Setting $s_2 = \hat{\Sigma}^{1/2}$ we have $s_2 s_2^T = \hat{\Sigma}$ and $s_2\sigma_{2n}s_2^T = \Sigma$. Thus the first two conditions of (14) (Lemma 3.3) are satisfied, and it remains to construct a covariance matrix γ_E satisfying $s_2\gamma_E s_2^T = Y$. Let s_2^+ be the Moore-Penrose pseudoinverse of s_2. We set

   $$\gamma_E = s_2^+ Y s_2^{+T} + P_{\ker(s_2)} \ , \tag{19}$$

   where $P_{\ker(s_2)}$ is the projection onto $\ker(s_2)$. Then

   $$s_2\gamma_E s_2^T = s_2 s_2^+ Y s_2^{+T} s_2^T = P_{\mathrm{im}(s_2)} Y P_{\mathrm{im}(s_2)}^T \tag{20}$$

where we used the fact that $s_2 P_{\ker(s_2)} = 0$ in the first identity and the properties of the Moore-Penrose-pseudoinverse (Lemma A.1) in the second step, and where we denoted the projection onto the range $\mathrm{im}(s_2)$ of s_2 by $P_{\mathrm{im}(s_2)}$.

Since $\mathrm{im}(Y) = \mathrm{im}(\hat{\Sigma})$ by assumption and $\mathrm{im}(\hat{\Sigma}) = \mathrm{im}(s_2 s_2^T) \subset \mathrm{im}(s_2)$, we have $P_{\mathrm{im}(s_2)} Y = Y$ and since $Y = Y^T$ is symmetric, it follows that

$$P_{\mathrm{im}(s_2)} Y P_{\mathrm{im}(s_2)}^T = Y \ .$$

Inserting this into (20) yields $s_2 \gamma_E s_2^T = Y$, as claimed (cf. (14)).

We next verify that γ_E is a valid covariance matrix. This is done using equation (4): we have

$$\begin{aligned} s_2^+ Y s_2^{+T} + P_{\ker(s_2)} &\geq (\hat{\Sigma}^{1/2})^+ (i\sigma_{2n} - iX\sigma_{2n}X^T)(\hat{\Sigma}^{1/2})^{+T} + P_{\ker(s_2)} \\ &= i(\hat{\Sigma}^{1/2})^+ \Sigma (\hat{\Sigma}^{1/2})^{+T} + P_{\ker(s_2)} \\ &= i(\hat{\Sigma}^{1/2})^+ \hat{\Sigma}^{1/2} \sigma_{2n} (\hat{\Sigma}^{1/2})^T (\hat{\Sigma}^{1/2})^{+T} + P_{\ker(s_2)} \\ &= i P_{\mathrm{im}(s_2^T)} \sigma_{2n} P_{\mathrm{im}(s_2^T)}^T + P_{\ker(s_2)} \end{aligned}$$

where we used (17) in the third step and introduced the projection $P_{\mathrm{im}(\hat{\Sigma}^{1/2})}$ onto the range of the symmetric matrix $s_2^T = (\hat{\Sigma}^{1/2})^T$ in the fourth step. Since s_2 is symmetric we have

$$P_{\ker(s_2)} = \mathbb{1}_{2n} - P_{\ker(s_2)^\perp} = \mathbb{1}_{2n} - P_{\mathrm{im}(s_2)} = \mathbb{1}_{2n} - P_{\mathrm{im}(s_2^T)} \ .$$

Using $\mathbb{1}_{2n} \geq i\sigma_{2n}$, we thus obtain

$$s_2^+ Y s_2^{+T} + P_{\ker(s_2)} \geq i P_{\mathrm{im}(s_2)} \sigma_{2n} P_{\mathrm{im}(s_2)}^T + i(\mathbb{1} - P_{\mathrm{im}(s_2)}) \sigma_{2n} (\mathbb{1} - P_{\mathrm{im}(s_2)}^T) = i\sigma_{2n}.$$

Here we used that $P_{\mathrm{im}(s_2)}$ commutes with σ_{2n} as a consequence of (15) and the fact that it is the projection onto the range of $\hat{\Sigma}^{1/2}$. This concludes the proof that (19) defines a valid covariance matrix.

2. The claim for $l > n$ then follows immediately by using the established claim for $l = n$: since $2n \geq \mathrm{rank}(\mathbb{1} - XX^T)$, there is a dilation $\Phi(\rho) = \mathrm{tr}_E(U(\rho \otimes \rho_E)U^*)$ involving n environment modes. For an arbitrary $(l-n)$-mode state $\rho_{\tilde{E}}$, we then have

$$\Phi(\rho) = \mathrm{tr}_{E\tilde{E}}((U \otimes \mathbb{1}_{\tilde{E}})(\rho \otimes (\rho_E \otimes \rho_{\tilde{E}}))(U \otimes \mathbb{1}_{\tilde{E}})^*) \ ,$$

providing us with a passive dilation using l modes.

3. Finally, consider the case $l < n$. Then we have $\operatorname{rank}(\hat{\Sigma}) \leq 2l$ by assumption. We can assume that $\operatorname{rank}(\hat{\Sigma}) = 2l$ without loss of generality (cf. Remark 3.1), since otherwise we can proceed as in step (2) to increase the number of environment modes.

   We exploit the form (16) of $\hat{\Sigma}^{1/2}$. Because $\hat{\Sigma}^{1/2}$ is symmetric (cf. (18)), we have $\mu^T = \mu$ and $\nu^T = -\nu$, hence the complex matrix $\hat{\Sigma}_{\mathbb{C}}^{1/2} := \mu + i\nu$ is Hermitian. We can thus diagonalise $\hat{\Sigma}_{\mathbb{C}}^{1/2}$ with a unitary $u \in U(n)$, which corresponds (see Lemma 2.1) to a matrix $o \in Sp(2n) \cap O(2n)$ such that $u\hat{\Sigma}_{\mathbb{C}}^{1/2}u^\dagger$ corresponds to $o\hat{\Sigma}^{1/2}o^T$. In particular,

$$o\hat{\Sigma}^{1/2}o^T = \operatorname{diag}(d_1, \ldots, d_l, \underbrace{0, \ldots, 0}_{n-l}, d_1, \ldots, d_l, \underbrace{0, \ldots, 0}_{n-l})$$

   This implies that $\hat{\Sigma}^{1/2}o^T$ has the form

$$\hat{\Sigma}^{1/2}o^T = \begin{pmatrix} A & 0_{2n\times(n-l)} & B & 0_{2n\times(n-l)} \end{pmatrix}$$

   for two matrices $A, B \in \mathbb{R}^{2n\times l}$. We now define s_2 to be the matrix where we erase the $2(n-l)$ zero columns, i.e. we choose

$$s_2 = \begin{pmatrix} A & B \end{pmatrix} \in \mathbb{R}^{2n\times 2l} .$$

   By construction, this implies that $s_2 s_2^T = \hat{\Sigma}$ as before, and since o^T commutes with σ_{2n} (Lemma 2.2), we also have $s_2\sigma_{2l}s_2^T = \sigma_{2n}s_2 s_2^T$. Again, γ_E is defined as in the case $l = n$ by (19) and we have a solution to the system (14) with $\gamma_E \geq i\sigma_{2l}$ by the same argument as in case 1.

□

3.3 Minimal dilations

In the following, we show that under the assumptions of Corollary 3.2, any pair of dilations are related by orthogonal symplectic matrices acting on the environment. More generally, let us define a *minimal dilation* as one with the least number of environment modes. We then have the following uniqueness property of minimal dilations.

Theorem 3.8. *Let $\Phi_{X,Y}$ be a passively dilatable n-mode Gaussian channel. Then*

(i) A dilation is minimal if and only if $l = \frac{1}{2}\operatorname{rank}(Y)$. There is a minimal dilation given by the construction of Theorem 3.1.

(ii) Let

$$S = \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix} \tag{21}$$

be the orthogonal symplectic matrix describing the passive Gaussian unitary associated with a minimal dilation. Then $\operatorname{rank}(s_2) = 2l = \operatorname{rank} Y$*. In particular,* $s_2 \in \mathbb{R}^{2n\times 2l}$ *is injective.*

(iii) Consider two minimal dilations

$$\Phi_{X,Y}(\rho) = \operatorname{tr}_E U(\rho \otimes \rho_E)U^* = \operatorname{tr}_E U'(\rho \otimes \rho'_E)U'^* \ ,$$

of $\Phi_{X,Y}$*, where* U, U' *are passive Gaussian unitaries on* $\mathcal{H}^{\otimes(n+l)}$*. Then there are two passive Gaussian unitaries* $\tilde{V}, V$ *on* $\mathcal{H}^{\otimes l}$ *such that*

$$U' = (\mathbb{1}_{\mathcal{H}^{\otimes n}} \otimes \tilde{V})U(\mathbb{1}_{\mathcal{H}^{\otimes n}} \otimes V) \qquad \text{and} \qquad \rho'_E = V^*\rho_E V \ .$$

Note that a statement analogous to (iii) was given in [2, Appendix D] for general (non-passive) dilations.

Remark 3.2. *Let us compare these statements to the results of [2, 3]. For a channel* $\Phi_{X,Y}$*, let* $l_{\min}^{mixed}(\Phi_{X,Y})$ *denote the minimal number of environment modes such that a dilation with a (potentially mixed) state of the environment exists. By explicit construction, it was shown in [2] (see also [3, Section 2]) that* $l_{\min}^{mixed}(\Phi_{X,Y}) \leq 2n - \operatorname{rank}(\Sigma)/2$*, where* Σ *is defined by* (14)*. This result was later improved to*

$$l_{\min}^{mixed}(\Phi_{X,Y}) \leq \operatorname{rank}(Y) - \operatorname{rank}(\Sigma)/2 \tag{22}$$

in [3], and this is conjectured to be optimal (a matching lower bound is not known, but see Remark 3.3). To compare to our results, assume that $\Phi_{X,Y}$ *is passively dilatable. Let* $l_{\min,passive}^{mixed}(\Phi_{X,Y})$ *denote the minimal number of environment modes such that a dilation with a passive unitary exists. By definition, we clearly have*

$$l_{\min}^{mixed}(\Phi_{X,Y}) \leq l_{\min,passive}^{mixed}(\Phi_{X,Y}) \ .$$

According to Theorem 3.8, we have

$$l_{\min,passive}^{mixed}(\Phi_{X,Y}) = \frac{1}{2}\operatorname{rank} Y \ . \tag{23}$$

But since $\operatorname{rank}(Y) \geq \operatorname{rank}(\Sigma)$ *(see e.g., [3, Eq. (10)] – this follows immediately from the positivity condition* (4)*), this means that*

$$l_{\min,passive}^{mixed}(\Phi_{X,Y}) = \operatorname{rank} Y - \frac{1}{2}\operatorname{rank} Y \leq \operatorname{rank}(Y) - \operatorname{rank}(\Sigma)/2 \ .$$

Thus our result is consistent with (22)*. We emphasize that in contrast to the case where passivity is not imposed on the dilating unitary, the exact minimal number* $l_{\min,passive}^{mixed}(\Phi_{X,Y})$ *of environment modes is known, i.e., given by expression* (23)*.*

Remark 3.3. *The authors of [2, 3] also consider dilations where the state ρ_E is pure. These are referred to as Stinespring dilations. Correspondingly, they consider the minimal number $l_{\min}^{pure}(\Phi_{X,Y})$ of environment modes for a Stinespring dilation with a pure Gaussian environment state ρ_E to exist. Imposing Gaussianity here is crucial to get a non-trivial problem, since any mixed state can be purified with only a single additional mode otherwise. By definition, we clearly have $l_{\min}^{mixed}(\Phi_{X,Y}) \leq l_{\min}^{pure}(\Phi_{X,Y})$. Improving an upper bound of [2], and by providing a new lower bound, the identity*

$$l_{\min}^{pure}(\Phi_{X,Y}) = \operatorname{rank}(Y - i\Sigma)$$

was shown in [3]. We have not considered the analogous question for passive dilations, since our focus is on establishing an equivalence with additive Gaussian channels (see Theorem 4.1). At least in one direction, the analysis of [3, Appendix B] should be useful: here the minimal number of modes needed to find a Gaussian purification of a generic multimode Gaussian state is computed.

Proof of Theorem 3.8. Statement (ii) of Theorem 3.1 implies that there is a dilation with $l = \frac{1}{2}\operatorname{rank}(Y)$ environment, and this number is minimal. This proves statement (i).

To prove statement (ii), fix a minimal dilation with orthogonal symplectic matrix S and covariance matrix γ_E. By (i), the number of environment modes is $\ell = \frac{1}{2}\operatorname{rank} Y$, i.e., $s_2 \in \mathbb{R}^{2n\times\operatorname{rank}(Y)}$ and $\gamma_E \in \mathbb{R}^{\operatorname{rank} Y\times\operatorname{rank} Y}$. By the minimality and (12), we have $2l = \operatorname{rank}(Y) = \operatorname{rank}(s_2\gamma_E s_2^T)$, but since $\gamma_E \geq i\sigma_{2l}$, the covariance matrix γ_E is full rank and it follows that $\operatorname{rank}(s_2) = 2l$. In particular, this implies that $s_2 \in \mathbb{R}^{2n\times 2l}$ is injective.

Finally, we can prove statement (iii): Consider two minimal dilations of $\Phi_{X,Y}$ with orthogonal symplectic matrices

$$S = \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix} \qquad \text{and} \qquad S' = \begin{pmatrix} s_1' & s_2' \\ s_3' & s_4' \end{pmatrix}$$

and covariance matrices γ_E and γ_E', respectively. In particular, $s_2, s_2' \in \mathbb{R}^{2n\times 2l}$ and

$$s_1 = s_1' = X \tag{24}$$

by (12). Using the orthogonality of S and S' (in the form (6)) therefore gives

$$s_2 s_2^T = s_2' s_2'^T \ . \tag{25}$$

Since s_2 is injective, $s_2^+ s_2 = \mathbb{1}_{2l}$ by the properties of the pseudoinverse. Multiplying (25) from the left by s_2^+ therefore gives $s_2^T = s_2^+ s_2' s_2'^T$ and multiplying this from the right with s_2^{+T} yields $s_2^T s_2^{+T} = s_2^+ s_2' s_2'^T s_2^{+T}$ which is equivalent to

$$s_2^+ s_2' (s_2^+ s_2')^T = \mathbb{1}_{2l} \ .$$

Hence

$$s_2^+ s_2' =: o \in O(2l) \tag{26}$$

is orthogonal. Multiplying Eq. (26) from the left by s_2 and using that $s_2 s_2^+ = P_{\mathrm{range}(s_2)}$ is the projection onto the range of s_2 we obtain $P_{\mathrm{range}(s_2)} s_2' = s_2 o$, hence

$$s_2' = s_2 o \tag{27}$$

because $P_{\mathrm{range}(s_2)} s_2' = s_2'$. The latter identity follows from the fact that the images of s_2 and s_2' coincide as a consequence of the assumption $s_2 s_2^T = s_2' s_2'^T$ and the fact that s_2^T and $s_2'^T$ are surjective (since s_2, s_2' are injective, as argued above).

Furthermore, using the symplecticity condition (6), we have

$$s_2 \sigma_{2l} s_2^T = s_2' \sigma_{2l} s_2'^T = s_2 o \sigma_{2l} o^T s_2^T \tag{28}$$

Since s_2 is minimal it is injective and hence s_2^T is surjective. Because of the injectivity of s_2 and the surjectivity of s_2^T, Eq. (28) implies

$$\sigma_{2l} = o \sigma_{2l} o^T \ ,$$

i.e., o is orthogonal symplectic, $o \in O(2l) \cap Sp(2l)$. Similarly, $Y = s_2 \gamma_E s_2^T = s_2' \gamma_E' s_2'^T$ by assumption, we have

$$\gamma_E' = o^T \gamma_E o \ . \tag{29}$$

using once again the injectivity of s_2 and s_2' (and correspondingly, the surjectivity of s_2^T and s_2^T).

Finally, we claim that S and S' only differ by an orthogonal symplectic matrix applied to the environment modes. Indeed, it follows from (24) and (27) that

$$S \begin{pmatrix} \mathbb{1}_{2n} & 0_{2n\times 2l} \\ 0_{2l\times 2n} & o \end{pmatrix} = \begin{pmatrix} s_1' & s_2' \\ s_3'' & s_4'' \end{pmatrix}$$

for some matrices $s_3'' \in \mathbb{R}^{2l\times 2n}$ and $s_4'' \in \mathbb{R}^{2l\times 2l}$. The second part of Lemma 2.3 thus implies that there is an orthogonal symplectic matrix $o' \in Sp(2l) \cap O(2l)$ acting on the l environment modes such that

$$\begin{pmatrix} \mathbb{1}_{2n} & 0_{2n\times 2l} \\ 0_{2l\times 2n} & o' \end{pmatrix} S \begin{pmatrix} \mathbb{1}_{2n} & 0_{2n\times 2l} \\ 0_{2l\times 2n} & o \end{pmatrix} = S' \ . \tag{30}$$

Combining (30) with (29) yields the claim. □

3.4 Passive channels

To conclude this section, we combine Theorem 3.1 and Theorem 3.8 to characterize passive channels. The latter are defined by having a dilation with a passive unitary U and an environment state ρ_E which is also passive. Here passivity of a state ρ_E is defined physically by the condition that ρ_E is the Gibbs state of a passive Hamiltonian H at some inverse temperature β, i.e., $\rho_E = e^{-\beta H}/\operatorname{tr}(e^{-\beta H})$. Mathematically, passivity of a state ρ_E is equivalent to the statement that its covariance matrix γ_E satisfies

$$[\gamma_E, \sigma_{2l}] = 0 \tag{31}$$

as argued in [7]. In other words, a passive channel is one which has no "hidden" squeezing: both the system-environment interaction and the state of the environment are associated with passive Hamiltonians. We have the following simple characterization of such channels:

Corollary 3.9. *Let $\Phi_{X,Y}$ be a passively dilatable Gaussian channel. Then the following are equivalent:*

(i) $[Y, \sigma_{2n}] = 0$.

(ii) $\Phi_{X,Y}$ *is passive.*

Proof. Suppose $\Phi_{X,Y}$ is passively dilatable. We first remark that any orthogonal symplectic matrix S as in (21) satisfies

$$s_2 \sigma_{2l} = \sigma_{2n} s_2 \ . \tag{32}$$

Indeed, this follows immediately using the block structure of S and $\sigma_{2(n+l)} = \sigma_{2n} \oplus \sigma_{2l}$ by taking the upper right block matrix of the identity $[S, \sigma_{2(n+l)}] = 0$.

We prove the two implications: (i)⇒(ii): Assume that $[Y, \sigma_{2n}] = 0$. Consider the minimal dilation constructed in Theorem 3.8, with orthogonal symplectic matrix S as in (21) and an environment state of ℓ modes with covariance matrix γ_E given by expression (19). According to Theorem 3.8, s_2 is injective, hence $\ker(s_2) = \{0\}$ and thus $\gamma_E = s_2^+ Y s_2^{+T}$. We will show that γ_E satisfies (31), which implies that $\Phi_{X,Y}$ can be passively dilated with a passive environment state ρ_E.

We use (32) to establish the identity

$$\sigma_{2l} s_2^+ = s_2^+ \sigma_{2n} P_{\mathrm{range}(s_2)} \ . \tag{33}$$

Indeed, we have

$$s_2^+ \sigma_{2n} P_{\mathrm{range}(s_2)} - \sigma_{2l} s_2^+ = s_2^+ \sigma_{2n} s_2 s_2^+ - \sigma_{2l} s_2^+ s_2 s_2^+$$

where we used the fact that $s_2 s_2^+ = P_{\mathrm{range}(s_2)}$ and $(s_2^+ s_2)s_2^+ = P_{\mathrm{range}(s_2^T)} s_2^+ = s_2^+$ by the properties of the pseudoinverse and the fact that s_2^T is surjective (as s_2 is injective). That is,

$$\begin{aligned} s_2^+ \sigma_{2n} P_{\mathrm{range}(s_2)} - \sigma_{2l} s_2^+ &= (s_2^+ \sigma_{2n} s_2 - \sigma_{2l} s_2^+ s_2) s_2^+ \\ &= (s_2^+ s_2 \sigma_{2l} - \sigma_{2l} s_2^+ s_2) s_2^+ \\ &= (P_{\mathrm{range}(s_2^T)} \sigma_{2l} - \sigma_{2l} P_{\mathrm{range}(s_2^T)}) s_2^+ = 0 \end{aligned}$$

where we used (32) in the second step and the fact that s_2^T is surjective (and thus $P_{\mathrm{range}(s_2^T)} = \mathbb{1}_{2l}$) in the last step. This establishes (33).

We will also need the transpose of (33), which reads

$$s_2^{+T} \sigma_{2l} = P_{\ker(s_2^T)^\perp} \sigma_{2n} s_2^{+T} \tag{34}$$

because $P^T_{\mathrm{range}(s_2)} = P_{\ker(s_2^T)^\perp}$. We can then compute

$$\begin{aligned} \sigma_{2l}\gamma_E &= \sigma_{2l} s_2^+ Y s_2^{+T} \\ &= s_2^+ \sigma_{2n} P_{\mathrm{range}(s_2)} Y s_2^{+T} && \text{by (33)} \\ &= s_2^+ \sigma_{2n} Y s_2^{+T} && \text{because } Y = s_2 \gamma_E s_2^T \\ &= s_2^+ Y \sigma_{2n} s_2^{+T} && \text{by the assumption } [Y, \sigma_{2n}] = 0 \\ &= s_2^+ Y P_{\ker(s_2^T)^\perp} \sigma_{2n} s_2^{+T} && \text{since } Y = s_2 \gamma_E s_2^T \\ &= s_2^+ Y s_2^{+T} \sigma_{2l} && \text{by (34)} \\ &= \gamma_E \sigma_{2l} \ . \end{aligned}$$

Thus $[\gamma_E, \sigma_{2l}] = 0$, as claimed.

(ii)⇒(i): Suppose $\Phi_{X,Y}$ is passive. Assume S is an orthogonal symplectic matrix and γ_E a covariance matrix of a passive state such that S and γ_E define a dilation of the channel $\Phi_{X,Y}$. Then $Y = s_2 \gamma_E s_2^T$ and thus

$$\begin{aligned} \sigma_{2n} Y &= s_2 \sigma_{2l} \gamma_E s_2^T && \text{by (32)} \\ &= s_2 \gamma_E \sigma_{2l} s_2^T && \text{because } \rho_E \text{ is passive, that is, (31)} \\ &= s_2 \gamma_E s_2^T \sigma_{2n} && \text{by the transpose of (32)} \\ &= Y \sigma_{2n} \ , \end{aligned}$$

hence $[Y, \sigma_{2n}] = 0$ as claimed. □

4 Passively dilatable channels are additive channels

Consider a (one-mode) channel of the form

$$\Phi(\rho) = V \left(\mathrm{tr}_E \, U_\lambda (W \rho W^* \otimes \rho_E) U_\lambda^* \right) V^* \ ,$$

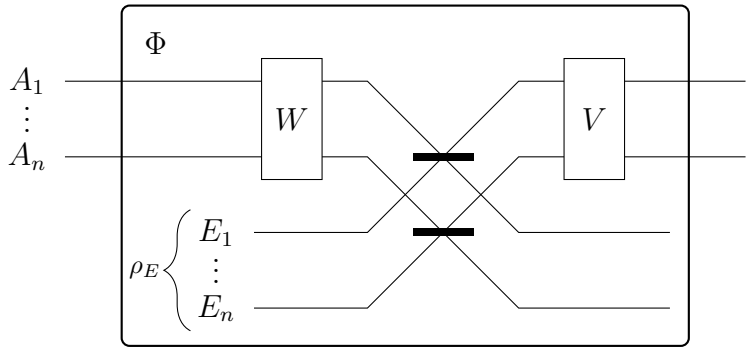

Figure 1: This figure shows how a general passively dilatable channel can be understood as an additive Gaussian channel composed with passive unitaries (two modes are drawn completely). This defines a normal form of passively dilatable channels.

where U_λ is the beamsplitter of transmissivity λ (see Example 3.2) and V, W are passive Gaussian (one-mode) unitaries. That is, Φ is obtained by applying passive unitaries to the input and output of an additive Gaussian channel. Since $\Phi(\rho) = \mathrm{tr}_E(U(\rho \otimes \rho_E)U^*)$ for $U = (V \otimes \mathbb{1}_E)U_\lambda(W \otimes \mathbb{1}_E)$, this channel is passively dilatable. Here we show the converse: any passively dilatable is equivalent (up to passive unitaries) to a (multi-mode) additive Gaussian channel. The following result is illustrated in Fig. 1.

Theorem 4.1. *Let $\Phi : \mathcal{B}(A_1 \dots A_n) \to \mathcal{B}(A_1 \cdots A_n)$ be a passively dilatable n-mode Gaussian channel. Then there is an n-mode Gaussian state $\rho_E = \rho_{E_1 \cdots E_n}$, n-mode Gaussian unitaries V, W and transmissivities $\lambda = (\lambda_1, \dots, \lambda_n) \in [0,1]^n$ such that for the multi-mode beamsplitter $U_\lambda = U_{\lambda_1}^{A_1 E_1} \otimes \cdots \otimes U_{\lambda_n}^{A_n E_n}$, we have*

$$\Phi(\rho) = V\,(\mathrm{tr}_E\, U_\lambda(W\rho W^* \otimes \rho_E)U_\lambda^*)\,V^* \qquad \textit{for all states } \rho\ .$$

Proof. Assume that $\Phi = \Phi_{X,Y}$ is specified by the pair (X, Y) of matrices. As in the proof of Theorem 3.1, consider $l = n$. Let (S, γ_E) be the dilation constructed in case 1 of the proof of the theorem, i.e., $S = \begin{pmatrix} s_1 & s_2 \\ s_3 & s_4 \end{pmatrix}$ satisfies

$$s_1 = X \qquad \text{and} \qquad s_2 = \hat{\Sigma}^{1/2} = (\mathbb{1} - XX^T)^{1/2} \tag{35}$$

and the covariance matrix γ_E is given by the expression (19). Since $[X, \sigma_{2n}] = 0$, we can decompose X as in Lemma 2.2. Let $D = (G_1 + iG_2)(X_1 + iX_2)(F_1 + iF_2)$ be the singular value decomposition of the complex matrix $X_1 + iX_2$. The matrix D is nonnegative but not necessarily full rank. By definiton and the isomorphism of Lemma 2.1, the unitaries $G_1 + iG_2$ and $F_1 + iF_2$ define passive symplectic elements $F, G \in Sp(2n) \cap O(2n)$. Define

$$\tilde{S} = \begin{pmatrix} G & 0 \\ 0 & \mathbb{1}_{2n} \end{pmatrix} S \begin{pmatrix} F & 0 \\ 0 & G^T \end{pmatrix} = \begin{pmatrix} Gs_1F & Gs_2G^T \\ s_3F & s_4G^T \end{pmatrix} =: \begin{pmatrix} \tilde{s}_1 & \tilde{s}_2 \\ \tilde{s}_3 & \tilde{s}_4\ . \end{pmatrix} \tag{36}$$

With (35) we obtain

$$\begin{array}{rcccl} \tilde{s}_1 & = & GXF & = & D \oplus D \\ \tilde{s}_2 & = & G(\mathbb{1}_{2n} - XX^T)^{1/2}G^T & = & (\mathbb{1}_{2n} - D^2 \oplus D^2)^{1/2} \ . \end{array} \tag{37}$$

Here we exploited that XX^T is equivalent to $(X_1+iX_2)(X_1+iX_2)^\dagger = (G_1+iG_2)^\dagger D^2(G_1+iG_2)$ under the isomorphism and hence $\mathbb{1}_{2n} - XX^T = G^T(\mathbb{1}_{2n} - D^2 \oplus D^2)G$. Since G is orthogonal we have $(\mathbb{1}_{2n} - XX^T)^{1/2} = G^T(\mathbb{1}_{2n} - D^2 \oplus D^2)^{1/2}G$.

We conclude from (36) that

$$s_1 = G^T\tilde{s}_1F^T \qquad \text{and} \qquad s_2 = G^T\tilde{s}_2G \ ,$$

i.e., the action of the channel on a covariance matrice γ is given by (cf. (11))

$$X\gamma X^T + Y = G^T\tilde{s}_1F^T\gamma F\tilde{s}_1^TG + G^T\tilde{s}_2G\gamma_E G^T\tilde{s}_2^TG \ .$$

Clearly, this means that the channel can be written as the composition

$$\Phi = \Phi_{s_1,s_2\gamma_E s_2^T} = \Phi_{G^T,0} \circ \Phi_{\tilde{s}_1,\tilde{s}_2\tilde{\gamma}_E\tilde{s}_2^T} \circ \Phi_{F^T,0} \ ,$$

where $\tilde{\gamma}_E = G\gamma_E G^T$ is a valid covariance matrix. It is clear from (37) and the fact that $(\tilde{S},\tilde{\gamma}_E)$ give a dilation that $\Phi_{\tilde{s}_1,\tilde{s}_2\tilde{\gamma}_E\tilde{s}_2^T}$ is an additive noise channel, hence the claim follows. □

A The Moore-Penrose pseudoinverse

In this appendix, we collect a few well-known facts about the Moore-Penrose pseudoinverse. Let $A \in \mathbb{R}^{k\times m}$ be a not necessarily invertible matrix. Using the singular value decomposition, we can find unitaries $U \in U(k), V \in U(m)$ and a diagonal matrix $D \in \mathbb{R}^{k\times m}$ with $A = UDV$. Define $A^+ = V^\dagger D^+ U^\dagger$ with $D^+ \in \mathbb{R}^{m\times k}$ and $D^+_{ii} = \frac{1}{D_{ii}}$ for all $D_{ii} \neq 0$ and zero otherwise. Then A^+ is called the *Moore-Penrose pseudoinverse.*

Lemma A.1. *Let $A \in \mathbb{R}^{k\times m}$ and let A^+ be its pseudoinverse. Then:*

1. *$P = AA^+$ is the orthogonal projection onto the range of A.*
2. *$Q = A^+A$ is the orthogonal projection onto the range of A^T.*

A proof can be found in any introductory book on linear algebra.

Acknowledgements

RK is supported by the Technische Universität München - Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement no. 291763. He also gratefully acknowledges support by DFG project no. KO5430/1-1. MI is supported by the Studienstiftung des deutschen Volkes.

References

[1] Samuel L. Braunstein. Squeezing as an irreducible resource. *Phys. Rev. A*, 71:055801, May 2005.

[2] Filippo Caruso, Jens Eisert, Vittorio Giovannetti, and Alexander S. Holevo. Multi-mode bosonic Gaussian channels. *New Journal of Physics*, 10(8):083030, 2008.

[3] Filippo Caruso, Jens Eisert, Vittorio Giovannetti, and Alexander S. Holevo. Optimal unitary dilation for bosonic Gaussian channels. *Phys. Rev. A*, 84:022306, Aug 2011.

[4] Jens Eisert and Michael M. Wolf. Gaussian quantum channels. In N. J. Cerf, G. Leuchs, and E. S. Polzik, editors, *Quantum Information with continuous variables of atoms and light*. World Scientific, USA, 2007.

[5] Alexander S. Holevo. One-mode quantum Gaussian channels: Structure and quantum capacity. *Problems of Information Transmission*, 43(1):1–11, 2007.

[6] Martin Idel, Daniel Lercher, and Michael M. Wolf. An operational measure for squeezing. arXiv:1607.00873.

[7] Daniel Lercher, Géza Giedke, and Michael M. Wolf. Standard super-activation for Gaussian channels requires squeezing. *New Journal of Physics*, 15(12):123003, 2013.

[8] Dusa McDuff and Dietmar Salamon. *Introduction to Symplectic Topology*. Oxford Science Publications, 1998.

[9] Michael Reck, Anton Zeilinger, Herbert J. Bernstein, and Philip Bertani. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.*, 73:58–61, Jul 1994.

[10] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, October 1948.

[11] Christian Weedbrook, Stefano Pirandola, Raúl García-Patrón, Nicolas J. Cerf, Timothy C. Ralph, Jeffrey H. Shapiro, and Seth Lloyd. Gaussian quantum information. *Rev. Mod. Phys.*, 84:621–669, May 2012.

Sinkhorn normal form for unitary matrices

M. Idel, M. Wolf September 26, 2016

Sinkhorn's theorem states that for any matrix $A \in \mathbb{R}^{n\times m}$ with positive entries there exist diagonal matrices D_1, D_2 such that $D_1 A D_2$ is doubly stochastic [5]. Recently, the question was posed whether a similar theorem can exist for unitary matrices [3]: For any unitary U, do there exist positive diagonal unitaries L, R such that LUR has row and column sums 1? We answer that question in the positive and also provide applications.

1 Sinkhorn for unitaries

The basic result is the following theorem:

Theorem 1.1. *For every unitary matrix $U \in U(n)$ there exist diagonal unitary matrices L, R such that LUR has unit row and column sums. The matrices L, R are not unique.*

The proof relies on the observation that the theorem is equivalent to the question whether there exists a vector $v = (e^{i\phi_1}, \dots, e^{i\phi_n})$ with $\phi_i \in [0, 2\pi)$ such that $Uv = (e^{i\psi_1}, \dots, e^{i\psi_n})$ or in other words, whether there exists a vector in the Clifford torus $\mathbb{C}P^{n-1}$ which is mapped to a vector in the Clifford torus by U. This in turn is immediately equivalent to a question in symplectic topology which was answered (among others) in [2].

2 Applications

One interesting consequence of this theorem is the following decomposition of unitary matrices:

Corollary 2.1. *Let $U \in U(n)$, then there exist diagonal unitary matrices $D_1, \dots, D_n$ and $\tilde{D}_1, \dots, \tilde{D}_{n-1}$ and a $\varphi \in [0, 2\pi)$ such that the first $i-1$ entries in each $D_i, \tilde{D}_i$ are equal to one and*

$$\begin{aligned} U = {} & D_1 F_n D_2 (\mathbb{1}_1 \oplus F_{n-1}) D_3 (\mathbb{1}_2 \oplus F_{n-2}) \cdots \\ & D_{n-1} (\mathbb{1}_{n-2} \oplus F_2) D_n (\mathbb{1}_{n-2} \oplus F_2^\dagger) \tilde{D}_{n-1} \cdots (\mathbb{1}_1 \oplus F_{n-1}^\dagger) \tilde{D}_2 F_n^\dagger \tilde{D}_1 e^{i\varphi}. \end{aligned}$$

In other words, any unitary can be decomposed into a product of diagonal unitaries and Fourier transforms on submatrices. This decomposition therefore also provides a simple design to implement Gaussian unitaries: If we know how to perform Fourier transforms on arbitrary many modes, we only need to add phase shifters to implement all unitary transformations.

3 Legal statement

The project was proposed by Michael Wolf. In all parts of this work I was significantly involved. At the time of publication, we were not aware that a similar proof (albeit for a different problem) had already been published in [4]. I only became aware of the proof after the two results were linked in [1].

References

[1] Ole Andersson and Ingemar Bengtsson. Cliffordtori and unbiased vectors, 2015. arXiv:1506.09062 [quant-ph].

[2] Paul Biran, Michael Entov, and Leonid Polterovich. Calabi quasimorphisms for the symplectic ball. *Communications in Contemporary Mathematics*, 06(05):793–802, 2004.

[3] Alexis De Vos and Stijn De Baerdemacker. Scaling a unitary matrix. *Open Systems & Information Dynamics*, 21(4), 2014.

[4] Kamil Korzekwa, David Jennings, and Terry Rudolph. Operational constraints on state-dependent formulations of quantum error-disturbance trade-off relations. *Phys. Rev. A*, 89:052108, 05 2014.

[5] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35(2):876–879, 1964.

Linear Algebra and its Applications 471 (2015) 76–84

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Sinkhorn normal form for unitary matrices

Martin Idel *, Michael M. Wolf

Zentrum Mathematik, M5, Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany

ARTICLE INFO

Article history:
Received 16 October 2014
Accepted 24 December 2014
Available online xxxx
Submitted by M. Bresar

MSC:
15A21
53D12

Keywords:
Matrix scaling
Unitary
Doubly-stochastic

ABSTRACT

Sinkhorn proved that every entry-wise positive matrix can be made doubly stochastic by multiplying with two diagonal matrices. In this note we prove a recently conjectured analogue for unitary matrices: every unitary can be decomposed into two diagonal unitaries and one whose row- and column sums are equal to one. The proof is non-constructive and based on a reformulation in terms of symplectic topology. As a corollary, we obtain a decomposition of unitary matrices into an interlaced product of unitary diagonal matrices and discrete Fourier transformations. This provides a new decomposition of linear optics arrays into phase shifters and canonical multiports described by Fourier transformations.

1. Introduction

For every $n \times n$ matrix A with positive entries there exist two diagonal matrices L, R such that LAR is doubly stochastic, i.e. the entries of each column and row sum up to one. This result was first obtained by Sinkhorn [8], who also gave an algorithm of how to compute L and R by iterated left and right multiplication of diagonal matrices.

Recently, De Vos and De Baerdemacker studied the same problem for unitary matrices [3]. They conjectured that for every $n \times n$ unitary U there exist two unitary diagonal

* Corresponding author. Tel.: +49 89 289 17017.
E-mail addresses: martin.idel@tum.de (M. Idel), wolf@ma.tum.de (M.M. Wolf).

http://dx.doi.org/10.1016/j.laa.2014.12.031
0024-3795/

matrices L, R such that LUR has all row and column sums equal to one. To support their conjecture, they construct an algorithm similar to the iteration procedure for matrices with positive entries from [8,9]. They also provide numerical evidence that the algorithm always converges to a unitary matrix with row and column sums equal to one.

The goal of this paper is to prove the conjecture of De Vos and De Baerdemacker that such a normal form always exists by reformulating the problem in terms of symplectic topology. It turns out that the reformulated problem is a special case of the Arnold (sometimes Arnold–Givental) conjecture on the intersection of Lagrangian submanifolds [6], which was solved for this case in [1,2]. More precisely, in Section 2 we show:

Theorem 2. *For every unitary matrix $U \in U(n)$ there exist two diagonal unitary matrices $L, R \in U(n)$ such that $A := LUR$ satisfies $\sum_j A_{ji} = \sum_j A_{ij} = 1$ for all $i = 1, \ldots, n$.*

For a given unitary $U \in U(n)$ the triple (L, R, A) is certainly not unique, since multiplying L by a global phase and R by its inverse does not change LAR. Hence, it makes sense to consider the decomposition $U = e^{i\varphi} L'AR'$, where L', R' are unitary diagonal such that $L'_{11} = R'_{11} = 1$ and $\varphi \in [0, 2\pi)$. In particular, for $U(2)$, a simple complete solution was given in [3] from which one can see that for every non-diagonal matrix, there are only two different A such that $e^{i\varphi} LAR = U$. For $n > 2$ the picture is less clear and the reformulation in terms of symplectic topology appears to give further insight into the freedom of the decomposition.

In addition to the Sinkhorn-type normal form above, in Section 3 we give several reformulations that might be interesting for applications, for instance regarding the decomposition of general $2n$-port linear optics devices into canonical multiports and phase shifters.

2. Sinkhorn-type normal form

In order to prove the decomposition theorem, we reformulate the problem of rescaling a unitary matrix into a problem in symplectic topology. For the reader's convenience, necessary results including elementary calculations and definitions are included in Appendix A. We only repeat the most important definitions for our reformulation. Recall that the complex projective space $\mathbb{C}P^n$ consists of all equivalence classes of $\mathbb{C}^{n+1}\backslash\{0\}$ w.r.t. $x \sim y \Leftrightarrow x = \lambda y$ with $\lambda \in \mathbb{C}\backslash\{0\}$.

Definition 1. The *Clifford Torus* is the n-dimensional torus embedded in $\mathbb{C}P^n$, i.e. the set of points

$$T^n := \left\{ [w_0, \ldots, w_n] \in \mathbb{C}P^n \;\middle|\; |w_0| = |w_1| = \ldots = |w_n| \right\}. \tag{1}$$

This torus, as shown in the appendix in Proposition 4, is a Lagrangian submanifold of the symplectic manifold $\mathbb{C}P^n$. We obtain the following connection to our normal form:

Lemma 1. *For any unitary $U \in U(n)$, there exist diagonal unitaries L and R such that $A := LUR$ has row and column sums equal to one if and only if the Clifford torus $T^{n-1} \subset \mathbb{C}P^{n-1}$ fulfills $T^{n-1} \cap UT^{n-1} \neq \emptyset$.*

Proof. Let $U \in U(n)$ be arbitrary but fixed. We first consider the usual torus $\mathbb{T}^n \subset \mathbb{C}^n$, i.e. the set of all vectors for which each component has modulus one:

$$\mathbb{T}^n := \left\{ \left(e^{i\phi_1}, \ldots, e^{i\phi_n}\right) \subset \mathbb{C}^n \;\middle|\; \phi_j \in \mathbb{R} \right\}$$

Let us first show that the existence of a normal form is equivalent to $\mathbb{T}^n \cap U\mathbb{T}^n \neq \emptyset$. For one direction, let $\varphi \in \mathbb{T}^n$ such that $U\varphi \in \mathbb{T}^n$, i.e. $\varphi \in \mathbb{T}^n \cap U\mathbb{T}^n$. Define the two diagonal matrices $R^{-1} := \operatorname{diag}(\varphi_1, \ldots, \varphi_n) \in U(n)$ and $L^{-1} := \operatorname{diag}((U\varphi)_i^{-1}) = \operatorname{diag}((\overline{U\varphi})_i) \in U(n)$. With $A := L^{-1}UR^{-1}$ and $e := (1, \ldots, 1)^T$ we obtain:

$$Ae = L^{-1}U\varphi = e$$

Likewise, since $\overline{A}e = Ae$ and A is unitary, we obtain

$$A^T e = A^T \overline{A} e = e$$

so that columns and rows of A sum up to one.

For the other direction, suppose $U = LAR$ is a decomposition as proposed. Then $\varphi := R^{-1}e \in \mathbb{T}^n$ and

$$U\varphi = LAR\varphi = LAe = Le \in \mathbb{T}^n$$

hence $U\varphi \in \mathbb{T}^n \cap U\mathbb{T}^n$.

The next step is to reformulate the problem using the Clifford torus. Clearly, $T^{n-1} \cap UT^{n-1} \neq \emptyset$ iff $(\lambda\mathbb{T}^n) \cap U\mathbb{T}^n \neq \emptyset$ for some $\lambda \in \mathbb{C} \setminus \{0\}$. Since U is norm preserving, any intersection requires $|\lambda| = 1$ so that

$$T^{n-1} \cap UT^{n-1} \neq \emptyset \quad \Leftrightarrow \quad \mathbb{T}^n \cap U\mathbb{T}^n \neq \emptyset. \qquad \square$$

One of the main conjectures in symplectic topology, the Arnold or Arnold–Givental conjecture, states that a Lagrangian submanifold and its image under a Hamiltonian isotopy intersect at least as often as the sum of the $\mathbb{Z}_2$-Betti-numbers. For T^n, this sum is not zero, thus, using Proposition 5, Arnold's conjecture states in particular that T^n should intersect with UT^n at least once. While the Arnold conjecture is wrong in all generality and most cases are unknown, there is a positive result to the weaker question whether the torus intersects with its displaced version (cf. [1,2]). In order to formulate this result, we need the following:

Definition 2. Let $(\mathcal{M}, \omega)$ be a closed symplectic manifold with Hamiltonian symplectomorphisms $\mathrm{Ham}(\mathcal{M})$. A Lagrangian submanifold $\mathcal{L} \subset \mathcal{M}$ is called *displaceable* by a Hamiltonian diffeomorphism, if there exists a $\psi \in \mathrm{Ham}(\mathcal{M})$ such that

$$\mathcal{L} \cap \psi\mathcal{L} = \emptyset.$$

The definition is slightly different from the one in [1], where the authors only consider nonempty open sets such that the restriction of ω to these sets is exact. However, they prove that the torus T^n is displaceable in the above definition, if and only if there exists an open neighborhood $\mathcal{V} \supset T^n$ such that $\omega|_{\mathcal{V}}$ is exact and $\mathcal{V}$ is displaceable. With this we can state the final and crucial ingredient in the proof of the normal form:

Theorem 1. *(See [1, Theorem 1.3].) The Clifford torus $T^n \subset \mathbb{C}P^n$ cannot be displaced from itself by a Hamiltonian isotopy.*

Because every unitary matrix defines a Hamiltonian isotopy (see Proposition 5 in the appendix), the theorem tells us in particular $T^n \cap UT^n \neq \emptyset$ for all unitaries $U \in U(n)$ so that together with Lemma 1 this proves the sought normal form:

Theorem 2. *For every unitary matrix $U \in U(n)$ there exist two diagonal unitary matrices $L, R \in U(n)$ such that $A := LUR$ fulfills $\sum_j A_{ji} = \sum_j A_{ij} = 1$ for all $i = 1, \ldots, n$.*

3. Equivalent normal forms for unitary matrices

To obtain equivalent normal forms, consider the $n \times n$ dimensional complex matrix F_n with entries $(F_n)_{kl} := \frac{1}{\sqrt{n}} \exp(\frac{2\pi i}{n} kl)$ with $k, l \in \{0, \ldots, n-1\}$, which is known as the *discrete Fourier transformation*. It is easy to see that $F_n^{-1} = F^\dagger$, hence $F_n \in U(n)$. If we denote the standard basis of $\mathbb{C}^n$ by $\{e_i\}_{i=0}^{n-1}$ and $e := (1, \ldots, 1)^T$, then

$$F_n e_0 = F_n^\dagger e_0 = \frac{e}{\sqrt{n}}.$$

Now let $A \in U(n)$ be such that $Ae = A^T e = e$. Then $F_n^\dagger A F_n e_0 = e_0$ and similarly, $(F_n^\dagger A F_n)^T e_0 = F_n A^T F_n^\dagger e_0 = e_0$, which shows that

$$F_n^\dagger A F_n = \begin{pmatrix} 1 & 0_{n-1}^T \\ 0_{n-1} & \tilde{U} \end{pmatrix}$$

where $0_{n-1} := 0 \in \mathbb{C}^{n-1}$ and $\tilde{U} \in U(n-1)$. Thus, given a unitary $U \in U(n)$, we know that there exists a decomposition

$$U = L F_n \begin{pmatrix} 1 & 0_{n-1}^T \\ 0_{n-1} & \tilde{U} \end{pmatrix} F_n^\dagger R \tag{2}$$

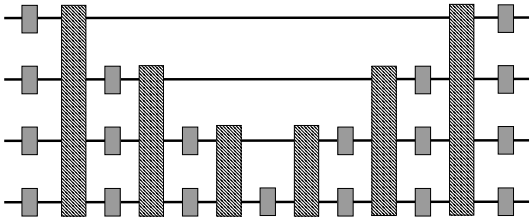

Fig. 1. In quantum optics, passive transformations on n modes are in one-to-one correspondence with $n \times n$ unitaries. Up to an overall phase, each unitary U admits a decomposition into $2(n-1)$ canonical multiports (which are independent of U and described by discrete Fourier transformations [hatched]) surrounded by $2n-1$ layers of single-mode phase shifters [grey]. Here, this is exemplified for $n = 4$.

with $\tilde{U} \in U(n-1)$ and diagonal $L, R \in U(n)$. We can now iterate the procedure by applying it to the $(n-1) \times (n-1)$-dimensional submatrix $\tilde{U}$ and obtain the corollary:

Corollary 1. *Let $U \in U(n)$, then there exist diagonal unitaries $D_1, \ldots, D_n$ and $\tilde{D}_1, \ldots, \tilde{D}_{n-1}$ and a $\varphi \in [0, 2\pi)$ such that the first $i-1$ entries in each D_i, $\tilde{D}_i$ are equal to one and*

$$\begin{aligned} U = {} & D_1 F_n D_2 (\mathbb{1}_1 \oplus F_{n-1}) D_3 (\mathbb{1}_2 \oplus F_{n-2}) \\ & \cdots D_{n-1} (\mathbb{1}_{n-2} \oplus F_2) D_n \big(\mathbb{1}_{n-2} \oplus F_2^\dagger\big) \tilde{D}_{n-1} \cdots \big(\mathbb{1}_1 \oplus F_{n-1}^\dagger\big) \tilde{D}_2 F_n^\dagger \tilde{D}_1 e^{i\varphi}. \end{aligned} \tag{3}$$

In other words any unitary can be decomposed into diagonal unitaries and discrete Fourier transformations in this way. This has an immediate application in quantum optics, where any $n \times n$ unitary corresponds to a passive transformation on n modes or a $2n$-multiport. In this scenario a diagonal unitary corresponds to a set of phase shifters, which are applied to the modes individually and the discrete Fourier transformation is known as canonical $2n$-multiport [5], which may be implemented by a symmetric fibre coupler. The structure of the corresponding decomposition is graphically depicted in Fig. 1.

Another version of the normal form is found by using that D is a diagonal matrix iff $FDF^\dagger$ is a *circulant matrix*, i.e. $(FDF^\dagger)_{i,j} =: \alpha_{i-j} \in \mathbb{C}$. Since the diagonal matrices form a group, so do the circulant matrices and we denote the group of $n \times n$ circulant matrices by $\mathrm{Circ}(n)$. Then:

Corollary 2. *Let $U \in U(n)$, then there exist $C_1, C_2 \in \mathrm{Circ}(n)$ and $\tilde{U} \in U(n-1)$ such that*

$$U = C_1 \operatorname{diag}(1, \tilde{U}) C_2. \tag{4}$$

Let us finally discuss the question of uniqueness of these decompositions and to this end come back to the original normal form

$$U = e^{i\varphi} D_1 A D_2, \tag{5}$$

where D_1, D_2 are unitary diagonal with $(D_i)_{11} = 1$ and A has row and column sums equal to 1. Counting parameters, using that the matrices A are isomorphic to $U(n-1)$ as proven above, we have:

$$1 + (n-1) + (n-1)^2 + (n-1) = n^2$$

parameters (cf. [3]). Hence, the number of parameters matches exactly the dimension of $U(n)$. Given a unitary $U = e^{i\varphi} D_1 A D_2$ as above, this means that it might be reasonable to expect only a discrete set of different decompositions or at least a discrete set of A that U can be scaled to. The exact number of different A can easily be seen to be two for the case $n = 2$ (cf. [3]), but already for $n = 3$ and $n = 4$, there is only a conjectured bound (6 and 20, cf. [7]).

In [2] it is proven that if T^n and UT^n intersect transversally, their number of distinct intersection points must be at least 2^n, which follows from general results in Floer-homology theory when applied to Lagrangian intersection theory. Since transversality is a generic property for intersections, one might therefore conjecture that for a generic unitary $U \in U(n)$ [2] implies a lower bound 2^{n-1} on the number of different normal forms. However, it is not true that we always have a discrete number of decompositions or (in contrast to the 2×2 case) at least a discrete number of A such that A has row and column-sums equal to one and $e^{i\varphi} LAR = U$. A counterexample is given by the Fourier transform in 4×4 dimensions, where we have for any $\varphi \in [0, 2\pi)$[1]:

$$\begin{aligned}\frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{i\varphi} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -e^{-i\varphi} \end{pmatrix} \\ \cdot \frac{1}{2}\begin{pmatrix} 1 & -ie^{i\varphi} & 1 & ie^{i\varphi} \\ e^{-i\varphi} & 1 & -e^{-i\varphi} & 1 \\ 1 & ie^{i\varphi} & 1 & -ie^{i\varphi} \\ -e^{-i\varphi} & 1 & e^{i\varphi} & 1 \end{pmatrix} \\ \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & ie^{-i\varphi} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -ie^{i\varphi} \end{pmatrix}\end{aligned} \tag{6}$$

After completion of this document, we learned that part of this section, in particular Corollary 1 were independently found in [4].

4. Conclusion

We have studied a variant of a Sinkhorn type normal form for unitary matrices. Its existence was conjectured in [3] and we give a nonconstructive proof. This means in

[1] We thank the anonymous referee for providing this counterexample.

particular that the question, whether the algorithm presented in [3] always converges for any set of starting conditions, remains open. Also, it would be nice to have an elementary proof of the fact that for any unitary matrix U we have $T^n \cap UT^n \neq \emptyset$. The decomposition is in not unique: We provided an example where, contrary to the 2×2-case, there is a one-parameter set of A as well as L and R, such that $LAR = U$. We suggested an argument that the number of different decompositions, if it is discrete, might grow exponentially. However this lower bound relies on a lower bound on Lagrangian intersections which holds only for transversal intersections.

Acknowledgements

We thank Michael Keyl for many helpful comments on the parts involving symplectic topology. M. Idel is supported by the Studienstiftung des Deutschen Volkes. M. Wolf acknowledges support from the CHIST-ERA/BMBF project CQC.

Appendix A. Symplectic preliminaries

This section introduces the definitions and results from symplectic topology beyond the first chapters of [6] needed to understand the basic reductions of the proof of Theorem 1 in [1].

A.1. Notation and basic definitions

To fix notation, a symplectic manifold will always be denoted by $\mathcal{M}$ and its symplectic form will be called ω. The group of *symplectomorphisms* of a symplectic manifold $(\mathcal{M}, \omega)$ will be denoted by $\mathrm{Symp}(\mathcal{M})$ and its *Hamiltonian symplectomorphisms* (i.e. all symplectomorphisms which are elements of the flow of a Hamiltonian vector field) will be denoted by $\mathrm{Ham}(\mathcal{M})$. We have the following characterization [6, Chapter 10]:

Proposition 3. *Let $(\mathcal{M}, \omega)$ be a closed symplectic manifold. If the manifold is simply connected (i.e. every loop is contractible)*

$$\mathrm{Ham}(\mathcal{M}) = \mathrm{Symp}_0(\mathcal{M})$$

where $\mathrm{Symp}_0(\mathcal{M})$ denotes the connected component of the identity of the whole group of symplectomorphisms.

In principle, the result also holds for arbitrary symplectic manifolds. One has to be more careful with non-compactly supported functions, but we can safely ignore these subtleties, since our manifold of interest will be closed.

Furthermore, let us recall that a *Lagrangian submanifold* $\mathcal{L}$ of a $2n$-dimensional symplectic manifold $(\mathcal{M}, \omega)$ is a smooth n-dimensional submanifold of $\mathcal{M}$ such that

$$T_p\mathcal{L}^{\varepsilon} := \left\{ X \in T_p\mathcal{M} \;\middle|\; \omega(X, Y) = 0 \ \forall Y \in T_p\mathcal{L} \right\} = T_p\mathcal{L} \quad \forall p \in \mathcal{L}$$

A.2. The Clifford-torus as a Lagrangian submanifold

We now study the Clifford torus as a special case of the Lagrangian submanifold of interest for our result.

Before proving that the Clifford torus is a Lagrangian submanifold, we need to specify the symplectic structure on $\mathbb{C}P^n$: Consider the map $\Phi : \mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{S}^{n+1} \subset \mathbb{C}^{n+1}$ via $z \mapsto z/|z|$. We will show that the pullback $\Phi^*\omega$ of the standard symplectic structure ω on $\mathbb{C}^{n+1}$ descends to a symplectic form ω_{FB} on $\mathbb{C}P^n$, the *standard symplectic structure* or *Fubini–Study form* of the complex projective space.

Proposition 4. *$\mathbb{C}P^n$, equipped with the Fubini–Study form is a $2n$-dimensional symplectic manifold and the Clifford torus is a Lagrangian submanifold thereof.*

Proof. Let us go through the construction in more detail and see, how it defines a symplectic form, e.g. a non-degenerate and closed 2-form on $\mathbb{C}P^n$. Throughout, we will consider the natural projection $\pi : \mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{C}P^n$.

Note that if $(x_0, y_0, \ldots, x_n, y_n)$ are the real coordinates of $\mathbb{R}^{2n+2} \cong \mathbb{C}^{n+1}$, we can use $(z_0, \bar{z}_0, \ldots, z_n, \bar{z}_n)$ as coordinates for any point $(z_0, \ldots, z_n) \in \mathbb{C}^{n+1}$ as well. Then the standard symplectic form reads

$$\omega = \sum_j dx^j \wedge dy^j = \frac{i}{2} \sum_j dz^j \wedge d\bar{z}^j$$

Considering the action of $\mathbb{C}^*$ on $\mathbb{C}^{n+1}$, we obtain $\omega_{\lambda \cdot z} = \frac{i}{2} \sum_j d(\lambda \cdot z^j) \wedge d\overline{(\lambda \cdot z^j)} = |\lambda|^2 \frac{i}{2} \sum_j dz^j \wedge d\bar{z}^j = |\lambda|^2 \omega_z$. Hence, if $\Phi : \mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{S}^{2n+1}$ is given by $z \mapsto z/|z|$, then $\Phi^*\omega$ will be invariant under the action of $\mathbb{C}^*$. This shows that $\Phi^*\omega$ descends to a well-defined 2-form ω_{FS} on $\mathbb{C}P^n$, by defining:

$$(\omega_{FS})_{\pi(p)}(d\pi X_{\pi(p)}, d\pi Y_{\pi(p)}) = \left(\Phi^*\omega\right)_p (X, Y)$$

The next step is to show non-degeneracy. For this, note that $\Phi^*\omega(X, Y) = 0 \; \forall Y$ if and only if $d\Phi X = 0$ pointwise, since ω is non-degenerate. But $d\Phi X = 0$ implies in particular $d\pi X = 0$ and hence, ω_{FS} as defined above is a non-degenerate 2-form.

Finally, we need to prove closedness. This can either be computed directly by considering coordinates, or by considering local sections of the projection π. Let $\{U_i\}_i$ be a cover of $\mathbb{C}P^n$ such that there exists a local section $\sigma_i : U_i \to \mathbb{C}^{n+1} \setminus \{0\}$. On each U_i we have $\omega_{FS} = \sigma_i^* \Phi^* \omega$. But then

$$d\omega_{FS} = d\left(\sigma_i^* \Phi^* \omega\right) = (\sigma_i \Phi)^* d\omega = 0$$

since d commutes with pullbacks and ω is closed. Since this holds on any patch U_i, $d\omega_{FS} = 0$ globally.

In addition, we need to see that the Clifford torus is a Lagrangian submanifold. It is easy to see that the Clifford torus is a submanifold of (real) dimension n, hence we only need to prove $(T_pT^n)^\varepsilon = T_pT^n \ \forall p \in T^n$. Given the canonical projection $\pi : \mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{C}P^n$, T^n is the image of π of the torus

$$\mathbb{T}^n := \left\{(z_0, \ldots, z_n) \;\middle|\; |z_0| = |z_1| = \ldots = |z_n| = 1\right\}$$

By inspection, we obtain for $p = (p_0, \ldots, p_n) \in \mathbb{C}^{n+1} \setminus \{0\}$:

$$T_p\mathbb{T}^n = \operatorname{span}\{p_i\partial_{\bar{p}_i} - \bar{p}_i\partial_{p_i} \mid i = 0, \ldots, n\} =: \operatorname{span}\left\{X_p^i \;\middle|\; i = 0, \ldots, n\right\}$$

Then $T_{\pi(p)}T^n$ will be spanned by $d\pi X^i_{\pi(p)}$.

Now, since already on the level of ω, we have $\omega_p(X_p^i, X_p^j) = 0$ for all $i, j \in \{0, \ldots, n\}$ and all $p \in \mathbb{C}^{n+1} \setminus \{0\}$, it is immediate that $(\omega_{FS})_{\pi(p)}(\pi_* X_p^i, \pi_* X_p^j) = 0$ for all i, j and for all $\pi(p) \in \mathbb{C}P^n$. Hence we have that $(T_pT^n)^\varepsilon \supseteq T_pT^n \ \forall p \in T^n$. Since equality then has to hold by dimensional analysis, we have T^n is a Lagrangian submanifold. □

Now consider the standard action of $U \in U(n+1)$ on $\mathbb{C}^{n+1}$. Note that U leaves ω invariant, since $\sum_i d(Uz)^i \wedge d\overline{Uz}^i = \sum_{ijk} U_{ij}\overline{U}_{ik} dz^j \wedge d\bar{z}^k = \sum_i dz^i \wedge d\bar{z}^i$. Furthermore, since U leaves the norm invariant by definition, we have that $U^*\omega_{FS} = \omega_{FS}$, where U^* is the pullback associated with the map U. This means that any unitary $U \in U(n+1)$ corresponds to a symplectomorphism of $\mathbb{C}P^n$. Since it is well-known that the complex projective space is simply connected and closed, its Hamiltonian symplectomorphism corresponds to its symplectomorphism. Hence:

Proposition 5. *We have $U(n+1) \subset \operatorname{Ham}(\mathbb{C}P^n, \omega_{FS})$, where the identification is achieved by considering the standard action of U on $\mathbb{C}^{n+1}$.*

References

[1] Paul Biran, Michael Entov, Leonid Polterovich, Calabi quasimorphisms for the symplectic ball, Commun. Contemp. Math. 06 (05) (2004) 793–802.

[2] Cheol-Hyun Cho, Holomorphic discs, spin structures, and Floer cohomology of the Clifford torus, Int. Math. Res. Not. 2004 (35) (2004) 1803–1843.

[3] Alexis De Vos, Stijn Baerdemacker, Scaling a unitary matrix, Open Syst. Inf. Dyn. 21 (4) (2014).

[4] Alexis De Vos, Stijn Baerdemacker, The synthesis of a quantum circuit, in: Proceedings of the 11th International Workshop of Boolean Problems, Freiberg University of Mining and Technology, 2014, pp. 129–136.

[5] K. Mattle, M. Michler, H. Weinfurter, A. Zeilinger, M. Zukowski, Non-classical statistics at multiport beam splitters, Appl. Phys. B Lasers Opt. 60 (1995) S111–S117.

[6] Dusa McDuff, Dietmar Salamon, Introduction to Symplectic Topology, Oxford Science Publications, 1998.

[7] V.S. Shchesnovich, Asymptotic evaluation of bosonic probability amplitudes in linear unitary networks in the case of large number of bosons, Int. J. Quantum Inf. 11 (2013).

[8] Richard Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices, Ann. Math. Stat. 35 (2) (1964) 876–879.

[9] Richard Sinkhorn, Paul Knopp, Concerning nonnegative matrices and doubly stochastic matrices, Pacific J. Math. 21 (2) (1967) 343–348.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 17, 2016

This Agreement between Martin Idel ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	3951311192217
License date	Sep 17, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Linear Algebra and its Applications
Licensed Content Title	Sinkhorn normal form for unitary matrices
Licensed Content Author	Martin Idel,Michael M. Wolf
Licensed Content Date	15 April 2015
Licensed Content Volume Number	471
Licensed Content Issue Number	n/a
Licensed Content Pages	9
Start Page	76
End Page	84
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Normal Forms and Squeezing in Continuous Variable Quantum Information Theory
Expected completion date	Sep 2016
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Requestor Location	Martin Idel Boltzmannstr. 3 Garching, 85748 Germany Attn: Martin Idel
Total	0.00 EUR

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:
15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.
16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following

clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript
    - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- after the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new

works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.8

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Sinkhorn review

M. Idel September 26, 2016

Since its first proofs in the 1960s, Sinkhorn's theorem has become a cornerstone in several areas of applied mathematics, ranging from transportation science and economics to biology. My own interest began during my Master's studies, where I proved a generalisation to positive maps. It later transpired that the theorem had already been proved in [2] albeit with a very different proof that seemed unrelated to my approach. We found this paper only shortly before we wanted to publish the results. Later in the same year, another paper appeared that did publish the results [1]. Learning from this experience, I started to write a review about the problem of matrix scaling, matrix balancing and equivalence scaling and its generalisations, in particular to positive maps.

1 Scope

The review gathers everything I could find about the mathematics concerning Sinkhorn's theorem. It starts with a historical section describing the origins of the question, before investigating many different approaches to prove the theorem in its original form. The fourth section gives an organised overview about maximal results achieved over the years. In the following two section, I first analyse related questions of DAD and DAD^{-1} scalings as well as other scalings (such as unitary scaling or complex matrix scalings) before discussing generalised approaches that cover multiple scalings. Two short sections are focused on complexity theoretic results for algorithms and applications. Finally, a large section is devoted to the problem of positive matrix scaling. It also contains a full proof of Gurvits' results, since I found the proofs difficult to follow.

2 Legal statement

This paper is a review and contains little to no nontrivial original results. I tried to give due credit wherever a section is inspired by existing partial reviews. The project developed over the years from my own interest. The original problem of finding a Sinkhorn analogue for positive maps was given to me by Michael Wolf for my Master's thesis [3], however the review project did not emerge before 2014.

References

[1] Tryphon T. Georgiou and Michele Pavon. Positive contraction mappings for classical and quantum schrödinger systems. *Journal of Mathematical Physics*, 56(3), 2015.

[2] Leonid Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69(3):448 – 484, 2004. Special Issue on {STOC} 2003.

[3] Martin Idel. On the structure of positive maps. Master's thesis, Ludwig-Maximilian Universität München, Technische Universität München, 2013.

A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps

Martin Idel*

Zentrum Mathematik, M5, Technische Universität München, 85748 Garching

Abstract

Given a nonnegative matrix A, can you find diagonal matrices D_1, D_2 such that D_1AD_2 is doubly stochastic? The answer to this question is known as Sinkhorn's theorem. It has been proved with a wide variety of methods, each presenting a variety of possible generalisations. Recently, generalisations such as to positive maps between matrix algebras have become more and more interesting for applications. This text gives a review of over 70 years of matrix scaling. The focus lies on the mathematical landscape surrounding the problem and its solution as well as the generalisation to positive maps and contains hardly any nontrivial unpublished results.

Contents

*martin.idel@tum.de

1. Introduction

It is very common for important and accessible results in mathematics to be discovered several times. Different communities adhere to different notations and rarely read papers in other communities also because the reward does not justify the effort. In addition, even within the same community, people might not be aware of important results - either because they are published in obscure journals, they are poorly presented in written or oral form or simply because the mathematician did not notice them in the surrounding sea of information. This is a problem not unique to mathematics but instead inherent in all disciplines with epistemological goals.

The scaling of matrices is such a problem that has constantly attracted attention in various fields of pure and applied mathematics[1]. Recently, generalisations have been studied also in physics to explore possibilities in quantum mechanics where it turns out that a good knowledge of the vast literature on the problem can help a lot in formulating approaches. This review tries to tell the mathematical story of matrix scaling, including algorithms and pointers to applications.

As a motivation, consider the following problem: Imagine you take a poll, where you ask a subset of the population of your country what version (if any) of a certain product they buy. You distinguish several groups in the population (for instance by age, gender, etc.) and you distinguish several types of product (for instance different brands of toothbrushs). From the sales statistics, you know the number of each product sold in the country and from the country statistics you know the number of people in different groups. Given the answers of a random sample of the population, how can you extrapolate results?

Central to a solution is the following innocuous theorem:

Theorem 1.1 (Sinkhorn's theorem, weak form Sinkhorn 1964)**.** *Given a matrix A with positive entries, one can find matrices D_1, D_2 such that $D_1 A D_2$ is doubly stochastic.*

The literature on Sinkhorn's theorem and its generalisations is vast. As we will see, there are some natural ways to attack this problem, which further explains why the different communities were often not aware of the efforts of their peers in other fields.

One of the main motivations for this review was a generalisation of Sinkhorn's theorem to the noncommutative setting of positive maps on matrix algebras:

[1]The term "Babylonian confusion" to describe the history of this problem was first used in Krupp 1979

Theorem 1.2 (Weak form of Gurvits 2003's generalisation to positive maps)**.** *Given a map $\mathcal{E} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ which maps positive semidefinite matrices to positive definite matrices one can find invertible matrices X, Y such that the map $\mathcal{E}'(\cdot) := Y\mathcal{E}(X \cdot X^\dagger)Y^\dagger$ is doubly stochastic, i.e.*

$$\mathcal{E}'(\mathbb{1}) = \mathbb{1}, \qquad \mathcal{E}'^*(\mathbb{1}) = \mathbb{1}$$

with the adjoint matrices $X^\dagger$ and the adjoint map $\mathcal{E}^*$.

Some results and approaches can be translated to this noncommutative setting, but many questions remain open and the noncommutativity of the problem makes progress difficult.

Very recently, a new generalisation of Sinkhorn's theorem to a noncommutative setting has appeared in Benoist and Nechita 2016.

The goal of this review is therefore threefold:

- Trace the historical developments of the problem and give credit to the many people who contributed to the problem.
- Illuminate the many approaches and connections between the approaches to prove Sinkhorn's theorem and its generalisations.
- Sketch the generalisation to positive maps and its history and highlight the questions that are yet unanswered and might be attacked using the knowledge from the classical version.

In addition, I will try to give a sketch of the algorithmic developments and pointers to the literature for applications. I will probably have forgotten and/or misrepresented contributions; comments to improve the review are therefore very welcome.

2. Notation and Preliminaries

Most of the concepts and notations discussed in this short section are well-known and can be found in many books. I encourage the reader to refer to this section only if some notation seems unclear.

We will mostly consider matrices $A \in \mathbb{R}^{n\times m}$. Such matrices are called *nonnegative* (*positive*) if they have only nonnegative (positive) entries. We denote by $\mathbb{R}^n_+$ ($\mathbb{R}^n_{+0}$) all vectors with only positive entries (nonnegative entries) and for any such $x \in \mathbb{R}^n_+$, $\mathrm{diag}(x)$ defines the diagonal matrix with x on its main diagonal, while $1/x \in \mathbb{R}^n_+$ defines the vector with entries $1/x_i$ for all i.

An important concept for nonnegative matrices is the pattern. The *support* or *pattern* of a matrix A is the set of entries where $A_{ij} > 0$. A subpattern of the pattern of A is then a pattern with fewer entries than the pattern of A. We write $B \prec A$ if B is a subpattern of A, i.e. for every $B_{ij} > 0$ we have $A_{ij} > 0$.

Finally, let us introduce irreducibility and decomposability. Details and connections to other notions for nonnegative matrices are explained in Appendix A. If A is nonnegative, then A is *fully indecomposable* if and only if there do not exist permutations P, Q such that

$$PAQ = \begin{pmatrix} A_1 & 0 \\ A_3 & A_2 \end{pmatrix} \tag{1}$$

where neither A_1 nor A_2 contain a zero row or column and $A_3 \neq 0$. The matrix is *irreducible*, if no permutation P can be found such that already PAP^T is of form (1). In particular, this implies that all fully indecomposable matrices are at least irreducible.

For positive vectors, we will not use the notation $x > 0$ to avoid confusion with the positive definite case: Especially in the second part of this review, we will be dealing mostly with *positive (semi)definite* matrices $A \in \mathbb{R}^{n\times n}$, which are symmetric matrices with only positive (nonnegative) eigenvalues and should not be confused with positive matrices. We also introduce the partial order $\geq$ for positive semidefinite matrices, where $A \geq B$ if and only if $A - B$ is positive semidefinite and $A > B$ if $A - B$ is positive definite.

When talking about positive maps, we will also adopt the notation that $\mathcal{M}_{n,m}$ denotes the complex $n \times m$ matrices, while the shorter $\mathcal{M}_n$ is used for complex $n \times n$ square matrices.

3. Different approaches to equivalence scaling

This section explores the historical development and current form of the mathematical landscape surrounding the following extension to Theorem 1.1:

Theorem 3.1. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix with nonnegative entries. Then for any vectors $r \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ with nonnegative numbers there exist diagonal matrices D_1 and D_2 such that*

$$D_1 A D_2 e = r$$
$$D_2 A^T D_1 e = c$$

if and only if there exists a matrix B with $Be = r$ and $B^T e = c$ and the same pattern as A. Here, $e = (1, \ldots, 1)^T$ which means that r contains the row sums of the scaled matrix and c contains the column sums.

Furthermore, if the matrix has only positive entries, D_1 and D_2 are unique up to a constant factor.

In Section 4, we give maximal formulations of this theorem. Some immediate questions emerge, such as: How to compute D_1, D_2 and the scaled matrix? Can this be generalised to arrays of higher dimension? All of these questions and many more have been answered in the literature.

Given a matrix $A \in \mathbb{R}^{n\times n}$ with positive entries and the task to scale A to given row sums r and column sums c, one is very naturally lead to the following approximation algorithm:

Algorithm 3.2 (RAS method)**.** Given $A \in \mathbb{R}^{m\times n}$, do:

- Multiply each row j of A with $r_j / \left(\sum_i A_{ij}\right)$ to obtain $A^{(1)}$ with row sums r.
- Multiply each column j of $A^{(1)}$ with $c_j / \left(\sum_i A_{ji}\right)$ to obtain $A^{(2)}$ with column sums c.
- If the row sums of $A^{(2)}$ are very far from r, repeat steps one and two.

If the algorithm converges, the limit B will be the scaled matrix. However, there is a priori no guarantee that D_1, D_2 exist, in which case we can only ask for *approximate scaling*, i.e. matrices D_1, D_2 such that $D_1 A D_2 \approx B$.

3.1. Historical remarks

The iterative algorithm 3.2 is extremely natural and it is therefore not surprising that it was rediscovered several times. It is at least known as *Kruithof's projection method* (Krupp 1979) or *Kruithof double-factor model* (especially in the transportation community; Visick et al. 1980), the *Furness (iteration) procedure* (Robillard and Stewart 1974), *iterative proportional fitting procedure (IPFP)* (Ruschendorf 1995), the *Sinkhorn-Knopp algorithm* (Knight 2008), the *biproportional fitting procedure* (in the case of $r = c = e$; Bacharach 1970) or the *RAS method* (especially in economics and accounting; Fofana, Lemelin, and Cockburn 2002). Sometimes, it is also referred to simply as *matrix scaling* (Rote and Zachariasen 2007), which is mostly used as the term for scalings of the form DAD^{-1}, or *matrix balancing*, which is mostly used for scalings to equal row and column sums. The algorithm is a special case of a number of other algorithms such as Bregman's balancing method (cf. Lamond and Stewart 1981) as we will see later on.

When was interest sparked in the RAS method and diagonal equivalence? The earliest claimed appearance of the model dates back to at least the 30s and Kruithof's use of the method in telephone forecasting (Kruithof 1937). At a similar time, according to Bregman 1967, the Soviet architect Sheleikhovskii considered the method. Sinkhorn 1964 claims that when he started to evaluate the method, it had already been proposed and in use. His example is the unpublished report Welch unknown. Bacharach 1970 acknowledges Deming and Stephan 1940 in transportation science, who popularised the RAS method in the English speaking communities.

None of these approaches seem to have been thoroughly justified. Bacharach notes that Deming and Stephan only propose an ad-hoc justification for using their method to study their problem, which turned out to be wrong (cf. Stephan 1942). He further claims that the first well-founded approach to use the RAS - this time in economics - was given by Richard Stone, who also coined the name "RAS model" (Bacharach cites

Stone 1962, although the name RAS must have occurred earlier as it already occurs in Thionet 1961 without attribution and explanation). However, one can argue that the first justified approach occurred earlier: Schrödinger 1931 had already posed a question regarding models of Brownian motion when given a priori estimates, which led to a similar problem. His approach was justified, albeit the ultimate justification in terms of large deviation theory needed to wait for the development of modern probability theory (cf. Georgiou and Pavon 2015). The problem boils down to solving a continuous analogue of Sinkhorn's theorem, which leads to the matrix problem using discrete distributions (essentially similar to Hobby and Pyke 1965) and was first attacked in Fortet 1940 using a fixed point approach similar to Algorithm 3.13.

However, none of the original papers provided a convergence proof with the possible exception of Fortet 1940[2]. As noted by Fienberg 1970, after Deming and Stephan provided their account, their community started to develop the ideas, but a proof was still lacking (Smith 1947; El-Badry and Stephan 1955; Friedlander 1961).

Summarising the last paragraphs, the RAS method was discovered independently for different reasons in the 30s to 40s, although none of the authors provided a proof (with the possible exception of Fortet). A more theoretical analysis developed in the 60s after Stone's results in economics (e.g. Stone 1962) and Sinkhorn 1964 in statistics and algebra. Since then, a large number of papers has been published analysing or applying Theorem 3.1. Every decade since the sixties contains papers where proving the theorem or an extension thereof is among the main results (examples are Sinkhorn 1964; Macgill 1977; Pretzel 1980; Borobia and Cantó 1998; Pukelsheim and Simeone 2009; Georgiou and Pavon 2015).

Many authors are aware of at least some other attempts, but only a few try to give an overview.[3] The situation is further complicated by a the fact that the technical answer to the question of scalability is tightly linked with the question of patterns, which has a rich history in itself, probably starting with Fréchet (overview of a long line of work in Fréchet 1960).

The last point is particularly interesting: In fact, one could summarise matrix scaling matrix scaling as follows: Given a nonnegative matrix A it is scalable to a matrix B fulfiling some constraints (mostly linear but some nonlinear constraints are allowed), a matrix is scalable with diagonal matrices (in different ways, mostly D_1AD_2 where D_1 and D_2 need not be independent) if and only if there exists a matrix C with the same pattern as A fulfiling the constraints.

[2]The notation and writing is very difficult to read today, so I am not entirely sure whether the proof is correct and captures the case we are interested in.

[3]This suggests once again that the problem had a very complicated history which also makes it difficult to find out whether a problem has already been solved in the past. Several authors have attempted more complete historical overviews such as Fienberg 1970; Macgill 1977; Schneider and Zenios 1990; Brown, Chase, and Pittenger 1993; Kalantari and Khachiyan 1996; Kalantari et al. 2008; Pukelsheim and Simeone 2009. In Rothblum and Schneider 1989, the authors claims that a colleague collected more than 400 papers on the topic of matrix scaling.

Today, proofs and generalisations of Theorem 3.1 and similar questions about scaling matrices in the form DAD or DAD^{-1} form a knot of largely interconnected techniques. We will now try to give an overview of these results and highlight their connections. A graphical overview is presented in Figure 3.1.

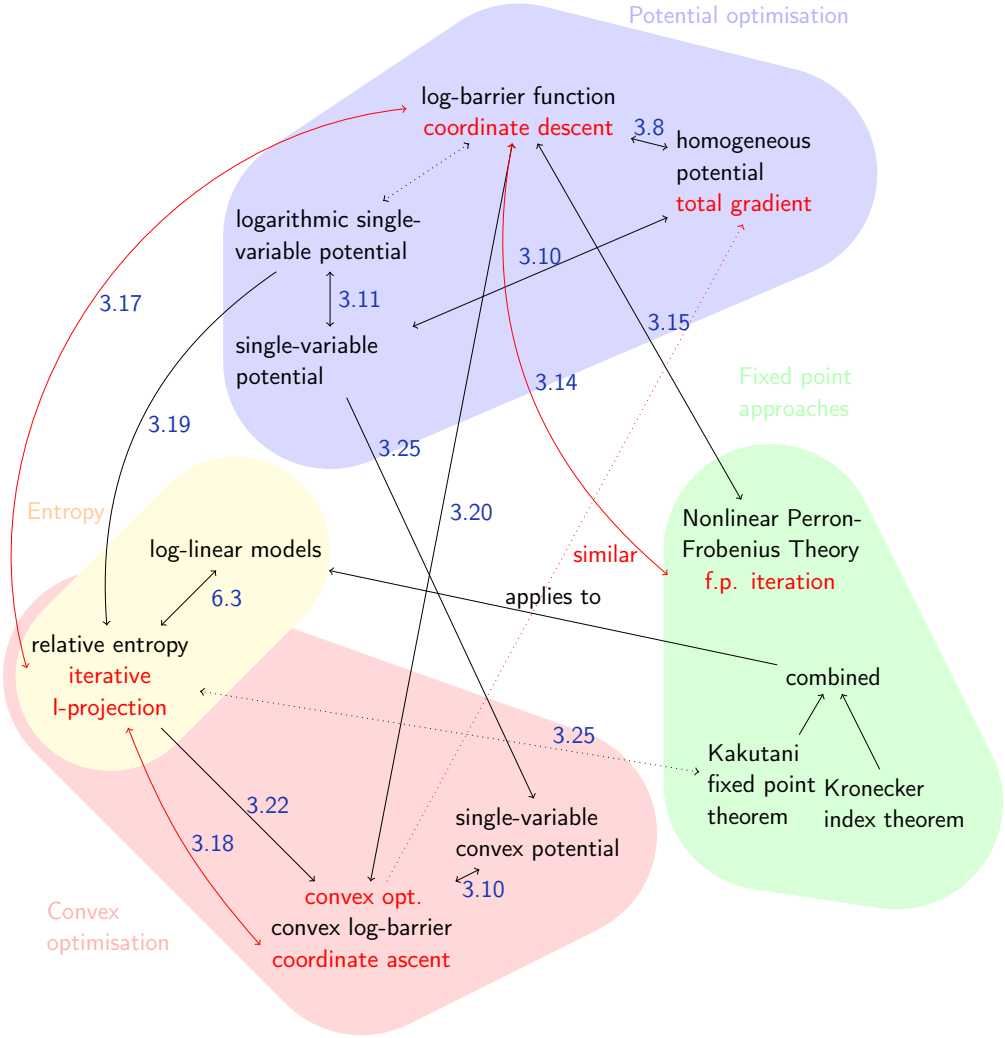

Figure 1: Connected approaches to prove Theorem 3.1 and their relationships. Red arrows and text denote natural algorithms and their connections.

3.2. The logarithmic barrier function

Potentials and barrier functions have been important in the study of matrix scaling since at least the unpublished results of Gorman 1963. Here, we largely follow Kalantari and Khachiyan 1996, who give a very lucid account about the interconnections between

different barrier function formulations for g.

Let $A \in \mathbb{R}^{n\times n}$ be a matrix with nonnegative entries and $r, c \in \mathbb{R}^n_+$. Define the *logarithmic barrier function*

$$g(x,y) = y^T Ax - \sum_{i=1}^{n} c_i \ln x_i - \sum_{i=1}^{n} r_i \ln y_i \tag{2}$$

If we take partial derivatives, we obtain

$$\begin{aligned} \partial_{y_i} g(x,y) &= Ax - r_i/y_i \\ \partial_{x_i} g(x,y) &= y^T A - c_i/x_i \end{aligned} \tag{3}$$

which implies that for any stationary point we have

$$\sum_j A_{ij} x_j y_i = r_i \qquad \sum_j A_{ji} x_i y_j = c_i$$

and setting $D_1 = \operatorname{diag}(y)$ and $D_2 = \operatorname{diag}(x)$ solves the scaling problem. Conversely, any scaling gives a stationary point of the logarithmic barrier function. In summary:

Lemma 3.3. *Given $A \in \mathbb{R}^{n\times n}$ nonnegative and two vectors $r, c \in \mathbb{R}^n_+$, then the matrix can be diagonally scaled to a matrix B with row sums r and column sums c if and only if the corresponding logarithmic barrier function (2) has a stationary point.*

According to Macgill 1977, this observation was first made by Gorman 1963 who also gave the first complete and correct proof. However, the paper only circulated privately. Gorman apparently did not consider this scaling function directly but used an approach similar or identical to the ones considered in convex geometry described in Section 3.5.

The potential barrier function can also be seen from the perspective of Lagrangian multipliers:

Lemma 3.4 (Marshall and Olkin 1968)**.** *Given $A \in \mathbb{R}^{n\times n}$ nonnegative and two vectors $r, c \in \mathbb{R}^n_+$, then the matrix can be diagonally scaled to a matrix B with row sums r and column sums c if and only if on the region*

$$\Omega := \left\{ (x,y) \middle| \prod_{i=1}^{m} x_i^{c_i} = \prod_{i=1}^{m} y_i^{r_i} = 1, x_i > 0, y_i > 0 \right\} \tag{4}$$

the function $y^T Ax$ is bounded away from zero and is unbounded whenever $\|x\|_\infty + \|y\|_\infty \to \infty$. The function $y^T Ax$ then attains a minimum defining D_1 and D_2.

This was used to prove our Theorem 3.1 in Marshall and Olkin 1968. We observe:

Observation 3.5. Lemma 3.4 and 3.3 are equivalent: The logarithmic barrier function is the Lagrange function of the optimisation problem in Lemma 3.4.

Now consider $g(x,y)$ for a fixed x. Since $(-\ln)$ is a convex function and x^TAy is linear in y, g is convex in y. The same holds for a fixed y, i.e. g is convex in both direction. It is then natural to consider the *coordinate descent algorithm* (for an introduction and overview see Wright 2015):

Algorithm 3.6. Given a nonnegative matrix A, take a starting point for g, e.g. $x_0 = y_0 = e$ and iterate:

1. For fixed y_n, find x_{n+1} by searching for the minimum of $g(x, y_n)$.
2. For fixed x_{n+1}, find y_{n+1} by searching for the minimum of $g(x_{n+1}, y)$.
3. Repeat until convergence.

It is possible to solve $\min_x g(x,y)$ or $\min_y g(x,y)$ analytically:

$$x_{n+1} = p/(Ay_n), \qquad y_{n+1} = q/(Ax_{n+1}).$$

This leads to the following observation:

Observation 3.7 (Kalantari and Khachiyan 1996)**.** Algorithm 3.6 and 3.2 are the same.

Proof. Define $D_n^{(1)} := \operatorname{diag}(y_n)$ and $D_n^{(2)} := \operatorname{diag}(x_n)$. Then we have $D_{n+1}^{(1)} A D_n^{(2)} e = r$ and $e^T D_n^{(1)} A D_n^{(2)} = c^T$, which implies that we perform successive row- and column normalisations as in the RAS method. □

Using the fact that the algorithm is a coordinate descend method, one can obtain a convergence proof including a discussion of convergence speed of this algorithm and a dual algorithm (Luo and Tseng 1992). See also Observation 3.17 for a discussion of coordinate ascent methods.

However, g is not jointly convex. For a purely (jointly) convex reformulation, consider the minimum for t along any line $g(tx, ty)$, where g is convex. If we define

$$k(x,y) := \min_{t>0} g(tx, ty) \tag{5}$$

minimising $k(x,y)$ is still equivalent to minimising $g(x,y)$. The corresponding k will be homogeneous and the domain for minimisation will in fact be compact.

Observation 3.8 (Kalantari and Khachiyan 1996)**.** We obtain:

$$k(x,y) = \min_{t>0} \left(t^2 y^T A x - 2n \ln t - \sum_{i=1}^n c_i \ln x_i - \sum_{i=1}^n r_i \ln y_i \right) \tag{6}$$

$$= \ln \left(\frac{(y^T A x)^n}{\prod_{i=1}^n x_i^{c_i} \prod_{j=1}^n y_j^{r_j}} \right) + n - n \ln(n) \tag{7}$$

hence minimising g is equivalent to minimising k.

This proves the following lemma:

Lemma 3.9. *Given a nonnegative matrix $A \in \mathbb{R}^{n\times m}$, it can be scaled to a matrix with row sums r and column sums c if and only if the minimum of $k(x,y)$ exists and is positive. The corresponding minima (x,y) define the diagonal matrices to achieve the scaling.*

The function k is also similar to Karmakar's potential function for linear programming and Algorithm 3.2 is the coordinate descent method for this function (Kalantari and Khachiyan 1996; Kalantari 1996).

Setting $y(x) = (Ax)^{-1}$, we arrive at another formulation of the problem. In the doubly stochastic case, this formulation is due to Djoković 1970; London 1971 and was later adapted to arbitrary column and row sums in Sinkhorn 1974[4]:

Lemma 3.10. *Let $A \in \mathbb{R}^{n\times n}$ be a nonnegative matrix. There exists a scaling to a matrix with row sums r and column sums c iff the infimum*

$$\inf\left\{\prod_{i=1}^{n}\left(\sum_{j=1}^{n}A_{ij}x_j\right)^{r_i}\middle|\prod_{i=1}^{n}x_i^{c_i}=1\right\} \tag{8}$$

is attained on $x,y \in \mathbb{R}_+$.

Observation 3.11. Note that the infimum is attained iff the infimum

$$\inf\left\{\sum_{i=1}^{n}r_i\ln\left(\sum_{j=1}^{n}A_{ij}x_j\right)\middle|\sum_{i=1}^{n}c_i\ln x_i=0\right\}$$

is attained. This is the formulation in Lemma 3.4.

Finally, let us sketch a proof using potential methods.

Sketch of proof of Theorem 3.1 (Potential version). We sketch a proof for arbitrary row and column sums based on the short proof of Djoković 1970 for doubly stochastic scaling: First assume that $A \in \mathbb{R}^{m\times n}$ is a positive matrix. Starting with equation (8) we define the function

$$f(x_1,\ldots,x_n) := \frac{\prod_{i=1}^{m}\left(\sum_{j=1}^{n}A_{ij}x_j\right)^{r_i}}{\prod_{i=1}^{n}x_i^{c_i}}$$

on the set of x_i with $x_i > 0$ and $\sum_i x_i = 1$. Consider an arbitrary point b on the boundary (i.e. $b_i = 0$ for at least one $i \in 1,\ldots n$). For $x_i \to b_i$, since $\prod_i x_i = 0$ and $\sum_j A_{ij}x_j \neq 0$

[4] later studied in Krupp 1979, who used an entropic approach for the generalised problem and in Berger and Kelley 1979, who used a direct convergence approach reminiscent of Sinkhorn and others.

always, we have that $f(x_1, \ldots, x_n) \to \infty$. Hence the function takes its minimum in the interior. At the minimum, the partial derivatives must vanish and we obtain:

$$\begin{aligned} 0 \overset{!}{=} \partial_{x_l} f = \prod_{i=1}^{m} \left(\frac{\left(\sum_{j=1}^{n} A_{ij} x_j\right)^{r_i}}{x_i^{c_i}} \right) \left(\sum_{k=1}^{m} \left(\frac{x_k^{c_k}}{\left(\sum_{p=1}^{n} A_{kj} x_j\right)^{r_k}} \right) \cdot \right. \\ \left. \left(\frac{r_k \left(\sum_{p=1}^{n} A_{kj} x_j\right)^{r_k - 1} A_{kl}}{x_k^{c_k}} - \frac{c_l \left(\sum_{p=1}^{n} A_{kj} x_j\right)^{r_k}}{x_l^{c_l+1}} \delta_{kl} \right) \right) \\ = \prod_{i=1}^{m} \left(\frac{\left(\sum_{j=1}^{n} A_{ij} x_j\right)^{r_i}}{x_i^{c_i}} \right) \left(\sum_{k=1}^{m} A_{kl} r_k \left(\sum_{p=1}^{n} A_{kj} x_j \right)^{-1} - \sum_{k=1}^{m} \frac{c_l x_k^{c_k}}{x_l^{c_l+1}} \delta_{kl} \right). \end{aligned}$$

If we take all conditions for $l = 1, \ldots n$, then this is equivalent to the condition

$$A^T(r/(Ax)) = c/x$$

which boils down to equations (3).

The more technical part for nonnegative matrices is a more careful analysis of what happens for nonnegative matrices that are not positive. For doubly stochastic matrices, we can use the fact that fully indecomposable matrices have a positive diagonal, which implies once again that $\prod_i \sum_j A_{ij} b_j \neq 0$. A similar argument can be made for arbitrary patterns, but we leave it out in this sketch. □

3.3. Nonlinear Perron-Frobenius theory

Another early approach uses nonlinear Perron-Frobenius theory which is essentially a very general approach to tackle fixed point problems for (sub)homogeneous maps on cones. A short overview is given in appendix B. The basic idea is given by:

Lemma 3.12 (Brualdi, Parter, and Schneider 1966)**.** *Given a nonnegative matrix $A \in \mathbb{R}^{n \times n}$, there exists a scaling of A to a matrix with row sums r and column sums c if and only if the following map has a fixed point $x > 0$:*

$$\begin{aligned} \mathbf{T} : \mathbb{R}^n &\to \mathbb{R}^n \\ \mathbf{T}(x) &= c/(A^T(r/(Ax))) \end{aligned} \tag{9}$$

This also suggests another simple algorithm:

Algorithm 3.13. Given a nonnegative matrix A. Let $x_0 = e$. Iterate until convergence:

$$x_{n+1} = \mathbf{T}(x_n). \tag{10}$$

The development of this idea that started with Menon 1967 and was used to provide a full proof of Theorem 3.1 in Brualdi, Parter, and Schneider 1966 for doubly stochastic matrices. Menon and Schneider 1969 consider arbitrary row- and column sums and give a complete study of the spectrum of the Menon-operator. Some contraction properties were used to give a direct proof of convergence of the RAS algorithm in Berger and Kelley 1979. The connection to Hilbert's projective metric, and therefore to "Nonlinear Perron-Frobenius theory" (cf. Lemmens and Nussbaum 2012), became clear later on and allowed to give upper bounds on the covergence speed of the RAS (Franklin and Lorenz 1989; Georgiou and Pavon 2015).

However, Menon was not the first to define the operator **T**: Looking closely at the arguments given in Fortet 1940, one can see the continuous version of **T**, which lead to an independent rediscovery of **T** and its connection to the Hilbert metric in Georgiou and Pavon 2015. Probably, Menon was not even the first to define the discrete version of the operator and to note that the existence of a fixed point can be seen by invoking Brouwer's fixed point theorem. This dates back to Thionet 1963; Thionet 1964, building on work about matrix patterns (Thionet 1961). According to Caussinus 1965, Thionet 1964 was also the first paper to conjecture the necessary and sufficient conditions for scalability[5]. The ideas where rediscovered another time in Balinski and Demange 1989b, where the authors used the fixed point argument to prove that a scaling exists and fulfils their axiomatic approach.

Let us connect the approach to Section 3.2. First note that the algorithm is nothing else but a slight variation of the RAS method:

Observation 3.14. Setting $y_{n+1} := r/(Ax_n)$ and $x_{n+1} := c/(A^T y_{n+1})$ we can immediately see that one iteration of Algorithm 3.13 is one complete iteration of the RAS method 3.2.

The connection with the logarithmic barrier method is also very close:

Observation 3.15. Any fixed point of the Menon operator defines a stationary point of the logarithmic barrier function (2) and vice versa.

Proof. Let A be a nonnegative matrix. The derivative conditions for the stationary points of (2) are given in equation (3), which are equivalent to:

$$Ax = r/y \quad y^T A = c/x \tag{11}$$

This implies immediately that $x = p/(A^T(q/Ax))$, hence x is a fixed point of **T**. Similarly, any positive fixed point immediately gives a scaling as a minimum of the logarithmic barrier function. □

This also proves Lemma 3.12.

[5]He also notes that the early history around Deming and Stephan 1940 is a little bit curious, since the authors claim to have a convergence proof but never publish it.

Sketch of proof of Theorem 3.1 (Nonlinear Perron-Frobenius theory version). Let us first assume A has only positive entries. Then $\mathbf{T}$ sends all vectors $x \in \mathbb{R}^n_{+0}$ to $\mathbb{R}^n_+$ hence using Brouwer's fixed point theorem $\mathbf{T}$ has a positive fixed point. Note that in order to apply Brouwer's fixed point theorem, we need to have a compact set. To achieve this, consider the operator $\tilde{\mathbf{T}}(x) = \mathbf{T}(x)/\sum_{i=1}^n \mathbf{T}(x)_i$.

For general nonnegative matrices A, one can extend $\mathbf{T}$ to be a map from $x \in \mathbb{R}^n_{+0}$ into itself (see also Appendix B) either by a general argument (see Theorem B.8) or by defining $\infty \cdot 0 = 0$ and $\infty \cdot c = \infty$ for all positive c. One can easily see that $\mathbf{T}$ will not send any entry to ∞.

However, it is not immediately clear when the fixed point is positive if A contains zero-entries. This is the main technical difficulty for a complete proof. Brualdi, Parter, and Schneider 1966 show that if A is fully indecomposable, $\mathbf{T}(x)$ has at least $k+1$ entries which are nonzero if x has exactly k entries which are nonzero, which immediately proves that the fixed point must be positive.

Upon closer observation, the map is contractive under Hilbert's metric and Banach's fixed point theorem immediately provides existence and uniqueness of the scaled matrix. The fixed point itself provides the diagonal of D_2. □

3.4. Entropy optimisation

Another approach, which underlies many justifications for applications, considers entropy minimisations under linear constraints. An overview of entropy minimisation and its relation to diagonal equivalence can be found in Brown, Chase, and Pittenger 1993, a broader overview about the relation of the RAS algorithm to entropy scaling with a focus on economic settings can be found in McDougall 1999.

To formulate the problem, we define the *Kullback-Leibler divergence*, *I-divergence* or *relative entropy*, which was first described in Kullback and Leibler 1951 (see also Kullback 1959) for two vectors $x, y \in \mathbb{R}^n_{+0}$:

$$D(x\|y) := \sum_{j=1}^{n} x_j \ln\left(\frac{x_j}{y_j}\right) \tag{12}$$

where we use the convention that the summand is zero if $x_j = y_j = 0$ and infinity if $x_j > 0, y_j = 0$. The relative entropy is nonnegative and zero if and only if $x = y$ and it is therefore similar to a distance measure. Given a set, what is the smallest "distance" of a point to this set in relative entropy? This is known as *I-projection* (cf. Csiszár 1975).

Let A be a nonnegative matrix and define

$$\Pi_1 := \{B | Be = r\}$$
$$\Pi_2 := \{B | e^T B = c^T\}.$$

We ask for the I-projection of A onto the set $\Pi_1 \cap \Pi_2$, i.e. we want to find A^* such that

$$D(A^*\|A) = \inf_{B \in \Pi_1 \cap \Pi_2} D(B\|A). \tag{13}$$

The connection to scaling was probably first used in Brown 1959, where the RAS method is used to improve an estimate for positive probability distributions of dimensions $2 \times 2 \times \ldots \times 2$ in the relative entropy measure (Brown cites Lewis 1959 as a justification for his approach, where the relative entropy is justified as a "closeness" measure). According to Fienberg 1970, this approach was later generalised to all multidimensional tables in Bishop 1967 based on some duality of optimisation by Good 1965[6]. Another early use of relative entropy occurs in Uribe, Leeuw, and Theil 1966 (see also Theil 1967), where it was noted without proof that the results were the same as the RAS.

A very natural approach to obtain A^* would be to try an iterative I-projection:

Algorithm 3.16. Let A be nonnegative.

- Let $A^{(0)} = A$.
- If n is even, find $A^{(n+1)}$ such that
$$D(A^{(n+1)} \| A^{(n)}) := \inf_{B \in \Pi_1} D(B \| A).$$
- If n is odd, find $A^{(n+1)}$ such that
$$D(A^{(n+1)} \| A^{(n)}) := \inf_{B \in \Pi_2} D(B \| A)$$
- Repeat the steps until convergence.

Observation 3.17 (cf. Csiszár 1975; Csiszár 1989)**.** The algorithms 3.16 and 3.2 are the same.

Proof. This was first shown in Ireland and Kullback 1968. We give a short argument based on Lagrangian multipliers restricted to column normalisation. Given $A \in \mathbb{R}^{n \times n}$, the Lagrangian for the problem is

$$L(B, \lambda) := D(B \| A) + \lambda_j \left(\sum_i A_{ij} - q_j \right).$$

Partial derivatives $\partial_{B_{ij}} L = 0$ and $\partial_{\lambda_j} L = 0$ lead to the system of equations:

$$\ln \left(\frac{B_{ij}}{A_{ij}} \right) + 1 + \lambda_j = 0 \qquad i, j = 1, \ldots, n$$
$$\sum_i A_{ij} - c_j = 0 \qquad j = 1, \ldots, n.$$

A solution is easily seen to be

$$B_{ij} = A_{ij} \frac{c_j}{\sum_k A_{kj}} \qquad i, j = 1, \ldots, n$$

The latter is the column renormalisation as in the RAS (Alg. 3.2). □

[6] Both references were not available to me.

This implies that if the iterated I-projection converges to the I-projection of (13), then matrix scalability solves equation (13). This was supposedly proved in Ireland and Kullback 1968[7] and Kullback 1968, but the proofs contain an error as pointed out in Csiszár 1975 (see also Brown, Chase, and Pittenger 1993). A corrected proof appeared in Csiszár 1975, however for some of the theorems it is not immediately clear whether more assumptions are needed as noted in Borwein, Lewis, and Nussbaum 1994.

In addition, the proof in Aaronson 2005 for positive matrices proves that the RAS converges using relative entropy as a "progress measure". He shows that it decreases under RAS steps to a unique stationary point. Another direct proof appeared in Franklin and Lorenz 1989.

At this point, let us make the following observation:

Observation 3.18 (Cottle, Duvall, and Zikan 1986)**.** The RAS method can also be seen as the coordinate ascent method to the dual problem of entropy minimisation.

This is justified as follows: When deriving the I-projections of each step of the algorithm, we set up the Lagrangian

$$L(B, \lambda) := D(B\|A) + \lambda_j \left(\sum_i A_{ij} - c_j \right)$$

and calculate its solution. This consists in explicitly solving the resulting equations for the Lagrangian multipliers λ_j. In this sense, the algorithm is not really a primal problem. This is also consistent with the nomenclature above: In Section 3.5 we see that the dual problem of entropy minimisation is a convex program that is basically just the (negative) logarithmic barrier function above. Since the RAS is the coordinate descent algorithm of this problem, it is the coordinate ascent method of the dual problem of entropy minimisation.

In other word, the justification of this observation is due to:

Observation 3.19 (Georgiou and Pavon 2015; Gurvits 2004)**.** Given a matrix $A \in \mathbb{R}^{n\times n}$ with nonnegative entries. Suppose there exist positive diagonal matrices such that $D_1 A D_2$ has row sums r and column sums c, then

$$-\ln\left(\inf\left\{ (\prod_{i=1}^n r_i \sum_{j=1}^n A_{ij}x_j) \middle| \prod_{i=1}^n x_i^{c_i} = 1 \right\} \right) = \inf\{D(B\|A) | Be = r, B^T e = c\} \tag{14}$$

and in particular, the minimum is the scaled matrix.

The proof of this observation will essentially follow from the results in Section 3.5.

Let us finish this section by giving another proof sketch of Sinkhorn's theorem:

[7] In Fienberg 1970, it is pointed out that a simplified version of this proof appeared in Dempster 1969, which however is unavailable to me.

Sketch of proof of Theorem 3.1 (Entropic version). We sketch the proof given in Csiszár 1975 restricted to our scenario, which is similar to the proof in Darroch and Ratcliff 1972 (see also Csiszár 1989 for a comment on the connection). We prove convergence of Algorithm 3.16, essentially by showing that the relative entropy of two successive iterations decreases to zero.

Given a nonnegative matrix A, assume that there exists a matrix $B \prec A$ with required row- and column sums. Otherwise, the relative entropy will always be infinite and the problem has no solution.

The crucial observation is that if A' is the I-projection of A onto Π, then for any $B \in \Pi$ we have

$$D(B\|A) = D(B\|A') + D(A'\|A). \tag{15}$$

This "Pythagorean identity" usually only holds with $\geq$. The equality case is a special case of the "minimum discrimination principle" (Kullback 1959; Kullback and Khairat 1966) and it is proven for constraints Π_i in Csiszár 1975. This equality leads to a very useful transitivity result (see also Ku and Kullback 1968) stating that if A has I-projection B on Π_i for some i and I-projection B' on Π, then B has I-projection B' on Π. This is not necessarily true in the general case.

Let A' be the I-projection of A onto Π. Denoting by $A^{(n)}$ the repeated I-projection as defined in Algorithm 3.16, repeated application of equation (15) shows

$$D(A'\|A) = D(A'\|A^{(n)}) + \sum_{i=1}^{n} D(A^{(n)}\|A^{(n-1)})$$

Therefore, the sequence $A^{(n)}$ lies in a bounded set and hence contains a convergent subsequence by compactness. However, we also have that $D(A^{(n)}\|A^{(n-1)}) \to 0$ for $n \to \infty$, which implies $\|A^{(n)} - A^{(n-1)}\|_\infty \to 0$ for $n \to \infty$, hence $A^{(n)}$ converges to some matrix A''. Clearly, $A'' \in \Pi$, since $A^{(2n)} \in \Pi_1$ and $A^{(2n+1)} \in \Pi_2$ for every $n \in \mathbb{N}$. Using the transitivity of the I-projection, A'' is the I-projection of $A^{(n)}$ for all n and equation (15) holds in the form

$$D(A''\|A^{(n)}) = D(A''\|A') + D(A'\|A^{(n)})$$

Since the first and last term converge to zero, $D(A''\|A') = 0$ and the I-projection A' is indeed the limit of Algorithm 3.16. □

A similar proof can be found in Brown, Chase, and Pittenger 1993.

3.5. Convex programming and dual problems

Recall the logarithmic barrier function g in equation (2) and that it is not jointly convex. However, it is very beneficial to make g convex for several reasons:

1. Convex programming is efficient in the complexity theoretic sense (Boyd and Vandenberghe 2004).

2. The duality theory for convex programming is very well developed and can lead to new algorithms (see littleO 2014 for a heuristic introduction and Rockafellar 1997; Boyd and Vandenberghe 2004 for a more careful analysis).

3. Uniqueness proofs can become simpler: A convex function has a unique minimum iff it is strictly convex at the minimum.

To obtain a convex program, one simply needs to substitute $x = (e^{\xi_1}, e^{\xi_2}, \ldots, e^{\xi_n})$ and $y = (e^{\eta_1}, e^{\eta_2}, \ldots e^{\eta_n})$ into g to obtain (Macgill 1977; Kalantari and Khachiyan 1996):

Lemma 3.20. *Given a nonnegative matrix $A \in \mathbb{R}^{n\times n}$, one can find diagonal matrices to scale A to a matrix with row-sum r and column sum c if and only if the function*

$$f(\xi, \eta) := \sum_{ij=1}^{n} A_{ij} e^{\eta_i + \xi_j} - \sum_{i=1}^{n} r_i \xi_i - \sum_{j=1}^{n} c_j \eta_j \tag{16}$$

attains its minimum on $\xi, \eta \in \mathbb{R}^n_{-0}$.

A proof based on this approach can be found in Bachem and Korte 1979.[8] We have already seen:

Observation 3.21. The convex programming formulation in Lemma 3.20 is equivalent to the logarithmic barrier function approach in Lemma 3.3.

Likewise, it can be shown:

Observation 3.22. The convex programming formulation 3.20 is the Wolfe dual (Macgill 1977; Krupp 1979) or Lagrangian dual (Balakrishnan, Hwang, and Tomlin 2004) of the entropy minimisation approach.

Proof. The entropy minimisation problem was given as:

$$\begin{aligned} &\inf_{B_{ij}} \sum_{ij} B_{ij} \ln(B_{ij}/A_{ij}) \\ &\text{s.t.} \sum_{i} B_{ij} = p_j \quad \sum_{j} B_{ij} = q_i \end{aligned}$$

This implies that the Wolfe dual is given by

$$\sup_{B_{ij}} \sum_{ij} B_{ij} \ln(B_{ij}/A_{ij}) + \sum_{j} u_j \left(\sum_{i} B_{ij} - p_j \right) + \sum_{i} v_i \left(\sum_{j} B_{ij} - q_i \right)$$

[8] In Bacharach 1970 it is also noted that the function is used in the approach by Gorman 1963 later to be simplified by Bingen 1965. Both papers are unavailable to me.

$$\text{s.t.} \ln(A_{ij}/B_{ij}) + 1 + u_j + v_i = 0 \quad \forall i,j$$
$$u, v \geq 0$$

The constrained can be rewritten as

$$B_{ij} = A_{ij} \exp(-1 - u_j - v_i)$$

and inserting this into the Wolfe dual function (see e.g. Bot and Grad 2010) we obtain:

$$\sup_{B_{ij}} \sum_{ij} B_{ij} \ln(B_{ij}/A_{ij}) + \sum_j u_j \left(\sum_i B_{ij} - p_j \right) + \sum_i v_i \left(\sum_j B_{ij} - q_i \right)$$
$$= \sup_{u,v} - \left(\sum_{ij} A_{ij} \exp(-1 - u_j - v_i) - \sum_j u_j p_j - \sum_i v_i q_i \right)$$

which is (up to the constant $\sum_{ij} A_{ij}/e$) the optimisation problem in 3.20. The calculation for the Lagrangian dual is similar (see Balakrishnan, Hwang, and Tomlin 2004). □

Another connection to the barrier function is to use *geometric programming:*

Observation 3.23. Minimisation of the logarithmic barrier function g is equivalent to

$$\min \ y^T A x$$
$$\text{s.t.} \prod_{i=1}^n x_i^{c_i} = 1, \prod_{i=1}^n y_i^{r_i} = 1$$

This is in standard form of a geometric program, which implies that a substitution $\xi = \ln(x), \eta = \ln(y)$ gives a convex program (Boyd and Vandenberghe 2004, Section 4.5.3).

This observation was made in Rothblum and Schneider 1989, which also gives necessary and sufficient conditions for a matrix to be scalable or approximately scalable.

As described in Kalantari and Khachiyan 1996, one can also reduce the problem to an unconstrained optimisation problem for only a single variable by taking the formulation of Lemma 3.10 and substituting $x = \exp(\xi)$ as above to obtain the minimising function

Lemma 3.24. *Given a nonnegative matrix $A \in \mathbb{R}^{n \times n}$, one can find diagonal matrices to scale A to a matrix with row sums r and column sums c if and only if the function*

$$f(\xi) = \sum_{j=1}^n r_j \ln \left(\sum_{i=1}^n A_{ij} e^{\xi_i} \right) - \sum_{i=1}^n c_i \xi_i. \tag{17}$$

attains its minimum on $\xi \in \mathbb{R}^n{}_0$.

Finally, let us return to entropy minimisation: Relative entropy is jointly convex and therefore a convex program. In fact, relative entropy is a special case of a broader class of functions called *Bregman divergences* which we will sketch in Section 6.5.

A proof of Theorem 3.1 using convex programming is often similar to the approach in Section 3.2. The advantage is that any critical point is automatically a minimum and one does not need to consider the boundary.

3.6. Topological (non-constructive) approaches

In the proof of Theorem 3.1 in Section 3.3, the result was achieved by Brouwer's fixed point theorem, but it is only one of many topological proofs.

For every nonnegative matrix A with a given pattern, we want to decide whether a scaling with prespecified row- and column sums exists. Assume that we also know the set of possible row- and column sums for a given pattern. In a sense, we therefore just have to prove that the map $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ defined via $(D_1, D_2) \mapsto D_1 A D_2$ hits all row- and column sums, or else: we need to see that the map

$$\begin{aligned} \phi' :& \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}^n_+ \times \mathbb{R}^n_+ \\ & (D_1, D_2) \mapsto (p, q) : \; p_i = \sum_j (D_1 A D_2)_{ij}, q_j = \sum_i (D_1 A D_2)_{ij} \end{aligned} \tag{18}$$

is onto. This is somewhat problematic, because the spaces involved are not compact, but by normalising both diagonal matrices and row- and column-sums, one can consider the map as a map from a compact space into itself. This approach was taken in Bapat 1982 for positive matrices (based on his thesis) and the map was shown to be surjective using a topological theorem, which Bapat claims is sometimes known as Kronecker's index theorem.[9] The case for general nonnegative matrices could only be covered by combining the approach with Raghavan 1984 (see Bapat and Raghavan 1989).

Raghavan 1984 uses yet another fixed point theorem (Kakutani's fixed point theorem of set-valued maps). Defining the set K of all matrices in $\mathbb{R}^{n\times m}$ with prescribed marginals and zero (sub)pattern of the a priori matrix A, he considers the map

$$\phi(H) = \{Z | Z \in K, \max_{Z'} \langle C(H), Z' \rangle = \langle C(H), Z \rangle \tag{19}$$

where $C(H)_{ij} = \log(A_{ij}/H_{ij})$ (if $A_{ij} > 0$, and 0 else), we take all matrices as vectors in $\mathbb{R}^{nm}$ and the usual scalar product. The fixed point theorem then implies that there exists H such that

$$\max_{Z' \in K} \langle C(H), Z \rangle = \langle C(H), H \rangle$$

and using the dual of this maximisation, one can show that it scales the matrix.

Observation 3.25. There is a simple connection to entropy minimalization, since $\langle C(H), H \rangle = D(H \| A)$.

However, we can also take the converse road: Instead of exploring the possibilites for every A, we can start with the set of matrices with prescribed row- and column sums

[9] I could not find any other instance of where the theorem is given that name. The theorem simply states that for any map $f : D^{n+1} \to D^{n+1}$, if f maps ∂D^{n+1} into itself and is of nonzero degree, then it must be surjective.

and matrix pattern $\mathcal{X}$ (call the set $\mathcal{M}(p,q,\mathcal{X})$) and map it to the set of all nonnegative matrices of pattern $\mathcal{X}$ (call it $\mathcal{M}(\mathcal{X})$) by diagonal equivalence, i.e. consider the map:

$$\begin{aligned}\psi :&\mathbb{R}^n_+ \times \mathbb{R}^n_+ \times \mathcal{M}(p,q,\mathcal{X}) \to \mathcal{M}(\mathcal{X})\\ &(D_1, D_2, B) \mapsto D_1 B D_2\end{aligned} \tag{20}$$

Again, it would be enough to show surjectivity. As such, it cannot be injective, because we can obviously shift a scalar from D_1 to D_2, hence we would at least have to restrict the first coordinate of D_1 to be 1. The resulting map ψ' is indeed a homeomorphism as shown in an overlooked paper of Tverberg 1976.

Another topological proof has recently been proposed in Friedland 2016. To describe the approach, note that the following two statements are equivalent:

1. There exist D_1, D_2 such that $D_1 A D_2$ has row sums r and column sums c.
2. There exist D_1', D_2' such that $D_1' A D_2'$ is a stochastic matrix with $D_1' A D_2' c = r$.

The proof is trivial, in fact $D_1' = D_1$ and $D_2' = D_2 \operatorname{diag}(1/c)$. In Friedland 2016, the author therefore restricts to stochastic matrices. To do this, he defines the following map:

$$\Phi_A :\mathbb{R}^n_+ \to \mathbb{R}^{n\times n}_+; \quad \Phi_A(x) = \operatorname{diag}(x) A / \operatorname{diag}(A^T x) \tag{21}$$

A quick calculation shows that $\Phi_A^T e = e$, hence the matrix is always stochastic. Hence given a nonnegative matrix A and row sums r and columns sums c, the question of scalability is equivalent to the question whether there exists an $x \in \mathbb{R}^n$ such that $\Phi_A(x)c = r$.

For positive matrices A and any $c \in \mathbb{R}^n_+$, Friedland 2016 now proves scalability by proving that the map $\Phi_{A,c} : x \to \Phi_A(x)c$ is continuous as a set from $\mathbb{R}^n_{+0} \cap \{v | \sum_i v_i = 1\}$ onto itself and a diffeomorphism from $\mathbb{R}^n_+ \cap \{v | \sum_i v_i = 1\}$ onto itself. The result is achieved using degree theory similar to Bapat and Raghavan 1989.

3.7. Other ideas

A very general approach to prove Theorem 3.1 was provided in Letac 1974, where matrix theorems are derived as a consequence of the following theorem:

Theorem 3.26 (Letac 1974)**.** *Let X be a finite set, $(\mu(x))_{x\in X}$ strictly positive numbers and $\mathcal{H}$ a fixed linear subspace of $\mathbb{R}^X$. Then there exists a unique (nonlinear) map from $\mathbb{R}^X \to \mathcal{H}$ denoted $f \mapsto h_f$ such that*

$$\sum_{x\in X} [\exp(f(x)) - \exp(h_f(x))] g(x) \mu(x) = 0$$

for all g in H.

Sinkhorn's theorem follows as an easy corollary:

Sketch of proof of Theorem 3.1, Letac 1974. First, let $X \subset \{1,\dots,m\} \times \{1,\dots n\}$, then we first define the following maps:

$$a : (\mathbb{R}^m, \mathbb{R}^n) \to \mathbb{R}^{m\times n}, \quad (\xi,\eta) \to (\xi_i + \eta_j)_{ij}$$
$$\pi : (\mathbb{R}^{m\times n}) \to \mathbb{R}^X, \quad (A_{ij})_{i=1,j=1}^{i=m,j=n} \to (A_{ij})_{(i,j)\in X}$$

The second is just the natural projection from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^X$.

Now let $A \in \mathbb{R}^{m\times n}$ be a nonnegative matrix and let $X := \{(i,j)|A_{ij} > 0\}$ be its pattern. We already know that the pattern is a necessary condition for scalability, hence we know that there exists a $B \in \mathbb{R}^{m\times n}$ with row sums r and column sums c. Given the pattern, we define $\mathcal{H} = \mathrm{ran}(\pi \circ a)$ the range of the composition of π and a.

Now let $F \in \mathbb{R}^X$ be the matrix with entries $F_{ij} := \log(B_{ij}/A_{ij})$ and apply the theorem to F, i.e. there exists a unique matrix $H \in \mathcal{H}$ such that

$$\sum_{ij} \exp(F_{ij}) A_{ij} G_{ij} = \sum_{ij} \exp(H_{ij}) A_{ij} G_{ij} \quad \forall G \in \mathcal{H}$$

But since $\exp(F_{ij})A_{ij} = B_{ij}$ and $H_{ij} = \xi_i + \eta_j$ for some $\xi \in \mathbb{R}^m$, $\eta \in \mathbb{R}^n$ by definition of $\mathcal{H}$, we have

$$\sum_{ij} B_{ij} G_{ij} = \sum_{ij} \exp(\xi_i) A_{ij} \exp(\eta_j) G_{ij} \quad \forall G \in \mathcal{H}$$

which implies that $\exp(\xi_i)A_{ij}\exp(\eta_j)$ has row sums r by taking $G = \pi \circ a(e_i, 0)$ and column sums c by taking $G = \pi \circ a(0, e_j)$ for the unit vectors $e_i \in \mathbb{R}^m$, $e_j \in \mathbb{R}^n$.

Clearly, the choice of (ξ,η) is unique up to $\ker(\pi \circ a)$, which can be made explicit and leads to the usual conditions. □

3.7.1. Geometric proofs

In principle, we have already two geometric interpretations of the RAS: First, the RAS is akin to iterated I-projections and second, the RAS is the application of a contractive mapping on a cone with a projective metric. Two other "geometric" proofs are known:

Fienberg 1970 shows that the RAS is a contractive mapping in the Euclidean metric using that the RAS preserves *cross-ratios* of a matrix. Given a matrix, the products

$$\alpha_{ijkl} := \frac{A_{ij}A_{kl}}{A_{il}A_{kj}} \tag{22}$$

remain invariant. This was first observed in Mosteller 1968, where it was used to justify the use of the RAS in statistical settings (see Section 8). Fienberg then follows that if one associates any positive matrix to a point of the simplex

$$S_{rc} = \{(A_{11},\dots,A_{1c},\dots,A_{r1},\dots,A_{rc})|\sum_{ij} A_{ij} = 1\}$$

by normalising the matrix, then any point reachable by diagonal equivalence scaling lies on a certain type of manifold inside the simplex. Using some structural knowledge of these manifolds he then shows that each full cycle of the RAS corresponds to a contraction mapping with respect to the Euclidean metric. The result is general enough to cover multidimensional tables, but in this simplicity handles only positive matrices.

There is an interesting connection: While the cross-ratios within the matrix remain constant, Hilbert's metric is also closely connected to cross-ratios. In fact, the contraction ratio is connected to the largest cross-ratio within the matrix and it is not finite if the matrix contains zeros. In that case, the matrix does not easily define a contraction in Hilbert's metric. The same holds true for Fienberg's proof.

Borobia and Cantó 1998 consider the column space of scaled matrices AS and notes that RAS is doubly stochastic if the columns are included in the convex hull of the columns of R^{-1} and the barycentre of the sets of the two columns coincide. The observation of the barycentre then leads them to a proof involving Brouwer's fixed point theorem once again. By some continuity argument, the proof can be extended to nonnegative matrices.

3.7.2. Other direct convergence proofs

Many papers contain direct convergence proofs, not least the original approach in Sinkhorn 1964 and the proof of the full result Sinkhorn and Knopp 1967 (another proof based on this approach is given in Pretzel 1980). The idea is to show that some seemingly unrelated quantity always converges. Often, this quantity turns out to be very much related to some potential barrier function or entropy and we already cited the approach in the corresponding section.

One different proof is the short convergence proof of Macgill 1977 establishing that $\sum_j A_{ij}^{(n)} / \sum_j A_{ij}^{(n-1)} \to 1$ and similarly $\sum_i A_{ij}^{(n)} / \sum_i A_{ij}^{(n-1)} \to 1$ for every i, j. This proof is in some sense derived from Bacharach's approach (Bacharach 1965; Bacharach 1970, see also Seneta 2006) and is very straightforward.[10] In parallel to Bacharach's earlier work Bacharach 1965 but not cited in his later Bacharach 1970, Caussinus 1965 proved the convergence of the RAS method in the general case of multidimensional matrices via the same idea which he attributes to Thionet 1964 (see the appendix of Caussinus 1965).

A second direct proof of convergence in Sinkhorn 1967, uses a norm difference as convergence measure. More precisely, he considers the map

$$\phi(x,y) = \max_i \left(r_i^{-1} \sum_j x_i A_{ij} y_j \right) - \min_i \left(r_i^{-1} \sum_j x_i A_{ij} y_j \right)$$

[10] Macgill also mentions yet another work that contains a proof of Theorem 3.1, namely Herrmann 1973, however no details are given beyond the fact that it contains also approximate scaling.

on the set of all $(x, y) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+$ with some boundedness condition on their entries and proves that $\phi(x, y) = 0$ is achieved by two positive vectors.

A third proof of direct and approximate scaling is given in Pukelsheim and Simeone 2009 by combining the approach of Bacharach with a simple L^1-error function borrowed from Balinski and Demange 1989b.

4. Equivalence scaling

Let us now collect maximal results. A similar but scarcely referenced collection of results was provided in Krupp 1979. We follow the cleaner presentation style of Rothblum and Schneider 1989. Starting with equivalence scaling, we have:

Theorem 4.1. *Let $A \in \mathbb{R}^{n\times m}$ be a nonnegative matrix and $r \in \mathbb{R}^n_+, c \in \mathbb{R}^m_+$. Then the following are equivalent:*

1. *There exist positive diagonal matrices D_1, D_2 such that $D_1 A D_2$ has row sums r and column sums c.*
2. *There exists a matrix B with row sums r and column sums c with the same pattern as A (Menon 1968; Brualdi 1968).*
3. *There exists no pair of vectors $(u, v) \in \mathbb{R}^n \times \mathbb{R}^m$ such that (Rothblum and Schneider 1989)*

$$\begin{aligned} u_i + v_j \geq 0 \quad \forall (i,j) \in \operatorname{supp}(A) \\ r^T u + c^T v \leq 0 \\ \text{either } u_i + v_j > 0 \text{ for some } (i,j) \in \operatorname{supp}(A) \text{ or } r^T u + c^T v < 0 \end{aligned}$$

4. *For every $I \subset \{1, \ldots, m\}, J \subset \{1, \ldots, n\}$ such that $A_{I^c J} = 0$ we have that*

$$\sum_{i \in I} r_i \geq \sum_{j \in J} c_j$$

   and equality holds if and only if $A_{I J^c} = 0$ (Menon and Schneider 1969).

5. *The RAS method converges and the product of the diagonal matrices of the iteration also converges to positive diagonal matrices (Sinkhorn and Knopp 1967).*

The equivalence of the first two items was essentially established in the proof sketches in section (3). The equivalence to the fourth item follows from the characterisation of matrix patterns (see appendix A) and the third follows from studying the geometric program 3.23.

For doubly stochastic scaling, using the classification of doubly stochastic patterns, we then know that scalability is equivalent to having total support (cf. Csima and Datta

1972). The scaling matrices D_1, D_2 are unique up to scalar multiplication if and only if A is fully indecomposable.

For approximate equivalence scaling, the results are similar. The only difference is that certain elements of A can become zero in the limit (which implies that elements of D_i must become zero and others infinite, hence the diagonal matrices cannot exist):

Theorem 4.2. *Let $A \in \mathbb{R}^{n\times m}$ be a nonnegative matrix and $r \in \mathbb{R}^n_+, c \in \mathbb{R}^m_+$. Then the following are equivalent:*

1. *For every $\varepsilon > 0$ there exist diagonal matrices $D_1,\ D_2$ such that $B = D_1 A D_2$ satisfies*
$$\|Be - r\| < \varepsilon, \|B^T e - c\| < \varepsilon$$
2. *There exists a matrix $A' \prec A$ such that A' is scalable to a matrix B with row sums r and column sums c.*
3. *There exists a matrix $B \prec A$ with row sums r and column sums c (Schneider and Saunders 1980).*
4. *There exists no pair of vectors $(u, v) \in \mathbb{R}^{n\times m}$ such that (Rothblum and Schneider 1989)*
$$\begin{aligned} u_i + v_j \geq 0 \quad \forall (i,j) \in \operatorname{supp}(A) \\ r^T u + c^T v < 0 \end{aligned}$$
5. *For every $I \subset \{1,\dots,m\}, J \subset \{1,\dots,n\}$ such that $A_{I^c J} = 0$ we have that*
$$\sum_{i\in I} r_i \geq \sum_{j\in J} c_j$$
6. *The RAS method converges (Sinkhorn and Knopp 1967 for the d.s. case).*

For doubly stochastic scaling, using the classification of doubly stochastic patterns, we have that approximate scalability is equivalent to A having support. Using Schneider and Saunders 1980, this is a trivial consequence of the fact that a matrix has total support if and only if it has doubly stochastic pattern and Proposition A.5[11].

The uniqueness conditions are also simple enough to state:

Theorem 4.3. *Let $A \in \mathbb{R}^{n\times m}$ be a nonnegative matrix and $r \in \mathbb{R}^m_+$, $c \in \mathbb{R}^n_+$. Then, A has at most one scaling.*

Furthermore, if there exist no permutations P, Q such that PAQ is a direct sum of block matrices, then D_1, D_2 are unique up to scalar multiples. Otherwise, the scaled matrices D_1, D_2 are only unique up to a scalar multiple in each block.

[11] One recent observation of this is in Bradley 2010. The observation has however already been made before such as in Achilles 1993

For doubly-stochastic scaling, this result already appears in Brualdi, Parter, and Schneider 1966. For the case of general marginals, it occurs in Menon 1968 and for general matrices and marginals in Hershkowitz, Rothblum, and Schneider 1988; Menon and Schneider 1969. The tools can also be applied to prove that the approximately scaled matrix is unique.

Let us now have a closer look at the difference between approximate scaling and equivalence scaling. What can be said about the convergence of the RAS?

Theorem 4.4 (Pretzel 1980, Theorem 1)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a matrix that is approximately scalable to a matrix with row sums r and column sums c. Let B be a matrix with row sums r and column sums c with maximal subpattern of A (i.e. the number of entries (i,j) such that $B_{ij} = 0$ and $A_{ij} > 0$ is minimal).*

Then A converges to a matrix $C \prec B$ and the same result holds for A' with $A'_{ij} = A_{ij}$ if $B_{ij} > 0$ and $A'_{ij} = 0$ else.

The continuity of the scaling can also be studied:

Theorem 4.5. *Let A be nonnegative and r, c be prescribed row- and column sums. Then the limit of the Sinkhorn iteration procedure is a continuous function of A on the space of matrices with r, c-pattern.*

When the scaling matrices are unique up to a scalar multiple, this also implies that the scaling is continuous in D_1, D_2.

The first proof of this result limited to the doubly-stochastic case was given in Sinkhorn 1972. The full result follows directly from the homeomorphism properties of the map (20) from Tverberg 1976. A discussion is also presented in Krupp 1979. Furthermore, the continuity can be achieved using arguments of Section 9.3.

Finally, let us mention another characterisation of equivalence scaling using *transportation graphs*.

Following Schneider and Zenios 1990, let $A \in \mathbb{R}^{m\times n}$ be a nonnegative matrix. Let $M = \{1,\dots,m\}, N = \{1,\dots,n\}$ and consider the bipartite graph with the bipartition given by the vertices M and N and the edges defined via $E = \{(i,j) : A_{ij} > 0\}$, directed from $i \in M$ to $j \in N$. Now we define a source S_1 that connects to each vertex in M, where the edges have capacity r_i (corresponding to the edge from S_1 to $i \in M$) and we define a sink S_2 that is connected from every vertex in N, where the edges have capacities c_j (see Fig. 2 for an example).

Then it is easy to see that the matrix is approximately scalable if and only if the maximum flow of this network is equal to $\sum_i r_i$. The flows along the edges E then define a matrix with the wanted pattern. The matrix is exactly scalable if and only if the maximum flow of this network is equal to $\sum_i r_i$ and every edge contains flow.

$$A = \begin{pmatrix} 0 & 3 & 4 & 1 \\ 2 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix}$$

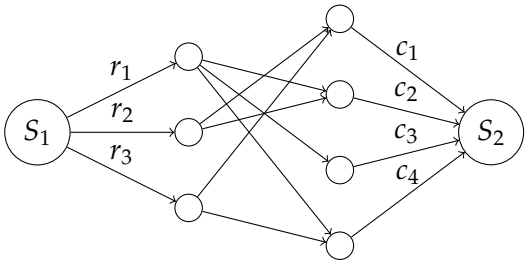

Figure 2: An easy example of the transportation graph for row sums r and column sums c corresponding to the pattern of the matrix A. This example is similar to an example in Schneider and Zenios 1990

5. Other scalings

The problem of equivalence scaling is closely connected to different forms of scalings, the most prominent ones asking for a diagonal matrix D such that DAD is row-stochastic or such that DAD^{-1} has equal row and column-sums.

Many modern approaches to matrix equivalence scaling are general enough to cover most of those different scalings (see Section 6).

5.1. Matrix balancing

Given a matrix A, does there exist a matrix D such that DAD^{-1} has equal column- and row sums? Clearly, this is a special case of D_1AD_2 scaling with a different set of constraints. We have the following characterisation:

Theorem 5.1. *Let $A \in \mathbb{R}^{n\times n}$ be a nonnegative matrix. Then the following are equivalent:*

1. *There exists a diagonal matrix D such that $B = DAD^{-1}$ fulfills $\sum_{i=1}^{n} B_{ij} = \sum_{i=1}^{n} B_{ji}$.*
2. *A is completely reducible or equivalently, a direct sum of irreducible matrices (Hartfiel 1971).*
3. *There exists B with the same pattern as A and $\sum_{i=1}^{n} B_{ij} = \sum_{i=1}^{n} B_{ji}$ (Letac 1974).*

The scaling of A is unique and D is unique up to scalars for each irreducible block of A.

The problem was first considered in Osborne 1960 in the context of preconditioning matrices (see Section 8) by proposing an algorithm and proving its convergence (and uniqueness). Grad 1971, building on Osborne's results, considers the matrix balancing method and provides an algorithm and convergence proof for completely reducible matrices. Unaware of the effort of Osborne and Grad, but considering "the analogue of [Sinkhorn's] result in terms of irreducible matrices" Hartfiel 1971 proves essentially the same result. His approach is based on a progress measure which is basically

the maximum difference of the row- and column sums. Letac 1974 provided an interpretation in terms of patterns. The same was later proved in Schneider and Saunders 1980; Golitschek, Rothblum, and Schneider 1983; Eaves et al. 1985 by yet different means.

Similar to the RAS method, one can propose a simple iterative approximation algorithm:

Algorithm 5.2 (Schneider and Zenios 1990)**.** Let $A \in \mathbb{R}^{n\times n}$ be nonnegative. Let $A^0 := A$. For $k = 0, 1, \ldots$ we define the steps

1. For $i = 1, \ldots, n$, let $u_i = \sum_{j=1}^n A_{ij}^k$ be the row sum and similarly v_i be the column sum. Then define p as the minimum index such that $|u_p - v_p|$ is maximal among $|u_i - v_i|$.
2. Define α_k such that $\alpha_k u_p = 1/\alpha_k v_p$.
3. Let $D = \operatorname{diag}(1, \ldots, 1, \alpha_k, 1, \ldots, 1)$ with α_k at the p-th position. Then define $A^{k+1} = DA^+D^{-1}$ and iterate.

According to Schneider and Zenios 1990, this algorithm is also similar to the proposed scheme in Osborne 1960. At any step, the p-th row is already correctly scaled, while all other rows change their scaling a bit. Note that unlike in the RAS method, the selection of the row and column to be scaled are done using norm differences. Given the results of Brown, Chase, and Pittenger 1993 that the RAS converges regardless of the order of column and row sum normalisations, a similar condition might also accelerate RAS convergence.

We have the following observation:

Proposition 5.3 (e.g. Schneider and Zenios 1990)**.** *The algorithm converges to a balanced matrix B. This matrix is also the unique minimiser of the function*

$$\sum_{i,j=1}^{n} \left(B_{ij} \ln \left(\frac{B_{ij}}{A_{ij}} \right) - B_{ij} \right) \tag{23}$$

subject to the balancing conditions.

Sketch of proof. The fact that the balanced matrix minimises the entropy functional can be seen by direct calculation (the minimiser must be a scaling of the original matrix and the balancing conditions ensure that the scaling is of the form DAD^{-1}).

A proof is similar to observation 3.17: Each step of the algorithm is an I-projection onto the set of matrices with only one row/column balancing constraint. Since the conditions are linear, the repeated projection will converge.

It remains to see that the order of the projections does not matter as long as all directions are chosen arbitrarily often. □

As with equivalence scaling, a graph version of this problem exists, this time using *transshipment graphs*. A nice description can be found in Schneider and Zenios 1990 (see also Figure 3): Given a nonnegative matrix $A \in \mathbb{R}^{n\times n}$, let $V = \{1, \dots, n\}$ and define the set of edges of the transshipment graph (V, E) by $E = \{(i, j)|A_{ij} > 0, i \neq j\}$. We can then add weights A_{ij} to any edge (i, j). A matrix is then balanced, if and only if the incoming flow at each vertex equals the outgoing flow.

$$A = \begin{pmatrix} 0 & 3 & 4 & 1 \\ 2 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 \\ 1 & 1 & 4 & 0 \end{pmatrix}$$

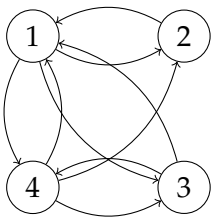

Figure 3: An easy example of the transshipment graph corresponding to the pattern of the matrix A similar to the example in Schneider and Zenios 1990.

5.2. DAD scaling

Another closely related problem is the question, whether given a nonnegative matrix A, there exists a single diagonal matrix D such that DAD has prespecified row- or column sums. A short but quite good overview is given in Johnson and Reams 2009.

Symmetric nonnegative matrices Let us first focus on the case where A is symmetric. It seems natural that this follows directly from Sinkhorn's theorem: If D_1AD_2 has equal row-sums and A is symmetric, so does D_2AD_1. By uniqueness of D_i up to scaling, this implies that one can choose $D_1 = D_2$. This was noted for example in Sinkhorn 1964.

The first discussion of the case of symmetric A can be traced back to the announcements Marcus and Newman 1961; Maxfield and Minc 1962[12]. A first proof for the case of positive matrices and doubly stochastic scaling was given in Sinkhorn 1964. Shortly later, Brualdi, Parter, and Schneider 1966 consider the case of doubly stochastic scaling for nonnegative matrices with positive main diagonal, while Csima and Datta 1972 shows that a doubly stochastic scaling exists if and only if there exists a symmetric doubly stochastic matrix with the same zero pattern if and only if the matrix has total support. This was extended in Brualdi 1974 to cover the case of arbitrary row sums giving the following theorem:

Theorem 5.4 (Brualdi 1974)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric nonnegative matrix. Then the following are equivalent:*

[12]This is covered in many papers, for instance Marshall and Olkin 1968.

1. *There exists a diagonal matrix D with positive entries such that DAD has row sums given by $r \in \mathbb{R}^n_+$.*
2. *There exists a symmetric nonnegative matrix B with the same pattern as A and row sums r.*
3. *For all partitions $\{I, J, K\}$ of $\{1, \dots, n\}$ such that $A(J \cup K, K) = 0$, $\sum_{i \in I} r_i \geq \sum_{i \in K} r_i$ with equality if and only if $A(I, I \cup J) = 0$.*

Furthermore, the scaling is unique.

The equivalence of 2. and 3. is given in Brualdi 1968. 1. follows from 2. using Sinkhorn's theorem and the reverse direction is proved via contradiction. Using the uniqueness in Sinkhorn's theorem then provides uniqueness for the scaling.

Note that the following observation gives a very simple proof of Theorem 3.1:

Observation 5.5. Let $A \in \mathbb{R}^{m \times n}$ be a matrix and $r \in \mathbb{R}^m_+, c \in \mathbb{R}^n_+$ be two prescribed vectors. Then A has an equivalence scaling if and only if the following symmetric matrix A'

$$A' = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \tag{24}$$

has a row-sum symmetric scaling to $(r')^T = (r^T, c^T)$.

Proof. First assume that there exist D_1, D_2 positive diagonal such that $D_1 A D_2$ fulfills

$$D_1 A D_2 e = r, \qquad D_2 A^T D_1 e = c.$$

Then setting $D' := \operatorname{diag}(D_1, D_2)$ we have

$$D' A' D' = \begin{pmatrix} 0 & D_1 A D_2 \\ D_2 A^T D_1 & 0 \end{pmatrix}$$

and clearly $D' A' D' e = (r^T, c^T)^T$.

Conversely, if A' has a row-sum symmetric scaling D', by an analogous argument A will have an equivalence scaling with row sums r and column sums c. □

This was already known in the 70s, maybe even earlier; explicit formulations include Rothblum, Schneider, and Schneider 1994; Knight 2008; Knight and Ruiz 2012. Note that the observation can easily be extended to not just row- and column sums, but all p-norms for $0 < p \leq \infty$ as considered in Section 6.7. It can also be extended to approximate scalings with the same proof. This implies:

Observation 5.6. Results from symmetric scaling for symmetric nonnegative matrices A can always be translated to cover equivalence scaling for arbitrary nonnegative matrices.

The other direction is not true, since clearly not all symmetric matrices are of the special form (24). However, it can still be beneficial to study equivalence scaling on its own, as many algorithms (e.g. the RAS) do not preserve symmetry.

Arbitrary symmetric matrices Theorem 5.4 can be generalised to cover matrices that are not necessarily nonnegative:

Theorem 5.7. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and $\lambda \in \mathbb{R}^n_+$ prescribed column sums. Then:*

1. *If A is positive semidefinite, then A is scalable if and only if A is strictly copositive (Kalantari 1990; Kalantari 1996).*

2. *Any principal submatrix of A (including A) is scalable if and only if A is strictly copositive (Johnson and Reams 2009).*

In general, at least one of the following two propositions is true (Kalantari 1996):

1. *The following set is not empty:*

$$\{x \in \mathbb{R}^n | x^T Ax = 0, x \geq 0, x \neq 0\} \tag{25}$$

2. *For all $\lambda \in \mathbb{R}^n_+$ with $\lambda > 0$ there exists a positive diagonal matrix D such that $DADe = \lambda$. In other words, for any set of prescribed row sums, there exists a scaling.*

More general conditions for scalability of arbitrary symmetric A can be found in Johnson and Reams 2009. We make a number of remarks concerning the results:

1. Another necessary condition for scalability (the matrix must be *diluted*) is provided in Livne and Golub 2004.

2. The question of equivalent conditions for the scalability of matrices remains open. However, these conditions might not have a very useful description, since scalability of arbitrary symmetric matrices is NP-hard (Khachiyan 1996[13]).

3. The second result implies in particular that if a matrix is strictly copositive, it is scalable, which was first proved in Marshall and Olkin 1968. Note that positive definite matrices are in particular strictly copositive, which means that this result encompasses the claimed proofs of scalability of completely positive matrices in Maxfield and Minc 1962. An elementary proof for matrices with strictly positive entries has recently appeared in Johnson and Reams 2009 based on an iterative procedure.

4. For doubly stochastic scaling, the alternative conditions of Kalantari 1996 can also be derived using linear programming duality and/or the hyperplane separation theorem using extremely general methods of duality in self-concordant cones (Kalantari 1998; Kalantari 1999; Kalantari 2005).

[13] This was conjectured also in Johnson and Reams 2009, who noted that deciding whether a matrix is (strictly) copositive is NP-complete according to Murty and Kabadi 1987. The authors seemed to have been unaware of the paper by Khachiyan. The alternative in Theorem 5.7 is also not very useful computationally, because deciding the emptiness of the set (25) is also NP-hard (Kalantari 1990, according to Kalantari 1996).

5. Scaling of the special class of Euclidean predistance matrices has been considered in Johnson, Masson, and Trosset 2005. It turns out that all such matrices are scalable.

6. Note that the equivalence conditions for positive semidefinite matrices can be strengthened. If a matrix is scalable and positive semidefinite,

$$\mu := \min\{x^T Ax | x \geq 0, \|x\|_2 = 1\}$$

can be bounded in terms of the matrix dimension (cf. Khachiyan and Kalantari 1992, where it is also noted that the scaling problem is related to linear programming).

Uniqueness of matrix scaling has also been studied:

Proposition 5.8. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and $\lambda \in \mathbb{R}^n_+$ prescribed row sums. Then*

1. *If A has two or more distinct scalings, then there exists a matrix D such that DAD has eigenvalues $+1$ and -1 (Johnson and Reams 2009).*

2. *For scalable positive definite matrices A there exist 2^n diagonal matrices D such that $DADe = \lambda$, one for each sign pattern of D (O'Leary 2003). In particular, scaling by positive diagonal matrices is unique.*

3. *If A is positive semidefinite, then if A is scalable to row sums r, the positive diagonal matrix is unique (Marshall and Olkin 1968).*

For the scaling of positive semidefinite matrices, upper and lower bounds on $\|D\|$ were derived in Khachiyan and Kalantari 1992; O'Leary 2003.

Johnson and Reams 2009 also note that for nonnegative matrices uniqueness holds in particular if A is primitive (including the case of positive matrices already covered in Sinkhorn 1964) or if A is irreducible and there does not exist a permutation P such that

$$PAP^T = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}.$$

It is also very simple to give an algorithm of RAS type for this problem, using the observation that a DAD scaling to row sums λ exists if and only if $ADe = r/(De)$. This implies that any scaling is a fixed point of the map $\mathbf{T}_{\text{sym}} : \mathbb{R}^n \to \mathbb{R}^n$ with $\mathbf{T}_{\text{sym}}(x) = r/(Ax)$.

Algorithm 5.9 (Knight 2008)**.** Let $A \in \mathbb{R}^{n\times n}$ be nonnegative and symmetric. For the algorithm, set $x_0 = e$ and iterate

$$x_{n+1} = \mathbf{T}_{\text{sym}}(x_n) \tag{26}$$

Nonsymmetric matrices If we do not restrict to symmetric matrices we can only hope to scale A to a matrix with given row-sums. The only notable result seems to be:

Proposition 5.10. *(Sinkhorn 1966) Let $A \in \mathbb{R}^{n\times n}$ be a positive matrix. Then there exists D such that DAD is stochastic.*

The theorem can be extended to cover arbitrary row sums. The first proof occurred in Sinkhorn 1966. Likewise, the proof in Johnson and Reams 2009 does not need symmetry of A.

5.3. Matrix Apportionment

Another scaling problem which is interesting particularly for its applications, is asking for an equivalence scaling, but with the added constraint that the resulting matrix have integer entries. This is important for instance when attributing votes to seats in a parliament and has been applied as early as 1997 (Balinski and González 1997, see also Pukelsheim and Schuhmacher 2004 for one of many explicit accounts for actual changes).

This problem, which is often called *matrix apportionment* has first been studied in Balinski and Demange 1989a; Balinski and Demange 1989b. Algorithms akin to the RAS method exist and others based on network flows can be obtained from Rote and Zachariasen 2007; an overview and many references can be found in Pukelsheim and Simeone 2009.

5.4. More general matrix scalings

This review has so far largely been concerned with nonnegative matrix scaling, with the exception of symmetric DAD scaling. This is understandable, as most of the applications concern nonnegative matrices. However, in view of completeness, let us mention a few of the (mostly quite recent) other cases of matrix scaling.

Arbitrary equivalence scaling While arbitrary D_1AD_2 scaling is interesting for real symmetric matrices, scalings of general real matrices have never sparked a similar amount of interest. It is merely known that the question whether or not a matrix is scalable is NP-hard (Khachiyan 1996) - a question that has also been considered for matrices over the algebraic numbers in Kalantari and Emamy-K 1997. Since the problem of nonnegative matrix scaling turns out to be equivalent to the existence of matrices with given pattern, it seems natural to ask whether the $(+,-,0)$-pattern of matrices with prescribed row- and column sums play a similar role. For positive diagonal scaling the sign pattern of the matrix cannot change and it is a necessary condition for scalability, which is not sufficient as shown in Johnson, Lewis, and Yau 2001. Nevertheless, the authors achieve a characterisation of general matrix patterns (generalising Brualdi 1968, see also Johnson and Stanford 2000; Eischen et al. 2002).

Complex matrices Let us first start with the definition

Definition 5.11. Let $A \in \mathbb{C}^{n\times m}$ be a complex matrix, then A is *doubly quasistochastic* if all sums and columns sum to one.

Note that in case all entries are nonnegative the matrix is doubly stochastic. For the rest of this section, let us restrict to square matrices. Quasistochasticity is interesting, because if A is quasistochastic, then $F_n^* A F_n e_1 = e_1$, where $e_1 = (1, 0, \ldots, 0)^T$ and F_n is the $n \times n$ discrete Fourier transformation. This is true since $F_n e_1 = e$ and e is an eigenvector of A by quasistochasticity. A doubly quasistochastic matrix A therefore satisfies that $F_n^* A F_n$ has e_1 as its first row and column. Repeating diagonal scalings and Fourier transform can then lead to new matrix decompositions.

The natural generalisation of DAD scaling would be D^*AD-scalings for positive semidefinite matrices. These were first studied in Pereira 2003 and later in Pereira and Boneng 2014. Observing that the proof of Marshall and Olkin 1968 extends to complex entries, the authors obtain already part of the following partial results:

Theorem 5.12 (Pereira and Boneng 2014)**.** *Let $A \in \mathbb{C}^{n\times n}$ be positive definite. Then there exist diagonal matrices D_1, D_2 such that $D_1 A D_2$ is doubly quasistochastic.*

Neither D_1, D_2 nor the scaled matrices are necessarily unique. However, there exists at most one scaling with positive matrices D_1, D_2.

The authors suggested that such scalings can be applied to generate highly entangled symmetric states. They furthermore conjectured that the number of such scalings would be upper-bounded, but this was disproved recently in Hutchinson 2016 by giving counterexamples for $n \geq 4$, which have infinitely many scalings. For $n = 3$, there exist at most four scalings. An RAS type algorithm can be obtained from the fact that an equivalent version of Observation 3.14 also holds in the complex case.

Unitary matrices For the subclass of unitaries, we proved the following theorem:

Theorem 5.13 (Idel and Wolf 2015)**.** *For every unitary matrix $U \in U(n)$ there exist diagonal unitary matrices D_1, D_2 such that $D_1 U D_2$ is doubly quasistochastic. Neither D_1, D_2 nor $D_1 U D_2$ are generally unique, in fact in some cases there may even be a continuous group of scalings.*

An algorithm how to obtain D_1, D_2 similar to the RAS method is given and studied in De Vos and De Baerdemacker 2014a, however its convergence is unknown.

The theorem was conjectured in De Vos and De Baerdemacker 2014a and used later (De Vos and De Baerdemacker 2014b; Idel and Wolf 2015) to prove that any unitary matrix can be considered as a product of diagonal unitary matrices and Fourier transforms on principal submatrices. Recently, it has also been applied to prove an analogue of the famous Birkhoff theorem for doubly-stochastic matrices (De Vos and De Baerdemacker 2016).

The proof of Theorem (5.13) boils down to noticing that a scaling exists if and only if there exists a vector x with $Ux = y$ and $|x_i| = |y_i| = 1$ for all $i = 1, \ldots, n$. This is a problem of symplectic topology in disguise and can be solved using a theorem in Biran, Entov, and Polterovich 2004. When we published the theorem in Idel and Wolf 2015 we were unaware of the fact that this proof had in principle already been found, since the equation $Ux = y$ with $|x_i| = |y_i| = 1$, which defines so called *biunimodular vectors* (see for instance Führ and Rzeszotnik 2015), also pops up in several other places. In this context, essentially the same proof was described in Lisi 2011. A first formal publication containing this proof was probably Korzekwa, Jennings, and Rudolph 2014 applying it to error-disturbance relations in quantum mechanics.

6. Generalised approaches

All of the approaches above can be generalised to some extend. Many can then incorporate also different scalings. With an eye towards matrix equivalence, we will attempt to see the different ways of generalisations and what can be gained. A quick summary can be found in Table 6.2.

6.1. Direct multidimensional scaling

Especially in transportation planning, equivalence scaling of arrays with three indices has been important from the beginning. Except for nonlinear Perron-Frobenius theory, the approaches can be readily generalised to this case. As already pointed out, Brown 1959 was the first to consider multidimensional scaling. According to Evans and Kirby 1974 (see also Evans 1970), Furness pointed out iterative scaling as a possible solution to certain transportation planning problems in the unpublished paper Furness 1962. Evans and Kirby themselves proved convergence in a limited scenario by extending the convex programming approach of equation (16) and proofs have been provided or pointed out in several other papers such as Fienberg 1970; Krupp 1979. The case of approximate multidimensional scaling is discussed in Brown, Chase, and Pittenger 1993.

For multidimensional exact or approximate scaling, the convergence results of Pretzel 1980 reflected in Theorem 4.4 still hold. In addition, the order in which we normalise any of the indices of the multidimensional array is irrelevant:

Theorem 6.1 (Brown, Chase, and Pittenger 1993 and comment in Brown 1959)**.** *Let A be an array with m indices (or dimensions) and let i_k be the dimension of the array that is scaled in the k-th step. If each element of $\{1, \ldots, m\}$ appears in the sequence $\{i_1, i_2, \ldots\}$ infinitely often, then the scaling converges to the limit of the cyclic RAS method, the I-projection of A.*

Type	**Base case**	D_1AD_2	DAD	DAD^{-1}	**multi-dim.**	**continuous**	**additional generalisation**
Algorithmic approaches	RAS-type algorithms	with row or column norm constraints (also max-sum)	✓	✓	-	-	-
Axiomatic approach	D_1AD_2 scaling with inequality constraints	✓	-	-	-	-	-
Convex optimisation	D_1AD_2, DAD scaling	✓	✓	-	-	-	also copositive matrices, complex scaling
Entropies	minimising relative entropy minimisation + (non)linear constraints	✓	(✓)	(✓)	✓	✓	special case of Bregman divergences; cross entropies; justifications
Letac's approach	no scaling: existence of some function	✓	✓	✓	-	-	completely different applications
Log linear models	scaling of prob. distributions $w_i = x_i \prod_j d_j^{C_{ij}}$ given C, x with constraints $Cw = b$	✓	✓	✓	✓	-	Different scalings defined via C
N-L Perron-Frobenius Theory	fixed points of homogeneous maps on cones	✓	✓	-	-	✓	Maps in different vector spaces (such as positive maps); infinite matrices
Truncated matrix scaling	$X = \Lambda DAD^{-1}$ balancing with $L \leq X \leq U$ entrywise	also inequalities	✓	✓	-	-	-

Table 1: This table gives an overview about possible approaches to matrix scaling, their application to various different scalings discussed in Section 6 and additional possible applications.

6.2. Log-linear models and matrices as vectors

Most of the ideas above use matrices as matrices, as sets of numbers with two indices. One can likewise consider just vectors of numbers and define columns and rows by defining partitions of the vectors. This approach has the advantage that the generalisation to multidimensional matrices is immediate. It was probably pioneered by Darroch and Ratcliff 1972, although Lamond and Stewart 1981 credit Murchland, who circulated his results later (Murchland 1977; Murchland 1978[14]). The approach was then taken on in Bapat and Raghavan 1989 (see also Bapat and Raghavan 1997, Chapter 6 for an overview and a more lucid presentation of their ideas). While Darroch and Ratcliff 1972 used an entropic approach, Bapat and Raghavan 1989 is based on a combination of optimisation and topological approaches as discussed in Section 3.6. The same theorem is also proved in Franklin and Lorenz 1989 in a very elementary fashion and in Rothblum 1989 using optimisation techniques.

The original goal of Darroch and Ratcliff 1972 was not to study matrix scaling but rather obtaining probability distributions using so called log-linear models. Given a positive (sub)probability distribution π over some finite index set I, a log-linear model is a probability distribution p such that

$$p_i = \pi_i D \prod_{s=1}^{d} D_s^{C_{si}} \tag{27}$$

which satisfies some constraints $\sum_{i \in I} C_{si} p_i = k_s$. Here, D and D_s have to be determined while C is given from the problem. The name derives from the fact that the solution is an exponential family of probability distributions.

Depending on the choice of C, one can write matrix balancing, equivalence scaling or DAD scaling as finding a log-linear model.

To achieve equivalence scaling with row-sums r and column sums s, consider for simplicity the case of a 2×3 matrix. Then C and b are given by

$$C = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} r_1 \\ r_2 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix}$$

and we define $y_1 = A_{11}, y_2 = A_{12}, \ldots, y_5 = A_{22}, y_6 = A_{23}$ (example from Bapat and Raghavan 1989; Rothblum 1989).

To achieve matrix balancing with row-sums equaling column sums, consider for simplicity the case 3×3, then C is given by

$$C = \begin{pmatrix} 0 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 1 & 1 & 0 \end{pmatrix} \tag{28}$$

[14] The papers were not available to me

and $b = 0$ and we order x, y again as before (example from Rothblum 1989).

We have the following theorem:

Theorem 6.2 (Bapat and Raghavan 1989)**.** *Let $C \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m_{+0}$. Let $K = \{v|Cv = b, v \geq 0\}$ be bounded. Let $x \in \mathbb{R}^n_{+0}$. Then there exists a $w \in K$ such that for some $D \in \mathbb{R}^n_+$ we have*

$$w_j = x_j \prod_{i=1}^{m} D^{C_{ij}}, \quad j = 1, \ldots, n$$

if and only if there exists a vector $y \in \mathbb{R}^n_{+0}$ with $y \in K$ and the same zero pattern as x.

Note that this is a major generalisation of scaling as the matrix C can contain any real numbers.

The limiting factor of the theorem is the boundedness of K. While the constraints in the case of matrix equivalence are bounded, the constraint set defined by (28) is not necessarily bounded. Rothblum 1989 applies a completely different proof which only works for positive matrices. However we can still apply Theorem 6.2: K is unbounded, because the matrix entries can become unbounded since we only want equal row and column sums but do not specify them further. We fix that by using

$$\tilde{C} = \begin{pmatrix} 0 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and $b_4 = 1$. The last row just implies that the sum of all matrix entries should be one which makes K a bounded set. A simple calculation then shows that this is equivalent to searching for a diagonal matrix D and a scalar d such that $dDAD^{-1}$ has equal row- and column sums and the sum of all matrix entries is one. Clearly, this is equivalent to matrix balancing and we can apply Theorem 6.2.

The connection to entropy minimisation is simple:

Lemma 6.3 (Darroch and Ratcliff 1972, Lemma 2)**.** *Given a positive (sub)probability distribution π, if a positive probability distribution p satisfying (27) and the linear constraints exists, then it minimises relative entropy $\sum_i p_i \log(p_i/\pi_i)$ subject to the linear constraints.*

Proof. The proof in Darroch and Ratcliff 1972 is a straightforward calculation and follows directly from Kullback and Khairat 1966. If q is a probability distribution satisfying the linear constraints, then

$$\begin{aligned} D(p||\pi) &= \sum_{i\in I} p_i (\log \xi + \sum_{s=1}^{d} C_{si} \log \xi_s) \\ &= \log \xi \left(\sum_{i\in I} p_i \right) + \sum_{s=1}^{d} \log \xi_s \left(\sum_{i\in I} C_{si} p_i \right) \end{aligned}$$

$$
\begin{aligned}
&= \log \xi \left(\sum_{i \in I} q_i \right) + \sum_{s=1}^{d} \log \xi_s \left(\sum_{i \in I} C_{si} q_i \right) \\
&= \sum_{i \in I} q_i \log(p_i / \pi_i) \\
&= D(q||\pi) - D(q||p)
\end{aligned}
$$

which implies the lemma by the nonnegativity of relative entropy. □

6.3. Continuous Approaches

Nonnegative matrices were always tied to joint probability distributions. Obviously, there is no reason to only study discrete probability distributions. The first such generalisation was obtained in Hobby and Pyke 1965. Also the basic theorems of Kullback 1968 and Csiszár 1975 are more general than counting measures (although both have problems with parts of their arguments, see Borwein, Lewis, and Nussbaum 1994).

As pointed out in Borwein, Lewis, and Nussbaum 1994, there are essentially two approaches to continuous versions, the entropy maximisation approach studied by Kullback and later Csiszár, and the approach via fixed point theorems or contractive ratios (one can see this as a precursor to nonlinear Perron-Frobenius theory) studied in, for instance, Fortet 1940; Nussbaum 1987; Nussbaum 1993. The natural continuous extension of the DAD theorem for symmetric matrices was studied in Nowosad 1966; Karlin and Nirenberg 1967 (via fixed points or iterative contractions). The most general results in Borwein, Lewis, and Nussbaum 1994 combine these two approaches. To give a flavour of their results, we cite

Theorem 6.4 (Borwein, Lewis, and Nussbaum 1994 Theorem 3.1)**.** *Given a finite measure spaces $\mu(s,t) = k(s,t)\,ds\,dt$ and marginal distributions $\alpha(s), \beta(t) \in L^1(dt/ds)$, consider the following minimisation problem:*

$$
\begin{aligned}
\min & \int_{S \times T} [u(x,y) \log(u(x,y)) - u(x,y)] k(s,t)\,ds\,dt \\
\text{s.t.} & \int_T u(s,t) k(s,t)\,dt = \alpha(s) \quad a.e. \\
& \int_S u(s,t) k(s,t)\,ds = \beta(t) \quad a.e.
\end{aligned} \tag{29}
$$

where $u \in L^1(dt, ds)$. Furthermore, we require $\int_S \alpha(s)\,ds = \int_T \beta(t)\,dt$. Then the minimisation problem has a unique optimal solution. If there exists a u_0 which fulfils the constraints and there exist $x_0 \in L^\infty$ and $y_0 \in L^\infty$ such that $\log u_0(s,t) = x(s) + y(t)$ almost everywhere, then u_0 is the unique solution. Conversely, if there exists a feasible solution u with $u > 0$ almost everywhere, then the unique optimal solution satisfies $u_0 > 0$ almost everywhere and there exist sequences $x_n \in L^\infty$ and $y_n \in L^\infty$ such that

$$
\lim_{n \to \infty} (x_n(s) + y_n(t)) = \log u_0(s,t) \quad a.e.
$$

This in fact also covers the approximate scaling case. Note that the results also extend to more than two marginals.

6.4. Infinite matrices

Instead of continuous functions, we can also consider infinite matrices. Results usually posit that row and column sums should be finite in some norm.

The first such result was obtained in Netanyahu and Reichaw 1969, which proves Theorem 3.1 in the case where the column and row sums are uniformly bounded in the l^1-norm and the matrix entries are uniformly bounded. The proof is reminiscent of Brualdi, Parter, and Schneider 1966 and Nonlinear Perron-Frobenius theory, using a fixed point argument involving Schauder's fixed point theorem.

Another approach was presented in Berger and Kelley 1979, where matrices that are infinite in one direction are studied (rows or columns are finite in a l^p-norm). Once again, the matrix entries must be bounded uniformly (in this case, in a l^p-norm) and convergence of the iterative algorithm to a unique solution is proved in certain topologies.

6.5. Generalised entropy approaches

As we saw in Section 3.4, we can write matrix scaling as the problem

$$\min_P D(P\|Q) \quad \text{s.t. } P \in \Pi$$

where Π is an intersection of linear constraints. This approach can be generalised in two ways: First, one could consider other functions than relative entropy but related to it or second, one can consider more general sets Π.

Telative entropy is a special case of *Bregman divergences*. These were originally introduced in Bregman 1967 and later named in Censor and Lent 1981. The idea is to study distance measures derived from functions $\phi : S \in \mathbb{R}^n \to \mathbb{R}$, which are defined on a closed convex set S, continuously differentiable and strictly convex. Then

$$\Delta_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle$$

behaves similarly to a metric, although it is not necessarily symmetric and obeys no triangle inequality. If one takes $\phi(x) = \sum_i (x_i \log(x_i) - x_i)$ (negative entropy modulo the linear term), then $\Delta_\phi(x, y) = \sum_i (x_i \ln(x_i/y_i) - x_i + y_i)$. This example was already studied in Bregman 1967 giving in addition an iterative algorithm to find the projections onto the minimum Bregman distance given linear constraints, which is a variant of the RAS method (see also Lamond and Stewart 1981).

Another way to generalise D is the basic observation underlying McDougall 1999: Relative entropy for matrices is equivalent to the sum of cross-entropies between matrix columns, where a *cross-entropy* of the column j of the matrices A, B is just

$D_j(A\|B) := \sum_i B_{ij} \log(B_{ij}/A_{ij})$. Instead of taking the sum of all cross-entropies, it might be justified to take weighted sums of cross-entropies. This is relevant in economic settings and, aside from McDougall 1999, was studied in e.g. Golan and Judge 1996; Golan, Judge, and Miller 1997[15].

On the other hand, we can work with relaxed constraints. This was covered in Brown, Chase, and Pittenger 1993: The extension of linear families of probability distributions is still covered by Csiszár 1975, while finding the I-projection for closed, convex but nonlinear constraints requires different means such as *Dykstra's iterative fitting procedure* (cf. Dykstra 1985).

6.6. Row and column sum inequalities scaling

Instead of wishing for matrices to have prespecified row and column sums, it might be interesting to consider cases where only lower and upper bounds on the row and column sums and the matrix entries are considered.

If we denote the set of all nonnegative matrices with row sums between $r^- \in \mathbb{R}^n_+$ and $r^+ \in \mathbb{R}^n_+$ and column sums between $c^- \in \mathbb{R}^n_+$ and $c^+ \in \mathbb{R}^n_+$ and total sum of its entries h by $R(r^-, r^+, c^-, c^+, h)$, then we can ask the question, whether for a given nonnegative $A \in \mathbb{R}^{n\times n}$, there exists $\delta > 0$ and D_1, D_2 diagonal matrices such that

$$B := \delta D_1 A D_2 \in R(r^-, r^+, c^-, c^+, h)$$

such that if $(D_1)_{ii} > 1$, then $\sum_j B_{ij} = r^-$ and $(D_1)_{ii} < 1$, then $\sum_j B_{ij} = r^+$ and the same conditions for D_2 and c. This problem was studied in Balinski and Demange 1989a; Balinski and Demange 1989b, where they call such a matrix a *fair share matrix*. The main purpose of the approach is described in Section 8.2. Using the arguments from nonlinear Perron-Frobenius theory (Section 3.3), they prove

Theorem 6.5 (Balinski and Demange 1989a)**.** *Let A be a nonnegative matrix. There exists a unique fair share matrix for A if and only if there exists a matrix $B \in R(r^-, r^+, c^-, c^+, h)$ with the same pattern as A.*

A different generalisation is called *truncated matrix scaling*. It is studied in Schneider 1989; Schneider 1990 and can also account for equivalence scaling. While the proofsuse a combination of optimisation techniques for an optimisation problem defined via entropy functionals, the motivation and interpretation uses graphs and transportation problems for graphs.

The problem considers matrix balancing and not equivalence scaling as its basic problem and then explains the connection. A problem consists of an ordered triple (A, L, U) of nonnegative matrices in $\mathbb{R}^{n\times n}$ satisfying

1. $0 \leq L \leq U \leq \infty$,

[15]References corrected but taken from McDougall 1999 as they were unavailable to me

2. There is a matrix X whose pattern is a subpattern of A such that X is balanced and $L \leq X \leq U$,

3. $U_{ij} > 0$ whenever $A_{ij} > 0$.

and asks for a diagonal matrix D and a nonnegative matrix Λ such that

1. $X = \Lambda DAD^{-1}$ is balanced and $L \leq X \leq U$,

2. X and Λ satisfy

$$\begin{cases} \Lambda_{ij} > 1 \;\Rightarrow\; X_{ij} = L_{ij} \\ \Lambda_{ij} < 1 \;\Rightarrow\; X_{ij} = U_{ij} \end{cases}$$

The conditions above are consistency conditions which are trivially necessary for the existence of D, Λ. One can easily see that for $L = 0$ and $U = \infty$ the problem is equivalent to matrix scaling, because $\Lambda = \mathbb{1}$. Note the similarity of the treatment of inequalities to the ideas of Balinski and Demange.

The connection to equivalence scaling is simple (cf. Schneider 1989): Given a nonnegative matrix A and row and column sums r, c, we start with the graph of Figure 2: we join the two vertices S_1 and S_2 into one vertex (call it S), keeping everything else fixed. We label the edges between the nodes with the corresponding matrix entries A_{ij}. A' is now the matrix corresponding to the graph.

Now we copy the graph twice and erase the weights of the edges and instead label the first graph by $l_{(i,j)}$ and the second by $u_{(i,j)}$ where

$$l_{(i,j)} := \begin{cases} 0 & \text{if } A_{ij} > 0 \\ r_i & \text{if } j = 0 \\ c_j & \text{if } i = 0 \end{cases} \qquad u_{(i,j)} := \begin{cases} \infty & \text{if } A_{ij} > 0 \\ r_i & \text{if } j = 0 \\ c_j & \text{if } i = 0 \end{cases}.$$

Now L' (U') is the matrix corresponding to the graph with labels $l_{(i,j)}$ ($u_{(i,j)}$). Finally, (A', L', U') is the triple for truncated matrix scaling.

The main result of the paper then includes:

Theorem 6.6 (Schneider 1989 Theorem 14 (part of it))**.** *Let (A, L, U) be a triple in $\mathbb{R}^{n\times n}$ fulfilling the consistency conditions 1.-3. above. Then the truncated matrix scaling has a solution satisfying the conditions 1. and 2. above if and only if there exists a balanced matrix X such that*

- $L \leq X \leq U$,
- $X_{ij} > 0$ *iff* $A_{ij} > 0$ *always.*

Once again, the answer is dominated by the pattern of the matrix and the conditions boil down to the usual conditions for similarity scaling. In a sense, this gives another explanation as to why both problems, equivalence scaling and matrix similarity, need pattern conditions for feasibility: They are both similar graph-related problems.

6.7. Row and column norm scaling

Instead of asking the question whether one can scale a matrix to prescribed row- and column sums, one can ask for a scaling to prescribed row- and column norms.

For the ∞-norm, this is discussed in Rothblum, Schneider, and Schneider 1994. Their proof relies on an algorithm for symmetric DAD scaling using Observation 5.5. In fact, an algorithm for the problem had already been studied for the symmetric case in Bunch 1971[16].

Theorem 6.7 (Rothblum, Schneider, and Schneider 1994)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a nonnegative matrix and $r \in \mathbb{R}^m_+$, $c \in \mathbb{R}^n_+$ be prescribed row and column maxima. Then the following are equivalent:*

1. *There exist diagonal matrices D_1 and D_2 such that D_1AD_2 has prescribed row and column maxima r and c.*
2. *There exists a matrix B with row and column maxima r and c with the same pattern as A.*
3. *There exists a matrix B with row and column maxima r and c with some subpattern of A.*
4. *The vectors r and c fulfil*

$$\max_{i=1,\dots m} r_i = \max_{j=1,\dots n} c_j \tag{30}$$

$$\max_{i\in I} r_i \leq \max_{j\in J^c} c_j \tag{31}$$

$$\max_{j\in J} c_j \leq \max_{i\in J^c} r_i \tag{32}$$

for every subsets $I \subset \{1,\dots,m\}$ and $J \subset \{1,\dots,n\}$ such that $A_{IJ} = 0$.

There are two further technical conditions given in Rothblum, Schneider, and Schneider 1994 as well as an algorithm that converges to the solution.

The usual equivalence scaling now corresponds to 1-norm scaling. For p-norms of row and columns with $0 < p < \infty$, it is shown that this problem reduces to 1-norm scaling in Rothblum, Schneider, and Schneider 1994. If $A^{(p)}$ denotes the entrywise power, we have:

Theorem 6.8 (Rothblum, Schneider, and Schneider 1994)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a nonnegative matrix and $r \in \mathbb{R}^m$, $c \in \mathbb{R}^n$. Then the following are equivalent:*

1. *There exist matrices D_1 and D_2 such that $D_1AD_2 = B$ has prescribed row and column p-norms r and c.*
2. *There exist matrices D_1 and D_2 such that $D_1^{(p)}A^{(p)}D_2^{(p)}$ has prescribed row and column sums $r^{(p)}$ and $c^{(p)}$.*

[16] Reference from Knight 2008 among others. I could not obtain the reference.

3. *There exists a matrix B with the same pattern as A and row and column sums given by $r^{(p)}$ and $c^{(p)}$.*

Hence the answer again reduces to a question of patterns. Likewise, the ε-scalability can immediately be transferred. Much weaker results were obtained in Livne and Golub 2004, where the problem of 2-norm scaling was studied for arbitrary (not necessarily nonnegative) matrices. A (fast) algorithm is also derived in Knight and Ruiz 2012.

6.8. Row and column product scaling

At this point, one might wonder what happens when replacing the row- and column sums by row- and column products. This has been treated in Rothblum and Zenios 1992, however it is not connected to entropy or maximum likelihood estimation, but instead to least square estimations, which is why we will not discuss the techniques here. However, this is interesting in light of the original justification of the RAS method in transportation planning by Deming and Stephan 1940. The results are simple:

Theorem 6.9 (Rothblum and Zenios 1992)**.** *Let A be a nonnegative matrix. Then the following are equivalent:*

1. *There exist positive diagonal matrices D_1 and D_2 such that D_1AD_2 has row and column products r_p and c_p.*
2. *There exists a matrix B with the same zero pattern as A and row and column products r_p and c_p.*

The scalded matrix D_1AD_2 is unique.

Furthermore, if A has no zero rows or columns, there always exists a matrix D such that DAD^{-1} has equal row and column products.

Note that in the case of matrix balancing to equal row and column products, the result is also the same: This is possible if and only if a balanced matrix with the same pattern exists which is always the case (cf. Rothblum and Zenios 1992, Theorem 5.2).

7. Algorithms and Convergence complexity

After the basic existence problems of matrix scaling were solved in the 60s to 80s, the focus shifted to algorithms and complexity theory in the 90s. The story is equally convoluted, not least because algorithmic complexity is difficult and not always well-defined in itself: One can decide to study worst case or average convergence speed, count algorithm steps or computational operations. Given the RAS method and the fact that it is a coordinate descent method for an intrinsically convex optimisation problem, which is amenable to a host of other techniques, the choice of a relevant class of algorithms is already not unique.

Since this review is geared more towards the mathematical aspects of the problem, our focus will lie on exact complexity results instead of proofs by example. Papers focussed on numerical aspects appeared as early as the late 70s, early 80s with Robillard and Stewart 1974 average convergence considerations, Bachem and Korte 1979 and Parlett and Landis 1982. A small overview about many of the recent developments can be found in Knight and Ruiz 2012.

7.1. Scalability tests

Most algorithms explicitly require that the matrix A be scalable (or positive). This means that we first need to check for scalability.

Proposition 7.1. *Let $r \in \mathbb{R}_+^m, c \in \mathbb{R}_+^n$ be two positive vectors with $\sum_i r_i = \sum_j c_j$. Let A be a nonnegative matrix, then one can check whether A is approximately scalable in polynomial time $\mathcal{O}(pq \log(q^2/p))$ with $q = \min\{m, n\}$ and p the number of nonzero elements in A.*

If $r \in \mathbb{Q}_+^m, c \in \mathbb{Q}_+^n$, then one can check for exact scalability in polynomial time of the same order.

The fact that approximate scalability can be efficiently checked was probably first seen in Linial, Samorodnitsky, and Wigderson 2000. A complete and well-readable proof giving explicit bounds appeared in Balakrishnan, Hwang, and Tomlin 2004, exact scalability can be found in Kalantari et al. 2008.

Sketch of Proof. We first follow the proof in Balakrishnan, Hwang, and Tomlin 2004, which uses the transportation graph described in Figure 2.

The matrix is approximately scalable iff the maximum flow of this network is equal to $\sum_i r_i$. The flows along the edges E then define a matrix with the wanted pattern. Such a network flow problem can be solved in time $\mathcal{O}(pq \log(q^2/p))$ with $q = \min\{m, n\}$ and p the number of nonzero elements in A (Ahuja et al. 1994).

In order to check for exact scalability, one has to check whether there exists a solution where each edge has a positive amount of flow (otherwise the entry would have to be reduced to zero). We can check for a solution to the maximum flow problem with minimum flow through each edge bigger than a prespecified value ε with the same costs as solving a maximum flow problem twice. Clearly, this does not help as, we would have to check scalability for any $\varepsilon > 0$.

However (following Kalantari et al. 2008) if r, c have only rational entries, we can find a number h such that hr, hc have only integer values. In this case, the flow problem has a solution iff there exists a matrix B with column sum hc and row sum hr where each positive entry fulfils $B \geq 1/|E|$, where $|E|$ denotes the number of edges in E.

This implies that it suffices to check for a solution with capacities hr, hc and minimum flow through each edge having prespecified value $1/(2|E|)$. □

For positive semidefinite matrices, scalability can also be checked easily:

Proposition 7.2 (Khachiyan 1996)**.** *Let $A \in \mathbb{R}^{n\times n}$ be positive semidefinite. Then A is scalable if and only if $Ax = 0$ and $e^T x = 1$ has no solution $x \geq 0$. This can be tested by a linear program.*

Proof. The formulation is already nearly in canonical form. We maximize $e^T x$ subject to the equality constraints $Ax = 0$ and $x \geq 0$. □

For arbitrary matrices scalability is mostly NP-hard (see Section 5.4).

7.2. The RAS algorithm

The RAS algorithm, being the natural algorithm to compute approximate scaling, is also the most studied algorithm. For the case of doubly stochastic matrices, it has long been known (cf. Sinkhorn 1967) that for positive matrices, the RAS converges linearly (sometimes called geometrically) in the l_∞ norm. Krupp 1979 gave a simple argument that the iteration will get better at any step. This can also be inferred from the fact that the RAS method is iterated I-projection onto a convex set using Csiszár 1975. Later, Franklin and Lorenz 1989 showed that the convergence is also linear in Hilbert's projective metric, while Soules 1991 showed linear convergence for all exactly scalable matrices basically in arbitrary vector norms, albeit without explicit bounds. Conversely, it was shown that only scalable matrices can have linear convergence meaning that the RAS converges sublinear for matrices with support that is not total (Achilles 1993). We have the following best bounds:

Theorem 7.3 (Knight 2008, Theorem 4.5)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a fully indecomposable matrix and denote by D_1, D_2 the diagonal matrices such that $D_1 A D_2$ is doubly stochastic. Let D_1^i and D_2^i be the diagonal matrices after the k-th step of the Sinkhorn iteration, there exists a $K \in \mathbb{N}$ such that for all $k \geq K$, in an appropriate matrix norm*

$$\|D_1^{i+1} \oplus D_2^{i+1} - D_1 \oplus D_2\| \leq \sigma_2^2 \|D_1^i \oplus D_2^i - D_1 \oplus D_2\| \tag{33}$$

where σ_2 is the second largest singular value of $D_1 A D_2$.

The proof of this theorem crucially relies on the fact that matrices with a doubly stochastic pattern are direct sums of primitive matrices (modulo row and column permutations). Hence it cannot easily be extended to matrices with arbitrary row and column sums if those matrix patterns allow for non-primitive matrices. The approach in Franklin and Lorenz 1989 for positive matrices can also be extended to arbitrary row and column sums.

We observe that the occurrence of the second singular value should not come as a big surprise: Given a stochastic matrix A, A^k converges to a fixed matrix and the convergence is dominated also by the gap between the largest singular value 1 and the second largest singular value of A.

For practical purposes, one then needs to work out how many operations are needed to obtain a given accuracy of the solution. The first such bounds can be derived from

the bounds in Franklin and Lorenz 1989. The main study of these questions was conducted in the early 90s and 2000s, starting with Kalantari and Khachiyan 1993. Let $A \in \mathbb{R}^{n\times n\times\ldots\times n}$ with d copies of $\mathbb{R}^n$ be a positive multidimensional matrix which can be scaled to doubly stochastic form, then they proved that the RAS takes at most

$$\mathcal{O}\left(\left(\frac{1}{\varepsilon}+\frac{\ln(n)}{\sqrt{d}}\right)d^{3/2}\sqrt{n}\ln\frac{V}{\nu}\right) \tag{34}$$

steps, where all matrix entries are in the interval $(\nu, V]$ and the maximal error is upper-bounded by ε. (Kalantari and Khachiyan 1993, Theorem 1). They also derive a bound for a randomised version of the RAS, where at each step, the direction of descent is selected randomly and once in a while, the whole error function is computed randomly. The expected runtime is then slightly lower.

In the case of positive matrix scaling, Kalantari et al. 2008 give better bounds covering also the case of inequality constraints as in Balinski and Demange 1989a. In particular, let $A \in \mathbb{R}^{n\times m}$ be a positive matrix with $\nu \leq A_{ij} \leq V$, let $N = \max\{n, m\}$, let $\rho = \max\{r_i, c_j\}$ and $h = \sum_{ij} A_{ij}$. Then the number of iterations needed to scale A to accuracy ε is of order

$$\mathcal{O}\left(\left(\frac{1}{\varepsilon}+\ln(hN)\right)\rho\sqrt{N}\left(\ln(\rho)+\ln\left(\frac{V}{\nu}\right)\right)\right).$$

The two results (specific bounds and asymptotic linear behaviour for convergence speed) imply that the RAS method has generally good convergence properties if the matrix is positive.

A fully polynomial time algorithm (i.e. without a factor involving the size of the matrix entries) for general marginals was given in Linial, Samorodnitsky, and Wigderson 2000 based on the RAS method with preprocessing. However, the algorithm scales with $\mathcal{O}(n^7\log(1/\varepsilon))$ for the general (r, c)-scaling and (using a different algorithm closer to the RAS) with $\mathcal{O}((n/\varepsilon)^2)$ for doubly-stochastic scaling.

In summary, the RAS method, while not fully polynomial by itself, can be tweaked in various ways to allow for fully polynomial algorithms. In addition, it has the advantage of being parallelisable as demonstrated in Zenios and Iu 1990; Zenios 1990. However, the scaling behaviour is not particularly fast in specific examples (see for instance Balakrishnan, Hwang, and Tomlin 2004; Knight and Ruiz 2012), in particular it doesn't seem to be very good at handling sparse matrices.

7.3. Newton methods

One of the first alternative algorithms to the RAS methods was provided in Marshall and Olkin 1968 as a minimisation of $x^T Ay$ using a modified Newton method as described in Goldstein and Price 1967 (it is not related how the equality constraints are introduced into the problem. This can be done using a C^2-penalty function).

Newton methods were also developed to solve the scaling problem for positive semidefinite matrices. They can either be seen as Newton's method applied to $x^T Ax$ for symmetric A (cf. Khachiyan and Kalantari 1992) or as Newton's method applied to the Sinkhorn iteration equation $x_{k+1} := e/(Ax_k)$ (cf. Knight and Ruiz 2012). Yet a different method was considered in Fürer 2004.

Kalantari 2005 shows that their algorithm converges in $\mathcal{O}(\sqrt{n}\ln(n/(\mu\varepsilon)))$ Newton iteration steps, where $\mu := \inf\{x^T Ax | x \geq 0\}$, if the matrix is scalable.

7.4. Convex programming

As noted in section 3.5, the convex programming formulation of the problem makes it amenable to a host of (polynomial time) techniques such as the ellipsoid method or interior point algorithms.

In the case of nonnegative matrices $A \in \mathbb{R}^{n\times n}$ with doubly stochastic marginals, a good bound was found in Kalantari and Khachiyan 1996, with operations of order

$$\mathcal{O}(n^4 \ln(n/\varepsilon)\ln(1/\nu)).$$

The bound uses ellipsoid methods. Later, the bounds were extended to cover generalised marginals in Nemirovski and Rothblum 1999 (also including the generalisation discussed in Rothblum 1989) specifically using ellipsoid methods for the convex optimisation formulation of equation (16). The first instance of an interior point algorithm was probably formulated in Balakrishnan, Hwang, and Tomlin 2004 applied to the entropy formulation. The authors find a strongly polynomial algorithm which scales better than Linial, Samorodnitsky, and Wigderson 2000 with $\mathcal{O}(n^6 \log(n/\varepsilon))$[17].

A different ansatz for an algorithm was used in Schneider 1990, where the author uses the duality in convex programming and a coordinate ascent algorithm for the dual problem of his truncated matrix scaling. This algorithm will then be some form of generalisation of the RAS method.

7.5. Other ideas

We give a short primer of other algorithms considered in the literature:

1. The first paper to develop new algorithms with a focus on speed and not only concepts was Parlett and Landis 1982, where a bunch of slightly different and optimised algorithms is derived.

2. An algorithm which is somewhat related to convex algorithms is considered in Kalantari 1996. It is a total gradient based, steepest descent algorithm for the homogeneous log-barrier potential.

[17] It seems that the authors were unaware of Kalantari et al. and Nemirovski and Rothblum 1999.

3. In Rote and Zachariasen 2007, using an algorithm for the matrix apportionment problem involving network flows and using ideas of Karzanov and McCormick 1997, they provide an algorithm where the number of iterations scales with

$$\mathcal{O}(n^3 \log n(\log(1/\varepsilon) + \log(n) + \log\log(V/\nu))). \tag{35}$$

Once again, $A_{ij} \in [\nu, V]$ for all i, j.

4. With ever larger matrices, it is sometimes infeasible to access each element of the matrix on its own, because the matrix is not stored in that form or processed somewhere else. This makes it interesting to consider algorithms that do not need access to all elements, such as the RAS method for nonnegative matrices. Algorithms that are "matrix free" in this sense were developed in Bradley 2010; Bradley and Murray 2011 for doubly stochastic scaling of positive semidefinte matrices.

5. Finally, let us mention that algorithms were also developed for infinity norm scaling (cf. Bunch 1971; Ruiz 2001; Knight, Ruiz, and Uçar 2014) and other norm scaling (cf. Ruiz 2001).

7.6. Comparison of the algorithms

A first comparison of several algorithms was performed in Schneider and Zenios 1990, however, the comparison is not really in terms of speed (for instance, all algorithms were implemented on different programming platforms), but in terms of useability.

While there have been many papers claiming superior convergence speed for their algorithm, the most comprehensive analysis has probably been achieved in Knight and Ruiz 2012, which is limited to doubly-stochastic scalings. In the paper, the authors compare the RAS method, a Gauss-Seidel implementation of the ideas of Livne and Golub 2004, and the fastest algorithm in Parlett and Landis 1982 with their own Newton-method algorithm. The test matrices are mostly large sparse matrices and the new algorithm is usually the fastest and most robust algorithm. The authors also claim that their Newton-based implementation is superior to Khachiyan and Kalantari 1992 and Fürer 2004. They also suggest that the algorithm should outperform the convex optimisation based algorithms, albeit a direct comparison to the most recent algorithm in Balakrishnan, Hwang, and Tomlin 2004 is missing, who only showed that their algorithm clearly outperforms the RAS method. Bradley and Murray 2011 also mention that their purely matrix free algorithm will outperform explicit methods such as those in Knight and Ruiz 2012 if accessing single elements in the matrix is actually slow.

In general, matrix scaling can today be done on a routine basis even for very large matrices.

8. Applications of Sinkhorn's theorem

The following problem can be encountered in many areas of applied mathematics (see also Schneider and Zenios 1990):

Problem 1. Let $A \in \mathbb{R}^{m \times n}$ be a nonnegative matrix. Find a matrix B which is close to A and which fulfills a set of linear inequalities, for instance

$$\sum_{j=1}^{n} A_{ij} = r_i, \qquad \sum_{i=1}^{m} A_{ij} = c_j.$$

We could also ask for balanced marginals or any other type of marginals. The problem is certainly not well-posed. What does "close" mean? This part of the review will try to give an overview why "close" means equivalence scaling in many applications. We will limit our attention mostly to the mathematical justification of matrix scalings, but I will try to give pointers to other literature.

This implies that we only consider "nearness" leading to equivalence scaling or matrix balancing as the result. In the literature, other nearest matrices have also been considered such as addition of small matrices (e.g. Bachem and Korte 1980). Schneider and Zenios 1990 describe network flow algorithms that allow for a wider variety of applications.

Matrix scaling has many different real world applications, which implies that it also needs different justifications. While statistical justifications exist, many applications argue with the simplicity of the method and the fact that it performs well in practice. These are valid arguments, but they are unsatisfactory from a mathematical point of view. In the two following subsections we collect mathematically rigorous (or partly rigorous) justifications and their history.

8.1. Statistical justifications

As seen, matrix scaling solves entropy minimisation with marginal constraints. This is one of the most powerful entries for justifications of equivalence scaling as the right model, since relative entropy has strong statistical justifications, mostly in the form of maximum entropy or minimum discrimination information - see Jaynes 1957 or later in Kullback and Khairat 1966 for a justification in physics, or Kullback 1959 and Gokhale and Kullback 1978 for a justification in statistics.

Another justification closely connected to entropy minimisation is *maximum likelihood* models. For instance, if given a set of distributions Q and an empirical i.i.d. sample P, then the maximum likelihood for P being a sample of Q is given by the minimal relative entropy (Csiszár 1989; Darroch and Ratcliff 1972). Max-Likelihood justifications for applications in contingency tables are given in Fienberg 1970; Good 1963.

A different class of justifications for the validity of the matrix scaling approach are arguments showing that matrix scaling conserves certain form of interactions within

the matrix. For instance, matrix scaling conserves cross products (Mosteller 1968) and so-called k-cycles (Berger and Kelley 1979, k-cycles are certain products of matrix and inverse matrix entries). Both can be desirable for modeling reasons.

Finally, let us mention that the original justification (matrix scaling is a least-square type optimisation) made in Deming and Stephan 1940 turned out to be wrong very quickly and was superseded by real least-square methods in Stephan 1942 and later in Friedlander 1961 or Carey, Hendrickson, and Siddharthan 1981. Those however are not the same as equivalence scaling (see Section 6.8).

8.2. Axiomatic justification

Another justification for matrix scaling, which is particularly useful for application in elections is given in Balinski and Demange 1989b. Instead of considering just any matrix "close" to the original estimate, we want this matrix to fulfil a set of axioms.

Let A be a nonnegative matrix, r_+, c_+ (r_-, c_-) be upper (lower) bounds to the row and column sums and $h > 0$ be a scalar. As in Section 6.6, we denote the set of all matrices B fulfiling the bounds $q := (r_-, r_+, c_-, c_+, h)$ with $h = \sum_{ij} B_{ij}$ by $R(q)$. For any matrix A and any set of bounds q, we search for a method $F(A, q)$ to allocate one out of potentially many matrices A' fulfiling q and the following axioms:

Axiom 1 *Excactness:* If $r_- = c_- = 0$ and $r_+ = c_+ = \infty$ then $A' = (h/\sum_{ij} A_{ij})A$

Axiom 2 *Relevance:* If q' is another set of bounds such that $R(q') \subset R(q)$ and there exists a possible $A' \in R(q')$, then $F(A, q') \subset F(A, q) \cap R(q')$.

Axiom 3 *Uniformity:* For any matrix A' with bounds q, if we construct a new matrix $A'^{,\prime}$ by exchanging any submatrix $A'_{I\times J}$ by another submatrix $B_{I\times J}$ which fulfils the same row and column sums minus the part of these bound allocated in $A'_{(I\times J)^c}$, then $A'' \in F(A, q)$.

Axiom 4 *Monotonicity:* If we have two matrices A, B with $A_{ij} \leq B_{ij}$ for all (i, j), then it also holds that $A'_{ij} \leq B'_{ij}$ for all possible allocations.

Axiom 5 *Homogeneity:* Suppose $r_- = r_+$ and $c_- = c_+$. Then, if two rows of A are proportional and are constrained to the same row sum, then the corresponding rows in A' are always equal.

Then Balinski and Demange 1989b show that equivalence scaling (the fair share matrix of Section 6.6) is the unique allocation method $F(A, q)$ for all nonnegative matrices A where $R(q)$ contains a matrix with the same pattern as A.

8.3. A primer on applications

We will only sketch applications here since a complete list and discussion is probably infeasible.

Transportation planning A natural problem in geography is connected to predicting flows in a traffic network. If one considers for example a network of streets in a city at rush hour and a number of workers that want to get home, it is important to know how the traffic will be routed through the network. This is to a large degree a problem of physical modeling and a number of methods have been developed in the last century (for a recent introduction and overview see Ortúzar and Willumsen 2011[18]).

For our purposes, the most interesting question results from estimating trip distribution patterns from prior or incomplete data. In a simplified model, one could consider only origin and destination nodes (e.g. home quarters and work areas), given by a nonnegative matrix A. While the matrix is known for one year, it might be necessary to predict the changes given that the amount of trips to and from one destination change.

Several papers have treated a justification of the RAS method in this case. For instance, Evans 1970 argues that the method provides a unique outcome and it is easier to handle and to compute than other methods (Detroit method, growth factor method,...). A discussion of trip distribution with respect to Problem 1 can be found in Schneider and Zenios 1990.

Contingency table analysis In many situations ranging from biology to economics, contingency tables need to be estimated from sample data. Contingency tables list the frequency distributions of events in surveys, experiments, etc. They are highly useful to map several variables and study their relations.

As a specific example, suppose a small census in Germany tries to estimate migration between the states. While the number of citizens is recorded, which means that the total net migration is known, it is not known where each individual migrant came from. From a small survey among migrants, how can one estimate the true table with correct marginals in the best possible way? If one does a maximum likelihood estimation, the result is once again matrix scaling (cf. Fienberg 1970; Plane 1982).

Social accounting matrices Social accounting matrices, or SAMs, are an old tool developed in Stone 1962 and later popularised in Pyatt and Thorbecke 1976 to represent the national account of a country. To date, it is an important aspect of national and international accounting (as a random example see Klose, Opitz, and Schwarz 2004 from the German national institute of statistics. An introduction to social accounting can also be found in Pyatt and Round 1985 and, from a short mathematical perspective, in Schneider and Zenios 1990).

The idea is to represent income and outcome of a national economy in a matrix. Often, good growth estimates are known for the row and column sums and certain estimates are known for individual cells. The account estimates are then often not

[18]The authors also discuss the RAS method in chapter 5. I am however not convinced by their claim that Bregman provided the best analysis of the mathematics of the problem.

balanced, which can be achieved using matrix balancing or matrix scaling. Justifications can be imported from statistics, most notably maximum likelihood.

Schrödinger bridges In Schrödinger 1931, the author considered the following setup: Suppose we have a Brownian motion and a model which we are very confident about. In an experiment we observe its density at two times t_0, t_1. Now suppose they differ significantly from the model predictions. How can we reconcile these observations by updating our model without discarding it completely?

This problem has been studied in a whole line of papers since then from Fortet 1940 to Georgiou and Pavon 2015. The minimum relative entropy approach can be justified using large deviations (see Ruschendorf 1995).

Decreasing condition numbers Given a system of linear equations $Ax = b$ with nonsingular A, solving it relies on the Gaussian elimination procedure, which is known to be numerically unstable for matrices with bad condition number $\kappa(A) := \|A\|_\infty \|A^{-1}\|_\infty$. In order to increase the stability, we have to modify A, for example by multiplying with diagonal matrices D_1, D_2 and considering $D_1 A D_2$. Given that linear systems are ubiquitous in numerical analysis, it is of paramount importance to know how best to precondition a matrix in order to minimise calculation errors (see for instance the survey Benzi 2002). The answer to this question is problem dependent. Particular problems, where equivalence scaling is helpful to go include integral controllability tests based on steady-state information and the selection of sensors and actuators using dynamic information (see Braatz and Morari 1994).

One of the first papers to consider minimisation of κ using diagonal scaling was Osborne 1960, who focused on matrix balancing instead of equivalence scaling (see also Livne and Golub 2004 and Chen and Demmel 2000 for sparse matrices). Since the condition number contains the maximum norm, it might be best to require balanced maximum rows and columns instead of balanced row sums as observed in Bauer 1963; Curtis and Reid 1972. This works particularly well for sparse matrices. Equivalence scaling has been studied as early as Householder 2006[19] and later in Olschowka and Neumaier 1996. If we use other p-norms in the definition of κ, a convex programming solution for minimising κ using equivalence scaling is provided in Braatz and Morari 1994.

Note that unlike in all applications studied so far, preconditioning a matrix is useful not only for nonnegative matrices. This is one reason why matrix balancing was studied for copositive and not simply nonnegative matrices. A very different measure of the "goodness" of scaling which might also be of numerical relevance was studied in Rothblum and Schneider 1980, where the authors solved the problem of matrix balancing of a matrix such that the ratio between the biggest and smallest element of the scaled matrix becomes minimial.

[19] Reference from Braatz and Morari 1994

Elections A very important application of equivalence scaling can be found in Voting: Given election results in a federal election, how can one best distribute the seats among the parties within the states such that each party and each state is represented according to the outcome of the election? Note that here, we need to adjust for natural numbers, which requires rounding (cf. Maier, Zachariassen, and Zachariasen 2010).

Early methods based on a discretised RAS method were developed in Balinski and Demange 1989a. The problem is very intricate in itself, because the justifications rely on what is perceived as "fair" and any method that is fair in some instances is unfair in others (see for instance the discussions in Balinski and González 1997; Pukelsheim and Schuhmacher 2004; for an overview, see Niemeyer and Niemeyer 2008).

Other applications Various other applications of equivalence scaling and matrix balancing exist such as:

1. A Sudoku Solver based on a stochastic algorithm based on the RAS was developed in Moon, Gunther, and Kupin 2009.
2. An algorithm to rank web-pages was developed in Knight 2008. The RAS allows to derive an algorithm similar in scope to the HITS algorithm (Kleinberg 1999).
3. The RAS method is analysed as a relaxed clustering algorithm in data mining (Wang, Li, and König 2010). However, it turns out that methods based on other Bregman-divergences are more favourable.
4. Given a (discretised) quantum mechanical time evolution, can we construct a local hidden variable model of its evolution corresponding to a deterministic stochastic transition matrix (Aaronson 2005)?
5. Given a Markov chain with a doubly stochastic transition matrix and given an estimate of the transition matrix, the best estimate of the real transition matrix is given by a scaled matrix (Sinkhorn 1964).
6. Regularising optimal transportation by an entropy penalty term such that it can be computed using the RAS, which is already much faster than optimal transport algorithms Cuturi 2013.

9. Scalings for positive maps

We have already seen that Sinkhorn scaling is interesting for classical Schrödinger bridges as well as for scaling transition maps of Markov processes, etc. From a physics perspective, all these applications are classical physics, transforming classical states (probability distributions) to classical states.

In quantum mechanics, the basic objects are *quantum states*. For finite dimensional systems (such as spin systems), these quantum states are positive semidefinite matrices

with unit trace. A *quantum operation* then maps states to states, i.e. it needs to be positive: If $A \geq 0$, then $\mathcal{T}(A) \geq 0$. In fact, this is not all that is required for quantum operations, but one actually needs $\mathcal{T}$ to be *completely positive* (for an overview about quantum operations and quantum channels, see Nielsen and Chuang 2000; Wolf 2012). (Completely) Positive trace-preserving maps are then the natural generalisation of stochastic matrices. This raises the question whether concepts as irreducibility and a Perron-Frobenius theorem exist also for quantum channels and indeed they do. A Perron-Frobenius analogue was probably first described in Schrader 2000, while the analogue for full indecomposability was first used in Gurvits 2004.

Let us define the concepts:

Definition 9.1. A positive map $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$ with $\mathcal{M}_d = \mathbb{C}^{d\times d}$ is called *irreducible* (as in Evans and Høegh-Krohn 1978; Farenick 1996) if for any nonzero orthogonal projection P such that

$$\mathcal{E}(P\mathcal{M}_d P) \subseteq P\mathcal{M}_d P \tag{36}$$

we have $P = \mathbb{1}$.

Likewise, it is called *fully indecomposable* if for any two nonzero orthogonal projections P, Q with the same rank such that

$$\mathcal{E}(P\mathcal{M}_d P) \subseteq Q\mathcal{M}_d Q \tag{37}$$

we have $P = Q = \mathbb{1}$.

Finally, a map is called *positivity improving* (the analogue to positive matrices) if for all $A \geq 0$, $\mathcal{E}(A) > 0$.

A lot of different characterisations have been found (see Appendix C).

Furthermore, let us define:

Definition 9.2. Let $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$ be a positive map. Then $\mathcal{E}$ is called *rank non-decreasing* if for all $A \geq 0$

$$\operatorname{rank}(\mathcal{E}(A)) \geq \operatorname{rank}(A). \tag{38}$$

It is called *rank increasing*, if the $\geq$ sign in equation (38) is replaced by a $>$.

The connections of Definitions 9.1 and 9.2 are explained in Appendix C.

Let us now define what we mean by scaling a positive map:

Definition 9.3. Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. We say that $\mathcal{E}$ is *scalable* to a doubly stochastic map, if there exist $X,\ Y \in \mathcal{M}_d$ such that

$$\mathcal{E}'(\cdot) := Y^\dagger \mathcal{E}(X \cdot X^\dagger) Y \tag{39}$$

is doubly stochastic (i.e. $\mathcal{E}'(\mathbb{1}) = \mathcal{E}'^*(\mathbb{1}) = \mathbb{1}$).
We call a positive map *ε-doubly stochastic*, if

$$\mathrm{DS}(\mathcal{E}) := \mathrm{tr}((\mathcal{E}(\mathbb{1}) - \mathbb{1})^2) + \mathrm{tr}((\mathcal{E}^*(\mathbb{1}) - \mathbb{1})^2) \leq \varepsilon^2 \tag{40}$$

We call $\mathcal{E}$ *ε-scalable* if there exists a scaling as in equation (39) to an ε-doubly-stochastic map $\mathcal{E}'$.

The error function DS, which is similar to an L^2-error function for matrices, will serve twofold: first, it defines approximate scalability (which can alternatively be defined by convergence of the RAS) and second, it defines a progress measure for convergence similar to error functions as considered in Balinski and Demange 1989b.

We can now state the full analogue of equivalence scaling to doubly stochastic form:

Theorem 9.4. *Given a positive map $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$, it is scalable to a doubly stochastic map iff there exist some matrices X and Y such that $Y\mathcal{E}(X \cdot X^\dagger)Y^\dagger$ is a direct sum of fully indecomposable maps.*

The scaling matrices are unique iff $\mathcal{E}$ is fully indecomposable.

The fact that fully indecomposable matrices are uniquely scalable was first proved in Gurvits 2003. His work built on earlier work in Gurvits and Samorodnitsky 2002; Gurvits 2002 (see also Gurvits 2004), based on a generalisation of the convex approach in equation (16) and the London-Djokovic approach in equation (8).

Recently, the problem was considered with the hope to apply it for unital quantum channels in Idel 2013 (which has never been formally published) and shortly afterwards in Georgiou and Pavon 2015 while trying to define and study "quantum" Schrödinger bridges. Both approaches use nonlinear Perron-Frobenius theory and thereby a generalisation of equation (11) to get a result. The approaches derived from classical results are discussed in Section 9.1.

Even earlier than Gurvits, a very limited version of the theorem was proven in Kent, Linden, and Massar 1999 (with subsequent generalisations) using an approach that does not derive from any of the classical approaches but instead uses the Choi-Jamiolkowski isomorphism. This is described in Section 9.2.

The extension of the theorem to necessary and sufficient conditions has as far as I know not been formally published[20].

Furthermore, we can state an analogue of approximate scaling, which to date has only been considered in Gurvits 2004:

Theorem 9.5. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. Then $\mathcal{E}$ is approximately scalable (i.e. ε-scalable for any $\varepsilon > 0$) if and only if $\mathcal{E}$ is rank non-increasing.*

An overview about different approaches and how they derive from existing approaches can be found in Figure 9.

[20] Gurvits actually claims a proof for the fact that a positive map is uniquely scalable iff it is fully indecomposable, but I did not understand how the only if part follows.

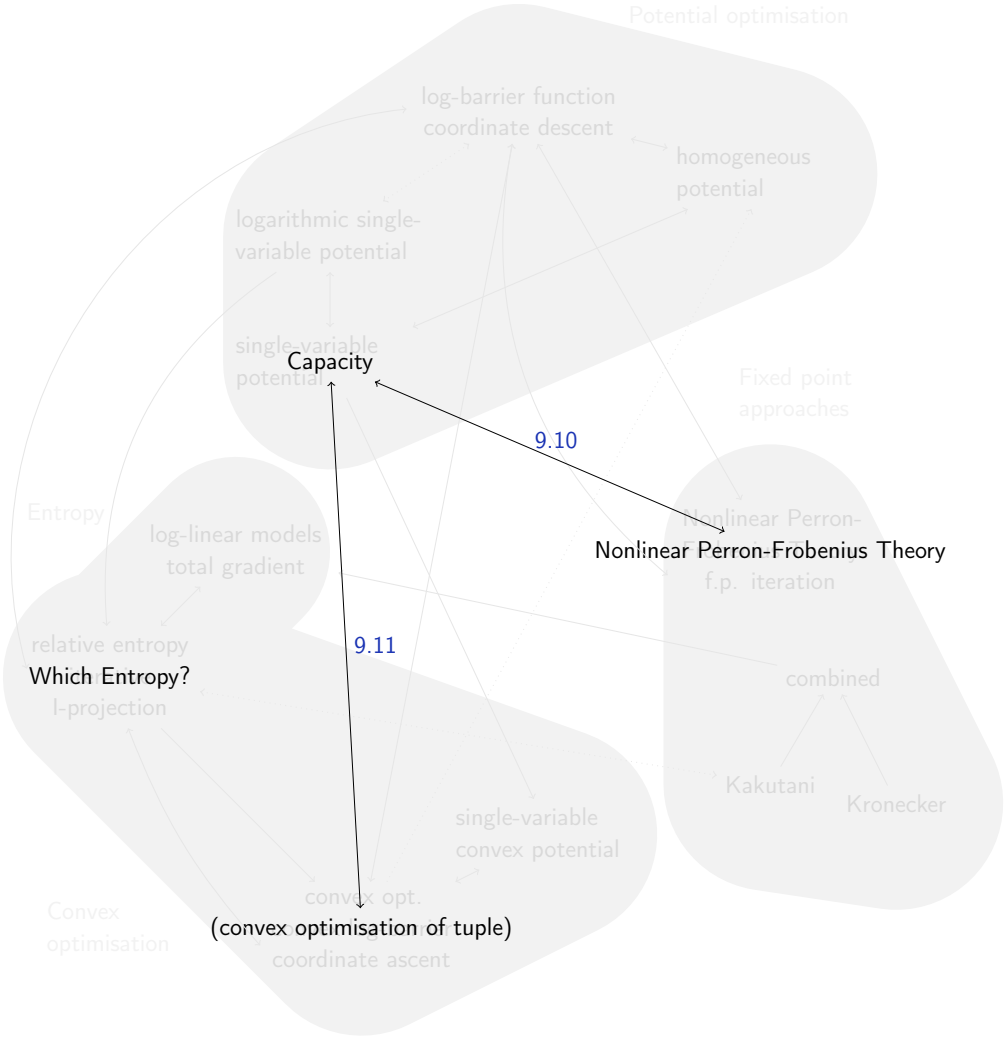

Figure 4: Approaches to positive map scalings and their connections. The classical approaches of Figure 3.1 are depicted in grey, while positive map approaches derived from classical approaches are overlayed in black.

9.1. Operator Sinkhorn theorem from classical approaches

We will now study how the theorems above were derived extending classical approaches to positive maps starting with an analogue of the RAS method for positive maps:

Algorithm 9.6. Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map.

1. Start with $\mathcal{E}_0 := \mathcal{E}$.
2. For each $i = 0, \ldots, n$ define:

$$\mathcal{E}_{2i+1}(\cdot) := \mathcal{E}_{2i}(\mathbb{1})^{1/2}\mathcal{E}_{2i}(\cdot)\mathcal{E}_{2i}(\mathbb{1})^{1/2} \tag{41}$$

$$\mathcal{E}_{2i+2}(\cdot) := \mathcal{E}_{2i+1}(\mathcal{E}_{2i+1}^*(\mathbb{1})^{1/2} \cdot \mathcal{E}_{2i+1}^*(\mathbb{1})^{1/2}) \tag{42}$$

3. Iterate till convergence

By construction, we iterate between trace-preserving (even) and unital (odd) maps.

9.1.1. Potential Theory and Convex programming

As stated, an approach along the lines of the London-Djokovic approach of equation (8) is found in Gurvits 2003; Gurvits 2004 (with methods of Gurvits 2002; Gurvits and Samorodnitsky 2000; Gurvits and Samorodnitsky 2002). Since the complete proofs are lengthy and scattered over several papers, we provide full proofs in Appendix D for the benefit of the reader. In this section, we only sketch the path of the proofs.

Recall that a matrix scaling exists iff the following is positive and the minimum is attained:

$$c(A) := \inf\left\{\prod_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_j \middle| \prod_{i=1}^{n} x_i = 1\right\} \tag{43}$$

Exchanging products with determinants and sums with traces, we obtain the following definition:

Definition 9.7. Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. Then define the *capacity* via

$$\mathrm{Cap}(\mathcal{E}) := \inf\{\det(\mathcal{E}(X)) | X > 0, \det(X) = 1\} \tag{44}$$

We will start with covering approximate scaling:

Approximate scalability The capacity is the right functional to study scaling:

Lemma 9.8 (Gurvits 2004)**.** *Let $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$ be a positive map. If $\mathrm{Cap}(\mathcal{E}) > 0$ then the RAS method of Algorithm 9.6 converges and $\mathcal{E}$ is ε-scalable for any $\varepsilon > 0$.*

The proof of this lemma uses the following observation: For any $C_1, C_2 > 0$ we have

$$\mathrm{Cap}(C_1 \mathcal{E}(C_2^\dagger \cdot C_2) C_1^\dagger) = \det(C_1 C_1^\dagger) \det(C_2 C_2^\dagger) \mathrm{Cap}(\mathcal{E}).$$

Then, a quick calculation shows that Algorithm 9.6 only decreases Cap using this equality. If $\mathrm{Cap}(\mathcal{E}) \neq 0$, one can then show that $\mathrm{DS}(\mathcal{E}_i) \to 0$ for $i \to \infty$.

Next, we need to see when the capacity is actually positive. To do this, for every unitary U we need to define the tuple

$$\mathbf{A}_{\mathcal{E},U} := (\mathcal{E}(u_1 u_1^\dagger), \ldots, \mathcal{E}(u_n u_n^\dagger)), \tag{45}$$

where u_i is the i-th column of U. This is done to connect the capacity with so called *mixed discriminants* (see also C), which are needed for the proof. In fact, we have:

Lemma 9.9. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map and $U \in U(n)$ a fixed unitary. Then defining*

$$\mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) := \inf \left\{ \det \left(\sum_i \mathcal{E}(u_i u_i^\dagger) \gamma_i \right) | \gamma_i > 0, \prod_{i=1}^n \gamma_i = 1 \right\}$$

where u_i are once again the rows of U, we have the following properties:

1. *Using the mixed discriminant M defined in Appendix C, we have*

$$M(\mathbf{A}_{\mathcal{E},U}) \leq \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) \leq \frac{n^n}{n!} M(\mathbf{A}_{\mathcal{E},U})$$

2. $\inf_{U \in U(n)} \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) = \mathrm{Cap}(\mathcal{E})$

Most of the proof is very technical and found in Gurvits and Samorodnitsky 2002. Some parts are explained in Appendix D.

Finally, this proves most of Theorem 9.5: We know that $\mathcal{E}$ is rank non-decreasing if and only if $\mathrm{Cap}(\mathcal{E})$ is positive. In that case, the RAS algorithm converges in which case the map is approximately scalable. For the other direction, one can use a simple contradiction argument: Any map close to a doubly stochastic map must be rank non-decreasing and as scaling does not change this property of a map, any approximately scalable map must be rank non-decreasing.

Exact scalability The capacity is also the correct generalisation for exact scaling:

Lemma 9.10 (Gurvits 2004)**.** *Let $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$ be a positive map. Then $\mathcal{E}$ is scalable to a doubly-stochastic map if and only if $\mathrm{Cap}(\mathcal{E}) > 0$ and the capacity can be achieved.*

The proof of this Lemma following Gurvits 2004 is given in Appendix D. The direction "Capacity is achieved $\Rightarrow$ the map is scalable", is proved by taking the Lagrangian and showing that at the minimum we have that

$$\nabla \ln(\det(\mathcal{E}(X))) = \mathcal{E}^*(\mathcal{E}(C)^{-1})$$

which implies scalability by Lemma 9.14. The converse direction is given by a direct calculation.

In order to prove that a map can be scaled to doubly stochastic form, one then needs to connect this lemma to full indecomposability of matrices. The proof is done using an argument involving strict convexity. Like the original London-Djokovic potential (8), the capacity is not a convex function, but one can make a substitution similar to Formulation (16) by considering the following function for any tuple $(A_i)_i$ of positive definite matrices:

$$f_A(\xi_1, \ldots, \xi_n) := \ln \det(e^{\xi_1} A_i + \ldots + e^{\xi_n} An). \tag{46}$$

Then we have:

Lemma 9.11. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map and given $U \in U(n)$, let $A = \mathbf{A}_{\mathcal{E},U}$. Then*

1. *f_A is convex on $\mathbb{R}^n$.*
2. *If $\mathcal{E}$ is fully indecomposable, then f_A is strictly convex on $\{\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n | \sum_i \xi_i = 0\}$.*

The proof is technical and uses mixed discriminants as well as results about them from Bapat 1989, which is why we only discuss it in the appendices. This is then used to prove

Lemma 9.12. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. If $\mathcal{E}$ is fully indecomposable, there exists a unique scaling of $\mathcal{E}$ to a doubly stochastic map.*

The idea is somewhat similar to the approximate Sinkhorn theorem: Since fully indecomposable maps are in particular rank non-decreasing, we know that the capacity is positive. For any $X > 0$, which is diagonalised by U, using the tuple $\mathbf{A}_{\mathcal{E},U}$, one can then see that $\det(\mathcal{E}(X)) = f_A(\log \lambda)$ with the eigenvalues λ of X. Showing that the infimum must lie inside a compact set then finishes the proof, since Lemma 9.11 implies existence and uniqueness of the minimum as f_A is strictly convex.

Using Lemma C.7, we can see then see that up to a unitary, every doubly stochastic map is a direct sum of fully indecomposable maps (much like doubly stochastic matrices are up to permutations a direct sum of fully indecomposable matrices, see Proposition A.5). Hence, any map which is a direct sum of fully indecomposable maps up to some scaling is clearly scalable to doubly stochastic maps. Sadly, the condition seems not very useful and the question remains open, whether one can simplify this condition.

However, that does in fact answer the question of unique scaling: Since the scaling for direct sums is not unique (we can always interchange summands), a map is uniquely scalable if and only if it is fully indecomposable.

9.1.2. Nonlinear Perron-Frobenius theory

The main idea of the alternative proofs in my Master's Thesis (Idel 2013) and the paper Georgiou and Pavon 2015 is to extend the Menon operator to positive semidefinite matrices:

Definition 9.13. Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive map, such that $\mathcal{E}(A), \mathcal{E}^*(A) > 0$ for all $A > 0$. Let $\mathcal{D}$ denote matrix inversion, then we define the following nonlinear map:

$$\begin{aligned}&\mathbf{T}_{\text{pos}} : \{A \in \mathcal{M}_n | A > 0\} \to \{A \in \mathcal{M}_n | A > 0\}\\ &\mathbf{T}_{\text{pos}}(\cdot) := \mathcal{D} \circ \mathcal{E}^* \circ \mathcal{D} \circ \mathcal{E}(\cdot)\end{aligned}$$

This map is well-defined and after normalisation, it sends positive definite matrices of trace one onto itself.

We can then reformulate the existence problem into a fixed point problem:

Lemma 9.14. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive map such that $\mathcal{E}(A), \mathcal{E}^*(A) > 0$ for all $A > 0$. Then there exist invertible $X, Y \in \mathcal{M}_n$ such that $Y^{-1}\mathcal{E}(X(\cdot)X^\dagger)Y^{-\dagger}$ is a doubly stochastic map if and only if $\mathbf{T}_{\text{pos}}$ has an eigenvector (a fixed point after normalisation) in the set of positive definite trace one matrices. Furthermore, X, Y can be chosen such that $X, Y > 0$.*

Proof. Let $\rho > 0$ be the positive definite eigenvector of $\mathcal{G}$. Then define $0 < \sigma := \mathcal{E}(\rho)$. Since ρ is an eigenvector, one immediately sees that $\mathcal{E}^*(\sigma^{-1}) = \lambda\rho^{-1}$ with $\lambda = \text{tr}(\mathcal{E}^*(\sigma_{-1}))^{-1}$. Now define $X := \sqrt{\rho}$ and $Y := \sqrt{\sigma}$ (i.e. $XX^\dagger = \rho$ and $YY^\dagger = \sigma$), then X, Y are positive definite and if we define the map:

$$\begin{aligned}&\mathcal{E}' : \mathcal{M}_n \to \mathcal{M}_n\\ &\mathcal{E}'(\cdot) := Y^{-1}\mathcal{E}(X(\cdot)X^\dagger)Y^{-\dagger}\end{aligned}$$

then a quick calculation shows $\mathcal{E}'(\mathbb{1}) = \mathbb{1}$ and $\mathcal{E}'^*(\mathbb{1}) = \mathbb{1}$:

$$\begin{aligned}\mathcal{E}'(\mathbb{1}) &= Y^{-1}\mathcal{E}(X(\mathbb{1})X^\dagger)Y^{-\dagger} = Y^{-1}\mathcal{E}(\rho)Y^{-\dagger}\\ &= Y^{-1}\sigma Y^{-\dagger} = Y^{-1}YY^\dagger Y^{-\dagger} = \mathbb{1}\end{aligned}$$

On the other hand, a similar calculation shows

$$\mathcal{E}'^*(\mathbb{1}) = X^\dagger\mathcal{E}^*(Y^{-\dagger}\mathbb{1}Y^{-1})X = \lambda\mathbb{1}$$

but since $\mathcal{E}'$ was shown to be unital, $\mathcal{E}'^*$ is trace-preserving and $\lambda = 1$ to begin with. Conversely, given X, Y as in the lemma, $XX^\dagger$ would be a fixed point of the Menon-operator. □

Note that this completes the proof of Lemma 9.10. We only have to see that the conditions at the minimum are met if and only if the Menon operator has a fixed point. This also provides the connection between the Menon operator and Gurvits' approach: As in the classical case, the conditions for a fixed point of the Menon operator are given by the Lagrange conditions of the London-Djokovic potential.

We observe that the Menon-operator, if it is defined, is a continuous, homogeneous positive map. This lets us give a proof of a weak form of the Operator Sinkhorn Theorem:

Proposition 9.15 (Idel 2013)**.** *Given a positive trace-preserving map $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ such that there exists an $\varepsilon > 0$ such that for all matrices $\rho \geq \varepsilon \mathbb{1}$ with unit trace it holds that $\mathcal{E}(\rho) \geq \frac{n\varepsilon}{1+(n-1)n\varepsilon}\mathbb{1}$, then we can find $X, Y > 0$ such that $Y^{-1}\mathcal{E}(X(\cdot)X^\dagger)Y^{-\dagger}$ is a doubly stochastic map.*

Proof. Let $\mathbf{T}_{\text{normpos}}(\cdot) := \mathbf{T}_{\text{pos}}(\cdot)/\operatorname{tr}(\mathbf{T}_{\text{pos}}(\cdot))$ is the normalised operator. Now assume that for all $\rho \geq \varepsilon\mathbb{1}$ with $\operatorname{tr}(\rho) = 1$, it holds $\mathcal{E}(\rho) \geq \delta\mathbb{1}$. In particular, if we call λ_{max} the maximal eigenvalue of $\mathcal{E}(\rho)$, then $\lambda_{max} \leq 1 - (n-1)\delta$. Hence we have:

$$\begin{aligned}
&\delta\mathbb{1} \leq \mathcal{E}(\rho) \leq (1-(n-1)\delta)\mathbb{1} \\
\Rightarrow \quad &\frac{1}{\delta}\mathbb{1} \geq \mathcal{D}(\mathcal{E}(\rho)) \geq \frac{1}{1-(n-1)\delta}\mathbb{1} \\
\Rightarrow \quad &\frac{1}{\delta}\mathbb{1} \geq \mathcal{E}^*(\mathcal{D}(\mathcal{E}(\rho))) \geq \frac{1}{1-(n-1)\delta}\mathbb{1} \\
\Rightarrow \quad &\delta\mathbb{1} \leq \mathcal{D}(\mathcal{E}^*(\mathcal{D}(\mathcal{E}(\rho)))) \leq (1-(n-1)\delta)\mathbb{1}
\end{aligned}$$

where we used the unitality of $\mathcal{E}^*$ in the third step. This implies

$$\mathbf{T}_{\text{normpos}}(\rho) \geq \frac{\delta}{1-(n-1)\delta}\mathbb{1}/n$$

Now we want $\frac{\delta}{1-(n-1)\delta} \geq \varepsilon n$, in which case the compact set of matrices $\{\rho > 0 | \operatorname{tr}(\rho) = 1, \rho \geq \varepsilon\mathbb{1}/n\}$ is mapped into itself, hence by Brouwer's fixed point theorem, we obtain a positive definite fixed point of $\mathcal{G}$. A quick calculation shows that this implies $\delta > n\varepsilon/(1+(n-1)n\varepsilon)$. □

Since a positive map $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ can always be converted into a trace-preserving map $\mathcal{E}'$ by setting $\rho := \mathcal{E}^*(\mathbb{1})$ and $\mathcal{E}'(\cdot) := \mathcal{E}(\sqrt{\rho^{-1}} \cdot \sqrt{\rho^{-1}})$, the assumption that $\mathcal{E}$ be trace-preserving is not really necessary. As a direct corollary, we obtain a similar result, which might be easier to use:

Corollary 9.16 (Idel 2013; Georgiou and Pavon 2015)**.** *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a trace-preserving and positivity improving map, then there exist maps $X, Y > 0$ such that $Y^{-1}\mathcal{E}(X(\cdot)X^\dagger)Y^{-\dagger}$ is a doubly stochastic map.*

As in the classical matrix case in Brualdi, Parter, and Schneider 1966, one idea to obtain necessary and sufficient criteria is to extend the map $\mathbf{T}_{\text{normpos}}$ to positive semidefinite matrices for all cases and then prove that there is a fixed point of the map inside the cone of positive definite matrices. However, we run into additional problems, since the cone of positive semidefinite matrices is not polyhedral (cf. Lemmens and Nussbaum 2012: there is no and cannot exist an equivalent version of theorem B.8; this does not preclude that an extension exists, but such a result must depend on the operator in question). Moreover, even if a continuous extension may be possible by using perturbation theory (for instance), the hardest part is to prove the existence of a fixed point inside the cone.

9.1.3. Other approaches and generalised scaling

We have seen that at least two classical approaches for proving that a matrix can be scaled to a doubly stochastic matrix can be extended to the quantum case without too much trouble (the proofs however might be more difficult): nonlinear Perron-Frobenius theory and the barrier function approach. In a sense, we have also seen convex programming approaches. An immediate question is whether one can extend the entropy approach. This was also asked in Gurvits 2004 and it is a major open question in Georgiou and Pavon 2015, since the motivation of Schrödinger bridges heavily relies on relative entropy minimisation. The answer is not clear since something like a quantum relative entropy is only used only on the level of matrices and a justification via the Choi-Jamiolkowski isomorphism is not immediate (see Section 9.2):

$$D(\rho\|\sigma) = \operatorname{tr}(\rho \log \rho - \rho \log \sigma). \tag{47}$$

Another question is how to extend the theorems from the doubly-stochastic map to cover arbitrary marginals, i.e. we want to scale a positive map $\mathcal{E}$ such that

$$\mathcal{E}(\rho) = \sigma, \qquad \mathcal{E}^*(\mathbb{1}) = \mathbb{1} \tag{48}$$

with some prespecified ρ, σ. For Gurvits' approach based on equation (8) this is not really straightforward, since it is unclear how to take appropriate powers of P, Q. For the approach via nonlinear Perron-Frobenius theory, this can be done to some degree:

Theorem 9.17 (Georgiou and Pavon 2015)**.** *Given a positivity improving map $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$ and two matrices $V, W > 0$ with $\operatorname{tr}(V) = \operatorname{tr}(W)$, there exist matrices $X, Y \in \mathcal{M}_d$ and a constant $\lambda > 0$ such that $\mathcal{E}'(\cdot) := Y\mathcal{E}(X \cdot X^\dagger)Y^\dagger$ fulfills*

$$\mathcal{E}'(V) = W$$
$$\mathcal{E}'^*(\mathbb{1}) = \mathbb{1}$$

Sketch of proof. The proof is a variation of the methods for the case $V = W = \mathbb{1}$. We consider the following Menon-type operator, which was essentially defined in Georgiou

and Pavon 2015:

$$\mathbf{T}_{\mathcal{E},V,W} := \mathbf{D}_1 \circ \mathcal{E}^* \circ \mathbf{D}_2 \circ \mathcal{E}$$

where

$$\mathbf{D}_1(\rho) = \rho^{-1/2} V^{-1} \rho^{-1/2}$$
$$\mathbf{D}_2(\rho) = (W^{1/2}(W^{-1/2}\rho^{-1}W^{-1/2})^{1/2}W^{1/2})^2$$

Step 1: Let $\mathcal{E}$ be positivity improving, then $\mathbf{T}_{\mathcal{E},V,W} : \overline{\mathcal{C}^d} \to \mathcal{C}^d$ is a well-defined, continuous, and homogeneous map. It is well-defined, since $\mathcal{E}$ maps $\overline{\mathcal{C}^d} \to \mathcal{C}^d$ and $\mathbf{D}_1$ and $\mathbf{D}_2$ send $\mathcal{C}^d \to \mathcal{C}^d$ if $V, W \in \mathcal{C}^d$. It is homogeneous, because $\mathcal{E}$ is linear and $\mathbf{D}_i(\lambda\rho) = \lambda^{-1}\mathbf{D}_i(\rho)$ for $i = 1, 2$. Finally, $\mathbf{D}_1$ is continuous as taking the square root of a positive definite matrix is continuous and matrix multiplication and inversion of positive definite matrices is continuous. Likewise, $\mathbf{D}_2$ is continuous and thus $\mathbf{T}_{\mathcal{E},V,W}$ as composition of continuous maps.

Step 2: We now claim that a scaling of $\mathcal{E}$ with as in the theorem with $X, Y > 0$ exists iff $\mathbf{T}_{\mathcal{E},V,W}$ has an eigenvector. This was observed in Georgiou and Pavon 2015 and is a straightforward but lengthy calculation.

Step 3: Finally, we can prove the existence of $X, Y > 0$ such that a scaling exists by invoking Brouwer's fixed point theorem. The map

$$\tilde{\mathbf{T}}(\cdot) : \overline{\mathcal{C}_1^d} \to \mathcal{C}_1^d \tilde{\mathbf{T}}(\cdot) := \mathbf{T}_{\mathcal{E},V,W}(\cdot) / \operatorname{tr}(\mathbf{T}_{\mathcal{E},V,W}(\cdot))$$

is a continuous, well-defined map, hence it has a fixed point. This is necessarily an eigenvector of $\mathcal{T}_{\mathcal{E},V,W}$, hence defines a scaling. □

The problem with this proof is that this Menon operator is no longer clearly a contraction mapping, hence uniqueness and convergence speed of the algorithm are not clear. Also, the obvious algorithm derived from this proof differs from the usual RAS algorithm. It is not clear how to remedy this or extend one of the other approaches. Also, this map has even worse prospect of being generalised to positive and not necessarily positivity improving maps. In any case, it is not immediately clear what the right necessary and sufficient conditions are. In the case of matrices, patterns were the important concept, but what is a pattern supposed to be for positive maps? One can always choose a basis and represent the map as matrix, but this is very much map-dependent and it is not clear what the correct interpretation will be.

Nevertheless, partial results for uniqueness have been achieved in Friedland 2016: The author proves that for positivity preserving maps $\mathcal{E}$, there exists a ball around $\mathbb{1}$ such that if V, W lie inside this ball, there exists a unique scaling of $\mathcal{E}$ to a trace-preserving positive map with $\mathcal{E}(V) = W$.

9.2. Operator Sinkhorn theorem via state-channel duality

Another formulation of the operator Sinkhorn theorem is given by the Choi-Jamiolkowski isomorphism. It states that given any positive map $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$, we have that

$$\tau_{\mathcal{E}} := (\mathrm{id} \otimes \mathcal{E})(\omega) \tag{49}$$

is a *block-positive* matrix (i.e. $\langle\phi_1|\langle\phi_2|\tau_{\mathcal{E}}|\phi_1\rangle|\phi_2\rangle \geq 0$ for all $|\phi_1\rangle, |\phi_2\rangle \in \mathcal{M}_n$). Here $\omega := 1/d \sum_{i,j=1}^{n} |ii\rangle\langle jj| \in \mathcal{M}_{n^2}$ is the so-called *maximally entangled state*. If $\mathcal{E}$ is completely positive, i.e. $\mathcal{E} \otimes \mathrm{id}_n$ is a positive map, then $\tau_{\mathcal{E}}$ is a positive semi-definite matrix. Now consider $X_1,\ X_2 \geq 0$ and $\mathcal{E}' := X_2^\dagger \mathcal{E}(X_1 \cdot X_1^\dagger) X_2$. We have

$$\tau_{\mathcal{E}'} = (X_1^{tr} \otimes X_2^\dagger)\tau_{\mathcal{E}}(X_1^{tr} \otimes X_2^\dagger)^\dagger \tag{50}$$

where we use $(\mathbb{1} \otimes X_1)\sum_i |ii\rangle = (X_1^{tr} \otimes \mathbb{1})\sum_i |ii\rangle$ and therefore

$$\tau_{\mathcal{E}'} = (\mathbb{1} \otimes X_2^\dagger)(\mathrm{id} \otimes \mathcal{E})((\mathbb{1} \otimes X_1)\omega(\mathbb{1} \otimes X_1)^\dagger)(\mathbb{1} \otimes X_2) \tag{51}$$

$$= (X_1^{tr} \otimes X_2^\dagger)(\mathrm{id} \otimes \mathcal{E})(\omega)(X_1^{tr\dagger} \otimes X_2) \tag{52}$$

Therefore, the task can be reformulated: Given a block positive matrix τ, find $X_1,\ X_2 \in \mathcal{M}_d$ such that

$$\tau' := (X_1 \otimes X_2)\tau(X_1 \otimes X_2)^\dagger$$

fulfils $\mathrm{tr}_2(\tau) = \mathrm{tr}_1(\tau) = \mathbb{1}/d$, where tr_i denotes the partial trace over the i-th system in $\mathcal{M}_d \otimes \mathcal{M}_d \equiv \mathcal{M}_{d^2}$. For $\tau \geq 0$ these operations are called *(local) filtering operations*. Often (c.f. Gittsovich et al. 2008; Wolf 2012), one asks for $X_1, X_2 \in SL(d)$ and the resulting trace being merely proportional to the identity, but this is of course just a normalisation.

We can then state an equivalent version of Sinkhorn scaling for positive map:

Proposition 9.18 (Kent, Linden, and Massar 1999; Leinaas, Myrheim, and Ovrum 2006; Verstraete, Dehaene, and De Moor 2001)**.** *Let $\rho \in \mathcal{M}_d \otimes \mathcal{M}_d$ be a positive definite density matrix. Then there exist matrices $X_1, X_2 \in \mathcal{M}_d$ such that*

$$(X_1 \otimes X_2)\rho(X_1 \otimes X_2)^\dagger = \frac{1}{d^2}\mathbb{1} + \sum_{k=1}^{k^2-1} \xi_k J_k^1 \otimes J_k^2 \tag{53}$$

where $\{J_k^1\}_k \subset \mathcal{M}_d$ and $\{J_k^2\}_k \subset \mathcal{M}_d$ form a basis of the traceless complex matrices and $\xi \in \mathbb{C}$ for the first and second tensor factor in $\mathcal{M}_d \otimes \mathcal{M}_d$ respectively.

Proof. Note that a positive definite ρ corresponds to a completely positive map, which maps positive semidefinite matrices to positive definite ones. In particular, the corresponding map is fully indecomposable. Hence by Theorem 9.4, there exists a scaling to a doubly stochastic map, which again corresponds to a positive definite $\tilde{\rho} \in \mathcal{M}_d \otimes \mathcal{M}_d$

such that $\mathrm{tr}_1(\tilde{\rho}) = \mathrm{tr}_2(\tilde{\rho}) = \mathbb{1}/d$. By construction, $\{\mathbb{1}/d, J_k^1\}_k$ and $\{\mathbb{1}/d, J_k^2\}$ form an orthonormal basis of $\mathcal{M}_d$, hence we can express $\tilde{\rho}$ as

$$\tilde{\rho} = \frac{1}{d^2}\mathbb{1} + \sum_{k=1}^{k^2-1} \xi_k J_k^1 \otimes J_k^2 + \sum_{k=1}^{k^2-1} \chi_k^1 \frac{\mathbb{1}}{d} \otimes J_k^1 + \chi_k^2 J_k^2 \otimes \frac{\mathbb{1}}{d}$$

with $\xi_k, \chi_k^1, \chi_k^2 \in \mathbb{C}$ for all k. Then

$$\begin{aligned}\mathrm{tr}_1(\tilde{\rho}) &= \frac{1}{d}\mathbb{1} + \sum_{k=1}^{k^2-1} \xi_k \,\mathrm{tr}(J_k^1) \otimes J_k^2 + \sum_{k=1}^{k^2-1} \left(\chi_k^1 \,\mathrm{tr}\left(\frac{\mathbb{1}}{d}\right) \otimes J_k^1 + \chi_k^2 \,\mathrm{tr}(J_k^2) \otimes \frac{\mathbb{1}}{d}\right) \\ &= \frac{1}{d}\mathbb{1} + \sum_{k=1}^{k^2-1} \chi_k^1 J_k^1 \overset{!}{=} \frac{1}{d}\mathbb{1}\end{aligned}$$

But then, since the J_k^1 are linearly independent, $\chi_k^1 = 0$ for all k. Likewise, $\chi_k^2 = 0$ for all k and we have the required normal form. □

The proposition has direct proofs and extensions to more than two parties (see for instance Verstraete, Dehaene, and De Moor 2002; Verstraete, Dehaene, and De Moor 2003; Wolf 2012). Here, it only uses the sufficient part of the criterion for scalability of positive maps, hence we can strengthen it to include parts of all block-positive matrices. Since, however, not all completely positive maps are fully indecomposable (e.g. the map $\mathcal{E} : \rho \to |\psi\rangle\langle\psi|$ for some vector $|\psi\rangle \in \mathbb{C}^d$ is not), it certainly does not extend to all states.

9.3. Convergence speed and stability results

Gurvits' proof already gives an estimate for the convergence speed of the scheme (see Theorem 4.7.3. in Gurvits 2004). Let us give an alternative proof using Hilbert's metric which is equivalent to the classical proof in Franklin and Lorenz 1989 and reminiscent of the convergence proof in Georgiou and Pavon 2015. Throughout the proof, we use several results from Appendix B, in particular the definition of Hilbert's projective metric d_H on the cone of positive semidefinite matrices and the definition of the contraction ratio γ in equation (68). To proceed, we define a metric on the space of positive maps that are scalable:

Definition 9.19. Let $\mathcal{E}, \mathcal{T} : \mathcal{M}_n \to \mathcal{M}_n$ be two positive maps such that $\mathcal{T}(\cdot) = Y\mathcal{E}(X \cdot X^\dagger)Y^\dagger$ for some positive matrices X, Y. Then

$$\Delta(\mathcal{E}, \mathcal{T}) = d_H(X, \mathbb{1}) + d_H(Y, \mathbb{1}) \tag{54}$$

defines a metric on the space of positive maps (two maps that cannot be scaled to each other have infinite distance).

A proof that this constitutes a metric may be found in Lemmens and Nussbaum 2012, Chapter 2. Recall the Sinkhorn iteration as defined in equations (41)-(42). For convenience we use a slightly different notation:

$$\begin{aligned}
\mathcal{E}^{(i)} &:= \mathcal{E}_{2i} \quad i > 0 \\
\mathcal{E}^{(i)\prime} &:= \mathcal{E}_{2i+1} \quad i \geq 0 \\
\rho^{(i)} &:= \mathcal{E}^{(i-1)}(\mathbb{1}) \quad i > 0 \\
\sigma^{(i)} &:= \mathcal{E}^{(i)\prime\dagger}(\mathbb{1}) \\
\mathcal{E}^{(0)} &:= \mathcal{E}.
\end{aligned}$$

Then:

Proposition 9.20. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positivity improving, trace preserving map. Let $\mathcal{T} := Y^{-1}\mathcal{E}(X \cdot X^\dagger)Y$ be the unique doubly stochastic scaling limit.*

$$\Delta(\mathcal{E}^{(k)}, \mathcal{T}) \leq \frac{\gamma^k}{1-\gamma}(d_H(\rho^{(1)}, e) + d_H(\sigma^{(1)}, e)) \tag{55}$$

$$\Delta(\mathcal{E}^{(k)\prime}, \mathcal{T}) \leq \frac{\gamma^k}{1-\gamma}(d_H(\rho^{(1)}, e) + d_H(\sigma^{(1)}, e)) \tag{56}$$

where $\gamma^{1/2} = \gamma^{1/2}(\mathcal{E})$ is the contraction ratio of equation (68). In particular, this implies via proposition B.6 (implying that here, $\gamma < 1$) that the convergence is geometric.

Proof. The proof is similar to the classical one in Franklin and Lorenz 1989. First recall the definition of Δ from Appendix B:

$$\Delta(\mathcal{E}) := \sup\{d_H(\mathcal{E}(\rho), \mathcal{E}(\sigma)) | \rho, \sigma \geq 0\}$$

Then $\Delta > 0$ but finite, since $\mathcal{E}$ is a positivity improving map and the maximum is attained. This is true, because it suffices to consider $d_H(\mathcal{E}(\rho), \mathcal{E}(\sigma))$ on the compact set $\{A \geq 0 | \|A\|_\infty = 1\}$ using Proposition B.6 (iii).

We first make the following observations:

$$d_H(\rho, \sigma) = d_H(\sigma^{-1/2}\rho\sigma^{-1/2}, \mathbb{1}) \quad \forall \rho, \sigma > 0 \tag{57}$$

$$d_H(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \leq \gamma^{1/2}(\mathcal{E}) d_H(\mathcal{E}(\rho), \mathcal{E}(\sigma)) \quad \forall \rho, \sigma > 0, \mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n \tag{58}$$

Equation (58) follows from the definition of $\gamma^{1/2}$. Equation (57) follows from the definition of M and m in the definition of the Hilbert metric and the fact that taking noncommutative inverses does not change positivity.

Let us now focus on $\gamma(\mathcal{E})$. Let $X, Y \in \mathcal{M}_n$ be invertible, then

$$\begin{aligned}
\Delta &= \sup\{d_H(Y^\dagger\mathcal{E}(X\rho X^\dagger)Y, Y^\dagger\mathcal{E}(X\sigma X^\dagger)Y) | \rho, \sigma \geq 0\} \\
&= \sup\{d_H(\mathcal{E}(X\rho X^\dagger), \mathcal{E}(X\sigma X^\dagger)) | \rho, \sigma \geq 0\}
\end{aligned}$$

$$= \sup\{d_H(\mathcal{E}(\tilde{\rho}), \mathcal{E}(\tilde{\sigma}))|\tilde{\rho}, \tilde{\sigma} \geq 0\}$$

using observation (58) and then X being invertible. In particular, this implies that for every $\gamma^{1/2}(\mathcal{E}^{(i)})$ we have a universal upper bound

$$\gamma^{1/2}(\mathcal{E}^{(i)}) < \tanh(\Delta/4). \tag{59}$$

Since $\Delta > 0$ but finite, this implies that we can upper bound each $\gamma(\mathcal{E}^{(i)})$ and $\gamma(\mathcal{E}'^{(i)})$ by some $\gamma < 1$. The rest is basically an iteration.

Consider $d_H(\rho^{(2)}, \mathbb{1})$. By definition, $\rho^{(2)} = \mathcal{E}^{(1)}(\mathbb{1}) = \mathcal{E}^{(1)\prime}((\sigma^{(1)})-1)$, and since all $\mathcal{E}^{(i)\prime}$ are unital:

$$d_H(\rho^{(2)}, \mathbb{1}) = d_H(\mathcal{E}^{(1)\prime}((\sigma^{(1)})^{-1}), \mathcal{E}^{(1)\prime}(\mathbb{1})) \leq \gamma^{1/2}(\mathcal{E}^{(1)\prime}) d_H(\mathbb{1}, \sigma^{(1)}) \tag{60}$$

where we used (58) and then (57). Similarly, since $\sigma^{(1)} = \mathcal{E}^{(0)\prime *}(\mathbb{1}) = \mathcal{E}^{(0)*}((\rho^{(1)})^{-1})$ and $\mathcal{E}^{(0)*}(\mathbb{1}) = \mathbb{1}$ by construction, we obtain:

$$d_H(\mathbb{1}, \sigma^{(1)}) = d_H(\mathcal{E}^{(0)*}(\mathbb{1}), \mathcal{E}^{(0)*}((\rho^{(1)})^{-1}) \leq \gamma^{1/2}(\mathcal{E}^{(0)*}) d_H(\rho^{(1)}, \mathbb{1}) \tag{61}$$

Combining (60) and (61) we obtain:

$$d_H(\rho^{(2)}, \mathbb{1}) \leq \gamma d_H(\rho^{(1)}, \mathbb{1}) \tag{62}$$

Similarly,

$$d_H(\sigma^{(2)}, \mathbb{1}) \leq \gamma d_H(\sigma^{(1)}, \mathbb{1}) \tag{63}$$

These are the key observations. Now using the definition of $\Delta(\cdot, \cdot)$ we obtain:

$$\begin{aligned} \Delta(\mathcal{E}^{(k)}, \mathcal{E}^{(k+1)}) &= d_H((\rho^{(k)})^{-1}, \mathbb{1}) + d_H((\sigma^{(k)})^{-1}, \mathbb{1}) \\ &\leq \gamma^{k-1}(d_H(\rho^{(1)}, \mathbb{1}) + d_H(\sigma^{(1)}, \mathbb{1})) \end{aligned}$$

Hence we have by the triangle inequality

$$\Delta(\mathcal{E}^{(0)}, \mathcal{E}^{(k+1)}) \leq \sum_{l=0}^{k-1} \gamma^l (d_H(\rho^{(1)}, \mathbb{1}) + d_H(\sigma^{(1)}, \mathbb{1}))$$

and therefore, if $\mathcal{T}$ denotes the limit of the Sinkhorn iteration, using the geometric series

$$\Delta(\mathcal{E}^{(0)}, \mathcal{T}) \leq \frac{1}{1-\gamma} (d_H(\rho^{(1)}, \mathbb{1}) + d_H(\sigma^{(1)}, \mathbb{1})) \tag{64}$$

$$\Delta(\mathcal{E}^{(k)}, \mathcal{T}) \leq \frac{\gamma^k}{1-\gamma} (d_H(\rho^{(1)}, \mathbb{1}) + d_H(\sigma^{(1)}, \mathbb{1})) \tag{65}$$

The other inequality for the maps $\mathcal{E}'$ follows from symmetric arguments. □

Note that in contrast to the classical case in Franklin and Lorenz 1989, because of the noncommutativity in equation (57), a simple extension to the general scaling of positivity improving maps seems not possible.

Next, we wish to generalise also the stability results. It seems natural that this should follow from the contraction results above:

Corollary 9.21. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be positivity improving, then the scaling is continuous in $\mathcal{E}$.*

Proof. Let $\mathcal{E}$ be a positivity improving map and $\mathcal{E}' = \mathcal{E} + \delta\mathcal{T}$ be a perturbation which is again positivity preserving, where $\mathcal{T}$ is a positive map with $\|\mathcal{T}\| = 1$ (for instance in the operator norm).

Then let X, Y be such that they scale $\mathcal{E}$ to a doubly stochastic map. This implies that

$$Y\mathcal{E}'(XX^\dagger)Y^\dagger = \mathbb{1} + \delta Y\mathcal{T}(XX^\dagger)Y^\dagger$$
$$X^\dagger\mathcal{E}'^*(Y^\dagger Y)X = \mathbb{1} + \delta X^\dagger\mathcal{T}^*(Y^\dagger Y)X$$

and the marginals are also close to $\mathbb{1}$. In fact, for any $\varepsilon > 0$ we can find $\delta > 0$ such that

$$d_H(Y\mathcal{E}'(XX^\dagger)Y^\dagger, \mathbb{1}) + d_H(X^\dagger\mathcal{E}'^*(Y^\dagger Y)X, \mathbb{1}), < \varepsilon$$

But then, by equation (64), we have that if $\mathcal{E}'^{\prime}$ is the scaling of $Y\mathcal{E}'(XX^\dagger)Y^\dagger$ to a doubly stochastic map, then

$$\Delta(Y\mathcal{E}'(XX^\dagger)Y^\dagger, \mathcal{E}'^{\prime}) \leq \frac{1}{1-\gamma}(d_H(\rho^{(1)}, \mathbb{1}) + d_H(\sigma^{(1)}, \mathbb{1})) < \frac{1}{1-\gamma}\varepsilon$$

Using the triangle inequality and the fact that $\Delta(\mathcal{E}, \mathcal{E}') < C\varepsilon$ for some constant C finishes the proof. □

As noted, both theorems can be extended to cover all exactly scalable positive maps using Gurvits' Theorem 4.7.3. of Gurvits 2004. Given the result for classical matrices, it seems natural that the convergence speed for rank non-decreasing but not exactly scalable matrices should not be geometric.

9.4. Applications of the Operator Sinkhorn Theorem

Let us finally mention applications of the operator version of Sinkhorn's theorem. The state-version of the theorem, since it can be seen as a normal form for states under local operations, has been applied in the study of states under LOCC operations (see for instance Kent, Linden, and Massar 1999; Leinaas, Myrheim, and Ovrum 2006).

The approximate operator version was developed to obtain polynomial-time algorithms (Sinkhorn scaling) for a problem known as "Edmond's problem". It asks the following question (Gurvits 2004): Given a linear subspace A of $\mathcal{M}_n$, does there exist a nonsingular matrix in V? The question can be asked also over different number fields

and in different contexts. It is particularly interesting, because it is related to rational identity testing over non-commutative variables as studied in Garg et al. 2015; Ivanyos, Qiao, and Subrahmanyam 2015. For further input we refer the reader to the extended and well-written review of the applications in Garg et al. 2015.

Finally, the exact scalability of fully indecomposable positive maps provided bounds on the *mixed discriminant* of matrix tuples, which is interesting to provide permanent bounds (Gurvits and Samorodnitsky 2000).

Acknowledgement I was first acquainted with the topic of matrix scaling through my Master's thesis supervisor Michael Wolf. We also worked together on unitary scaling and I think him for his valuable input. I am supported by the Studienstiftung des deutschen Volkes.

References

Aaronson, Scott (2005). "Quantum computing and hidden variables". In: *Phys. Rev. A* 71 (3). DOI: 10.1103/PhysRevA.71.032325.

Achilles, Eva (1993). "Implications of convergence rates in Sinkhorn balancing". In: *Linear Algebra and its Applications* 187, pp. 109 –112. DOI: 10.1016/0024-3795(93)90131-7.

Ahuja, Ravindra K. et al. (1994). "Improved Algorithms for Bipartite Network Flow". In: *SIAM Journal on Computing* 23.5, pp. 906–933. DOI: 10.1137/S0097539791199334.

Bacharach, Michael (1965). "Estimating Nonnegative Matrices from Marginal Data". In: *International Economic Review* 6.3, pp. 294–310. ISSN: 00206598, 14682354.

— (1970). *Biproportional Matrices and Input-Output Change*. Cambridge University Press.

Bachem, Achim and Bernhard Korte (1979). "On the RAS-algorithm". In: *Computing* 23.2, pp. 189–198. DOI: 10.1007/BF02252097.

— (1980). "Minimum norm problems over transportation polytopes". In: *Linear Algebra and its Applications* 31, pp. 103 –118. DOI: 10.1016/0024-3795(80)90211-6.

Balakrishnan, H., Inseok Hwang, and C. J. Tomlin (2004). "Polynomial approximation algorithms for belief matrix maintenance in identity management". In: *Decision and Control, 2004. CDC. 43rd IEEE Conference on*. Vol. 5, pp. 4874 –4879. DOI: 10.1109/CDC.2004.1429569.

Balinski, Michel and Gabrielle Demange (1989a). "Algorithms for proportional matrices in reals and integers". In: *Mathematical Programming* 45.1-3, pp. 193–210. DOI: 10.1007/BF01589103.

— (1989b). "An axiomatic approach to proportionality between matrices". In: *Mathematics of Operations Research* 14.4, pp. 700–719. DOI: 10.1287/moor.14.4.700.

Balinski, Michel and Victoriano Ramírez González (1997). "Mexican electoral law: 1996 version". In: *Electoral Studies* 16.3, pp. 329–340. DOI: 10.1016/S0261-3794(97)00025-5.

Bapat, Ravindra B. (1982). "D1AD2 theorems for multidimensional matrices". In: *Linear Algebra and its Applications* 48.0, pp. 437 –442. DOI: 10.1016/0024-3795(82)90125-2.

— (1989). "Mixed discriminants of positive semidefinite matrices". In: *Linear Algebra and its Applications* 126.0, pp. 107 –124. DOI: 10.1016/0024-3795(89)90009-8.

Bapat, Ravindra B. and T. E. S. Raghavan (1989). "An extension of a theorem of Darroch and Ratcliff in loglinear models and its application to scaling multidimensional matrices". In: *Linear Algebra and its Applications* 114 - 115.0. Special Issue Dedicated to Alan J. Hoffman, pp. 705 –715. DOI: 10.1016/0024-3795(89)90489-8.

— (1997). *Nonnegative Matrices and Applications*. Cambridge Books Online. Cambridge University Press. ISBN: 9780511529979.

Bauer, F. L. (1963). "Optimally scaled matrices". In: *Numerische Mathematik* 5.1, pp. 73–87. DOI: 10.1007/BF01385880.

— (1965). "An elementary proof of the Hopf inequality for positive operators". In: *Numerische Mathematik* 7.4, pp. 331–337. DOI: 10.1007/BF01436527.

Benoist, Tristan and Ion Nechita (2016). *On bipartite unitary matrices generating subalgebra-preserving quantum operations*. arXiv:1608.05811v1 [quant-ph].

Benzi, Michele (2002). "Preconditioning Techniques for Large Linear Systems: A Survey". In: *Journal of Computational Physics* 182.2, pp. 418 –477. DOI: 10.1006/jcph.2002.7176.

Berger, Marc A. and C. T. Kelley (1979). "A variational equivalent to diagonal scaling". In: *Journal of Mathematical Analysis and Applications* 72.1, pp. 291 –304. DOI: 10.1016/0022-247X(79)90290-7.

Bhatia, Rajendra (1996). *Matrix Analyis*. Springer. ISBN: 978-1-4612-0653-8.

Bingen, F (1965). "Simplification de la Démonstration d'un Théorème de MWM Gorman". In: *Université Libre de Bruxelles, duplicated*.

Biran, Paul, Michael Entov, and Leonid Polterovich (2004). "Calabi quasimorphisms for the symplectic ball". In: *Communications in Contemporary Mathematics* 06.05, pp. 793–802. DOI: 10.1142/S0219199704001525.

Birkhoff, Garrett (1957). "Extensions of Jentzsch's Theorem". In: *Transactions of the American Mathematical Society* 85.1, pp. 219–227. DOI: 10.1090/S0002-9947-1957-0087058-6.

Bishop, Y. M. M. (1967). "Multidimensional contingency tables: cell estimates". PhD thesis. Harvard University.

Borobia, Alberto and Rafael Cantó (1998). "Matrix scaling: A geometric proof of Sinkhorn's theorem". In: *Linear Algebra and its Applications* 268.0, pp. 1 –8. DOI: 10.1016/S0024-3795(97)00010-4.

Borwein, Jonathan M., Adrian Stephen Lewis, and Roger D. Nussbaum (1994). "Entropy minimization, *DAD* problems, and doubly stochastic kernels". In: *Journal of Functional Analysis* 123.2, pp. 264–307. DOI: 10.1006/jfan.1994.1089.

Bot, Radu Ioan and Sorin-Mihai Grad (2010). "Wolfe duality and Mond-Weir duality via perturbations". In: *Nonlinear Analysis: Theory, Methods & Applications* 73.2, pp. 374 –384. DOI: 10.1016/j.na.2010.03.026.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press. ISBN: 0521833787.

Braatz, Richard D. and Manfred Morari (1994). "Minimizing the Euclidean Condition Number". In: *SIAM Journal on Control and Optimization* 32.6, pp. 1763–1768. DOI: 10.1137/S0363012992238680.

Bradley, Andrew M. (2010). "Algorithms for the Equilibration of Matrices and Their Application to Limited-Memory Quasi-Newton Methods". PhD thesis. Stanford ICME.

Bradley, Andrew M. and Walter Murray (2011). *Matrix-free approximate equilibration*. arXiv preprint arXiv:1110.2805.

Bregman, L. M. (1967). "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR Computational Mathematics and Mathematical Physics* 7.3, pp. 200 –217. DOI: 10.1016/0041-5553(67)90040-7.

Brown, David T. (1959). "A note on approximations to discrete probability distributions". In: *Information and Control* 2.4, pp. 386 –392. DOI: 10.1016/S0019-9958(59)80016-4.

Brown, Jack B., Phillip J. Chase, and Arthur O. Pittenger (1993). "Order independence and factor convergence in iterative scaling". In: *Linear Algebra and its Applications* 190.0, pp. 1 –38. DOI: 10.1016/0024-3795(93)90218-D.

Brualdi, Richard A. (1968). "Convex sets of non-negative matrices". In: *Canadian Journal of Mathematics* 20, pp. 144–157. DOI: 10.4153/CJM-1968-016-9.

— (1974). "The *DAD* theorem for arbitrary row sums". In: *Proc. Amer. Math. Soc.* 45, pp. 189–194. DOI: 10.1090/S0002-9939-1974-0354737-8.

Brualdi, Richard A., Seymour V. Parter, and Hans Schneider (1966). "The diagonal equivalence of a nonnegative matrix to a stochastic matrix". In: *Journal of Mathematical Analysis and Applications* 16.1, pp. 31 –50. DOI: 10.1016/0022-247X(66)90184-3.

Bunch, James R. (1971). "Equilibration of Symmetric Matrices in the Max-Norm". In: *J. ACM* 18.4, pp. 566–572. DOI: 10.1145/321662.321670.

Carey, Malachy, Chris Hendrickson, and Krishnaswami Siddharthan (1981). "A Method for Direct Estimation of Origin/Destination Trip Matrices". In: *Transportation Science* 15.1, pp. 32–49. DOI: 10.1287/trsc.15.1.32.

Caussinus, Henri (1965). "Contribution à l'analyse statistique des tableaux de corrélation". In: *Annales de la Faculté des sciences de Toulouse : Mathématiques* 29, pp. 77–183.

Censor, Yair and Arnold Lent (1981). "An iterative row-action method for interval convex programming". In: *Journal of Optimization Theory and Applications* 34.3, pp. 321–353. DOI: 10.1007/BF00934676.

Chen, Tzu-Yi and James W. Demmel (2000). "Balancing sparse matrices for computing eigenvalues". In: *Linear Algebra and its Applications* 309.1, pp. 261 –287. DOI: 10.1016/S0024-3795(00)00014-8.

Cottle, Richard W., Steven G. Duvall, and Karel Zikan (1986). "A Lagrangean relaxation algorithm for the constrained matrix problem". In: *Naval Research Logistics Quarterly* 33.1, pp. 55–76. DOI: 10.1002/nav.3800330106.

Csima, J. and B. N. Datta (1972). "The DAD theorem for symmetric non-negative matrices". In: *Journal of Combinatorial Theory, Series A* 12.1, pp. 147 –152. DOI: 10.1016/0097-3165(72)90090-8.

Csiszár, Imre (1975). "I-divergence geometry of probability distributions and minimization problems". In: *The Annals of Probability*, pp. 146–158.

— (1989). "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling". In: *The Annals of Statistics*, pp. 1409–1413.

Curtis, A. R. and J. K. Reid (1972). "On the Automatic Scaling of Matrices for Gaussian Elimination". In: *IMA Journal of Applied Mathematics* 10.1, pp. 118–124. DOI: 10.1093/imamat/10.1.118.

Cuturi, Marco (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*, pp. 2292–2300.

Darroch, J. N. and D. Ratcliff (1972). "Generalized Iterative Scaling for Log-Linear Models". In: *The Annals of Mathematical Statistics* 43.5, pp. 1470–1480. DOI: 10.1214/aoms/1177692379.

De Vos, Alexis and Stijn De Baerdemacker (2014a). "Scaling a unitary matrix". In: *Open Systems & Information Dynamics* 21.4. DOI: 10.1142/S1230161214500139.

— (2014b). "The Synthesis of a Quantum Circuit". In: *Proceedings of the 11th International Workshop of Boolean Problems*. Freiberg University of Mining and Technology, pp. 129–136.

— (2016). "The Birkhoff theorem for unitary matrices of prime dimension". In: *Linear Algebra and its Applications* 493, pp. 455 –468. DOI: 10.1016/j.laa.2015.12.005.

Deming, W. Edwards and Frederick F. Stephan (1940). "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known". In: *Ann. Math. Statist.* 11.4, pp. 427–444. DOI: 10.1214/aoms/1177731829.

Dempster, A. P. (1969). "Some theory related to fitting exponential models". Research Report S-4, Department of Statistics, Harvard Univ.

Djoković, D. (1970). "Note on Nonnegative Matrices". In: *Proceedings of the American Mathematical Society* 25.1, pp. 80–82. DOI: 10.1090/S0002-9939-1970-0257114-X.

Dykstra, Richard L. (1985). "An Iterative Procedure for Obtaining I-Projections onto the Intersection of Convex Sets". In: *The Annals of Probability* 13.3, pp. 975–984. DOI: 10.2307/2243723.

Eaves, B. Curtis et al. (1985). "Line-sum-symmetric scalings of square nonnegative matrices". In: *Mathematical Programming Essays in Honor of George B. Dantzig Part II*. Ed. by Richard W. Cottle. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 124–141. ISBN: 978-3-642-00921-1. DOI: 10.1007/BFb0121080.

Eischen, Ellen E. et al. (2002). "Patterns, linesums, and symmetry". In: *Linear Algebra and its Applications* 357.1, pp. 273 –289. DOI: 10.1016/S0024-3795(02)00417-2.

El-Badry, M. A. and F. F. Stephan (1955). "On Adjusting Sample Tabulations to Census Counts". In: *Journal of the American Statistical Association* 50.271, pp. 738–762. DOI: 10.1080/01621459.1955.10501964.

Evans, A.W. (1970). "Some properties of trip distribution methods". In: *Transportation Research* 4.1, pp. 19 –36. DOI: 10.1016/0041-1647(70)90072-9.

Evans, David E. and Raphael Høegh-Krohn (1978). "Spectral Properties of Positive Maps on C*-Algebras". In: *Journal of the London Mathematical Society* s2-17.2, pp. 345–355. DOI: 10.1112/jlms/s2-17.2.345.

Evans, Suzanne P. and Howard R. Kirby (1974). "A three-dimensional Furness procedure for calibrating gravity models". In: *Transportation Research* 8.2, pp. 105 –122. DOI: 10.1016/0041-1647(74)90037-9.

Eveson, Simon P. and Roger D. Nussbaum (1995). "An elementary proof of the Birkhoff-Hopf theorem". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 117 (01), pp. 31–55. DOI: 10.1017/S0305004100072911.

Farenick, Douglas R. (1996). "Irreducible Positive Linear Maps on Operator Algebras". In: *Proceedings of the American Mathematical Society* 124.11, pp. 3381–3390. DOI: 10.1090/S0002-9939-96-03441-7.

Fienberg, Stephen E. (1970). "An Iterative Procedure for Estimation in Contingency Tables". In: *The Annals of Mathematical Statistics* 41.3, pp. 907–917. DOI: 10.1214/aoms/1177696968.

Fofana, Ismael, André Lemelin, and John Cockburn (2002). "Balancing a social accounting matrix". In: *Laval: Centre de Recherche en Économie et Finances Appliquées (CREFA) Université Laval*.

Fortet, R. (1940). "Résolution d'un système d'equations de M. Schrödinger". In: *J. Math. Pure Appl.* IX, pp. 83 –105.

Franklin, Joel and Jens Lorenz (1989). "On the scaling of multidimensional matrices". In: *Linear Algebra and its Applications* 114 - 115.0. Special Issue Dedicated to Alan J. Hoffman, pp. 717 –735. DOI: 10.1016/0024-3795(89)90490-4.

Fréchet, M. (1960). "Sur les tableaux dont les marges et des bornes sont données". In: *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 28.1/2, pp. 10–32. DOI: 10.2307/1401846.

Friedland, Shmuel (2016). *On Schrödinger's bridge problem*. arXiv:1608.05862v1 [math-ph].

Friedlander, D. (1961). "A Technique for Estimating a Contingency Table, Given the Marginal Totals and Some Supplementary Data". In: *Journal of the Royal Statistical Society. Series A (General)* 124.3, pp. 412–420. DOI: 10.2307/2343244.

Frobenius, Ferdinand Georg (1912). "Über Matrizen aus nicht negativen Elementen". In: *Sitzungsbericht Königl. Preuss. Akad. Wiss.* Pp. 456–477. DOI: 10.3931/e-rara-18865.

Führ, Hartmut and Ziemowit Rzeszotnik (2015). "On biunimodular vectors for unitary matrices". In: *Linear Algebra and its Applications* 484.0, pp. 86 –129. DOI: 10.1016/j.laa.2015.06.019.

Fürer, Martin (2004). "Quadratic convergence for scaling of matrices". In: *ALENEX/ANALC*, pp. 216–223.

Furness, K. P. (1962). "Trip forecasting". Paper presented at a seminar on the use of computers in traffic planning. London, unpublished.

Garg, Ankit et al. (2015). *A deterministic polynomial time algorithm for non-commutative rational identity testing with applications*. arXiv:1511.03730v2 [cs.CC].

Georgiou, Tryphon T. and Michele Pavon (2015). "Positive contraction mappings for classical and quantum Schrödinger systems". In: *Journal of Mathematical Physics* 56.3, 033301. DOI: 10.1063/1.4915289.

Gittsovich, O. et al. (2008). "Unifying several separability conditions using the covariance matrix criterion". In: *Phys. Rev. A* 78 (5), p. 052319. DOI: 10.1103/PhysRevA.78.052319.

Gokhale, D. V. and Solomon Kullback (1978). "The minimum discrimination information approach in analyzing categorical data". In: *Communications in Statistics - Theory and Methods* 7.10, pp. 987–1005. DOI: 10.1080/03610927808827687.

Golan, Amos and George Judge (1996). "Econometric Methodology, Part I: Recovering information in the case of underdetermined problems and incomplete economic data". In: *Journal of Statistical Planning and Inference* 49.1, pp. 127 –136. DOI: 10.1016/0378-3758(95)00033-X.

Golan, Amos, George Judge, and Douglas Miller (1997). *Maximum entropy econometrics: Robust estimation with limited data*. Chichester (United Kingdom) John Wiley and Sons.

Goldstein, A. A. and J. F. Price (1967). "An effective algorithm for minimization". In: *Numerische Mathematik* 10.3, pp. 184–189. DOI: 10.1007/BF02162162.

Golitschek, Manfred, Uriel G. Rothblum, and Hans Schneider (1983). "A conforming decomposition theorem, a piecewise linear theorem of the alternative, and scalings of matrices satisfying lower and upper bounds". In: *Mathematical Programming* 27.3, pp. 291–306. DOI: 10.1007/BF02591905.

Good, I. J. (1963). "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables". In: *The Annals of Mathematical Statistics* 34.3, pp. 911–934. DOI: 10.1214/aoms/1177704014.

— (1965). "The estimation of probabilities: An essay on modern Bayesian methods". MIT Research Monograph No. 30.

Gorman, W. M. (1963). "Estimating trends in Leontief matrices: a note on Mr. Bacharach's paper". In: *Nuffield College, Oxford*.

Grad, J. (1971). "Matrix Balancing". In: *The Computer Journal* 14.3, pp. 280–284. DOI: 10.1093/comjnl/14.3.280.

Gurvits, Leonid (2002). *Quantum Matching Theory (with new complexity theoretic, combinatorial and topological insights on the nature of the Quantum Entanglement)*. arXiv:0201022 [quant-ph].

— (2003). "Classical Deterministic Complexity of Edmonds' Problem and Quantum Entanglement". In: *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*. STOC '03. New York, NY: ACM, pp. 10–19. ISBN: 1-58113-674-9. DOI: 10.1145/780542.780545.

— (2004). "Classical complexity and quantum entanglement". In: *Journal of Computer and System Sciences* 69.3. Special Issue on {STOC} 2003, pp. 448 –484. DOI: 10.1016/j.jcss.2004.06.003.

Gurvits, Leonid and Alex Samorodnitsky (2000). "A Deterministic Polynomial-time Algorithm for Approximating Mixed Discriminant and Mixed Volume". In: *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*. STOC '00. New York, NY: ACM, pp. 48–57. ISBN: 1-58113-184-4. DOI: 10.1145/335305.335311.

— (2002). "A Deterministic Algorithm for Approximating the Mixed Discriminant and Mixed Volume, and a Combinatorial Corollary". In: *Discrete & Computational Geometry* 27.4, pp. 531–550. DOI: 10.1007/s00454-001-0083-2.

Hartfiel, Darald J. (1971). "Concerning diagonal similarity of irreducible matrices". In: *Proceedings of the American Mathematical Society* 30.3, pp. 419–425.

Herrmann, D. (1973). "Some Results on Uniqueness and Existence of Constrained Matrix Problems". PhD thesis. Faculty of Commerce, University of Birmingham.

Hershkowitz, Daniel, Uriel G. Rothblum, and Hans Schneider (1988). "Classifications of Nonnegative Matrices Using Diagonal Equivalence". In: *SIAM Journal on Matrix Analysis and Applications* 9.4, pp. 455–460. DOI: 10.1137/0609038.

Hetyei, G. (1964). "2 x l-es téglalapokkal lefedhetö idomokröl". In: *Pécsi Tanárképzö Föisk. Tud. Közl*, pp. 351–368.

Hobby, Charles and Ronald Pyke (1965). "Doubly stochastic operators obtained from positive operators". In: *Pacific Journal of Mathematics* 15.1, pp. 153–157. DOI: 10.2140/pjm.1965.15.153.

Householder, Alston S. (2006). *The Theory of Matrices in Numerical Analysis*. First published in 1964. Dover Publications. ISBN: 978-0486449722.

Hutchinson, George (2016). "On the cardinality of complex matrix scalings". In: *Special Matrices* 4.1, pp. 141–150. DOI: 10.1515/spma-2016-0014.

Idel, Martin (2013). "On the structure of positive maps". MA thesis. Ludwig-Maximilian Universität München, Technische Universität München.

Idel, Martin and Michael M. Wolf (2015). "Sinkhorn normal form for unitary matrices". In: *Linear Algebra and its Applications* 471, pp. 76 –84. DOI: 10.1016/j.laa.2014.12.031.

Ireland, C. T. and Solomon Kullback (1968). "Contingency tables with given marginals". In: *Biometrika* 55, pp. 179–189. DOI: 10.1093/biomet/55.1.179.

Ivanyos, Gábor, Youming Qiao, and K. V. Subrahmanyam (2015). *Non-commutative Edmonds' problem and matrix semi-invariants*. arXiv:1508.00690v2 [cs.DS].

Jaynes, Edwin T. (1957). "Information theory and statistical mechanics". In: *Physical review* 106.4, p. 620. DOI: 10.1103/PhysRev.106.620.

Johnson, Charles R., Suzanne A. Lewis, and Donald Y. Yau (2001). "Possible line sums for a qualitative matrix". In: *Linear Algebra and its Applications* 327.1, pp. 53 –60. DOI: 10.1016/S0024-3795(00)00306-2.

Johnson, Charles R., Robert D. Masson, and Michael W. Trosset (2005). "On the diagonal scaling of Euclidean distance matrices to doubly stochastic matrices". In: *Linear Algebra and its Applications* 397, pp. 253 –264. DOI: 10.1016/j.laa.2004.10.023.

Johnson, Charles R. and Robert Reams (2009). "Scaling of symmetric matrices by positive diagonal congruence". In: *Linear and Multilinear Algebra* 57.2, pp. 123–140. DOI: 10.1080/03081080600872327.

Johnson, Charles R. and David P. Stanford (2000). "Patterns that allow given row and column sums". In: *Linear Algebra and its Applications* 311.1, pp. 97 –105. DOI: 10.1016/S0024-3795(00)00071-9.

Kalantari, Bahman (1990). "Canonical problems for quadratic programming and projective methods for their solution". In: *Proc. AMS Conference on Mathematical Problems Arising from Linear Programming, 1988, in Contemp. Math.* Pp. 243–263. ISBN: 1-58113-184-4.

— (1996). "A theorem of the alternative for multihomogeneous functions and its relationship to diagonal scaling of matrices". In: *Linear Algebra and its Applications* 236.0, pp. 1 –24. DOI: 10.1016/0024-3795(94)00162-6.

— (1998). *Scaling dualities and self-concordant homogeneous programming in finite dimensional spaces*. Tech. rep. DIMACS Technical Report 98-37. Department of Computer Science, Rutgers University.

— (1999). *Scaling dualities and self-concordant homogeneous programming in finite dimensional spaces*. Tech. rep. Technical Report LCSR-TR-359, Department of Computer Science. Department of Computer Science, Rutgers University.

— (2005). *Matrix scaling dualities in convex programming*. Tech. rep. Department of Computer Science, Rutgers University.

Kalantari, Bahman and M.R. Emamy-K (1997). "On linear programming and matrix scaling over the algebraic numbers". In: *Linear Algebra and its Applications* 262, pp. 283 –306. DOI: 10.1016/S0024-3795(97)80036-5.

Kalantari, Bahman and Leonid Khachiyan (1993). "On the rate of convergence of deterministic and randomized RAS matrix scaling algorithms". In: *Operations Research Letters* 14.5, pp. 237 –244. DOI: 10.1016/0167-6377(93)90087-W.

— (1996). "On the complexity of nonnegative-matrix scaling". In: *Linear Algebra and its Applications* 240.0, pp. 87 –103. DOI: 10.1016/0024-3795(94)00188-X.

Kalantari, Bahman et al. (2008). "On the complexity of general matrix scaling and entropy minimization via the RAS algorithm". In: *Mathematical Programming* 112.2, pp. 371–401. DOI: 10.1007/s10107-006-0021-4.

Karlin, S. and L. Nirenberg (1967). "On a Theorem of P. Nowosad". In: *Journal of Mathematical Analysis and Applications* 17.1, pp. 61 –67. DOI: 10.1016/0022-247X(67)90165-5.

Karzanov, Alexander V. and S. Thomas McCormick (1997). "Polynomial Methods for Separable Convex Optimization in Unimodular Linear Spaces with Applications". In: *SIAM Journal on Computing* 26.4, pp. 1245–1275. DOI: 10.1137/S0097539794263695.

Kent, Adrian, Noah Linden, and Serge Massar (1999). "Optimal Entanglement Enhancement for Mixed States". In: *Phys. Rev. Lett.* 83 (13), pp. 2656–2659. DOI: 10.1103/PhysRevLett.83.2656.

Khachiyan, Leonid (1996). "Diagonal matrix scaling is NP-hard". In: *Linear Algebra and its Applications* 234.0, pp. 173 –179. DOI: 10.1016/0024-3795(94)00099-9.

Khachiyan, Leonid and Bahman Kalantari (1992). "Diagonal Matrix Scaling and Linear Programming". In: *SIAM Journal on Optimization* 2.4, pp. 668–672. DOI: 10.1137/0802034.

Kleinberg, Jon M. (1999). "Hubs, Authorities, and Communities". In: *ACM Comput. Surv.* 31.4es. DOI: 10.1145/345966.345982.

Klose, Manfred, Alexander Opitz, and Norbert Schwarz (2004). "Sozialrechnungsmatrix für Deutschland". In: *Wirtschaft und Statistik*, pp. 605–620.

Knight, Philip A. (2008). "The Sinkhorn-Knopp Algorithm: Convergence and Applications". In: *SIAM J. Matrix Anal. Appl.* 30.1, pp. 261–275. DOI: 10.1137/060659624.

Knight, Philip A. and Daniel Ruiz (2012). "A fast algorithm for matrix balancing". In: *IMA Journal of Numerical Analysis*. DOI: 10.1093/imanum/drs019.

Knight, Philip A., Daniel Ruiz, and Bora Uçar (2014). "A Symmetry Preserving Algorithm for Matrix Scaling". In: *SIAM Journal on Matrix Analysis and Applications* 35.3, pp. 931–955. DOI: 10.1137/110825753.

Korzekwa, Kamil, David Jennings, and Terry Rudolph (2014). "Operational constraints on state-dependent formulations of quantum error-disturbance trade-off relations". In: *Phys. Rev. A* 89 (5), p. 052108. DOI: 10.1103/PhysRevA.89.052108.

Kruithof, R. (1937). "Telefoonverkeersrekening". In: *De Ingenieur* 52, E15–E25.

Krupp, R. S. (1979). "Properties of Kruithof's Projection Method". In: *The Bell System Technical Journal* 58.2, pp. 517–538. DOI: 10.1002/j.1538-7305.1979.tb02231.x.

Ku, Harry H. and Solomon Kullback (1968). "Interaction in multidimensional contingency tables: an information theoretic approach". In: *J. Res. Nat. Bur. Standards* 72, pp. 159–199.

Kullback, Solomon (1959). *Information Theory and Statistics*. John Wiley & Sons.

— (1968). "Probability Densities with Given Marginals". In: *The Annals of Mathematical Statistics* 39.4, pp. 1236–1243. DOI: 10.1214/aoms/1177698249.

Kullback, Solomon and M. A. Khairat (1966). "A Note on Minimum Discrimination Information". In: *Ann. Math. Statist.* 37.1, pp. 279–280. DOI: 10.1214/aoms/1177699619.

Kullback, Solomon and R. A. Leibler (1951). "On Information and Sufficiency". In: *Ann. Math. Statist.* 22.1, pp. 79–86. DOI: 10.1214/aoms/1177729694.

Lamond, B. and N. F. Stewart (1981). "Bregman's balancing method". In: *Transportation Research Part B: Methodological* 15.4, pp. 239 –248. DOI: 10.1016/0191-2615(81)90010-2.

Leinaas, Jon Magne, Jan Myrheim, and Eirik Ovrum (2006). "Geometrical aspects of entanglement". In: *Phys. Rev. A* 74 (1), p. 012313. DOI: 10.1103/PhysRevA.74.012313.

Lemmens, Bas and Roger Nussbaum (2012). *Nonlinear Perron-Frobenius Theory*. Cambridge University Press.

Letac, Gerard (1974). "A Unified Treatment of some Theorems on Positive Matrices". In: *Proc. Amer. Math. Soc.* 43.1, pp. 11–17. DOI: 10.1090/S0002-9939-1974-0338037-8.

Lewis, P. M. (1959). "Approximating probability distributions to reduce storage requirements". In: *Information and Control* 2.3, pp. 214 –225. DOI: 10.1016/S0019-9958(59)90207-4.

Linial, Nathan, Alex Samorodnitsky, and Avi Wigderson (2000). "A Deterministic Strongly Polynomial Algorithm for Matrix Scaling and Approximate Permanents". In: *Combinatorica* 20.4, pp. 545–568. DOI: 10.1007/s004930070007.

Lisi, Sam (2011). *Given two basis sets for a finite Hilbert space, does an unbiased vector exist?* Mathematics Stack Exchange. (version: 2011-04-05). URL: http://math.stackexchange.com/q/29819.

littleO, (http://math.stackexchange.com/users/40119/littleo) (2014). *Please explain the intuition behind the dual problem in optimization.* Mathematics Stack Exchange. URL: http://math.stackexchange.com/q/624633.

Livne, Oren E. and Gene H. Golub (2004). "Scaling by Binormalization". In: *Numerical Algorithms* 35.1, pp. 97–120. DOI: 10.1023/B:NUMA.0000016606.32820.69.

London, David (1971). "On matrices with a doubly stochastic pattern". In: *Journal of Mathematical Analysis and Applications* 34.3, pp. 648–652. DOI: 10.1016/0022-247X(71)90104-1.

Lovász, László and M. D. Plummer (2009). *Matching Theory*. AMS Chelsea Publishing Series. AMS Chelsea Pub. ISBN: 9780821847596.

Luo, Z. Q. and P. Tseng (1992). "On the convergence of the coordinate descent method for convex differentiable minimization". In: *Journal of Optimization Theory and Applications* 72.1, pp. 7–35. DOI: 10.1007/BF00939948.

Macgill, Sally M. (1977). "Theoretical properties of biproportional matrix adjustments". In: *Environment and Planning A* 9.6, pp. 687–701. DOI: 10.1068/a090687.

Maier, Sebastian, Petur Zachariassen, and Martin Zachariasen (2010). "Divisor-based biproportional apportionment in electoral systems: A real-life benchmark study". In: *Management Science* 56.2, pp. 373–387. DOI: 10.1287/mnsc.1090.1118.

Marcus, M. and M. Newman (1961). "The permanent of a symmetric matrix". In: *Notices of the A.M.S.* 8.

Marshall, Albert W. and Ingram Olkin (1968). "Scaling of matrices to achieve specified row and column sums". In: *Numerische Mathematik* 12.1, pp. 83–90. DOI: 10.1007/BF02170999.

Maxfield, J. and H. Minc (1962). "A doubly stochastic matrix equivalent to a given matrix". In: *Notices of the A.M.S.* 9.

McDougall, Robert (1999). *Entropy Theory and the RAS are friends*. GTAP Working Paper No. 06 (300).

Menon, M. V. (1967). "Reduction of a Matrix with Positive Elements to a Doubly Stochastic Matrix". In: *Proc. Amer. Math. Soc.* 18.2, pp. 244–247. DOI: 10.1090/S0002-9939-1967-0215873-6.

— (1968). "Matrix links, an extremization problem, and the reduction of a non-negative matrix to one with prescribed row and column sums". In: *Canad. J. Math.* 20, pp. 225–232. DOI: 10.4153/CJM-1968-021-9.

Menon, M. V. and Hans Schneider (1969). "The spectrum of a nonlinear operator associated with a matrix". In: *Linear Algebra and its Applications* 2.3, pp. 321 –334. DOI: 10.1016/0024-3795(69)90034-2.

Moon, Todd K., Jacob H. Gunther, and Joseph J. Kupin (2009). "Sinkhorn solves sudoku". In: *Information Theory, IEEE Transactions on* 55.4, pp. 1741–1746. DOI: 10.1109/TIT.2009.2013004.

Mosteller, Frederick (1968). "Association and Estimation in Contingency Tables". In: *Journal of the American Statistical Association* 63.321, pp. 1–28. DOI: 10.1080/01621459.1968.11009219.

Murchland, J. D. (1977). "The multiproportional problem". Manuscript JDM-263, draft 1, University College London Transport Studies Group.

— (1978). "Applications, history and properties of bi- and multi-proportional models". Unpublished Seminar at London School of Economics, London.

Murty, Katta G. and Santosh N. Kabadi (1987). "Some NP-complete problems in quadratic and nonlinear programming". In: *Mathematical Programming* 39.2, pp. 117–129. DOI: 10.1007/BF02592948.

Nemirovski, Arkadi and Uriel Rothblum (1999). "On complexity of matrix scaling". In: *Linear Algebra and its Applications* 302-303.0, pp. 435 –460. DOI: 10.1016/S0024-3795(99)00212-8.

Netanyahu, E. and M. Reichaw (1969). "A Theorem on Infinite Positive Matrices". In: *Proceedings of the American Mathematical Society* 20.1, pp. 13–15. DOI: 10.1090/S0002-9939-1969-0236203-1.

Nielsen, Michael and Isaac Chuang (2000). *Quantum Computation and Quantum Information*. Camebridge University Press.

Niemeyer, Horst F. and Alice C. Niemeyer (2008). "Apportionment methods". In: *Mathematical Social Sciences* 56.2, pp. 240 –253. DOI: 10.1016/j.mathsocsci.2008.03.003.

Nowosad, P. (1966). "On the integral equation $Kf = 1/f$ arising in a problem in communication". In: *Journal of Mathematical Analysis and Applications* 14.3, pp. 484 –492. DOI: 10.1016/0022-247X(66)90008-4.

Nussbaum, Roger D. (1987). "Iterated Nonlinear Maps and Hilbert's Projective Metric: A Summary". In: *Dynamics of Infinite Dimensional Systems*. Ed. by Shui-Nee Chow and Jack K. Hale. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 231–248. ISBN: 978-3-642-86458-2. DOI: 10.1007/978-3-642-86458-2_23.

— (1993). "Entropy Minimization, Hilbert's Projective Metric, and Scaling Integral Kernels". In: *Journal of Functional Analysis* 115.1, pp. 45 –99. DOI: 10.1006/jfan.1993.1080.

O'Leary, Dianne P. (2003). "Scaling symmetric positive definite matrices to prescribed row sums". In: *Linear Algebra and its Applications* 370, pp. 185 –191. DOI: 10.1016/S0024-3795(03)00387-2.

Olschowka, Markus and Arnold Neumaier (1996). "A new pivoting strategy for Gaussian elimination". In: *Linear Algebra and its Applications* 240, pp. 131 –151. DOI: 10.1016/0024-3795(94)00192-8.

Ortúzar, Juan de Dios and Luis G. Willumsen (2011). *Modelling Transport*. 4th ed. Wiley. ISBN: 978-0-470-76039-0.

Osborne, E. E. (1960). "On Pre-Conditioning of Matrices". In: *J. ACM* 7.4, pp. 338–345. DOI: 10.1145/321043.321048.

Panov, A. (1985). "On mixed discriminants connected with positive semidefinite quadratic forms". In: *Soviet Mathematis - Doklady* 31.

Parlett, B. N. and T. L. Landis (1982). "Methods for scaling to doubly stochastic form". In: *Linear Algebra and its Applications* 48.0, pp. 53 –79. DOI: 10.1016/0024-3795(82)90099-4.

Pereira, Rajesh (2003). "Differentiators and the geometry of polynomials". In: *Journal of Mathematical Analysis and Applications* 285.1, pp. 336 –348. DOI: 10.1016/S0022-247X(03)00465-7.

Pereira, Rajesh and Joanna Boneng (2014). "The theory and applications of complex matrix scalings". In: *Special Matrices* 2.1. DOI: 10.2478/spma-2014-0007.

Perron, Oskar (1907). "Zur Theorie der Matrices". In: *Mathematische Annalen* 64.2, pp. 248–263. DOI: 10.1007/BF01449896.

Plane, David A. (1982). "An information theoretic approach to estimation of migration flows". In: *Journal of Regional Science* 22.4, pp. 441–456. DOI: 10.1111/j.1467-9787.1982.tb00769.x.

Pretzel, Oliver (1980). "Convergence of the iterative scaling procedure for non-negative matrices". In: *J. London Math. Soc.* 21.2, pp. 379,384. DOI: 10.1112/jlms/s2-21.2.379.

Pukelsheim, F. and C. Schuhmacher (2004). "Das neue Zürcher Zuteilungsverfahren für Parlamentswahlen". In: *Aktuelle Juristische Praxis, Pratique Juridique Actuelle* 13, pp. 505–522.

Pukelsheim, Friedrich and Bruno Simeone (2009). *On the iterative proportional fitting procedure: Structure of accumulation points and L1-error analysis*. Preprint. URL: http://www.dss.uniroma1.it/sites/default/files/vecchie-pubblicazioni/RT_7_2009_Pukelsheim.pdf.

Pyatt, Graham and Jeffery I. Round (1985). *Social accounting matrices : a basis for planning*. Washington DC. The World Bank. URL: http://documents.worldbank.org/curated/en/1985/09/439689/social-accounting-matrices-basis-planning.

Pyatt, Graham and E. Thorbecke (1976). *Planning Techniques for a better Future (A WEP study)*. International Labour Office. ISBN: 978-9221015529.

Raghavan, T. E. S. (1984). "On pairs of multidimensional matrices". In: *Linear Algebra and its Applications* 62.0, pp. 263 –268. DOI: 10.1016/0024-3795(84)90101-0.

Robillard, Pierre and Neil F. Stewart (1974). "Iterative numerical methods for trip distribution problems". In: *Transportation Research* 8.6, pp. 575–582. DOI: 10.1016/0041-1647(74)90034-3.

Rockafellar, R. T. (1997). *Convex Analysis*. Convex Analysis. Princeton University Press. ISBN: 9780691015866.

Rote, Günter and Martin Zachariasen (2007). "Matrix scaling by network flow". In: *Symposium on Discrete Algorithms: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Vol. 7. 09, pp. 848–854. ISBN: 9780898716245.

Rothblum, Uriel G. (1989). "Generalized scalings satisfying linear equations". In: *Linear Algebra and its Applications* 114, pp. 765 –783. DOI: 10.1016/0024-3795(89)90492-8.

Rothblum, Uriel G. and Hans Schneider (1980). "Characterizations of optimal scalings of matrices". In: *Mathematical Programming* 19.1, pp. 121–136. DOI: 10.1007/BF01581636.

— (1989). "Scalings of matrices which have prespecified row sums and column sums via optimization". In: *Linear Algebra and its Applications* 114 - 115.0. Special Issue Dedicated to Alan J. Hoffman, pp. 737 –764. DOI: 10.1016/0024-3795(89)90491-6.

Rothblum, Uriel G., Hans Schneider, and Michael H. Schneider (1994). "Scaling Matrices to Prescribed Row and Column Maxima". In: *SIAM Journal on Matrix Analysis and Applications* 15.1, pp. 1–14. DOI: 10.1137/S0895479891222088.

Rothblum, Uriel G. and Stavros A. Zenios (1992). "Scalings of matrices satisfying line-product constraints and generalizations". In: *Linear Algebra and its Applications* 175, pp. 159 –175. DOI: 10.1016/0024-3795(92)90307-V.

Ruiz, Daniel (2001). *A scaling algorithm to equilibrate both rows and columns norms in matrices*. Tech. rep. RAL-TR-2001-034. Computational Science and Engineering Department, Atlas Centre, Rutherford Appleton Laboratory.

Ruschendorf, Ludger (1995). "Convergence of the iterative proportional fitting procedure". In: *The Annals of Statistics*, pp. 1160–1174.

Samelson, Hans (1957). "On the Perron-Frobenius theorem". In: *Michigan Math. J.* 4.1, pp. 57–59. DOI: 10.1307/mmj/1028990177.

Sanz, Mikel et al. (2010). "A Quantum Version of Wielandt's inequality". In: *IEEE Transactions on Information Theory* 56 (9). DOI: 10.1109/TIT.2010.2054552.

Schneider, H. and B. D. Saunders (1980). "Applications of the Gordan-Stiemke Theorem in Combinatorial Matrix Theory". In: *Combinatorics 79*. Ed. by Peter L. Hammer. Vol. 9. Annals of Discrete Mathematics. Elsevier, pp. 247 –. DOI: 10.1016/S0167-5060(08)70073-6.

Schneider, Hans (1977). "The concepts of irreducibility and full indecomposability of a matrix in the works of Frobenius, König and Markov". In: *Linear Algebra and its Applications* 18.2, pp. 139–162. DOI: 10.1016/0024-3795(77)90070-2.

Schneider, Michael H. (1989). "Matrix scaling, entropy minimization, and conjugate duality. I. existence conditions". In: *Linear Algebra and its Applications* 114-115.0. Special Issue Dedicated to Alan J. Hoffman, pp. 785–813. DOI: 10.1016/0024-3795(89)90493-X.

Schneider, Michael H. (1990). "Matrix scaling, entropy minimization, and conjugate duality (II): The dual problem". In: *Mathematical Programming* 48.1-3, pp. 103–124. DOI: 10.1007/BF01582253.

Schneider, Michael H. and Stavros A. Zenios (1990). "A comparative study of algorithms for matrix balancing". In: *Operations Research* 38.3, pp. 439–455. DOI: 10.1287/opre.38.3.439.

Schrader, R. (2000). "Perron-Frobenius Theory for positive maps on trace ideals". In: *Mathematical Physics in Mathematics and Physics: Quantum and Operator Algebraic Aspects*. Ed. by Robert Longo. American Mathematical Society. ISBN: 978-0-8218-2814-4.

Schrödinger, Erwin (1931). Sonderausgabe a. d. Sitz.-Ber. d. Preuß. Akad. d. Wiss., Phys.-math. Klasse. Verlag W. de Gruyter, Berlin.

Seneta, E. (2006). *Nonnegative Matrices and Markov Chains*. Springer. ISBN: 978-0-387-32792-1.

Sinkhorn, Richard (1964). "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices". In: *Annals of Mathematical Statistics* 35.2, pp. 876–879. DOI: 10.1214/aoms/1177703591.

— (1966). "A relationship between arbitrary positive matrices and stochastic matrices". In: *Canad. J. Math.* 18, pp. 303–306. DOI: 10.4153/CJM-1966-033-9.

— (1967). "Diagonal Equivalence to Matrices with Prescribed Row and Column Sums". In: *The American Mathematical Monthly* 74.4, pp. 402–405. DOI: 10.2307/2314570.

— (1972). "Continuous Dependence of A in the D_1AD_2 Theorems". In: *Proceedings of the American Mathematical Society* 32.2, pp. 395–398. DOI: 10.1090/S0002-9939-1972-0297792-4.

— (1974). "Diagonal Equivalence to Matrices with Prescribed Row and Column Sums. II". In: *Proceedings of the American Mathematical Society* 45.2, pp. 195–198. DOI: 10.2307/2040061.

Sinkhorn, Richard and Paul Knopp (1967). "Concerning Nonnegative Matrices and Doubly Stochastic Matrices". In: *Pacific Journal of Mathematics* 21.2, pp. 343–348. DOI: 10.2140/pjm.1967.21.343.

Smith, John H. (1947). "Estimation of Linear Functions of Cell Proportions". In: *The Annals of Mathematical Statistics* 18.2, pp. 231–254. DOI: 10.1214/aoms/1177730440.

Soules, George W. (1991). "The rate of convergence of Sinkhorn balancing". In: *Linear Algebra and its Applications* 150.0, pp. 3 –40. DOI: 10.1016/0024-3795(91)90157-R.

Stephan, Frederick F. (1942). "An Iterative Method of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known". In: *The Annals of Mathematical Statistics* 13.2, pp. 166–178. DOI: doi:10.1214/aoms/1177731604.

Stone, R. (1962). "Multiple classifications in social accounting". In: *Bulletin de l'institut International de Statistique* 39.3, pp. 215–33.

Theil, H. (1967). *Economics and information theory*. Studies in mathematical and managerial economics. North-Holland Pub. Co.

Thionet, Pierre (1961). "Sur le remplissage d'un tableau à double entrée". In: *Journal de la société française de statistique* 102, pp. 331–345.

— (1963). "Sur certaines variantes des projections du tableau d'échanges inter- industriels". In: *Bull. Inst. Int. Stat.* 40, pp. 119–132. ISSN: 0373-0441.

— (1964). "Note sur le remplissage d'un tableau à double entrée". In: *Journal de la société française de statistique* 105, pp. 228–247.

Tverberg, Helge (1976). "On Sinkhorn's representation of nonnegative matrices". In: *Journal of Mathematical Analysis and Applications* 54.3, pp. 674 –677. DOI: 10.1016/0022-247X(76)90186-4.

Uribe, Pedro, C. G. de Leeuw, and H. Theil (1966). "The Information Approach to the Prediction of Interregional Trade Flows". In: *The Review of Economic Studies* 33.3, pp. 209–220. ISSN: 00346527, 1467937X.

Verstraete, Frank, Jeroen Dehaene, and Bart De Moor (2001). "Local filtering operations on two qubits". In: *Phys. Rev. A* 64 (1), p. 010101. DOI: 10.1103/PhysRevA.64.010101.

— (2002). "Lorentz singular-value decomposition and its applications to pure states of three qubits". In: *Phys. Rev. A* 65 (3), p. 032308. DOI: 10.1103/PhysRevA.65.032308.

— (2003). "Normal forms and entanglement measures for multipartite quantum states". In: *Phys. Rev. A* 68 (1), p. 012103. DOI: 10.1103/PhysRevA.68.012103.

Visick, George et al. (1980). "A modification to Kruithof's double-factor method". In: *Transportation Research Part B: Methodological* 14.4, pp. 307–318. DOI: 10.1016/0191-2615(80)90011-9.

Wang, Fei, Ping Li, and Arnd Christian König (2010). "Learning a Bi-Stochastic Data Similarity Matrix". In: *2010 IEEE International Conference on Data Mining*, pp. 551–560. DOI: 10.1109/ICDM.2010.141.

Welch, Lloyd (unknown). "Unpublished Report of the Institute of Defense Analysis". Princeton, New Jersey.

Wolf, Michael M. (2012). "Quantum Channels and Operations, Guided Tour". lecture notes.

Wright, Stephen J. (2015). "Coordinate descent algorithms". In: *Mathematical Programming* 151.1, pp. 3–34. DOI: 10.1007/s10107-015-0892-3.

Zenios, Stavros A. (1990). "Matrix Balancing on a Massively Parallel Connection Machine". In: *ORSA Journal on Computing* 2.2, pp. 112–125. DOI: 10.1287/ijoc.2.2.112.

Zenios, Stavros A. and Siu-Leong Iu (1990). "Vector and parallel computing for matrix balancing". In: *Annals of Operations Research* 22.1, pp. 161–180. DOI: 10.1007/BF02023052.

A. Preliminaries on matrices

Sinkhorn's theorem is closely related to irreducibility and notions connected to it.

Therefore, we first recall the following charactersation of irreducible matrices.

Proposition A.1. *Let $A \in \mathbb{R}^{n\times n}$ be nonnegative. The following are equivalent:*

1. *A is irreducible.*
2. *The digraph associated to A is strongly connected.*
3. *For each i and j there exists a k such that $(A^k)_{ij} > 0$.*
4. *For any partition $I \cup J$ of $\{1, \dots, n\}$, there exists a $j \in J$ and an $i \in I$ such that $A_{ij} \neq 0$.*

An overview about these and similar properties can be found in Schneider 1977. Graph related properties are proven in Brualdi 1968.

Let us now describe a graph for any matrix: Given a nonnegative matrix $A \in \mathbb{R}^{n\times n}$ and any partition $I \cup J = \{1, \dots, 2n\}$, let $I \cup J = V$ and $E = \{(i,j) | A_{ij} > 0\} \subset I \times J$ be the vertices and edges of the (bipartite) Graph $G_A := (V, E)$.

Definition A.2. A bipartite graph $G = (V, E)$ has a *perfect matching*, if it contains a subgraph where the degree of any vertex is exactly one, i.e. any vertex is matched with exactly one other vertex.

Note that this definition is not dependent on the size of the entries of A, it only depends on whether an entry is positive or zero.

Proposition A.3 (Brualdi, Parter, and Schneider 1966 Lemma 2.3)**.** *Let $A \in \mathbb{R}^{m \times n}$ be nonnegative. The following are equivalent:*

1. *A is fully indecomposable,*
2. *PAQ is fully indecomposable for all permutations P, Q,*
3. *There exist permutations P, Q such that PAQ is irreducible and has a positive main diagonal.*
4. *For any $(i, j) \in E$ the edge set of the bipartite graph G_A for A, there exists a perfect matching in G_A containing this edge.*

Sketch of proof. The equivalence $(1) \Leftrightarrow (2)$ is obvious and $(1) \Leftrightarrow (3)$ is done in Brualdi, Parter, and Schneider 1966. The direction $\Rightarrow$ follows from the Frobenius-König theorem (cf. Bhatia 1996, Chapter 2). The converse direction follows from a short contradiction proof.

Finally, $(1) \Leftrightarrow (4)$ follows essentially from Theorem 4.1.1 in Lovász and Plummer 2009, which was first observed in Hetyei 1964[21]. □

Since multiplication of positive diagonal matrices from the right and from the left does not change the pattern of a matrix, having a matrix that has the required row and column sums is an easy necessary condition for scalability.

Let us now consider the special case of doubly stochastic matrices in more detail. We define:

Definition A.4. Let $A \in \mathbb{R}^{n \times n}$ be a nonnegative matrix. Then A has *total support* if it is nonzero and for every $A_{ij} > 0$ there exists a permutation σ such that $\sigma(i) = j$ and $\prod_{k=1}^{n} A_{\sigma(k)k} \neq 0$. In other words, A has total support if any nonzero element lies on a positive diagonal (Sinkhorn 1967).

Furthermore, A has *support*, if there exists a positive diagonal, i.e. there exists an A_{ij} such that for some permutation σ with $\sigma(i) = j$ we have $\prod_{k=1}^{n} A_{\sigma(k)k} \neq 0$.

Proposition A.5. *Let $A \in \mathbb{R}^{n \times n}$ be a nonnegative matrix. The following are equivalent:*

1. *After independent permutations of rows and columns, A is a direct sum of fully indecomposable matrices.*

[21] In Hungarian. Reference taken from Lovász and Plummer 2009

2. *A has a doubly-stochastic pattern.*

3. *A has total support*

Furthermore, A has support if and only if there exists a matrix B with a subpattern of A with total support.

Sketch of Proof. For $1 \Leftrightarrow 2$ we follow the proof in Brualdi, Parter, and Schneider 1966.

Let A be doubly stochastic. Since we can permute rows and columns independently, we can assume that A is of the form

$$A := \begin{pmatrix} A_1 & 0 & \dots & 0 \\ A_{21} & A_2 & \dots & 0 \\ \vdots & & & \vdots \\ A_{k1} & A_{k2} & \dots & A_k \end{pmatrix}$$

for some $k \in \mathbb{N}$. All A_i are either 1×1 zero-matrices or fully indecomposable (otherwise iterate). Since A is doubly stochastic one can quickly see that $A_{ij} = 0$ for all $i < j$. Furthermore, no A_i can be zero, because this would then result in a zero-row. Hence, A can be decomposed as a direct sum of fully indecomposable maps.

For $2 \Leftrightarrow 3$ see Sinkhorn 1967.

Finally, consider a matrix B with total support. Clearly, if it is a submatrix of some other matrix A, then A will have support, since any element $A_{ij} > 0$ which is contained in B will lie on a nonzero diagonal. Conversely, if A has support, setting any element A_{ij} which does not lie on a positive diagonal to zero produces a matrix that has total support. □

B. Introduction to Nonlinear Perron-Frobenius theory

The basic result underlying Perron-Frobenius theory is an old theorem from Perron 1907 and Frobenius 1912 stating:

Theorem B.1. *Let $A \in \mathbb{R}^{n \times n}$ be a nonnegative, irreducible matrix with spectral radius $\rho(A)$. Then $\rho(A) > 0$ is a nondegenerate positive eigenvalue of A with a one-dimensional eigenspace consisting of a vector x with only positive components.*

The theorem was later interpreted geometrically in Birkhoff 1957; Samelson 1957, where the authors noted that it follows using Hilbert's projective metric and contraction principles. From then on, it was slowly extended to (not necessarily linear) operators, which lies at the heart of nonlinear Perron-Frobenius theory. The connection to matrix scaling became clear int the 60s to 80s and parts of each theory have developed alongside each other since. Probably the best current reference on the topic is Lemmens and Nussbaum 2012. In the following, we sketch some of the most important ideas surrounding the theory:

Recall that given a topological vector space $\mathcal{V}$, a *cone* is a set $\mathcal{C} \subset \mathcal{V}$ such that for all $v \in \mathcal{C}$, $\alpha v \in \mathcal{C}$ for all $\alpha > 0$. A *convex cone* is a cone that contains all convex combinations. By definition, it is equivalent to say that $\mathcal{C}$ is a convex cone if and only if $\alpha v + \beta w \in \mathcal{C}$ for all $v, w \in \mathcal{C}$ and all $\alpha, \beta > 0$. A cone is called *solid*, if it contains an interior point in the topology of the vector space and *closed* if it is closed in the given topology. It is called *polyhedral*, if it is the intersection of finitely many closed half-spaces (cf. Rockafellar 1997).

An easy way to construct a convex, solid cone is by using a partial order $\geq$. Then the set $\mathcal{C}$ defined via $v \in \mathcal{C} \Leftrightarrow v \geq 0$ is a closed convex cone (see Rockafellar 1997 for the connection between ordered vector spaces and convex cones). Given two cones $\mathcal{C}, \mathcal{K}$ which are defined by a partial order, we call a map $\mathbf{T} : \mathcal{C} \to \mathcal{K}$ *order-preserving* or *monotonic*, if for $v \geq w$ we have $\mathbf{T}(v) \geq \mathbf{T}(w)$. We call it *strongly order-preserving*, if for every $v \geq w \in \mathcal{V}$ we have $\mathbf{T}(v) - \mathbf{T}(w) \in \mathcal{V}' > 0$. Last but not least, we call $\mathbf{T}$ *homogeneous*, if $\mathbf{T}(\alpha v) = \alpha \mathbf{T}(v)$ for all $\alpha \in \mathbb{R}$.

As is the case with classical Perron-Frobenius theory, the spectral radius is the crucial notion. For general maps between cones, there are several definitions, which turn out to be the same for most purposes, hence we restrict to one such notion:

Definition B.2 (Lemmens and Nussbaum 2012, Chapter 5.2)**.** Let $\mathcal{K}$ be a solid closed cone in a finite dimensional vector space with a fixed norm $\|\cdot\|$ and $f : \mathcal{K} \to \mathcal{K}$ a continuous homogeneous map. Define the *cone spectral radius* as

$$r_{\mathcal{K}}(f) := \sup\{\limsup_{m\to\infty} \|f^m(x)\|^{1/m} | 0 \neq x \in \mathcal{K}\}$$

The first crucial observation is that the spectral radius is actually attained for homogeneous order-preserving maps (which is not the case for general maps):

Theorem B.3 (Lemmens and Nussbaum 2012, Cor. 5.4.2)**.** *Let $\mathcal{K}$ be a solid closed cone in a finite dimensional vector space $\mathcal{V}$. If $f : \mathcal{K} \to \mathcal{K}$ is a continuous, homogeneous, order-preserving map, then there exists $x \in \mathcal{K} \setminus \{0\}$ with $f(x) = r_{\mathcal{K}}(f)x$*

Note that the theorem does not tell us whether the eigenvector lies inside the cone or on its boundary. The next and very powerful theorem does also not settle existence, but if existence is known, then it assures uniqueness and convergence:

Theorem B.4 (Lemmens and Nussbaum 2012, Thm. 6.5.1)**.** *Let $\mathcal{K}$ be a solid closed cone in a finite dimensional vector space $\mathcal{V}$ and let $\varphi \in \mathcal{K}^*$ the dual cone. If $f : \mathrm{int}(\mathcal{K}) \to \mathrm{int}(\mathcal{K})$ is a homogeneous strongly order-preserving map and there exists a $u \in \mathrm{int}(\mathcal{K})$ with $\varphi(u) = 1$ such that $f(u) = ru$, then*

$$\lim_{k\to\infty} \frac{f^k(x)}{\varphi(f^k(x))} = u \tag{66}$$

for all $x \in \mathrm{int}(\mathcal{K})$.

This theorem is essentially due to the fact that the map is contractive in what is known as Hilbert's projective metric of the cones. Once the attractiveness is established, uniqueness follows immediately, since if we had another fixed point $v \in \text{int}(\mathcal{K})$, then $f^k(v) = v$ would imply a contradiction to equation 66. Since the attractiveness can be used to estimate convergence speeds (see Franklin and Lorenz 1989), we include the relevant proposition due to Birkhoff (Birkhoff 1957).

Definition B.5 (Lemmens and Nussbaum 2012, Chapter 2.1)**.** Let $\mathcal{K} \subset V$ be a closed, convex, solid cone in some vector space, then for any $x, y \in \mathcal{K}$ such that $x \leq \alpha y$ and $y \leq \beta x$ for some $\alpha, \beta > 0$, define

$$M(x/y; \mathcal{K}) := \inf\{\beta > 0 | x \leq \beta y\}$$
$$m(x/y; \mathcal{K}) := \sup\{\alpha > 0 | \alpha y \leq x\}$$

Then we can define *Hilbert's projective metric* as

$$d_H(x, y, \mathcal{K}) := \ln\left(\frac{M(x/y)}{m(x/y)}\right) \tag{67}$$

We will leave out $\mathcal{K}$, when it is clear from the context. Furthermore, we set $d_H(0,0) = 0$ and $d_H(x,y) = \infty$ if d_H is otherwise not well-defined.

We have the following properties:

Proposition B.6 (Lemmens and Nussbaum 2012 Proposition 2.1.1)**.** *Let $\mathcal{K} \subset V$ be a closed, convex, solid cone in some vector space V. Then d_H satisfies:*

(i) $d_H(x,y) \geq 0$ *and* $d_H(x,y) = d_H(y,x)$ *for all* $x, y \in \mathcal{K}$.

(ii) $d_H(x,z) \leq d_H(x,y) + d_H(y,z)$ *for all* $x, y, z \in \mathcal{K}$ *such that the quantities are well-defined.*

(iii) $d_H(\alpha x, \beta y) = d_H(x,y)$ *for all* $x, y \in \mathcal{K}$ *and* $\alpha, \beta > 0$.

Note that the first two properties show that d_H is indeed a metric and the third property shows why it is called a *projective metric*.

Proposition B.7 (Birkhoff 1957)**.** *Let $\mathcal{K} \subset V$ be a bounded, closed, convex and solid cone in some vector space V. Let $\mathcal{E} : \mathcal{K} \to \mathcal{K}$ be a linear map, then*

$$\gamma^{1/2}(\mathcal{E}) := \sup\left\{\frac{d_H(\mathcal{E}(x), \mathcal{E}(y))}{d_H(x,y)} \middle| x, y \in \mathcal{K}\right\} \leq \tanh(\Delta/4) \tag{68}$$

where $\Delta := \max\{d_H(\mathcal{E}(x), \mathcal{E}(y)) | x, y \in \mathcal{K}\}$.

Furthermore, $\gamma^{1/2}(\mathcal{E} \circ \mathcal{F}) \leq \gamma^2(\mathcal{E})\gamma^2(\mathcal{F})$ *and* $\gamma^{1/2}(\mathcal{E}^*) = \gamma^2(\mathcal{E})$.

A proof can be found in Birkhoff 1957; Bauer 1965. The result can be extended to much more general scenarios, see also Eveson and Nussbaum 1995 and references therein. One can actually show that equality holds, i.e. $\tanh(\Delta/4)$ is also attained, but this is not important here.

With all this machinery, we still need to prove existence of a fixed point in the interior of $\mathcal{P}$. The general theory for proving that a fixed point lies in the interior is weak and we generally have to prove it "by hand".

For, the Menon operator we have an additional problem: It is at first only well-defined on the interior of the cone of positive semidefinite matrices. A natural question is whether it can be extended to cover the closed cone as well. For matrices, this is covered by the following theorem:

Theorem B.8 (Lemmens and Nussbaum 2012, theorem 5.1.5)**.** *Let $\mathcal{C}, \mathcal{K}$ be cones and $\mathbf{S} : \mathcal{C} \to \mathcal{K}$ be an order-preserving, homogeneous map. If $\mathcal{C}$ is solid and polyhedral, then there exists a continuous, order-preserving, homogeneous extension $\overline{\mathbf{S}} : \overline{\mathcal{C}} \to \overline{\mathcal{K}}$.*

C. Preliminaries on Positive Maps

In this section, let $\mathcal{C}^n \subset \mathcal{M}_n$ be the cone of positive definite matrices (elements will also be written as $A > 0$) with its closure $\overline{\mathcal{C}^n}$, the cone of positive semidefinite matrices (elements will also be written as $A \geq 0$). Likewise, a subscript $_1$ at any of the cones denotes the bounded subset of unit trace matrices. Positive maps on cones are elements form a cone themselves, the dual cone, which will be denoted by $(\mathcal{C}^n)^*$ and $(\overline{\mathcal{C}^n})^*$.

Let us start with irreducible maps. Many different characterizations exist, which we recall for the reader's convenience:

Proposition C.1. *For a positive, linear map $T : \mathcal{M}_d \to \mathcal{M}_d$ the following properties are equivalent:*

1. *T is irreducible,*
2. *if $P \in \mathcal{M}_d$ is a Hermitian projector such that $T(P\mathcal{M}_d P) \subset P\mathcal{M}_d P$ then $P \in \{0, \mathbb{1}\}$,*
3. *for every nonzero $A \geq 0$ we have $(\mathrm{id} + T)^{d-1}(A) > 0$,*
4. *for every nonzero $A \geq 0$ and every strictly positive $t \in \mathbb{R}$ we have $\exp(tT)(A) > 0$,*
5. *There does not exist a nontrivial orthogonal projection P s.th. $\mathrm{tr}(T(P)(\mathbb{1} - P)) = 0$.*

Most of these properties are well-known. A proof can be found in Wolf 2012.

As with matrices in the original Sinkhorn theorem, irreducibility is not the right characterization to work with, since given an irreducible map $\mathcal{E}$ and two $X, Y > 0$, $Y\mathcal{E}(X.X^\dagger)Y^\dagger$ is not necessarily irreducible. Before giving a characterization of fully indecomposable maps, we will study rank non-decreasing maps.

Definition C.2 (Gurvits 2004)**.** To every positive map $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ and any unitary $U \in U(n)$, we associate the *decoherence operator* $\mathcal{E}_U$ via:

$$\mathcal{E}_U(X) := \sum_i \mathcal{E}(u_i u_i^\dagger) \operatorname{tr}(X u_i u_i^\dagger) \tag{69}$$

where u_i is the i-th row of U. Furthermore, we associate to every decoherence operator the tuple

$$\mathbf{A}_{\mathcal{E},U} := (\mathcal{E}(u_1 u_1^\dagger), \ldots, \mathcal{E}(u_n u_n^\dagger)) \tag{70}$$

This will be important during the proof of the Sinkhorn scaling, because every map $\mathcal{E}$ will be associated to the mixed discriminants of its decoherence operators:

Definition C.3. Let $(A_i)_i$ be an n-tuple with $A_i \in \mathcal{M}_n$, then

$$M(A_1, \ldots, A_n) := \frac{\partial^n}{\partial x_1 \ldots \partial x_n} \det(x_1 A_1 + \ldots + x_n A_n)|_{x_1,\ldots,x_n=0} \tag{71}$$

is called the *mixed discriminant*.

Then we have the following characterization of rank non-decreasing maps, which is essentially due to Gurvits 2004:

Proposition C.4. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. Then the following expressions are equivalent:*

(i) *$\mathcal{E}$ is rank non-decreasing.*

(ii) *$\mathcal{E}_U$ is rank non-decreasing for any unitary $U \in U(n)$.*

(iii) *For any $U \in U(n)$, if $(A_i)_i := \mathbf{A}_{\mathcal{E},U}$, then*

$$\operatorname{rank}\left(\sum_{i \in \mathcal{S}} A_i\right) \geq |\mathcal{S}| \qquad \forall \mathcal{S} \subseteq \{1, \ldots, n\}$$

(iv) *For any $U \in U(n)$, $M(\mathbf{A}_{\mathcal{E},U}) > 0$.*

(v) *$\mathcal{E}'(\cdot) := Y^\dagger \mathcal{E}(X \cdot X^\dagger) Y$ is rank non-decreasing for any X, Y of full rank.*

The proofs that (i), (ii), (iii) and (v) are equivalent are essentially the same as for the fully indecomposable case in C.6. It remains to show the equivalence of (v) with (i). This was done in Panov 1985.

We can define what will turn out as a measure of being indecomposable for a tuple of matrices:

Definition C.5. Let $A := (A_i)_i$ be an n-tuple of matrices $A_i \in \mathcal{M}_n$ and denote by A^{ij} the tuple where A_i is substituted by A_j. Then define:

$$\overline{M}(A) := \min_{i \neq j} M(A^{ij}) \tag{72}$$

the minimal mixed discriminant.

For fully decomposable maps, we have the following characterization (part of which is already present in Gurvits 2004):

Proposition C.6. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. Then the following expressions are equivalent:*

(i) $\mathcal{E}$ *is fully indecomposable*

(ii) $\mathcal{E}^*$ *is fully indecomposable*

(iii) For all singular, but nonzero $A \geq 0$, $\operatorname{rank}(\mathcal{E}(A)) > \operatorname{rank} A$.

(iv) Property (iii) holds for $Y\mathcal{E}(X \cdot X^\dagger)Y^\dagger$ *for every* $X, Y > 0$.

(v) There do not exist nontrivial orthogonal projections P, Q *of the same rank such that* $\operatorname{tr}(\mathcal{E}(P)(\mathbb{1} - Q)) = 0$.

(vi) $\mathcal{E}_U$ *is fully indecomposable for all* $U \in U(n)$.

(vii) For any $U \in U(n)$, *if* $(A_i)_i := \mathbf{A}_{\mathcal{E},U}$, *then*

$$\operatorname{rank}\left(\sum_{i \in \mathcal{S}} A_i\right) > |\mathcal{S}| \qquad \forall \mathcal{S} \subset \{1, \ldots, n\}, 0 < |S| < n$$

(viii) $\overline{M}(A_{\mathcal{E},U}) > 0$ *for all* $U \in U(n)$.

Furthermore, when this is satisfied, $\mathcal{E}$ *and via* (v) *also* $\mathcal{E}^*$ *map the open cone* $\mathcal{C}^n$ *into itself. Note also that the properties (vi) to (viii) are also equivalent for any fixed unitary.*

Proof. (i) $\to$ (iii): let $\mathcal{E}$ be fully indecomposable and assume there was a nonzero $A \geq 0$, with $\operatorname{rank}(\mathcal{E}(A)) \leq \operatorname{rank} A$. Since the kernels are vector spaces, this implies we can find a unitary matrix U transforming the basis such that $\ker(\mathcal{E}(A)) \supseteq U \cdot \ker A$. Thus:

$$\begin{aligned} & \ker(\mathcal{E}(A)) \supseteq U \cdot \ker A \\ \Leftrightarrow \quad & \ker(\mathcal{E}(A)) \supseteq \ker UAU^\dagger \\ \Leftrightarrow \quad & \ker(U^\dagger \mathcal{E}(A) U) \supseteq \ker A \end{aligned}$$

The latter implies $\operatorname{supp}(U^\dagger \mathcal{E}(A) U) \subseteq \operatorname{supp} A$. Let P be the projection onto the support of A. By assumption, $P \neq \{0, \mathbb{1}\}$ since A is nonzero and singular. For any positive

matrix B with $BP = B$, we have $\operatorname{supp} B \subseteq \operatorname{supp} A$. Hence, there exists a constant $r > 0$ such that $A \geq rB$. Then $\mathcal{E}(A) \geq r\mathcal{E}(B)$ and $\operatorname{supp}(\mathcal{E}(B)) \subseteq \operatorname{supp}(\mathcal{E}(A))$. But this implies via linearity $\operatorname{supp}(Q\mathcal{E}(\mathcal{M}_n)Q) \subseteq \operatorname{supp}(P\mathcal{M}_n P)$, where $Q := UPU^\dagger$ is an orthogonal projection.
(iii) $\leftrightarrow$ (iv): Given (iii), the claim follows immediately from the fact that since $X, Y > 0$, the matrix ranks are not changed. For any nonzero and singular A we have $\operatorname{rank}(A) = \operatorname{rank}(XAX^\dagger)$. By assumption, for any nonzero, singular $A \geq 0$ we have $\operatorname{rank}(\mathcal{E}(A)) > \operatorname{rank}(A)$, $\operatorname{rank}(\mathcal{E}(XAX^\dagger)) > \operatorname{rank}(XAX^\dagger)$ and hence

$$\operatorname{rank}(Y\mathcal{E}(XAX^\dagger)Y^\dagger) > \operatorname{rank}(A)$$

(iii) $\rightarrow$ (i): Note that given P, Q of the same rank such that $\mathcal{E}(P\mathcal{M}_n P) \subseteq Q\mathcal{M}_n Q$, we have in particular $\mathcal{E}(P) = QAQ$ for some $A \in \mathcal{M}_n$. Since Q is of the same rank as P, $\operatorname{rank}(\mathcal{E}(P)) = \operatorname{rank}(QAQ) \leq \operatorname{rank} P$, which is a contradiction.
(v) $\rightarrow$ (i): Note that if $\mathcal{E}(P\mathcal{M}_n P) \subseteq Q\mathcal{M}_n Q$, then in particular $\operatorname{tr}(\mathcal{E}(P)(\mathbb{1} - Q)) = 0$ since $(\mathbb{1} - Q)$ is the orthogonal complement of Q.
(i) $\rightarrow$ (v): Let $A \geq 0$. By positivity of $\mathcal{E}$, we have

$$\begin{aligned} 0 \leq \operatorname{tr}(\mathcal{E}(PAP)(\mathbb{1} - Q)) &= \operatorname{tr}(AP\mathcal{E}^*(\mathbb{1} - Q)P) \\ &\leq \|A\|_\infty \operatorname{tr}(P\mathbb{1}P\mathcal{E}^*(\mathbb{1} - Q)) = \|A\|_\infty \operatorname{tr}(\mathcal{E}(P)(\mathbb{1} - Q)) = 0 \end{aligned}$$

Hence in particular $\operatorname{supp}(\mathcal{E}(PAP)) \subseteq \operatorname{supp}(Q)$ and $\mathcal{E}(P\mathcal{M}_n P) \subseteq Q\mathcal{M}_n Q$.
(i) $\leftrightarrow$ (ii): This equivalence follows directly from (iv) by expressing Q and P in terms of the projections onto the orthogonal complements.

The remaining equivalences (i) $\leftrightarrow$ (vi),(vii),(viii) can be found in Gurvits 2004 (with proofs scattered throughout the earlier papers by the same author). We just repeat them here using our notation for the reader's convenience.

(iii) $\rightarrow$ (vi): By definition, $\mathcal{E}_U = \mathcal{E} \circ \mathcal{U}$ where $\mathcal{U}(X) = \sum_i \operatorname{tr}(Xu_i u_i^\dagger)$. Obviously, $\mathcal{U}$ is doubly stochastic, hence rank non-decreasing. Therefore, if (iii) holds for $\mathcal{E}$, it must also hold for $\mathcal{E} \circ \mathcal{U}$.

(vi) $\rightarrow$ (iii): Let $\mathcal{E}_U$ be fully indecomposable for all unitaries U and assume that $\operatorname{rank}(\mathcal{E}(X)) \leq \operatorname{rank}(X)$ for some $X \geq 0$ with $0 < \operatorname{rank}(X) < n$. Let U be the unitary that diagonalizes X, then $\mathcal{E}_U(X) = \mathcal{E}(X)$, hence $\mathcal{E}_U$ is not fully indecomposable. This is a contradiction.

(vi) $\Leftrightarrow$ (vii): Let $\mathcal{T}_U$ be fully indecomposable. Then

$$\begin{aligned} \operatorname{rank}(X) &< \operatorname{rank}(\mathcal{E}_U(X)) \\ &= \operatorname{rank}\left(\sum_{i=1}^n \mathcal{E}(u_i u_i^\dagger) \operatorname{tr}(Xu_i u_i^\dagger)\right) \\ &= \operatorname{rank}\left(\sum_{\substack{1 \leq i \leq n \\ \operatorname{tr}(Xu_i u_i^\dagger) \neq 0}} \mathcal{E}(u_i u_i^\dagger)\right) \end{aligned}$$

Note that if $S := \{i | \operatorname{tr}(X u_i u_i^\dagger)\}$, then $\operatorname{rank}(X) \leq |S|$ and hence follows the claim. For the other direction, we can use the same idea.

(vii) $\leftrightarrow$ (viii): Let $A := \mathbf{A}_{\mathcal{E},U}$ for all unitary U fulfill (vii). Define A^{ij} as the tuple where the i-th element is replaced by the j-th. Note that

$$\operatorname{rank}\left(\sum_{k\in\mathcal{S}} A_k^{ij}\right) \geq \operatorname{rank}\left(\sum_{k\in\mathcal{S}\setminus\{j\}} A_k\right) \geq |\mathcal{S}|$$

for any $\mathcal{S} \subset \{1,\ldots,n\}$, where the last inequality follows from the fact that $\mathcal{E}$ is fully indecomposable by assumption. Hence, from the proposition C.4 we know that the mixed discriminant of A^{ij} cannot vanish, i.e. $M(A^{ij}) > 0$. Minimizing over $i \neq j$ and the compact $U(n)$ gives $\tilde{M}(\mathbf{A}_{\mathcal{E},U}) > 0$.

Conversely, let $A := \mathbf{A}_{\mathcal{E},U}$ not fulfill (vii) for some unitary U, i.e. for some $\mathcal{S} \subset \{1,\ldots,n\}$ with $0 < \mathcal{S} < n$ we have $\operatorname{rank}\left(\sum_{k\in\mathcal{S}} A_k\right) \leq |\mathcal{S}|$. Let $i \in \mathcal{S}, j \notin \mathcal{S}$, then for the tuple $A^{(ij)}$ as before, we have:

$$\operatorname{rank}\left(\sum_{k\in\mathcal{S}\cup\{j\}} A_k^{ij}\right) = \operatorname{rank}\left(\sum_{k\in\mathcal{S}} A_k\right) < |\mathcal{S}| + 1 = |\mathcal{S}\cup\{j\}|$$

But then, by proposition C.4, $M(A^{ij}) = 0$ and hence also $\overline{M} = 0$. □

This proposition shows in particular that any fully indecomposable map is *primitive*: For any unit trace $\rho \geq 0$, $\mathcal{E}^d(\rho) > 0$. Note that the converse might not be true. By the characterization of primitive maps (Sanz et al. 2010, Theorem 6.7), this implies that each fully indecomposable map has only one fixed point.

Lemma C.7. *If $\mathcal{T}$ is a doubly-stochastic positive linear map, then there exists a unitary matrix U such that $U\mathcal{T}(\cdot)U^\dagger$ admits a set of orthogonal projections $\{P_i\}_i$ such that $\sum_i P_i = \mathbb{1}$, $P_iP_j = \delta_{ij}P_i$ and $U\mathcal{T}(P_i\mathcal{M}_dP_i)U^\dagger \subseteq P_i\mathcal{M}_dP_i$. Furthermore, the restriction of $U\mathcal{T}(\cdot)U^\dagger$ to $P_i\mathcal{M}_dP_i$ is fully indecomposable for every i.*

Proof. Note that for an arbitrary unitary $U > 0$ the maps $\mathcal{T}(U(\cdot)U^\dagger)$ and $U\mathcal{T}(\cdot)U^\dagger$ are still doubly-stochastic. Let P, Q be a nontrivial Hermitian projector decomposing $\mathcal{T}$, i.e. $\operatorname{tr}(\mathcal{T}(P)(\mathbb{1}-Q)) = 0$ by proposition C.6 (if no such projector exists, we are finished). Then we have:

$$\begin{aligned} 0 &= \operatorname{tr}(\mathcal{T}(P)(\mathbb{1}-Q)) = \operatorname{tr}(P\mathcal{T}^*(\mathbb{1}-Q)) \\ &= \operatorname{tr}(P) - \operatorname{tr}(P\mathcal{T}^*(Q)) = \operatorname{tr}(Q) - \operatorname{tr}(Q\mathcal{T}(P)) \\ &= \operatorname{tr}(Q\mathcal{T}(\mathbb{1}-P)) \end{aligned}$$

where we used that $\mathcal{T}$ is doubly-stochastic in the second and last equality and in between we only used the cyclicity and linearity of the trace as well as the fact that P

and Q have equal rank and thus their traces equal. This means that if P reduces $\mathcal{T}$, then also $\mathbb{1} - Q$ reduces $\mathcal{T}$, i.e.

$$\begin{aligned} & \mathcal{T}(P\mathcal{M}_n P) \subset Q\mathcal{M}_n Q \\ \Rightarrow \quad & \mathcal{T}((\mathbb{1}-P)\mathcal{M}_n(\mathbb{1}-P)) \subset (\mathbb{1}-Q)\mathcal{M}_n(\mathbb{1}-Q) \end{aligned}$$

Since the two projections P, Q are of the same rank, there exists a unitary matrix U such that $Q = UPU^\dagger$. This implies that $\mathcal{T}'(\cdot) = U\mathcal{T}(\cdot)U^\dagger$ is reducible by P and $(\mathbb{1} - P)$, which implies that $\mathcal{T}'$ is a direct sum of maps defined on $P\mathcal{M}_n P$ and $(\mathbb{1} - P)\mathcal{M}_n(\mathbb{1} - P)$.

We obtain these maps by setting

$$\begin{aligned} \mathcal{T}_1' &:= \mathcal{T}'(P \cdot P)|_{P\mathcal{M}_n P} \\ \mathcal{T}_2' &:= \mathcal{T}((\mathbb{1}-P)\cdot(\mathbb{1}-P))|_{(\mathbb{1}-P)\mathcal{M}_n(\mathbb{1}-P)}. \end{aligned}$$

By construction, $P, (\mathbb{1} - P)$ are the identities on the respective subspaces and the maps are therefore doubly stochastic, i.e. $\mathcal{T}_1'(\mathbb{1}_{P\mathcal{M}_n P}) = \mathcal{T}(P) = P = \mathbb{1}_{P\mathcal{M}_n P}$ (and for $\mathcal{T}_2'$ equivalently).

If the restricted maps are not fully indecomposable, we can iterate the procedure, thereby going over to $\mathcal{T}''(\cdot) = (U_1 \oplus U_2)\mathcal{T}'(\cdot)(U_1 \oplus U_2)$ and so forth, which will terminate after finitely many steps, since the ranks of the projections involved have to decrease, thus giving a map $\tilde{\mathcal{T}}(\cdot) = \tilde{U}\mathcal{T}(\cdot)\tilde{U}^\dagger$, which admits the stated decomposition. □

D. Gurvits' proof of scaling and approximate scaling

This appendix provides the details of Gurvits' approach, hence it does not contain original material. For easier readability, we repeat all Lemmata.

D.1. Approximate scalability

We need a way to study scalability:

Definition D.1. Let $C_1, C_2 \in \mathcal{M}_n$ and $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ a positive, linear map. Then we define a *locally scalable functional* to be a map $\varphi \in \overline{\mathcal{C}^d}^*$ such that

$$\varphi(C_1\mathcal{E}(C_2^\dagger \cdot C_2)C_1^\dagger) = \det(C_1C_1^\dagger)\det(C_2C_2^\dagger)\varphi(\mathcal{E}) \tag{73}$$

A locally scalable functional will be called *bounded*, if $|\varphi(\mathcal{E})| \leq f(\mathrm{tr}(\mathcal{E}(\mathbb{1})))$ for some function f.

Locally bounded functionals are the right tools to study scalability:

Proposition D.2. *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. Given a bounded locally scalable functional φ such that $\varphi(\mathcal{E}) \neq 0$, the Sinkhorn-iteration procedure converges:*

$$\mathrm{DS}(\mathcal{E}_n) \to 0 \quad n \to \infty \tag{74}$$

Proof. We follow Gurvits 2004. Recall the definitions of the Sinkhorn iteration in equations (41)-(42). Because of property (73), we have

$$\varphi(\mathcal{E}_{i+1}) = a(i)\varphi(\mathcal{E}_i)$$
$$a(i) = \begin{cases} \det(\mathcal{E}_i^*(\mathbb{1}))^{-1} & \text{if } i \text{ odd} \\ \det(\mathcal{E}_i(\mathbb{1}))^{-1} & \text{if } i \text{ even} \end{cases}$$

Let i be even. Note that $\mathcal{E}_i$ is trace-preserving for i even, hence $\mathrm{tr}(\mathcal{E}_i(\mathbb{1})) = n$. Let $s_j^{(i)}$ be the singular values of $\mathcal{E}_i(\mathbb{1})$ and observe:

$$|\det(\mathcal{E}_i(\mathbb{1}))| = \prod_{j=1}^{n} s_j^{(i)} \leq \frac{1}{n}\sum_{j=1}^{n} s_j^{(i)} = \frac{1}{n}\mathrm{tr}(\mathcal{E}_i(\mathbb{1})) = 1 \tag{75}$$

using the arithmetic-geometric mean inequality (AGM). Similarly, for i odd, $\mathcal{E}_i$ is unital, hence $\mathrm{tr}(\mathcal{E}_i^*(\mathbb{1})) = n$ and we can use the AGM inequality again to obtain that $a(i) \geq 0$ for all $i \geq 0$ and therefore

$$|\varphi(\mathcal{E}_{i+1})| \geq |\varphi(\mathcal{E}_i)|$$

and thus, as φ was assumed to be bounded, $|\varphi(\mathcal{E}_i)|$ converges to some value $c \leq f(\mathrm{tr}(\mathcal{E}(\mathbb{1})))$.

It remains to prove that for $|\varphi(\mathcal{E})| \neq 0$, $\mathrm{DS}(\mathcal{E}_i)$ converges to zero for $i \to \infty$. The idea is of course that if $|\varphi(\mathcal{E})| \neq 0$, then $|a(i)|$ converges to one and thus $\mathcal{E}_i(\mathbb{1})$ and $\mathcal{E}_i^*(\mathbb{1})$ converge to $\mathbb{1}$.

To make this more formal, since $|\varphi(\mathcal{E})|$ converges, for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $d \geq N$:

$$\begin{aligned} & ||\varphi(\mathcal{E}_d)| - |\varphi(\mathcal{E}_{d+1})|| \leq \varepsilon \\ \Leftrightarrow \quad & \left||\varphi(\mathcal{E}_d)| - \frac{1}{|a(i)|}|\varphi(\mathcal{E}_d)|\right| \leq \varepsilon \\ & \Leftrightarrow \quad |a_i| \geq \frac{1}{1+\varepsilon|\varphi(\mathcal{E}_d)|^{-1}} \geq \frac{1}{1+\varepsilon|\varphi(\mathcal{E})|^{-1}} \end{aligned}$$

where we used that $|\varphi(\mathcal{E}_d)|$ increases monotonically in the last inequality.

Let us now only consider i even. Then we have just seen that

$$\frac{1}{1+\varepsilon|\varphi(\mathcal{E})|^{-1}} \leq \det(T_i(\mathbb{1})) \leq 1$$

hence, for $i \geq N$ even, we have:

$$\mathrm{DS}(\mathcal{E}_i) = \mathrm{tr}((\mathcal{E}_i(\mathbb{1}) - \mathbb{1})^2) = \sum_{j=1}^{n}(s_j^{(i)} - 1)^2$$

where the $s_j^{(i)}$ are the singular values of $\mathcal{E}_i(\mathbb{1})$. If we can upper bound the last quantity by $\tilde{\varepsilon}(\varepsilon)$, we are done. This is an exercise in using logarithms:

Since $\mathcal{E}_i$ is trace-preserving as i is even, $s_j^{(i)} \leq d$ for all i. If we set $\alpha := \frac{(n-1)-\ln(n)}{(n-1)^2}$, then by strict concavity of the logarithm,

$$\ln(x) \leq (x-1) - \alpha(x-1)^2 \qquad x \leq d$$

since $\ln(d) = (x-1) - \alpha(x-1)^2$ and $\ln(1) = 0$. But then:

$$\begin{aligned} 0 \leq \sum_{j=1}^{n} (s_j^{(i)} - 1)^2 &\leq \sum_{j=1}^{n} \left(\frac{s_j^{(i)} - 1}{\alpha} - \frac{\ln(s_j^{(i)})}{\alpha} \right) \\ &= -\sum_{i=1}^{n} \frac{\ln(s_j^{(i)})}{\alpha} \\ &= -\frac{1}{\alpha} \ln(\prod_{i=1}^{n} s_j^{(i)}) \leq -\frac{1}{\alpha} \ln(1-\varepsilon) \\ &\leq \frac{\varepsilon}{\alpha} \end{aligned}$$

where we used that $\sum_{j=1}^{n} s_j^{(i)} = \mathrm{tr}(\mathcal{E}_i(\mathbb{1})) = n$. But $\frac{\varepsilon}{\alpha} \to 0$ for $i \to \infty$.

Exchanging $\mathcal{E}_i$ with $\mathcal{E}_i^*$ gives the same reasoning for odd i. In total, we get that for any $\varepsilon > 0$ exists an $N \in \mathbb{N}$ such that for all $d \geq n$

$$0 \leq \mathrm{DS}(\mathcal{E}_d) \leq \frac{\varepsilon}{\alpha}$$

hence $\mathcal{DS}(\mathcal{E}_i) \to 0$ for $i \to \infty$. □

Lemma D.3. Cap *is a bounded locally scalable functional.*

Proof. Note that for

$$\begin{aligned} &\inf\{\det(C_2^\dagger \mathcal{E}(C_1 X C_1^\dagger) C_2) | X > 0, \det(X) = 1\} \\ &\quad = \inf\{\det(C_2^\dagger) \det(C_2) \det(\mathcal{E}(C_1 X C_1^\dagger)) | X > 0, \det(X) = 1\} \\ &\quad = \det(C_2^\dagger C_2) \inf\{\det(\mathcal{E}(C_1 X C_1^\dagger)) | X > 0, \det(X) = 1\} \\ &\quad = \det(C_2^\dagger C_2) \inf\{\det(\mathcal{E}(\tilde{X})) | X > 0, \det(\tilde{X}) = \det(C_1) \det(C_1^\dagger) \det(X), \det(X) = 1\} \\ &\quad = \det(C_2^\dagger C_2) \det(C_1^\dagger C_1) \inf\{\det(\mathcal{E}(\tilde{X})) | \tilde{X} > 0, \det(\tilde{X}) = 1\} \end{aligned}$$

hence Cap is a locally scalable functional. Via the AGM inequality, we have

$$0 \leq \mathrm{Cap}(\mathcal{E}) \leq \det(\mathcal{E}(\mathbb{1})) \leq \left(\frac{\mathrm{tr}(\mathcal{E}(\mathbb{1}))}{n} \right)^{\frac{1}{n}}$$

hence Cap is bounded. □

This gives a proof of Lemma 9.8.

Lemma D.4 (Lemma 9.9 of the main text). *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map and $U \in U(n)$ a fixed unitary. Then defining*

$$\mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) := \inf\left\{ \det\left(\sum_i \mathcal{E}(u_i u_i^\dagger)\gamma_i \right) | \gamma_i > 0, \prod_{i=1}^n \gamma_i = 1 \right\}$$

where u_i are once again the rows of U, we have the following properties:

1. *Using the mixed discriminant M, we have*

$$M(\mathbf{A}_{\mathcal{E},U}) \leq \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) \leq \frac{n^n}{n!} M(\mathbf{A}_{\mathcal{E},U})$$

2. $\inf_{U \in U(n)} \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) = \mathrm{Cap}(\mathcal{E})$

Proof. The first part of the lemma is one of the main results of Gurvits and Samorodnitsky 2002. Since the proof is long due to many technicalities, we leave it out here.

The second part gives the relation between the two capacities. Let $\{X_d\}_d$ with $\det(X_d) = 1$ and $X_d > 0$ be such that $\det(\mathcal{E}(X_d)) \to \mathrm{Cap}(\mathcal{E})$, $d \to \infty$. Then there exist unitaries $U_d \in U(n)$ such that $U_d X_n U_d^\dagger$ is diagonal with diagonal entries $\lambda_i^{(d)}$. By construction,

$$\det\Big(\sum_{1 \leq i \leq n} \mathcal{E}((u_d)_i (u_d)_i^\dagger)\lambda_i^{(d)} \Big) = \det(\mathcal{E}(X_d))$$

where (u_d) are again the columns of U_d. Hence

$$\inf_{U \in U(n)} \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) \leq \mathrm{Cap}(\mathcal{E})$$

Likewise, we can construct a sequence of U_d such that $\mathrm{Cap}(\mathbf{A}_{\mathcal{E},U_d})$ converges to the infimum and we can construct a sequence $(\gamma_{(k)}^{(d)})_k$ with $(\gamma_{(k)}^{(d)})_i > 0$ for each U_d such that

$$\det\left(\sum_{i=1}^n \mathcal{E}((u_d)_i (u_d)_i^\dagger)(\gamma_{(k)}^{(d)})_i \right) \to \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U_d}) \qquad \text{for } k \to \infty$$

Taking the diagonal sequence $\gamma_{(d)}^{(d)}$ we obtain a sequence converging to $\inf \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U})$. Finally, define $X_k = U_k \operatorname{diag}(\gamma_{(k)}^{(k)}) U_k^\dagger$, then $X_k > 0$ and

$$\det(\mathcal{E}(X_k)) = \det\left(\sum_{i=1}^n \mathcal{E}((u_d)_i (u_d)_i^\dagger)(\gamma_{(k)}^{(d)})_i \right)$$

and hence

$$\mathrm{Cap}(\mathcal{E}) \leq \inf_{U \in U(n)} \mathrm{Cap}(\mathbf{A}_{\mathcal{E},U})$$

after taking the limit $k \to \infty$. □

Finally, we can write down the Operator Sinkhorn theorem (Theorem 9.5 in the main text):

Theorem D.5 (Approximate Operator Sinkhorn Theorem, Gurvits 2004 Theorem 4.6)**.** *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. Then $\mathcal{E}$ is ε-scalable for all $\varepsilon > 0$ iff $\mathcal{E}$ is rank non-decreasing.*

Proof. We mostly need to combine the lemmas. By lemma D.3, the capacity is a bounded, locally scalable functional, which implies by proposition D.2 that $\mathrm{DS}(\mathcal{E}_i)$ converges, if $\mathrm{Cap}(\mathcal{E}) > 0$. Now, by lemma 9.9,

$$\mathrm{Cap}(\mathcal{E}) = \inf\{\mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) | U \in U(n)\}$$

Since $U(n)$ is compact, it suffices to show that for every U, $\mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) > 0$. Again, by lemma 9.9,

$$\mathrm{Cap}(\mathbf{A}_{\mathcal{E},U}) \geq M(\mathbf{A}_{\mathcal{E},U})$$

but $M(\mathbf{A}_{\mathcal{E},U}) > 0$ for every U if and only if $\mathcal{E}$ is rank non-decreasing by proposition C.4. Hence, $\mathrm{DS}(\mathcal{E}_i)$ converges for rank non-decreasing maps.

Now suppose that $\mathcal{E}$ is a positive map such that in the Sinkhorn iteration, $\mathrm{DS}(\mathcal{E}_i)$ converges. Then, for some $i \in \mathbb{N}$, $\mathrm{DS}(\mathcal{E}_i) < \frac{1}{n}$. We claim that then $\mathcal{E}_i$ is rank non-decreasing and by consequence, also $\mathcal{E}$ is rank non-decreasing via proposition C.4.

To see this, assume $\mathcal{E}(\mathbb{1}) = \mathbb{1}$ and $\mathcal{E}^*(\mathbb{1}) = \mathbb{1} + E$, where E is Hermitian and $\mathrm{tr}(E^2) \leq 1/n$. We can do this, because this is exactly what $\mathcal{E}_i$ looks like for i big enough such that $\mathrm{DS}(\mathcal{E}_i) < \frac{1}{n}$ and i is odd. Given a matrix $U \in U(n)$ and the corresponding $A := \mathbf{A}_{\mathcal{E},U}$, we have that

$$\sum_{i=1}^{n} A_i = \mathcal{E}(\mathbb{1}) = \mathbb{1}$$

Likewise, for every i:

$$\mathrm{tr}(A_i) = \mathrm{tr}(A_i \mathbb{1}) = \mathrm{tr}(u_i u_i^\dagger T^*(\mathbb{1})) = 1 + \mathrm{tr}(u_i u_i^\dagger E) =: 1 + \delta_i \tag{76}$$

But by assumption,

$$\begin{aligned} \sum_{i=1}^{n} |\delta_i|^2 &\leq \sum_{i,j=1}^{n} |\,\mathrm{tr}(u_i u_j^\dagger E)|^2 \\ &= \sum_{i,j=1}^{n} \langle u_i | E | u_j \rangle \langle u_j | E^\dagger | u_i \rangle \\ &= \mathrm{tr}(E^2) \leq \frac{1}{n} \end{aligned} \tag{77}$$

Now, suppose that $\mathcal{E}$ is not rank non-decreasing. Then, by proposition A.5 (vii), there is a U such that $\mathbf{A}_{\mathcal{E},U}$ fulfills

$$\operatorname{rank}\left(\sum_{i=1}^{k} A_i\right) < k$$

for some $0 < k < n$. Note that, since $\mathcal{E}$ is positive, $A_i \geq 0$, hence $H := \sum_{i=1}^{k} A_i$ fulfills $0 \leq H \leq \mathbb{1}$. As $\operatorname{rank}(H) \leq k-1$, we have $\operatorname{tr}(H) \leq k-1$. From equation (76), we obtain

$$\operatorname{tr}(H) = \sum_{i=1}^{k} A_i = k + \sum_{i=1} k\delta_i$$

Using equation (77), by the Cauchy Schwarz inequality,

$$\sum_{i=1}^{k} |\delta_i| \leq \sqrt{k/n} < 1$$

which contradicts $\operatorname{tr}(H) \leq k-1$, hence $\mathcal{E}$ must be rank non-decreasing. □

D.2. Exact scalability

Lemma D.6 (Lemma 9.10 of the main text)**.** *Let $\mathcal{E} : \mathcal{M}_d \to \mathcal{M}_d$ be a positive map. Then $\mathcal{E}$ is scalable to a doubly-stochastic map if and only if $\operatorname{Cap}(\mathcal{E}) > 0$ and the capacity can be achieved.*

Proof. Suppose there exists $C > 0$ with $\det(\mathcal{E}(C)) = \operatorname{Cap}(\mathcal{E})$. The Lagrangian of the capacity is

$$\mathcal{L}(X) := \ln(\det(\mathcal{E}(X))) + \lambda \ln(\det(X))$$

with the Lagrangian multiplier $\lambda \in \mathbb{R}$. Therefore, the minimum fulfills

$$\nabla \ln(\det(\mathcal{E}(X)))|_{X=C} = (-\lambda)\nabla \ln(\det(X))|_{X=C} \tag{78}$$

We claim that the conditions are equivalent to

$$\mathcal{E}^*((\mathcal{E}(C))^{-1})^{-1} = C^{-1} \tag{79}$$

Let E_{ij} be the usual matrix unit, then

$$\begin{aligned}(\nabla \ln(\det(\mathcal{E}(X)))|_{X=C})_{jk} &= \frac{\partial}{\partial E_{jk}} \ln\left(\sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} \mathcal{E}(C)_{i\sigma(i)}\right) \\ &= \frac{1}{\det(\mathcal{E}(C))} \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) \frac{\partial}{\partial E_{jk}} \prod_{i=1}^{n} \mathcal{E}(C)_{i\sigma(i)}.\end{aligned}$$

Noting that

$$\partial E_{jk}\mathcal{E}(C)_{i\sigma(i)} = \mathrm{tr}(E_{\sigma(i)i}\frac{\partial \mathcal{E}(C)}{\partial E_{jk}}) = \mathrm{tr}(\mathcal{E}^*(E_{\sigma(i)i})E_{jk}) = \mathcal{E}^*(E_{\sigma(i)i})_{jk}$$

we have

$$\begin{aligned}
(\nabla \ln(\det(\mathcal{E}(X)))|_{X=C})_{jk} &= \frac{1}{\det(\mathcal{E}(C))}\sum_{\sigma\in S_n}\mathrm{sgn}(\sigma)\sum_{l=1}^{n}\mathcal{E}^*(E_{\sigma(l)l})_{jk}\prod_{i\neq l}\mathcal{E}(C)_{i\sigma(i)} \\
&= \mathcal{E}^*\left(\frac{1}{\det(\mathcal{E}(C))}\sum_{\sigma\in S_n}\mathrm{sgn}(\sigma)\sum_{l=1}^{n}E_{\sigma(l)l}\prod_{i\neq l}\mathcal{E}(C)_{i\sigma(i)}\right)_{jk} \\
&= \mathcal{E}^*\left(\frac{1}{\det(\mathcal{E}(C))}\sum_{m,n=1}^{n}\left(\sum_{\sigma(m)=n\in S_n}\mathrm{sgn}(\sigma)(-1)^{n-m}\prod_{i\neq m}\mathcal{E}(C)_{i\sigma(i)}\right)E_{mn}\right)_{jk} \\
&= \mathcal{E}^*(\mathcal{E}(C)^{-1})_{jk}
\end{aligned}$$

where in the last step we use Cramer's rule. For $\mathcal{E} = \mathrm{id}$ we obtain the right hand side of equation (79) from equation (78), hence follows the claim. It now follows from Lemma 9.14 that any C fulfilling Equation (79) defines a scaling.

Conversely, suppose $\tilde{\mathcal{E}}(\cdot) = C_1\mathcal{E}(C_2^\dagger C_2)C_1^\dagger$ is a doubly stochastic map. Since $\tilde{\mathcal{E}}$ is unital, the eigenvalues of $\tilde{\mathcal{E}}(X)$ are majorized by the eigenvalues of X (cf. Wolf 2012 Theorem 8.8). Majorization stays invariant under strictly increasing functions (cf. Bhatia 1996, Chapter 1), hence we have (λ_i being the eigenvalues of X and $\lambda_i^{\tilde{\mathcal{E}}}$ the eigenvalues of $\tilde{\mathcal{E}}(X)$):

$$\sum_i -\ln(\lambda_i^{\tilde{\mathcal{E}}}) \leq \sum_i -\ln(\lambda_i)$$

which is equivalent to $\det(\tilde{\mathcal{E}}(X)) \geq \det(X)$. Hence, a doubly stochastic map is in particular determinant increasing. Obviously, equality is attained at $X = \mathbb{1}$. But then:

$$\begin{aligned}
\det(\mathcal{E}(X)) &= |\det(C_1)|^{-2}|\det(C_2)|^{-2}\det(\tilde{\mathcal{E}}(X)) &(80)\\
&\geq |\det(C_1)|^{-2}|\det(C_2)|^{-2}\det(X). &(81)
\end{aligned}$$

A quick calculation shows that $X = C_2^\dagger C_2/\det(C_2^\dagger C_2)^{1/n}$ attains equality in Equation (81). This then necessarily minimises the capacity. □

Lemma D.7 (Lemma 9.11 of the main text)**.** *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map and given $U \in U(n)$, let $A = \mathbf{A}_{\mathcal{E},U}$. Then*

1. *f_A is convex on $\mathbb{R}^n$.*
2. *If $\mathcal{E}$ is fully indecomposable, then f_A is strictly convex on $\{\xi = (\xi_1,\ldots,\xi_n) \in \mathbb{R}^n | \sum_i \xi_i = 0\}$.*

Proof. We follow the proof of Gurvits and Samorodnitsky 2002. Given a tuple A of positive definite matrices, one can show (Bapat 1989):

$$\det(e^{\xi_1}A_1 + \ldots + e^{\xi_n}A_n) = \sum_{r\in P_n} t_r e^{(\xi,r)} \tag{82}$$

where $(\cdot,\cdot)$ denotes the general inner product, P_n is the set of n-tuples of integers $r_i \geq 0$ such that $\sum_i r_i = n$ and

$$t_r := \frac{1}{r_1!\ldots r_n!} M(\overbrace{A_1,\ldots,A_1}^{r_1},\ldots,\overbrace{A_n,\ldots,A_n}^{r_n}) \tag{83}$$

This implies that we can rewrite f_A:

$$f_A(\xi_1,\ldots,\xi_n) = \ln\det(e^{\xi_1}A_1 + \ldots + e^{\xi_n}) = \ln\left(\sum_{r\in P_n} t_r e^{(\xi,r)}\right)$$

It is well known that for positive matrices this is a convex function, but let us follow the proof of Gurvits and Samorodnitsky 2002 here.

Let $f =: \ln g$. We need to prove that $\nabla^2 f$, the Hessian, is positive (semi)definite. By definition, $\nabla^2 f = \frac{1}{g^2}(g(\nabla^2 g) - (\nabla g)(\nabla g)^{tr})$, hence it is sufficient that $g(\nabla^2 g) \geq (\nabla g)(\nabla g)^{tr}$.

Note that for any $v \in \mathbb{R}^n$ we have $\nabla e^{(\xi,v)} = e^{(\xi,v)}\cdot v$ and $\nabla^2 e^{(\xi,v)} = e^{(\xi,v)}vv^{tr}$, where vv^{tr} is positive definite. Therefore:

$$\begin{aligned} g(\nabla^2 g) - (\nabla g)(\nabla g)^{tr} &= \sum_{r\in P_n} t_r e^{(\xi,r)} \cdot \sum_{s\in P_n} t_r e^{(\xi,s)} ss^{tr} - \sum_{r,s\in P_n} t_r t_s e^{(\xi,r+s)} rs^{tr} \\ &= \frac{1}{2}\sum_{r,s\in P_n} t_r t_s e^{(\xi,r+s)}(r-s)(r-s)^{tr} \geq 0 \end{aligned}$$

hence the Hessian of f is positive semi-definite and therefore f is convex.

Now, assume that $\mathcal{E}$ is fully indecomposable, hence the tuple $A := \mathbf{A}_{\mathcal{E},U}$ fulfills proposition A.5 (vii) and (viii) for all $U \in U(n)$. In particular, if A^{ij} is the tuple A with the j-th entry being replaced by the i-th. entry. Then $M(A^{ij}) > 0$ in particular. Note that $M(A^{ij}) = 2t_{r_{ij}}$ by equation (83), where

$$(r_{ij})_k := \begin{cases} 2 & k = i \\ 0 & k = j \\ 1 & \\ \text{else} & \end{cases}.$$

Then,

$$\nabla^2 f \geq \frac{1}{g^2}\sum_{r,s\in P_n} t_r t_s e^{(\xi,r+s)}(r-s)(r-s)^{tr}$$

$$\geq \frac{1}{8g^2} \sum_{i\neq j,k\neq l} M(A^{ij})M(A^{kl})e^{(\xi,r_{ij}+s_{kl})}(r_{ij}-r_{kl})(r_{ij}-r_{kl})^{tr}$$
$$\geq \frac{cM^2}{8g^2} \sum_{i\neq j,k\neq l} (r_{ij}-r_{kl})(r_{ij}-r_{kl})^{tr}$$

where $c := \min_{i\neq j\neq k\neq l} e^{(\xi,r_{ij}-r_{kl})}$ and $M := \min_{i\neq j} M(A^{ij}) > 0$ by proposition A.5 (viii).

We only need to consider $\sum_{i\neq j,k\neq l}(r_{ij}-r_{kl})(r_{ij}-r_{kl})^{tr} =: S$ and show that this is a positive definite matrix on the hyperplane H. Using the usual matrix units E_{mn} we can write:

$$S := \sum_{i\neq j\neq k\neq l} (E_{ii}+E_{jj}+E_{kk}+E_{ll}+2(E_{il}+E_{jk}-E_{ik}-E_{jl}-E_{kl}))$$

We find that $S_{ii} = (n-1)(n-2)(n-3)$, since only the first four summands contribute to the diagonal terms. For the off-diagonal terms, note that in $2(E_{il}+E_{jk}-E_{ik}-E_{jl}-E_{kl})$, all unordererd combinations of i,j,k,l occur, twice with a positive sign and four times with a negative. Hence we obtain $(n-2)\cdot(n-3)$ terms with either E_{ij} or E_{ji} that are not cancelled by other terms and therefore $S_{ij} = -(n-2)(n-3)$. In short:

$$S = (n-1)(n-2)(n-3)\begin{pmatrix} 1 & \frac{1}{n-1} & \cdots & \frac{1}{n-1} \\ \frac{1}{n-1} & 1 & \cdots & \frac{1}{n-1} \\ \vdots & & \ddots & \vdots \\ \frac{1}{n-1} & \frac{1}{n-1} & \cdots & 1 \end{pmatrix}$$

Note that the image of S is just the hyperplane H and it is easy to see that S is a multiple of the projection onto the hyperplane S. Therefore, $\nabla^2 f$ is strictly convex on H. □

Finally, we obtain the theorem:

Lemma D.8 (Lemma 9.12 of the main text)**.** *Let $\mathcal{E} : \mathcal{M}_n \to \mathcal{M}_n$ be a positive, linear map. If $\mathcal{E}$ is fully indecomposable, there exists a unique scaling of $\mathcal{E}$ to a doubly stochastic map.*

Proof. Recall that one can show (Bapat 1989):

$$\det(e^{\xi_1}A_1 + \ldots + e^{\xi_n}A_n) = \sum_{r\in P_n} t_r e^{(\xi,r)} \tag{84}$$

where $(\cdot,\cdot)$ denotes the general inner product, P_n is the set of n-tuples of integers $r_i \geq 0$ such that $\sum_i r_i = n$ and $t_r := \frac{1}{r_1!\ldots r_n!} M(\overbrace{A_1,\ldots,A_1}^{r_1},\ldots,\overbrace{A_n,\ldots,A_n}^{r_n})$.

Suppose $X \geq 0$, $\det(X) = 1$ and $\mathcal{E}$ is fully indecomposable. Let $U \in U(n)$ diagonalize X with eigenvalues $\gamma_i = e^{\xi_i}$. Assume the eigenvalues are ordered $\gamma_1 \geq \ldots \geq \gamma_n$.

Observe that then $\det(\mathcal{E}(X)) \leq \det(\mathcal{E}(\mathbb{1}))$ is equivalent to say that $f_A(\xi) \leq f_A(0)$, where $A = \mathbf{A}_{\mathcal{E},U}$. We know:

$$\begin{aligned}\det(A_1 + \ldots + A_n) &\geq \det(\gamma_1 A_1 + \ldots + \gamma_n A_n) \\ &= \sum_{r \in P_n} t_r e^{(\xi, r)} \geq \frac{1}{2} \sum_{i \neq j} M_{ij} e^{(\xi, r_{ij})}\end{aligned}$$

where we use that certainly for all $i \neq j \in \{1, \ldots, n\}$, $r_{ij} :=$ with $r_k = 1$ for all $k \neq i, j$ and $r_i = 2$, $r_j = 0$ is a valid n-tuple where the coefficient $t_r = \frac{1}{2} M^{ij}$. Since all the terms in the sum of equation 84 are positive, we can just leave out all other r. By definition, $\overline{M}(\mathcal{E}) \leq M^{ij}$ for every A, hence:

$$\begin{aligned}\det(A_1 + \ldots + A_n) &\geq \frac{1}{2} \overline{M}(\mathcal{E}) \sum_{i \neq j} e^{(\xi, r_{ij})} \\ &\geq \frac{1}{2} \overline{M}(\mathcal{E}) e^{\max_{i \neq j} (\xi_i - \xi_j)} \\ &\geq \frac{1}{2} \overline{M} \frac{\gamma_1}{\gamma_N}\end{aligned}$$

where we used that $(\xi, r_{ij}) = \sum_{k \neq j} \xi_k + \xi_i = \xi_i - \xi_j$ since $\sum_i \xi_i = 0$. Since $\det(A_1 + \ldots + \ldots) = \det(\mathcal{E}(\mathbb{1}))$, we have

$$\frac{\gamma_1}{\gamma_n} \leq \frac{2 \det(T(\mathbb{1}))}{M(A)} \leq \frac{2 \det(\mathcal{E}(\mathbb{1}))}{\overline{M}} < \infty$$

from the lemma above. But then, the infimum must be attained on the compact subset $\{\det(X) = 1 | \gamma_1 \leq \frac{2 \det(\mathcal{E}(X))}{\overline{M}}\}$. Therefore, also for the capacity $\mathrm{Cap}(\mathcal{E})$ the infimum can be considered on a compact subset of $\{\det(X) = 1\}$ and is then attained. Uniqueness is ensured by the strict convexity of f_A. □